Computer Science & Information Technology 130

David C. Wyld,
Natarajan Meghanathan (Eds)

Computer Science & Information Technology

9th International Conference on Natural Language Processing (NLP 2020)
November 21~22, 2020, Zurich, Switzerland

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

ISSN: 2231 - 5403
ISBN: 978-1-925953-28-2
DOI: 10.5121/csit.2020.101401- 10.5121/csit.2020.101426

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The 9th International Conference on Natural Language Processing (NLP 2020) November 21~22, 2020, Zurich, Switzerland, 9th International Conference on Software Engineering and Applications (JSE 2020), International Conference on Machine Learning Techniques (MLTEC 2020), 11th International conference on Database Management Systems (DMS 2020), International Conference on Networks & IOT (NeTIOT 2020), 9th International Conference on Information Technology Convergence and Services (ITCS 2020), 9th International Conference on Signal & Image Processing (SIP 2020), 7th International Conference on Foundations of Computer Science & Technology (CST 2020) and 7th International Conference on Artificial Intelligence & Applications (ARIA 2020) was collocated with 9th International Conference on Natural Language Processing (NLP 2020). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The NLP 2020, JSE 2020, MLTEC 2020, DMS 2020, NeTIOT 2020, ITCS 2020, SIP 2020, CST 2020 and ARIA 2020 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, NLP 2020, JSE 2020, MLTEC 2020, DMS 2020, NeTIOT 2020, ITCS 2020, SIP 2020, CST 2020 and ARIA 2020 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the NLP 2020, JSE 2020, MLTEC 2020, DMS 2020, NeTIOT 2020, ITCS 2020, SIP 2020, CST 2020 and ARIA 2020.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld,
Natarajan Meghanathan (Eds)

General Chair	Organization
David C. Wyld,	Southeastern Louisiana University, USA
Natarajan Meghanatha,	Jackson State University, USA

Program Committee Members

Abd El-Aziz Ahmed,	Cairo University, Egypt
Abdulhamit Subasi,	Effat University, Saudi Arabia
Adeyanju Sosimi,	University of Lagos, Nigeria
Adil BROURI,	Moulay Ismail University, Morocco
Ahmed Farouk AbdelGawad,	Zagazig University, Egypt
Ahmed Korichi,	University of Ouargla, Algeria
Ahmed Lbath,	University of Grenoble, France
Akhil Gupta,	Lovely Professional University, India
Alexander Gelbukh,	Instituto Politécnico Nacional, Mexico
Amal azeroual,	Mohammed V University, Morocco
Ammar A. Aldair,	University of Basrah,Iraq
Ana Luísa Varani Leal,	University of Macau, China
Anand Nayyar,	Duy Tan University, Vietnam
Anis Tissaoui,	University of Jendouba, Tunisia
Anouar Abtoy,	Abdelmalek Essaadi University, Morocco
Antony P J,	A J Institute of Engineering and Technology, India
Anwar Basha H,	SRM Institute of Science and Technology, India
Aravinda c v,	NMAM Institute of Technology, India
Archit Yajnik,	Sikkim Manipal University, India
Arthur,	Universidade Federal de Santa Catarina, Brazil
Assia DJENOUHAT,	University Badji Mokhtar Annaba, Algeria
Athanasios Vasilakos,	Lulea University of Technology, Sweden
Azida Zainol,	University of Jeddah, Saudi Arabia
B. K. Tripathy,	VIT, Vellore, India
Baki Koyuncu,	Istanbul gelisim University, Turkey
Basavaraj S. Anami,	KLE Institute of Technology, India
Benmohammed M,	University of Constantine, Algeria
Bilal H. Abed-alguni,	Yarmouk University, Jordan
Boo-Hyung Lee,	KongJu National University, South Korea
Boulmakoul,	Hassan II University, France
Chengliang Huang,	Ryerson University, Canada
D. P. Acharjya,	VIT, Vellore, India
Dadmehr Rahbari,	The University of Qom, Iran
Daniel Ekpenyong Asuquo,	University of Uyo, Nigeria
David Obdrzalek,	Charles University, Czech Republic
Diab Abuaiadah,	Waikato Institute of Technology, New Zealand
Dibya Mukhopadhyay,	University of Alabama at Birmingham, USA
Dinesh N. Patil,	KITS, India
Diptendu Sinha Roy,	National Institute of Technology, India
Dongping Tian,	Baoji University of Arts and Sciences, China
Dorra Driss,	University of Sfax, Tunisia
El-Sayed M. El-Horbaty,	Ain Shams University, Egypt
Emilio Jimenez Macias,	University of La Rioja, Spain

Faouzia Benabbou,	University Hassan II of Casablanca, Morocco
Farukuzzaman Khan,	Islamic University, Bangladesh
Farzin Piltan,	University of Ulsan, Korea
Fernando Zacarias Flores,	Universidad Autonoma de Puebla, Mexico
Filiz Cele,	Istanbul Aydin UNniversity, Turkey
Gammoudi Aymen,	University of Tunis, Tunisia
Giuseppe Bruno,	Bank of Italy, Italy
H L Shashirekha,	Mangalore University, India
Hala Abukhalaf,	Palestine Polytechnic University, Palestine
Hanan Salam,	University of Pierre and Marie Curie, France
Hang Su,	Politecnico di Milano, Italy
Hao-En Chueh,	Chung Yuan Christian University, Taiwan
Heldon Jose,	Professor of Integrated Faculties of Patos, Brazil
Himani Mittal,	GGDSD college, India
Hossein Jadidoleslamy,	MUT University, Iran
Huaming Wu,	Tianjin University, China
Ibrahim Abu El-Khair,	Minia University, Egypt
Isa Maleki,	Islamic Azad University, Iran
Islam Atef,	Alexandria university, Egypt
Jalel Akaich,	Institut Superieur de Gestion de Tunis, Tunisia
Javid Taheri,	Karlstad University,Sweden
Jay Pavagadhi,	CEO at Olak, Former Apple, USA
Jesuk Ko,	Universidad Mayor de San Andres (UMSA), Bolivia
Jingsha He,	Beijing University of Technology, China
John Tass,	University of Patras, Greece
Jose Alfredo F. Costa,	Federal University, Brazil
Jun Zhang,	South China University of Technology, China
Kadu Machado,	Brasilia University - UnB, Brazil
Kaveh,	Iran University of Science and Technology, Iran
Kayhan Erciyes,	Izmir University, Turkey
Kazuyuki Matsumoto,	Tokushima University, Japan
Ke-Lin Du,	Concordia University, Canada
Kiran Sharma,	Northwestern University, United States of America
Kshira Sagar Sahoo,	Vnrvjiet, Hyderabad, India
Le Nguyen Quoc Khanh,	Taipei Medical University, Taiwan
lfredo Cuzzocrea,	University of Trieste, Italy
Lien-Fa Lin,	Kao Yuan University, Taiwan
lsraa Shaker Tawfic,	Ministry of Science and Technology, Iraq
Luisa Maria Arvide Cambra,	University of Almeria, Spain
Lutz Schubert,	University of Ulm, Germany
M V Ramana Murthy,	Osmania University,India
Mahdi Mazinani,	IAU Shahreqods, Iran
Mahmood Hashemi,	Beijing University of Technology, China
Maissa HAMOUDA,	SETIT & ISITCom, University of Sousse, Tunisia
Manyok Chol David,	University of Juba, South Sudan
Marco Javier Suarez Baron,	Uptc, Colombia
Marco Palomino,	University of Plymouth, UK
María Hallo,	Escuela Politécnica Nacional, Ecuador
Mario Versaci,	DICEAM - University Mediterranea, Italy
Masoud Rashidinejad,	Queens University, Canada
Md Forhad Rabbi,	Curtin University, Australia
Mehdi Gheisari,	Guangzhou University, China

Mehdi Nezhadnaderi,	Islamic Azad University, Iran
Milad Moradi Vastegani,	Medical University of Vienna, Austria
Mohamed Fakir,	university Sultan Moulay Slimane, Morocco
Mohamed Saad AZIZI,	Moulay-Ismail University, Morocco
Mohamed Tounsi,	Prince Sultan University, Saudi Arabia
Mounir ZRIGUI,	Monastir University, Tunisia
Muhammad Atif Bilal,	Jilin University, China
Mu-Song Chen,	University of Texas, Taiwan
Nadia Abd-Alsabour,	Cairo University, Egypt
Nahlah Shatnawi,	Yarmouk University, Jordan
naors y anad alsaleem,	university of al-hamdaniya, iraq
Narinder Singh,	Punjabi University, India
Nasrin Akhter,	University Putra Malaysia, Malaysia
Natheer K Gharaibeh,	Taibah University, Saudi Arabia
Natheer Khlaif Gharaibeh,	Taibah University, Saudi Arabia
Nikola Ivkovic,	University of Zagreb, Croatia
Niloofar rastin,	Shiraz University, Iran
Nizar Aifaoui,	LGM, ENIM, Tunisia
Norisma Idris,	University of Malaya, Malaysia
Omar Al-harbi,	Jazan University University, Saudi Arabia
Osama Rababah,	The University of Jordan, Jordan
Panagiotis Fotaris,	University of Brighton, UK
Paulo Quaresma,	University of Évora, Portugal
Picky Butani,	Shubh Solutions LLC, USA
Piotr Malak,	University of Wroclaw, Poland
Pranita Mahajan,	SIES Graduate School Of Technology, India
Punnoose A K,	Flare Speech Systems, India
R. Ragupathy,	Annamalai University, India
Raed Ibraheem Hamed,	University of Anbar, Iraq
Ramadan Elaiess,	University of Benghazi, Libya
Ramakrishnan Raman,	Higher Colleges of Technology, UAE
Ritu Sharma,	Himachal Pradesh University Shimla, India
Rohola Zandie,	University of Denver, USA
Salem Nasri,	National School of Engineering, Tunisia
Sandeep Chaurasia,	Manipal University, India
Sarfraz,	Kuwait University, Kuwait
Sasikumar Gurumurthy,	Vit University, India
Satish Gajawada,	IIT Roorkee, India
Seppo Sirkemaa,	University of Turku, Finland
Shahzad Ashraf,	Hohai University, China
Sharathyh Kumar,	Mit Mysore, India
Shinde Pravin,	Mumbai University, India
SivaKumar PV,	VNR VJIET, India
Smain Femmam,	UHA University, France
Somdip Dey,	University of Essex, UK
Sridharan D,	Anna University, India
Stefano Michieletto,	University of Padova, Italy
Sunil Vadera,	University of Salford, UK
Susmita Gupta,	Indian Institute of Technology, India
Swati Nikam,	Savitribai Phule Pune University, India
Tanik Saikh,	Indian Institute of Technology Patna, India
Tanzila Saba,	Prince Sultan University, Saudi Arabia

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

9th International Conference on Natural Language Processing (NLP 2020)

9th International Conference on Software Engineering and Applications (JSE 2020)

International Conference on Machine Learning Techniques (MLTEC 2020)

11th International conference on Database Management Systems (DMS 2020)

International Conference on Networks & IOT (NeTIOT 2020)

9th International Conference on Information Technology Convergence and Services (ITCS 2020)

9th International Conference on Signal & Image Processing (SIP 2020)

7th International Conference on Foundations of Computer Science & Technology (CST 2020)

7th International Conference on Artificial Intelligence & Applications (ARIA 2020)

NEWS ARTICLE TEXT CLASSIFICATION AND SUMMARY FOR AUTHORS AND TOPICS

Aviel J. Stein[1], Janith Weerasinghe[2],
Spiros Mancoridis[1] and Rachel Greenstadt[2]

[1]College of Computing and Informatics, Drexel University,
Philadelphia, Pennsylvania, USA
[2]Tandon School of Engineering, New York University, New York, USA

ABSTRACT

News articles are important for providing timely, historic information. However, the Internet is replete with text that may contain irrelevant or unhelpful information, therefore means of processing it and distilling content is important and useful to human readers as well as information extracting tools. Some common questions we may want to answer are "what is this article about?" and "who wrote it?". In this work we compare machine learning models for evaluating two common NLP tasks, topic and authorship attribution, on the 2017 Vox Media dataset. Additionally, we use the models to classify on a subsection, about ~20%, of the original text which show to be better for classification than the provided blurbs. Because of the large number of topics, we take into account topic overlap and address it via top-n accuracy and hierarchical groupings of topics. We also consider edge cases in authorship by classifying on inter-topic and intra-topic author distributions. Our results show that both topics and authors readily identifiable consistently perform best when using neural networks rather than support vector, random forests, or naive Bayes classifiers, although the latter methods perform acceptably.

KEYWORDS

Natural Language Processing, Topic Classification, Author Attribution, Summarization, Machine Learning

1. INTRODUCTION

The Internet is full of information, and a large part of it is text and images. Images are fast for humans to process but text takes more time. Natural Language Processing (NLP) techniques use statistical and computation driven methods to analyze large bodies of text. One of the most common forms of text online is a news article. In Section 2, we discuss related work in NLP. Two common tasks for NLP scientists is either authorship or topic classification. Authorship classification can be useful for plagiarism or detecting fake accounts and topic classification can be helpful for sorting or searching a dataset. The 2017 Vox Media is an understudied dataset that has advantages over other contemporary news article datasets in terms of the number of articles as well as labeled topics and authors. Most studies only explore one of these tasks, so one advantage of this work is that we explore both side-by-side in the same context, and, thus, showing that they are comparable techniques. Another item we explore is how extractive summaries of text can help distill important information from larger texts for either human or model consumption. These NLP techniques are helpful for many academic and industrial applications as off-the-shelf, open-source tools have become more reliable and accessible.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 01-12, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101401

Because contexts may differ, it is important to have baselines and reusable datasets to compare results or build models for transfer learning. One such dataset is the “20 newsgroup text dataset”, which contains around 18,000 articles on 20 topics and does not include author labels. By contrast, Vox Media published a dataset that includes approximately 23,000 articles covering 186 topics and 817 authors. The Vox Media dataset [1] was published in 2017 and has received surprising little attention from the NLP community.

In Section 3, we discuss what methods we use to extract features and classify text. Text classification generally relies on machine learning to provide high accuracy results when applied to large data sources. For our two classification tasks, authorship attribution and topic classification, we extracted several types of features such as word n-gram, term frequency inverse document frequency (TFIDF), and part of speech (PoS) features but found that n-gram word count resulted in the best performance. We perform classification with various common machine learning models (see Section 3.3). Text summarization is performed by distilling the most important pieces of the text to a suitable degree of the original text. We used word frequency as the words score in each sentence and found the sentence score by averaging the score of all the words in a sentence, baring stop words. We constructed two types of dataset for topic and author, the dense dataset contained 10 classes each with 300 samples and the sparse dataset contained 50 classes each with 50 samples.

In Section 4, we perform the experiments demonstrating NLP efficacy for Vox articles. After performing the classifications, we inspected our models by performing confusion matrix and feature analysis, to understand how the classification may be affected by a confluence of signals. Previous work [2] on the Vox Media dataset explored the use of unsupervised learning to identify topics and categories of articles. This is a good approach, since several of the topics are closely related (e.g., politics vs. politics and policy). We also used some unsupervised approaches to explore what kind of commonalities the texts exhibited regardless of their labeled class. To account for this, we used a top-n accuracy and 2-layer hierarchical approach. We categorize similar topics as into groups and first classify on the main topic, then categorize the sub-topics within each category. To account for authorship possible edge cases, such as all authors writing about the same topic or each author never writing about the same topic more than once, we also constructed inter-topic and intra-topic datasets and found that in both cases the authorship signal is still strong, sometimes stronger than the topic signal. Generally, though, authors tend to write about the same topics as they have in the past. For the 10-class dataset we attained 74% accuracy topic attribution and 86% accuracy author attribution. Using the same methods to extract features from the summaries of the 10-class dataset, we obtained 60% and 53% accuracy for topics and authors respectively. Summaries retained the authorship signal because they consist of a subsection of sentences from the original text. these summaries contained valuable information for machine learning models than the original summary, or “blurb”, provided by the dataset. Correcting for topic overlap, with top-n and hierarchical models we can attain topic attribution between 83%-87%. We also considered inter-topic and intra-topic authorship attribution and found that with similar conditions to the dense dataset, in this case 8 authors with 300 samples each, authorship can be attributed with up to 92% accuracy in inter-topic. Intra-topic is a little harder, with only 50 samples each and 8 authors, it scores 68% accuracy. Finally, in Section 5, we consider the limitations of these approaches, discuss the implications of our work, and suggest ways that future research can improve upon them.

2. RELATED WORK

2.1. Topic Classification

There are many aspects of text that can be attributed beyond topics as well such as classifying news based on bias [3] (see Figure 1) and credibility [4] as well as detect fake news [5]. For example, one approach classifies news articles based on their source and attributed to Fox, Vox, or PBS with at best 94% accuracy, but is it because of the text's style or the content signal? [6] The approach used by Yirey et al. [7] focuses on distinguishing between articles on Finance, Stocks, Education, and Environment and scores around and with a similar number of articles per topic. However, one drawback was that the dataset had to be well balanced. Another use of topic analysis is tracking topics that a user may be interested in and can help suggest future articles for the user to read.[8] This of process has to do two things, 1) track articles a user reads, and 2) identify the topics articles. Topics of past articles will likely be similar to topics in future articles. An open question is whether a user likes articles written by certain sources or authors.

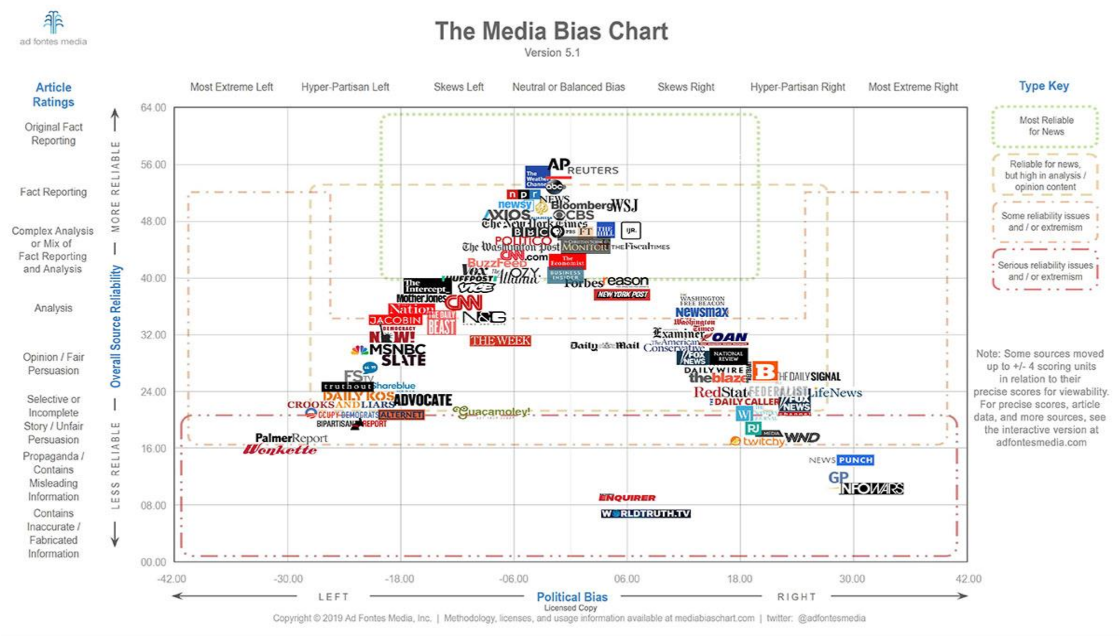


Figure 1. News organizations by political bias and overall reliability. Vox articles skew to the left of the political spectrum and are generally a "complex analysis or mix of fact reporting and analysis"

2.2. Authorship Classification

A related problem involves attributing individual authorship to documents. While this is not strictly an NLP task, as it has also been applied to other things where authorship is relevant such as art[9], music[10], source code[11], *etc.*, it is most prevalent with text. Authorship classification can be used to determine if someone plagiarized or helped preserve the anonymity of the author. Researchers performed deep learning authorship attribution on a dataset of 10 authors and lengthy articles achieving 95% accuracy, and, on shorter articles, 77% accuracy.[12] However, in the case of a news organization, they may have tens or hundreds of authors, so it may not be as robust in those circumstances. In less edited and smaller text portions, white prints scored 95% accuracy on eBay comments.[13] The level of professional editing could influence how much style is present. Another difference between this and other text corpora, is that the authors of articles likely pass their work through editors, and also likely have a style guild. This may create organizational signal or decrease authorship signal.

2.3. Text Summarization

Text summarization is the processes of generating a condensed document that retains the meaning and important information from the original text source. We generate a summary by using a small fraction of the text from the original. The two main ways to generate summaries are extractive and abstractive. Abstractive is harder and requires sophisticated learning and NLP approaches to produce novel phrasing. We choose to go with extractive because they filter for the most important sentence and are easy and flexible to construct. There are also two different kinds of summaries, inductive and informative. Inductive tend to be very short (~5% of original text) and informative are longer (~20% of original text). [14,15,16] In their survey of existing summarizing methods, they compare these methods but take the categories for granted.

3. METHODS

This section describes the data, feature extraction, machine learning classifying models, and summary methods used for our results. We choose to use balanced datasets (*i.e.* those with approximately the same number of samples per class) for the sake of visualizations, though the results remain about the same with natural distributions. We used common strategies for feature extraction including term frequency inverse document frequency (TFIDF), n-gram, and PoS. Additionally, we compared different machine learning algorithms to see how they performed under a variety of conditions. We then summarized the text by using a reductive model.

3.1. Data & Preprocessing

We start by considering the Vox Media 2017 dataset[1]. Most authors have fewer than 50 articles, yet authors with more than 50 articles account for 91% of all articles published. Similar, most topics have fewer than 50 articles. There are also several articles written by multiple authors and some author's names are clearly pseudonyms, for example "A #Never Trump Delegate". Also, many of the topics are related, which we deal with in Section 4.2. To deal with this skewed data, we construct two curated subsections of the data containing balanced number of articles per class (*i.e.,* author or topic). One contains 10 classes, each with 300 articles, the other contain 50 classes, each with 50 articles. Having the dataset balanced is useful for dissecting the results in the confusion matrices (Figures 3-4), but do not significantly improve overall accuracy. We also choose to ignore topics such as "Life", "Identities", and "The Latest" because we found that they tend to act as a miscellaneous category for Vox instead of focusing on a topic. We also filtered out the topics "Xpress" and "Vox Sentence" which tend to have very short articles, which makes them unsuitable for this task in addition to often being vague. This dataset has many favorable features such as being well curated for machine learning, including author and topic labels, and including inductive summary, which most other datasets such as *20 news organization* do not have.

3.2. Features

Machine learning models use features from the text to learn the class signatures. To this end, we extracted three types of features. First, we use word count and word bigram count. We also use word and word bigram TFIDF. We also use Natural Language Tool Kit's (NLTK) built in Tree bank Word Tokenizer and tagger to do PoS. We also limit the number of features in order to train the models more efficiently. We ignore features that are either very common or very rare as they are prone for bloating or over fitting, specifically by limiting features with term frequency between 0.01 and 0.99. We use the Random Forest (RF) model for feature importance evaluation. We also exclude all non-alphabetic characters besides spaces and periods. We exclude some features that are artifacts of the web embedding. Finally, we use the RF feature importance metric to look at what features are most important for distinguishing classes. Though we tried various features types, we found that n-gram term frequency performed best, while also not causing over fitting and use the same feature construction parameters for both authorship and topic.

3.3. Classifier Models

Discrete classification is a machine learning task with many classifiers readily available. We use several different learning algorithms and techniques as a comparative opportunity. We use naïve bayes (NB), decision trees (DT),RF, support vector classifiers (SV), and multi-layer perceptron neural networks (NN). These learning algorithms are supported by many open source libraries. We use a one vs. all (OvA) strategy with the NN to get slightly better results.

3.4. Text Summarization

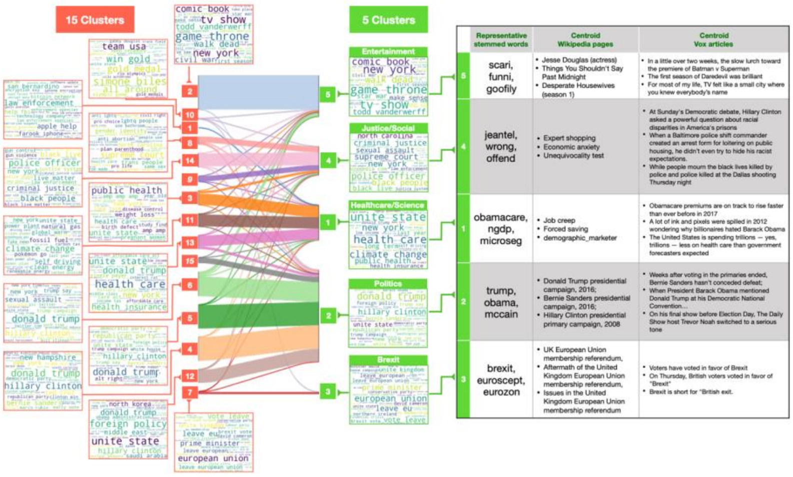
To generate the text summaries, we used the NLTK sentence and word tokenizer and for removing stop words (*i.e.,* common words).Then, we tokenize at the sentence and word levels; filter out the stop words; and calculate the frequency of every non-stop word. We then score a sentence by the sum of the frequency score of the words, divided by the length of the sentence. We are then able to score the sentences and keep the sentences with high scores according to a threshold. Since this method filters for the most salient sentences and does not create new sentence structures or introduce new words or phrases it is an extractive rather than abstractive method. More complex summary methods could be applied to these texts, but we will consider that for future work. Informative summaries are generally around 20% the length of the original text[14], therefore we give about 10% margin on either side and remove 70%-90% of the original text. The threshold for doing this varies on the number of sentences and length of the document. To address this, we apply the summarization method several times until it converges to within the desired margin. If it cannot converge, we discard the document. The reason for not being able to converge is likely from sentences not falling into the threshold we set, which could be because the sentences are too long or the text is too short. The occurrence is rare and likely does not artificially inflate the results. The blurbs that are included in the dataset are on average (2.1+/-0.7%) the length of the article for our experiments, which is on the low end of acceptable for inductive summaries.

$$Sentence\ Importance = \frac{\Sigma(Word\ Frequency)}{Number\ of\ Words\ in\ Sentence}$$

Equation 1. Simple sentence importance calculation

3.5. Unsupervised Techniques

Text that is found in the wild is messy. The Vox Media dataset has the advantage of being well sorted and with pre-assigned labels, but even with this advantage, it has characteristics that make classification difficult. For example, many of the labels are closely associated and topics that vary in size and breadth. This is why initial research on this data focused on unsupervised approaches to topic clustering and lacked direct accuracy results as included in work. Related work showed that there are topics that emerge from the data such as, politics, entertainment, *etc* (see figure 2) and found that there are various numbers of clusters that have good coherence.[2] Other related work uses Latent Dirichlet Allocation (LDA) [17] to analyze how topics words compare to those of high importance to RF classification. LDA works by comparing word and topic distributions. By comparison, RF ranks the words by importance by finding the words which most separate topics.



	Representative stemmed words	Centroid Wikipedia pages	Centroid Vox articles
5	scari, funni, goofily	• Jesse Douglas (actress) • Things You Shouldn't Say Past Midnight • Desperate Housewives (season 1)	• In a little over two weeks, the slow lurch toward the premiere of Batman v Superman • The first season of Daredevil was brilliant • For most of my life, TV felt like a small city where you knew everybody's name
4	jeantel, wrong, offend	• Expert shopping • Economic anxiety • Unequivocality test	• At Sunday's Democratic debate, Hillary Clinton asked a powerful question about racial disparities in America's prisons • When a Baltimore police shift commander created an arrest form for loitering on public housing, he didn't even try to hide his racist expectations. • While people mourn the black lives killed by police and police killed at the Dallas shooting Thursday night
1	obamacare, ngdp, microseg	• Job creep • Forced saving • demographic_marketer	• Obamacare premiums are on track to rise faster than ever before in 2017 • A lot of ink and pixels were spilled in 2012 wondering why billionaires hated Barack Obama • The United States is spending trillions — yes, trillions — less on health care than government forecasters expected
2	trump, obama, mccain	• Donald Trump presidential campaign, 2016; • Bernie Sanders presidential campaign, 2016; • Hillary Clinton presidential primary campaign, 2008	• Weeks after voting in the primaries ended, Bernie Sanders hasn't conceded defeat; • When President Barack Obama mentioned Donald Trump at his Democratic National Convention... • On his final show before Election Day, The Daily Show host Trevor Noah switched to a serious tone
3	brexit, euroscept, eurozon	• UK European Union membership referendum, • Aftermath of the United Kingdom European Union membership referendum, • Issues in the United Kingdom European Union membership referendum	• Voters have voted in favor of Brexit • On Thursday, British voters voted in favor of "Brexit" • Brexit is short for "British exit.

Figure 2. Sankey Diagram from [2] demonstrates an unsupervised approach clustering Vox Media articles by topic with 15 clusters to 5 clusters. Also included are root words and examples of wiki pages and Vox article titles.

4. Experiments & Results

The results of our work are a comparative analysis of author vs. topic classification with full text and summaries. For the author and topic comparison we use the same models, features, and dataset structure, though the individual articles may differ. Summaries were generated from the articles directly. We also include unsupervised learning to get an intuitive understanding of the data, such as scatter of article clusters and list of topic words. Finally, we consider how to handle problems with topic overlap and edge cases for authorship.

4.1. Authorship vs. Topic Classification

We start with stylometry. We can detect the style of an author statistically by doing the feature extraction as described previously. In the dense dataset, we trained with articles of the top 10 most prolific authors of Vox. With the dense dataset, we used 80% for training and saved 20% for testing, and attained up to 84% accuracy using a neural network with the OvA (NN_OvA); though the other methods also behave fairly well for this task. With the Sparse dataset we have some loss in signal but still strong considering there we are classifying 50 authors with 70% accuracy, whereas the baseline for guessing is 2%.

Table 1**.** Authorship attribution accuracy with various Machine Learning models on with n-gram word counts features.

Model	**Dense % Accuracy**	**Sparse % Accuracy**
Naïve Bayes	81	64
Decision Trees	53	30
Random Forest	74	51
Support Vector	74	32
Neural Network	83	51
Neural Network One-vs-All	**86**	**70**

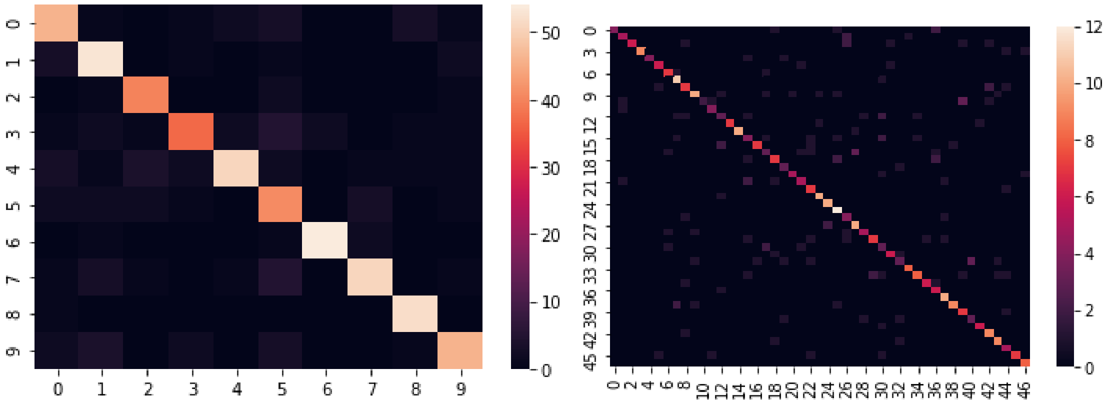


Figure 3.Authorship confusion matrix for dense using NN_OvA model (left) and sparse using NN_OvA model (right).

Topics can be classified with between 62%-74% accuracy. Therefore, given similar information, topics are 10% less accurate with dense information and 8% less accurate with sparse information. Looking at the confusion matrix of topics with dense information, it appears that one topic tends to dominate, and in the sparse dataset there appears to be two that were misclassified as each other. Where as in the authorship case, the errors are more scattered. Additionally to compare this approach to other work with fewer number of topics, such as in [7] and fewer articles we were able to score 90% accuracy in distinguishing between “Politics & Policy”, “Science & Health”, “Culture”, and “Business & Finance”.

Table 2. Topic classification accuracy with various Machine Learning models on with n-gram word counts features.

Model	**Dense %Accuracy**	**Sparse % Accuracy**
Naïve Bayes	73	61
Decision Trees	55	45
Random Forest	70	62
Support Vector	65	38
Neural Network	72	53
Neural Network One-vs-All	**74**	**62**

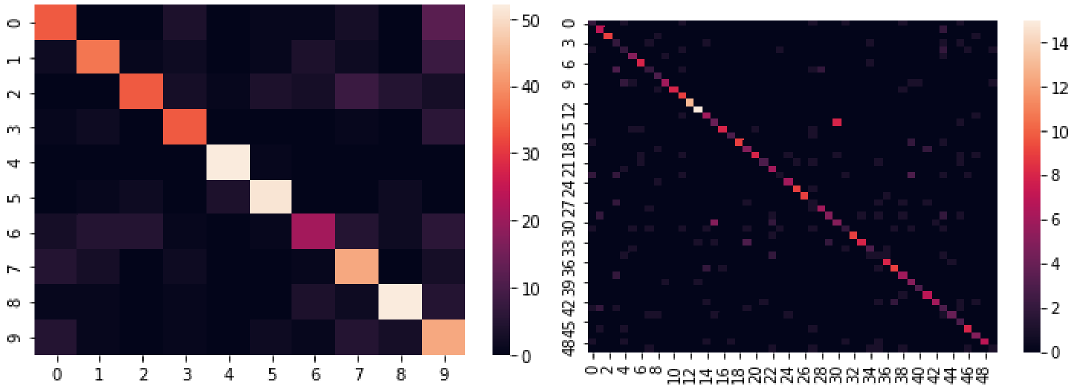


Figure 4. Topic confusion for dense using NN_OvA model (left) and sparse for dense using NN_OvA model (right).

We consider two kinds of summaries: inductive and informative. The Vox Media dataset comes with a short "blurb" for each article which makes for good comparison with our generated summaries. Each blurb is ~2% the length of the original text. The summaries we generate end up being ~17% the length of the original text. As can be seen in Table 3, summaries lose about 25% authorship signal and 20% topic signal. The scores that they get are still far above random guessing for both author and topic classification, though some useful information can be lost. This may be because the text becomes more general. So in terms of evaluating the quality of the summary, it clear that the longer summaries that we generate are better suited for learning models. This approach to summaries can be helpful if the amount of data is large by reducing it by 80% but the models work better with access to the full text. There should be continued exploration into abstractive summaries for this same task. The hope is that by abstracting information in novel ways (*i.e.*, not verbatim from the text) that salient information could be condensed more effectively.

Table 3. Classification Results on Summaries using NN_OvA model

Data Source	**Dense % Accuracy**	**Sparse % Accuracy**
Blurb for Authorship	19	8
Generated for Authorship	60	42
Blurb for Topic	21	5
Generated for Topic	53	41

4.2. Unsupervised Insights

The prior analysis focused on supervised learning with handcrafted labels as given by Vox Media organization. However, unsupervised learning can provide insights into trends within the data. We first consider how feature importance as learned from the RF compares to related words extracted by LDA. The LDA groups words by certain components and gives the top words for each component. The top 10 words for top 10 topics for each dataset and the feature importance from of the top 100 words from the RF. They share many of the same words with high importance such "Trump", "health", and "people". There are two potential issues we see from this, 1 (as addressed in Section 4.3), there are topics that overlap, 2 (as addressed in Section 4.4), authorship seems to be tied to topic in some way.

4.3. Adjusting for Topic Overlap

As we explored in the ways to address topic overlap, such as hierarchical and top-n approaches, the unsupervised language models may be more indicative of patterns within a of body of text. Therefore, we provide the reader with some visuals of an unsupervised view of the data. We use principle component analysis to visualize the data based on author and topic and then use k-means clustering with 10 clusters to show how that fits the data.

One of the problems we noticed with the labeled topics is that some are general or closely related to other topics. To address this, we ease the classification by a top-n and a hierarchical approach for the sparse topic data. This is the case that makes the most sense because there are enough categories of things that could be conflated. Using the top-5 topics brings the accuracy from 62% up to 87%. To do the hierarchical model, we have to manually select topics of each group. Informed by [2], and descriptions online, we group 5 super-topics, each with a number of subtopics (see Table 4) and include 800-1000 samples each. Using the machine learning models we can get 84% accuracy.

Table 4. Five general topics and associated subtopics

General Topics	**Total # Articles**	**Subtopics**
Politics	5479	Politics and Policy, Politics, Mike Pence, Ted Cruz, Congress, Hillary Clinton, Marco Rubio, Donald Trump, Jeb Bush, Bernie Sanders, Mischiefs of Faction
Health	1173	Health Care, Infectious Disease, Obama Care, Science & Health
Environment, Technology & Business	1084	Energy and Environment, Grist, New Money, Apple, Transportation, Space, Business & Finance, Technology, Labor Market
Social Issues	848	LGBTQ, Identities, Race in America, Marriage Equality
Entertainment	1442	Books, Game of Thrones, Movies, Culture, Music, Episode of the Week, Star Wars, Reviews

This approach differs from the top-n approach because we had to manually choose groupings, whereas with top-n, each article may have a different top grouping and still score correctly. This approach has the advantage that one can specify the topic and subtopics but works at a comparable level for a much smaller range of topics and needs more data.

4.4. Stylometry via Intra-topic and Inter-topic Authorship Classification

We suspected that there may be confusion in the signal between topics and author. As mentioned in [18] there may be irrelevance by correlated features which, when under unfortunate circumstance, cause highly confident incorrect classifications. Their example uses rotating images in the MNIST data. The concern in our case is that the topic may be indicative of the author. After all, some authors specialize in topics so instead of style detection, maybe it is a sort of article detector. While our goal is authorship classification, it is not strictly style. But to address this, we also consider trying to detect style by containing samples from within a topic. We choose to run experiments for The Latest, Donald Trump, and Politics and Policy because they had enough authors with enough articles each to do comparable experiments. The topics had between 8-10 authors with 60-300 (see Table 5). For the experiment on “The Latest”, which most resembles the dense dataset, it scores even higher, but this may be because it is a miscellaneous category. Whereas the experiment on “Donald Trump” yields 64%, maybe because it had fewer samples or because it was more specific. We also performed an intra-topic experiment where authors were allowed only one sample per topic. For this experiment we had 8 authors with 50 samples each and it scored 68% accuracy. It would appear that authorship is actually easier to detect within a topic, but it can be detected whether the author focuses on one topic or writes about many with consistent accuracies.

Table 5. This shows how well authorship stylometry works within topics and the #articles indicates the number of articles per author.

Topic	Number of Authors	Number of Articles per Author	Accuracy
The Latest	8	300	92%
Donald Trump	9	60	64%
Politics & Policy	10	60	81%

5. LIMITATION, DISCUSSION, & FUTURE WORK

This work provides insight into common NLP techniques and tools for a large unexplored dataset. It shows what kind of accuracy to expect from a dataset in which the text was well edited but large and shows that state-of-the-art accuracy can be achieved for authorship and topics. We also suggest ways of dealing with topic overlap in new contexts and discuss handcrafted vs. naturally occurring groupings. We hope these results are helpful for other NLP researchers in the pursuit of linguistic knowledge and that future research use it to enlighten their search and find better ways to achieve similar goals. We also demonstrate that a reductive approach to text summarization retains both authorship and topic signal to some degree. However, other summarization approaches could be explored in this context for interesting results.

5.1. Discussion

We see that we can use topic and author signals to classify documents. One concern that was raised is how these signals conflate. It is my belief that they are inexorable intertwined with regard to authorship. An example of this concern is that if an author writes a lot about a specific topic, what is classifier picking up on? So, for example, in [19] they use stylometry to test for plagiarism using student academic papers as their corpora. But since academic papers are required to be novel and are usually about very specific topics, it is not clear that they are not picking up on authorship or topic similarity.

There has been some work on how to know whether or not to trust your classifier when there is a "data shift". Their method deals with classifying the MNIST dataset and rotating the images. However changing between perhaps non-independent classification, there may be no way to disentangle with certainty,[18] or one may need to be aware of out of distribution changes in the data.[20] However, with unaltered data, this is generally not a problem but is necessary for generative models adversarial models. That being said, it is unclear what is the degree of Vox's editing signature that is included in the signal. An interesting question is, if the topic signal would shift when the text was edited to imitate someone else's writing. Changing the phrasing of sentences can throw off these types of attribution. Tools like ParChoice [21] retain semantics while changing specific words. These types of adversarial should be considered for creating robust models. Another way of evaluating the semantics in these cases would be to make sure they still do well in the topic classification cases. If that fails it is likely that they are changing the meaning rather than the style.

5.2. Limitations

There are some ways that this work is limited. It focuses on only articles from Vox, but could be expanded to include articles from other news sources. It also only uses simple machine learning methods, but more advanced neural networks and architectures could be used. Additionally, we could use other methods for generating summaries. We also focus just on English texts, but could apply these techniques to other languages as well.

5.3. Future Work

We measure the efficacy of summary generation for machine classification contexts. The method we used, called extractive, is useful because it is fast, flexible and easy to use. Summaries can also be generated using different means [3], which might result in different or better outcomes depending on the task. Methods involving generation rather than reduction [20] showed that one can adjust for domain data and could be considered for generative methods. Perhaps this could be used to improve text generating GANs. Using these methods on sources with multiple label introduces an interesting multioutput problem.

In addition to improving tasks explored here, there are other interesting pursuits one could take with this data or similar data, to explore where the learning is transferable. There is the concern of various signals being present or biasing the text. Work related to privacy and anonymity are often a concern when it comes to identifying individuals. It is important to be aware that these methods are largely used as supporting forensic evidence rather than absolute truth. However, this could also be used for good if we can use it to debias text or use multiple texts to form a multisource summary.

6. CONCLUSIONS

This work explores the new and rich news article data set provided by Vox Media for the NLP community. We demonstrate that state-of-the-art classification approaches with off-the-shelf language and learning tools are well suited for news articles, even though they may have been edited. We provide direct comparison between style and topic features and show that author attribution can score between 70%-86% accuracy for groups between 10 and 50 authors and between 62%-74% for 10 to 50 topics. The topic accuracy gap can be compensated for, when considering topic overlap in grey areas such as comparing topics like political figures and general politics. We compare top-n and hierarchical topics and combing methods to increase the score to 87%. Additionally, we show that simple extractive summarization techniques retain both authorship and topic signal and show how this compares to human generated abstractive summaries.

ACKNOWLEDGEMENTS

We want to thank Vox Media for providing the data used in this work and for their commandment to adapting journalism to the digital age. Additionally, this work was reviewed and encouraged by the Army Research Laboratory and Funded by the Auerbach Berger Chair in Cybersecurity held by Spiros Mancoridis, at Drexel University

References

[1] Vox Media, Workshop for Data Science + Journalism (DS+J) (2017)

[2] Altuncu, M. Tarik, Sophia N. Yaliraki, and Mauricio Barahona. "Content-driven, unsupervised clustering of news articles through multiscale graph partitioning." arXiv preprint arXiv:1808.01175 (2018).

[3] Kiesel, Johannes, et al. "Semeval-2019 task 4: Hyperpartisan news detection." Proceedings of the 13th International Workshop on Semantic Evaluation. 2019.

[4] Bountouridis, Dimitrios, et al. "Explaining credibility in news articles using cross-referencing." SIGIR workshop on ExplainAble Recommendation and Search (EARS). 2018.

[5] Khan, Junaed Younus, et al. "A benchmark study on machine learning methods for fake news detection." arXiv preprint arXiv:1905.04749 (2019)

[6] Lee, Stephen M. "Variation in Political News." (2019).

[7] Suh, Yirey, et al. "A comparison of oversampling methods on imbalanced topic classification of Korean news articles." Journal of Cognitive Science 18.4 (2017): 391-437.

[8] Kaur, Kamaldeep, and Vishal Gupta. "A survey of topic tracking techniques." Int J 5 (2012).

[9] Hughes, James M., et al. "Empirical mode decomposition analysis for visual stylometry." IEEE transactions on pattern analysis and machine intelligence 34.11 (2012): 2147-2157.

[10] Mara, Michael. "Artist attribution via song lyrics." (2014).

[11] Caliskan-Islam, Aylin, et al. "De-anonymizing programmers via code stylometry." 24th {USENIX} Security Symposium ({USENIX} Security 15). 2015.

[12] Ramnial, Hoshiladevi, Shireen Panchoo, and Sameerchand Pudaruth. "Authorship attribution using stylometry and machine learning techniques." Intelligent Systems Technologies and Applications. Springer, Cham, 2016. 113-125.

[13] Abbasi, Ahmed, and Hsinchun Chen. "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace." ACM Transactions on Information Systems (TOIS) 26.2 (2008): 1-29.

[14] Tas, Oguzhan, and Farzad Kiyani. "A Survey Automatic Text Summarization." PressAcademia Procedia 5.1 (2007): 205-213

[15] Nenkova, Ani, and Kathleen McKeown. "A survey of text summarization techniques." Mining text data. Springer, Boston, MA, 2012. 43-76.

[16] Babar, S. A., and Pallavi D. Patil. "Improving performance of text summarization." Procedia Computer Science 46 (2015): 354-363.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993–1022

[18] Ovadia, Yaniv, et al. "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift." Advances in Neural Information Processing Systems. 2019.

[19] Ramnial, Hoshiladevi, Shireen Panchoo, and Sameerchand Pudaruth. "Authorship attribution using stylometry and machine learning techniques." Intelligent Systems Technologies and Applications. Springer, Cham, 2016. 113-125.

[20] Ren, Jie, et al. "Likelihood ratios for out-of-distribution detection." Advances in Neural Information Processing Systems. 2019.

[21] Gröndahl, Tommi, and N. Asokan. "Effective writing style imitation via combinatorial paraphrasing." arXiv preprint arXiv:1905.13464 (2019).

CADA-FVAE-GAN: ADVERSARIAL TRAINING FOR FEW-SHOT EVENT DETECTION

Xiaoxiang Zhu, Mengshu Hou, Xiaoyang Zeng and Hao Zhu

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

ABSTRACT

Most supervised systems of event detection (ED) task reply heavily on manual annotations and suffer from high-cost human effort when applied to new event types. To tackle this general problem, we turn our attention to few-shot learning (FSL). As a typical solution to FSL, cross-modal feature generation based frameworks achieve promising performance on images classification, which inspires us to advance this approach to ED task. In this work, we propose a model which extracts latent semantic features from event mentions, type structures and type names, then these three modalities are mapped into a shared low-dimension latent space by modality-specific aligned variational autoencoder enhanced by adversarial training. We evaluate the quality of our latent representations by training a CNN classifier to perform ED task. Experiments conducted on ACE2005 dataset show an improvement with 12.67% on F1-score when introducing adversarial training to VAE model, and our method is comparable with existing transfer learning framework for ED.

KEYWORDS

Event Detection, Few-Shot Learning, Cross-modal generation, Variational autoencoder, GAN

1. INTRODUCTION

As an essential subtask for IR, ED aims at identifying the event triggers in the text and assigning the pre-defined event types to each of the triggers. There are 33 types of events according to the ACE2005 corpus, such as "*Attack*", "*Transport*", "*Die*" etc. For instance, in the sentence "Tuesday's southern Philippines airport blast", "blast" is the trigger of event "*Attack*", ED system should identify the word "blast" and categorize it to the corresponding event type.

ED task is usually modeled as the multi-classification problem in the traditional supervised methods. These methods suffer from the heavy reliance on manual annotations and features specific to the particular event types, which makes it difficult to handle new or unseen types without additional human annotation efforts. In order to overcome this challenge, we model ED task with transfer learning approaches.

Few-shot learning framework, as a typical solution for transfer learning, which enables models to handle classification task for new classes of examples, gives us a valuable inspiration. The goal of FSL is to learn transferable knowledge from training examples (*seen* classes) to test examples (*unseen* classes), with only a few examples moved from test into training examples. Zero-shot learning (ZSL) is another framework similar to FSL, where the classes of training and test

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 13-24, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101402

examples are absolutely disjoint. Both zero- and few-shot learning approaches typically exploit semantic knowledge to achieve transferability. [1-4] improves zero-shot predictions of images with semantic knowledge learned from unconstructed text description. Neural Snowball [5] is a few-shot relation extraction (RE) framework transferring semantic knowledge from existing relations to new ones. [6] designs a hybrid attention-based neural model to improve noisy few-shot relation classification (RC) by grasping external knowledge. [7] applies ZSL to event extraction problem by learning a generic mapping function of event types and mapping both event mentions (trigger and context) and types into a shared semantic space.

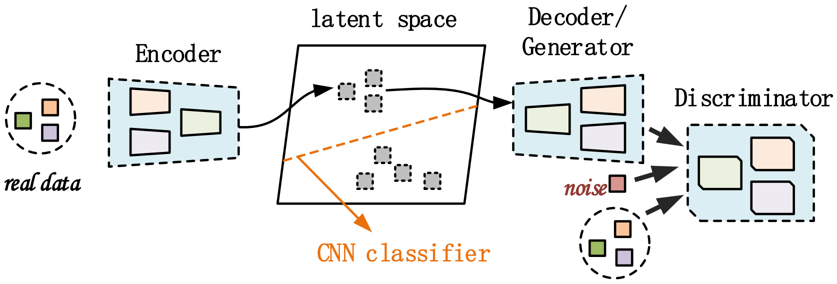


Figure 1. A review of our CADA-fVAE-GAN model

Our CADA-fVAE-GAN is a cross-modal framework utilizing VAE as its key module for feature generation. Recently, cross-modal deep learning model has received much attention.[8] proposes a ZSL classification framework, on which image features and its descriptive text of categories are mapped into a shared semantic word vector space. [9] and [10] jointly learn multi-modal representations with distribution alignment in their latent space. [11] uses VAE-based generative model to perform generalized ZSL/FSL via images classification problem, by mapping multi-modal samples into a shared latent low-dimension feature space, which achieves encouraging results. Significantly, [11] indicates that latent features constructed by VAE are semantically interpretable for classification. Therefore, we decide to advance the model to few-shot learning for event detection framework, and study the generality of latent features inferred by VAE encoder. The superiority of VAE lies in variational inference, while its generation performance is not comparable to some powerful generative models such as GAN. In some degree, this is partly due to the fact that VAE is not capable to encode high-quality latent representations. To alleviate this problem, [12] develops a conditional generative model with the combination of VAE and GAN, which learns highly discriminative features for downstream task. [13] introduces adversarial training to VAE for better variational inference. [14] combines VAE with GAN and utilizes learned features in data space for better measurement of similarities.

Inspired by adversarial training of GAN and its attractive usage in NLP [15-18], we fold the generator of GAN and the decoder of VAE into one, realizing the sharing of neural parameters and training process. We notice that the input for our model consists of three modalities, including event mention structures parsed by AMR, event type structures and type name embeddings, which are not considered as high-level representations. And experimental results indicate that using abstract features extracted from these modalities by CNN will degrade the interpretability of latent representations constructed by VAE encoder. However, a CNN classifier is able to extract valuable features from latent representations with original modalities as input, and produce acceptable classification results for our ED task. Figure 1 is a brief review of our CADA-fVAE-GAN model. In summary, our main contributions is three-fold:

1. We apply VAE-based generative model to few-shot learning of event detection for the first time, and demonstrate the transferability of latent representations constructed by VAE.
2. We combine VAE with GAN to improve the quality of latent representations and the transferability of the model via adversarial training.
3. Experiments conducted on ACE2005 dataset achieve ideal results, which demonstrates the effectiveness of our proposed model, and indicates some promising direction for further research on ED problem.

Next, in Section 2, we discuss several representative works on event detection, including traditional feature based methods and recent neural network based methods. In Section 3, we explore the architecture of our CADA-fVAE-GAN model, and then we perform ablation study to evaluate each module in Section 4. Finally, Section 5 concludes our work and discusses possible future scope for further research.

2. LITERATURE REVIEW

Recently, much more attention has been attracted on event extraction. Traditional methods are mainly based on feature-learning. [22-25] reply on fine-annotated textual features to identify the types of event triggers. With the emergence of deep neural network, [26-28] exploit convolution neural network (CNN) to construct higher-granularity informative representations through stacked convolution layers, which prove the feasibility of CNN on event detection. Moreover, as a typical sequential model, recurrent neural network (RNN) is equipped with the qualities to perform sentence classification task. [29] extracts syntactic relations by constructing dependency bridges over Bi-LSTM. [30] introduces document-level information to bidirectional RNN and alleviating the complexity for inference rules. [31] builds document embeddings and supervised attention to enhance event trigger identification and classification. [32] combines CNN and Bi-LSTM to extract informative representations for event detection. Recent large pre-trained language model such as BERT [33] and ELMo [34] also attract some researchers: [35] using a transition-based framework and BERT embeddings, [17] using BERT based encoders and adversarial training mechanism, [36] using a Bi-LSTM with BERT token representations, [16] introducing an incremental learning framework with ELMo word representations.

3. METHODOLOGY

We model event detection as a multi-classification task. The inputs for our model include three modalities: event mention structures constructed by AMR and their corresponding type structures and type name embeddings pre-defined in ACE2005 corpus. So we give three VAEs, one for each modality. To improve reconstruction quality of VAE, we collapse decoder in VAE and generator in GAN into one, by sharing neural network parameters and training process. In our model, decoders not only need to reconstruct low-dimension latent representations, which is the basic function of VAE, but also play a role of generator in GAN. Therefore, a randomized noise is constructed as additional input for each decoder/generator. Eventually, M+1(M is the number of modalities, namely 3 in this work) outputs are produced from each decoder/generator, which will be treated as *fake* data for the input of discriminator. And the original input for VAE encoder will be fed into discriminator as *real* data. Discriminator takes its responsibility of differentiating *fake* data and *real* data. The detailed architecture of the model is shown as Figure 2, and a brief training procedure is listed in Algorithm 1.Intuitively, due to the introduction of GAN, VAE decoder will be improved to produce higher-quality reconstructions after adversarial training process. In this way, VAE encoder is expected to construct latent features with richer semantic meanings, which is beneficial to downstream task.

3.1. AMR Semantic Graph Encoder

FSL usually needs high-quality class representations to learn a robust transfer learning model. Following [7], we take advantage of AMR to build semantic graph by identifying event triggers and arguments (such as *Time*, *Location*, *Person*, etc). For instance, the AMR-parsed event mention structure of the sentence “1994 civil war in Rwanda, where government-led militia slaughtered an estimated 800,000 opposition,…” and “Toefting transferred to Bolton in February

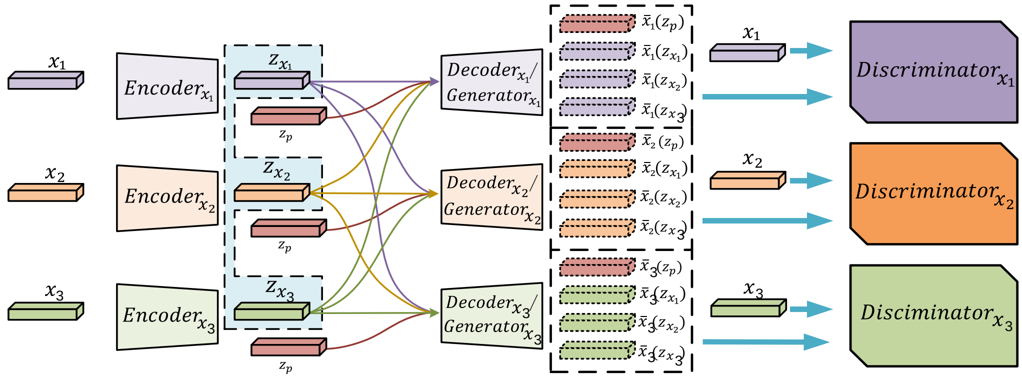


Figure 2. Architecture CADA-fVAE-GAN model

Algorithm 1 Training Procedure

```
Input:
    $M = 3$ is the number of modalities
    $N_{epoch} = 80$ is the number of training epoch
    $N_d$ is the iterations to train discriminator in each epoch
    $\omega = 1000$ is the weighting factor for $\mathcal{L}_G$
for epoch in $1, 2, \ldots, N_{epoch}$ do
    Sample $x_1, x_2, \ldots, ...x_M$ from $M$ modalities
    # Train discriminator
    for $n$ in $1, 2, \ldots, N_d$ do
        Use VAE encoder to map $x_1, x_2, \ldots, ...x_M$ into latent space as $z_{x_1}, z_{x_2}, \ldots, z_{x_M}$
        Sample a random noise $z_p$ from $\mathcal{N}(0, I)$
        Use VAE decoder to produce reconstructions $\bar{x}_1(z_{x_i}), \bar{x}_2(z_{x_i}), \ldots, \bar{x}_M(z_{x_i}), \bar{x}_i(z_p)$ for
        all $i$ in $1, 2, \ldots, M$
        Compute $\mathcal{L}_D$ with $\bar{x}_1(z_{x_i}), \bar{x}_2(z_{x_i}), \ldots, \bar{x}_M(z_{x_i}), \bar{x}_i(z_p)$ and $x_i$
        $w \leftarrow Adam(\nabla \mathcal{L}_D, w)$
    end for
    # Train VAE like a generator
    Use VAE encoder to map $x_1, x_2, \ldots, ...x_M$ into latent space as $z_{x_1}, z_{x_2}, \ldots, z_{x_M}$
    Use VAE decoder to produce reconstructions $\bar{x}_1(z_{x_i}), \bar{x}_2(z_{x_i}), \ldots, \bar{x}_M(z_{x_i}), \bar{x}_i(z_p)$ for all
    $i$ in $1, 2, \ldots, M$
    Compute $\mathcal{L}_{fVAE}, \mathcal{L}_{CA}, \mathcal{L}_{DA}$
    Use decoder as generator to compute $\mathcal{L}_G$
    $w \leftarrow Adam(\nabla \mathcal{L}_{CADA-fVAE} + \omega \mathcal{L}_G, w)$
end for
```

2002 from German club Hamburg.” are shown in the Figure 3 (top). Figure 3 (bottom) also shows their pre-defined event type structures in ACE2005 dataset. Considering the shared semantic meaning between an event trigger and its type name, and the similarity between mention structure and its type structure, we exploit multi-modal VAEs to map three modalities into a shared latent space, then extract their semantic representations.

3.2. Preprocessing for Multi-Modal Structural Features

According to the learned event mention structures, we represent each edge in the directed graph as a tuple u $=< w_1, \epsilon, w_2 >$, where w_1, w_2 denote word entities at endpoints, ϵ denote the AMR relation between w_1 and w_2, such as <war, :mod, civil>. For each event mention structure, we fix the number of binary relations to r, then map w_1 and w_2 to their word embeddings V_{w_1} and V_{w_2} $\in$ R^d, where d is dimension of word embeddings. Then V_{w_1} and V_{w_2} are concatenated and we can get a matrix $M_u \in R^{2d\times \mathrm{r}}$ representing all the relations in the event mention. Assume that $M_\epsilon \in R^{2d\times 2\mathrm{d}}$ is the matrix representation of AMR relation ϵ, then $M_\epsilon \times M_u$ is the composition representation for the event mention structure.

For the type structure, each edge in the graph is represented as v $=< a, b >$. The number of such tuple in each type structure is also fixed to r. Concatenate embeddings of word entities a and b, namely V_a and $V_b \in R^d$, we get the matrix representation of type structure $M_v \in R^{2d\times \mathrm{r}}$.

As for the event type name denoted by t, we simply use its word embedding $V_t \in R^d$.

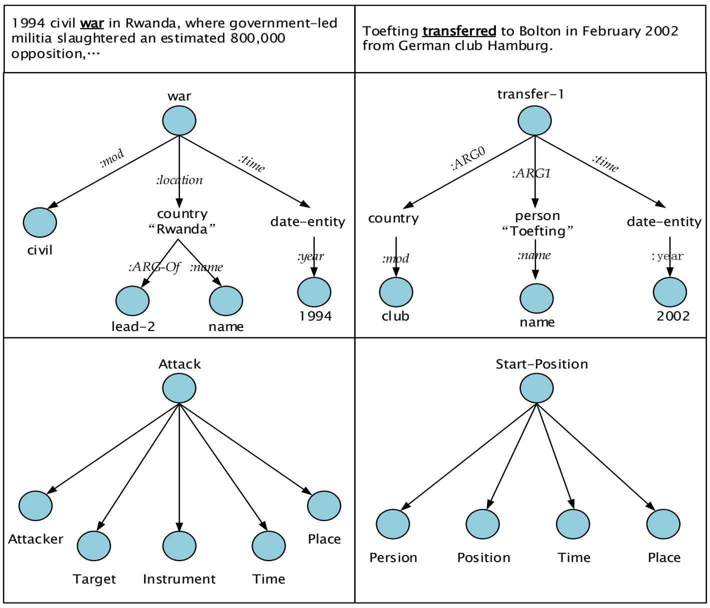


Figure 3. Examples of Event Mention Structure and Type Structures

In AMR graph, each edge with the keyword “:<arg-name>” represents the semantic relation between two word entities or concepts, including affiliation, coreference, category definition and target orientation. The root node is usually the central word in a sentence, such as event name or an action name

3.3. Adversarial Latent Features Generating Network

In this section, we propose a VAE-based feature generation model with cross- distribution-aligned and adversarial training for event detection, which is called CADA-fVAE-GAN. The model exploits adversarial training to strengthen VAE and few-shot learning process.

3.3.1. Basic VAE

Variational autoencoder (VAE) is a typical generative neural network consisting of an encoder and a decoder. VAE encoder maps the given data into a latent feature space, and the decoder reconstruct latent features and maps them back to original data space. Different from trivial autoencoder, VAE is skilled at inferring the true conditional probability distribution of latent variables z$\sim p_\theta(z|x)$. VAE performs this by approximating a closest posterior distribution$q_\phi(z|x) \sim \mathcal{N}(\mu, \Sigma)$, and minimize their variational distance. The objective function of trivial VAE is written as:

$$\mathcal{L}(\theta, \phi; \mathrm{x}) = \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x) \parallel p_\theta(z)]$$

where the first RHS term is reconstruction loss, the second term is Kullback-Leibler divergence(KLD) between $q_\phi(z|x)$and $p_\theta(z)$, which can be written as followed in Gaussian case:

$$D_{KL}[q_\phi(z|x)||p_\theta(z)] = -\frac{1}{2}\sum_{j=1}^{J}(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2)$$

where J is dimensionality of z, μ_j and σ_j denote each element of mean and s.d. evaluated at datapoint j.

In addition, when reconstructing original samples from latent variables z, we can adopt a reparameterization trick as followed:

$$\tilde{z} = \mu_z + \sigma_z \odot \varepsilon, \qquad \text{where } \varepsilon\sim\mathcal{N}(0, \mathrm{I})$$

3.3.2. Cross- and Distribution-Aligned VAE

In our few-shot event detection framework, for unseen classes, only category descriptions and a few mention samples are provided to training set. Therefore, it is necessary for the model to have the capability of cross-modal generalization. Namely, one modality-specific encoder/decoder is expected to encode/decode another modalities with high-quality. For the better performance, we exploit β-VAE [19]. Since each modality have its specific VAE, such that x_1 for event mention structure, x_2 for type structure and x_3 for type name embedding, so the final loss for our basic fVAE is:

$$\mathcal{L}_{fVAE} = \sum_{i}^{M} \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta\left(x^{(i)}\middle|z^{(i)}\right)\right] - \beta D_{KL}\left[q_\phi\left(z^{(i)}\middle|x^{(i)}\right) \parallel p_\theta\left(z^{(i)}\right)\right]$$

By weighting KLD with β, we can produce better reconstructions than trivial VAE. Now we introduce constraint to cross-modal reconstruction for every modality-specific VAE. As is depicted in Figure 2, each decoder should learn to reconstruct latent representations from other $\mathrm{M} - 1$ modalities, which leads to our cross-aligned loss:

$$\mathcal{L}_{CA} = \sum_{i}^{M} \sum_{j \neq i}^{M} \left\| x^{(i)} - DEC_j(E_i(x^{(i)})) \right\|^2$$

where E_i and DEC_j denote objective functions of encoder for modality i , and decoder for modality j, respectively.

Furthermore, distributions of different latent variables are aligned by minimizing their Wasserstein distance [20]. The Wasserstein distance between two Gaussian distributions is:

$$W_{i,j} = [\left\|\mu_i - \mu_j\right\|^2 + tr(\Sigma_i) + tr\left(\Sigma_j\right) - 2(\Sigma_i^{\frac{1}{2}} \Sigma_i \Sigma_j^{\frac{1}{2}})^{\frac{1}{2}}]^{\frac{1}{2}}$$

Since covariance matrices constructed by encoder are diagonal and commutative, we can simplify this equation to:

$$W_{i,j} = (\left\|\mu_i - \mu_j\right\|^2 + \left\|\Sigma_i - \Sigma_j\right\|^2)^{\frac{1}{2}}$$

So the total loss of distribution-aligned for M modalities is:

$$\mathcal{L}_{DA} = \sum_{i}^{M} \sum_{j \neq i}^{M} W_{i,j}$$

We combine the basic VAE loss$\mathcal{L}_{fVAE}$ with cross- and distribution-aligned:

$$\mathcal{L}_{CADA-fVAE} = \mathcal{L}_{fVAE} + \zeta \mathcal{L}_{CA} + \vartheta \mathcal{L}_{DA}$$

where ζ and ϑ respectively weight cross- and distribution-aligned loss.

3.3.3. Adversarial Training

Our model aims at providing an enlightening perspective to the semantic representation of latent features via a classification task in the NLP field. Intuitively, the higher quality of reconstructions indicates the more interpretable latent representations. Under the constraint of VAE objective function, improving decoder will accordingly improve encoder. Moreover, it has been shown that combining VAE and GAN leads to better generation results [12-14]. Inspired by the superiority of adversarial training strategy, we decide to link a discriminator following VAE decoder, and decoder plays a role of generator in GAN.

WGAN [21] has been proved to have better theoretical properties than the vanilla GAN, for which we choose WGAN in our model. According to the architecture shown as Figure 2, the losses of generators and discriminators are:

$$\mathcal{L}_G = \sum_{i}^{M} \mathbb{E}[D_i(G_i(z_p^{(i)}))] + \sum_{i}^{M} \sum_{j}^{M} \mathbb{E}[D_i(G_i(z_{x^{(j)}}))]$$

$$\mathcal{L}_D = \sum_{i}^{M} \mathbb{E}[D_i(x^{(i)})] - \sum_{i}^{M} \mathbb{E}\left[D_i\left(G_i\left(z_p^{(i)}\right)\right)\right] - \sum_{i}^{M} \sum_{j}^{M} \mathbb{E}\left[D_i\left(G_i(z_{x^{(j)}})\right)\right] - \sum_{i}^{M} \lambda \mathbb{G}_i$$

where G_i and D_i are generator and discriminator for modality i, $z_p^{(i)}$ is random noise sampled from $\mathcal{N}(0, \mathrm{I})$ for modality i, and $z_{x^{(i)}}$ is latent representation for modality i.$\mathbb{G}_i$ is gradient penalty for modality i, with a penalty coefficient λ:

$$\mathbb{G}_i = \mathbb{E}\left[\left(\left\|\nabla_{\hat{x}^{(i)}_{z_p}} D_i\left(\hat{x}^{(i)}_{z_p}\right)\right\|_2\right)^2\right] + \sum_{j}^{M} \mathbb{E}[(\|\nabla_{\hat{x}^{(j)}} D_i(\hat{x}^{(j)})\|_2)^2]$$

where $\hat{x}^{(i)}_{z_p} = x^{(i)} + \alpha(\tilde{x}^{(i)}_{z_p} - x^{(i)})$, $\hat{x}^{(j)} = x^{(j)} + \alpha(\tilde{x}^{(j)} - x^{(j)})$ with $\alpha \sim U(0,1)$,$x^{(i)}$ and $x^{(j)}$ are the real sample for modality i and j respectively, $\tilde{x}^{(i)}_{z_p}$ is reconstructed from random noise $z_p^{(i)}$, $\tilde{x}^{(j)}$ is reconstruction for modality j.

Final objective function is:

$$\min_{CADA-fVAE,G} \max_{D} \mathcal{L}_{CADA-fVAE} + \omega\mathcal{L}_G + \mathcal{L}_D$$

where ω is the weighting factor.

3.3.4. Implementation details

All encoders and decoders are implemented as MLPs with one hidden layer, which will not degrade performance. On the one hand, AMR graph abstractly represents event mention. On the other hand, a CNN classifier is used to predict event types, and higher-level semantic representations will be obtained further. More hidden layers lose key information. We find that 1560 hidden units for event mention structure encoders and 1660 for decoders produce better results in our work. The encoder of type name embeddings and type structures have 1450 hidden units and 665 for decoders.

The dimension of VAE latent space is 120. Each discriminator is implemented as MLP with one hidden layer and 1450 units, whose output is activated by a Sigmoid. Following [11], gradient penalty coefficient λ is set to 10. We find that $\omega = 1000$ works well on ACE2005 dataset. During the training of 80 epochs, $\zeta$is increased from epoch 6 to epoch 22 by a rate of 0.54 per epoch, while ϑ is increased from epoch 21 to 75 by 0.044 per epoch. As is suggested by [21], we update decoder/generator every 5 discriminator iterations. All modules including classifier are trained using Adam optimizers, with learning rate = 1.5e-4 for VAE and 5e-5 for discriminators. CNN classifier is trained for one epoch with learning rate=1e-3 and CrossEntropyLoss as its criterion.CNN classifier is implemented with two one-dimension convolution layers, each of which contains a ReLU and a MaxPool1d. Final predictions are produced by a fully connected layer.

4. Experiments

4.1. Settings

ACE2005 dataset defines 33 event types, on which experiments are conducted to evaluate the performance of the model. Training set contains the top-10 most popular event types (*Attack, Transport, Die, Meet, Sentence, Arrest-Jail, Transfer-Money, Elect, Transfer-Ownership, End-Position*) as seen types, and the remaining 23 types are selected as unseen types, which are included in the test set. In the few-shot learning, n examples of event mention features per class are moved from the test to the training set, where n is set to 2. We use P (Precision), R (Recall), F1-score and H (Harmonic mean) as performance metrics. Note that:

$$\mathrm{P} = \frac{TP}{TP + FP}, \quad \mathrm{R} = \frac{TP}{TP + FN}, \quad \mathrm{F1} = \frac{2PR}{P + R}$$

When evaluating the quality of a multi-classification task, TP is # of true positives, FP is # of false positives and FN is # of false negatives.

$$\mathrm{H} = \mathbb{E}\left(\sum_{i} acc_i\right),$$

where acc_i is the accuracy that samples ofi^{th} event type are predicated correctly in the given sequence.

4.2. Ablation Study

In this section, we analyze crucial building modules in the proposed model by disabling each of them, respectively.

fVAE is the baseline only using β-VAE without cross-aligned, distribution-aligned and adversarial training.

CA-fVAE,**DA-fVAE**, **CADA-fVAE** are the baseline models usingβ-VAE with cross-aligned, distribution-aligned and both of them.

CADA-fVAE-GAN combines WGAN with CADA-fVAE to improve the quality of latent representations constructed by VAE framework.

We can draw conclusions from Table 1 that both cross-aligned and distribution-aligned improve the performance. The cross-alignment works better than distribution-alignment (36.21% vs. 31.59% on F1-score, 49.52% vs. 46.39% on H), and more outstanding results are produced by compositing two tricks. Moreover, the introduction of GAN further improves the performance. Compared with CADA-fVAE, our CADA-fVAE-GAN increases the test results by 0.13% on P, 6.78% on R, 12.67% on F1 and 2.38% on H. Ablation study shows the adversarial training leads the encoder of VAE to producing higher-quality latent representations by improving the VAE decoder directly, under the restraint of VAE objective function.

Table 1. Results of ablation study.

	P (%)	**R (%)**	**F1 (%)**	**H (%)**
fVAE	31.06	40.08	24.96	40.10
CA-fVAE	39.92	48.26	36.21	49.52
DA-fVAE	39.58	45.72	31.59	46.39
CADA-fVAE	42.90	45.54	37.84	52.04
CADA-fVAE-GAN	**43.21**	**52.36**	**50.51**	**54.42**

4.3. Model Comparision

In this section, we show that the proposed few-shot learning model achieves comparable performance with existing transfer learning framework for ED.We compare our method with the following baseline:

Transfer: [7] design a transferable architecture for event extraction, using CNN to generate vector representations for the event mention and event type structure. The top-10 most popular event types in ACE2005 chosen by us as *seen* types are the same as [7]. Table 2 shows the performance.

Table 2. Event trigger classification performance on unseen ACE2005 event types.

	P (%)	**R (%)**	**F1 (%)**	**H (%)**
Transfer	**75.50**	36.30	49.10	-
CADA-fVAE-GAN	43.21	**52.36**	**50.51**	54.42

CADA-fVAE-GAN exploits latent representations encoded by VAE, which is proved to be comparable with CNN representations generated by **Transfer**. Conclusions can be drawn from the above results that VAE+GAN could be used to generate features for ED task in transfer learning situations.

5. CONCLUSIONS

In this paper we propose a few-shot event detection model named CADA-fVAE-GAN, which introduces adversarial training to variational autoencoders (VAE). To improve the performance of VAE, cross- and distribution alignment are exploited. With cross-aligned latent distributions and reconstructions, latent representations are enriched by more interpretable semantic meaning. Moreover, adversarial training provided by WGAN strengthens VAE encoder indirectly. Experiments conducted on ACE2005 dataset demonstrate the transferability of low-dimension latent semantic knowledge constructed by VAE and the effectiveness of adversarial training.

Future scope of the research is suggested to be focused on generalization improvements. Specifically, few-shot learning of event detection can be advanced to zero-shot, generalized few-shot and generalized zero-shot, which are of more practical value

6. ACKNOWLEDGEMENT

This research is supported by the grants from the National Key Research and Development Program of China (No. 2019YFB1705601)

REFERENCES

[1] Frome, Andrea, et al. "DeViSE: A Deep Visual-Semantic Embedding Model." neural information processing systems (2013): 2121-2129.

[2] Lampert, Christoph H., Hannes Nickisch, and Stefan Harmeling. "Learning to detect unseen object classes by between-class attribute transfer." computer vision and pattern recognition (2009): 951-958.

[3] Reed, Scott, et al. "Learning Deep Representations of Fine-Grained Visual Descriptions." computer vision and pattern recognition (2016): 49-58.

[4] Zhang, Li, Tao Xiang, and Shaogang Gong. "Learning a Deep Embedding Model for Zero-Shot Learning." computer vision and pattern recognition (2017): 3010-3019.

[5] Gao, T., Han, X., Xie, R., Liu, Z., Lin, F., Lin, L., & Sun, M. (2020). Neural Snowball for Few-Shot Relation Learning. In AAAI (pp. 7772-7779).

[6] Gao, Tianyu, et al. "Hybrid Attention-based Prototypical Networks for Noisy Few-Shot Relation Classification." national conference on artificial intelligence (2019): 6407-6414.

[7] Huang, Lifu, et al. "Zero-shot transfer learning for event extraction." arXiv preprint arXiv:1707.01066 (2017).

[8] Socher, Richard, et al. "Zero-Shot Learning Through Cross-Modal Transfer." neural information processing systems (2013): 935-943.

[9] Tsai, Yaohung Hubert, Liangkang Huang, and Ruslan Salakhutdinov. "Learning Robust Visual-Semantic Embeddings." international conference on computer vision (2017): 3591-3600.

[10] Mukherjee, Tanmoy, Makoto Yamada, and Timothy M. Hospedales. "Deep matching autoencoders." arXiv preprint arXiv:1711.06047 (2017).

[11] Schonfeld, Edgar, et al. "Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders." computer vision and pattern recognition (2019): 8247-8255.

[12] Xian, Yongqin, et al. "F-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning." computer vision and pattern recognition (2019): 10275-10284.

[13] Yu, Xianwen, et al. "VAEGAN: A Collaborative Filtering Framework based on Adversarial Variational Autoencoders.." international joint conference on artificial intelligence (2019): 4206-4212.

[14] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016, June). Autoencoding beyond pixels using a learned similarity metric. In International conference on machine learning (pp. 1558-1566).

[15] Zhu, Yizhe, et al. "A Generative Adversarial Approach for Zero-Shot Learning from Noisy Texts." computer vision and pattern recognition (2018): 1004-1013.

[16] Lu, Yaojie, et al. "Distilling Discrimination and Generalization Knowledge for Event Detection via Delta-Representation Learning." meeting of the association for computational linguistics (2019): 4366-4376.

[17] Wang, Xiaozhi, et al. "Adversarial Training for Weakly Supervised Event Detection." north american chapter of the association for computational linguistics (2019): 998-1008.

[18] Hong, Yu, et al. "Self-regulation: Employing a Generative Adversarial Network to Improve Event Detection." meeting of the association for computational linguistics (2018): 515-526.

[19] Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M., ... & Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. international conference on learning representations.

[20] Givens, Clark R., and Rae Michael Shortt. "A class of Wasserstein metrics for probability distributions." Michigan Mathematical Journal 31.2 (1984): 231-240

[21] Adler, Jonas, and Sebastian Lunz. "Banach Wasserstein GAN." neural information processing systems (2018): 6755-6764.

[22] Ahn, D. (2006, July). The stages of event extraction. In Proceedings of the Workshop on Annotating and Reasoning about Time and Events (pp. 1-8).

[23] Ji, Heng, and Ralph Grishman. "Refining Event Extraction through Cross-Document Inference." meeting of the association for computational linguistics (2008): 254-262.

[24] Gupta, Prashant, and Heng Ji. "Predicting Unknown Time Arguments based on Cross-Event Propagation." meeting of the association for computational linguistics (2009): 369-372.

[25] Riedel, Sebastian, Limin Yao, and Andrew Mccallum. "Modeling relations and their mentions without labeled text." european conference on machine learning (2010): 148-163.

[26] Nguyen, Thien Huu, and Ralph Grishman. "Event Detection and Domain Adaptation with Convolutional Neural Networks." international joint conference on natural language processing (2015): 365-371.

[27] Chen, Yubo, et al. "Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks." international joint conference on natural language processing (2015): 167-176.

[28] Nguyen, Thien Huu, and Ralph Grishman. "Modeling Skip-Grams for Event Detection with Convolutional Neural Networks." empirical methods in natural language processing (2016): 886-891.

[29] Sha, Lei, et al. "Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[30] Duan, Shaoyang, Ruifang He, and Wenli Zhao. "Exploiting Document Level Information to Improve Event Detection via Recurrent Neural Networks." international joint conference on natural language processing (2017): 352-361.

[31] Zhao, Yue, et al. "Document Embedding Enhanced Event Detection with Hierarchical and Supervised Attention." meeting of the association for computational linguistics (2018): 414-419.

[32] Liu, Yingchi, et al. "Document Information Assisted Event Trigger Detection." international conference on big data (2018): 5383-5385.

[33] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[34] Peters, Matthew E., et al. "DEEP CONTEXTUALIZED WORD REPRESENTATIONS." north american chapter of the association for computational linguistics (2018): 2227-2237.

[35] Zhang, Junchi, et al. "Extracting Entities and Events as a Single Task Using a Transition-Based Neural Model." international joint conference on artificial intelligence (2019): 5422-5428.

[36] Sims, Matthew, Jong Ho Park, and David Bamman. "Literary Event Detection." meeting of the association for computational linguistics (2019): 3623-3634.

AUTHORS

Xiaoxiang Zhu is Master's Degree in reading, at the Department of computer science and Engineering, University of Electronic Science and Technology (UESTC), Chengdu, China. He received the B.S. degrees in UESTC in 2018. His interests focus on distributed storage and computing.

Mengshu Hou is a professor in the school of computer science & engineering at the University of Electronic Science and Technology of China (UESTC). He received the M.S and Ph.D degrees in 2002 and 2005 respectively from the UESTC. His current research interests include distributed database management system, machine learning applied to natural language processing, and big data analytics.

Xiaoyang Zeng is currently a Ph.D. at the Department of computer science and Engineering, University of Electronic Science and Technology (UESTC), Chengdu, China. He received the B.S. degrees in Southwest Petroleum University in 2018, and passed the successive master-doctor program and is studying in UESTC. His research interests focus on natural language processing and text mining.

Hao Zhu is an engineer in the Information Center at the University of Electronic Science and Technology of China (UESTC). He received the B.S and M.S degrees in 2002 and 2006 respectively from the UESTC. His current research interests include management informatization, data visualization, and big data analysis

SEMANTIC MANAGEMENT OF ENTERPRISE INFORMATION SYSTEMS THROUGH ONTOLOGIES

Valentina Casola and Rosario Catelli

Department of Electrical Engineering and Information Technologies (DIETI),
University of Naples Federico II, Naples, Italy

ABSTRACT

This article introduces a model for cloud-aware enterprise governance with a focus on its semantic aspects. It considers the need for Business-IT/OT and Governance-Security alignments. The proposed model suggests the usage of ontologies as specific tools to address the governance of each IT/OT environment in a holistic way. The concrete utilization of ISO and NIST standards allows to correctly structure the ontological model: in fact, by using these well-known international standards it is possible to significantly reduce terminological and conceptual inconsistencies already in the design phase of the ontology. This also brings a considerable advantage in the operational management phase of the company certifications, congruently aligned with the knowledge structured in this manner. The semantic support within the model suggests further possible applications in different departments of the company, with the aim of evaluating and managing projects in an optimal way, integrating different but important points of view of stakeholders.

KEYWORDS

Cloud, Enterprise, Governance, Information management, Ontology, Semantic systems

1. INTRODUCTION

For most organisations, data and the technology that supports this data represent their most precious assets but also most underestimated. Information Technology (IT) is on the whole considered as an utility of a corporation [5] and its inherent intricacy requests the introduction of assorted assessment, management, and governance models and therefore led to the explosive growth of the discipline of information systems (IS) research throughout the last thirty years [28]. Such models aim to be each typically applicable and capable to handle specific corporate issues, e.g., cloud manufacturing-based product life cycle management [27]. The requirement for assurance regarding the worth of IT, the management of IT-related risks and augmented needs for management over data are currently considered strategic and creation of value possible through their alignment [32]. Value, risk and management are at the heart of IT governance, whose responsibility lies with both management and the board of directors, and consists of leadership, organisational structures and processes that ensure that the company's IT supports and extends the organisation's strategies and objectives .

The increasing complexity of IT is linked to a twofold aspect: on the one hand, IT becomes increasingly pervasive within companies, embracing both their core business and staff services, and on the other hand, the way in which IT services are delivered becomes increasingly intricate and difficult to manage (just think of the gradual transition from on-premise software to that

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 25-40, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101403

provided in the cloud, which in turn can be IaaS, PaaS, Saas and so on). Therefore, meeting the increasingly restrictive demands coming from the various business areas is challenging and will affect the time needed to deliver the services (for example, to verify the security according to corporate policies) and the costs related to them (the assessment of which becomes increasingly difficult).

For instance, to address queries associated with price and scaling capability, several organisations are considering driving their IT from resource-based approach to service-based approach, with the flexibility to scale the IT capability up and down as requested, that is a cloud based computing approach. Starting with the concept of economies of scale, the sharing of converging resources and infrastructures is at the heart of the concept of cloud computing. In order to obtain a shared but secure environment it is necessary to apply a correct governance strategy based on a model capable of taking into account the increasingly numerous forms through which cloud services are provided, in full respect of the needs of stakeholders, customer contracts and regulatory, legal and privacy aspects, perhaps unifying IT management and governance models [3].

This poses specific challenges on enterprise information systems. To enhance interoperability among all enterprise stakeholder and in order to avoid any misunderstanding, it is needed to be very careful at the semantic level to get clear red from company level right down to technical level and to provide clear objectives. For these reasons, our work explores the possibilities related to the creation and improvement of a domain ontology building methodology whose aim is to bright words and terminologies used among enterprises employees and information systems.

The remainder of the work is structured as follows: we summarise background and state of the art in Section 2 and Section 3 respectively, than we illustrate our methodology, an use case and some considerations in Section 4, finally we outline conclusion and future work in Section 5.

2. BACKGROUND

The cloud paradigm is progressively transforming into a mainstream paradigm and is considered as a serious subject of analysis in computer science. As a result, “cloud computing” is becoming a watchword within the company. The recognition of digital devices therefore the current use of the web translates into an ever-increasing demand for cloud computing. Cloud computing allows huge economies of scale in IT service delivery, but together it faces a variety of challenges. Benefits that are primarily related to it include rapid deployment, pay-per-use, lower prices, scalability, rapid provisioning, fast elasticity, ubiquitous network access, increased resiliency, protection against network attacks, low-cost disaster recovery and data storage solutions, on-demand security checks, real-time detection of system intrusions and rapid restart of services. Therefore, the move to cloud services makes users more efficient, facilitates collaboration with their colleagues, and provides continuous access to alternative digital services [19]. However, cloud applications, like all alternative technologies, face several sensitive issues related to the risks they introduce into the business [36].

For example, moving information to the cloud while offering great convenience for users because they do not have to worry about the complexity of direct management of storage infrastructure hardware [33], the Data Storage as a Service (DSaaS) introduces many challenges to information security (e.g. CIA) that should be addressed, and historically, information security issues are the domain of IT governance.

Although cloud services share infrastructures to produce compliant and guaranteed IT services, IT governance is required to ensure:

- governance framework setting and maintenance;
- benefits delivery;
- risk optimisation; – resource optimisation; – stakeholder engagement.

In order to meet these IT governance requirements in a cloud-based environment, it is always necessary to provide a complete asset inventory of all objects from completely different business departments, and related to your technology (servers, software, switches, etc..), your infrastructure (rooms, buildings, racks, air conditioning systems, power supply units, etc..), and your organisation (service level agreements, contracts, data, suppliers, people, roles and responsibilities, organisational units, etc..), and their relationships and inter dependencies.

Within the space of analysis of data systems and also in management science it is a typical approach to try to identify a model that correctly represents a specific problem and then use the model in order to produce relevant recommendations. A model that aspires to capture the complexity of contemporary computing and data management should apply a holistic view and at the same time build on existing best practices within specific areas (e.g., economic, legal, political and technological), and should therefore address these aspects:

- a medium-sized organisation generally has thousands of technical and infrastructural objects to inventory;
- the model should be able to offer data regarding the possession of objects, also considering possible interrelationships (e.g. the information element X is a component of document Y which is maintained by department Z);
- the model should jointly offer information regarding regulatory compliance and therefore think of a multifaceted reading (for example, regulatory needs could rely on the legal entity of the service provider and/or the location of a specific infrastructure element, the characteristics and/or possession of the information element).

3. State of the Art

This section presents the required definitions and concepts in the areas of cloud governance and enterprise architecture management.

3.1. Cloud Governance

Cloud governance is a natural extension of IT governance [25], but to date there are mainly two approaches used by the industry to manage cloud services. The first considers cloud providers as common service providers aiming to manage them with typical approaches adopted for non-cloud service providers. This approach involves that part of the industry expects these cloud providers to acquire the role of globally reliable mediators for the type of service provided, as ascertained by [21]. This implies an increasing demand for specific certifications (e.g. ISO 9001, ISO 27001) as a guarantee of quality and security for the services offered. The second approach instead, more pointed towards internal management, aims to enhance IT governance by making it ”cloud-aware”, shifting the focus towards the adoption of IT governance frameworks, such as ITIL and COBIT, that are sufficiently updated to keep taking into account the disruption of the cloud world and that can act as a gateway to its better integration within the corporate world.

3.2. Enterprise Architecture Management

Enterprise Architecture Management (or EAM) is a "management practice that establishes, maintains and uses a coherent set of guidelines, architecture principles and governance regimes that provide direction and practical help in the design and development of an enterprise's architecture to achieve its vision and strategy" [1] so it aims to model all relevant components of a corporation and their relationships with many objectives.

The use of a holistic model that includes IT governance and cloud aware EAM methods leads to some advantages:

- organisations can have compelled to maintain a distinctive inventory and a rule set to confirm compliance with several normative and standard requirements;
- it is a viable and simple way to establish strong and resilient cloud governance;
- greater focus on corporate goal that reduces the conflict among heterogeneous stakeholder teams because the right formalisation guarantees traceable and repeatable results, facilitating the division of labour among the stakeholder teams [13] [14].

On the other hand, a "model" approach also has disadvantages:

- it depicts the enterprise architecture as a snapshot in time, not offering reiterative process support for future architecture solutions and tests against different scenarios;
- it is prohibitively time-intensive to keep updated and leaves too much room for error as changes to the architecture occur unchecked and isolated in the heads of small groups of architecture specialists.

To bring the highly distributed knowledge of the contributing stakeholders should be a main objective of EAM. For this reason, successful enterprise architecture programs are approached from a management perspective as opposed to a modelling perspective, and planning tools should support not only the modelling of architecture, but also the creation of roll-out and implementation plans for continuous improvement over time. In this way is possible the support of collaboration in a wide group of stakeholders from both business and IT (C-level, IT strategists, planning teams, technology implementer and business analysts) who contribute to the EA management and planning process. In this way EAM will support sustainable business strategy realisation.

4. A Novel Domain Ontology Building Methodology

Our methodology tries to sketch the optimal way to ensure the best possible solution to domain ontology building problem. Despite several methods are improving their capabilities to find and describe with enough generality high level domain concepts, building the so-called upper ontologies (e.g. SUMO, BFO, CCO and so on), there is a void we need to fill in to overcome limitations left behind (or below from the point of view of an ontologist). The main reason is that the ontology building problem, whatever you say, is a domain and specific problem that arise from the bottom, from the need to give an answer to a question in your specific research field and, in that moment, it shows itself like an instrument. An ontology is firstly useful when it is able to clearly define a common vocabulary for researcher who need to share information in a specific domain. Secondly, it includes machine-interpretable definitions of basic concepts in the domain and relations among them [18]. Nonetheless, there are always several viable alternatives to model a domain, but the specific implemented method is too often left apart and only the final ontology is shown. The development of an ontology can be divided in 7 steps [18]:

1. determine the domain and scope of the ontology;
2. consider reusing existing ontologies;
3. enumerate important terms in the ontology;
4. define the classes and the class hierarchy;
5. define the properties of classes - slots;
6. define the facets of the slots;
7. create instances.

Our methodology suggests some practical hints about these steps. First and foremost, the brainstorming activity around the first step is not trivial. Great importance should be done to the definition of the limits of you work to avoid any bias towards sub fields of your domain which are not the focus you had in mind at the beginning. The same approach should be applied looking into existing ontologies: the reuse can be more useful if you are able to extract concepts and terms from .owl files and understand if a smarter way to use them is possible. A chance not considered until now and proposed by us consists in leveraging something existing to enumerate important terms in the ontology: ISO always defines terms and definitions vocabularies across its operating domains, although as we will see, some incongruities must be fix. ISO vocabularies give us an important advantage through the path of the ontology development: they avoid to us to forget the overview out of our domain (so we will be ready to create links and bridges to the external domains) and strengthen the initial audit activity needed to enumerate terms. Left apart step 7, that is mostly practical, step 4 to 6 are really intertwined. In out method we suggest a multi-dictionary approach to get out of them: that we name “Ontological Research” consists in a research activity through ISO, NIST and Cambridge Dictionary. On the one hand, it is possible to extract a technical meaning from standards and understand where and why they want to go, but it is on your own to refine their meanings and put them in the correct way. The main problem using only technical standards is due to their circular definition: they usually define something referring themselves to something else and so on, until you are at the beginning again. On the other hand, it is possible to extract a semantic meaning from Cambridge Dictionary to mix and match what standards say, trying to solve their circularity. To build a working ontology, the knowledge engineer must understand where is needed to stop circularity and find the best elementary definition: it is true that this could be extremely subjective, however standards and vocabularies can lead you and help you through the path. It follows an illustrative figure of our methodology in Figure 1.

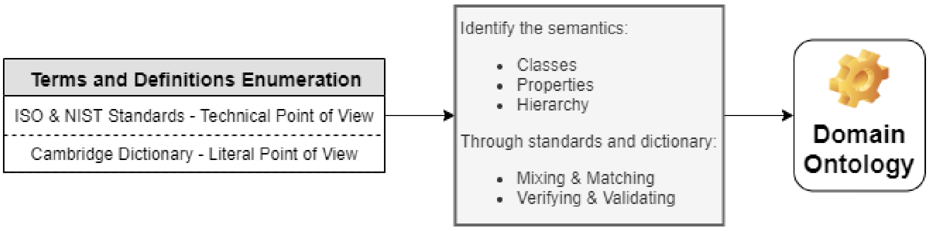

Terms and Definitions Enumeration
ISO & NIST Standards - Technical Point of View
Cambridge Dictionary - Literal Point of View
Identify the semantics:
Classes
Properties
Hierarchy
Through standards and dictionary:
Mixing & Matching
Verifying & Validating
Domain Ontology

Figure 1. Overview of our methodology

4.1. Use Case: Information Security Domain ontology

We decided to focus our work on the Information Security domain, leveraging the terms and definition from ISO/IEC 27000:2018. Because of the novelty of the method, we decided to avoid the reuse of any existing ontology to test thoroughly the possibilities of our method. We have identified “Process” as high-level concept and decided to build a class around it. ISO/IEC 27000:2018 (and also NISTIR 8053) defines a process as a “set of interrelated or interacting activities which transforms inputs into outputs”: for this reason we decided to include two related operations, receivesInput() and providesOutput() functions. After that, we have move on the "Activity" concept: ISO/IEC/IEEE 15288:2015 defines an activity as a "set of cohesive tasks of a process" so we were addressed to find out the meaning of the "Task" concept. Here several drawbacks arose because there are at least three conflicting definitions as follows:

- ISO/IEC/IEEE 15288:2015 states that a task is "required, recommended, or permissible action, intended to contribute to the achievement of one or more outcomes of a process";
- ISO 9241-11:2018 states that a task is "set of activities undertaken in order to achieve a specific goal";
- NIST SP 800-181 states that "a task is a specific piece of work that, combined with other identified tasks, composes the work in a specific specialty area or work role".

Which ones could be acceptable? Why? As highlighted in Figure 2 there were several problems to solve, so we started analysing "Task" definitions one by one:

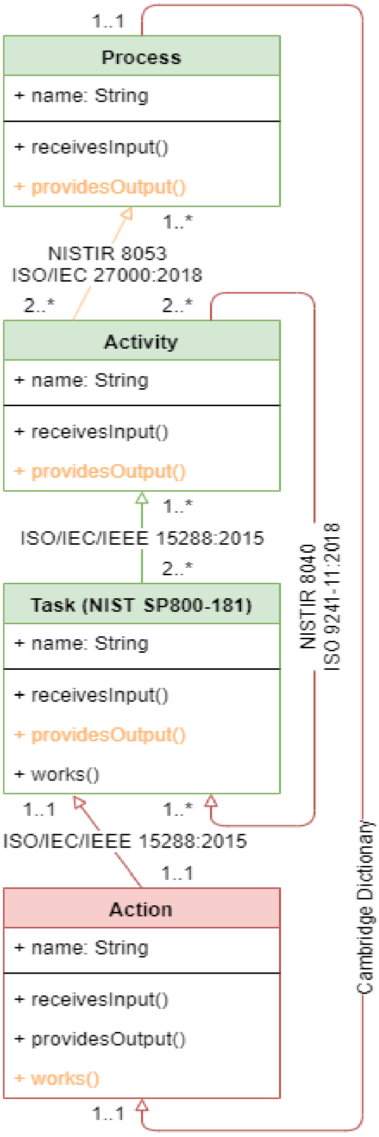


Figure 2. UML-based representation of "Process" class creation

- ISO/IEC/IEEE 15288:2015 leverages the concept of “Action” and speaks about the “Outcome” of a process (is it the same of “output” indicated by ISO/IEC 27000:2018?). Unfortunately, we did not find any standards that define the concept of “Action”. But Cambridge Dictionary does: it defines the concept of “Action” as “the process of doing something, especially when dealing with a problem or difficulty”. This suggested us to overcome this standard definition of “Task” in order to avoid circularity;
- ISO 9241-11:2018 creates circularity, using the already defined “Activity” concept, so we decide to overcome this definition also (but what is a “goal”?);
- NIST SP 800-181 gives us an enough atomic definition of the concept of “Task”, identifying it as “a specific piece of work”. Here, the boundaries of our ontology impose us not to go deeper to avoid any lack of focus on the main topic that is “Information Security”.

Nonetheless, the concept of “Action” defined by ISO/IEC/IEEE 15288:2015 inspired us to search for “Outcome” concept. Our first thought was related to the same concept of “Process”: its ISO definition assumes quantitative characteristic, promoting an ungluing from the industrial logic where processes are used to “create value” and not only more objects. Standards don’t define “Output”, “Outcome”, “Result” and “Effect” so we used Cambridge Dictionary:

- output is “an amount of something produced by a person, machine, factory, country, etc...” and this confirmed our first thought was not so wrong;
- outcome is “a result or effect of an action, situation, event, etc...”;
- result is “something that happens or exists because of something else”;
- effect is “the result of a particular influence”.

And then we also searched for “Goal” and surprisingly we found it in NISTIR 8040 – ISO 9241-11:2018 defined as “intended outcome”. Clearly, it was something missing. To promote the “Process” to a qualitative characteristic we decided to significantly change one of its operation: providesOutput() became deliversResults(). The semantic agreement among all the others concept was found also adding these definitions:

- Cambridge Dictionary states the a “situation” is “the set of things that are happening and the conditions that exist at a particular time and place”;
- ISO Guide 73:2009 states that an “event” is an “occurrence or change of a particular set of circumstances”.

Subsequently we decide to treat “Action”, “Situation” and “Event” as object of type “Process”, “Effect” as object of type “Result”. The last problem was about “Outcome”. Analysing one of the notes attached to the definition of “Event” we were directed through a possible solution: it states that “An event without consequences can also be referred to as [...]”. Because of the note we have considered a “consequence” like something with negative connotation although the definition of “Consequence” on ISO/IEC 15026 as “effect (change or non-change), usually associated with an event or condition or with the system and usually allowed, facilitated, cause, prevented, changed, or contributed to by the event, condition, or system” could be considered neutral, while “Outcome” (linked to “Goal” also) like something with positive connotation. Hence, we added two operations to our “Result” class, “hasOutcome” and “hasConsequence”, inherited by that particular “Result-object” that is “Effect”. A sorted “Process” class is represented in Figure 3.

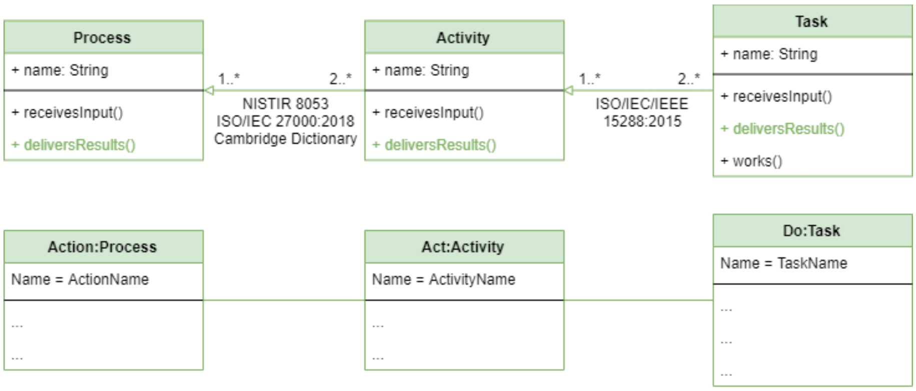


Figure 3. UML-based representation of Process class and related Action-object

4.2. Considerations Over Modern EAM Challenges

In proposing such a model of governance, we have to face several challenges that we could summarise as follows:

1. reduction of the points of view;
2. partiality of the points of view;
3. integration of the EAM within the processes.

The first problem is the identification of points of view that we could consider relevant and that are usually related to stakeholders. But in doing so, we risk excessively narrowing the field of vision: in fact, it is necessary to take into account the points of view of the application situations on which we plan to map the model. This step is fundamental in order to balance the reduced capacity of stakeholders to precisely outline their needs and use cases: this limit is often one of the main reasons behind the failure of a business project. Application situations have a direct impact on the layers needed to build the EAM model, its inter dependencies and the level of granularity of its requirements. For example, if we were to focus on a ”business” perspective, within a model we would certainly find structural objectives, processes and IT elements, but the granularity with which the latter would be defined would certainly be inadequate when we would be from an ”IT” perspective, more focused on the analysis of hardware/software systems and their performance. And even more, this limit would tend to emerge if we put ourselves from an “Information Security” perspective, so that the elements at stake must guarantee certain levels of confidentiality, integrity and confidentiality. Therefore, the identification of relevant points of view and their objectives is a precondition for the correct definition of the desired elements, interrelationships and granularity.

Secondly, care must be taken not to consider only part of the perspectives mentioned above. In fact, what happens in the real world in an ascertained way is due to ”historical” reasons that lead companies to partial problem-solving (think, for example, of the necessary integration of legacy tools within processes that are the subject of the Information Security perspective). In such cases, the challenge for companies is to be able to link these systems together in a coherent way to the perspective considered in order to make them easily manageable.

The third point to consider in order to maintain a consistent EA is to ensure that it is integrated into relevant processes such as change management and the tools needed to manage the data of the same. It may seem trivial, but hold-up or lock-in effects due to the pre-existing (or proprietary) tools perfectly run within the company routines are common and can inhibit the acceptance of new tools and approaches proposed. If the degree of acceptance of what is new is low for this, then the full achievement of the benefits proposed by the new EAM will fail. In particular, the migration of data from legacy tools to new ones often leads to inconsistencies, losses or contradictions, all the greater as the different elements, interrelations and granularity at stake from different perspectives are different. Solving these types of problems requires a very significant effort in terms of human resources. But using semantic tools it is possible, while confining each perspective only to the data relevant to it, to avoid information overflow and to be able to make more informed decisions, also improving access control both for the analysis of the functions and for the editing of the data.

In the following, a novel way to manage EA models is introduced. It aims to overcome the restrictions mentioned using semantic technologies. The ontologizing of The Enterprise Ontology can help to redefine "the notion of architecting" as stated in [16]:

"In the context of high levels of complexity and uncertainty, the notion of causality often breaks down. Often, one can only assume that everything is in relationship with everything else. Consequently, understanding the ramifications of changes such as disruptive technologies and new architecture models (i.e. cloud computing, outsourcing) is often almost impossible. New resources such as contextual data of customers will have to be used effectively to gain a competitive edge. To face such challenges, the notion of architecting will surely have to be redefined.".

The following conceptual model tries to overcome the above restrictions and is based on a matrix containing six horizontal and three vertical layers (see Figure 4).

The six horizontal layers are:

1. Business Level
2. Integration Level
3. Core Level
4. Staff Level
5. Infrastructure Level
6. IT/OT Level

The three vertical levels are:

1. Information Security
2. Computer and System Security
3. Network and Physical Security

Horizontal layers provide the general paradigm of alignment between company and IT/OT ensuring consistency between business objectives, operations and IT infrastructure. The vertical layers provide the general paradigm of alignment between security, compliance and governance from the point of view of information objects and specific requirements, thus roles and responsibilities. Below is an overview of the main aspects of the model, which will form the basis for the elaboration of the semantic aspects of the model, which will be presented in the next section.

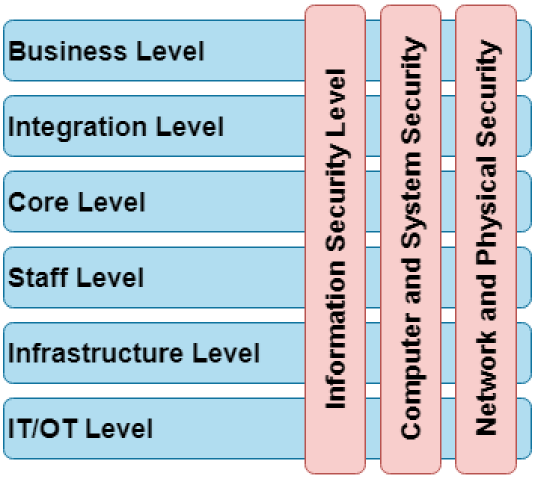

Business Level
Integration Level
Core Level
Staff Level
Infrastructure Level
IT/OT Level
Information Security Level
Computer and System Security
Network and Physical Security

Figure 4. Overview of the EAM model

4.2.1. The Paradigm of Business-IT/OT Alignment

The paradigm of alignment between Business and IT/OT is implemented in six horizontal levels.

1. Business Level.

It defines the company's global focus, including its mission and strategy, as well as its business model.

2. Integration Level.

It defines the final products to be presented on the market at regional level (EMEA, APAC, etc...), interfacing with the Core Level (of which it defines internal contracts and service levels) and orienting its efforts in order to respond to the needs traced by the Business Level.

3. Core Level.

It deals with the coordination of the Core Departments, which are the organisational units at the heart of the corporate mission and defines the organisational and technical needs.

4. Staff Level.

It provides support to the Core Level, integrating the non-core business functions in a harmonic and ready-to-use way.

5. Infrastructure Level.

It allocates the available resources (software, hardware and network from an IT point of view; but also buildings, local air conditioning, energy and physical access systems) to the staff units as needed (cloud, on premise).

6. IT/OT Level.

It defines the available resources, IT and OT, and manages their organisation (e.g. nodes, type and degree of virtualisation, components, connections and barriers).

4.2.2. The Paradigm Of Governance-Security Alignment

The Governance-Security alignment is embodied within the three vertical levels.

1. Information Security Level.

It establishes roles, responsibilities and accountability with regard to data, information and requirements like relevancy of standards (e.g. ISO 27001), specific governance necessities, yet as business or application-driven necessities (e.g. confidentiality, integrity, availability, reliability, dependability).

2. Computer and System Security Level.

It defines security technical needs (user accesses, disk encryption, etc.) on the basis of internal necessities and external requests as pointed out by Information Security Level. It evaluates and manages cloud computing doable deployment models within the company.

3. Network and Physical Security Level.

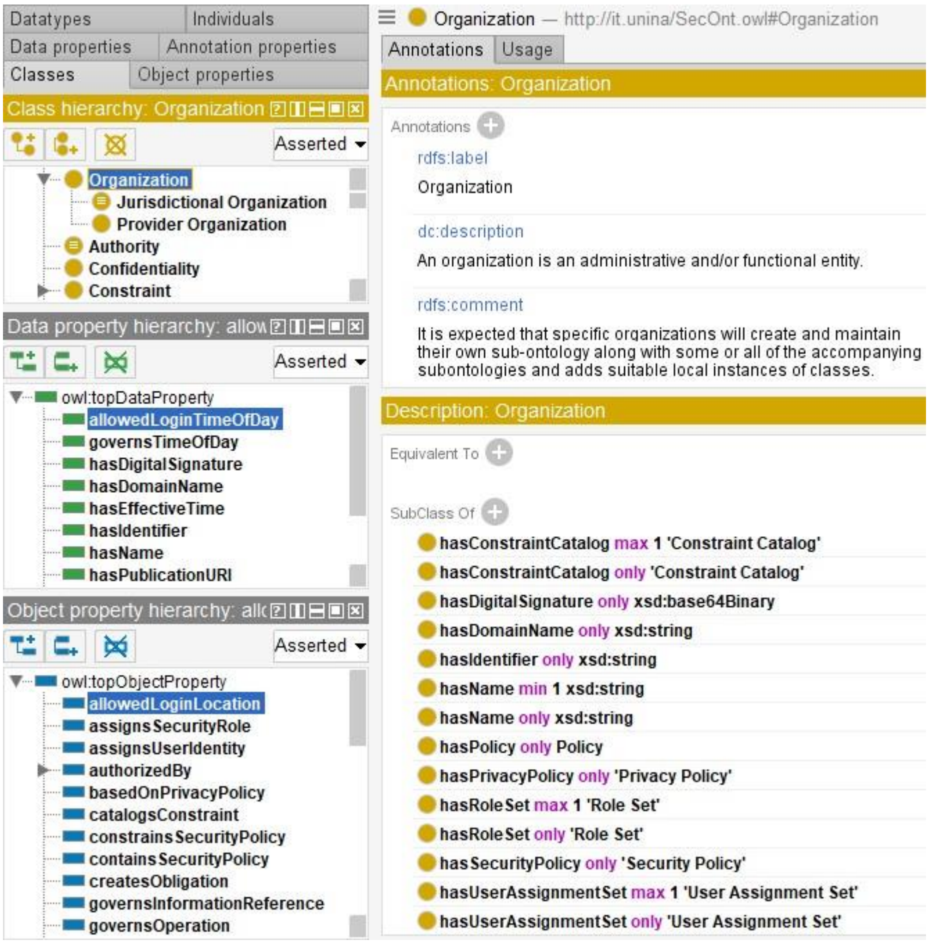
It defines the security of the perimeter within which each asset operates on the basis of internal necessities and external requests as pointed out by Information Security Level, but can act independently if necessary in case of emergency to increase the speed of response.

4.3. Ontologized Eam Model

Ontologies are a means of formally shaping the structure of a [26]. They provide a shared understanding of certain domains that can be communicated between people and application systems [8]. Ontologies aim to determine “semantic agreements”, reducing language ambiguity and knowledge variations between agents, which can cause errors, misunderstandings and inefficiencies [2]. Given their importance, ontologies are seen as the cornerstone of many promising technologies such as, for example, the semantic web and related data, reporting an abundant implementation in literature like [9],[22]. The scope of IT/OT service management is very far from this trend, although there are several attempts to use ontologies in some areas such as the life cycle of cloud services [15], software system development and IT service management processes [30], quality of service - security metrics [6], facilitation of operational procedures in public administration [24], service management in the Internet of Things [23] or IT service management for business-IT integration [31].

4.3.1. An Ontology-Based Approach

The application of ontologies in the various fields requires careful reflection on the basic mappings between the higher ideas and the application cases. Starting from [31] there is an ontological approach for the establishment of a scientific technique that allows to implement the ontology approach in an extremely easy and well-defined way, thus supporting its use. The reference standard related to the development of ontologies is the web ontology Language (OWL) defined by the World Wide Web Consortium (W3C). The OWL allows the use of various logical formalisms to mechanically process domain information by providing value-added reasoning services to classify individuals, verify the consistency of information bases and deduce new types of information within the taxonomy. The ontologies outlined under the OWL embrace classes as sets of individuals, individuals as examples of classes and properties as binary relationships between individuals. Among the different ontological development environments, the best known is certainly Protege, an open source program that allows both the development of ontologies and their visualisation, although in the case of complex ontologies it is more appropriate to employ specialised tools in visualisation such as OWL-VisMod [10]. The idea of representing a collection of terms as enterprise related ontology has been projected over fifteen years [29] and there are timid attempts to represent corporate governance in semantic ways, particularly in the fields of information security [7] and IT governance [4], hence the application of enterprise design governance in public administration [20]. Of explicit interest is that the pioneering work of [34] in the field of compliance management, which fits well as a starting point for reflection within the modelling of the Information Security Level. The ontology introduced is based on what is available at the state of the art and seeks to improve and combine it into a single coherent and global ontology for the EAM, extending a number of approaches presented in [34] following an approach supported by [31] and drawing on [17]. An extract of the ontology is shown in Figure 5.


Datatypes
Individuals
Data properties
Annotation properties
Classes
Object properties
Class hierarchy: Organization
Asserted
Data property hierarchy: allow
Asserted
Object property hierarchy: allc
Asserted
Organization — http://it.unina/SecOnt.owl#Organization
Annotations
Usage
Annotations: Organization
Annotations
Description: Organization
Equivalent To
SubClass Of

Organization
Jurisdictional Organization
Provider Organization
Authority
Confidentiality
Constraint

owl:topDataProperty
allowedLoginTimeOfDay
governsTimeOfDay
hasDigitalSignature
hasDomainName
hasEffectiveTime
hasIdentifier
hasName
hasPublicationURI

owl:topObjectProperty
allowedLoginLocation
assignsSecurityRole
assignsUserIdentity
authorizedBy
basedOnPrivacyPolicy
catalogsConstraint
constrainsSecurityPolicy
containsSecurityPolicy
createsObligation
governsInformationReference
governsOperation

rdfs:label
Organization

dc:description
An organization is an administrative and/or functional entity.

rdfs:comment
It is expected that specific organizations will create and maintain their own sub-ontology along with some or all of the accompanying subontologies and adds suitable local instances of classes.

hasConstraintCatalog max 1 'Constraint Catalog'
hasConstraintCatalog only 'Constraint Catalog'
hasDigitalSignature only xsd:base64Binary
hasDomainName only xsd:string
hasIdentifier only xsd:string
hasName min 1 xsd:string
hasName only xsd:string
hasPolicy only Policy
hasPrivacyPolicy only 'Privacy Policy'
hasRoleSet max 1 'Role Set'
hasRoleSet only 'Role Set'
hasSecurityPolicy only 'Security Policy'
hasUserAssignmentSet max 1 'User Assignment Set'
hasUserAssignmentSet only 'User Assignment Set'

Figure 5. An extract of the proposed ontology

4.3.2. Verification

The verification of a specific ontology requires two phases: the characterisation of the models of ontology up to isomorphism and the indication that these models correspond to the structures supposed for the ontology [11]. These two phases are often conducted using approaches such as the theory of reducibility [12]. But due to the lack of alternative ontologies based on OWL, it is not possible to conduct a test in the sense of phase two at the time of writing. In the first step we must demonstrate that a theorem concerning the connection between the classes of ontology models and thus the class of the supposed structures is often replaced by a theorem concerning the connection between ontology (a theory) and thus theory by axiomizing the supposed structures [11]. This requires that this axiomatization be already identified. Overall, the approach of [12], [11] represents a possible way to formally verify ontology, but the lack of alternative ontologies makes comparison impossible. For this reason, once we have analysed the most common classification approaches within the IT governance domain, we think of the ISO 27001 standard as sufficiently close for a meaningful comparison, since ISO documents specify roles and responsibilities (present in a part of our model) within a certification scheme.

5. Conclusions and Future Works

The presented model integrates approaches from EAM, IT governance and cloud computing so as to produce a holistic governance framework, with a look at the semantic aspects considering information and knowledge objects, roles, as well as necessities. It also attributes exactly the skills for secure management of cloud environments. The Business-IT/OT alignment builds on existing works within the space of EA and extends them by many necessary aspects, detailing more clearly the layers of the model and introducing pregnant semantic relationships between components of the layers. The aspect of the point of view is considered in more details and the integration of the EAM within the processes is deepened, while the governance paradigm follows best practices of literature. The latter extends them applying more recent approaches providing an additional clear view on the relationships between information objects, roles, and specific necessities from the point of view of application domains and client teams. The presented semantic consideration provides a viable framework for the facilitation of state-of-the-art semantic approaches within the additional development of the model. Provided that previous frameworks are modelled using well-known ontologies, this allows authors to formalise the model in a very structured manner and to arrange it for future automatic reasoning and future ontology verification.

The given model with its paradigms may be a viable approach to fill the gap between the standard views of Business – IT/OT alignment and Governance – Security alignment, on the one hand, and integrating the not-postponeable revolution of the cloud, on the other hand. Thus, once completely developed, it will function as a methodological framework for both cloud-aware company and cloud-suppliers, ensuring an improved synergy between requests and offers.
Yet, the work on the model continues to be in its early stages. The queries of precise mappings and reflections between the various levels of the model are still open. Authors expect that a nonstop focus on semantic aspects and also the application of semantic methods can give a possible path to refine the model in this regard. Beside the work on semantic aspects, more work in the context of the model is targeted on 2 main areas. First and foremost, the doable relationships and also the degree of granularity ought to be developed from the point of view of each relevant perspective. Secondly, the compliance issue must be specified a lot of in-depth and to be extended with the potential to produce specific recommendations supported by both Business and IT/OT levels (e.g., an amendment request as a part of the change management): specific extensions of the approach leveraging COBIT and ITIL frameworks could be considered. Finally, the development of a tool based on semantic technologies will follow closely the model growth.

Acknowledgements

This document has been produced thanks to a doctoral grant funded by Hitachi Rail STS S.p.A. (formerly Ansaldo STS S.p.A.), and in particular by the CSAC (Cyber Security and Assurance Control) Department, to which the authors sincerely express their gratitude for the valuable support provided.

REFERENCES

[1] Ahlemann, F., Stettiner, E., Messerschmidt, M., Legner, C.: Strategic enterprise architecture management: challenges, best practices, and future developments. Springer Science & Business Media (2012)

[2] Blanco, C., Lasheras, J., Fernández-Medina, E., Valencia-García, R., Toval, A.: Basis for an integrated security ontology according to a systematic review of existing proposals. Computer Standards & Interfaces 33(4), 372–388 (2011)

[3] Bounagui, Y., Mezrioui, A., Hafiddi, H.: Toward a unified framework for cloud computing governance: An approach for evaluating and integrating it management and governance models. Computer Standards & Interfaces 62, 98–118 (2019)

[4] Brandis, K., Dzombeta, S., Haufe, K.: Towards a framework for governance architecture management in cloud environments: A semantic perspective. Future Generation Computer Systems 32, 274–281 (2014)

[5] Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation computer systems 25(6), 599–616 (2009)

[6] Charuenporn, P., Intakosum, S.: Qos-security metrics based on itil and cobit standard for measurement web services. J. UCS 18(6), 775–797 (2012)

[7] Ekelhart, A., Fenz, S., Klemen, M.D., Weippl, E.R.: Security ontology: Simulating threats to corporate assets. In: International Conference on Information Systems Security. pp. 249–259. Springer (2006)

[8] Fensel, D.: Ontologies. In: Ontologies, pp. 11–18. Springer (2001)

[9] Garcia-Crespo, A., Colomo-Palacios, R., Gomez-Berbis, J.M., Ruiz-Mezcua, B.: Semo: a framework for customer social networks analysis based on semantics. Journal of Information Technology 25(2), 178–188 (2010)

[10] García-Peñalvo, F.J., Colomo-Palacios, R., Garcìa, J., Theròn, R.: Towards an ontology modeling tool. a validation in software engineering scenarios. Expert Systems with Applications 39(13), 11468–11478 (2012)

[11] Grüninger, M.: Verification of the owl-time ontology. In: International Semantic Web Conference. pp. 225–240. Springer (2011)

[12] Grüninger, M., Hahmann, T., Hashemi, A., Ong, D.: Ontology verification with repositories. In: FOIS. pp. 317–330. No. 209 (2010)

[13] Jonkers, H., Lankhorst, M., Van Buuren, R., Hoppenbrouwers, S., Bonsangue, M.,Van Der Torre, L.: Concepts for modeling enterprise architectures. International Journal of Cooperative Information Systems 13(03), 257–287 (2004)

[14] Jonkers, H., Lankhorst, M.M., ter Doest, H.W., Arbab, F., Bosma, H., Wieringa,R.J.: Enterprise architecture: Management tool and blueprint for the organisation. Information systems frontiers 8(2), 63 (2006)

[15] Joshi, K., Finin, T., Yesha, Y.: Automating cloud services lifecycle through semantic technologies (May 26 2016), uS Patent App. 14/550,264

[16] Lapalme, J., Gerber, A., Van der Merwe, A., Zachman, J., De Vries, M., Hinkelmann, K.: Exploring the future of enterprise architecture: A zachman perspective. Computers in Industry 79, 103–113 (2016)

[17] Mense, A., Blobel, B.: Hl7 standards and components to support implementation of the european general data protection regulation. European Journal for Biomedical Informatics 13(1), 27–33 (2017)

[18] Noy, N.F., McGuinness, D.L., et al.: Ontology development 101: A guide to creating your first ontology (2001)

[19] Park, S.C., Ryoo, S.Y.: An empirical investigation of end-users’ switching toward cloud computing: A two factor theory perspective. Computers in Human Behavior 29(1), 160–170 (2013)

[20] Peristeras, V., Mocan, A., Vitvar, T., Nazir, S., Goudos, S.K., Tarabanis, K.: Towards semantic web services for public administration based on the web service modeling ontology (WSMO) and the governance enterprise architecture (GEA). na (2006)

[21] Petruch, K., Stantchev, V., Tamm, G.: A survey on it-governance aspects of cloudcomputing. International Journal of Web and Grid Services 7(3), 268–303 (2011)

[22] Rodríguez-González, A., Colomo-Palacios, R., Guldris-Iglesias, F., Gómez-Berbís, J.M., García-Crespo, A.: Fast: Fundamental analysis support for financial statements. using semantics for trading recommendations. Information Systems Frontiers 14(5), 999–1017 (2012)

[23] Sammarco, C., Iera, A.: Improving service management in the internet of things. Sensors 12(9), 11888–11909 (2012)

[24] Savvas, I., Bassiliades, N.: A process-oriented ontology-based knowledge management system for facilitating operational procedures in public administration. Expert Systems with Applications 36(3), 4467–4478 (2009)

[25] Stantchev, V., Stantcheva, L.: Extending traditional it-governance knowledge towards soa and cloud governance. International Journal of Knowledge Society Research (IJKSR) 3(2), 30–43 (2012)

[26] Sure, Y., Staab, S., Studer, R.: Ontology engineering methodology. In: Handbookon ontologies, pp. 135–152. Springer (2009)

[27] Talhi, A., Fortineau, V., Huet, J.C., Lamouri, S.: Ontology for cloud manufacturing based product lifecycle management. Journal of Intelligent Manufacturing 30(5), 2171–2192 (2019)

[28] Thomas, O.: Understanding the term reference model in information systems research: history, literature analysis and explanation. In: International Conference on Business Process Management. pp. 484–496. Springer (2005)

[29] Uschold, M., King, M., Moralee, S., Zorgios, Y.: The enterprise ontology. The knowledge engineering review 13(1), 31–89 (1998)

[30] Valiente, M.C., García-Barriocanal, E., Sicilia, M.A.: Applying ontology-based´ models for supporting integrated software development and it service management processes. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(1), 61–74 (2011)

[31] Valiente, M.C., Garcia-Barriocanal, E., Sicilia, M.A.: Applying an ontology approach to it service management for business-it integration. Knowledge-Based Systems 28, 76–87 (2012)

[32] Van Grembergen, W., De Haes, S.: Enterprise governance of information technology: achieving strategic alignment and value. Springer Publishing Company, Incorporated (2009)

[33] Wang, C., Wang, Q., Ren, K., Cao, N., Lou, W.: Toward secure and dependable storage services in cloud computing. IEEE transactions on Services Computing 5(2), 220–232 (2011)

[34] Yip, F., Wong, A.K.Y., Parameswaran, N., Ray, P.: Rules and ontology in compliance management. In: 11th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2007). pp. 435–435. IEEE (2007)

[35] Zarrabi, F., Pavlidis, M., Mouratidis, H., Islam, S., Preston, D.: A meta-model for legal compliance and trustworthiness of information systems. In: International Conference on Advanced Information Systems Engineering. pp. 46–60. Springer (2012)

[36] Zissis, D., Lekkas, D.: Addressing cloud computing security issues. Future Generation computer systems 28(3), 583–592 (2012)

AUTHORS

Valentina Casola is Associate Professor at the Department of Electrical Engineering and Information Technologies of the University of Naples Federico II. She graduated in Electronic Engineering with honors in 2001 and received her PhD in Electronic Engineering in 2004. Since 2005 she has taught several courses at the Faculty of Engineering, including: "Electronic Computers I", "Programming I" and "Secure System Design". Her research activities are both theoretical and practical and mainly concern safety assessment methodologies and design methodologies for secure distributed systems. These activities are carried out in collaboration with other academic institutions and international companies in numerous projects. Valentina Casola is author of numerous publications in journals and in international conferences and is member of program committees of numerous international conferences.

Rosario Catelli is a PhD student at the Department of Electrical Engineering and Information Technologies of the University of Naples Federico II. He started his PhD in Hitachi Rails STS then moved to the Institute for High Performance Computing and Networking (ICAR), which is part of the National Research Council. He is currently working in the field of natural language processing.

INJECTING EVENT KNOWLEDGE INTO PRE-TRAINED LANGUAGE MODELS FOR EVENT EXTRACTION

Zining Yang[1], Siyu Zhan[1], Mengshu Hou[1],
Xiaoyang Zeng[1] and Hao Zhu[2]

[1]School of Computer Science and Engineering, University of Electronic
Science & Technology of China, Chengdu, China
[2]Information Center, University of Electronic
Science & Technology of China, China

ABSTRACT

The recent pre-trained language model has made great success in many NLP tasks. In this paper, we propose an event extraction system based on the novel pre-trained language model BERT to extract both event trigger and argument. As a deep-learning-based method, the size of the training dataset has a crucial impact on performance. To address the lacking training data problem for event extraction, we further train the pre-trained language model with a carefully constructed in-domain corpus to inject event knowledge to our event extraction system with minimal efforts. Empirical evaluation on the ACE2005 dataset shows that injecting event knowledge can significantly improve the performance of event extraction.

KEYWORDS

Natural Language Processing, Event Extraction, BERT, Lacking Training DataProblem

1. INTRODUCTION

One Common task of Information Extraction (IE) is event extraction (EE) which aims to detect whether the text has mentioned some real-world events and if so, classifying event types and identifying event arguments. An example sentence and its event annotation in the ACE2005 [1] dataset has been provided in Figure 1. With the increasing amount of text data, EE is becoming an increasingly important component in many natural language processing (NLP) applications for decision making, risk analysis, and system monitoring.

Deep learning has been proven efficient and obtains the state-of-the-art result for event extraction task. As a kind of supervised learning approach, its performance is highly dependent on the quality and quantity of the training data. Generally, to achieve better performance, a neural network involves more parameters and therefore needs more data to converge without over fitting. However, labeling training data is not only time-consuming and laborious but also requires professional domain knowledge, which limits the size of the available corpus. For example, the ACE2005 corpus only has a total of 599 documents which is a very small quantity for the task to extract 33 predefined events and their arguments with 36 predefined roles.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 41-50, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101404

The common idea of current solutions is data expansion technology, which generates more labeled training data from external corpus and uses both original and generated data for model training. We argue that the data generating method is hard for event extraction because events typically have a complex structure: an event can be mentioned by different triggers, different events have different arguments with different roles. To avoid this problem, instead of generating training data explicitly, we directly use the unlabeled corpus to inject event knowledge into our event extraction system by the novel pre-trained language model, which can be regarded as implicitly expand training data.

Concretely, we first build an event extraction system based on the pre-trained language model to extract both event trigger and event argument as our baseline. And then build an unlabeledevent training dataset from a large corpus which is then being used to further train the language model to inject the event knowledge to the event extraction system. Compared to the baseline, our method achieves approximately 2% improvement for both trigger and argument classification.

The paper is organized as follows. Section 2 presents related works, along with a special focus on pre-trained language model based on which we build our event extraction system with the help of external event corpus in section 3.The event corpus construction details and evaluation settings are introduced in section 4. Section 5 concludes the paper.

Sentence: *Leung* was **hired** by the *FBI* and **paid** almost *$2 million* over *20 years* to spy on the *Chinese*.			
EVENT 0:		**EVENT 1:**	
Event Type	Personnel: Start-Position	**Event Type**	Transaction: Transfer-Money
Trigger	hired	**Trigger**	paid
Arguments	Person: Leung Entity: FBI	**Arguments**	Giver: FBI Money: $2 million Recipient: Leung Time: 20 years

Figure 1. An example sentence of ACE2005 dataset, there are two event mentions: Start-Position event triggered by hired and Transfer-Money event triggered by paid. Each event has some entities (underlined words or phrases) as its arguments with specific role.

2. RELATED WORK

2.1. Event Extraction

A variety of methods have been used for event extraction task. The pattern matching technique manually constructing event patterns with the help of professional knowledge. [2] and [3] are very early and typical pattern-based extraction system. Traditional feature-based machine learning algorithms are also widely used for event extraction task. These approach first extract feature from training text to train classifiers, then applying the classifiers for new text. [4] formulate the event extraction as a structured learning problem, and proposed a joint extraction algorithm integrating local and global features into a structured perceptron model to predict triggers and arguments simultaneously. [5] proposed a cross-entity event extraction model that exploited utilize global information as global features together with sentence-level features to train classifier. Recently, neural based deep learning method is becoming mainstream for event extraction. Deep learning can help to reduce the difficulties of feature engineering. Benefit from the well-designed network structure and the depth of network layers, it can typically achieve better performance than traditional machine learning algorithms. DMCNN [6] utilize a variant of convolution neural network called dynamic multi-pooling CNN to extract features and event automatically. JRNN [7] adopts bidirectional recurrent neural network (RNN) to jointly extract event trigger and arguments. JMEE [8] propose an event extraction framework that extract features using bidirectional long short-term memory (LSTM) networks, and capture the global relationship by graph convolutional network (GCN) with attention mechanism.

A large and growing body of literature has investigated how to improve the extraction accuracy from a small set of labeled dataset. Utilize the bootstrapping [9] and active learning strategy [10] is challenging for event extraction as it is hard to evaluate the classification confidence for the generated event structure. Some methods expand data from knowledge bases (KBs, such as FrameNet [11][12][13], WordNet [14]) based on a set of hypotheses which is complicated and hard to cover the many different types of events.

2.2. Pretrained Language Model

Pre-trained language models have made great success in recent years and been a standard part of many NLP tasks. It adopts a two stages strategy: pre-trained on the massive unlabeled corpus to learn general contextualized representations with linguistic information of language and then fine-tune on a specific downstream task. For downstream tasks, pre-trained language model can be regarded as an encoder that encodes each token of the original text into a vector with contextual and semantic information which has been proved to be very effective and helpful to the downstream task. The Generative Pre-trained Transformer (GPT) [15] by OpenAI builds a unidirectional language model (LM) based on the transformer and firstly introduces the fine-tuning approach. Bidirectional Encoder Representations from Transformers (BERT) [16]overcome the unidirectionality constraint through a new training object called mask language model (MLM) and introduce the next sentence prediction (NSP) training object to obtain sentence representation.

The BERT language model is pretrained using the general English corpus, while the downstream tasks usually require some task-specific knowledge. However, very little research has been done to solve this domain mismatch problem. BioBERT [17] and SciBERT [18] shows pre-training with in-domain data are very efficient for biomedical and science domain tasks. [19] uses product knowledge to further training BERT for Review Reading Compression (RRC) task. [19] and [20] use in-domain data to improve the performance of Aspect-Target Sentiment Classification (ATSC) task. In [21], physiology, government and psychology knowledge are used to further train BERT to improve the Short Answer Grading task. Inspired by the aforementioned work, we leverage in domain event knowledge to improve the event extraction performance.

3. METHODOLOGY

This section describes how we build the event extraction system and inject event knowledge based on the BERT pretrained language model.

We extract event trigger and argument in a pipelines mode though two BERT fine-tune strategy respectively: token classification and sentence pair classification.

3.1. Event Trigger Extraction through Token Classification

Given a sentence and a set of predefined event types, trigger extraction aims to find the phrase in the sentence that most clearly express an event occurrence, and identify the event subtypes. This can be seen as a simple sequence labeling task. We encode the input by BERT as a single sentence and feed the contextual representation (BERT's last hidden layer) of each token to a classifier to assign an event type. Besides 33 event subtypes defined by ACE2005, we use an extra "None" label to denote that a token does not trigger any event so that we can identify and classify triggers at the same time. We adopt the IO tagging because a trigger may across more than one token and two triggers hardly appear in adjacent positions.

3.2. Argument Extraction Through Sentence Pair Classification

Argument extraction is relatively more complicated. Following [4] and [8], we directly use the gold annotations for entities. In a sentence consist of words$\{w_1, w_2, \ldots, w_n\}$, some of the words are trigger words T: $\{w_{t1}, w_{te}, \ldots, w_{tk}\}$ with corresponding event type and some of the words are entity mention E: $\{w_{e1}, w_{e2}, \ldots, w_{ej}\}$ as argument candidates, argument extraction aims to identify if the candidate entity is an argument of event triggered by the trigger words, and if so, recognize its role.

[22] explores constructing an auxiliary sentence as extra BERT input for Aspect-Based Sentiment Analysis (ABSA) task: predict sentiment polarity of each target's aspects in a sentence which is similar to our argument extraction task. Their experiment demonstrates that converting a single sentence classification task to several sentence pair classification tasks can significantly improve the performance for the ABSA task. They discuss that their method can be seen as exponentially expanding the corpus. Inspired by their work, we also adopt this method to our system for argument extraction.

We treat the argument extraction task for a sentence as several multiclass classification problems: given a sentence s, events triggered by T and candidates entities E, predict the role over the full set of trigger-entity pairs. Table 1 shows the examples used to extract arguments for the example sentence in Figure 1. There are 37 roles in total. ACE2005 defines 36 different argument roles (e.g. place, person). We use an extra 'None' label to indicate that the entity is not the argument of a given event so that we can identify and classify arguments simultaneously). For each trigger-entity pair, we first build a simple auxiliary pseudo-sentence. For example, the generated sentence for the trigger-entity pair (paid, FBI) is "paid - FBI". We use the sentence pair (the original English sentence and the generated auxiliary sentence) as BERT input. Follow the BERT convention, one special classification token "[CLS]" is added as the first token, and two "[SEP]" tokens are inserted between two sentences and appended to the end respectively.The final BERT input tokens s for this example is "[CLS] Leung was hired by the FBI and paid almost $2 million over 20 years to spy on the Chinese. [SEP] paid - FBI [SEP]". We use BERT to encode the constructed input sentence and get the last hidden layer $h \in \mathbb{R}^{L\times H}$($H$ is the hidden size of BERT and L is the sequence length) as the contextual embedding:

$$\boldsymbol{h} = \boldsymbol{BERT}(s) \tag{1}$$

We use the "[CLS]" token's embedding in last hidden layer (denoted as $h_{[CLS]} \in \mathbb{R}^{H}$) to predict the argument role. The predicted argument role distribution is defined as:

$$\boldsymbol{z} = \boldsymbol{softmax}\left(\boldsymbol{W_e} h_{[CLS]} + \boldsymbol{b_e}\right) \tag{2}$$

Where $W_e \in \mathbb{R}^{K\times H}$, $b_e \in \mathbb{R}^{K}$ are weights and bias for event type e. As different event type has a different set of arguments, we use separate argument classifiers for each event type so that the argument classifier can utilize the event type information.

For each sentence, the argument classification error is defined as the average of all the cross-entropy between the gold and our predicted arguments role distribution:

$$\mathcal{L}_{arg} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} z_{n,k} log(\hat{z}_{n,k}) \quad (3)$$

N is the total number of the trigger-entity pairs in the sentence. K is the total number of argument roles.$z_{n,k} \in \{0,1\}$ denote the gold role for the entity of the event, $\hat{z}_{n,k}$ is our model output.

Table 1. multiclass classification problems for arguments extraction

Trigger	Event Type	Entity	Role (Label)
hired	Start-Position	Leung	Person
hired	Start-Position	FBI	Entity
hired	Start-Position	$2 million	None
hired	Start-Position	20 years	None
hired	Start-Position	Chinese	None
paid	Transfer-Money	Leung	Recipient
paid	Transfer-Money	FBI	Giver
paid	Transfer-Money	$2 million	Money
paid	Transfer-Money	20 years	Time
paid	Transfer-Money	Chinese	None

3.3. Inject Event Knowledge by Further Pretrain BERT

To inject event knowledge to the BERT model, starting from the original BERT checkpoint which is trained on general English corpus (BooksCorpus and Wikipedia), we further pre-train it by in-domain corpus as an intermediate step before fine-tuning it for our event extract system described in 3.1 and 3.2.

Two training objects are used to further pretrain the BERT model: Mask Language Model (MLM) and Next Sentence Prediction (NSP).

For MLM task, 15% random tokens in the original sentence is masked (80% of which is replaced by special token “[mask]”, another 10% of which is replaced by a random token and the remind 10% is unchanged). The model is trained to predict masked tokens.

For NSP task, given a sentence pair (A, B), the model is trained to determine whether they are adjacent (sentence B is the actual next sentence that follows sentence A).

4. EXPERIMENT

4.1.Data Set and metric

We utilize the ACE2005 dataset to evaluate our event extraction system. Following previous data split convention [4][5], we use 40 newswire documents as testset, 30 randomly documents as development set, and remaining 529 documents as training set. We also adopt the following criteria to evaluate the extraction performance as previous work [4][6][7][8][12]:

A trigger is correct if its event subtype and offsets match those of a reference trigger.

An argument is correctly identified if its event subtype and offsets match those of any of the reference argument mentions.

An argument is correctly identified and classified if its event subtype, offsets, and argument role match those of any of the reference argument mentions.

We report individual micro precision, recall and f1 score on the test set for trigger/arguments identification/classification. The precision (Equation 4) is the ratio between correct predictions for all events and all predictions reported by the model. The recall (Equation 5) is the ratio between correct predictions for all events and all trigger/arguments that should be identified/classified. The f1 score (Equation 6) is the harmonic mean between the precision and the recall.

$$\boldsymbol{precision} = \frac{true\ positives}{true\ positives\ +\ false\ positives} \tag{4}$$

$$\boldsymbol{recall} = \frac{true\ positives}{true\ positives\ +\ false\ negatives} \tag{5}$$

$$\boldsymbol{F_1} = 2 * \frac{precison * recall}{precison\ +\ recall} \tag{6}$$

4.2. Hyperparameters and Details of Fine-Tune

We utilize the BERT-base to build our baseline model. Fine-tuning is performed on a single GPU with batch size 32. We set the maximum BERT sequence length to 256. Shorter sequences are padded and no sequences exceed this limit. We train the model using Adam optimizer at learning rate 2e-5 with weight decay 0.01 until converge.

4.3. Event Corpus

In this section, we describe how we build the event corpus for further pre-training BERT. We notice that almost half of the original data in ACE2005 comes from newswire and broadcast news. And as an event extraction data set, it contains a wide range of topics and event types. Therefore, to cover all the ACE2005 events, we utilize the New York Times Annotated Corpus [23] which contains over 1.8 million articles written and published by the New York Times to build our event corpus. NYT is a very large dataset, pretraining with all the data requires a lot of computing resources which can be very expensive. On the other hand, not all articles in NYT involves useful topics that can help improve the performance of ACE2005 task (for example, many articles are related to company report, biographical information, eta) Therefore, we preprocess the NYT corpus by manually selecting articles related to ACE-defined event types. Concretely, each article in NYT corpus is released with metadata and the “descriptors” field specifies a list of descriptive terms corresponding to subjects mentioned in the article, many subjects in NYT corpus have a very strong relation with the ACE predefined event subtypes. We screened the news documents with the most similar topics to each event type to form our corpus, see Table 2 for details.

We ended up with 290409 articles, including 150M words in total as our event corpus. Notice that the total article number is slightly smaller than the sum of all subjects (321356) because some articles may have several different subjects.

Table 2. Components of event corpus

ACE2005 Data Set		NYT Corpus	
Event Type	**Event Subtype**	**Selected Subjects**	**#article**
Life	Be-Born、Marry、Divorce、Injure、Die	weddings and engagements	43848
		deaths	24486
		murders and attempted murders	12804
		accidents and safety	10690
Movement	Transport	armament, defense and military forces	11309
Transaction	Transfer-Ownership、Transfer-Money	finances	26342
Personnel	Start-Position、End-Position、Elect、Nominate	suspensions, dismissals and resignations	16392
		appointments and executive changes	25227
		elections	23668
Contact	Meet、Phone-Write	united states international relations	20390
Conflict	Demonstrate、Attack	civil war and guerrilla warfare	15657
		bombs and explosives	5583
		demonstrations and riots	7750
Justice	Acquit、Charge-Indict、Arrest-Jail、Release-Parole、Sue、Convict、Appeal、Sentence、Trial-Hearing、Fine、Execute、Extradite、Pardon	suits and litigation	23808
		decisions and verdicts	5188
		trials	5381
Business	Merge-Org、Start-Org、Declare-Bankruptcy、End-Org	mergers, acquisitions and divestitures	32903
		reform and reorganization	9930

4.4. Hyperparameters and Details of Further Pre-Training

We create training examples using our event corpus with dupe factor 5, each example consists of a pair of sentences with some tokens masked for MLM and NSP object. The maximum sequence length is 256 which is consistent with the fine-tuning stage. Start from the original BERT checkpoint, the model is further pre-trained on a cloud TPU for 200k steps of batch size 384 at learning rate 2e-5.

4.5. Effect of Event Knowledge

Table 3 shows the effect of event knowledge. The event extraction system based on original pre-trained BERT already achieves a fairly considerable score (73.3% f1 score on trigger classification and 58.4% f1 score on argument classification). After further updating the model through the event corpus, we observed the model (denoted by **Event BERT**) achieve better performance over all metrics on both trigger and argument identification/classification task. It gains 1.8% f1 score improvement on trigger classification and 2.3% f1 score improvement on argument classification which shows the benefits of having in-domain event knowledge.

Table 3. Effect of event knowledge.

Models	**Trigger Identification**	**Trigger Classification**	**Argument Identification**	**Argument Classification**
	P R F1	**P R F1**	**P R F1**	**P R F1**
BERT	78.0 75.7 76.8	74.5 72.3 73.3	60.7 64.1 62.4	56.7 60.1 58.4
EventBERT	78.1 78.0 78.1	75.2 75.0 75.1	62.6 64.6 63.6	59.7 61.8 60.7

5. Conclusion and Future Work

In this paper, we propose an event extraction system based on the pre-trained language model for both event trigger and argument extraction. We explore a new way of using external corpus. An elaborately constructed event corpus is built to improve the ACE2005 event extraction task by further pretraining the BERT language model. Experimental results show that our method is very effective and achieve around 2% improvement while avoiding designing complex event generation processes and rules.

We believe the idea of injecting in-domain knowledge by further pretraining the BERT can be helpful to other different NLP tasks especially for which generating extra training data is hard and painful. However, one major limitation is that a corpus that contain specific in-domain knowledge is required for each different task. For ACE2005 event extraction task, building such a corpus is easy as the ACE2005 dataset involves just common topic. But this is not the case for many other tasks that involves specialized fields knowledge or lacks relative resources.

Therefore, one possible direction for future work is to minimize the cost of constructing the knowledge corpus when applying our method to other tasks. One way to achieve it would be to transfer knowledge from one task to another so that we can reuse the knowledge corpus. It is to be verified that our model can also improve some similar task like KBP event extraction task.

Acknowledgments

This research is supported by the grants from the National Key Research and Development Program of China (No. 2019YFB1705601).

References

[1] Walker, C., Strassel, S.,Medero, J.,& Maeda, K. (2006). ACE 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia.

[2] Riloff, E. (1993). Automatically Constructing a Dictionary for Information Extraction Tasks. Proceedings of the Eleventh National Conference on Artificial Intelligence (pp. 811–816).

[3] Cao, K., Li, X., Ma, W., & Grishman, R. (2018). Including New Patterns to Improve Event Extraction Systems. FLAIRS Conference.

[4] Li, Q., Ji, H.,& Huang, L. (2013, 8). Joint Event Extraction via Structured Prediction with Global Features. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 73–82).

[5] Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G.,& Zhu, Q. (2011, 6). Using Cross-Entity Inference to Improve Event Extraction. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 1127–1136)

[6] Chen, Y., Xu, L., Liu, K., Zeng, D.,& Zhao, J. (2015, 7). Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

[7] Nguyen, T. H., Cho, K., & Grishman, R. (2016, 6). Joint Event Extraction via Recurrent Neural Networks. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 300–309).

[8] Liu, X., Luo, Z.,& Huang, H. (2018, 10). Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 1247–1256).

[9] Abney, S. (2002). Bootstrapping. Proceedings of the 40th annual meeting of the association for computational linguistics, (pp. 360–367).

[10] Liao, S., & Grishman, R. (2011, 11). Using Prediction from Sentential Scope to Build a Pseudo Co-Testing Learner for Event Extraction. Proceedings of 5th International Joint Conference on Natural Language Processing (pp. 714–722).

[11] Li, W., Cheng, D., He, L., Wang, Y., & Jin, X. (2019). Joint Event Extraction Based on Hierarchical Event Schemas From FrameNet. IEEE Access, 7, 25001-25015.

[12] Liu, S., Chen, Y., He, S., Liu, K.,& Zhao, J. (2016, 8). Leveraging FrameNet to Improve Automatic Event Detection. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2134–2143).

[13] Chen, Y., Liu, S., Zhang, X., Liu, K.,& Zhao, J. (2017, 7). Automatically Labeled Data Generation for Large Scale Event Extraction. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 409–419).

[14] Araki, J., & Mitamura, T. (2018, 8). Open-Domain Event Detection using Distant Supervision. Proceedings of the 27th International Conference on Computational Linguistics (pp. 878–891).

[15] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Improving language understanding by generative pre-training.

[16] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, 6). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186).

[17] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H.,& Kang, J. (2019, 9). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. (J. Wren, Ed.) Bioinformatics.

[18] Beltagy, I., Lo, K.,& Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. SciBERT: A Pretrained Language Model for Scientific Text.

[19] Xu, H., Liu, B., Shu, L.,& Yu, P. (2019, 6). BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 2324–2335).

[20] Rietzler, A.,Stabinger, S.,Opitz, P., & Engl, S. (2020, 5). Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. Proceedings of The 12th Language Resources and Evaluation Conference (pp. 4933–4941).

[21] Sung, C.,Dhamecha, T.,Saha, S., Ma, T., Reddy, V.,& Arora, R. (2019, 11). Pre-Training BERT on Domain Resources for Short Answer Grading. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6071–6075).

[22] Sun, C., Huang, L., & Qiu, X. (2019, 6). Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 380–385).

[23] Sandhaus, E. (2008). The new york times annotated corpus. Linguistic Data Consortium, Philadelphia, 6, e26752.

AUTHORS

Zining Yang is a postgraduate student at the School of Computer Science and Engineering at University of Electronic Science & Technology of China (UESTC), Chengdu, China. He received his B.S degree also at UESTC in 2015. His main research interests include natural language processing, data storage, and data mining.

Siyu Zhan is currently an associate professor at the School of Computer Science and Engineering at University of Electronic Science and Technology of China (UESTC). He was a visiting scholar at the Electrical and Computer Engineering Department at Virginia Polytechnic Institute and State University (Virginia Tech) on 2007 and at Computer Science Department at Wayne State University on 2017. His interests include distributed computer system, machine learning, wireless communications, networking and software engineering.

Mengshu Hou is a professor in the School of Computer science &Engineering at the University of Electronic Science and Technology of China (UESTC). He received the M.S and Ph.D. degrees in 2002 and 2005 respectively from the UESTC.

Xiaoyang Zeng is currently a Ph.D. at the Department of computer science and Engineering, University of Electronic Science and Technology (UESTC), Chengdu, China. He received the B.S. degrees in Southwest Petroleum University in 2018, and passed the successive master-doctor program and is studying in UESTC. His research interests focus on natural language processing and text mining.

Hao Zhu is an engineer in the Information Center at the University of Electronic Science and Technology of China (UESTC). He received the B.S and M.S degrees in 2002 and 2006 respectively from the UESTC. His current research interests include management informatization, data visualization, and big data analysis.

ARABIC LOCATION NAME ANNOTATIONS AND APPLICATIONS

Omar ASBAYOU

Department of LEA, Lumière University, CRTT, Lyon 2, France

ABSTRACT

This paper show how location named entity (LNE) extraction and annotation, which makes part of our named entity recognition (NER) systems, is an important task in managing the great amount of data. In this paper, we try to explain our linguistic approach in our rule-based LNE recognition and classification system based on syntactico-semantic patterns. To reach good results, we have taken into account morpho-syntactic information provided by morpho-syntactic analysis based on DIINAR database, and syntactico-semantic classification of both location name trigger words (TW) and extensions. Formally, different trigger word sense implies different syntactic entity structures. We also show the semantic data that our LNE recognition and classification system can provide to both information extraction (IE) and information retrieval(IR).The XML database output of the LNE system constituted an important resource for IE and IR. Future project will improve this processing output in order to exploit it in computer-assisted Translation (CAT).

KEYWORDS

Location name annotations, Location named entities, Information retrieval, Information extraction

1. INTRODUCTION

Geographic document pose a problem to IR. For this reason, it is very important to particularly exploit lexical entitiesreferring to location instances (location named entities/proper names). The provided data and classification are important in IR especially in location proper name contextualisation in text. Text geo-parsing, which means identifying place names in corpus, is one of the experimental approaches in Digital Humanities, particularly in Geospatial Humanities.I. Gregory (2015) emphasises the importance of spatial humanities and the necessity of the creation of corresponding databases for geographic information.J.L. Leidner, (2007) worked on LNE for geographic information system (GIS). Text geo-parsing is part of location named entity recognition (LNER), which is a subfield in NLP. A great deal of research has been done on named entity recognition (NER) in general. However, this task, as we will particularly show in this paper, is a sub-task that can actually play an important role in many natural language processing (NLP) applications, especially in IR. This is actually realised by providing annotated index, which is an XML database (LNER system output) enriched with semantic annotations, which can provide a NE class filter to research engine. It also plays a vital role in ontology enrichment for semantic web. LNE interrelations and their relation with other NE of different classes (PERSON, ORGANISATION, EVENT, TIME etc.) represent an interesting resource for IR and IE etc. Generally, NE are lexical entities with very important informative value. M. Asharef*et.al* [2012], for example, used NE extraction from crime text. The recognition and classification of these entities can provide valuable economic, social and political information

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 51-65, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101405

associated to locations. This is due to what M. Herrmann [2008] calls ‘‘referential unicity’’ to define and characterise NE and particularly proper names. « LOCATION », beside « PERSON », « ORGANISATION », « EVENT » etc. is one of the most important NE classes. Fine annotations of LNE are central in many tasks such as improving geographic text and general corpus analysis. They offer the possibility to associate different entity classes to different location sub-classes.

2. Methodology: Linguistic Description

LNE belongs to different lexical and lexico-syntactic categories:

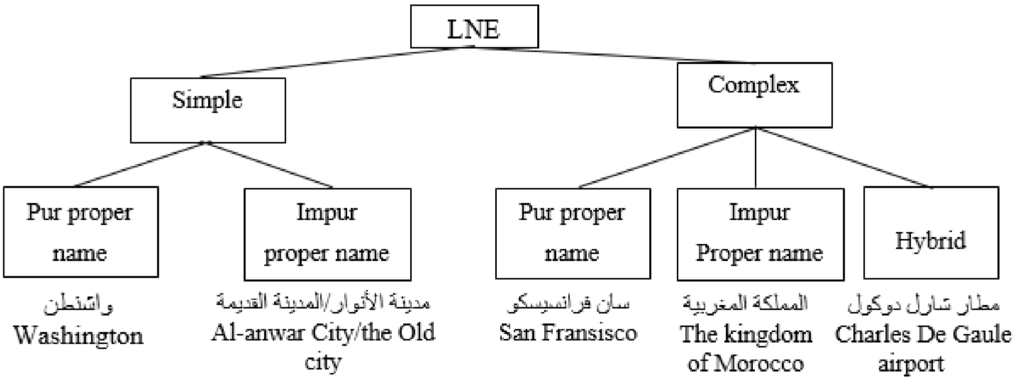


Figure 1.LNE lexico-syntactic categories

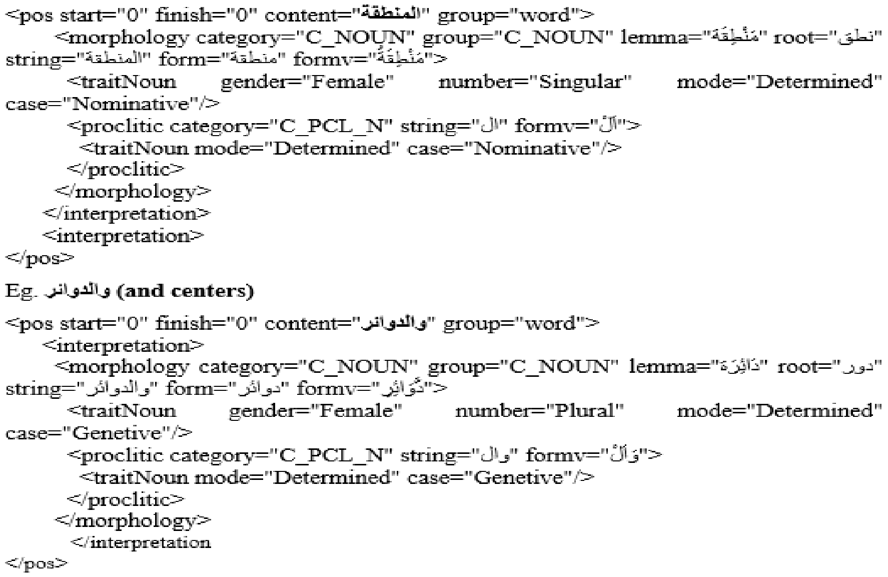
Our rule-based system takes into account these structure properties in their recognition and classification by fine annotations. Therefore, we used a set of lexical, syntactic, and semantic classifications to build correct syntactico-semantic rules of our system.

2.1. Lexical Information

2.1.1.Morpho-syntactic analysis

Our LNER system makes part of our rule-based NER system, which is processing of six levels. We are not going to expose a detailed description of our system here. Level 0 applies the morpho-syntactic analysis which is the first and basic step in our system. We use a morpho-syntactic analysis system based on DIINAR, a rich Arabic lexical database constructed by many researchers: J. Dichy, from Lumière university Lyon 2, and A. Braham from Manouba University Tunisia, (linguistic aspect), M. Hassoun, ENSSIB Lyon, Research Institute for Computer Science and Telecommunication in Tunis, and S. Ghazali from High Institute of Language in Tunis(computer science aspect).This database provides our morpho-synyactic analysis system with rich morpho-syntactic data. The morpho-syntactic-analysis represents the first pre-treatment step for our LNE extraction system (Level 0). For example, the words المنطقة (the region) and والدوائر (and centres) aremorpho-syntactically analysed as follows:

E.g.1:**المنطقة (the region)**



```
<pos start="0" finish="0" content="المنطقة" group="word">
      <morphology category="C_NOUN" group="C_NOUN" lemma="مَنْطِقَة" root="نطق"
string="المنطقة" form="منطقة" formv="مَنْطِقَةْ">
       <traitNoun     gender="Female"     number="Singular"     mode="Determined"
case="Nominative"/>
       <proclitic category="C_PCL_N" string="ال" formv="اَلْ">
        <traitNoun mode="Determined" case="Nominative"/>
       </proclitic>
      </morphology>
     </interpretation>
     <interpretation>
</pos>
Eg. والدوائر (and centers)
<pos start="0" finish="0" content="والدوائر" group="word">
     <interpretation>
      <morphology category="C_NOUN" group="C_NOUN" lemma="دَائِرَة" root="دور"
string="والدوائر" form="دوائر" formv="دَوَائِر">
       <traitNoun     gender="Female"     number="Plural"     mode="Determined"
case="Genetive"/>
       <proclitic category="C_PCL_N" string="وال" formv="وَالْ">
        <traitNoun mode="Determined" case="Genetive"/>
       </proclitic>
      </morphology>
        </interpretation
</pos>
```

Figure 2. Examples of morpho-syntactic analysis

This Lexical resource constitute the base for our syntactic rules in that it provides lexical information that are basic data in our NE recognition and classification system.

2.1.2 Semantic classification

Our sub-classification of the class « «LOCATION» is based on many approaches: field, organisation location, organisation building, geographic location, address and facility.

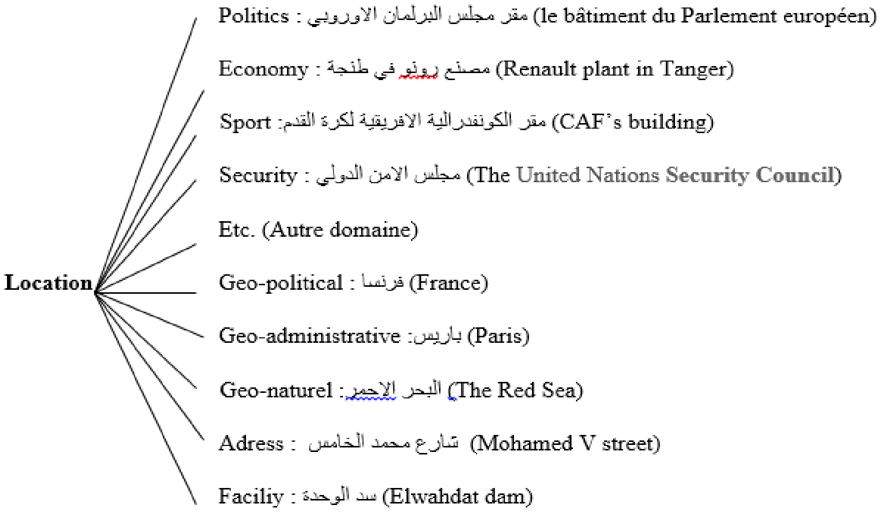
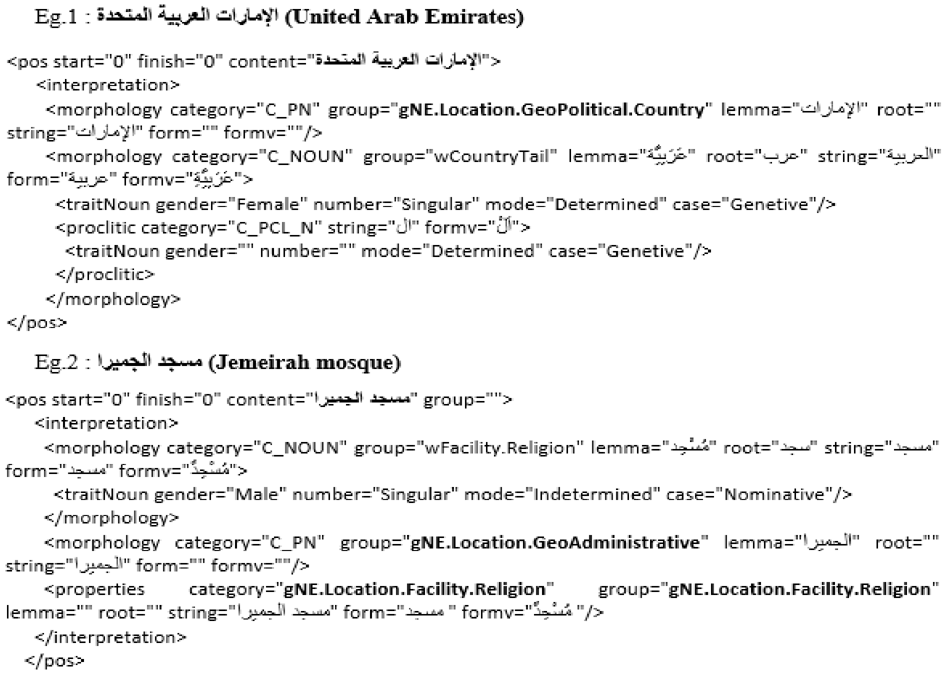

Location
Politics : مقر مجلس البرلمان الاوروبي (le bâtiment du Parlement européen)
Economy : مصنع رونو في طنجة (Renault plant in Tanger)
Sport :مقر الكونفدرالية الافريقية لكرة القدم (CAF's building)
Security : مجلس الامن الدولي (The United Nations Security Council)
Etc. (Autre domaine)
Geo-political : فرنسا (France)
Geo-administrative :باريس (Paris)
Geo-naturel :البحر الاحمر (The Red Sea)
Adress : شارع محمد الخامس (Mohamed V street)
Faciliy : سد الوحدة (Elwahdat dam)

Figure 3. Location sub-categories

The figure above show that LOCATION class can be defined by:

- Field (politics, security, sport, economy .etc.): it functions as a distinctive feature based on a well-defined semantic field classification.
- Geography: this allows us to distinguish geographic proper names form other types of location. This sub-class is subdivided on many subclasses : geopolitical, which includes « country », « city », « region », « department » etc. and geo-natural, in which we put « sea », « mountain », « river » etc.
- Facility: this information is associated to different facility proper names (dam, motorway, stadium etc.).

This information is provided by the recognised LNE constituent information and expressed by fine semantic annotations. The figure bellow show how LNE are annotated by our system and the classifying information provided by these semantic annotations (for visibility, I put the entity and the annotation in bold):



Eg.1 : الإمارات العربية المتحدة **(United Arab Emirates)**

```
<pos start="0" finish="0" content="الإمارات العربية المتحدة">
   <interpretation>
    <morphology category="C_PN" group="gNE.Location.GeoPolitical.Country" lemma="الإمارات" root=""
string="الإمارات" form="" formv=""/>
    <morphology category="C_NOUN" group="wCountryTail" lemma="عَرَبِيَّةٌ" root="عرب" string="العربية"
form="عربية" formv="عَرَبِيَّةٌ">
     <traitNoun gender="Female" number="Singular" mode="Determined" case="Genetive"/>
     <proclitic category="C_PCL_N" string="ال" formv="أَلْ">
      <traitNoun gender="" number="" mode="Determined" case="Genetive"/>
     </proclitic>
    </morphology>
</pos>
```

Eg.2 : مسجد الجميرا **(Jemeirah mosque)**

```
<pos start="0" finish="0" content="مسجد الجميرا" group="">
   <interpretation>
    <morphology category="C_NOUN" group="wFacility.Religion" lemma="مُسْجِد" root="سجد" string="مسجد"
form="مسجد" formv="مُسْجِدٌ">
     <traitNoun gender="Male" number="Singular" mode="Indetermined" case="Nominative"/>
    </morphology>
    <morphology category="C_PN" group="gNE.Location.GeoAdministrative" lemma="الجميرا" root=""
string="الجميرا" form="" formv=""/>
    <properties category="gNE.Location.Facility.Religion" group="gNE.Location.Facility.Religion"
lemma="" root="" string="مسجد الجميرا" form="مسجد " formv="مُسْجِدٌ "/>
   </interpretation>
  </pos>
```

Figure 4. Examples of LNE annotation output

We note that the second annotation example contains two LNE categories: gNE.Location.GeoAdministrative and gNE.Location.Facility.Religion. Based on entity constituents translated by our lexical semantic classification (e.g. wFacility.Religion), each of these annotations provide some location information (location, geo-administrative, facility, religion). To morpho-syntactic data, we add and semantic information .We classified the linguistic entities involved in LNE structure, in our syntactico-semantic rule system, according to different semantic (semantic fields) and conceptual relations (TW classes). We put this TW lexicon into different categories according to common semantic, conceptual and lexico-syntactic criteria.

A. TW (Trigger words) :

TW are word marking NE initial position. For example, مجلس (council), جمعية (association), هيئة (committee) etc. are *generic TW* belonging to the class LOCATION/ORGANISATION. These are distinguished form *specific TW* like قنصلية (consulate) ,مطار (airport) for syntactico-semantic reasons. These semantically different TW classes belong to different LOCATION sub-classes since they participate in different syntactico-semantic patterns (rules). In Level 1, we added this lexico-semantic information to morpho-syntactic output of level 0.

B. Proper names :

We have also enrichedLevel 1 with a set of:
- ***Lists of sub-category locations***: these lists of simple LNE of countries, cities, rivers, mountains etc. aims to enrich le lexical database. For example:

a.المغرب(Morocco), فرنسا (France) etc. =>gNE.LOCATION.GeoPolitical.Country
b. باريس (Paris), الرباط(Rabat) etc. =>gNE.LOCATION.GeoAdministrative.City

- ***Syntactic entities:*** syntactic entities are extracted and annotated using our syntactico-semantic rules combining constituents.

a. الحدود الجنوبية الشرقية (The *southeastern frontiers)* =>gNE.LOCATION.GeoPolitical
*b.*الملعب الاولمبي الياباني (The Japanese Olympic Stadium) =>gNE.LOCATION.Facility.Sport
c. مدينة أكادير (the city of Agadir) =>gNE.LOCATION.GeoAdministrative.City

To illustrate the results of Level 1 we suggest the following two sentences, in which different colors mark different entity classes and sub-classes extracted in this level by our NE recognition and classification system:

Sentence 1:
التقى الرئيس الروسيفلاديمير بوتين رئيسالاستخبارات العامة السعوديةالأميربندر بن سلطانأمس الثلاثاء في موسكو

Sentence 2:
أكدرئيسالمجلس الوطنيلتنظيمالقطاع الخصوصي وتشجيع المبادرات الشيخ سلمان بن علي بهذا الخصوص على أن تعمل على زيادةمساهمةالطاقة المتجددة في خليطالطاقة الكلية

In these two examples, the NER system extracted and classified the LNE:موسكو (Moscow) (gNE.LOCATION.GeoAdministrative.Country), in sentence 1, and the complex TW المجلس الوطني (the national council) in of the NE المجلس الوطنيلتنظيمالقطاع الخصوصي وتشجيع المبادرات. LOCATION/ORGANISATION ambiguity is resolved in the following levels exploiting contextual elements such as the prepositionsفي (in) and ORGANISATION attributes such asرئيس (president) in *رئيسالمجلسالوطني* لحقوق الإنسان (the president of the National Council of Human Rights) which disambiguates the NE class into gNE_PERSON_Society. Here, we combine ‘‘*organisation*’’ with ‘‘*person function*’’ attribute.

Our study deals with NE linguistic specificities and the elements involved in their syntactico-semantic structures. These linguistic data highlight several levels of analysis.

2.2. Syntactico-Semantic Information

Our study starts from the principles that TW represent the head of the noun phrase NE (NP)and it is accompanied with one or several modifiers or complements. Second, Each LNE class has defined set of attributes (generally extracted in Level 1 and 2), which are put in a well-determined distribution in NE extraction patterns. Third, Complex LNE structures are composed of many linguistic entities combined in a defined order (immediate constituent analysis). Fourth, fine annotation depends on trigger word and extension information. Therefore, in the syntactico-semantic level, we will shed light on two important aspects: NE structure constituents and NE class attributes.

2.2.1. LNE Structure Constituents

LNE constituents are crucial in LNE constituent combinations and classification. The structure of NE is divided into two parts: trigger words and extensions.

A. Trigger word :

We have taken into account different perspectives in constructing the typology of trigger words:

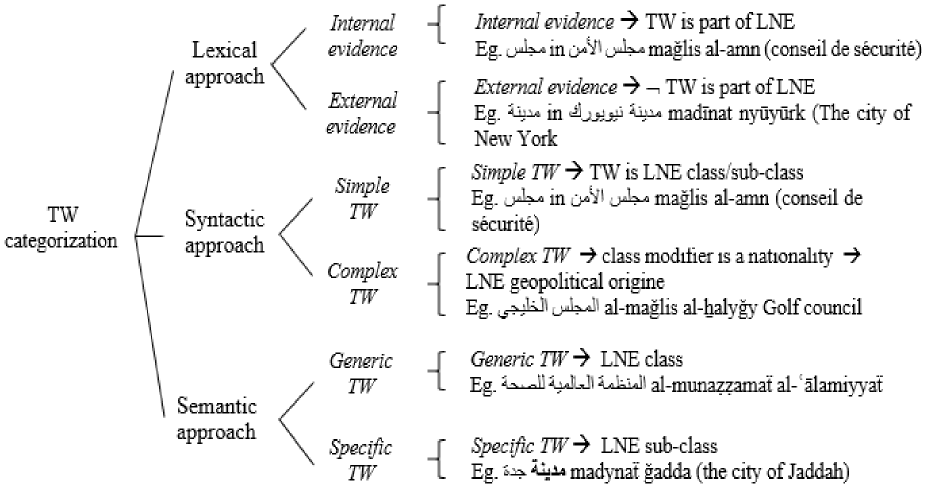

TW categorization
Lexical approach
Syntactic approach
Semantic approach
Internal evidence
External evidence
Simple TW
Complex TW
Generic TW
Specific TW
Internal evidence → TW is part of LNE
Eg. مجلس in مجلس الأمن mağlis al-amn (conseil de sécurité)
External evidence → ¬ TW is part of LNE
Eg. مدينة in مدينة نيويورك madīnat nyūyūrk (The city of New York
Simple TW → TW is LNE class/sub-class
Eg. مجلس in مجلس الأمن mağlis al-amn (conseil de sécurité)
Complex TW → class modifier is a nationality → LNE geopolitical origine
Eg. المجلس الخليجي al-mağlis al-ẖalyğy Golf council
Generic TW → LNE class
Eg. المنظمة العالمية للصحة al-munaẕẕamaẗ al-ʿālamiyyaẗ
Specific TW → LNE sub-class
Eg. مدينة جدة madynaẗ ğadda (the city of Jaddah)

Figure 5.Trigger word typology

As the figure shows, TW sub-categorization provides interesting information for LNE recognition and classification.

B. NE extensions :

The NE extensions concern the morpho-syntactic or syntactic entities occurring after trigger words, they, mark the frontiers of the extracted NE and specify their sub-classes. Information provided by the extension is very useful in solving the problems of NE frontiers and classification.

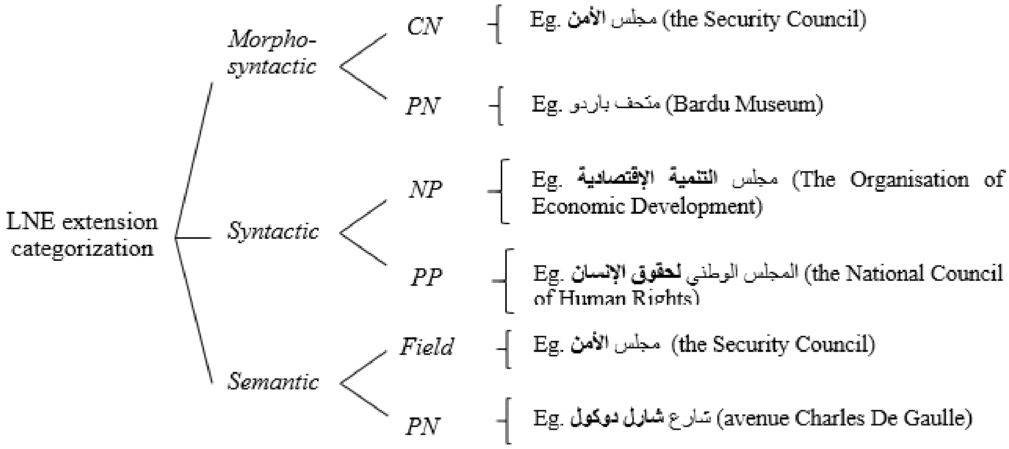

LNE extension categorization
Morpho-syntactic
CN
Eg. مجلس الأمن (the Security Council)
PN
Eg. متحف باردو (Bardu Museum)
Syntactic
NP
Eg. مجلس التنمية الإقتصادية (The Organisation of Economic Development)
PP
Eg. المجلس الوطني لحقوق الإنسان (the National Council of Human Rights)
Semantic
Field
Eg. مجلس الأمن (the Security Council)
PN
Eg. شارع شارل دوكول (avenue Charles De Gaulle)

Figure 6. LNE extension categorisation

2.2.2. LNE class semantic attributes and relation with NE

LNE semantic attributes are classes that participate in the formation of their syntactico-semantic patterns. They have values, which are recognised and represented by different types of linguistic entities classified by our NER system. For example, the LNE attribute « person proper name» in شارع شارل دوكولšāriʿšārldūgul (Charles De Gaulle) has the value «شارل دوكول » šārldūgul (Charles De Gaulle). LNE attribute position is « after TW » and LNE are, in their turn, are attributes in other named entities.

A. Attributes after TW

In this case, class attribute values do not change LNE class. The formal description is a semantic feature structure: LNE « x attribute » has « y value ». For example, the LNE شارع شارل دوكولšāriʿ šārl dūgul (Avenue Charles De Gaulle) « person proper name attribute» has « شارلدوكول **šārl** dūgul (Charles De Gaulle) value ».

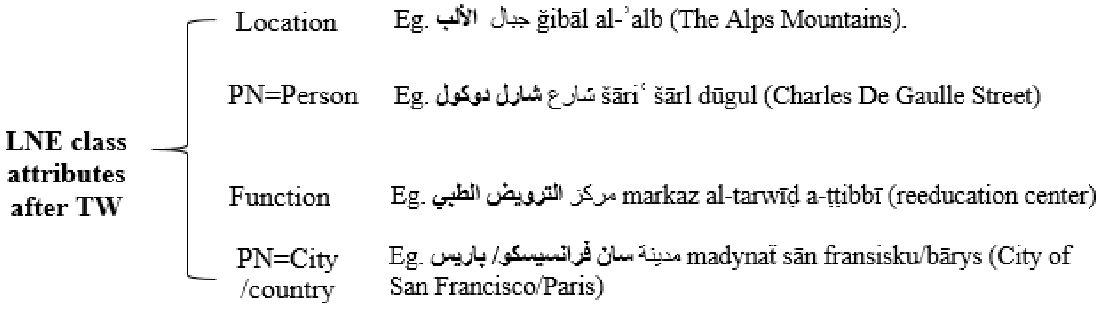

LNE class attributes after TW
Location
Eg. جبال الألب ǧibāl al-ʾalb (The Alps Mountains).
PN=Person
Eg. شارع شارل دوكول šāriʿ šārl dūgul (Charles De Gaulle Street)
Function
Eg. مركز الترويض الطبي markaz al-tarwīḍ a-ṭṭibbī (reeducation center)
PN=City /country
Eg. مدينة سان فرانسيسكو/ باريس madynaẗ sān fransisku/bārys (City of San Francisco/Paris)

Figure 7. LNE class attributes after TW

B. LNE=NE attributes

The LNE do not have attributes before the TW but can be attributes of other NE classes.Nonetheless, they should not be lost within other NE in which they are constituents providing some information such as event/location relation (a and b) and event/organisation relation (c):

- EVENT NE attribute ATW = GEOPOLITICAL LOCATION NE

a) مؤتمر **جنيف** (Geneva Meeting) is EVENT NE with «location attribute» whose value is جنيفğunif (Geneva).

b) الدورة التاسعة والعشرين للألعاب الأوليمبية في **الصين** (The twenty-ninth Olympic Games 2008 in China) is EVENT NE with «location attribute» whose value is **الصين**(China).

- ORGANISATION NE attribute BTW = LOCATION BUILDING NE

c) **مبنى** مجلس الأمن الدولي (The international Security Council Building) is ORGANISATION NE with «location building attribute» whose value is the whole NE **مبنى** مجلس الأمن الدولي (The international Security Council Building) because the attribute is before the ORGANISATION NE TW.This contributes in solving the problem of some cases of metonymy typical of ORGANISATION NE.

C. The role of « nationality » modifier and of « geopolitical location » complement

« Nationality » and « geopolitical location » are geo political (location) attributes in NE whose value is respectively an adjective (modifier) or geopolitical proper name (complement); both entities denote a geo political information of the NE in which they are constituent after TW.

a. رئيس الحكومة **المغربية** سعد الدين العثماني (the **Moroccan** government president saʿd a-ddīn al-ʿuṯmānī) =>gNE_PERSON_Politics

b. الرئيس **الامريكي** دونالد ترامب(The **American** President Donald Trump) =>gNE_PERSON_Politics

c. البنك المركزي **الأوروبي**(the **European** Central Bank)

d. *مهرجان **مراكش** الدولي للفيلم 2019* (Marrakech International Film Festival 2019)

From annotations provided by our NER system, geopolitical origin is one of the information that can be extracted:

Table 1. NE semantic constituents

NE	Annotation	Class	Field	Geopolitical origine	Proper name	Date
الرئيس الامريكي دونالد ترامب	gNE_PERSON_Politics	PERSON	Politics	American (America)	دونالد ترامب	…
رئيس الحكومة المغربية سعد الدين العثماني	gNE_PERSON_Politics	PERSON	Politics	المغربية (Moroccan/ morocco)	سعد الدين العثماني	…
البنك المركزي الأوروبي	gNE_ORGANISATION_Finance	ORGANISATION	Finance	الأوروبي (European, Europe)	البنك المركزي الأوروبي	…
مهرجان مراكش الدولي 2019 للفيلم	gNE_EVENT_Art	EVENT	Art	*مراكش* (Marrakech)	*مهرجان مراكش الدولي للفيلم*	*2019*

Location information provided by « nationality » as well as city and country names can be exploited in to enrich different kind of databases for many purposes and to establish relation with other NE classes.

3. RESULTS AND EVALUATION

The result of our LNER system is an XML database with annotated place proper names. Here are some examples:

```
<pos start="0" finish="0" content="نفق الشندغة" group="">
<interpretation>
<morphology category="C_NOUN" group="wFacility" lemma="نُفُق" root="نفق" string="نفق"
form="نفق" formv="نُفُقَ">
<traitNoun gender="Female" number="Singular" mode="Annexion" case="Accusative"/>
</morphology>
<morphology category="C_PN" group="cPN" lemma="الشندغة" root="" string="الشندغة" form=""
formv=""/>
<properties category="gNE.Location.Facility" group="gNE.Location.Facility" lemma="" root=""
string="نفق الشندغة" form="نفق "formv="نُفُقَ "/>
</interpretation>
</pos>
```

```
<pos start="0" finish="0" content="الجمهورية التونسية">
<interpretation>
<morphology category="C_NOUN" group="gNE.Location.GeoPolitical.Country"
lemma="جُمْهُورِيَّة" root="جمهر" string="الجمهورية" form="جمهورية" formv="جُمْهُورِيَّةَ">
<traitNoun gender="Female" number="Singular" mode="Determined" case="Accusative"/>
<proclitic category="C_PCL_N" string="ال" formv="اَلْ">
<traitNoun gender="" number="" mode="Determined" case="Accusative"/>
</proclitic>
</morphology>
</pos>
```

```
<pos start="0" finish="0" content="دبي">
<interpretation>
<morphology category="C_VB" group="gNE.Location.GeoAdministrative.City"
lemma="دَبَّ/يَدَبُّ" root="دبب" string="دبي" form="دبي" formv="دَبِّي">
<traitVerb pronoun="2PFS" tens="IMP_SPL_ACT"/>
</morphology>
<properties category="gNE.Location.GeoAdministrative.City"
group="gNE.Location.GeoAdministrative" lemma="دَبَّ/يَدَبُّ" root="دبب" string="دبي" form="دبي"
formv="دَبِّي">
<traitVerbpronoun="2PFS" tens="IMP_SPL_ACT"/>
</properties>
</interpretation>
</pos>
```

```
<posstart="0" finish="0" content="حديقة الخور" group="">
<interpretation>
<morphology category="C_NOUN" group="wFacility" lemma="حَدِيقَة" root="حدق" string="حديقة"
form="حديقة" formv="حَدِيقَةٌ">
<traitNoun gender="Female" number="Singular" mode="Indetermined" case="Nominative"/>
</morphology>
```

<morphology category="C_NOUN" group="**gNE.Location.GeoAdministrative**" lemma="خَوَر" root="خور" string="**الخور**" form="خور" formv="خَوَرِ">

<traitNoun gender="Male" number="Singular" mode="Determined" case="Genetive"/>

<proclitic category="C_PCL_N" string="ال" formv="اَلْ">

<traitNoun gender="" number="" mode="Determined" case="Genetive"/>

</proclitic>

</morphology>

<properties category="**gNE.Location.Facility**" group="gNE.Location.Facility" lemma="" root="" string="حديقة الخور" form="حديقة الخور" formv="حَدِيقَةُ اَلْخَوَرِ">

<traitNoun gender="Male" number="Singular" mode="Determined" case="Genetive"/>

</properties>

</interpretation>

</pos>

<pos start="0" finish="0" content="**لمصرف الإمارات**">

<interpretation>

<morphology category="C_NOUN" group="wEconomicOrgIndet" lemma="مُصَرِّف" root="صرف" string="لمصرف" form="مصرف" formv="مٌّصُرفَ">

<traitNoun gender="Male" number="Singular" mode="Indetermined" case="Nominative"/>

<proclitic category="C_PCL_N" string="ل" formv="لَ">

<traitNoun gender="" number="" mode="Indetermined" case="Nominative"/>

</proclitic>

</morphology>

<morphology category="C_PN" group="**gNE.Location.GeoPolitical.Country**" lemma="الإمارات" root="" string="الإمارات" form="" formv=""/>

<properties category="**gNE.Organisation.Economy**" group="gNE.Organisation.Economy" lemma="" root="" string="لمصرف الإمارات" form="لمصرف "formv="لَمٌّصُرفَ "/>

</interpretation>

</pos>

<pos start="0" finish="0" content="**جبال حجر**" group="">

<interpretation>

<morphology category="C_NOUN" group="wGeoNaturalLocation" lemma="جِبَال" root="جبل" string="جبال" form="جبال" formv="جِبَالٌ">

<traitNoun gender="Male" number="Singular" mode="Indetermined" case="Nominative"/>

</morphology>

<morphology category="C_NOUN" group="cNoun" lemma="حِجْر" root="حجر" string="حجر" form="حجر" formv="حِجْرَ">

<traitNoun gender="None" number="Singular" mode="Annexion" case="Accusative"/>

</morphology>

<properties category="**gNE.Location.GeoNatural**" group="gNE.Location.GeoNatural" lemma="" root="" string="جبال حجر" form="جبال حجر" formv="جِبَالٌ حِجْرَ">

<traitNoun gender="None" number="Singular" mode="Annexion" case="Accusative"/>

</properties>

</interpretation>

</pos>

<pos start="0" finish="0" content="**جزيرة مسندم**" group="">

<interpretation>

<morphology category="C_NOUN" group="wGeoNaturalLocation" lemma="جَزِيرَة" root="جزر" string="جزيرة" form="جزيرة" formv="جَزِيرَةٍ">

<traitNoun gender="Female" number="Singular" mode="Indetermined" case="Genetive"/>

```
</morphology>
<morphology category="C_PN" group="cPN" lemma="مسندم" root="" string="مسندم" form=""
formv=""/>
<properties category="gNE.Location.GeoNatural" group="gNE.Location.GeoNatural"
lemma="" root="" string="جزيرة مسندم" form="جزيرة "formv="جَزِيرَةٍ "/>
</interpretation>
</pos>
```

```
<pos start="0" finish="0" content="الساحل الإماراتي" group="">
<interpretation>
<morphology category="C_NOUN" group="wCardinalPoint" lemma="سَاحِل" root="سحل"
string="الساحل" form="ساحل" formv="سَّاحِلِ">
<traitNoun gender="Male" number="Singular" mode="Determined" case="Genetive"/>
<proclitic category="C_PCL_N" string="ال" formv="اَلْ">
<traitNoun gender="" number="" mode="Determined" case="Genetive"/>
</proclitic>
</morphology>
<morphology category="wNationalityMasculin" group="wNationalityMasculin"
lemma="الإماراتي" root="" string="الإماراتي" form="" formv=""/>
<properties category="gNE.Location.GeoPolitical" group="gNE.Location.GeoPolitical"
lemma="" root="" string="الساحل الإماراتي" form="الساحل "formv="اَلْسَّاحِلِ "/>
</interpretation>
</pos>
```

```
<pos start="0" finish="0" content="مطار شارل دوكول" group="">
<interpretation>
<morphology category="C_NOUN" group="wFacility" lemma="مَطَار" root="مطر" string="مطار"
form="مطار" formv="مَطَارٌ">
<traitNoun gender="Male" number="Singular" mode="Indetermined" case="Nominative"/>
</morphology>
<morphology category="gNE.Person" group="gNE.Person.PN" lemma="شارل دوكول " root=""
string="شارل دوكول " form="" formv=""/>
<properties category="gNE.Location.Facility" group="gNE.Location.Facility" lemma="" root=""
string="مطار شارل دوكول" form="مطار  "formv="مَطَارٌ "/>
</interpretation>
</pos>
```

Figure 8. Examples of our LNE recognition andannotation

Our LNE system made good results. We used two corpora for evaluation:*ANERCorp* (154 674 words), which is available online, and French Press Agency (FPA) *Corpus* (30 000 words).Here is the evaluation table (recall, precision and f-measure):

$$\text{Recall} = \frac{\text{Number of correctly annotated entities} \times 100}{\text{Number of entities in the corpus}}$$

$$\text{Precision} = \frac{\text{Number of correctly annotated entities} \times 100}{\text{Numberof annotated entities}}$$

$$\text{F-mesure} = \frac{2\text{*recall *precision}}{\text{Recall + precision}}$$

Table 2. LNER results

Corpus	NE in the corpus	correctly annotated NE	Recall	Precision	F-measure
ANERCorp	4008	3709	92,53 %	96,46%	94,45
FPA *Corpus*	2523	2344	94,96 %	97,82 %	96,36%

We exploited the result of our LNER system in the construction of Techlimed research engine filtering by different LNE class (geo-political, geo-administrative, facility, geo-natural), and different fields (politics, economy, social, sport, justice, health, science, religion, administration, etc.). Figure 9 is the research engine interface showing a cloud of NE including LNE recognised by our system:

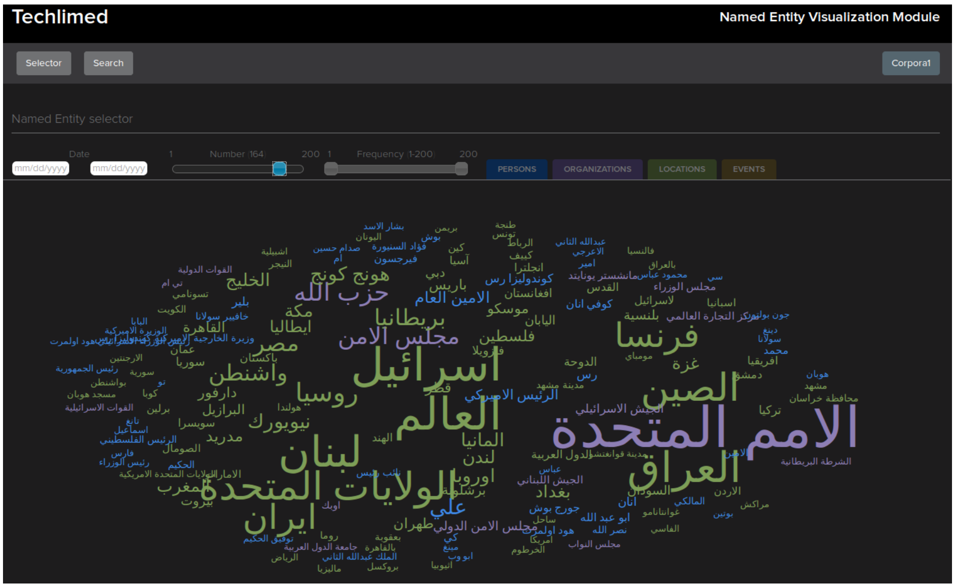
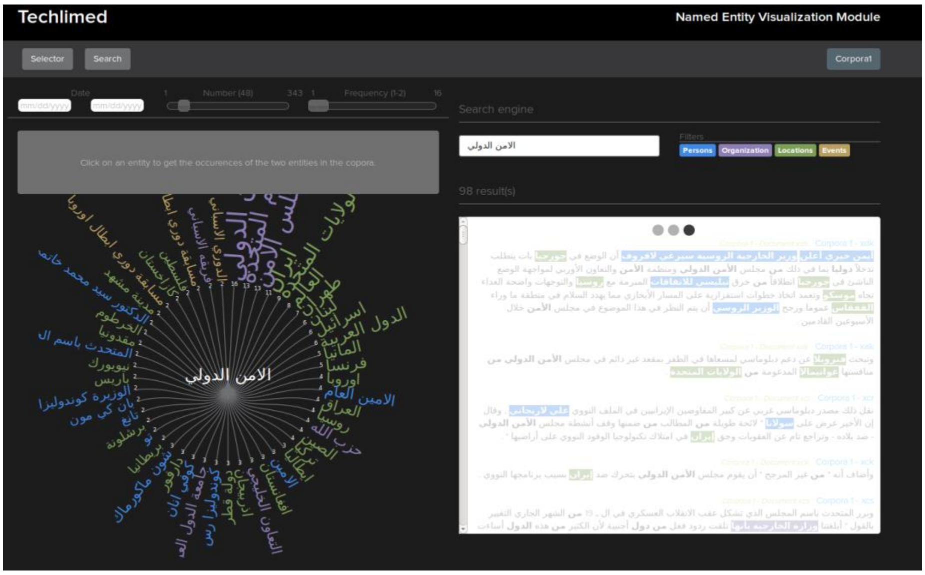


Figure 9. The LNE output in Techlimed research engine

The most frequent are bigger and clearer (bigger size): e.g.الأمم المتحدة (United Nation), الولايات المتحدة (USA), اسرائيل (Israel), فرنسا (France), الصين (China), ايران (Iran) etc. The less frequent are smaller: e.g. مجلس الأمن الدولي (the International Security Council), جامعة الدول العربية (the Arab League), مركز التجارة العالمية (the International Trade Centre), سويسرا (Switzerland) etc. The research engine uses the output of our system of NE extraction and classification to contextualise the query. The figures below (research engine pages) show some aspects of our LNER contribution in IR:



الامن الدولي

أمين خبري أعلن وزير الخارجية الروسية سيرغي لافروف أن الوضع في جورجيا بات يتطلب تدخلاً دوليا بما في ذلك من مجلس الأمن الدولي ومنظمة الأمن والتعاون الأوربي لمواجهة الوضع الناشئ في جورجيا انطلاقاً من خرق تبليسي للاتفاقات المبرمة مع روسيا والتوجهات واضحة العداء تجاه موسكو وتعمد اتخاذ خطوات استفزازية على المسار الأبخازي مما يهدد السلام في منطقة ما وراء القفقاس عموما ورجح الوزير الروسي أن يتم النظر في هذا الموضوع في مجلس الأمن خلال الأسبوعين القادمين

وتبحث فنزويلا عن دعم دبلوماسي لمسعاها في الظفر بمقعد غير دائم في مجلس الأمن الدولي من منافستها غواتيمالا المدعومة من الولايات المتحدة .

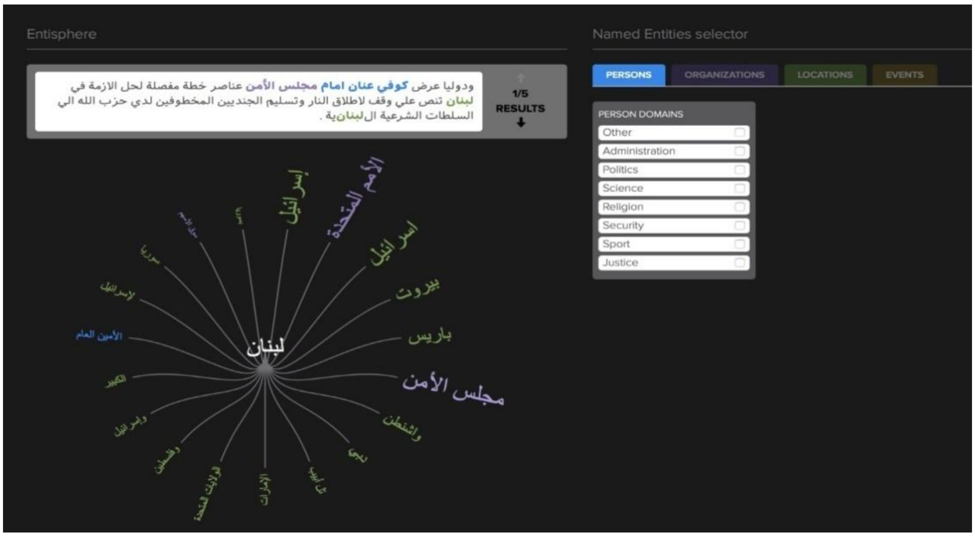
نقل ذلك مصدر دبلوماسي غربي عن كبير المفاوضين الإيرانيين في الملف النووي علي لاريجاني . وقال إن الأخير عرض على سولانا " لائحة طويلة من المطالب من ضمنها وقف أنشطة مجلس الأمن الدولي - ضد بلاده - وتراجع تام عن العقوبات وحق إيران في امتلاك تكنولوجيا الوقود النووي على أراضيها " .

وأضاف أنه " من غير المرجح " أن يقوم مجلس الأمن الدولي بتحرك ضد إيران بسبب برنامجها النووي .

ويرر المتحدث باسم المجلس الذي تشكل عقب الانقلاب العسكري في الـ 19 من الشهر الجاري التغيير بالقول " أبلغتنا وزارة الخارجية بأنها تلقت ردود فعل من دول أجنبية لأن الكثير من هذه الدول أساءت

Figure 10. Research engine contextualisation of the query الامن الدولي (the international security) with NE extracted by our system

The figure shows the obtained results of the query الامن الدولي (the international security)**.** The diagram in the retrieved page on the left of the screen shows that the query is associated not only with LNE but also with the rest of NE recognised by our system; and the click on any NE on the left diagram gives access to the text with the corresponding context on the right side of the page. The research engine also uses the NE classification to filter by classes and subclasses. The following figure is an example a query contextualisation:



ودوليا عرض كوفي عنان امام مجلس الأمن عناصر خطة مفصلة لحل الازمة في لبنان تنص علي وقف لاطلاق النار وتسليم الجنديين المخطوفين لدي حزب الله الي السلطات الشرعية اللبنانية .

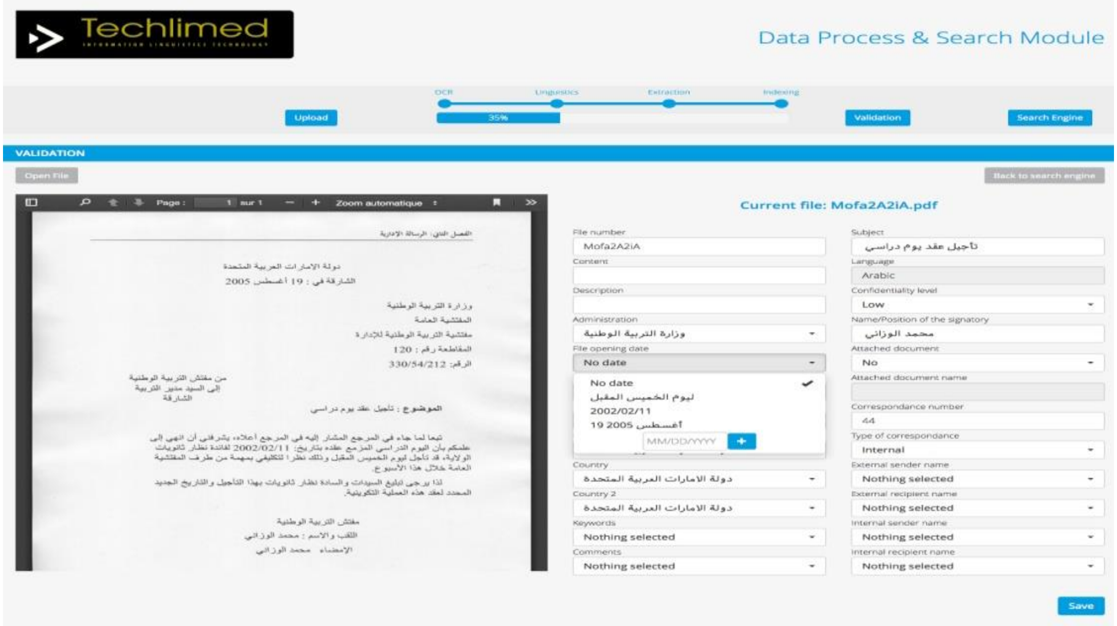
PERSON DOMAINS

- Other □
- Administration □
- Politics □
- Science □
- Religion □
- Security □
- Sport □
- Justice □

Figure 11. Contextualisation of the query لبنان (Lebanon) and filtering by NE classes and subclasses

As we can see, we can filter by class (e.g. PERSON, ORGANISATION, LOCATION, and EVENT) and by field (politics, science, religion, sport, economy etc.).

We also developed an information extraction application using the output of our LNE extraction and classification system. The figures bellow shows the contribution of our NE recognition system in information extraction from administrative letters:



الفصل الثاني: الرسالة الإدارية

دولة الإمارات العربية المتحدة
الشارقة في : 19 أغسطس 2005

وزارة التربية الوطنية
المفتشية العامة
مفتشية التربية الوطنية للإدارة
المقاطعة رقم : 120
الرقم: 330/54/212

من مفتش التربية الوطنية
إلى السيد مدير التربية
الشارقة

الموضوع : تأجيل عقد يوم دراسي

تبعا لما جاء في المرجع المشار إليه في المرجع أعلاه، يشرفني أن انهي إلى علمكم بأن اليوم الدراسي المزمع عقده بتاريخ: 2002/02/11 لفائدة نظار ثانويات الولاية، قد تأجل ليوم الخميس المقبل وذلك نظرا لتكليفي بمهمة من طرف المفتشية العامة خلال هذا الأسبوع.

لذا يرجى تبليغ السيدات والسادة نظار ثانويات بهذا التأجيل والتاريخ الجديد المحدد لعقد هذه العملية التكوينية.

مفتش التربية الوطنية
اللقب والاسم : محمد الوزاني
الإمضاء محمد الوزاني

Current file: Mofa2A2iA.pdf

File number: Mofa2A2iA
Content:
Description:
Administration: وزارة التربية الوطنية
File opening date: No date
- No date ✓
- ليوم الخميس المقبل
- 2002/02/11
- 19 2005 أغسطس
- MM/DD/YYYY +

Country: دولة الامارات العربية المتحدة
Country 2: دولة الامارات العربية المتحدة
Keywords: Nothing selected
Comments: Nothing selected

Subject: تأجيل عقد يوم دراسي
Language: Arabic
Confidentiality level: Low
Name/Position of the signatory: محمد الوزاني
Attached document: No
Attached document name:
Correspondance number: 44
Type of correspondance: Internal
External sender name: Nothing selected
External recipient name: Nothing selected
Internal sender name: Nothing selected
Internal recipient name: Nothing selected

Figure 12 : Information extraction based on our NE extraction and classification

Here les NE annotation are used to fill in the form: administration, country, country2 etc. We can use the same LNE system in any other IE system.

4. CONCLUSION

This paper shows our LNR system and the importance of a linguistic approach in this task. The process is composed of many processing levels going form lexical to syntactico-semantic analysis. The LNE syntactic and semantic information are vital to their extraction and classification. We exploited the output of our system within Techlimed in the information retrieval and extraction applications. We integrated the obtained fine annotations made of LNE classes and subclasses in the systems of indexations for an efficient information retrieval and extraction. We can also extend our analysis to sentence annotation using verb classification as predicates. The objective is, to exploit the output enriched by these semantic annotations, to develop the project of the extraction of relations between different extracted NE classes within sentences. That is to say, wecan extract and classify predicate relationship between sentence NE arguments. The obtained results can be used in ontology enrichment and semantic analysis of sentences for many NLP applications like CAT and Controlled Arabic.

REFERENCES

[1] Asharef, M., Omar, N., Albared, M. (2012). "Arabic NE recognition in crime documents". In *Journal of Theoretical and Applied Information Technology*, Vol. 44. N°. 1, pp. 1-6.

[2] Attia, M., Toral, A., Tounsi, L., Monachini M., Van Genabith, J. (2010). "An automatically built NE lexicon for Arabic". In *LREC 2010, 7th conference on International Language Resources and Evaluation*. Valletta, Malta.

[3] Beaudet S. (2002).Extraction et analyse sémantique automatique des entités spatiales géographiques. Intership report for computer science master.

[4] Bodenhamer, D.J., Harris, T.M., and Corrigan, J., (2013). "Deep mapping and the spatial humanities". In *International Journal of Humanities and Arts Computing*, 7 (1–2), 170–175.

[5] Cooper, D., Donaldson, C., and Murrieta-Flores, P., eds., (2016). *Literary mapping in the digital age. Digital research in the arts and humanities*. London, UK: Routledge.

[6] Daille B., Fourour N., Morin E. (2000). ‘‘Catégorisation des noms propres : une étude en corpus’’. In *Cahiers de Grammaire*, Vol 25, pp. 115-129.

[7] Ehrmann, M. (2008). Les entités nommées, de la linguistique au TAL: statut théorique et méthodes de désambiguïsation. Thesis (PhD). Paris 7 University.

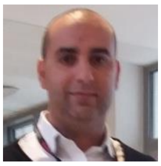
[8] Ehrmann, M., Jacquet, G. (2006). ‘‘Vers une double annotation des entités nommées’’ . In *Traitement Automatique du Langues*, Vol. 47, pp. 63-88.

[9] El Maarouf, I., Villaneau, J., Rosset, S. (2011). ‘‘Extraction de patrons sémantiques appliquées à la classification d'entités nommées’’. In *TALN*. Montpellier.

[10] Gregory, I., Donaldson, D., Murrieta-Flores, P., Rayson, P., (2015). ‘‘Geoparsing, GIS, and textual analysis: current developments in spatial humanities research’’.In *International Journal of Humanities and Arts Computing*, 9 (1), 1–14. doi:10.3366/ijhac.2015.0135.

[11] Leidner, J.L., (2007). Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names. Thesis (PhD). The University of Edinburgh.

AUTHOR



I am **Omar ASBAYOU**, a teacher in Lumière University Lyon 2, and a memberin CRTT laboratory. My research focuses on Arabic language processing. I was an engineer researcher in Techlimed, a company specialised inThe automatic processingof Arabic

JOINT EXTRACTION OF ENTITY AND RELATION WITH INFORMATION REDUNDANCY ELIMINATION

Yuanhao Shen and Jungang Han

School of computer science and Technology, Xi`an University of Posts and Telecommunications, Xi`an, China

ABSTRACT

To solve the problem of redundant information and overlapping relations of the entity and relation extraction model, we propose a joint extraction model. This model can directly extract multiple pairs of related entities without generating unrelated redundant information. We also propose a recurrent neural network named Encoder-LSTM that enhances the ability of recurrent units to model sentences. Specifically, the joint model includes three sub-modules: the Named Entity Recognition sub-module consisted of a pre-trained language model and an LSTM decoder layer, the Entity Pair Extraction sub-module which uses Encoder-LSTM network to model the order relationship between related entity pairs, and the Relation Classification sub-module including Attention mechanism. We conducted experiments on the public datasets ADE and CoNLL04 to evaluate the effectiveness of our model. The results show that the proposed model achieves good performance in the task of entity and relation extraction and can greatly reduce the amount of redundant information.

KEYWORDS

Joint Model, Entity Pair Extraction, Named Entity Recognition, Relation Classification, Information Redundancy Elimination.

1. INTRODUCTION

Extraction of entity and relation, a core task in the field of Natural Language Processing (NLP), can automatically extract the entities and their relations from unstructured text. The results of this task play a vital role in various advanced NLP applications, such as knowledge map construction, question answering, and machine translation.

Supervised extraction of entity and relation usually uses a pipelined or joint learning approach. The pipelined approach treats the extraction task as two serial sub-tasks: named entity recognition [1] and relation classification. The relation classification sub-task first pairs the identified entities according to some pairing strategy, and then classifies the relationships between the entities. Due to the small number of entities that are related, the pipelined model usually generates a large number of pairs of unrelated entities during the pairing phase. Besides, the method also suffered from error propagating and paying little attention to the relevance of the two sub-tasks. To tackle the problems, researchers have conducted a lot of research on the joint learning and achieved better results. Joint Learning refers to extracting entities and classifying relations by one joint model. The joint models usually adopt three research ideas: parameter sharing [2], [3], [4], multi-head selection [5], [6], [7], and table filling [8], [9], [10]. These ideas take advantage of the relevance of sub-tasks to mitigate the error propagation, but still have to deal with the redundant

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 67-81, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101406

information of unrelated entity pairs. Eberts *et al.* [11] proposed a span-based joint model that relies on the pre-trained Transformer network BERT as its core. The model achieved excellent performance but still suffered from the redundancy problem. Zheng *et al.* [12] proposed a method that uses a novel labeling mechanism to convert the extraction task into a sequence labeling task without generating redundant information, but is unable to handle the overlapping relations.

To solve the information redundancy problem and overlapping relation problem described above, we propose a joint model that can handle the sub-tasks of named entity recognition (NER), entity pair extraction (EPE), and relationship classification (RC) simultaneously. The NER sub-task uses the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model [13] to generate word vectors, and takes into account the long-distance dependence of entity labels. The EPE sub-task first uses the proposed Encoder-LSTM network to directly extract the multiple sets of related entity pairs from the sample, then identifies the subject entity and the predicate entity in each entity pair. This approach avoids generating the redundant entity pairs in traditional methods, and also works for overlapping relationship. The RC sub-task uses the traditional relation classification method but taking more abundant and reasonable information as its inputs to improve the performance of classification. In order to solve the problem of information loss between sub-modules and strengthen the interaction between sub-tasks, we designed and added the Connect&LayerNorm layer between sub-modules. We conducted experiments on the public datasets ADE and CoNLL04 to evaluate the effectiveness of our model. The results show that the proposed model achieves good results, and at the same time the model can greatly reduce the amount of redundant information. Compared with other methods, our NER sub-module and RC sub-module have achieved excellent performance. Compared with the traditional LSMT network, the proposed Encoder-LSTM network achieves a significant improvement in performance.

The remainder of the paper is structured as follows. In section 2, we review the related work of named entity recognition, relation classification, and joint extraction tasks. In section 3, we introduce the joint entity and relation extraction model we proposed in detail. In section 4, we first describe the detailed information about the experimental setup, then introduce the experimental results, and analyze the redundancy problem and overlapping relations in detail. Finally, we give the conclusions in Section 5.

2. RELATED WORKS

2.1. Named Entity Recognition

As a basic task in the field of NLP, NER is to identify the named entities. At present, NER has matured in several research directions. Statistical machine learning-based methods [14], [15], [16] require feature engineering and rely more on corpora. Deep learning-based methods [2], [17], [18] can learn more complex features because of their excellent learning ability. Such methods usually use CNN or RNN to learn sentence features, and then use methods such as conditional random files (CRF) to decode the dependencies between labels, and finally identify the entity label of each token. Deep learning-based methods have also been tried to combine with pre-trained language models such as BERT and achieved excellent performance [19].

2.2. Relation Classification

The RC task is a hot research direction in the information extraction task, and its purpose is to determine the category of relationship between two entities in the sentence. Traditional RC methods [20] have good performance on corpora in specific fields, but they rely too much on NLP tools and require a lot of time to design and extract effective features. Due to the advantages of easy learning of complex features, methods based on deep learning [21], [22], [23], [24], [25] have also been widely studied and used by researchers. This type of method uses the original sentence information and the information indicating the entity as inputs to a CNN or RNN to learn the features of a sentence, and finally classifies the constructed relation vector. In recent years, methods based on the combination of deep learning and attention mechanisms have gained significant improvement in performance [26], [27].

2.3. Joint Entity and Relation Extraction

The original intention of the method based on joint learning is to overcome the shortcomings of the pipeline-based method. In the early research, feature-based systems [28] can handle two sub-tasks at the same time, but they rely heavily on the features generated by NLP tools and have the problem of propagation errors. To overcome the problems, some methods based on deep learning have been proposed. In 2016, Gupta *et al.* [9] proposed a Table Filling Multi-Task Recurrent Neural Network (TF-MTRNN) model which simplifies the NER and RC tasks into the Table Filling. In 2017, Zheng *et al.* [3] improved the work of [2], proposed a joint model that does not use NLP tools, and solved the problem of long-distance dependence of entity labels. In 2017, Zheng *et al.* [12] proposed a novel labeling mechanism that converts entity and relation extraction task into a sequence labeling task. This method does not generate redundant information. In 2018, to solve the problem of overlapping relations, Bekoulis *et al.* [5] proposed an end-to-end joint model, which treats the extraction task as a multi-head selection problem, so that each entity can judge the relation with other entities. In 2019, Eberts *et al.* [11] proposed a span-based model that achieves the SOTA performance in the field of joint extraction of entity and relation. This model abandons the traditional BIO/BIOU annotation method and consists of three parts: span classification, spam filtering, and relation classification.

Based on the above research, we propose a joint extraction method for information redundancy elimination. Compared with feature-based methods, this method does not require any additional manual features and NLP tools. Compared with previous methods based on deep learning, our method avoid generating redundant information and can handle the overlapping relations.

3. MODEL

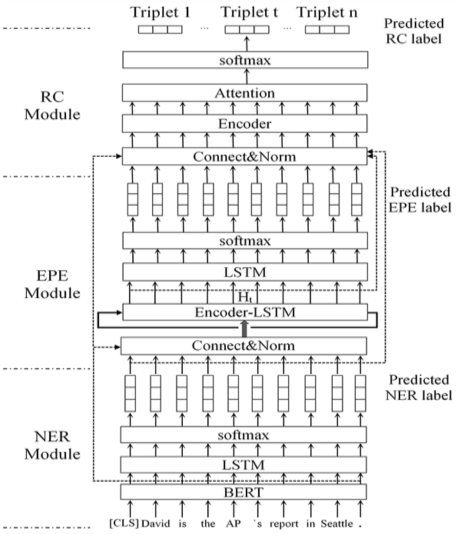


Figure 1. The framework of joint extraction model.

The joint model we proposed consists of three modules: NER module, EPE module, and RC module, as shown in Fig. 1. The NER module identifies the entity label of each token in the text. The EPE module takes sentences and entity labels as inputs, to extracts multiple related entity pairs, and identifies the subject entity and predicate entity for each pair of entities. The RC module classifies the relations.

3.1. Named Entity Recognition

The essence of the NER task is sequence labeling, which assigns a label to each token in the sentence. As shown in Fig. 1, the NER module of the proposed model includes a pre-trained BERT model for generating word vectors, an LSTM decoding layer for solving label dependencies [3], and a softmax layer. At first, the NER module inputs the constructed input vector to the BERT model [13] and obtains the word vector of the sentence. The set of word vectors can be expressed as $S=\{w_1,\cdots,w_t,w_{t+1},\cdots,w_l\}\in R^{l\times d}$, where w_t is the d-dimensional word vector of the t-th word and l is the fixed length of samples. Next, S is inputted to the LSTM decoding layer to perform the following calculation:

$$y_t = LSTM(w_t) \tag{1}$$

where $y_t \in \mathbf{R}^d$, the output of the t-th unit of the decoding layer. Finally the predicted probability of each label of each token of the sentence is obtained through the softmax layer. The predicted probability is expressed as $N=\{p_1,\cdots,p_t,p_{t+1},\cdots,p_l\}\in R^{l\times n_t}$, where n_t stands for the number of entity labels in the NER module. The loss function of a single sample of this module can be expressed as:

$$L_{ner} = -\sum_{j=1}^{l}\sum_{i=1}^{n_t} Y_{ji}\cdot\log(N_{ji}) \tag{2}$$

where $Y\in R^{l\times n_t}$ is the label of a single sample in the NER module.

Considering the correlation between sub-tasks, we use the original sentence information and the prediction information of the label as the input of the EPE module, denoted as $Z\_connect=\{z_1,\cdots,z_t,z_{t+1},\cdots,z_l\}\in R^{l\times(n_t+d)}$, where $z_t=\left[w_t;p_t\right]$. In addition, we perform LayerNrom [29] processing on the combined input, which is expressed as:

$$Z = LayerNorm(Z\_connect) \tag{3}$$

3.2. Entity Pair Extraction

The EPE task is designed to extract multiple pairs of related entities from the inputted sentence. As shown in Fig. 1, the EPE module consists of an Encoder-LSTM network, an LSTM decoding layer, and a softmax layer. Retrieving the pairs of related entities from the sample in a specific order can get a unique sequence, in the form of [(subject entity, predicate entity), ... , (subject entity, predicate entity)]. When the search order is from left to right, the sequence corresponding to the input sample of Fig. 1 takes the form of [(David, AP), (AP, Seattle)]. The order of the sequence is not dependent on whether or not there are overlapping relations among the entities. It is easy to find that the current element pays more attention to the information of the previous element, so we need to retain more new information in each recurrent unit. The addition of new memory in GRU is limited by the old memory, and the update gate in LSTM independently controls how much information in added to the new memory, and the LSTM network can alleviate the problem of gradient disappearance in the traditional RNN model with the long sequence.

Based on the above analysis, the EPE module first uses the Encoder-LSTM network to model the order of the sequence. The output of each recurrent unit of the Encoder-LSTM network is a sentence encoding that contains a pair of related entities. Our proposed Encoder-LSTM network consists of the Encoder structure in Transformer and the LSTM network. The design purpose of the network is to use the Encoder to improve the ability of the recurrent unit to model sentences. The design idea of the network is similar to ConvLSTM [30]. The structure of the Encoder-LSTM network is shown in Fig. 2.

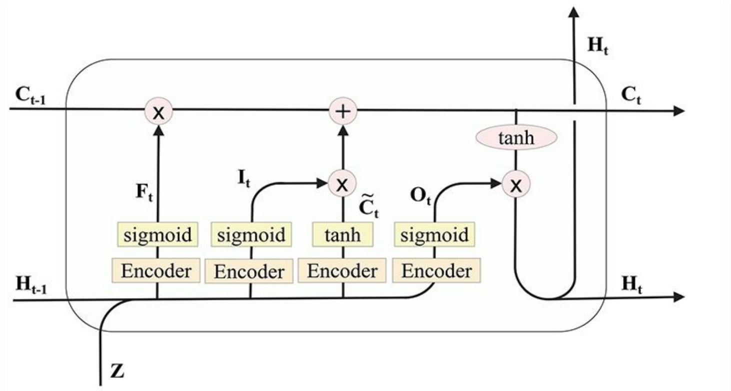


Figure 2. The structure of Encoder-LSTM network.

The network is composed of four parts: the input gate I_t, the forget gate F_t, the output gate O_t, and $\tilde{C}_t$, each of which has its own sentence coding structure $Encoder_i$, $Encoder_f$, $Encoder_o$, and $Encoder_c$ respectively. The calculation details of the Encoder-LSTM network are as follows:

$$I_t = \sigma(Encoder_i([Z;H_{t-1}]) \cdot W_i + b_i) \tag{4}$$

$$F_t = \sigma(Encoder_f([Z;H_{t-1}]) \cdot W_f + b_f) \tag{5}$$

$$O_t = \sigma(Encoder_o([Z;H_{t-1}]) \cdot W_o + b_o) \tag{6}$$

$$\tilde{C}_t = \tanh(Encoder_c([Z;H_{t-1}]) \cdot W_c + b_c) \tag{7}$$

$$C_t = I_t * \tilde{C}_t + C_{t-1} * F_t \tag{8}$$

$$H_t = O_t * \tanh(C_t) \tag{9}$$

where $t = 0,1,\ldots,n$, n stands for the number of related entity pairs being extracted, which is the hyperparameter of the model. C_t and H_t are the state and output of the current recurrent unit respectively. C_{t-1} and H_{t-1} are the state and output before current unit respectively.

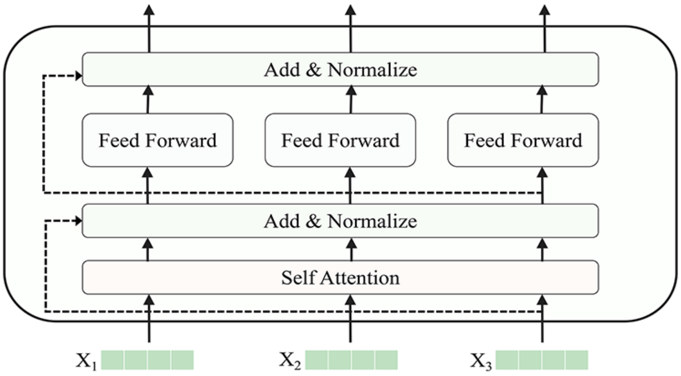


Figure 3. The structure of Encoder.

The function $Encoder_*$ represents the Encoder structure in the Transformer model [29]. The structure of Encoder is shown in Fig. 3. The output of the Encoder-LSTM network is n sets of sentence encoding $H=\{H_1,\ldots,H_t,\ldots,H_n\}\in R^{n\times l\times d_w}$, where d_w is the dimension of the hidden layers of the networks.

The relation type of the entity pair is determined by both the types of the subject entity and the predicate entity. Just knowing the categories of two entities is not sufficient to determine the relationship of the entity pair. Therefore, the EPE module should be able to identify the subject-predicate label of entities in the sentence encoding. The EPE module takes H as input and predicts the subject-predicate label of entities through the LSTM decoding layer and the softmax layer. The prediction probability of the subject-predicate label is expressed as $M=\{M_1,M_2,\cdots,M_n\}\in R^{n\times l\times n_d}$. The loss function of a single sample of this module can be expressed as:

$$L_{epe}=-\sum_{k=1}^{n}\sum_{j=1}^{l}\sum_{i=1}^{n_d}Y_{kji}\cdot\log(M_{kji}) \tag{10}$$

where $Y\in R^{n\times l\times n_d}$ is the subject-predicate label of a single sample, and n_d is the number of subject-predicate labels in the EPE module.

3.3. Relation Classification

The goal of the RC module is to classify the relations of entity pairs that have been specified by subject-predicate labels. As shown in Fig. 1, this module consists of Encoder structure, Attention mechanism, and softmax layer.

The input of the traditional RC task contains not only sentence encoding information but also position information indicating two entities. This is different from the RC task of the previous joint method that only uses inter-entity sentence information [3] or two tokens as input information [5]. To improve the performance of the RC task, we adopt the idea of Position Feature [23] and Position Indicators [24], and use the predicted subject-predicate labels M_t of entities as the position indicator of two entities. In addition, in order to strengthen the interaction between sub-tasks and solve the problem of information loss between sub-tasks , the input of the RC task also includes the information of NER module. Finally, the RC task takes the concatenation of the sentence encoding H_t , the predicted subject-predicate label M_t , the predicted entity label N, and the word vectors S as the input, which can be expressed as $[M_t;H_t;N;S] \in R^{l\times(d_w+n_d+d+n_t)}, t \in 1,2,\ldots n$. Next we perform LayerNorm [29] processing on the input.

$$LN_t = LayerNorm([M_t;H_t;N;S]) \tag{11}$$

To improve the performance, the RC module first uses the Encoder structure to learn sentence features.

$$L_t = Encoder_r(LN_t) \tag{12}$$

then the features are processed by the Attention mechanism [31] to get the relation vector.

$$r_t = \text{Attention}(L_t) \tag{13}$$

where $r_t \in \mathbf{R}^{d_w+n_d}$ represents the relation vector of the t-th related entity pair. Finally, the module obtains the predicted probability $P = \{p_1,\cdots,p_t,\cdots,p_n\} \in R^{n\times n_r}$ of the relation category through the softmax layer, where $p_t \in \mathbf{R}^{n_r}$ is the prediction probability of the relation category of the t-th entity pair, and n_r is the number of relation categories. The loss function of a single sample of this module can be expressed as:

$$L_{rc} = -\sum_{j=1}^{n}\sum_{i=1}^{n_r} Y_{ji} \cdot \log(P_{ji}) \tag{14}$$

where $Y \in R^{n\times n_r}$ is the relation label of a single sample.
Different from the traditional joint model, our model performs the task of entity and relation extraction with three sub-modules, and the final loss is the sum of the three parts: $L_{all} = L_{ner} + L_{epe} + L_{rc}$.

4. EXPERIMENT AND ANALYSIS

4.1. Experimental Setting

DATASET: We conducted experiments on two datasets: (i) Adverse Drug Events, ADE dataset [32]， and (ii) the CoNLL04 dataset [33]. ADE: The dataset includes two entity types Drug and Adverse-Effect and a single relation type Adverse-Effect. There are 4272 sentences and 6821 relations in total and similar to previous work [11], we remove ~120 relations with overlapping entities. Since there are no official test set, we evaluate our model using 10-fold cross-validation similar to previous work [11]. The final results are displayed in F1 metric as a macro-average across the folds. We adopt strict evaluation setting to compare to previous work [5], [11], [34], [35]. CoNLL04: The dataset contains four entity types (Location, Organization, Person, Other) and five relation types (Kill, Live_in, Located_in, OrgBased_in, Work_for). For the dividing rules of the dataset, the experiment follows the method defined by Gupta *et al.* [9]. The original 1441 samples are divided into the training set, the validation set, and the test set, with 910, 243, and 288 samples respectively. We adopt relaxed evaluation setting to compare to previous work [5], [9], [10]. We measure the performance by computing the average F1 score on the test set.

BASELINES: The baselines we used are recent methods for the ADE dataset and the CoNLL04 dataset. Method Li *et al.* (2016) [34] and method Li *et al.* (2017) [35] have achieved good results on the ADE corpus using a joint model based on parameter sharing. Methods Gupta *et al.*(2016)

[9] and Adel&Schütze(2017) [10] formulate joint entity and relation extraction as a table-filling problem. Method Bekoulis *et al.* (2018) [5] employ a bidirectional LSTM to encode words and use a sigmoid layer to output the probability of a specific relation between two words that belong to an entity. Method Eberts *et al.* (2019) [11]proposed a span-based joint model that relies on the pre-trained Transformer network BERT as its core and achieves the best results.

METRICS: To compare with the previous research, the experiment will evaluate the performance of the three sub-tasks by the values of Precision, Recall, and F1-measure. We use two different settings to evaluate performance, namely strict and relaxed. In the strict setting, an entity is considered correct if the boundaries and the type of the entity are both correct; an entity pair is considered correct if the boundaries and the type of the subject entity and the predicate entity are both correct and the argument entities are both correct; a relation is correct when the type of the relation and the argument entity pair are both correct. In the relaxed setting, the experiment will assume that the boundary of the entities is known, an entity is considered correct if the type of any token of the entity is correctly classified; an entity pair is correct when the type of any token of the subject entity and the predicate entity are both correct and the argument entities are both correct; a relation is correct when the type of the relation and the argument entity pair are both correct. The formulas for Precision, Recall, and F1 are as follows.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F_1\text{ - }measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{17}$$

HYPERPARAMETERS: The experiment uses the language Python, the TensorFlow libraries, and the pretrained BERT model of cased_L-12_H-768_A-12 to implement the joint model. For our training on the ADE dataset, the learning rate, the batch size, and the number of iterations are 0.00002, 8, and 40 respectively. The fixed length of the sentence is 128. The value of Dropout is varied for modules and ranging from 0.3 to 0.5. The number of hidden layer units in the Encoder-LSTM network is 96, and the hyperparameter n is 3. The number of layers, the number of heads in Encoder-LSTM network are 2, 4 respectively. We adjusted the hyperparameters of the model for different datasets. The experiment was conducted on an Nvidia DGX-1 server equipped with 8 TeslaV100 GPUs with 128GB of memory per GPU.

4.2. Results

The final experimental results are shown in Table 1. The first column indicates the considered dataset. The second column is the comparable previous methods and ours. The results of the NER task (Precision, Recall, F1) are shown in the next three columns, then follows the results of EPE and RC task. Since the EPE task is proposed for the first time in this paper, there are no comparable results for this task. The last column gives the average F1 of all sub-tasks (Overall F1).

Table 1. Comparisons with the different methods.

Dataset	Methods	NER			EPE			RC			Overall
		P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	F1
ADE	Li *et al.* (2016)	79.50	76.60	79.50	-	-	-	64.00	62.90	63.40	71.45
	Li *et al.* (2017)	82.70	86.70	84.60	-	-	-	67.50	75.80	71.40	78.00
	Bekoulis *et al.* (2018)	84.72	88.16	86.40	-	-	-	72.10	77.24	74.58	80.49
	Eberts *et al.* (2019)	89.26	89.26	89.25	-	-	-	77.77	79.96	78.84	84.05
	Proposed(Encoder-LSTM)	90.54	92.69	91.60	80.17	80.03	80.10	77.63	80.03	78.81	83.50
CoNLL04	Gupta *et al.* (2016)	88.50	88.90	88.80	-	-	-	64.40	53.10	58.30	73.60
	Adel&Schütze (2017)	-	-	82.10	-	-	-	-	-	62.50	72.30
	Bekoulis *et al.* (2018)	93.41	93.15	93.26	-	-	-	72.99	63.37	67.01	80.14
	Proposed(Encoder-LSTM)	91.87	96.45	94.11	68.42	67.16	67.78	66.42	63.23	64.79	75.56

For the ADE dataset, we can observe that in the NER task, the Proposed(Encoder-LSTM) method achieves the best performance. The macro-F1 value of this method is 2.5% higher than that of the Eberts *et al.* (2019) method. In the EPE task, the macro-F1 value of the Proposed(Encoder-LSTM) method is 83%. In the RC task, the Proposed (Encoder-LSTM) method has significantly improved the macro-F1 value compared to the Li *et al.* (2016) method, Li *et al.* (2017) method, and Bekoulis *et al.* (2018) method, and has similar performance compared to the Eberts *et al.* (2019) method.

Considering the results in the CoNLL04 dataset, we can observe that the Proposed(Encoder-LSTM) method achieves the best results in the NER task. Compared with method Bekoulis *et al.* (2018), the Proposed(Encoder-LSTM) method has a significant improvement in F1 value. In the EPE task, the F1 value of the Proposed(Encoder-LSTM) method is 67.78%. In the RC task, the Proposed(Encoder-LSTM) method achieves good results. Compared with method Adel&Schütze(2017), the F1 value of the Proposed(Encoder-LSTM) method is increased by about 2.3%.

It can been seen from the results that our model has achieved excellent performance on both NER and RC modules, but the overall performance of our model is similar to the comparison methods. The reason for the above phenomenon is that the performance of EPE module has become the bottleneck of the overall performance of the model. It can be noticed that there are differences in the performance of the model on the two datasets. After analysis, this is related to the number of samples containing multiple related entity pairs in the dataset. Because our model extracts entity pairs by learning the order relationship of related entity pairs, the ADE dataset can provide more effective data than the CoNLL04 dataset.

Table 2. Ablation tests on the ADE dataset.

Settings	NER F1(%)	EPE F1(%)	RC F1(%)	Overall F1(%)
Proposed	91.59	80.09	78.91	83.50
- LSTM Decoder	91.32	79.73	78.81	83.28
- Connect&LayerNorm	91.56	78.51	76.66	82.24
- Encoder-LSTM	91.04	76.24	74.81	80.69

We conduct ablation tests on the ADE dataset reported in Table 2 to analyze the effectiveness of the Encoder-LSTM network and other components in the model. The performance of the model decreases (~0.2% in terms of Overall F1 score) when we remove the LSTM decoder layer. This shows that the LSTM Decoder layer can strengthen the ability of model to learn the dependency between entity tags [3]. The performance of EPE and RC tasks decreases (~1.2%) when we remove the Connect&LayerNorm layer of the RC module and only use the predicted subject-predicate labels and the sentence encoding as inputs for the RC task. This shows that the predicted entity labels and the word vectors provide meaningful information for the RC component and this approach can solve the problem of information loss between subtasks. There is also a reasonable explanation that this approach is similar to the residual structure [29], which can alleviate the problem of gradient disappearance. Finally we conduct experiments by removing the Encoder-LSTM network and substituting it with a LSTM network. This approach leads to a slight decrease in the F1 performance of the NER module, while the performance of the EPE task and the RC task decreased by about 2%. This happens because the Encoder structure in the Encoder-LSTM network can improve the ability of recurrent units to model sentences.

Table 3. Model performance for different hyperparameter values.

Hyper-parameters	value	NER F1(%)	EPE F1(%)	RC F1(%)	Overall F1(%)
Encoder-layer	2	91.46	78.86	77.20	82.51
	3	91.59	80.09	78.81	83.50
	4	91.59	78.62	77.36	82.52
	32	91.25	77.58	75.75	81.52
hidden size	64	90.73	78.42	77.17	82.10
	96	91.59	80.09	78.81	83.50
	128	91.60	78.96	77.43	82.66

We also evaluated the impact of different hyperparameter values in the Encoder-LSTM network on model performance. Table 3 show the performance of our model on the ADE dataset for different values of Encoder layer and hidden size hyperparameters in Encoder-LSTM network, respectively. It can be observed that the model achieves the best performance with the encoder layers of 3 and the hidden size of 96.

4.3. Analysis of Redundancy and Overlapping Relation

The redundancy problem means that the model generates and has to evaluate a large number of unrelated entity pairs. The method we proposed directly extracts the pairs of related entities from the samples, without producing redundant information in the traditional sense. In order to solve the problem of different numbers of triples in different samples, our method uses the hyperparameter n to specify the number of related entity pairs extracted in each sample, but this approach leads to the inevitable generation of redundant sentence coding in the EPE module.

Because the redundancy of the model is proportional to the number of times the model classifies the relationships, we use this number to evaluate and compare the redundancy of different models. The method proposed by Miwa *et al.* [8] labels m(m-1)/2 cells in the entity and relation table to predict possible relationships, where m is the sentence length. The method by Zheng *et al.* [3] and Bekoulis *et al.* method [5] first identify entities, and then classify the relationships between each pair of entities, so these two methods classify the relationships k^2 times, where k is the number of identified entities. Our method directly extracts the related entity pairs and then classifies the relationships of each entity pair. Therefore, the number of times our method classifies the relationships is equal to the number n of related entity pairs extracted by the model, and n is the hyperparameter of our model. Based on the above analysis, we obtain a statistical table of the number of times the model classifies the relationships, as shown in Table 4.

Table 4. Redundancy of different models.

Methods	times
Miwa&Sasaki(2014)	$m(m$-1)/2
Zheng *et al.*(2017)	k^2
Bekoulis *et al.*(2018)	k^2
Proposed	n

The parameter m, k and n in the Table 4 stand for the sentence length, the number of entities, and the hyperparameter of our model respectively.

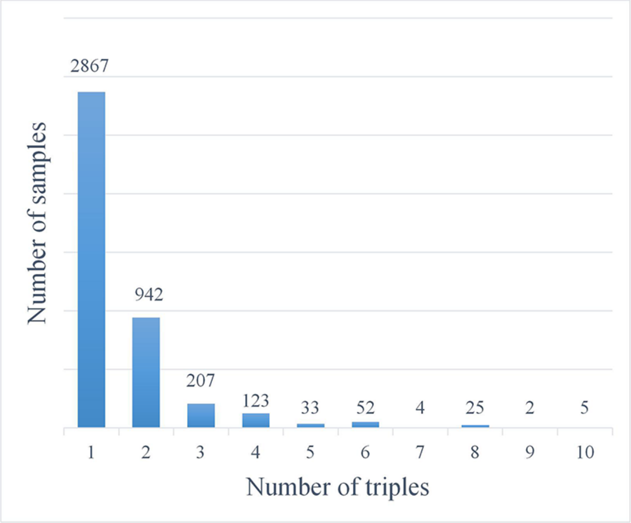


Figure 4. Distribution histogram of the number of triples in the ADE dataset.

After analysis, more than 99% of the word pairs are irrelevant [9]. About 45% of the samples contain more than 3 entities, and the related entity pairs only account for a small part of all entity pairs. As shown in Fig. 4, about 77% of the samples contain only one triple, and about 96% of the samples contain no more than three triples. For example, assuming the input sample is shown in Fig. 1, then m, k, and n take the value of 128, 3, and 3 respectively. The number of times of Miwa&Sasaki(2014) method, Zheng *et al.* (2017) method, Bekoulis *et al.* (2018) method, and our method are 8128, 9, 9, and 3 respectively. Therefore, if the value of n is appropriately selected, the redundancy of the proposed method is much smaller than that of other methods.

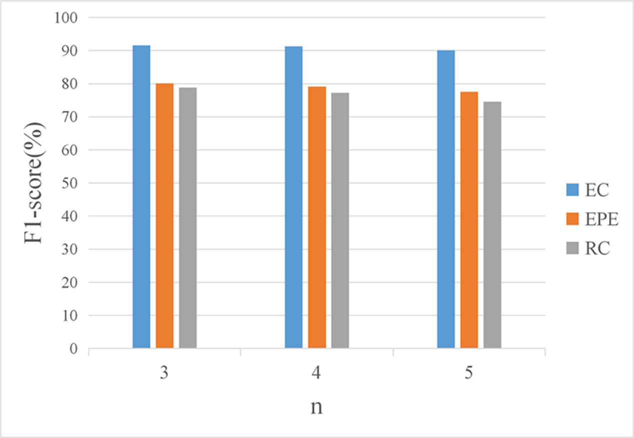


Figure 5. Model performance for different values hyper-parameter n.

Since the redundancy of our model depends on the value of n, to evaluate the impact of redundancy on performance, we conduct experiments based on different values of n, and the results are shown in Fig. 5. It can be observed that the model has the best overall performance when the hyperparameter n is 3. The change of the value of n has little effect on the performance of the NER module and the EPE module. As the value of n increases, the performance of the RC module and the EPE module decreases significantly. After analysis, this phenomenon is related to the distribution of the number of triples in the sample. Theoretically, as the value of n increases, the EPE module can better model the sequence information of related entity pairs. However, it can be seen from Fig. 4 that there are very few useful data when n is greater than 3. At this time, the increase of the value of n not only cannot help the learning of the EPE module, but also seriously interferes with the training of the model. Based on the above analysis, the choice of n value should depend on the distribution of the number of triples in the sample. If the samples in the corpus contain sufficient related entity pairs, our model will perform better, otherwise our model will perform not well.

There are two types of overlapping relations [36]. The first type is that an entity has relations with multiple other entities. Our EPE module uses the order information of the sequence of related entity pairs to extract entity pairs. This type of overlapping relations does not affect the unique order of the sequence. Therefore, the proposed method works well with such situation. The second type of overlapping relations refers to the multiple relationships between one entity pair. Since this situation does not exist in the ADE dataset and the CoNLL04 dataset, we treat the RC task as a multiclass classification task to evaluate which relationship category the entity pair belongs to. Specifically, our model uses the softmax function as the activation function of the output layer, and the categorical cross-entropy as the loss function. If we need to deal with the second kind of overlapping relations, we can treat the RC task as a multilabel classification task, such as the Bekoulis method [5], to evaluate the various relationships that may exist in the entity pair. Specifically, our model uses the sigmoid function as the activation function of the output layer, and uses binary cross-entropy as the loss function.

5. Conclusion

We have presented the joint extraction model based on entity pair extraction with information redundancy elimination. The model first extracts multiple sets of sentence encoding from the sample, then identifies the subject entity and the predicate entity in each set of sentence encoding, and finally classifies the relationship between the two entities. We also propose the Encoder-LSTM network, which improves the ability of recurrent units to model sentences. By conducting experiments on the ADE dataset and the CoNLL04 dataset, we verified the effectiveness of the method and evaluated the performance of the model. Compared with other joint extraction methods, our method solves the problem of redundancy of unrelated entity pairs while achieving excellent performance, and can handle the cases with overlapping relationships.

Since the performance of our EPE module limits the overall model, as the future work we will try to optimize the solution of the EPE. And we plan to verify the proposed method on more actual datasets.

Acknowledgements

The work is partially supported by the Shaanxi Key Laboratory of Network Data Analysis and Intelligence Processing. Our deepest gratitude also goes to the anonymous reviewers for their suggestions for improving this paper.

References

[1] J. Li, A. Sun, J. Han, C. Li “A Survey on Deep Learning for Named Entity Recognition,” IEEE *Transactions on Knowledge and Data Engineering*, vol. pp, no. 99, pp. 1–1, Mar. 2020.

[2] M. Miwa and M. Bansal, “End-to-end relation extraction using LSTMs on sequences and tree structures,” in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, Berlin, Germany, 2016, pp. 1105–1116.

[3] S. Zheng, Y. Hao, D. Lu, H. Bao, J. Xu, H. Hao, and B. Xu, “Joint entity and relation extraction based on a hybrid neural network,” *Neurocomputing*, vol. 257, pp. 59–66, Sep. 2017.

[4] Z. Peng, S. Zheng, J. Xu, Z. Qi, and X. Bo, “Joint Extraction of Multiple Relations and Entities by Using a Hybrid Neural Network,” in *Proc. Lect. Notes Comput. Sci.*, Nanjing, China, 2017, pp. 135–146.

[5] G. Bekoulis, J. Deleu, T. Demeester, and C. Develder, “Joint entity recognition and relation extraction as a multi-head selection problem,” *Expert Systems with Applications*, vol. 114, pp. 34–45, Dec. 2018.

[6] G. Bekoulis, J. Deleu, T. Demeester, and C. Develder, “An attentive neural architecture for joint segmentation and parsing and its application to real estate ads,” *Expert Systems with Applications*, vol. 102, pp. 100–112, Jul. 2018.

[7] X. Zhang, J. Cheng, and M. Lapata, “Dependency parsing as head selection,” in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguist.*, Valencia, Spain, 2017, pp. 665–676.

[8] M. Miwa and Y. Sasaki, “Modeling joint entity and relation extraction with table representation,” in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, Doha, Qatar, 2014, pp. 1858–1869.

[9] P. Gupta, H. Schütze, and B. Andrassy, “Table filling multi-task recurrent neural network for joint entity and relation extraction,” in *Proc. Int. Conf. Comput. Linguist.*, Osaka, Japan, 2016, pp. 2537–2547.

[10] H. Adel and H. Schütze, “Global normalization of convolutional neural networks for joint entity and relation classification,” in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, Copenhagen, Denmark, 2017, pp. 1723–1729.

[11] M. Eberts, A. Ulges, “Span-based Joint Entity and Relation Extraction with Transformer Pre-training,” *arXiv*, 2019.

[12] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu, “Joint extraction of entities and relations based on a novel tagging scheme,” in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, Vancouver, Canada, 2017, pp. 1227–1236.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018.

[14] F. Alam and I. Asiful, , “A proposed model for Bengali named entity recognition using maximum entropy markov model incorporated with rich linguistic feature set,” in *ACM Int. Conf. Proc. Ser.*, Dhaka, Bangladesh, 2020.

[15] L.Gong, X. Liu, X. Yang, L. Zhang, Y. Jia, and R. Yang, “CBLNER: A Multi-models Biomedical Named Entity Recognition System Based on Machine Learning,” in *Lect. Notes Comput. Sci.*, Nanchang, China, 2019, pp. 51–60.

[16] A. Anandika, S. Mishra, “A study on machine learning approaches for named entity recognition,” in *Proc. -Int. Conf. Appl. Mach. Learn.*, Bhubaneswar, India, 2019, pp. 153–159.

[17] H. Wei *et al.*, “Named Entity Recognition From Biomedical Texts Using a Fusion Attention-Based BiLSTM-CRF,” *IEEE Access*, vol. 7, pp. 73627–73636, June. 2019.

[18] S. Zhang, Y. Shen, J. Gao, J. Chen, J. Huang, and S. Lin, “A Multi-domain Named Entity Recognition Method Based on Part-of-Speech Attention Mechanism,” in *Commun. Comput. Info. Sci.*, Kunming, China, 2019, pp. 631–644.

[19] Z. Dai, X.Wang, P. Ni, Y. Li, G. Li, and X. Bai, “Named entity recognition using bert bilstm crf for chinese electronic health records,” in *Proc. Int. Congr. Image Signal Process., BioMed. Eng. Inf.*, Huaqiao, China, 2019, pp. 1–5.

[20] B. Rink and S. Harabagiu, “UTD: Classifying semantic relations by combining lexical and semantic resources,” in *Proc. Int.Workshop Semant. Evaluation*, Uppsala, Sweden, 2010, pp. 256–259.

[21] X. Guo, H. Zhang, H. Yang, L. Xu and Z. Ye, "A Single Attention-Based Combination of CNN and RNN for Relation Classification," *IEEE Access*, vol. 7, pp. 12467–12475, 2019.

[22] C. Zhang *et al.*, “Multi-Gram CNN-Based Self-Attention Model for Relation Classification,” *IEEE Access*, vol. 7, pp. 5343–5357, 2019.

[23] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, “Relation classification via convolutional deep neural network,” in *Proc. COLING - Int. Conf. Comput. Linguist.*, Dublin, Ireland, 2014, pp. 2335–2344.

[24] D. Zhang and D. Wang, “Relation classification via recurrent neural network,” *CoRR*, vol. abs/1508.01006, 2015.

[25] X. Huang, J. Lin, W. Teng and Y. Bao, “Relation classification via CNNs with Attention Mechanism for Multi-Window-Sized Kernels,” in *Proc. IEEE Adv. Inf. Technol., Electron. Autom. Control Conf.,* Chengdu, China, 2019, pp. 62-66.

[26] L. Wu, H. Zhang, H. Yang, Y. Yang, X. Liu, and K. Gao, “Dynamic Prototype Selection by Fusing Attention Mechanism for Few-Shot Relation Classification,” in *Lect. Notes Comput. Sci.*, Phuket, Thailand, 2020, pp. 431–441.

[27] S. Zhang, D. Zheng, X. Hu, and M. Yang, “Multi-Channel CNN Based Inner-Attention for Compound Sentence Relation Classification,” *IEEE Access*, vol. 7, pp. 141801–141809, 2019.

[28] Q. Li and H. Ji, “Incremental joint extraction of entity mentions and relations,” in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, Baltimore, MD, United states, 2014, pp. 402–412.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Adv. neural inf. proces. syst.*, Long Beach, CA, United states, 2017, pp. 5999–6009.

[30] X. Shi *et al.*, “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting,”in *Conf. Advances in Neural Information Processing Systems*, Montreal, QC, Canada, 2015, pp. 802 - 810.

[31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. Int. Conf. Mach. Learn.*, Lile, France, 2015, pp. 2048–2057.

[32] Gurulingappa *et al.* “Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports,”*Biomedical Informatics*, vol. 45, no. 5, pp. 885–892, 2012.

[33] D. Roth and W. Yih, “A linear programming formulation for global inference in natural language tasks,” in *Proceedings of the Eighth Conference on Computational Natural Language Learning*, Boston, MA, USA, 2004, pp. 1–8.

[34] F. Li, Y. Zhang, M. Zhang, and D. Ji, “Joint Models for Extracting Adverse Drug Events from Biomedical Text,”in *Conf. Int. Joint Conf. Artif. Intell.*, New York, NY, United states, 2016, pp. 2838 - 2844.

[35] F. Li, M. Zhang, G. Fu, and D. Ji, “A neural joint model for entity and relation extraction from biomedical text,”*Bmc Bioinformatics*, vol. 18, no. 1, pp. 198, 2017.

[36] S. Wang, Y. Zhang, W. Che, and T. Liu, “Joint extraction of entities and relations based on a novel graph scheme,” in *Proc. IJCAI Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, 2018, pp. 4461–4467.

AUTHORS

Yuanhao Shen

He received the B.E. degree from Xi`an University of Posts and Telecommunications, China, 2018. He is currently pursuing the master’s degree in the College of Computer Science and Technology. Hisl research interests include natural language processing and deep learning.

Jungang Han

He is a professor at Xi’an University of Posts and Telecommunications. He is the author of two books, and more than100 articles in the field of computer science. His current research interests include artificial intelligence, deep learning for medical image processing.

DOMAIN-TRANSFERABLE METHOD FOR NAMED ENTITY RECOGNITION TASK

Vladislav Mikhailov[1, 2] and Tatiana Shavrina[1,2]

[1]Sberbank, Moscow, Russia
[2]Higher School of Economics, Moscow, Russia

ABSTRACT

Named Entity Recognition (NER) is a fundamental task in the fields of natural language processing and information extraction. NER has been widely used as a standalone tool or an essential component in a variety of applications such as question answering, dialogue assistants and knowledge graphs development. However, training reliable NER models requires a large amount of labelled data which is expensive to obtain, particularly in specialized domains. This paper describes a method to learn a domain-specific NER model for an arbitrary set of named entities when domain-specific supervision is not available. We assume that the supervision can be obtained with no human effort, and neural models can learn from each other. The code, data and models are publicly available.

KEYWORDS

Named Entity Recognition, BERT-based Models, Russian Language

1. INTRODUCTION

Named Entity Recognition (NER) is a rapidly developing NLP task that aims to extract mentions of entities from texts and label them with the predefined semantic types such as PER (person), LOC (geographical location), ORG (organization), etc. Regarding the definition of named entity (NE), we follow [1] who classify NEs into two main categories: generic entities (e.g., person, geographical location and organization) and domain-specific entities (e.g., genes and terms). The increasing need to process a large amount of data facilitates the development of NER models, particularly in specialized fields that exhibit an extensive variety of NEs. Nevertheless, a corpus of high-quality NER annotations is difficult and expensive to obtain since it requires domain expertise. In the practical setting, little or no domain supervision is available, specifically for the Russian language.

Hence, we exploit a method to construct a domain-specific labelled dataset for NER without human supervision. The dataset is then used to train a domain-specific NER model for extraction of knowledge graph (KG) entities in the downstream task of question answering (QA). The methodology includes the following subtasks. (1) Construction of a domain-specific entity vocabulary using the Wikipedia category graph and a set of seed categories. (2) Construction of a domain-specific text corpus from the corresponding Wikipedia articles and publicly available resources. (3) Training a general-domain NER model for preliminary text annotation. (4) Implementation of the morphology-based algorithm for automatic text annotation using the entity vocabulary. (5) Assembly of domain-specific NER dataset from annotations obtained with the general-domain NER model and the morphology-based algorithm. (6) Training the final domain-specific NER model.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 83-92, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101407

We consider that the methodology is domain-transferable and can be modified to depend upon the target domain and available data. We conduct experiments in the domain of Russian History which is represented by a rich set of both generic (historical figures, names of battles, wars and cultural properties) and domain-specific entities (historical terms, concepts and phenomena). Besides, we report the results of the NER model on an additional test set which consists of 1,795 manually annotated samples from the Unified State History Exam.

2. Related Works

Domain-specific NER using external knowledge has been a subject of recent research. The main idea is to use publicly available text data to extract features for NER models [2] or combine texts with non-linguistic data such as game states [3]. Current state-of-the-art NER models incorporate external knowledge from BERT-based language models [4]. Nevertheless, the direct application of pre-trained language models in the domain-specific scenario may result in unsatisfactory performance due to shift in word distribution or an insufficient representation of domain knowledge in the pretraining corpora. The problem has been alleviated in the scientific and biomedical domains for English. SciBERT [5] and BioBERT [6] achieved significant improvement over original BERT [7] on a number of downstream tasks within the domains. However, a vast amount of domain texts is required to train domain-specific language models. Another approach involves training NER models using dictionaries from domain-specific KGs [8]. The model achieved strong performance competitive to supervised benchmarks. Still, there is a lack of publicly available domain-specific KGs and domain-specific NER datasets for the Russian language that can be used as a source of domain knowledge.

A common NER solution is to fine-tune a BERT-based model on the available supervision. However, the target domain usually differs from the pre-training corpus which may result in the unsatisfactory performance of the model on the downstream task. Recently, unsupervised domain adaptation of language models has shown quality improvement in a number of NLP tasks, including sequence labelling [9]. [10] study unsupervised domain adaptation of BERT in the limited labelled and unlabelled data scenarios. The results report that fine-tuning of the pre-trained language model even on a small amount of domain data (1,000 samples) before training on the downstream task improves performance. In our work, we apply the domain adaptation procedure to compare the performance of the trained BERT-based NER models. We now describe the data collection pipeline, automatic annotation procedure and the model training.

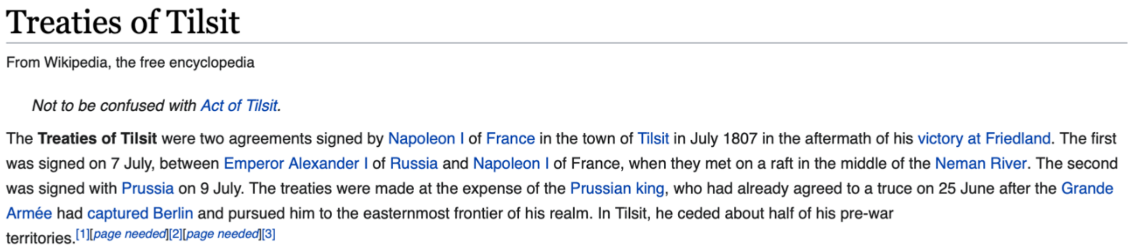
3. Method

We investigate a setting when no domain-specific supervision and annotation resources are available. We proceed from the assumption that a model can learn from another model, and domain-specific texts can be annotated without human effort using only an entity vocabulary. Section 3.1 describes the construction of the domain-specific entity vocabulary V and domain-specific text corpus D using publicly available resources. We also train a general-domain NER model on an available general-domain NER dataset. We refer to this model as RuBERT-general and use it as a source of external knowledge particularly about generic entities (see Section 3.2.1). Besides, we developed a morphology-based annotation algorithm over V which serves as a source of domain knowledge (see Section 3.2.2). Each text from D was annotated with RuBERT-general and the morphology-based annotation algorithm. We then unified the annotations obtained from the previous steps (see Section 3.2.3). The general-domain NER dataset is further combined with the assembled domain-specific NER. This allows us to train a model that is aware of both domain and general knowledge. Finally, we train the domain-specific NER model on the resulting annotations. We describe the training procedure and the results of the experiments in Section 4. Section 5 presents the discussion of the method and outlines future work. Section 6 highlights the main contribution of this work and draws the conclusion.

3.1. Data Collection

Data collection pipeline consists of two stages: construction of a domain-specific entity vocabulary V and construction of a domain-specific text corpus D. If a primary V is not available, it is important to conduct the collection procedure thoroughly for the model to learn relevant domain-specific entities as well as contexts in which they can occur.

In the first stage, we used Wikipedia API to retrieve a list of titles of Wikipedia articles related to the domain of Russian History. We traversed the Wikipedia graph over a set of seed categories, e.g. ”History of Russia, by periods”, “Battles involving Kievan Rus’”, ”Wars involving Russia” and etc. For each title in the retrieved list, we parsed a text of the corresponding article, a summary, a list of categories and a list of interlinks in the article. Each interlink and Wikipedia title are considered to be NEs and added to V. We collected statistics on how frequently each NE ∊ V is referred to in the Wikipedia articles. We also built a frequency vocabulary for the list of categories for each NE. NEs with the interlink frequency of lower than 2 and the category frequency of lower than 3 are discarded from V. Furthermore, we filtered V with a set of seed words for category names such as “USSR”, “war”, “battle”, “culture”, “politics” and etc. Figure 1 shows an example of NEs for the Wikipedia article “Treaties of Tilsit”, where the interlinks are underlined and coloured in blue. The interlink frequency of the NE “Treaties of Tilsit” is 202. The NE relates to the following categories: “19th-century treaties”, “Napoleonic Wars treaties”, “1807” and etc.



Treaties of Tilsit

From Wikipedia, the free encyclopedia

Not to be confused with Act of Tilsit.

The **Treaties of Tilsit** were two agreements signed by Napoleon I of France in the town of Tilsit in July 1807 in the aftermath of his victory at Friedland. The first was signed on 7 July, between Emperor Alexander I of Russia and Napoleon I of France, when they met on a raft in the middle of the Neman River. The second was signed with Prussia on 9 July. The treaties were made at the expense of the Prussian king, who had already agreed to a truce on 25 June after the Grande Armée had captured Berlin and pursued him to the easternmost frontier of his realm. In Tilsit, he ceded about half of his pre-war territories.[1][*page needed*][2][*page needed*][3]

Figure 1. A summary of the Wikipedia article "Treaties of Tilsit". Each interlink and its corresponding title are added to the domain-specific entity vocabulary.

We additionally retrieved NEs from publicly available structured resources such as glossaries, educational books and tasks from the Unified State History Exam to enrich **V**. Consider the examples of the NEs in the resulting **V**. Generic NEs include “Christianization of Kievan Rus'”, “Vladimir the Great”, “Seven Years’ War”, “Saint Petersburg”, “Battle of Borodino” and etc. Domain-specific entities include “Oprichnina”, “Streltsy”, “Bondhold”, “Time of Troubles” and etc. An important note should be made that the quality of the **V** depends on the depth of the Wikipedia graph traversal due to a number of ambiguous titles and a large degree of cross-reference in the articles. We experimented with the traversal depth values of 1 and 2. The first vocabulary consists of nearly 17,000 NEs, while the second one includes 95,000 NEs, mostly redundant ones. The quality of the resulted vocabularies was validated manually. Towards a better quality of the domain NEs, we used the first vocabulary in our experiments.

In the second stage, we constructed a domain-specific corpus **D** which consists of texts from publicly available resources such as books about Russian History (autobiographical fiction, educational books, documentaries and series of lectures), the Wikipedia article summaries, the

Unified State History Exam variants and tests on Russian History. The size of the corpus is 5.65M tokens. We used **D** to obtain annotations with the general-domain NER model and the morphology-based algorithm. Besides, we used **D** for domain adaptation of the final NER model. We provide details in the next section.

3.2. Automatic Annotation Procedure

3.2.1. General-domain Annotation

RuBERT [11] is a monolingual BERT model for the Russian language which outperforms multilingual BERT over a number of NLP tasks for Russian including NER (http://docs.deeppavlov.ai/en/master/features/models/ner.html). In our experiments, we use RuBERT architecture for the **RuBERT-general** model and for the domain-specific NER model. Note that these are two different models. We obtained **RuBERT-general** by fine-tuning RuBERT model on WikiNER [12]. WikiNER is a general-domain NER dataset which has proved its quality for the NER task and is widely used in multiple languages. The dataset consists of 204,778 samples in the Inside-Outside-Beginning (IOB) scheme with 4 semantic labels: "PER", "LOC", "ORG" and "MISC". The data was randomly split into 163,822 train samples, 20,478 dev samples and 20,478 test samples. We trained the model for 5 epochs with default parameters using Hugging Face NLP-library [13]. Evaluation results of the general-domain NER model on the dev and test sets are presented in Table 1.

Table 1. Evaluation of the RuBERT-general model.

	Precision	**Recall**	**F1**
Dev	0.910	0.914	0.912
Test	0.913	0.916	0.914

Each text ∈ **D** was annotated with **RuBERT-general**. Figure 2 shows an example of the text "Ivan The Terrible introduced oprichnina. " and its **RuBERT-general** annotation. w of indices ∈ {0, …, 4} refers to a word token. "B" (begin) prefix denotes the first token of an entity mention and "I" (inside) prefix corresponds to the tokens following it.

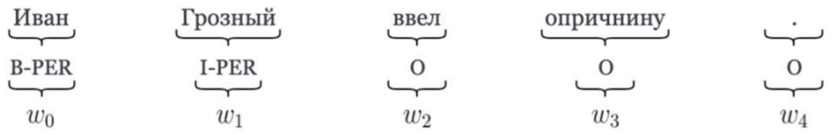


Figure 2. An example of the text annotated with the RuBERT-general model.

3.2.2. Domain-specific Annotation

To provide annotations on domain-specific entities, we implemented a morphology-based algorithm for automatic text annotation using only **V**. The algorithm is closely related to [14]. The work proposes assembly of a general-domain NER dataset, called SESAME, using DBpedia as a source of NEs & their semantic types and Wikipedia as a document collection. A basic idea is to detect mentions of DBpedia entities in a Wikipedia text based on the exact match. Each entity mention (i.e. a character span) is tagged with its corresponding semantic type from DBpedia. Figure 3 shows an example of the annotation. The DBpedia entities "John Smith" and

“Rio de Janeiro” are tagged with their corresponding labels “PER” and “LOC” in the IOB format. w of indices ∊ {0, ..., 7} denotes a word token, while s corresponds to a sentence token. In contrast, we construct the entity vocabulary from scratch. Semantic labels are the general-domain NER predictions or the “MISC” label assigned with the morphology-based algorithm. We now describe the annotation procedure with the algorithm.

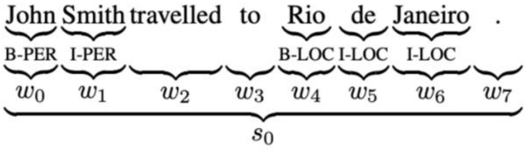

John Smith travelled to Rio de Janeiro .
B-PER I-PER B-LOC I-LOC I-LOC
w_0 w_1 w_2 w_3 w_4 w_5 w_6 w_7
s_0

Figure 3. An example of the text annotation algorithm by [14].

In an offline step, we split each text into sentences using rusenttokenize library, a rule-based sentence segmenter for Russian (https://pypi.org/project/rusenttokenize). Each NE ∊ **V** and each text from the corpus was tokenized with Spacy Russian Tokenizer (https://github.com/aatimofeev/spacy_russian_tokenizer) and lemmatized with pymorphy2 [15]. Next, we identify mentions of NEs ∊ **V** in the processed input text. Each mention is tagged with the *“MISC”* label in the IOB format to mark the boundaries of the entity. The remaining tokens are tagged with *“O”* (outside). The overall scheme is outlined in Figure 4, where *t* corresponds to the input text and *e* denotes the detected mentions of entities from **V**.

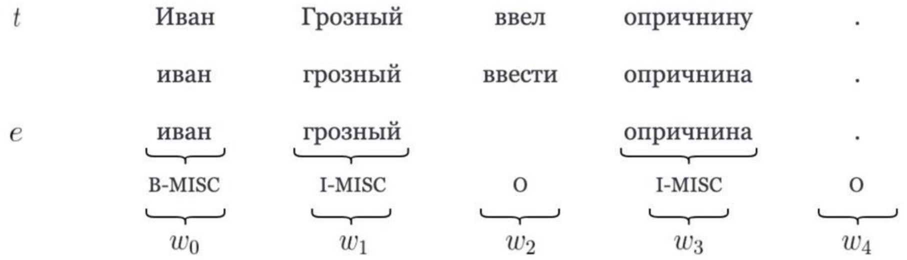

t Иван Грозный ввел опричнину .
иван грозный ввести опричнина .
e иван грозный опричнина .
B-MISC I-MISC O I-MISC O
w_0 w_1 w_2 w_3 w_4

Figure 4. A graphical structure for annotation procedure with the morphology-based algorithm.

Note that one can use a general-domain KG such as Wikidata to map instances of NEs with their corresponding labels to enrich the semantic label inventory, e.g. “city” ⇒ “LOC”. Another option for the algorithm modification is to manually annotate all entities in **V** only once, and then assign the labels to all entity mentions. For instance, “Ivan The Terrible“ ⇒ “B-PER I-PER” or "Treaties of Tilsit” ⇒ “B-MISC I-MISC I-MISC”. During the annotation procedure, the corresponding set of predefined labels is then can be assigned to each entity mention.

3.2.3. Annotation Unification

Therefore, each text ∊ **D** has two annotations. The next step is to unify the obtained annotations. If an entity mention is annotated by both the **RuBERT-general** model and the morphology-based algorithm, we prefer the **RuBERT-general** annotation to the morphology-based one. This allows for preserving relevant semantic labels for generic NEs. The unification procedure is illustrated in Figure 5, where the NE "Ivan The Terrible" (w_0 and w_1) is tagged with the “B-PER” and “I-PER” labels by **RuBERT-general**. The NE "oprichnina" (w_3) is tagged with the “MISC” label by the morphology-based algorithm. The remaining tokens are tagged with the “O” label. Annotation obtained with the **RuBERT-general** model is marked as general-domain annotation. Annotation obtained with the morphology-based algorithm is referred to as domain-specific annotation. Annotation unification corresponds to the result of the procedure.

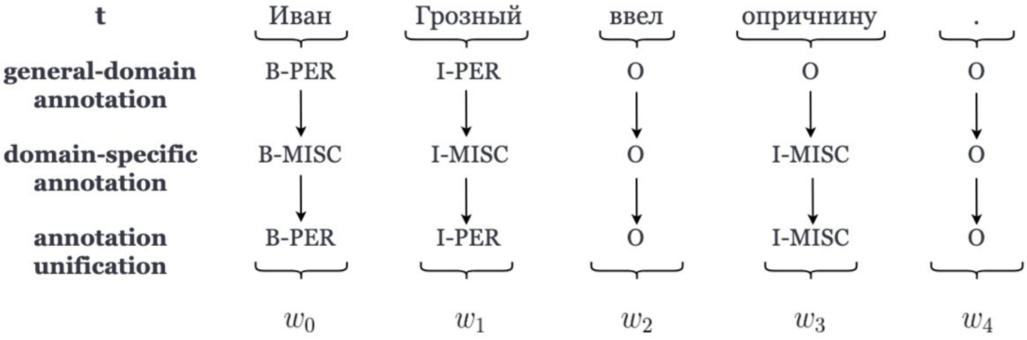
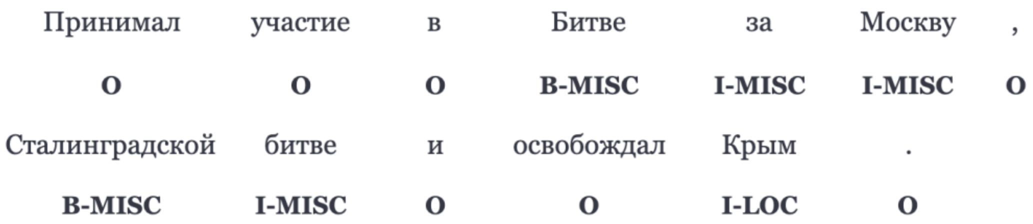


Figure 5. An overall scheme for the annotation unification procedure.

After the unification procedure, we discarded duplicates and samples that are labelled with only the “O” tag. The size of the filtered NER dataset is 160 410 samples. Since the targeted domain exhibits both generic and domain-specific entities, we combined the domain-specific NER dataset with the WikiNER dataset. The total size of the assembled NER dataset is 402,027 samples. Consider an example of the dataset sample "Took part in the Battle of Moscow, the battle of Stalingrad and liberated the Crimea." and its corresponding annotation (see Figure 6).



Принимал	участие	в	Битве	за	Москву	,
O	**O**	**O**	**B-MISC**	**I-MISC**	**I-MISC**	**O**
Сталинградской	битве	и	освобождал	Крым	.	
B-MISC	**I-MISC**	**O**	**O**	**I-LOC**	**O**	

Figure 6. A sample from the constructed NER dataset.

4. EXPERIMENTS AND RESULTS

In our experiments, we train three BERT-based models on the constructed NER dataset: (1) **BERT-original**, a multilingual BERT model; (2) **RuBERT-original**, a monolingual BERT model for Russian; and (3) **RuBERT-adapted**, an adapted monolingual BERT model for Russian. Specifically, we apply domain adaptation to RuBERT over **D** before training for the NER task. We compare the performance of the three models. To obtain the **RuBERT-adapted** model, we fine-tuned RuBERT language model on **D** using Hugging Face library. The model was trained with default parameters for 12 epochs. The best perplexity achieved is 11.2. Notably, fine-tuning of the language models on domain-specific data may lead to perplexity decrease while increasing the downstream task performance [16]. We trained **BERT-original**, **RuBERT-original** and **RuBERT-adapted** models via Hugging Face library with default parameters for 5 epochs. The training results are shown in Table 2.

Table 2. Evaluation of BERT-original, RuBERT-original and RuBERT-adapted models on the dev and test sets.

	Dev			**Test**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BERT-original	0.908	0.908	0.908	0.907	0.906	0.907
RuBERT-original	**0.913**	**0.913**	**0.912**	**0.914**	**0.916**	**0.915**
RuBERT-adapted	0.892	0.896	0.894	0.891	0.897	0.894

Knowing that the annotation samples in the collected NER dataset may contain automatic annotation errors or be inconsistent, we manually annotated extra 1,795 samples from the Unified State History Exam. The additional test set allows to assess the performance of the models reliably, disregarding this potential deficiency. The models are evaluated over a full inventory of semantic labels. We computed Precision, Recall and F1-score for each label by taking their average over the weighted number of instances. We present the results for each model in Tables 3, 4 and 5. B and I refer to the beginning and the inside prefixes respectively. AVG is the weighted average metric. Support corresponds to the number of instances.

Table 3. Evaluation of BERT-original on the additional test set over a weighted inventory of semantic labels.

	B				**I**					
Metric/Label	**LOC**	**MISC**	**ORG**	**PER**	**LOC**	**MISC**	**ORG**	**PER**	**O**	**AVG**
Precision	0.76	0.58	0.63	0.87	0.89	0.75	0.74	0.93	0.75	0.78
Recall	0.52	0.47	0.42	0.88	0.89	0.49	0.55	0.90	0.95	0.78
F1	**0.62**	0.52	**0.50**	0.87	**0.89**	0.59	0.63	0.92	0.84	0.76
Support	96	478	105	345	559	1157	196	617	2717	6270

Table 4. Evaluation of RuBERT-original on the additional test set over a weighted inventory of semantic labels.

	B				**I**					
Metric/Label	**LOC**	**MISC**	**ORG**	**PER**	**LOC**	**MISC**	**ORG**	**PER**	**O**	**AVG**
Precision	0.75	0.58	0.53	0.91	0.88	0.68	0.57	0.80	0.80	0.77
Recall	0.49	0.49	0.33	0.71	0.87	0.51	0.56	0.94	0.94	0.78
F1	0.59	0.53	0.41	0.80	0.88	0.58	**0.64**	0.86	**0.96**	0.77
Support	96	478	105	345	559	1157	196	617	2717	6270

Table 5. Evaluation of RuBERT-adapted on the additional test set over a weighted inventory of semantic labels.

	B				**I**					
Metric/Label	**LOC**	**MISC**	**ORG**	**PER**	**LOC**	**MISC**	**ORG**	**PER**	**O**	**AVG**
Precision	0.79	0.60	0.57	0.94	0.91	0.74	0.64	0.91	0.82	0.80
Recall	0.46	0.59	0.35	0.86	0.84	0.59	0.55	0.96	0.94	0.81
F1	0.59	**0.59**	0.44	**0.90**	0.87	**0.65**	0.59	**0.93**	0.88	**0.80**
Support	96	478	105	345	559	1157	196	617	2717	6270

BERT-original demonstrates a slight improvement over **RuBERT-original** model and the best F1-score over “B-LOC”, “B-ORG” and “I-LOC” labels. **RuBERT-original** achieves the best F1-score over “I-ORG” and “O” labels. **RuBERT-adapted** shows performance gain over “B-MISC”, ”I-MISC”, “B-PER” and “I-PER” semantic labels which results in +2 Precision score, +3 Recall score and +3 F1-score.

In our work, we aim at extracting mentions of KG entities from texts. In particular, we do not apply semantic labels of NEs, but this can be a useful feature for an entity linking model. Hence, we additionally evaluate the performance over the weighted inventory of 3 semantic labels: “B-MISC”, “I-MISC” and “O”. The results are shown in Table 6, where Prec. refers to Precision and S corresponds to Support.

Table 6. Evaluation of BERT-original, RuBERT-original and RuBERT-adapted models on the additional test set over a weighted inventory of unified semantic labels.

	BERT-original			**RuBERT-original**			**RuBERT-adapted**			
Label	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**S**
B-MISC	0.86	0.74	0.79	0.87	0.62	0.76	0.89	0.78	**0.83**	1024
I-MISC	0.93	0.76	0.84	0.88	0.79	0.83	0.91	0.82	**0.86**	2529
O	0.77	0.95	0.85	0.80	0.94	0.86	0.82	0.94	**0.88**	2717
AVG	0.85	0.84	0.83	0.84	0.84	0.83	0.87	0.87	**0.87**	6270

In the unified label setting, **BERT-original** performs on par with **RuBERT-original**. **RuBERT-adapted** achieves performance gains over all semantic labels which results in +3 Precision score, +3 Recall score and + 4 F1-score as compared to **BERT-original** and **RuBERT-original**.

5. Discussion

The proposed method to train a domain-specific NER model has received a satisfactory performance with no human effort. The availability of Wikipedia in multiple languages and versatility of the Wikipedia graph may potentially allow for transferability to new domains and across languages. However, a number of drawbacks need to be solved for better performance and generalization of the method:

- The constructed entity vocabulary is not guaranteed to be free of noise. The main reason for irrelevant entities is a large degree of cross-reference in Wikipedia articles. This may cause redundant label predictions in the inference step. The drawback can be alleviated by extracting mentions of entities using a curated list of Wikipedia sections, additional vocabulary filtering or validation by annotators.
- The constructed entity vocabulary suffers from incompleteness. This may be due to the following reasons. First, not all of the interlinks (i.e. NEs) and entity aliases are highlighted in Wikipedia articles. Second, an interlink always refers to one Wikipedia title (i.e. the same surface form of an NE) resulting in a low lexical variability of the entity vocabulary. This can be solved by querying a KG for a set of NE aliases.
- Despite a rich variety of domains covered in Wikipedia, domain-specific knowledge may not be sufficiently represented in the document collection as well as similar publicly available resources. This necessitates the extra data mining for the model to learn relevant contexts.
- Another problem is lemmatization quality. In some cases, incorrect lemmas are obtained due to word ambiguity and the rich inflectional morphology of Russian. Besides, the morphology-based algorithm only partially covers the inventory of semantic labels. Although the majority of such cases are solved during the unification procedure, the remaining part still leads to annotation inconsistency. Hence, the model may get "confused" during the training and inference steps. A solution for this is to postprocess the assembled dataset based on the token-label frequency or use a KG to map ontological types to the corresponding semantic labels.

We believe that the method can be applied in domain-oriented areas, e.g. processing legal documents, educational dialogue systems and QA systems over KGs. Besides, for languages such as English, the method may potentially be transferred with a better performance achieved due to a variety of the available resources that may be used to automatically obtain the supervision.

6. CONCLUSIONS

This paper introduces a method to learn a domain-specific NER model for an arbitrary set of named entities without domain-specific supervision available. The code and models used in the experiments can be found at https://github.com/vmkhlv/histqa-domain-ner. The method is based on the semi-supervised approach. Specifically, a document collection can be automatically annotated using natural language pre-processing tools and a domain-specific entity vocabulary which can be constructed from scratch. Besides, we assume that neural models can learn from each other. Pre-trained language models can be used for training a NER model that is aware of both external and domain knowledge. We empirically show that BERT-based models trained over our method receive satisfactory performance with no human effort. However, a number of drawbacks need to be solved to gain performance. Future work is to be dedicated to quality improvement and exploring the transferability of the method across multiple domains and languages. The latter can be obtained thanks to versatility of Wikipedia.

REFERENCES

[1] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering.

[2] Deheng Ye, Zhenchang Xing, Chee Yong Foo, Zi Qun Ang, Jing Li, and Nachiket Kapre. 2016. Software-specific named entity recognition in software engineering social content. In 2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER), volume 1, pages 90–101. IEEE.

[3] Suzushi Tomori, Takashi Ninomiya, and Shinsuke Mori. 2016. Domain specific named entity recognition referring to the real world by deep neural networks. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 236–242.

[4] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A unified mrc framework for named entity recognition. arXiv preprint arXiv:1910.11476.

[5] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

[6] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[8] Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. arXiv preprint arXiv:1809.03599.

[9] Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. arXiv preprint arXiv:1904.02817.

[10] Alexandre Rochette, Yadollah Yaghoobzadeh, and Timothy J Hazen. 2019. Unsupervised domain adaptation of contextual embeddings for low-resource duplicate question detection. arXiv preprint arXiv:1911.02645.

[11] Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. arXiv preprint arXiv:1905.07213.

[12] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. Artificial Intelligence, 194:151–175.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cis-tac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

[14] Daniel Menezes, Ruy Milidiu, and Pedro Savarese. 2019. Building a massive corpus for named entity recognition using free open data sources. In 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), pages 6–11. IEEE.

[15] Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In International Conference on Analysis of Images, Social Networks and Texts, pages 320–332. Springer.

[16] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

CHINESE MEDICAL QUESTION ANSWER MATCHING BASED ON INTERACTIVE SENTENCE REPRESENTATION LEARNING

Xiongtao Cui and Jungang Han

College of Computer and Engineering, Xi'an University of Posts and Telecommunications, Xi'an, China

ABSTRACT

Chinese medical question-answer matching is more challenging than the open-domain question-answer matching in English. Even though the deep learning method has performed well in improving the performance of question-answer matching, these methods only focus on the semantic information inside sentences, while ignoring the semantic association between questions and answers, thus resulting in performance deficits. In this paper, we design a series of interactive sentence representation learning models to tackle this problem. To better adapt to Chinese medical question-answer matching and take the advantages of different neural network structures, we propose the Crossed BERT network to extract the deep semantic information inside the sentence and the semantic association between question and answer, and then combine with the multi-scale CNNs network or BiGRU network to take the advantage of different structure of neural networks to learn more semantic features into the sentence representation. The experiments on the cMedQA V2.0 and cMedQA V1.0 dataset show that our model significantly outperforms all the existing state-of-the-art models of Chinese medical question answer matching.

KEYWORDS

Question answer matching, Chinese medical field, interactive sentence representation, deep learning

1. INTRODUCTION

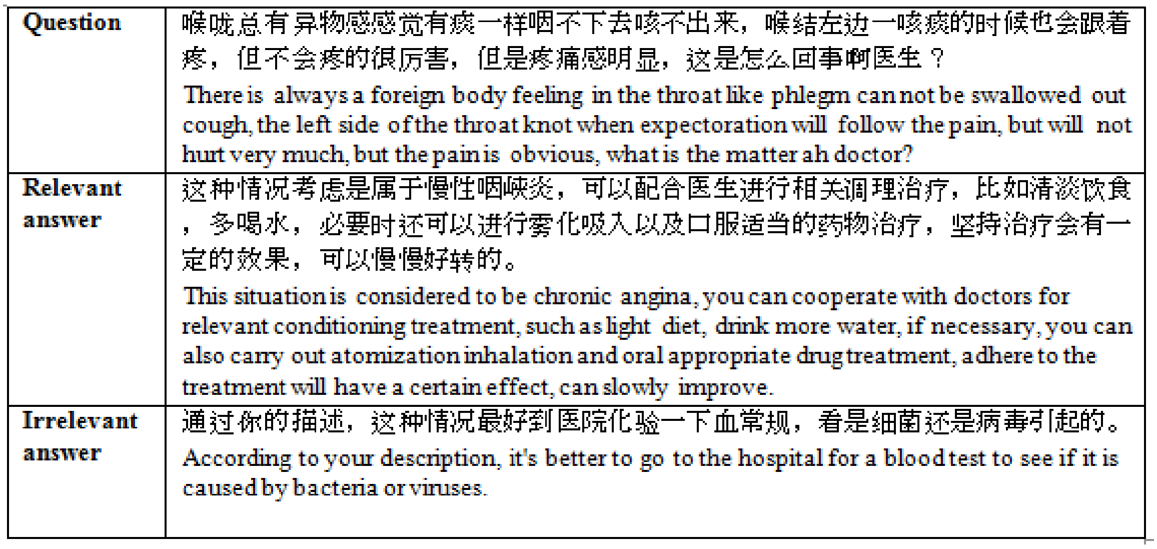
In recent years, more and more patients seek answers through the online medical health community, which brings time convenience to patients and accumulates a large number of medical questions and answers data. The deep learning method can learn knowledge from the huge questions-answers dataset before automatically answer the question raised by patients[1], which not only shortens the waiting time of patients in queue, but also reduces the workload of doctors.

This paper focuses on the study of Chinese medical question-answer matching, which is a crucial step to automatically answer patient questions. For example, as shown in Table 1, for a patient's question, the question-answer matching is to select the most matched relevant answer from the candidate answer set. The Chinese question-answer matching in the field of professional medicine is more challenging than in the open-domain [2]. Due to the differences between medical and open-domain in thesaurus and word interpretation, existing word segmentation tools inevitably produce errors in the Chinese language processing of medical texts, which reduce the accuracy of question-answer matching. Zhang et al. [3] proposed a character-level embedding

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 93-109, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101408

method to effectively solve the problem of word segmentation on medical text, and proposed a multi-scale interactive network framework in the later study to mine the semantic information and semantic association of medical questions and answers [4]. However, their model has limited performance in capturing semantic information and semantic association, which makes it difficult to proceed to practical application.

Table 1 An example of Chinese medical question-answer matching



Question	喉咙总有异物感感觉有痰一样咽不下去咳不出来，喉结左边一咳痰的时候也会跟着疼，但不会疼的很厉害，但是疼痛感明显，这是怎么回事啊医生？ There is always a foreign body feeling in the throat like phlegm can not be swallowed out cough, the left side of the throat knot when expectoration will follow the pain, but will not hurt very much, but the pain is obvious, what is the matter ah doctor?
Relevant answer	这种情况考虑是属于慢性咽峡炎，可以配合医生进行相关调理治疗，比如清淡饮食，多喝水，必要时还可以进行雾化吸入以及口服适当的药物治疗，坚持治疗会有一定的效果，可以慢慢好转的。 This situation is considered to be chronic angina, you can cooperate with doctors for relevant conditioning treatment, such as light diet, drink more water, if necessary, you can also carry out atomization inhalation and oral appropriate drug treatment, adhere to the treatment will have a certain effect, can slowly improve.
Irrelevant answer	通过你的描述，这种情况最好到医院化验一下血常规，看是细菌还是病毒引起的。 According to your description, it's better to go to the hospital for a blood test to see if it is caused by bacteria or viruses.

In response to the above problems, we design a series of interactive sentence representation learning models. In these models, we propose an crossed BERT [5] network, which is a modification of the Siamese [6] structure using the BERT network. This makes the question and answer pay attention to each other's semantic information in the process of model learning, and sentence representation obtains more information features. Then, we add multi-scale CNNs [3] or bidirectional GRU [7] network into the model, to take the advantages of different neural network structures. Due to the strict professional requirements for answering medical questions in Chinese, we chose to conduct experiments on the cMedQA V2.0 and cMedQA V1.0 dataset. The experimental results show that our models significantly outperform the existing state-of-the-art models. The top-1 accuracy on development dataset and test dataset of cMedQA V2.0 dataset is improved by 9.2% and 10.1% respectively, and the top-1 accuracy on development dataset and test dataset of cMedQA V1.0 was improved by 10.2% and 9.8% respectively.

The other parts of this paper are organized as follows: Section 2 introduces the research work related to this paper; Section 3 proposes a series of models that we designed; Section 4 describes the data set cMedQA V2.0 and cMedQA V1.0 and analyses the experimental results; Section 5 summarizes our study.

2. Related Work

We will briefly introduce the recent research works on the application of deep learning technology to question-answer matching in general fields and the professional medical field.

2.1. General Field

Early question answering methods such as logical rules [8] [9], information retrieval [10-12], and matching-based [13-15] only mined the shallow text information without extracting the deep semantic information of the text.

In recent years, more and more researchers have begun to focus on deep learning methods to mine deep text features of the text. Hu et al. [16] proposed a convolutional neural network model, which captures rich multi-level features within sentences to match two sentences. Qiu et al. [17] used a convolutional neural tensor network to model the interaction of two sentences through a tensor layer. Yin et al. [18] proposed an attention convolution neural network to model sentence pairs. Wang et al. [19] and Tan et al. [20] used an LSTM network to capture the order information of sequences while encoding sentences. Chen et al. [21] described a model based on position attention recurrent neural network to incorporate the word position of context into the attention representation. Wang et al. [22] added external attention information to hidden representation based on the recurrent neural network to obtain sentence representation containing attention. Tran et al. [23] presented a multi-hop attention mechanism, which uses multiple attention steps to learn the representation of the candidate answers.

The aforementioned work takes the advantages of neural network in extracting deep-level semantic features in the general field for question answering. However, answering questions in the specific medical field needs special study to void performance decline.

2.2. Medical Field

Compared with the general field, there is only a small amount of research work in Chines medical question answering. Perhaps the special Chinese language structure and the medical expertise complicated the problem.

Zhang et al [3] constructed the data set cMedQA V1.0 for Chinese medical question answering and proposed a multi-scale convolutional neural network model based on character embedding to extract text semantic information. They performed experiments on cMedQA V1.0, showed that character embedding and multis-cale convolutions were more advantageous than statistical rule methods.

Ye et al. [24] proposed a multi-layer composite convolutional neural network model, which stacks multiple convolutional neural networks together and extracts the characteristics of questions and answers from each layer, thus enriching the information of the final representation vector. They performed experiments on cMedQA V1.0 datasets and achieved the state-of-the-art performance at the time.

Zhang et al. [25] proposed a hybrid model of CNN and GRU neural networks for Chinese medical question-answer selection. Their model combines the advantages of different structures of two neural networks, thus achieving the most advanced performance on cMedQA V1.0 datasets.

Zhang et al. [4] proposed a method of incorporating attention mechanisms into multi-scale convolutional networks to focus on the interaction of questions and answers. And they constructed the cMedQA v2.0 data set, which is the optimization and update of cMedQA V1.0. The experimental results on the two data sets showed the advanced performance of their methods and that cMedQA v2.0 can better adapt to the complex neural network model.

Tian et al. [26] constructed a Chinese medical Q&A corpus called ChiMed and proposed a baseline model based on CNN and LSTM to validate this data set. He et al.[27] constructed a large-scale Chinese medicine question-answer dataset called webMedQA and proposed the convolution semantic clustering representation method to solve the question-answer matching problem.

The aforementioned methods have increasingly improved the performance in Chinese medical question-answer matching, but they are still limited in mining complex deep semantic information and the semantic association of question-answer pairs. Therefore, we aim to design more complex neural network models to overcome the limitation.

3. MODELS

The similarity between the sentence representations of questions and answers can measure their matching relationship, but it is only limited in the semantic information inside the sentence. Interactive sentence representation can focus on the connection between the questions and answers sentences to improve the accuracy of matching. Therefore, we designed four interactive sentence representation learning models with different architecture to extract more sentence features and capture connections of questions and answers. First, we construct a Siamese structured network model for question-answer matching using the BERT network, in which the question and the answer are represented as vectors of the same length for cosine similarity calculation. Then, we modified the neural network part of the previous model to a Crossed BERT network, which not only contains deep information features in the sentence representation, but also learns the semantic relationship between the question and the answer. Finally, we add a multi-scale CNNs network or bidirectional GRU network to the neural network part of the previous model, and their advantages in network structure can further extract more useful information features. In the following subsections, we will describe more technical details of these models.

3.1. Siamese BERT Model

The Siamese [6] BERT network model for question-answer matching is shown in Figure 1. The Siamese structure of this model includes two branches which share weights and parameters. The questions and answers will be represented as vectors of the same length for similarity calculation.

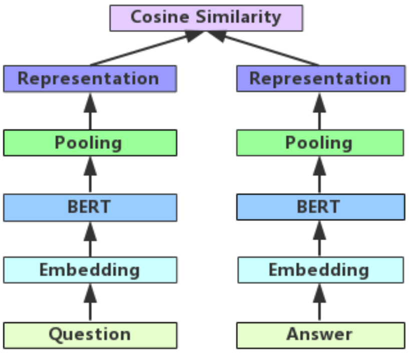


Figure 1. Siamese BERT network model.

The model uses the BERT network to learn sentence representation, as shown in Figure 2. We use position embedding, segment embedding, and token embedding of sentence sequences as inputs to the BERT network, where token is equivalent to Chinese characters. The character

[CLS] is inserted into the sequence of sentences as the first token of a sentence and the character [SEP] as the last token of a sentence. The BERT network consists of multiple bidirectional transformer [29] encoder layers. In each layer, there is a multi-head self-attention sublayer, which pays attention to the connection between two words at any position, as shown in Figure 3. The output of the BERT network is the context encoding of the input sequence, which is denoted by following equation:

$$H = BERT(E_0, E_1, ..., E_n) \tag{1}$$

where $H = [T_0, T_1, \dots, T_n]$, and T_i are context representation of each token. They are input to the mean pooling layer to extract useful information, shown as:

$$P = Pool_{mean}(H) \tag{2}$$

where P denotes the pooled output.

We use the pooling of BERT network output as a sentence representation. If the sentence representations of questions and answers are expressed as q and a respectively, then the cosine

similarity for calculating the association between q and a is shown as:

$$Sim(q,a) = \cos(q,a) = \frac{\|q \bullet a\|}{\|q\| \bullet \|a\|} \tag{3}$$

where ||·|| stands for vector length.

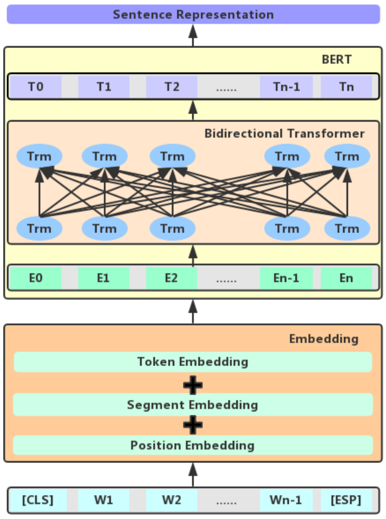


Figure 2. BERT for Sentence Representation.

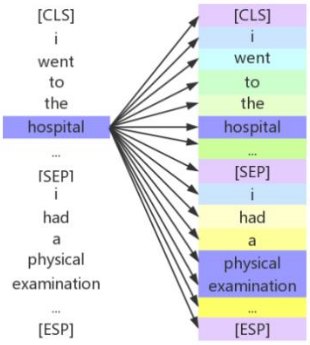


Figure 3. Self-attention sublayer.

3.2. Crossed BERT Siamese Model

The above model can extract deep-seated sentence representations of questions and answers separately, but the semantic correlation between them is ignored. Sergey et al. [28] proposed a 2-Channel network structure in the application of comparing the similarity of image patches, which inspired us to design the Crossed BERT Siamese network model, as shown in Figure 4.

In this model, where bidirectional transformer of two BERT networks interact with each other, as shown in Figure 5. Compared with the previous model, the neural network part was modified to an crossed BERT network. Therefore, the token output of question will pay attention to the Chinese characters in the answers and enrich the sentence representation of the questions. The token output calculation formula for the BERT of the questions is as follows:

$$T_{q_i} = Trm^{q_i}\left(E_{q_1}, E_{q_2}, \dots, E_{q_n}, E_{a_1}, E_{a_2}, \dots, E_{a_n}\right) \tag{4}$$

where T_{q_i} is i-th output of the BERT network in the question, and *Trm* is the bidirectional transformer. Likewise, the token output of the answer also pay attention to the semantic features in the question.

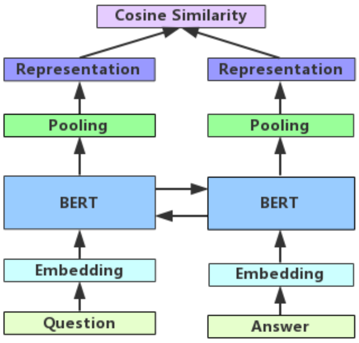


Figure 4. The Crossed BERT Siamese network model.

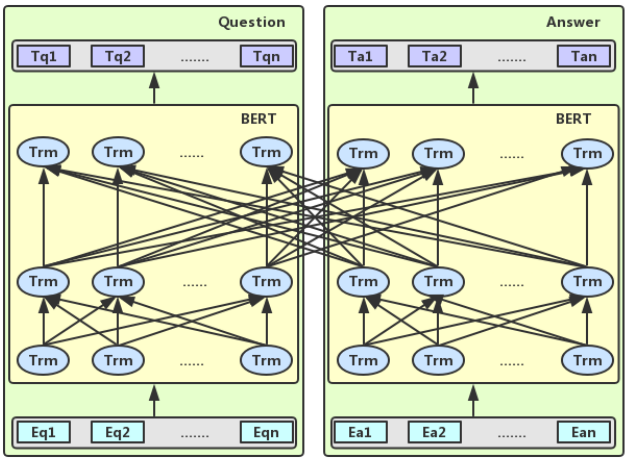


Figure 5. Crossed BERT network.

3.3. Crossed BERT Siamese Multi-Scale CNNs Model

Zhang et al [25] developed a hybrid model using CNN and GRU, combining the advantages of different neural network structures. Their method inspired us to design a hybrid model architecture of Crossed BERT Siamese network and Multi-Scale Convolutional Neural Networks (Multi-Scale CNNs), as shown in Figure 6.

In this model, the multi-scale CNNs network uses a series of convolution kernels of different sizes in convolution operations, each of which extracts n-gram features in sentences, as shown in Figure 7. Given a sequence C=[t_0,t_1,…,t_{l-k_i+1}] and convolution kernel size set K = $\{k_1, k_2, \dots, k_s\}$, where the convolution output of the i-th convolution kernel k_i is shown as：

$$O_j^{k_i} = f(W_j^{k_i} \circ [t_0, t_1, \dots, t_{l-k_i+1}] + b^{k_i}) \qquad (5)$$

where $O_j^{k_i} \in R^{l-k_i+1}$, $f(\cdot)$ is the activation function, $W_j^{k_i}$ are the matrix of weight parameters, vector b^{k_i} is bias parameters, and $W^\circ C$ is matrix multiplication. The number of convolution

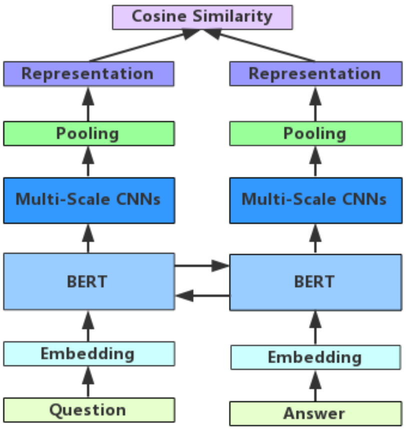


Figure 6. The Crossed BERT Siamese multi-scale CNNs.

kernel is expressed as N, and the output of multi-scale CNNs network layer is $O^{k_i} = [O_0^{k_i}, O_1^{k_i}, \dots, O_N^{k_i}]$. After that, we choose the max pooling to extract the useful feature information after convolution, the maximum value is more sensitive to the combined matrix features. Shown as:

$$p^{k_i} = [\max(O_0^{k_i}), \max(O_1^{k_i}), \dots, \max(O_N^{k_i})] \tag{6}$$

Next, the convolution outputs of different scales are concatenated and expressed as $P = [p^{k_1}, p^{k_2}, \dots, p^{k_s}]$.

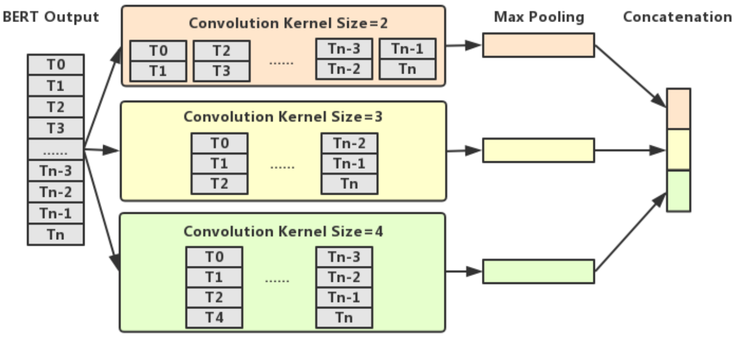


Figure 7. The multi-scale CNNs.

We represent the output of BERT network as H=[T_0,T_1,…,T_n]. Then, H are input into the multi-scale CNNs network as shown in Figure 7. Finally, the output of multi-scale CNNs, P is used as sentence representation for cosine similarity calculation.

3.4. Crossed BERT Siamese BiGRU Model

CNN network is good at mining the static features of local position in sentences, but it is difficult to extract the order information of Chinese characters in sentences. The recurrent neural network (RNN) can capture sequence information in sentences. However, RNN may have problems with gradient disappearance and gradient explosion during model training [30]. The GRU network not only solves the problems of RNN, but also simplifies long-term short-term memory (LSTM) network and improves the model computing performance [7]. Therefore, we add a bidirectional GRU (BiGRU) network layer to the Crossed Siamese BERT model, as shown in Figure 8.

The hidden state of GRU network is shown in Figure 9. The hidden layer updates it state h_t as shown below:

$$r_t = \sigma(W_r \bullet [h_{t-1}, x_t]) \tag{7}$$

$$z_t = \sigma(W_z \bullet [h_{t-1}, x_t]) \tag{8}$$

$$\tilde{h}_t = \tanh(W \bullet [r_t * h_{t-1}, x_t]) \tag{9}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{10}$$

where z_t, r_t,σ, and W are update gate, reset gate, sigmoid activation function and weight parameters respectively. The value of update gate ranges 0 to 1, which determines the memory of the previous hidden state in the current hidden state. The reset gate controls the amount of information entered into the current hidden state from the previous hidden state.

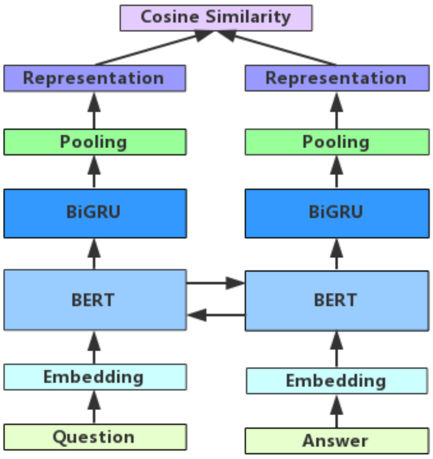


Figure 8. The Crossed BERT Siamese BiGRU.

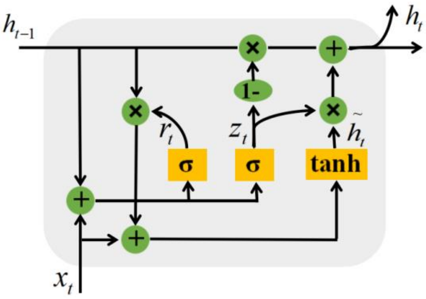


Figure 9. The hidden state of GRU network.

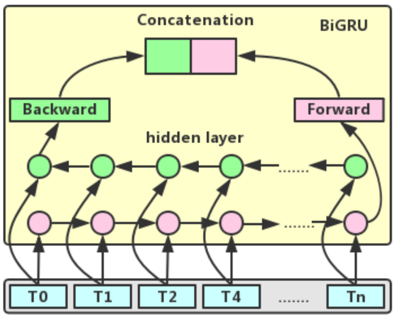


Figure 10. Network Structure of BiGRU network.

In this model, first inputs the question and answer to the BERT network to extract the features H=[T_0,T_1,…,T_n]. Then, H are input into the BiGRU network layer of the two branches respectively. The network structure of BiGRU is shown in Figure 9. The forward output of the GRU network is $\vec{h}$ and the backward output is $\overleftarrow{h}$, then the BiGRU output is h = $\vec{h} \parallel \overleftarrow{h}$. The output of BiGRU network layer can be shown as：

$$G = BiGRU\,(H) \tag{11}$$

Where $H \in R^{n \times 2h_d}$, and h_d is hidden layer dimension. Then, we average pooling the output of the BiGRU network layer, shown as:

$$P = Pool_{mean}(G) \tag{12}$$

Finally, we use P as sentence representation for cosine similarity calculation.

3.5. Objective Function

In this paper's model, we mark each tuple training data as (q_i, a_i^+, a_i^-), where the relevant answer to question q_i is a_i^+ and the irrelevant answer is a_i^-. The goal of model training is to maximize the cosine similarity between q_i and a_i^+, and also minimize the cosine similarity between q_i and a_i^-. We use margin loss function [31] as the training objective function of the model, which is defined as

$$L = \max\{\,0, M - Sim(q_i, a_i^+) + Sim(q_i, a_i^-)\} \tag{13}$$

where the margin value M is a constant and represents the distance between a_i^+ and a_i^-. *Sim(·)* is the similarity of cosine. If the value of loss function is 0, then $M < |Sim(q_i, a_i^+) - Sim(q_i, a_i^-)|$. After that, we use fixing weight decay regularization in Adam (AdamW) [32] algorithm to update the training parameters of the model, which improves the generalization ability of the adaptive gradient algorithm. The algorithm will automatically reduce the learning rate along with the increasing of training time.

4. Experiments

4.1. Dataset

We use the Chinese medical questions and answers dataset cMedQA V2.0 and cMedQA V1.0 to verify the effectiveness of our model in medical question answer matching task. The cMedQA V2.0 dataset was constructed by Zhang et al. [4] and was derived from an online Chinese medical health community(*http://www.xywy.com/*) which is provided by real users. The average number of characters for questions and answers in the cMedQA V2.0 dataset is 49 characters and 101 characters respectively, and the specific statistical results are shown in Table 2. The cMedQA V1.0 dataset is the initial version of cMedQA V2.0, which was collected by Zhang et al [3] and the detailed statistics are shown in Table 3.

4.2. Metrics

We used top-k accuracy (ACC@K) to evaluate the performance of our model, which is defined as

$$ACC@K = \frac{1}{N}\sum_{i=1}^{N} 1[a_i \in c_i^k] \tag{14}$$

where c_i^k is the set of top-k answers with the highest similarity to the question q_i,which belongs to the candidate answer set. The expression $1[\bullet] \rightarrow \{0,1\}$ denote a mapping that if the value in square brackets is true then the mapped value is 1, otherwise 0.

Each question in the cMedQA V2.0 dataset or cMedQA V1.0 dataset has 100 candidate answers, and we used top-1(ACC@1) to evaluate the performance of our model. The random selection has an accuracy of only 1%, so this is a very stringent measurement.

Table 2. The statistics of cMedQA V2.0 dataset.

	Question	**Answer**	**Average Characters Per Question**	**Average Characters Per Answer**
Train	100,000	188,490	48	101
Development	4,000	7,527	49	101
Test	4,000	7,552	49	100
Total	108,000	203,569	49	101

Table 3. The statistics of cMedQA V1.0 dataset.

	Question	**Answer**	**Average Characters Per Question**	**Average Characters Per Answer**
Train	50,000	94,134	120	212
Development	2,000	3,774	117	216
Test	2,000	3,835	119	211
Total	54,000	101,743	119	212

4.3. Baselines

To evaluate the performance of our model, we use the baseline models of the related studies as follows：

- **Single-CNN**: The model of Siamese structure, there is only one size convolution kernel to handle the question answer matching task.
- **Multi-Scale CNNs:** The model in which different scales of convolution are used in Siamese network to capture deep semantic information of questions and answers [3].
- **Multi-Level Composite CNNs:** The model proposed by Ye et al. [24] to extract intermediate features from each layer of convolution of Multi-layer CNNs, not just the superposition of multi-layer networks.
- **BiGRU**: The model in which the Siamese structure of the BiGRU network was used to capture the deep semantics and dependencies of question-and-answer pairs.
- **BiGRU Multi-Scale CNNs**: The model proposed by Zhang et al. [25] with multiple network hybrid structures and achieved state-of-the-art performance on cMedQA V1.0 datasets. The model takes the output of BiGRU as the input of multi-scale CNNs. It can capture not only local location invariant features but also sequence and dependent information.
- **BiGRU Shortcuts Multi-Scale CNNs Interactive**: The multi-scale interaction model proposed by Zhang et al [4] designed with shortcuts connection. The output of BiGRU and the previous embedded vectors are sent to the multi-scale CNNs, and then the attention interaction matrix is generated as the weight of the pooled output vector.

4.4. Experimental parameters

Our model is built using Pytorch framework. We conducted the experiments using DGX-1 deep learning server from Nvidia Corporation. In order to reduce the training time, we use 80% of the training dataset for model training.

All of our models used the pre-trained BERT model, which is the Chinese version of Google's BERT-Base model [5]. We pre-trained BERT-Base model using the Chinese Medical Corpus. The BERT-Base model has 12 transformer layers, 768 hidden states and 12 heads with self-attention, totaling 110M parameters. We process the input of the model, which treats the length of the sequence of questions and answers into a fixed length of 150 Chinese characters. If the length is less than 150 Chinese characters, padding the remaining positions with zero. If there are more than 150 Chinese characters in the sentence, it will be truncated. The feature maps for each convolution scale of Multi-Scale CNNs are 500. The output of BiGRU in each direction is 200 in dimensions. The margin value M of the loss function is 0.1. The initial learning rate is 2e-5.

4.5. Results

The experimental results of our model on the cMedQA V2.0 and cMedQA V1.0 dataset are shown in row -12 of Table 4. The Crossed BERT Siamese multi-scale CNNs use the convolution kernel with size 2 and 3. Dev (%) is the top-1 accuracy of development set. Test (%) is top-1 accuracy (ACC@1) of the validation set.

Models in rows 1 to 8 of table 4 list the performance of baseline models. Among the first three single network models, Multi-Scale CNNs achieves the highest computational score for model evaluation, which can capture semantic feature information with different size of granularity. The baseline models below the three models are multi-layer network structures. Compared with the previous three single network models, the evaluation scores were slightly improved. The combination of BiGRU and CNN shows that the multi-network layer model is feasible. Multi-Level Composite CNNs extract feature information from each convolution layer, enriching the final vector representation feature. The BiGRU network shortcuts Multi-Scale CNNs interactive model not only mix the advantages and disadvantages of the previous model, but also focuses on the interaction of questions and answers in the training process, so it improves the performance of the model.

The models in row 9 to 12 are four Crossed sentence representation network models proposed in this paper. Compared with Siamese BERT model, the performance of the Crossed BERT Siamese model is significantly improved. This indicates that the relationship between sentences should be paid attention in the question answer matching task. Comparing rows 10 and 11, the top-1 accuracy of the Crossed BERT BiGRU model is higher. This demonstrates that BiGRU has an advantage in compensating for the deficiencies of BERT network.

Table 4. Top-1 accuracy (ACC@1) results of model.

Model	**cMedQA V2.0**		**cMedQA V1.0**	
	Dev(%)	**Test(%)**	**Dev(%)**	**Test(%)**
CNN	67.6	67.8	64.0	64.5
BiGRU	68.9	68.7	64.9	66.7
Multi-Scale CNNs[3]	70.0	70.9	65.4	64.8
BiGRU-CNN	69.5	70.0	-	-
CNN-BiGRU	67.9	67.7	-	-
BiGRU Multi-Scale CNNs[25]	-	-	68.4	68.4

Multi-Level Composite CNNs[24]	70.4	70.1	65.6	66.2
BiGRU Shortcuts Multi-Scale CNNs Interactive[4]	72.1	72.1	66.1	67.1
Siamese BERT	78.3	78.6	75.1	75.2
Crossed BERT Siamese	80.5	80.4	77.3	77.9
Crossed BERT Siamese Multi-Scale CNNs	80.2	80.7	77.5	77.6
Crossed BERT Siamese BiGRU	**81.3**	**82.2**	**78.6**	**78.2**

The above experimental results show that our series of models have better performance than all the baseline models, especially the Crossed BERT Siamese BiGRU model. An important reason for the improved performance of the Chinese medical question-answer matching model is our proposed Crossed BERT network. This also illustrates that question-answer matching requires not only abundant sentence features, but also relevant information between question-answer pairs.

4.6. Discussion

4.6.1. Pre-training

In this paper, the pre-training model we used is a Chinese version of the BERT-Base model provided by Google [5]. The BERT-Base model uses Chinese corpus training in the general field. In medical field, text structure is more complex and word combination is diversified. Therefore, we collected a large number of medical texts to construct a medical corpus and used it to pre-train the BERT-Base model.

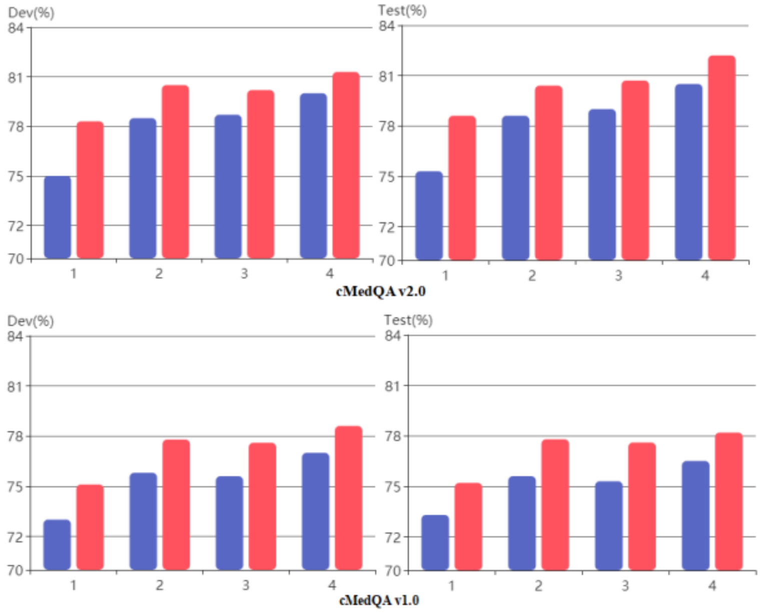


Figure 11. Results of pre-training experiment.

On the cMedQA V2.0 and cMedQA V1.0 dataset, we conducted a comparative experiment on the without pre-training BERT-Base model and Chinese Medical pre-training BERT model, as shown in Figure 11. The blue histogram represents the without pre-training BERT-Base model. The red histogram represents Chinese Medical pre-training BERT model. Dev (%) is the top-1 accuracy of development set. Test (%) is top-1 accuracy (ACC@1) of the validation set. The abscissa 1 represents the Siamese BERT model, the abscissa 2 represents the Crossed BERT

Siamese model, the abscissa 3 represents the Crossed BERT Multi-Scale CNNs model, and the abscissa 4 represents the Crossed BERT Siamese BiGRU model.

The experimental results depicted in figure 11 show that Chinese Medical pre-training BERT model performs better than the without pre-training BERT based model. Therefore, to use the pre-training model for sentence representation learning for Chinese medical texts, it is necessary to take into account the differences between the medical field and the general field and adapting to current language tasks.

4.6.2. Sentence representation of BERT network

In order to adapt the BERT network to sentence representation learning, we have carried out experiments and discussions on the output of the BERT network in the Siamese BERT model. We choose three methods of BERT network output commonly used for language tasks:

1) First Token: Take the first token of the last layer of BERT network as output directly;
2) Mean Token: The average pooling of all tokens at the last layer of the BERT network is used as output;
3) Mean Useful Token: The padding position of the last layer of the BERT network is covered with zero, and then the useful tokens average pooling is used as the output.

The experimental results on cMedQA V2.0 dataset and cMedQA V1.0 dataset are shown in Table 6. The first column represents the output category of the BERT network in the Siamese Bert model. The Following column, Dev(%) is the top-1 accuracy of development set. Test(%) is top-1 accuracy(ACC@1) of the validation set.

Table 6. Results of BERT network output.

BERT Output	**cMedQA V2.0**		**cMedQA V1.0**	
	Dev(%)	**Test(%)**	**Dev(%)**	**Test(%)**
First Token	73.45	73.95	70.39	70.95
Mean Token	77.30	77.78	74.05	74.14
Mean Useful Token	**78.28**	**78.58**	**75.15**	**75.25**

The above experimental results show that mean useful token of BERT performs better on cMedQA V2.0 dataset and cMedQA V1.0 dataset. Therefore, the application of BERT network in sentence representation learning needs more abundant and useful features.

4.6.3. Error Analysis

Our proposed model is significantly superior to previous state-of-the-art models and achieves better performance on cMedQA V1.0 and cMedQA V2.0 datasets. However, in the practical application of automatic answering to medical questions, the guaranteed accuracy is still not achieved. Three reasons may lead to this problem. The first reason is the uneven distribution of experimental datasets and the more complex sentence semantics in the medical field. The second reason is the error generated while collecting data sets manually. The third reason is that a question may correspond to multiple answers, and our model only matches a single answer, which affects the final performance of our model.

5. CONCLUSIONS

In this paper, we propose a series of sentence representation learning models for Chinese medical question-answer matching. They can not only capture deeper semantic information from question sentences and answer sentences, but also integrate the association information into the final representation vector. The experimental results on the cMedQA V2.0 and cMedQA V1.0 dataset show that our model achieves better performance than that of all the baseline models.

Zhang et al [33] proposed a new graph neural network graph BERT, which improved the performance and computational efficiency of the traditional graph neural network. Li et al [34] designed a graph matching network which is used to calculate the similarity between two graph structure data. Inspired by their research, as the future work, we will try to transform unstructured text data into graph structure similar to a knowledge map. The data of graph structure is easier for a machine to understand, so as to improve the accuracy of our model in answering questions. And we will collect more medical question-and-answer data sets to improve the performance of the model.

REFERENCES

[1] Feng, Minwei & Xiang, Bing & Glass, Michael & Wang, Lidan & Zhou, Bowen. (2015). Applying Deep Learning to Answer Selection: A Study and An Open Task. 10.1109/ASRU.2015.7404872.

[2] Liu, Xin & Chen, Qingcai & Deng, Chong & Zeng, Huajun & Chen, Jing & Li, Dongfang & Tang, Buzhou.(2018). LCQMC:A Large-scale Chinese Question Matching Corpus. Proceedings of the 27th International Conference on Computational Linguistics, pp. 1952–1962.

[3] Zhang, Sheng & Zhang, X. & Wang, H. & Cheng, J. & Li, P. & Ding, Z.. (2017). Chinese Medical Question Answer Matching Using End-to-End Character-Level Multi-Scale CNNs. Applied Sciences (Switzerland). 7. 10.3390/app7080767.

[4] Zhang, Sheng & Zhang, Xin & Wang, Hui & Guo, Lixiang & Liu, Shanshan. (2018). Multi-Scale Attentive Interaction Networks for Chinese Medical Question Answer Selection. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2883637.

[5] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [Online].Available:https://arxiv.org/abs/1810.04805

[6] Bromley, Jane & Guyon, Isabelle & Lecun, Yann & Scklnger, Eduard & Shah, Roopak. (1993). Signature Verification using a Siamese Time Delay Neural Network. International Conference on Neural Information Processing Systems Morgan Kaufmann Publishers Inc.

[7] Cho, Kyunghyun & van Merriënboer, Bart & Gulcehre, Caglar & Bougares, Fethi & Schwenk, Holger & Bengio, Y.. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 10.3115/v1/D14-1179.

[8] Ranjan, Prakash , and R. C. Balabantaray . "Question answering system for factoid based question." 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I) IEEE, 2016.

[9] Jain, Sonal & Dodiya, Tripti. (2014). Rule Based Architecture for Medical Question Answering System. 10.1007/978-81-322-1602-5_128.

[10] Choi, Sungbin , et al. "Semantic concept-enriched dependence model for medical information retrieval." Journal of Biomedical Informatics 47.2(2014):18-27.

[11] Ben Abacha, Asma & Zweigenbaum, Pierre. (2012). Medical question answering: Translating medical questions into SPARQL queries. ACM SIGHIT International Health Informatics Symposium (IHI 2012). 10.1145/2110363.2110372.

[12] Ben Abacha, Asma & Zweigenbaum, Pierre. (2015). MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. Information Processing & Management. 51. 10.1016/j.ipm.2015.04.006.

[13] Yikang Shen, Wenge Rong, Zhiwei Sun. (2015). Question/Answer Matching for CQA System via Combining Lexical and Sequential Information.

[14] Sarrouti M, Ouatik El Alaoui S. A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. J Biomed Inform. 2017;68:96-103. doi:10.1016/j.jbi.2017.03.001

[15] ShekharJangid, Chandra, K Vishwakarma, Santosh, and I Lakhtaria, Kamaljit. "Ad-hoc Retrieval on FIRE Data Set with TF-IDF and Probabilistic Models." International Journal of Computer Applications 93.9(2014):22-25.

[16] Hu, Baotian & Lu, Zhengdong & Li, Hang & Chen, Qingcai. (2015). Convolutional Neural Network Architectures for Matching Natural Language Sentences. Advances in Neural Information Processing Systems. 3. (NIPS)

[17] Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In: Proceedings of international joint conferences on artificial intelligence(IJCAI), pp 1305–1311.

[18] Yin, Wenpeng & Schütze, Hinrich & Xiang, Bing & Zhou, Bowen. (2016). ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. Transactions of the Association for Computational Linguistics. 4. 259-272. 10.1162/tacl_a_00097.

[19] Wang, Di & Nyberg, Eric. (2015). A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering. 707-712. 10.3115/v1/P15-2116.

[20] Tan, Ming & Dos Santos, Cicero & Xiang, Bing & Zhou, Bowen. (2015). LSTM-based Deep Learning Models for Non-factoid Answer Selection. CoRR Vol. abs/1511.04108 (2015).

[21] Chen, Qin & Hu, Qinmin & Huang, Xiangji & He, Liang & An, Weijie. (2017). Enhancing Recurrent Neural Networks with Positional Attention for Question Answering. 993-996. 10.1145/3077136.3080699.

[22] Wang, Bingning & Liu, Kang & Zhao, Jun. (2016). Inner Attention based Recurrent Neural Networks for Answer Selection. 1288-1297. 10.18653/v1/P16-1122.

[23] Nam Khanh Tran & Claudia Niedereée. (2018). Multihop Attention Networks for Question Answer Matching. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18). Association for Computing Machinery, New York, NY, USA, 325–334. DOI:https://doi.org/10.1145/3209978.3210009

[24] Ye, Dong , et al. "Multi-level Composite Neural Networks for Medical Question Answer Matching." 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC) IEEE Computer Society, 2018.

[25] Zhang, Yuteng , et al. "Chinese medical question answer selection via hybrid models based on CNN and GRU." Multimedia Tools and Applications (2019).

[26] Tian, Yuanhe & Ma, Weicheng & Xia, Fei & Song, Yan. (2019). ChiMed: A Chinese Medical Corpus for Question Answering. 250-260. 10.18653/v1/W19-5027.

[27] He, Junqing & Fu, Mingming & Tu, Manshu. (2019). Applying deep matching networks to Chinese medical question answering: A study and a dataset. BMC Medical Informatics and Decision Making. 19. 10.1186/s12911-019-0761-8.

[28] Zagoruyko, Sergey & Komodakis, Nikos. (2015). Learning to Compare Image Patches via Convolutional Neural Networks. 10.1109/CVPR.2015.7299064

[29] Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need. [Online].Available: https://arxiv.org/abs/1706.03762

[30] Sundermeyer, Martin & Schlüter, Ralf & Ney, Hermann. (2012). LSTM Neural Networks for Language Modeling.

[31] LeCun Y, Chopra S, Hadsell R, Ranzato M, Huang F. A tutorial on energy-based learning. Predicting Structured Data. 2006.

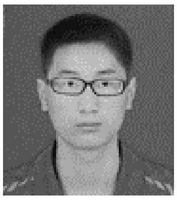
[32] Ilya Loshchilov and Frank Hutter, 'Decoupled Weight Decay Regularization', arXiv, cs.LG, 2017, 1711.05101.

[33] Jiawei Zhang and Haopeng Zhang and Congying Xia and Li Sun, (2020), Graph-BERT: Only Attention is Needed for Learning Graph Representations. [Online]. Available:https://arxiv.org/abs/2001.05140.

[34] Yujia Li and Chenjie Gu and Thomas Dullien and Oriol Vinyals and Pushmeet Kohli, (2020), Graph Matching Networks for Learning the Similarity of Graph Structured Objects. [Online]. Available: https://arxiv.org/abs/1904.12787.

AUTHORS

Xiongtao Cui



He received his bachelor's degree in network engineering from Xi'an University of Posts and telecommunications in 2018, and is currently pursuing a master's degree in big data processing and high performance computing. His research interests include machine learning, deep learning and natural language processing.

Jungang Han

He is a professor at Xi’an University of Posts and Telecommunications. He is the author of two books, and more than100 articles in the field of computer science. His current research interests include artificial intelligence, deep learning for medical image processing.

A PATTERN-MINING DRIVEN STUDY ON DIFFERENCES OF NEWSPAPERS IN EXPRESSING TEMPORAL INFORMATION

Yingxue Fu[1, 2] and Elaine Uí Dhonnchadha[2]

[1]School of Computer Science, University of St Andrews, Scotland, UK
[2] Center for Language and Communication Studies, Trinity College
Dublin, Dublin 2, Ireland

ABSTRACT

This paper studies the differences between different types of newspapers in expressing temporal information, which is a topic that has not received much attention. Techniques from the fields of temporal processing and pattern mining are employed to investigate this topic. First, a corpus annotated with temporal information is created by the author. Then, sequences of temporal information tags mixed with part-of-speech tags are extracted from the corpus. The TKS algorithm is used to mine skip-gram patterns from the sequences. With these patterns, the signatures of the four newspapers are obtained. In order to make the signatures uniquely characterize the newspapers, we revise the signatures by removing reference patterns. Through examining the number of patterns in the signatures and revised signatures, the proportion of patterns containing temporal information tags and the specific patterns containing temporal information tags, it is found that newspapers differ in ways of expressing temporal information.

KEYWORDS

Pattern Mining, TKS algorithm, Temporal Annotation, Tabloids and Broadsheets

1. INTRODUCTION

Newspapers are broadly categorized into two types: broadsheets and tabloids. As different newspapers target different audiences, the newspapers may use different words and sentence structures in their reports and thus have distinctive styles. Meanwhile, since newspapers aim at reporting events in a timely manner and things evolve with time, newspapers typically contain more temporal information than the other kinds of texts. Although the stylistic differences of newspapers have been studied thoroughly [2, 3, 4, 5], how newspapers differ in expressing temporal information remains an under-explored topic. This may be attributed to the simplified view of temporal information in natural language texts. The research on automatic processing of temporal information, however, shows the complex nature of temporal information.

Therefore, it would be of some interest and significance to investigate whether different newspapers express temporal information in different ways. As temporal information can be annotated automatically with temporal processing tools developed by the research community, it is possible to incorporate temporal information into the syntactic analysis of news articles.

Tabloid articles tend to be shorter and less serious in content compared to broadsheets. Therefore, it may be legitimate to presume that tabloids contain less temporal information than broadsheets

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 111-129, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101409

and that they avoid using explicit temporal expressions. However, evidence supporting this assumption has not been found.

To investigate this question, we create a corpus of newspaper articles annotated with temporal information so that sequences containing temporal information tags can be extracted. Inspired by previous research [1], we use the method of mining part-of-speech (POS) skip-gram patterns from the sequences and deriving signatures for the different newspapers.

Skip-gram modeling [6] is a technique proposed to solve the problem of data sparsity in natural language processing (NLP). Skip-grams are sequences of tokens that are similar to fixed-length n-grams but allow a gap of a user-defined size between adjacent tokens, for example, the pattern “hit ball” can be extracted from the sentence “I hit the tennis ball” with a gap of two.

Although the parameter setting of the pattern mining algorithm and the steps of deriving signatures of the newspapers are similar to previous research [1], the purpose of our research is different. This method [1] has been used to test the effectiveness of the POS skip-gram patterns in the task of authorship attribution, in other words, to test if the extracted POS skip-gram patterns can be used as a stylistic feature to characterize an author’s work. In contrast, our focus is to compare the skip-gram patterns containing temporal information tags in the signatures of different newspapers so that a better understanding about how newspapers differ in expressing temporal information can be obtained. As research shows that the POS skip-gram patterns that form the signature of an author are effective in capturing the style of an author, we delve into the signatures of the newspapers and pay attention to the patterns that are formed by temporal information tags to investigate our research question.

Contrary to our preconception, our analysis shows that the Daily Mirror, generally described as a tabloid, contains a greater proportion of temporal information in its signature and the temporal information tends to be expressed with explicit temporal expressions, which differs from the Guardian, which is typically described as a broadsheet.

It is worth mentioning that in spite of the wide application of neural networks in natural language processing tasks, considering our research question, the neural approach is a less desirable option because of the difficulty in observing the process and interpreting the results.

The next section introduces related work in the fields of temporal processing, pattern mining and the application of pattern mining for authorship attribution. Section 3 discusses the methodology, including the details of corpus creation and annotation, the algorithm for pattern mining and the steps for obtaining the signatures of the newspapers. Section 4 presents the results and discussion. Section 5 concludes the paper and points out future work.

2. RELATED WORK

To create the corpus annotated with temporal information, it is necessary to understand the techniques for processing temporal information in NLP. A review of the research on automatic extraction of temporal expressions and temporal relations and automatic identification of events is presented.

2.1. Automatic Extraction of Temporal Information

Generally speaking, temporal information in natural language texts can be embodied in three ways: a) temporal expressions: at 10:30, on Christmas Day, recently and the like; b) tense and aspect of verbs, such as goes, went, had gone, is eating, has been eating; c) temporal relations, for instance, the explosion happened soon after he got out of the theater.

The annotation of temporal information has been standardized under the ISO-TimeML scheme [7]. EVENT, TIMEX3, SIGNAL and LINK are the four major tags in this annotation scheme. EVENT denotes things that happen or occur and may be related to temporal expressions or involve temporal relations, for example, "the car crash" in the sentence "he was killed in the car crash yesterday" is an event associated with the temporal expression "yesterday".

TIMEX3 marks up explicit temporal expressions which may have the attributes of "duration", "date", "time" and "set". For example, in the sentence "the rain lasted for two weeks", "two weeks" has the attribution of "duration"; in the sentence "George was born on December 12, 1979", the temporal expression "December 12, 1979" has the attribute of "date"; "three years ago" in the sentence "he left the village three years ago" has the attribute of "time"; and in the sentence "I visited her twice a week that year", "twice" has the attribute of "set".

SIGNAL is used to annotate function words which reveal the connection between temporal objects, such as "before", "during", and "when".

LINK is a general tag for temporal relations. The annotation of temporal relations depends on the extraction and annotation of temporal expressions and events. Under the TimeML annotation scheme, LINK can be divided into three types:

- TLINK represents temporal relations between events or between an event and a temporal expression, which are rooted theoretically in the 13 temporal relationships [8], e.g. "before", "equal", "meet", "overlap", "during", "start" and "finish";
- SLINK represents a subordinate relationship between two events or between an event and a signal, where a introducing relation is typically present, for example, the relation between "wanted" and "leave" in the sentence "Mary wanted John to leave his family", and the relation between "regret" and "wear" in "Mary regrets that she didn't wear high heels that day";
- ALINK represents the relationship between an aspectual event and its argument event, such as "stop talking", "keep reading" and "starts to rain" [9].

Generally speaking, temporal information extraction is implemented using rule-based methods, machine learning techniques, or a hybrid of the two [10]. HeidelTime [11] and SUTime [12] are two of the best performing tools for temporal expression extraction, and both use rule-based methods which make normalization of temporal expressions easier. HeidelTime is the best performing system in TempEval-2 (http://semeval2.fbk.eu/semeval2.php?location=tasksT5) for extracting temporal expressions. Each temporal expression is viewed as a three-tuple consisting of a temporal expression, the type of the temporal expression (i.e., one of four types: date, time, duration and set), and the normalized value of the temporal expression. The goal is to extract the temporal expression and assign the type and calculate the normalized value correctly. SUTime is a rule-based temporal tagger built on regular expression patterns. Three types of rules are applied: text regex rules which are applied first to map simple regular expressions over characters or tokens to temporal representations; compositional rules that map regular expressions over chunks that are formed by tokens and temporal objects to temporal representations, (these rules being applied iteratively after the text regex rules); and filtering rules which discard ambiguous expressions that are likely to be non-temporal expressions, e.g. rules designed for polysemous words, such as "fall" [12]. The dependence on patterns for extraction of temporal expressions suggests a close connection between patterns and temporal expressions.

Machine learning techniques have also been proposed as a method of inferring the temporal relation linking a main clause and a subordinate clause attached to it. This is an example of learning temporal relations with machine learning techniques. There are also models [14, 15] which use hybrid methods. As indicated by [10], both rule-based approaches and systems implemented using machine learning algorithms typically rely on grammatical and syntactical attributes, such as POS tags, tense and so on.

Due to the dependence of techniques for automatic processing of temporal information on syntactic patterns of texts, we mix POS tags with temporal information tags for analyzing the stylistic features of newspaper articles.

2.2. Research on Pattern Mining

The second major part of our theoretical background is pattern mining. Pattern mining is one of the fundamental tasks of data mining and it consists of finding interesting, useful or unexpected patterns in a database [16]. This field originates from research [17] that provides an algorithm for solving the problem of discovering patterns of items bought by customers at a store, which may be used for commercial purposes, such as product catalogue design, add-on sales, store layout and customer segmentation based on purchasing patterns.

Under the framework of data mining, the temporal information in a dataset is used in the data preparation step for ordering items for subsequent pattern mining. In this case, the temporal aspect is not assigned with much weight. In the case of mining meaningful patterns from the database of customer purchase records, each transaction is considered as an itemset and the transactions are sorted based on the time of the transactions before a pattern mining algorithm is applied [18].

Temporality in pattern mining becomes an important consideration under the framework of temporal pattern mining. It is noticed that the correlation degree between some terms or patterns can change with time. [19] illustrates this point with the pattern formed by “Hillary Clinton” and “Candidate” which are correlated more strongly in 2008 than the pattern formed by “Hillary Clinton” and “Secretary of State”, while “Hillary Clinton” and “Secretary of State” form a more prominent pattern in 2009 than the pattern formed by “Hillary Clinton” and “Candidate”. The length of time interval is considered as an important factor in the prediction task [20] and the pattern discovery task [21]. [22] point out a problem with vanilla sequential pattern mining that does not take time gaps into account: if most of the customers buy B after A, and C after B, the manager can use this pattern to promote B when a customer purchases A and promote C when a customer buys B. However, if the time intervals between the purchases are not known, improper product recommendation would occur. This happens when customers purchase B several days after A, and buy C a certain amount of time after buying B, rather than buying B immediately after A and buying C with the same time gap after purchasing A. To solve this problem, a time-gap sequential pattern mining algorithm is proposed in their paper.

The above study demonstrates how the temporal aspect is handled in data mining. As temporal expressions are generally not expressed by a single word, the phrase is a more suitable unit for study. Events and temporal relations can be treated as patterns in a text. When temporal patterns are added in pattern mining, some interesting patterns may be obtained. However, the combination of temporal information and syntactic patterns in texts has not yet explored.

2.3. Pattern Mining for Authorship Attribution

Just as pattern mining algorithms can be used to identify the customers' purchasing habits, they can also be used for characterizing the distinct writing styles of different authors. [23] apply pattern mining techniques to e-mail forensic analysis. As suggested by the authors, the two most widely used machine learning algorithms for authorship attribution, Decision Trees and Support Vector Machines (SVM), have limitations in forensic investigations. When a decision tree is built, a decision node is constructed based only on the local information of one attribute, and the combined effect of several features is not captured. A second drawback of using Decision Trees in this task is that the same set of attributes is used for all the suspects, which could be manipulated for supporting wrong arguments. The problem with the SVM is that it is a learning function that works like a black box whose result is difficult to interpret for forensic purposes. This property makes SVM a less desirable choice.

Therefore, the authors present a method based on a pattern mining algorithm to extract the unique write-print of each suspect. Write-print is a term which denotes the patterns that can uniquely capture the writing style of an individual. The Apriori algorithm [17] is adopted in the pattern mining step. Only a pattern that appears above a pre-set threshold frequency, which is defined as 'support', is considered to be frequent. The support value is calculated as the percentage of emails that contain the pattern in the training set of a suspect author. The second step is to filter out common frequent patterns so that a pattern in the write-print of one suspect does not appear in the write-print of another. A write-print perfectly matches a test email if the test email contains every pattern in the write-print. Since frequent patterns may vary in occurrence frequency, and the more frequent patterns are generally more important than the others, the support of a frequent pattern is taken into account in the function for calculating the similarity between the write-print and the test email. This paper demonstrates the advantage of using pattern mining algorithms for forensic use over using machine learning algorithms.

The use of frequent POS skip-grams for authorship attribution is explored in [1]. The basic idea of this paper is to combine POS skip-grams and a top-k sequential pattern mining algorithm for the authorship attribution task. POS skip-grams are constructed in a similar way as POS n-grams except for the fact that a gap is allowed between adjacent POS tags, thus introducing an additional parameter representing the size of the gap. With the top-k sequential pattern mining algorithm, only the most frequent POS skip-grams are considered. The data for the experiments, comprising 30 books written by 10 authors, is taken from the Gutenberg Project (https://www.gutenberg.org). Each author is represented by three texts. Since it is believed that an author not only writes in his own style but also shares common patterns with the other authors, in order to obtain the unique signature of the author, patterns in the union of the other authors , i.e. reference patterns, will be removed from the initially obtained signature of the author. The Pearson correlation coefficient is chosen to measure the correlation between an anonymous test text and the signature of each author. It is shown that using POS skip-grams provides better performance than using POS bigrams and trigrams. The influence of the parameters of the top-k pattern mining algorithm on the overall performance and on the classification accuracy for each author is also studied.

It can be seen from the above papers that applying pattern mining algorithms for authorship attribution typically includes the steps of extracting more representative patterns from a text, finding the unique patterns of each author, and using a function for calculating the correlation/similarity between a test file and the set of unique patterns of each author which is called a write-print or a signature. Compared with the models for authorship attribution implemented using machine learning algorithms, the approach based on pattern mining has some advantages, such as being capable of discovering unique patterns for each author, which can serve as more credible evidence in forensic scenarios. As our research question requires the analysis of the patterns containing temporal information tags, the research on applying pattern mining techniques for authorship attribution can be a source of inspiration.

3. METHODOLOGY

To study the stylistic differences between different newspapers, a corpus is created by the authors.

3.1. Corpus Creation

The first step is to create a corpus annotated with temporal information. 1200 news articles are collected from four online newspapers: BBC (https://www.bbc.com), the Guardian (https://www.theguardian.com/uk), the Independent (https://www.independent.co.uk), and the Daily Mirror (https://www.mirror.co.uk). To reduce the influence of text categories on the result, texts classified under the categories of sports, politics and science & technology are collected in the same number. The publishing time of the news articles ranges from January 2020 to May 2020.

3.2. Corpus Annotation

The TARSQI Toolkit (TTK) is a suite of temporal processing modules for automatic temporal and event annotation of natural language texts [24]. TTK allows multiple linguistic annotation tasks to be performed, including tokenization, lemmatization, chunking, POS tagging, sentence and phrase boundary detection, temporal expression annotation and temporal relation annotation. It integrates several modules for temporal processing, including: the PreProcessor for tokenization, POS tagging and chunking, which is actually implemented by the TreeTagger [25]; GUTime for extracting temporal expressions; Evita for extracting events; Slinket for modal parsing; S2T for temporal repercussions of modal relations; Blinker for opportunistic pattern-based parsing of temporal relations; Classifier which is a MaxEnt classifier trained on the TimeBank corpus for identifying temporal relations between previously recognized events and temporal expressions in a text [24]; and Link Merger for ensuring consistency of all the temporal relations.

As the latest release of TTK is mainly written in Python 2 which has been declared End of Life in 2020, the source code of TTK released on GitHub cannot run without manual correction and not all the modules can work normally. Based on the tagger's performance in the pilot experiment, only modules that work are selected. Hence, PreProcessor, GUTime, Evita, Slinket and S2T are used in the annotation process. The statistics of the corpus is presented below.

Table 1. Statistics of the corpus.

	BBC	The Guardian	The Independent	The Daily Mirror
<s>	10262	11013	8627	8438
<lex>	218498	254613	196453	175833
<vg>	18316	21696	16679	14686
<ng>	20127	23175	18082	15843
<EVENT>	14946	17612	13750	12073
<TIMEX3>	2487	2501	1957	2004
<SLINK>	753	764	537	560
<TLINK>	674	667	454	492
<ALINK>	0	0	0	0

It can be seen that ALINK is not recognized, which may be attributed to the exclusion of some modules of TTK. The meaning of most of the tags in Table 1 has been explained in section 2.1. Apart from the typical tags specified under ISO-TimeML annotation scheme, such as <EVENT>, <TIMEX3>, <SLINK>, <TLNK> and <ALINK>, there are other tags generated by the PreProcessor of TTK: <s> represents the marker of a sentence (without the closing tag </s>); <lex> denotes tokens; <vg> means verb phrase; and <ng> represents noun phrase.

From the statistics, it can be seen that the Guardian ranks the first in terms of the number of sentences, followed by BBC, the Independent and the Daily Mirror. Articles from the Guardian, which is classified as a broadsheet, are longer than the others, and articles from the Daily Mirror, which is generally described as a tabloid, are the shortest of all. The same trend can be found in the statistics of tokens, verb phrases, noun phrases and EVENT, which may be explained by the fact that these aspects are closely related to the low-level linguistic tasks, such as POS tagging. As to the numbers of temporal expressions represented by <TIMEX3> and temporal relations including <SLINK> and <TLINK>, the Daily Mirror exceeds the Independent.

3.3. The TKS Algorithm

The Top-K Sequential (TKS) pattern mining algorithm is used in the pattern mining step because it outperforms TSP which is the current state-of-art algorithm for the same task by more than an order of magnitude in terms of execution time and memory usage [26].

In the field of sequential pattern mining, the task of finding the most frequent sequential patterns is associated with the question of how to define the threshold value for being “frequent” so as to obtain an appropriate number of patterns. If too many patterns are discovered, the patterns might be less representative of the data and the computational costs are high both for the algorithm and further processing, while if too few patterns are found, some interesting or important patterns might be missed.

Therefore, to reduce the difficulty of the problem, the question of mining the most frequent sequential patterns is redefined as mining the top-k sequential patterns, where k is a user-defined number of sequential patterns to be discovered.

Before the details of the TKS algorithm are explained, some concepts may have to be clarified. A dataset can be formally defined as $S=\{s_1, s_2, s_3. . . .s_i\}$, where $s_1...s_i$ are sequences. A set of items I may be defined as $I=\{i_1, i_2, i_3,...i_m\}$ and an itemset t is a set of items that belong to I, such as$\{i_1\}$ or $\{i_2, i_3\}$. Each sequence may contain one or more itemsets, for instance, $s_1 =\{t_1, t_2 , t_3\}$. A k-item sequence means a sequence $s=\{t_1, t_2, t_3, . . . t_k\}$, where t_n is an itemset for $1\leq n\leq k$. The *support* denotes the number of sequences in S that contain a specific pattern. It can also be expressed as the ratio of sequences that contain the pattern with respect to the total number of sequences in the database.

Each sequence can be considered either as a sequence-extended sequence or an itemset-extended sequence. Sequence-extension, which is also referred to as s-extension, means generating a new pattern by appending a new itemset after the existing itemsets of a sequence. For example, for s_1=({a}, {b}, {c}) and s_2=({a}, {b}, {c}, {d, e}), s_2 is an s-extension of s_1. Itemset-extended sequence, which is also called i-extension, means generating a new pattern by adding a new item to the last itemset of a sequence. For instance, for s_1=({a}, {b}, {c}) and s_2=({a}, {b}, {c, d, e}), s_2 is an i-extension of s_1.

The TKS algorithm employs a vertical database representation and the basic candidate-generation procedure of SPAM [27]. The meaning of vertical database representation may be understood in this way: given a database with *m* items and *s* sequences, each sequence may be identified with a unique ID and each of the *m* items may be represented separately by its presence in the itemsets of the sequence. Table 2 gives an illustration of horizontal database representation from which the vertical database representation may be derived.

Table 2. An example of horizontal database representation.

SID	Sequence
1	({a, b}, {c})
2	({a, c}, {a, d})
3	({c, d})

In Table 2, SID denotes the id of each sequence in the database, and the unique items are {a, b, c, d}. In the sequence with the sequence ID 1, *a* appears in itemset 1, *b* appears in itemset 1, *c* appears in itemset 2, and *d* does not appear in any itemset. When the above table is turned into vertical representation, four tables will be generated so that each of the items in {a, b, c, d} is represented by its presence in the itemsets of the respective sequence:

Table 3. Vertical representation of *a*.

SID	Itemsets
1	1
2	1, 2
3	

As can be seen from Table 3, item *a* appears in the first itemset in sequence 1 and the first and second itemsets in sequence 2 but does not appear in sequence 3.

Table 4. Vertical representation of *b*.

SID	Itemsets
1	1
2	
3	

As indicated by Table 4, item *b* only appears in the first itemset in sequence 1.

Table 5. Vertical representation of *c*.

SID	Itemsets
1	2
2	1
3	1

As indicated by Table 5, item *c* appears in the second itemset in sequence 1, the first itemset in sequence 2 and the first itemset in sequence 3.

Table 6. Vertical representation of *d*.

SID	Itemsets
1	
2	2
3	1

As indicated by Table 6, item *d* appears in the second itemset in sequence 2 and the first itemset in sequence 3.

The database is scanned only once to obtain the vertical database representation and calculate the support of each item. Starting with the items, candidate patterns obtained through s-extension and i-extension are searched. If the support of a candidate pattern generated in this way surpasses a pre-set threshold, the candidate pattern will be used as basis for generating further candidate patterns through s-extension and i-extension. Infrequent patterns will not be extended to form frequent patterns, which is called the Apriori property [26].

TKS and SPAM are similar in terms of the vertical database representation and basic candidate generation procedure described above. However, TKS redefines the frequent pattern mining problem as discovering top-k sequential patterns and new strategies are introduced.

The initial threshold is set to 0. Then the basic candidate generation procedure is applied. When a pattern is found, it is added to the list of patterns which are ordered based on the supports of the patterns. When *k* sequential patterns are found, the threshold is raised to the support of the pattern with the lowest support in the list, so that patterns with supports lower than the threshold will not be considered. The process continues until no more patterns can be found. In this way, the problem of mining the most frequent sequential patterns is turned into the task of mining the top-k sequential patterns.

The second strategy is to extend the most promising sequential patterns first [26]. This strategy means that among the set of patterns that can be extended to form new patterns, the pattern with the highest support is extended first. In this way, the most promising patterns are found first and the threshold will be increased faster, thereby improving the efficiency of the algorithm.

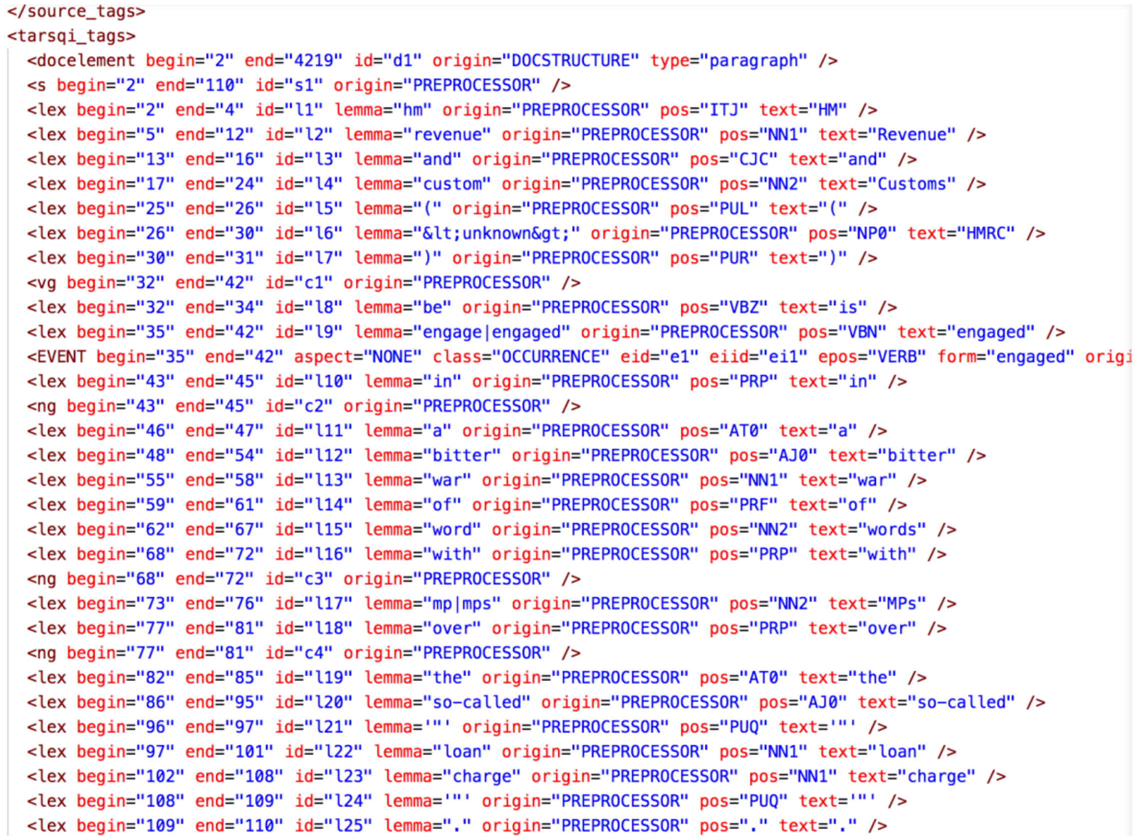
The third strategy is to discard infrequent items in the generation of candidate patterns [26]. With the increase of the threshold, the items whose supports are below the threshold are not considered and if a sequence contains a single infrequent item, the item will be recorded in a hash table and skipped when patterns are extended.

A special structure called a precedence map (PMAP) is introduced. Its basic form is <j, n, s> for s-extension and <j, n, i> for i-extension. For example, if an item has PMAP <e, 3, s>, it means that the item is followed by e in three sequences of the database by means of s-extension.

The application of the TKS algorithm for pattern mining is implemented using the Sequential Pattern Mining Framework (SPMF), which is an open-source data mining library offering implementations of more than 55 data mining algorithms [28]. Compared with other open source data mining libraries such as Weka (https://www.cs.waikato.ac.nz/ml/weka/), Mahout (http://mahout.apache.org/) and Knime (http://www.knime.org/), SPMF specializes in frequent pattern mining.

3.4. The Implementation

As the files tagged with TTK are saved as xml files, the xml.etree.ElementTree module (https://docs.python.org/3/library/xml.etree.elementtree.html) is used for parsing the files.



```
</source_tags>
<tarsqi_tags>
  <docelement begin="2" end="4219" id="d1" origin="DOCSTRUCTURE" type="paragraph" />
  <s begin="2" end="110" id="s1" origin="PREPROCESSOR" />
  <lex begin="2" end="4" id="l1" lemma="hm" origin="PREPROCESSOR" pos="ITJ" text="HM" />
  <lex begin="5" end="12" id="l2" lemma="revenue" origin="PREPROCESSOR" pos="NN1" text="Revenue" />
  <lex begin="13" end="16" id="l3" lemma="and" origin="PREPROCESSOR" pos="CJC" text="and" />
  <lex begin="17" end="24" id="l4" lemma="custom" origin="PREPROCESSOR" pos="NN2" text="Customs" />
  <lex begin="25" end="26" id="l5" lemma="(" origin="PREPROCESSOR" pos="PUL" text="(" />
  <lex begin="26" end="30" id="l6" lemma="&lt;unknown&gt;" origin="PREPROCESSOR" pos="NP0" text="HMRC" />
  <lex begin="30" end="31" id="l7" lemma=")" origin="PREPROCESSOR" pos="PUR" text=")" />
  <vg begin="32" end="42" id="c1" origin="PREPROCESSOR" />
  <lex begin="32" end="34" id="l8" lemma="be" origin="PREPROCESSOR" pos="VBZ" text="is" />
  <lex begin="35" end="42" id="l9" lemma="engage|engaged" origin="PREPROCESSOR" pos="VBN" text="engaged" />
  <EVENT begin="35" end="42" aspect="NONE" class="OCCURRENCE" eid="e1" eiid="ei1" epos="VERB" form="engaged" origi
  <lex begin="43" end="45" id="l10" lemma="in" origin="PREPROCESSOR" pos="PRP" text="in" />
  <ng begin="43" end="45" id="c2" origin="PREPROCESSOR" />
  <lex begin="46" end="47" id="l11" lemma="a" origin="PREPROCESSOR" pos="AT0" text="a" />
  <lex begin="48" end="54" id="l12" lemma="bitter" origin="PREPROCESSOR" pos="AJ0" text="bitter" />
  <lex begin="55" end="58" id="l13" lemma="war" origin="PREPROCESSOR" pos="NN1" text="war" />
  <lex begin="59" end="61" id="l14" lemma="of" origin="PREPROCESSOR" pos="PRF" text="of" />
  <lex begin="62" end="67" id="l15" lemma="word" origin="PREPROCESSOR" pos="NN2" text="words" />
  <lex begin="68" end="72" id="l16" lemma="with" origin="PREPROCESSOR" pos="PRP" text="with" />
  <ng begin="68" end="72" id="c3" origin="PREPROCESSOR" />
  <lex begin="73" end="76" id="l17" lemma="mp|mps" origin="PREPROCESSOR" pos="NN2" text="MPs" />
  <lex begin="77" end="81" id="l18" lemma="over" origin="PREPROCESSOR" pos="PRP" text="over" />
  <ng begin="77" end="81" id="c4" origin="PREPROCESSOR" />
  <lex begin="82" end="85" id="l19" lemma="the" origin="PREPROCESSOR" pos="AT0" text="the" />
  <lex begin="86" end="95" id="l20" lemma="so-called" origin="PREPROCESSOR" pos="AJ0" text="so-called" />
  <lex begin="96" end="97" id="l21" lemma='"' origin="PREPROCESSOR" pos="PUQ" text='"' />
  <lex begin="97" end="101" id="l22" lemma="loan" origin="PREPROCESSOR" pos="NN1" text="loan" />
  <lex begin="102" end="108" id="l23" lemma="charge" origin="PREPROCESSOR" pos="NN1" text="charge" />
  <lex begin="108" end="109" id="l24" lemma='"' origin="PREPROCESSOR" pos="PUQ" text='"' />
  <lex begin="109" end="110" id="l25" lemma="." origin="PREPROCESSOR" pos="." text="." />
```

Figure 1. Illustration of an annotated file

Fig. 1 shows a part of an annotated file. To keep the original structure of the annotation and the syntactic patterns generated by the PreProcessor of TTK, the principle for extracting the tags is to keep the tags of <s>, <ng>, <vg>, <EVENT> and <TIMEX3>, extract the POS tag of the corresponding <lex>, and keep the stop as "." in the output sequence. The stop is required by SPMF so that the sentences can be treated as different sequences rather than treating the whole text as one sequence, which might cause misleading results. As <EVENT> is generally associated with a noun or verb which is important for maintaining the grammaticality of the text, the method of extracting <EVENT> is to extract both the tag of <EVENT> and the POS tag of the token annotated with <EVENT>. The sequences comprised by the above tags extracted from each news article are saved into separate files with the required extension of ".text" so that SPMF can recognize the file as a text document.

The following sentence is taken from a news article from BBC: "Team Ineos have withdrawn from all races until 23 March following the death of sporting director Nico Portal and the "very uncertain situation" surrounding the coronavirus outbreak." (Original text: https://www.bbc.com/sport/cycling/51737966).

The corresponding sequence extracted from the annotated file is as follows: "s NN1 NN2 VBB VBN vg EVENT VBN PRP ng DT0 NN2 PRP ng TIMEX3 ng CRD NP0 VBG vg EVENT VBG AT0 NN1 PRF AJ0 NN1 NP0 NP0 CJC AT0 PUQ AJ0 AJ0 NN1 PUQ VBG vg EVENT VBG AT0 NN1 NN1". (See a complete list of POS tags at http://ucrel.lancs.ac.uk/bnc2/bnc2guide.htm)

The second step is to extract signatures for BBC, the Guardian, the Independent, and the Daily Mirror. As each news article is independent, their top-k POS skip-gram patterns are extracted separately. The TKS algorithm has the following parameters: the number *k* of POS skip-gram patterns to be found for each article, the minimum pattern length *minlen*, the maximum pattern length *maxlen*, and the maximum gap between the POS tags *g* (when *g* is set to zero, the skip-grams are the same as fixed-length n-grams). Based on the previous experiment results [1], the parameters are set as follows: *k*=250, *minlen*=1, *maxlen*=2 and *g*=1. The extracted patterns are saved separately for each news article.

To test the performance of the authorship attribution method, 75% of the files containing the extracted POS skip-gram patterns of each newspaper are used as training data and 25% are used as test data. The initial signature of a newspaper is formed by the patterns that appear in the extracted patterns of all or the majority of the news articles of the newspaper. For example, the initial signature of BBC is obtained by calculating the intersection of the top-250 POS skip-gram patterns of 225(=300*0.75) news articles.

The POS skip-gram patterns of a newspaper are obtained by concatenating the extracted top-250 POS skip-gram patterns of all the 225 news articles of the newspaper. Then pandas.dataframe API (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html) is used to read the files and find the duplicate patterns from the concatenated file, which are the patterns that the 225 files of patterns have in common. Because there are no repetitive patterns in a single file, if a pattern appears 225 times in the concatenated file, it must be a pattern that the 225 files of extracted patterns have in common.

Authors write in their own styles but they may share common structures with the other authors writing in the same language. To make sure that the signature obtained for each newspaper is unique, the common patterns between this newspaper and the others will be removed from the signature obtained in the above step. For this purpose, the notion of reference patterns is used. The reference patterns for a newspaper are formed by concatenating the POS skip-gram patterns of the other three newspapers. For example, the reference patterns of the Independent are formed by concatenating the top-250 POS skip-gram patterns of the other three newspapers. The final revised signature of a newspaper is then obtained by removing from its initial signature all patterns that also occur in its reference patterns.

4. Results & Discussion

The numbers of patterns extracted during the implementation steps are presented below.

Table 7. Numbers of patterns extracted in the steps of implementation.

	Number of patterns extracted from the training data	**Number of reference patterns**	**Number of patterns in the initial signature**	**Number of patterns in the revised signature**
BBC	55059	164422	14744	269
The Guardian	56744	162737	16666	203
The Independent	54611	164870	14290	203
The Daily Mirror	53067	166414	13221	62

Table 7 shows the numbers of patterns selected in each step of implementation. The numbers of patterns extracted from the training data of the four newspapers do not differ much. The number of reference patterns is, by the definition of the reference patterns, influenced by the total numbers of patterns discovered from the training data of the other three newspapers. Therefore, the number of reference patterns of the Daily Mirror is the largest, followed by the Independent, BBC, and the Guardian, that is to say, it follows a reverse trend compared with the number of patterns extracted from the training data.

As far as the number of patterns in the initial signature is concerned, the previous trend still holds: if the news articles from a newspaper are generally longer, the number of patterns that they have in common will be greater, and therefore, a greater number of patterns can be found in their initial signature.

When the reference patterns are removed from the initial signatures, the numbers of patterns that remain show a different trend. The number of patterns contained in the revised signature of BBC is the largest, followed by the Guardian, the Independent, and the Daily Mirror, which suggests that BBC has more unique patterns and the Daily Mirror has the smallest number of unique patterns. Through examination of the texts, it is found that articles from BBC contain more subheadings and links to related articles, while articles from the Daily Mirror contain fewer these types of texts, which may be one of the reasons for the greater number of unique patterns in the revised signature of BBC.

Moreover, it is noteworthy that even though the Guardian has the largest number of patterns in its initial signature and the smallest number of reference patterns to be removed from its initial signature, the number of unique patterns in its revised signature is the same as the Independent, which means that a considerable amount of patterns shared by the files in the training data of the Guardian do not have the discriminative power for telling this newspaper and the others apart.
As we are interested in finding out if different newspapers express temporal information in different ways, the patterns that involve temporal information will be examined for comparing the four newspapers in this aspect.

Since <s>, <ng>, <vg> and other tags are not directly related to temporal information and temporal relations are not completely recognized by the tagger, statistics are generated only for patterns containing tags of <EVENT> and <TIMEX3> in the following table.

Table 8. Ratios of patterns containing temporal information tags with respect to the total numbers of patterns in the revised signatures.

	Number of patterns containing temporal information tags	**Number of patterns in the revised signature**	**Ratio(%)**
BBC	33	269	12.3
The Guardian	25	203	12.3
The Independent	29	203	14.3
The Daily Mirror	22	62	35.5

From Table 8, it can be seen that the revised signature of the Daily Mirror has a greater ratio of patterns containing temporal information tags than the other three newspapers, which seems to be contrary to our preconception. However, as temporal expressions and events generally form the essential elements in the development of a topic, if they are given explicitly, the readers can grasp the core information more quickly, which makes it easier to condense the report into a shorter article. This explanation is similar to the explanation [29] for the more frequent use of numbers in tabloids. The proportions of patterns containing temporal information tags of the other three newspapers do not differ much.

To see if the four newspapers can be distinguished by some patterns in particular, the patterns containing temporal information tags are presented below. The patterns are understood in the following way: as an example, in a pattern “2 -1 3 -1 #SUP: 5”, the pattern is formed by 2 followed by 3, -1 is used to separate the items from each other, “#SUP:” denotes *support*, and the number after “#SUP:” is the value of the support of the pattern. This follows the format of the output of the SPMF interface.

Table 9. Patterns containing temporal information tags in the revised signature of BBC.

BBC						
1	EVENT	-1	#SUP:	78		
2	vbn	-1	EVENT	-1	#SUP:	18
3	EVENT	-1	vbd	-1	#SUP:	37
4	TIMEX3	-1	#SUP:	30		
5	EVENT	-1	#SUP:	74		
6	TIMEX3	-1	ng	-1	#SUP:	30
7	EVENT	-1	vbd	-1	#SUP:	36
8	EVENT	-1	#SUP:	65		
9	vbn	-1	EVENT	-1	#SUP:	27
10	TIMEX3	-1	ng	-1	#SUP:	29
11	TIMEX3	-1	#SUP:	29		
12	EVENT	-1	vbg	-1	#SUP:	36
13	TIMEX3	-1	ng	-1	#SUP:	30
14	EVENT	-1	vbn	-1	#SUP:	35
15	EVENT	-1	vbd	-1	#SUP:	37
16	EVENT	-1	#SUP:	73		
17	vg	-1	EVENT	-1	#SUP:	67
18	TIMEX3	-1	#SUP:	30		
19	EVENT	-1	vbn	-1	#SUP:	40
20	vg	-1	EVENT	-1	#SUP:	68
21	EVENT	-1	#SUP:	70		
22	TIMEX3	-1	#SUP:	29		
23	TIMEX3	-1	ng	-1	#SUP:	29
24	EVENT	-1	#SUP:	47		
25	EVENT	-1	#SUP:	66		
26	vg	-1	EVENT	-1	#SUP:	59
27	EVENT	-1	vbn	-1	#SUP:	28
28	EVENT	-1	vbn	-1	#SUP:	45
29	EVENT	-1	#SUP:	64		
30	vbn	-1	EVENT	-1	#SUP:	24
31	vg	-1	EVENT	-1	#SUP:	58
32	vg	-1	EVENT	-1	#SUP:	49
33	vg	-1	EVENT	-1	#SUP:	74

As can be seen from Table 9, eight of the 33 patterns in the revised signature of BBC are formed with the <TIMEX3> tag. Recall that the <TIMEX3> tag annotates explicit temporal expressions such as “today” and “on December 20, 1980”. Compared with the Guardian and the Independent below, this is a much greater ratio.

Table 10. Patterns containing temporal information tags in the revised signature of the Guardian.

The Guardian						
1	vg	-1	EVENT	-1	#SUP:	60
2	EVENT	-1	vbz	-1	#SUP:	25
3	EVENT	-1	#SUP:	62		
4	vg	-1	EVENT	-1	#SUP:	56
5	EVENT	-1	#SUP:	61		
6	EVENT	-1	vbd	-1	#SUP:	33
7	EVENT	-1	vbd	-1	#SUP:	25
8	EVENT	-1	vbd	-1	#SUP:	25
9	vbn	-1	EVENT	-1	#SUP:	21
10	EVENT	-1	vbn	-1	#SUP:	38
11	EVENT	-1	vbd	-1	#SUP:	25
12	EVENT	-1	vbd	-1	#SUP:	25
13	EVENT	-1	vbd	-1	#SUP:	31
14	EVENT	-1	vbd	-1	#SUP:	25
15	vbn	-1	EVENT	-1	#SUP:	28
16	EVENT	-1	vbn	-1	#SUP:	41
17	EVENT	-1	vbg	-1	#SUP:	37
18	EVENT	-1	#SUP:	69		
19	EVENT	-1	#SUP:	68		
20	EVENT	-1	vbz	-1	#SUP:	38
21	vg	-1	EVENT	-1	#SUP:	65
22	EVENT	-1	vbz	-1	#SUP:	38
23	vg	-1	EVENT	-1	#SUP:	80
24	EVENT	-1	vbg	-1	#SUP:	32
25	EVENT	-1	#SUP:	85		

In the revised signature of the Guardian, no patterns containing the tag <TIMEX3> are found, which suggests that the temporal expressions are not used in a distinctive way in the news articles of the Guardian. Among the 25 patterns, there are seven pattern comprised of the tag <EVENT> followed by the <vbd> tag.

Table 11. Patterns containing temporal information tags in the revised signature of the Independent.

The Independent						
1	EVENT	-1	vbd	-1	#SUP:	34
2	EVENT	-1	vbz	-1	#SUP:	31
3	EVENT	-1	#SUP:	60		
4	TIMEX3	-1	ng	-1	#SUP:	23
5	EVENT	-1	vbg	-1	#SUP:	44
6	vg	-1	EVENT	-1	#SUP:	148
7	EVENT	-1	vbd	-1	#SUP:	114
8	EVENT	-1	vbn	-1	#SUP:	61
9	EVENT	-1	#SUP:	168		
10	TIMEX3	-1	#SUP:	23		
11	vbn	-1	EVENT	-1	#SUP:	33
12	EVENT	-1	vbg	-1	#SUP:	30
13	EVENT	-1	#SUP:	63		
14	vg	-1	EVENT	-1	#SUP:	57
15	EVENT	-1	vbd	-1	#SUP:	32
16	vbn	-1	EVENT	-1	#SUP:	19
17	EVENT	-1	vbz	-1	#SUP:	31
18	EVENT	-1	vbn	-1	#SUP:	43
19	EVENT	-1	#SUP:	77		
20	EVENT	-1	vbn	-1	#SUP:	25
21	EVENT	-1	#SUP:	55		
22	vg	-1	EVENT	-1	#SUP:	50
23	EVENT	-1	vbn	-1	#SUP:	37
24	vbn	-1	EVENT	-1	#SUP:	25
25	EVENT	-1	vbd	-1	#SUP:	35
26	EVENT	-1	#SUP:	79		
27	vg	-1	EVENT	-1	#SUP:	72
28	EVENT	-1	vbd	-1	#SUP:	32
29	vg	-1	EVENT	-1	#SUP:	51

The revised signature of the Independent contains two patterns formed by <TIMEX3>, which indicates that slightly more temporal expressions are used in articles from the Independent than the Guardian. However, when compared with BBC and the Daily Mirror, the temporal expressions are used much less frequently in the Independent.

Table 12. Patterns containing temporal information tags in the revised signature of the Daily Mirror.

The Daily Mirror						
1	TIMEX3	-1	ng	-1	#SUP:	28
2	EVENT	-1	#SUP:	50		
3	EVENT	-1	vbn	-1	#SUP:	48
4	vbn	-1	EVENT	-1	#SUP:	34
5	TIMEX3	-1	#SUP:	28		
6	EVENT	-1	vbn	-1	#SUP:	32
7	TIMEX3	-1	ng	-1	#SUP:	31
8	TIMEX3	-1	#SUP:	31		
9	vg	-1	EVENT	-1	#SUP:	44
10	EVENT	-1	#SUP:	5		
11	vg	-1	EVENT	-1	#SUP:	4
12	TIMEX3	-1	#SUP:	35		
13	EVENT	-1	vbn	-1	#SUP:	33
14	EVENT	-1	#SUP:	50		
15	TIMEX3	-1	ng	-1	#SUP:	35
16	vbn	-1	EVENT	-1	#SUP:	23
17	EVENT	-1	vbg	-1	#SUP:	29
18	TIMEX3	-1	ng	-1	#SUP:	21
19	TIMEX3	-1	#SUP:	21		
20	vg	-1	EVENT	-1	#SUP:	46
21	EVENT	-1	vbz	-1	#SUP:	26
22	EVENT	-1	#SUP:	48		

Among the 22 patterns in the revised signature of the Daily Mirror, eight are patterns containing the tag of <TIMEX3>, which is an indicator that the Daily Mirror tends to use more explicit temporal expressions in its news articles than the other newspapers. In combination with Table 8, it may be concluded that the news articles of the Daily Mirror, generally described as a tabloid, contain a greater proportion of temporal information and the temporal information is expressed with explicit temporal expressions, which makes it easier to convey the essentials of a piece of news within limited space. In contrast, articles from the Guardian and the Independent are more likely to convey temporal information implicitly. BBC lies in the middle of this spectrum of explicitness and implicitness.

In the authorship attribution step, an experiment following the method used in [1] is performed. The accuracies for BBC, the Guardian, the Independent and the Daily Mirror are 90.7%, 6.7%, 0% and 0%, respectively. In the experiment [1], the success ratio for the classification task on Catharine Trail is equal to zero under the same setting as in the present experiment. It is explained in [1] that some authors are harder to identify because some of them attempt to hide their identities with deliberate variations in style, and therefore the patterns in their signatures are heterogeneous, or because some authors write about the daily routine life, which makes their writing share much in common with the other authors. Except for BBC, all the other newspapers are not classified with high accuracy, which may be attributed to the fact that BBC contains more subheadings and links to related articles and this feature can be more easily characterized by skip-gram patterns.

5. CONCLUSIONS & FUTURE WORK

In this study, the focus is to find if different newspapers express temporal information in different ways and how they differ in terms of the amount of temporal information and the specific patterns containing temporal information tags.

From the perspective of the number of patterns in the revised signatures, it can be seen that the number of patterns contained in the revised signature of BBC is the largest, followed by the Guardian, the Independent and the Daily Mirror, which suggests that BBC has more unique patterns and the Daily Mirror has the smallest number of unique patterns. This may be attributed in part to the fact that articles from BBC contain more subheadings and inserted titles of articles related to a topic than the other newspapers while articles from the Daily Mirror, constrained by space, normally do not contain these types of texts. The POS skip-gram patterns can capture this aspect.

As articles from the Guardian, generally described as a broadsheet, are longer, more patterns are mined under the same setting of the algorithm and therefore, the initial signature of the Guardian contains the greatest number of patterns. However, even though the reference patterns of the Guardian are the smallest in number, when the initial signature of the Guardian is revised by moving the reference patterns, the patterns that remain in its revised signature are not the largest in number, which means that a considerable amount of patterns shared by the files in the training data of the Guardian do not have the discriminative power for telling this newspaper and the others apart and the language use of the Guardian is possibly not so distinctive from the other newspapers in the dataset.

From the perspective of the ratio of patterns containing temporal information tags to the number of patterns in the revised signatures, it can be seen that the revised signature of the Daily Mirror which is generally described as a tabloid has a greater ratio of patterns containing temporal information tags than the other three newspapers, which may run counter to some preconceptions. However, since temporal expressions and events generally form the essential elements of a piece of news, if they are given explicitly, the readers can grasp the core information more quickly, which makes it easier to condense a report into a shorter article. This result is consistent with the research [29] which shows that tabloids use more numbers than broadsheets for similar reasons.
As far as the specific patterns containing temporal information tags are concerned, it can be concluded that articles from the Daily Mirror, typically described as a tabloid, contain a greater proportion of temporal information and the temporal information is expressed more frequently with explicit temporal expressions than the other newspapers.

Due to the constraint of the current version of TTK, only tags of <EVENT> and <TIMEX3> are analyzed and studied. If the temporal information can be annotated fully, it is likely that a more complete picture of the different ways in expressing temporal information can be obtained.
In future work, experiment can be carried out using other methods for the authorship attribution task. Meanwhile, different datasets can be used for testing the results of this study, and how adjusting the parameter setting of the pattern mining algorithm influences the result remains a question that needs further investigation.

REFERENCES

[1] Pokou, Y. J. M., Fournier-Viger, P., & Moghrabi, C. (2016) "Authorship attribution using small sets of frequent part-of-speech skip-grams", *The Twenty-Ninth International Flairs Conference*.

[2] Crystal, D. & Davy, D. (2016) *Investigating English Style*, Routledge.

[3] Fowler, R. (2013) *Language in the News: Discourse and Ideology in the Press*, Routledge.

[4] Bagnall, N. (1993) *Newspaper language*, Routledge.

[5] Timuçin, M. (2010) "Different language styles in newspapers: An investigative framework", *Journal of Language and Linguistic Studies*, Vol. 6, No. 2, pp104–126.

[6] Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. (2006) "A closer look at skip-gram modelling", *LREC*, Vol. 6, pp 1222–1225.

[7] Pustejovsky, J., Lee, K., Bunt, H., & Romary, L. (2010) "ISO-TimeML: An International Standard for Semantic Annotation", *LREC*, Vol. 10, pp394–397.

[8] Allen, J. F. (1983) “Maintaining knowledge about temporal intervals”, *Communications of the ACM,* Vol. 26, No.11, pp832–843.

[9] Pustejovsky, J., Ingria, R., Saurí, R., Castaño, J. M., Littman, J., Gaizauskas, R. J., Setzer, A., Katz, G., & Mani, I. (2005) “The Specification Language TimeML” [Online]. Available: http://www.timeml.org/timeMLdocs/timeMLspec.pdf.

[10] Wang, W., Kreimeyer, K., Woo, E. J., Ball, R., Foster, M., Pandey, A., Scott, J., & Botsis, T. (2016) “A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports”, *Journal of Biomedical Informatics,* Vol. 62, pp78– 89.

[11] Strötgen, J. & Gertz, M. (2010) “Heideltime: High quality rule-based extraction and normalization of temporal expressions”, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp321–324.

[12] Chang, A. X. & Manning, C. D. (2012) “Sutime: A library for recognizing and normalizing time expressions”, *LREC*, Vol. 2012, pp3735–3740.

[13] Lapata, M. & Lascarides, A. (2004) “Inferring sentence-internal temporal relations”, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL*, pp153–160.

[14] Yang, Y.-L., Lai, P.-T., & Tsai, R. T.-H. (2014) “A hybrid system for temporal relation extraction from discharge summaries”, *International Conference on Technologies and Applications of Artificial Intelligence*, pp379–386.

[15] Chang, Y.-C., Dai, H.-J., Wu, J. C.-Y., Chen, J.-M., Tsai, R. T.-H., & Hsu, W.-L. (2013) “Tempting system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries”, *Journal of Biomedical Informatics*, Vol. 46, ppS54–S62.

[16] Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., & Thomas, R. (2017) “A survey of sequential pattern mining”, *Data Science and Pattern Recognition*, Vol. 1, No. 1, pp54–77.

[17] Agrawal, R., Imieliński, T., & Swami, A. (1993) “Mining association rules between sets of items in large databases”, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp207–216.

[18] Maylawati, D., Aulawi, H., & Ramdhani, M. (2018) “The concept of sequential pattern mining for text”, *IOP Conference Series: Materials Science and Engineering,* Vol. 434, No.012042, doi:10.1088/1757-899X/434/1/012042.

[19] Hoonlor, A. (2011) *Sequential patterns and temporal patterns for text mining*. Rensselaer Polytechnic Institute.

[20] Hirate, Y. & Yamana, H. (2006) “Sequential pattern mining with time intervals”, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp775–779.

[21] Laxman, S. & Sastry, P. S. (2006) “A survey of temporal data mining”, *Sadhana,* Vol. 31, No. 2, pp173–198.

[22] Yen, S.-J. & Lee, Y.-S. (2013) “Mining non-redundant time-gap sequential patterns”, *Applied Intelligence*, Vol. 39, No.4, pp727–738.

[23] Iqbal, F., Binsalleeh, H., Fung, B. C., & Debbabi, M. (2010) “Mining writeprints from anonymous e-mails for forensic investigation”, *Digital Investigation,* Vol. 7, No. 1-2, pp56–64.

[24] Verhagen, M. & Pustejovsky, J. (2012) “The Tarsqi Toolkit”, *LREC*, pp2043–2048.

[25] Schmid, H. (1995) “Improvements in part-of-speech tagging with an application to German”, *Proceedings of the ACL SIGDAT-Workshop*, pp47–50.

[26] Fournier-Viger, P., Gomariz, A., Gueniche, T., Mwamikazi, E., & Thomas, R. (2013) “TKS: efficient mining of top-k sequential patterns”, *International Conference on Advanced Data Mining and Applications*, pp109–120.

[27] Ayres, J., Flannick, J., Gehrke, J., & Yiu, T. (2002) “Sequential pattern mining using a bitmap representation”, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp429–435.

[28] Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., & Tseng, V. S. (2014) “Spmf: a java open-source pattern mining library”, *The Journal of Machine Learning Research,* Vol. 15, No. 1, pp3389–3393.

[29] Li, Y., Zhang, D., & Wanyi, D. (2014) “A case analysis of lexical features in english broadsheets and tabloids”, *International Journal of English Linguistics,* Vol. 4, No. 4, p115-122.

AUTHORS

Yingxue Fu is a Ph.D. candidate at the University of St Andrews, UK. She studied under the program of M.Phil. in Speech and Language Processing at Trinity College Dublin in 2019-2020.

Dr Elaine Uí Dhonnchadha is Assistant Professor in Computational Linguistics and M.Phil Coordinator for the Centre for Language and Communication Studies. Prior to joining Trinity College, Elaine worked as a researcher in Institiúid Teangeolaíochta Éireann (The Linguistics Institute of Ireland), as a Lecturer in Dublin City University and as a Systems Analyst and Programmer in a number of software development companies.

MULTI-LAYER ATTENTION APPROACH FOR ASPECT BASED SENTIMENT ANALYSIS

Xinzhi Ai[1], Xiaoge Li[1], Feixiong Hu[2], Shuting Zhi[1] and Likun Hu[2]

[1]School of Computing, Xi’an University of Posts
and Telecommunications, Xi’an, China, 710121
[2]Tencent Technology (Shenzhen) Co., Ltd, China, 518057

ABSTRACT

Based on the aspect-level sentiment analysis is typical of fine-grained emotional classification that assigns sentiment polarity for each of the aspects in a review. For better handle the emotion classification task, this paper put forward a new model which apply Long Short-Term Memory network combine multiple attention with aspect context. Where multiple attention mechanism (i.e., location attention, content attention and class attention) refers to takes the factors of context location, content semantics and class balancing into consideration. Therefore, the proposed model can adaptively integrate location and semantic information between the aspect targets and their contexts into sentimental features, and overcome the model data variance introduced by the imbalanced training dataset. In addition, the aspect context is encoded on both sides of the aspect target, so as to enhance the ability of the model to capture semantic information. The Multi-Attention mechanism (MATT) and Aspect Context (AC) allow our model to perform better when facing reviews with more complicated structures. The result of this experiment indicate that the accuracy of the new model is up to 80.6% and 75.1% for two datasets in SemEval-2014 Task 4 respectively, While the accuracy of the data set on twitter 71.1%, and 81.6% for the Chinese automotive-domain dataset. Compared with some previous models for sentiment analysis, our model shows a higher accuracy.

KEYWORDS

Aspect-level sentiment analysis, Multiple attention mechanism, LSTM neural network

1. INTRODUCTION

Aspect sentiment analysis is one of the important parts of natural language processing (NLP), which provides fine-grained classification of emotions [1]. Such as, the following sentence “Large memory, but performance was poor”, where the polarity of the aspect “memory” shows positive, the polarity of another aspect “performance” reveals negative. It's complicated to judge the polarity of such a multi-aspect review if information about the objective of the aspect is ignored. This is a common challenge in general sentiment classification tasks.

The key of aspect sentiment analysis is to model appropriate context features for an aspect target in a comment. Traditional approaches mostly with manual features to train classifiers. In contrast, neural networks can learn feature representations from given data without manual feature engineering effectively. However, these models largely rely on information from additional analysis such as dependency parsing, which may accumulate errors during the sentiment analysis. It has been proposed that differentiated the both sides context of aspect target based on their locations relative to the aspect word and applied Bi-LSTM networks to encode the both side context respectively. However, the model proposed may not capture the sentiment information

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 131-142, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101410

that is far from the aspect target in the comment. Prior work has applied the advantage of the attention module to solve aspect sentiment analysis problems. The model automatically generates aspect-to-text and text-to-aspect bidirectional attention [7]. In general, prior models with attention module only consider the semantic feature for an aspect target or context, which fails to take into account other features affecting the accuracy of the aspect sentiment analysis, such as context location and class balancing.

The above deficiencies, a model named LSTM-MATT-AC [11] is proposed, which introduce Multiple attention combining mechanisms (MATT) and aspect context module(AC). Where multi-attention mechanism consists of three parts: content attention, location attention, and class attention. The content attention is responsible for considering the semantic information related to the aspect targets in the comments; the location attention mainly focuses on the relative location information between the aspect targets and their context; the class attention is introduced to overcome the classification model data variance caused by an imbalance in the training data. And adopted the aspect context module encodes the bidirectional information of the aspect targets into the sentiment features. Finally, the performance of the model is evaluated on four different language datasets, including Chinese and English. The evaluation results show that the model is widely applicable. In addition, the results verify that our model is insensitive to language.

The others of this paper is arranged as below, the second section introduces some related work; the third part gives an overview of LSTM-MATT-AC model and a detailed description of location, content and class attention; A large number of experiments are carried out in the fourth part, and the results prove the validity of the new model, and last section concludes the paper and presents direction for future improvement.

2. RELATED WORK

2.1. Aspect Sentiment Classification

Traditional methods used sentiment classifiers with expensive hand-crafted features. Most prior work consisted of two steps: (1) A sentiment lexicon was built. For example, use existing dependency relation triple parsers; the dependency parsing knowledge between sentiment words and aspect words; use WordNet's annotation information; Also constructing domain specific sentiment lexicon based on the space similarity of sentiment embedding that contain semantic information. (2) The appropriate classifier was selected. Which contains three ways: one is Naive Bayes, the another is Support Vector Machine (SVM), or combine these two classifiers. Recently, an increasing number of researchers have used neural networks for aspect sentiment analysis. Some typical models that are commonly used include the recursive neural network, and the tree-LSTMs neural network [5]. These models utilize the syntax structures of comments to encode the hierarchical grammatical information of the comments. This work suggests that word embedding and deep neural network could utilize syntactic and semantic structures in comments for aspect sentiment analysis without manual intervention.

2.2. Attention Mechanism

The essence of attention module is to retain useful information, filter out irrelevant information [7], and overcome the limitation of poor performance of recurrent neural network (RNN) in feature coding of long sequence texts. This method associates the target object with each feature in the text, and obtains the corresponding attention probability distribution. In 2014, successfully solved the problem of image classification by introducing the attention mechanism into the RNN [8]. Subsequently, neural networks with attention mechanisms became a heated research topic.

With the advancement of research, many subtasks of natural language processing have got good results using attention mechanisms, such as textual entailment recognition, text abstraction extraction, speech recognition, and machine reading comprehension. Prior work also validated the effectiveness of attention mechanisms in aspect sentiment analysis tasks.

3. MODEL

3.1. Task Definition

First, the sentiment analysis tasks, we need to give a fixed sentence of length n $s=\{w_1,\ldots,w_l,\ldots,w_r,\ldots,w_n\}$ and an aspect target $a=\{w_l,\cdots,w_r\}$. For each word w_i, we obtain an embedding $v_i \in R^{d_w}$ from $L \in R^{d_w \times |V|}$, where $|V|$ refers to size of the vocabulary, d_w is used to indicate dimension of embedding word.

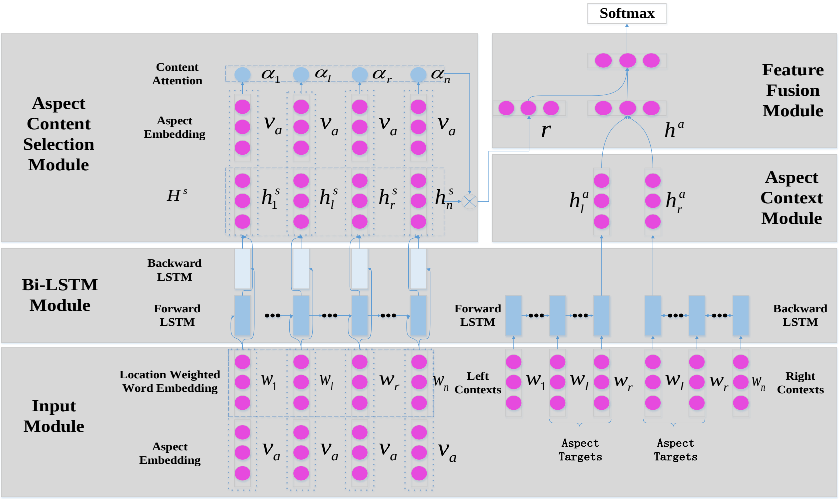


Figure 1. The architecture of LSTM-MATT-AC model

3.2. An Overview of Model

Input module: The first module is input embedding which includes aspect, location-weighted word and aspect context, in which aspect embedding need to map each aspect word into embedding $v_a \in R^{d_w}$, while location-weighted word embedding refers to integrating the location distribution, and generate corresponding weight embedding, and aspect context embedding splits sentence word embedding sequence of into two parts on both sides of the context.

Bi-LSTM module: The aspect embedding and the location-weighted word embedding are combined as input of the left Bi-LSTM module. We obtain the forward hidden states $\overrightarrow{H}^s \in R^{d_h \times n}$ and backward hidden states $\overleftarrow{H}^s \in R^{d_h \times n}$ from Bi-LSTM simultaneously. Concatenating $\overrightarrow{H}^s$ and $\overleftarrow{H}^s$ generate output hidden states $H^s \in R^{2d_h \times n}$, where d_h refers to the dimension of the hidden states.

$$H^s = \overrightarrow{H}^s \| \overleftarrow{H}^s \qquad (1)$$

Then embed context of the aspect into the Bi-LSTM neural network to the right of the Bi-LSTM module. Similar to the left part, the hidden state in two directions will be generated.

$$\overrightarrow{H}^a = \overrightarrow{LSTM}([v_1, v_2, \cdots, v_l, \cdots v_r]) \qquad (2)$$

$$\overleftarrow{H}^a = \overleftarrow{LSTM}([v_1, v_2, \cdots v_r, \cdots v_{n+1}, v_n]) \qquad (3)$$

Aspect content selection module: The module main task that hidden state is multiplied by the content attention distribution $\alpha \in R^{1\times n}$, to achieve the feature embedding which is highly relevant to aspect targets.

$$r = H^s \alpha^T \qquad (4)$$

Aspect context module: The feature embedding $h^a \in R^{2d_h}$ through concatenate the hidden states $\overrightarrow{h}^a \in R^{d_h}$ and $\overleftarrow{h}^a \in R^{d_h}$ is generated.

Feature fusion: By $r \in R^{2d_h}$ and $h^a \in R^{2d_h}$ are integrated with the expression with the following equation to come into being the final feature embedding $h^* \in R^{d_h}$.

$$h^* = \tanh(W_r r + W_a h^a) \qquad (5)$$

3.3. Location Attention

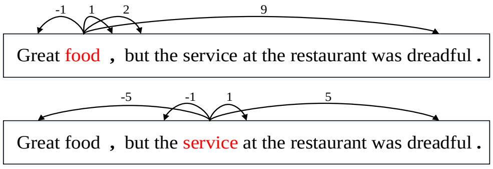

-1
1
2
9
Great food , but the service at the restaurant was dreadful .
-5
-1
1
5
Great food , but the service at the restaurant was dreadful .

Figure 2. The relative distances between words and aspect targets in sentences

For sentiment analysis tasks, the emotional relative with the distance between the target and the context provides crucial information. In the sentence shown in Figure 2, “great” is the sentiment word for the aspect target "food", and “great” is closer to the aspect target “food” than to “dreadful”. However, for another aspect "service", the relative distance to the relevant sentiment word "dreadful" and the irrelevant sentiment word "great" is the same. This phenomenon poses a threat to the simple location weight calculation. In this paper, citing the punctuation-based algorithm proposed by Han [11] to be calculated.

3.4. Content Attention

According to the aspect target, the hidden states H^s concatenation with the aspect embedding v_a to produce content attention distribution α.

$$M = \tanh\left(W_h H^s; W_v v_a \otimes e_n\right) \quad (6)$$

$$\alpha = soft\max\left(w^T M\right) \quad (7)$$

Where $M \in R^{(d_h+d_w)\times n}$, $\alpha \in R^{1\times n}$, $W_h \in R^{d_h \times d_h}$, $W_v \in R^{d_w \times d_w}$ and $w \in R^{d_w+d_h}$ are projection parameters, $v_a \otimes e_n = [v_a, v_a, \cdots, v_a]$ means that v_a is repeatedly concatenated n times, and e_n refers to column vector with n 1s.

3.5. Class Attention

To avoid the bias problem of the LSTM-MATT-AC model, this paper introduces the loss function [3] into the class attention, which penalizes the misclassification of underrepresentation more seriously. In addition, in the case of extremely unbalanced training data, smoothing factor φ is introduced to smooth the weights, which may lead to very high class attention weights.

$$\omega_i^c = \frac{\max(c)}{c_i + \varphi \times \max(c)} \quad (8)$$

Where c is the list containing the number of data points under each class, and c_i refers to the amount of data points under the i-th class. For the data from the Restaurant, Automobile and Laptop domains, we set the parameter $\varphi = 0$, for the Twitter domain data, set the parameter $\varphi = 0.1$

.

3.6. Model Training

The end-to-end training model is established by using the back propagation, and the cross entropy loss is selected as the loss function. To avoid over fitting, we add L2 regularization to the loss function [10].

$$loss = -\sum_i \sum_j y_i^j \log \hat{y}_i^j + \lambda \|\theta\|^2$$

(9)Where y_i^j is the correct sentiment, $\hat{y}_i^j$ indicates the prediction result of the model, i refers to index of review data, j means index of class of the final sentiment, λ as L2 regularization items, and θ present parameter set.

4. EXPERIMENTS

4.1. Dataset

New proposed model LSTM-MATT-AC on four data sets were evaluated, as shown in Table 2. Firstly, we evaluated our model on three more English datasets from SemEval-2014 Task 4 and Twitter. For different languages have considerably different expressions for sentiments. After that, a dataset from the Chinese Automobile domain were used for comparison.

Table 2. Details of the experimental datasets

Data Set	Train			Test			Total
	Positive	Neutral	Negative	Positive	Neutral	Negative	
Restaurant	2148	631	894	720	194	209	4796
Automobile	953	57	528	295	48	144	2025
Laptop	989	468	868	337	167	131	2960
Twitter	1561	3127	1560	173	346	1743	8510

4.2. Hyper parameters Setting

In our evaluation, we selected the training data of 20% randomly as the testing data to tune a set of hyper parameters. Word embedding is initialized by the 300-dimensional size Glove vectors [7]. We use U (-0.01, 0.01) to uniformly distribute random initialization of off-vocabulary words. The LSTM hidden state dimension is set as 150. All parameters of the random initialization, equal distribution of U (-0.05, 0.05). Our model for the batch of 25 samples of training.

4.3. Model Comparison

We have four domain-specific datasets will be in the field of the new proposed model and the following baseline model are compared.

a) TD-LSTM (Target-dependent LSTM) using two LSTM network context is modelled as a target. To connect two LSTM network finally hide status for the sentiment prediction.
b) TC-LSTM [12] (Target-connection LSTM) extends the TD-LSTM, which combines aspect embedding and the embedding of each word.
c) AT-LSTM (Attention-based LSTM) Based on the attention of the LSTM (attention - -based LSTM) by LSTM network context modelling to target first, and then combining embedded LSTM hide status and ways, to generate attention to weight.
d) ATAE-LSTM [14]is an aspect-oriented embedded attention-based LSTM developed on the basis of attention-LSTM.
e) AOA-LSTM [6] (attention-over-attention LSTM) First, aspects and text are modelled simultaneously through Bi-LSTM. Then, the fusion between the aspect target and the text representation produces bidirectional attention.

4.4. Result and Discussion

Next apply accuracy metric to confirm results and performance analysis of the model classifier, which defined as follows:

$$acc = \frac{num_{cor}}{num_{all}} \quad (10)$$

Where num_{cor} represents the amount of correctly predicted samples, num_{all}represents the total amount of samples and acc measures the percentage of the correctly predicted samples in all sample data [11].

4.4.1. 3-class Sentimental Classification

Table 3. Comparison results: accuracy of the 3-class sentimental classification
the Best performances in bold

Model	Restaurant	Laptop	Twitter	Automobile
TD-LSTM	0.756	0.681	0.646	0.786
TC-LSTM	0.763	0.710	0.680	0.792
AT-LSTM	0.762	0.689	0.672	0.788
ATAE-LSTM	0.772	0.687	0.681	0.795
AOA-LSTM	**0.812**	0.745	—	—
LSTM-MATT-AC	0.806	**0.751**	**0.711**	**0.816**

Compared with the unidirectional LSTM used by the ATAE-LSTM, the LSTM-MATT-AC utilizes the Bi-LSTM to better encode the context semantic features of the comments. The aspect content attention is appended to the output layer of the Bi-LSTM to generate the weight distribution of the semantic correlation between the aspect and context, and then the weighted total of the hidden states from the Bi-LSTM generates feature embedding that is closely related with aspect target. To address the model data variance caused by the class imbalance in the training data, we introduce the class attention mechanism into the loss function of the model for capture deeper sentiment features from the class with less data.

For aspect phrases, different words contribute to the aspect expression differently. Attention mechanism that phrase is an important part of the model is able to focus on, so as to promote the accuracy of sentiment classification [11]. In general, our model achieves better results on datasets of four different domains and two different languages than other models. The high performance of our model shows that it has a good domain migration ability and is language-insensitive.

4.4.2.2-class sentimental classification

For further compare the performance of new model with other baseline models in aspect sentiment analysis, we conducted another experiment with only the data of "Positive" and "Negative" in four datasets. The experimental results are shown in the following table.

Table 4.Comparison results: accuracy of 2-class sentimental classification. the Best performances in bold

Model	Restaurant	Laptop	Twitter	Automobile
ASGCN	0.806	0.755	0.721	—
AEN-BERT[15]	0.831	0.799	—	—
TD-LSTM	0.892	0.868	0.830	0.869
TC-LSTM	0.894	0.853	0.858	0.878
AT-LSTM	0.892	0.874	0.858	0.873
ATAE-LSTM	0.909	**0.876**	0.867	0.882
LSTM-MATT-AC	**0.910**	0.870	**0.868**	**0.897**

Table 4 reveals that the accuracy of the models were significantly improved after the "Neutral" data was removed. We investigated the "Neutral" reviews in the four datasets, and found that most of the "Neutral" comments are objective statements about aspect targets, and do not contain sentiment expressions from reviewers. For example, the review "It takes about 2 hours to be served our 2 courses." does not express reviewer’s sentiment tendency. Another finding is that some "Positive" and "Negative" comments are too implicit about the sentiment of aspect targets and the syntactic structures are relatively complex, resulting in the model not able to accurately capture the sentiment features in the training phase, thus misclassifying the reviews into "Neutral" class. In conclusion, the objectivity of "Neutral" review expression and the ambiguity of "Neutral" sentiment class itself lead to "Neutral" data greatly affecting the accuracy of model. The accuracies of the LSTM-MATT-AC on the Restaurant, Twitter and Automobile datasets are higher than that of other models, reaching 91%, 86.8%, and 89.7% respectively. Our experiments show that the LSTM-MATT-AC can better deal with aspect sentiment classification of different domains and languages.

We conducted another experiment to proof the validity of the multi-attention module, such as the below models to compare: (1) the LSTM-NL-AC model which has no location attention compared with model presented in this paper; (2) the LSTM-NC-AC model which has no class attention compared with the new proposed model. For this experiment, we just kept "Positive" and "Negative" part data from the four domain-specific datasets, as a binary sentiment classification experiment. The evaluation results are shown in Figure 3. The overview of the four datasets is shown in Figure 4.

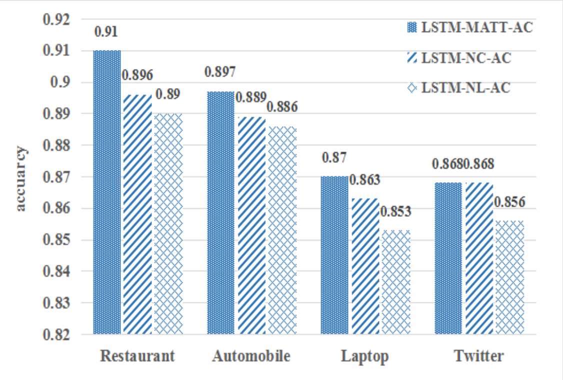


Figure 3.The impact of the multi-attention module on the classification accuracy

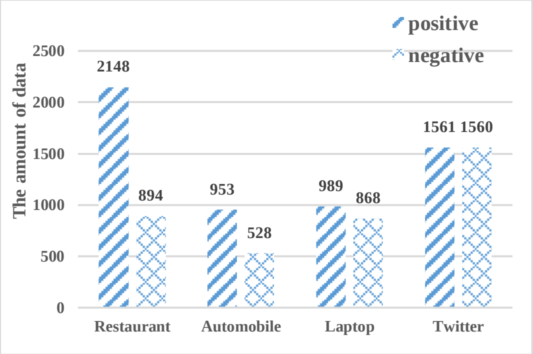


Figure 4. Data overview of the two-classification tasks

From figure 3, the LSTM-MATT-AC model reaches optimal performance. After the addition of location attention, the model at least 1.1% improvement in the datasets compared with the no location attention model; With the class attention, other accuracy rates have improved, except that twitter remained consistent. Figure 4 shows that class distributions of Restaurant, Laptop and Automobile datasets are imbalanced compared with Twitter dataset. Furthermore, we observed that, compared with the location attention, class attention to improve model performance contribution. To study the impact of aspect context module the accuracy of aspect sentiment classification, we constructed two models named LSTM-MATT-1 and LSTM-MATT-2, replacing the aspect context module with and respectively. Then the three LSTM-MATT models (i.e., LSTM-MATT-1, LSTM-MATT2, and LSTM-MATT-AC) were used to conduct another aspect sentiment analysis experiment on four domain-specific datasets. The evaluation results are shown in Figure 5.

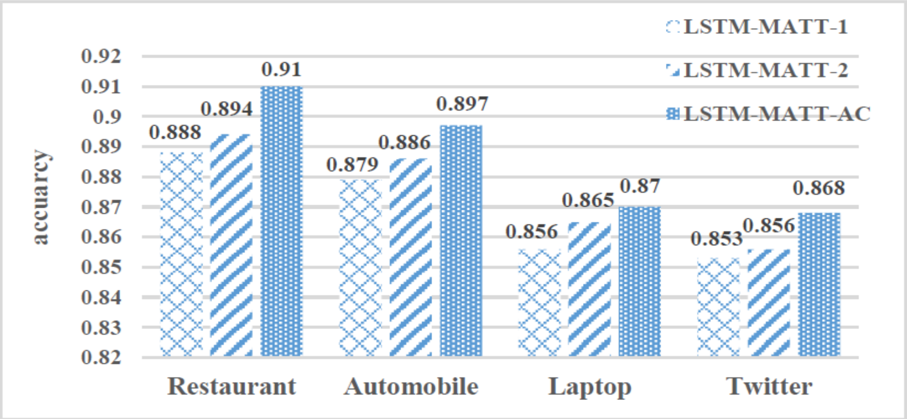


Figure 5. The impact of different aspect context processing methods on classification accuracy

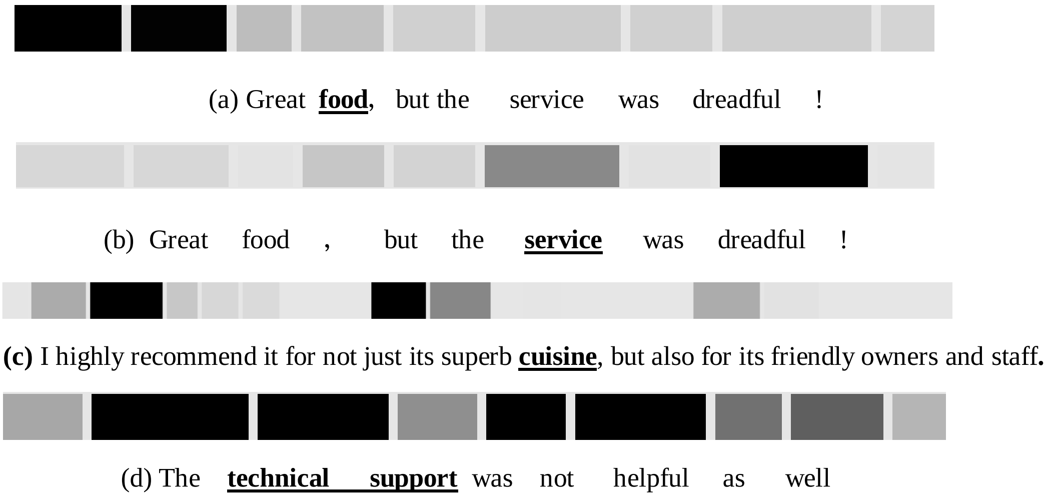
Figure 5 shows that the classification accuracy of the LSTM-MATT-2 is generally higher than that of the LSTM-MATT-1 on four domain-specific datasets. This indicates that encodes richer aspect context information than, and also verifies that the LSTM with three threshold mechanisms can effectively learn long-distance semantic dependencies in reviews. The accuracy of the LSTM-MATT-AC is higher than that of the LSTM-MATT-2, with improvements of 1.79%, 1.24%, 0.58%, and 1.4% in the Restaurant, Automobile, Laptop, and Twitter domains respectively. The result shows that aspect context module, which takes an aspect target as a target object and encodes the left and right context of aspect word with Bi-LSTM network, can make the LSTM-MATT-AC generate more accurate sentiment features with the aspect, and ultimately raise the accuracy of aspect sentiment classification.

Table 5.Attention weights of the word sequence in different models

Model	great	food	,	but	the	**service**	was	dreadful	!
LSTM-NP-AC	0.2880	0.0311	0.0311	0.0237	0.0269	0.0643	0.0463	0.4606	0.0280
LSTM-NC-AC	0.1783	0.1705	0.0885	0.0769	0.0771	0.0755	0.0874	0.2000	0.0458
LSTM-MATT-AC	0.0043	0.0021	0.0007	0.0012	0.0004	0.1250	0.0012	0.8691	0.0006

To further study the effect of the location and class attention on the aspect sentiment classification, we extracted a sample from the Restaurant dataset. We then compared the attention weights of word sequences among LSTM-MATT-AC, LSTM-NL-AC and LSTM-NC-AC. Table 5 shows that for the aspect target "service", the LSTM-NL-AC without location attention assigns the sentiment words "great" and "dreadful" a higher weight, while the LSTM-MATT-AC can accurately identify the correct sentiment word "dreadful" and increase its weight from 0.4606 to 0.8691. This weight fluctuation shows that the model with location attention can effectively capture the influence of words on the aspect target through the relative location information between an aspect target and each word in a review, enhance the model's content attention layer to recognize correct sentiment words, and thus generate more accurate aspect sentiment features through more reasonable weight allocation. On the other hand, because the number of positive reviews from the Restaurant dataset is about 2.4 times the number of negative reviews, the class is imbalanced. For the aspect target "service" whose sentiment polarity is negative in the instance, we find that the LSTM-NC-AC assigns weights to each word more evenly in the clause "but the service was dreadful!", and the sum of the weights allocated to the clause is 0.5169, which is much lower than the sum of the 0.9949 assigned to the clause by the LSTM-MATT-AC. After analysing the attention weights of word sequence, we found that the model will likely be biased when the training data is imbalanced. From the perspective of attention distribution, we find that the model reduces the attention to the class clauses with less data, which is mainly manifested in the lower weight sum given to clauses and the more average weight assigned to each word in the clauses. The above demonstrates the validity of the new model that introduces location attention and class attention in aspect sentiment analysis.

4.5. Visualizing Attention Mechanism



(a) Great **food**, but the service was dreadful !

(b) Great food , but the **service** was dreadful !

(c) I highly recommend it for not just its superb **cuisine**, but also for its friendly owners and staff**.**

(d) The **technical support** was not helpful as well

Figure 6. Attention visualizations. the color darkness reveals the importance degree

In Figure 6, we list four examples from the Restaurant and Laptop datasets, and visualizing the weight of each word's attention to determine which has the greatest effect on the emotional polarity of the aspect. The darkness of the colour indicates the grade of the attention weight [20]. The darker the colour is, the more important the word is. In the first two examples, the comment “Great food, but the service was dreadful!” contains aspect of the “food” and another aspect “service”. However, the emotional polarity of the aspect "food" reveals positive. However, the polarity of the "service" represents opposite, as two clauses with opposite sentiment polarity are connected by the word "but". The results of the attention visualizations in Figure 6 (a) and Figure 6 (b) show that our model accurately finds the corresponding sentiment words "great" and "dreadful" of aspect words "food" and "service" under the premise of identifying the turning sentences and thus predicts the correct sentiment polarity of the two aspect targets. In the third example, the correct polarity of the aspect word "cuisine" is positive in comment "I highly recommend it for not just its superb cuisine, but also for its friendly owners and staff". Although there is a negative word "not" in front of the word "cuisine", our model recognizes that the "not" does not mean negative in this review and therefore accurately classifies the sentiment of the aspect target “cuisine”. The last example, the polarity of the aspect "technical support" is negative in comment "The technical support was not helpful as well". Compared to the third example, our model recognizes that "not" is a negation of the sentiment word "helpful" that is positive and therefore accurately predicts the correct emotional polarity of the aspect.

5. Conclusion

This paper proposed a new model LSTM-MATT-AC to tackle challenges in aspect sentiment analysis. The model leveraged the multi-attention mechanism and aspect context, where multiple attention makes the new model to concentrate the relative between aspect targets and their context words from different perspectives. And accurately represent the impact of each context words on the aspect targets. Meanwhile, because different aspect targets have different contexts in a review, the aspect context module of the LSTM-MATT-AC independently encodes the both sides context of the aspect targets through Bi-LSTM. Evaluations on four domain-specific datasets indicate that new proposed model has improved from the accuracy compared to models from prior work, validating the high-performance of the multi-attention mechanism and aspect context on aspect sentiment analysis.

As our model averages the each word embedding in an aspect phrase, the features of important words in the aspect phrase may not be captured effectively. We plan to study how the introduction of the aspect of future work about the attention mechanism.

Acknowledgements

The paper was supported by the Shaanxi science and technology innovation foundation under Grant NO. 2019PT-12, the Graduate Innovation Foundation Project of Xi’an University of Posts & Telecommunications (CXJJLY2019094) and Tencent Technology (Shenzhen)Co., Ltd.

References

[1] Fan X, Li X, Du F, et al. Apply word vectors for sentiment analysis of APP reviews[C]// International Conference on Systems and Informatics. IEEE, 2017:1062-1066.

[2] Ilmania, Abdurrahman, S. Cahyawijaya and A. Purwarianti, "Aspect Detection and Sentiment Classification Using Deep Neural Network for Indonesian Aspect-Based Sentiment Analysis," 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, 2018, pp. 62-67.

[3] X. Fang and J. Tao, "A Transfer Learning based Approach for Aspect Based Sentiment Analysis," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 478-483.

[4] R. S. Silva, R. R. M. U. A. Rathnayaka, P. M. A. K. Wijeratne, M. T. N. Deshani, Y. H. P. P. Priyadarshana and D. Kasthurirathna, "Ontological Approach for Aspect-Based Sentiment Analysis," 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 2019, pp. 106-111.

[5] Ruder S, Ghaffari P, Breslin J G. A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis[J]. 2016.

[6] Rocktäschel T, Grefenstette E, Hermann K M, et al. Reasoning about Entailment with Neural Attention[J]. 2015.

[7] Ma D, Li S, Zhang X, et al. Interactive Attention Networks for Aspect-Level Sentiment Classification[J]. 2017.

[8] Liang B, Liu Q, Xu J, et al. Aspect-based sentiment analysis based on multi-attention CNN[J].Journal of Computer Research and Development,2017,54(08):1724-1735.

[9] Tang D, Qin B, Liu T. Aspect Level Sentiment Classification with Deep Memory Network[J]. 2016.

[10] Baziotis C, Pelekis N, Doulkeridis C. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis[C]// International Workshop on Semantic Evaluation. 2017:747-754.

[11] Han H, Li X, Zhi S, et al. Multi-Attention Network for Aspect Sentiment Analysis[C]// Proceedings of the 2019 8th International Conference on Software and Computer Applications. 2019:22-26.

[12] Lingling Song, Yazhou Zhang, Yuexian Hou. Convolutional neural network with pair-wise pure dependence for sentence classification[C]// 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), 2018:78-95.

[13] Zhai Penghua, Zhang Dingyi. Bidirectional-GRU Based on Attention Mechanism for Aspect-level Sentiment Analysis[C]// Proceedings of the 2019 11th International Conference on Machine Learning and Computing - ICMLC, 2019:56-64.

[14] Zhang, C., Li, Q., Song, D. (2019). Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. The 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP2019). 3-7 November 2019, Hong Kong, China.

[15] Song, Y. Wang J, Jiang T, Liu Z and Rao Y. 2019. Attentional encoder network for targeted sentiment classification. arXiv preprint arXiv:1902.09314.

[16] Huang, Binxuan & Carley, Kathleen M. (2018). Parameterized Convolutional Neural Networks for Aspect Level Sentiment Classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL.

AUTHORS

Xinzhi Ai received B.Eng. in electrical and electronic engineering from Hao Jing College of Shaanxi University of Science & Technology.in 2018, she is pursuing a master’s degree at Xi’an University of Posts and Telecommunications. As a research assistant at text mining Lab., Her research interests include deep learning, data mining and aspects sentimental analysis.

Xiaoge Li is a Professor of Computer Science at Xi’an University of Posts and Telecommunications. In 2000, He was with Cymfony, NY., USA engaging in research and development of information Extraction and Question & answering system. His currently research include natural language processing, Machine learning and knowledge Graphing.

Feixiong Hu received his B.Eng. in Computer Engineering from Xi’an University of Posts and Telecommunications in 2012. He is currently the leader of Artificial intelligence for IT operations intelligent operation (AIOps) of Shenzhen Tencent Computer System Co., Ltd. His current research interests are big data and machine learning.

Shuting Zhi received her B.Eng. (2016) and M.Eng. (2018) in Computer Engineering from Xi’an University of Posts and Telecommunications. Since 2018, she has joined Xiaomi Corporation, China. She has been working on the projects of xiaoai knowledge platform, word sense ambiguity and sentimental analysis.

Likun Hu received the master's degree in computer system structure from Northeast Forestry University in 2017. Currently, He is working in Shenzhen Tencent Computer System Co., Ltd., as a senior operations engineer of Shenzhen Tencent Computer System Co., Ltd., China. He has being engaged in development of big data analysis and log Anomaly Detection.

A TOPOLOGICAL METHOD FOR COMPARING DOCUMENT SEMANTICS

Yuqi Kong[1], Fanchao Meng[1] and Ben Carterette[2]

[1]Department of Computer & Information Sciences,
University of Delaware, Newark, USA
[2]Spotify, Greenwich Street, New York, USA

ABSTRACT

Comparing document semantics is one of the toughest tasks in both Natural Language Processing and Information Retrieval. To date, on one hand, the tools for this task are still rare. On the other hand, most relevant methods are devised from the statistic or the vector space model perspectives but nearly none from a topological perspective. In this paper, we hope to make a different sound. A novel algorithm based on topological persistence for comparing semantics similarity between two documents is proposed. Our experiments are conducted on a document dataset with human judges' results. A collection of state-of-the-art methods are selected for comparison. The experimental results show that our algorithm can produce highly human-consistent results, and also beats most state-of-the-art methods though ties with NLTK.

KEYWORDS

Topological Graph, Document Semantics Comparison, Natural Language Processing, Information Retrieval, Topological Persistence

1. INTRODUCTION

The problem focused in this paper is Document Semantics Comparison Problem as follows. Given two human readable, fairly long documents in similar lengths (and temporarily only in English), a real value to reflect the similarity between the two documents is desired. This problem is one of the fundamental problems lying at the heart of Natural Language Processing (NLP) and Information Retrieval (IR). To date, this problem has been attacked majorly from the statistical perspective, such as TF-IDF [1] [2] based methods, and vector space model (VSM) [3] methods. The former category directly utilizes TF-IDF information combined with statistical techniques to design the methods, while the latter category emphasizes and represents the relationships between words or constituents via VSM models, most of which are constructed still based on statistical information. In this paper, several state of the art and concrete such methods are selected for comparison. In Section 2, these methods will be briefly explained.

The methods proposed in this paper sheds a light from a different perspective, topology. More specifically, our methods are utilizing topological persistence [4] to represent the relationship between any two given documents, then the semantics similarity is computed from this representation. Our methods started with a very natural motivation. That is, if two documents have similar semantics, then they must have a relatively larger amount of relationships reflecting this similarity. Then a core task to formulate this similarity is to represent the relationships between the two documents. A document can be considered as a concrete carrier of its semantics. It consists of a collection of words and the relationships between these words. These relationships

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 143-151, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101411

are ciphered in the grammars and the conventions in human languages. From a computational perspective, our arsenal to represent these relationships is actually quite limited. Then we need to be creative to take advantage of our weapons. Parse trees [5], constituency-based and dependency-based, are powerful tools to represent word relationships within a sentence. To measure the similarity or distance between any two words, a number of candidates are also ready to be picked. Then utilizing these two types of tools, we are able to partially construct the relationships in two documents. In other words, within a document, words in a sentence are organized in a connected component via a parse tree, and words in different document can be connected via checking the similarity or distance between them. In this way, a combinatorial graph is constructed, in which every path from word to another reflects a direct or indirect relationship, and the collection of such relationships can consequently reflect a portion of the relationship between the two documents. Then how to extract features from this graph becomes the key to define the similarity of semantics.

Topological persistence represents and extracts features of topological spaces from the algebraic topological space, and in algebraic topology language, it is also a one-dimensional abstract simplicial complex [4]. Our methods compute homology groups [4] and the corresponding topological persistence [6] for representing the relationships between the documents. The final similarity scores are computed based on the topological persistence. Again, in an intuitive way, the topological persistence for a given dimension contains a set of “holes” which have birth and death [6] so that their lifetime can be measured. The longer the lifetime, the more important the “hole”. Moreover, in our case, a “hole” represents a piece of a sentence in one document has a relatively strong semantic relationship with a piece of another sentence in the other document. The background of topological persistence will be introduced in Section 3.

The major contribution of this paper is a novel algorithm to compute semantic similarity between documents. For the experimentation, we compare the similarity scores produced by our algorithm with those given by human judges, and also, we compare our algorithm with a collection of state-of-the-art methods. The experimental results strongly support that firstly our algorithm can produce similarity scores highly consistent with human judges; meanwhile it has better performance than most methods selected though ties with NLTK. Section 4 and 5 will present our algorithm and the experiments respectively. Discussion and conclusion follow in section 6 and 7.

2. RELATED WORK

The first method that we have particular interest is Doc2vec [7], which is a deep learning model designed base on Word2vec [8]. The essential idea of this method is actually not complicated. Since Word2vec is a model to represent words, then why not we add another vector (for paragraph) to represent document. Then the authors of Word2vec throw in another model named as Distributed Memory version of Paragraph Vector (DMPV) [8] which acts as a memory, in a rough way to explain, memorizing the topic of document. This model has demonstrated some significant progress on several NLP tasks such as topic extraction and sentiment analysis [8]. However, topic extraction and sentiment analysis are not equivalent tasks to document semantics comparison problem, since the latter problem is concentrating on the base level semantics (without any implications, metaphors, ironies or any other such concerns). Then it is also interesting to know if this state-of-the-art method would also work well on our problem.

The second work that has been selected is a state-of-the-art concrete software library for NLP research and development, named as NLTK [9]. In this library, a vector space model-based method [11] for comparing document semantics can be found. Its general idea is nothing more a classic one. that is, TF-IDF combined with cosine similarity, and the vector space model in

NLTK is actually constructed via utilizing TF-IDF. However, it optimizes the procedure a lot via some statistic techniques, and in this way, its performance has been having a good reputation.

The third word is Text2vec [12]. This is one of the newest implementation libraries for NLP, and it contains four methods for comparing document semantics. They are Jaccard similarity [13], Cosine similarity (similar as the one provided by NLTK), Cosine with LSA [14] and Euclidean distance with LSA. These methods are all implemented based on vector space model and statistics. The most interesting point is that a couple of methods are combined with LSA which has been playing an important role in topic extraction.

We are not aware of any published work to date on use of topological persistence on representation of semantics and comparisons except a paper published in 2013 by Xiaojin Zhu [15]. The author shows a relatively preliminary application of persistent homology in natural language processing. The objective of this paper is use persistence to distinguish simple rhythmic literatures, and child or adolescent writings. Specifically, in the methods proposed in this paper, each paragraph is formulated as a bag-of-words vector, and the datasets are nursery rhymes, and child and adolescent writings; moreover, the use of topological persistence is limited to the number of "holes". Our methods to represent and compare semantics will be designed for more sophisticated and general cases.

3. THEORETICAL BACKGROUND

In this section, we provide a brief introduction to fundamentals of homology theory and persistence. Intuitively, the major task of homology theory is to describe "holes" in a geometric space from the algebraic perspective, and persistence describe how persistent these "holes" are.

We start with the formal definition of the most important concept for formulating geometric spaces from the algebraic perspective, namely abstract simplicial complex.

In an abstract simplicial complex, a p-simplex is a p-dimensional simplex (e.g. a line segment is a 1-simplex, or a triangle convex hull is a 2-simplex). A p-chain is a formal sum of a set of p - simplices, written as $\sum_{k} \alpha_k \sigma_k$, where α_k is a co-efficient in the ground field $\mathbb{F}$ and σ_k is a simplex. The p-cycles and the p-boundary are all p-chains. They are defined by the boundary operator, denoted by ∂_p. The boundary operator maps p-chains to $(p-1)$-boundaries. For example, given a triangle-shape convex hull, the boundary operator takes this convex hull and returns the triangle consisting of three-line segments without the interior of the convex hull. The resulting triangle is called the boundary of the convex hull. A p-cycle is a p-chains to whom apply the boundary operator will return zero. In other words, $\partial_2 = 0$. This property of boundary operators is called the chain complex property [4] [16]. A p-homology-classes is a set of p-cycles equivalent to one another, and the equivalence relation is defined in the way that if two p-cycles $\mathcal{C}_p^i$ and $\mathcal{C}_p^j$ are equivalent then $\mathcal{C}_p^i - \mathcal{C}_p^j$ is a p-boundary. A p-homology-group is the set of p-homology-classes computed from a complex. In a formal way, the p-homology-group is defined as $\mathcal{H}^p = \frac{\mathcal{Z}_p}{\mathcal{B}_p}$, where $\mathcal{Z}_p$ is the p-cycle group, and $\mathcal{B}_p$ is the p-boundary group.

Equivalently, the homology *group* [4] is also defined as $\mathcal{H}^p = \frac{\mathcal{K}er(\partial_p)}{\mathcal{I}m(\partial_{p+1})}$.

A filtration is a sequence of indexed sets attached to the abstract simplicial complex, where each simplex is assigned with a filtration value [6] indicating the moment when this simplex is about to appear in the sequence, the birth of a p-homology-class is the earliest moment when any of its representative $\mathcal{p}$-cycle appears in the sequence, the death of a $\mathcal{p}$-homology-class is the earliest moment when a $\mathcal{p}+1$-cycle containing the exact set of vertices of any of this $\mathcal{p}$-homology-class's $\mathcal{p}$-cycle appears (N.B. this moment can be infinity), and the lifetime of a $\mathcal{p}$-homology-class is the distance between its birth and death. A visualization of the collection of the set of $\mathcal{p}$-homology-classes with birth-death pairs computed from an abstract simplicial complex with a filtration is called a persistence diagram.

4. Algorithms

Given:

Two English documents, $\mathrm{D_i}$, $\mathrm{D_j}$.
A predetermined parameter, $\theta_t \in [0,1]$.
A stopwords list, $\mathrm{W_s}$

Seek:

A real value reflecting the semantic similarity between $\mathrm{D_i}$ and $\mathrm{D_j}$.

Step 1: For each sentence, $\mathrm{S_{ik}} \in \mathrm{D_i}$, and each sentence, $\mathrm{S_{jh}} \in \mathrm{D_j}$, where $\mathrm{k, h}$ are indices for the two sentences respectively, compute their dependency-based parse trees.

Step 2: Do tokenization and lemmatization on each parse tree, (i.e. prune non-word terminals and convert words to lemmas).

Step 3: For each term pair $(\mathrm{t_{ik}^p}, \mathrm{t_{jh}^q})$, where $\mathrm{t_{ik}^p} \in \mathrm{S_{ik}}$, $\mathrm{t_{jh}^q} \in \mathrm{S_{jh}}$, and $\mathrm{t_{ik}^p}$, $\mathrm{t_{jh}^q} \notin \mathrm{W_s}$, and $\mathrm{p, q}$ are indices for the two terms respectively, compute the word similarity between $\mathrm{t_{ik}^p}$ and $\mathrm{t_{jh}^q}$ via utilizing Wordnet [10] LIN similarity [17]. We denote this similarity $\tau_t(\mathrm{t_{ik}^p}, \mathrm{t_{jh}^q})$, and $\tau_t(\mathrm{t_{ik}^p}, \mathrm{t_{jh}^q}) \in [0,1]$.

Step 4: θ_w is taken as a threshold for the word similarities. For each $\tau_t(\mathrm{t_{ik}^p}, \mathrm{t_{jh}^q})$, if $\tau_t(\mathrm{t_{ik}^p}, \mathrm{t_{jh}^q}) \geq \theta_t$, then the two terms $\mathrm{t_{ik}^p}$ and $\mathrm{t_{jh}^q}$ are considered as two vertices and placed into an empty graph, and also an edge between the two terms is created. The weight on this edge is $\tau_t(\mathrm{t_{ik}^p}, \mathrm{t_{jh}^q})$.

Step 5: For each parse tree obtained in Step1, it is a graph, in which the vertices are the terms and the edges are determined by the tree. For each edge in this graph, set its weight to1. Then union all such graphs obtained from the parse tree with all resulting graphs from Step 4. The final graph will be an undirected and weighted graph, denoted by $\mathcal{G}_{\mathrm{ij}}$.

Step 6: The graph obtained from Step 5 is a one-dimensional abstract simplicial complex, denoted by Σ_{ij}^1. Given this abstract simplicial complex, compute the homology group for dimension one, denoted by $\mathcal{H}_{\mathrm{ij}}^1$.

Step 7: Set the filtration value for each simplex in Σ_{ij}^1 via utilizing the weights in $\mathcal{G}_{\mathrm{ij}}$ (i.e. the filtration value for an edge, $(\mathrm{t_{ik}^p}, \mathrm{t_{jh}^q})$, which is a one-dimensional simplex is set to $1 - \tau_t(\mathrm{t_{ik}^p}, \mathrm{t_{jh}^q})$, and the filtration values for the two corresponding vertices which are zero-dimensional simplices on this edge are all set to the same values.

Step 8: Given the homology group, $\mathcal{H}_{\mathrm{ij}}^{1}$, and the abstract simplicial complex, Σ_{ij}^{1}, obtained from Step 6, and the filtration values for the simplices obtained from Step7, compute the topological persistence for dimension one, denoted by $\mathcal{P}_{\mathrm{ij}}^{1}$.

Step 9: In $\mathcal{P}_{\mathrm{ij}}^{1}$, for each homology class, denoted by $[\mathrm{c_l}]$, its birth is determined by the minimum filtration value of the simplices in $\mathrm{c_l}$; and its death is determined by the maximum filtration value which is equal 1. Compute the lifetime of $[\mathrm{c_l}]$ which is equal to $\min_{(\mathrm{t_u^l},\mathrm{t_v^l})\in \mathrm{c_l}} \{\tau_t(\mathrm{t_u^l},\mathrm{t_v^l})\}$, where $(\mathrm{t_u^l},\mathrm{t_v^l})$ is a one-dimensional simplex in $\mathrm{c_l}$ (which is also an edge in Σ_{ij}^{1}).

Step 10: The final similarity between $\mathrm{D_i}$, and $\mathrm{D_j}$ is the sum of all lifetimes of the homology classes in $\mathcal{P}_{\mathrm{ij}}^{1}$, (i.e. $\sum_{[\mathrm{c_l}]\in\mathcal{P}_{\mathrm{ij}}^{1}} \min_{(\mathrm{t_u^l},\mathrm{t_v^l})\in \mathrm{c_l}} \{\tau_t(\mathrm{t_u^l},\mathrm{t_v^l})\}$), obtained from Step 9.

5. EXPERIMENTATION

Design: The goal of this experiment is evaluating the performance of Algorithm (TopoSem) proposed in Section 4. TopoSem will be compared to human judges. The performance of TopoSem should reflect how competent this algorithm can compare semantics of two documents as human judges. A dataset containing a collection of English documents will be utilized. For each pair of documents, human judges determine if this pair of documents have similar semantics, and provide a score to measure their similarity. Taking these similarity scores, two groups of document pairs can be constructed. One group contains all pairs that are determined by human judges as similar in semantics, and the other contains dissimilar pairs. Then TopoSem is applied to both groups to give each document pair in the two groups a semantic similarity score. If these scores given by TopoSem agree on the two groups, then the performance of TopoSem is considered as positive. Essentially, this experiment is a classification task, where the two classes are determined by the two groups, and TopoSem will be tested on classifying document pairs collected from the two groups into the two classes.

Settings: The dataset in use is provided by Michael D. Lee [18] which contains 50 documents selected from the Australian Broadcasting Corporation's news mail service. The lengths of documents vary from 51 to 126 words, and cover a number of broad topics. The documents in this dataset have been evaluated by human judges. For each pair of documents, there is an average of scores from human judges ranging from 1 to 5, where 1 indicates highly unrelated, and 5 indicates highly related. Besides this 50-document dataset, Michael D. Lee also provides an additional dataset containing 300 background documents which are in average longer than the 50 documents.

To utilize WordNet [10] Lin [17] to compare word meanings, an information content database needs to be specified. What is selected in this experiment is SemCor provided by WordNet 3.0.

Group 1, denoted by $\mathbb{G}_1$, of document pairs is constructed by collection all pairs that are scored $\leq$ 2.5, and Group 2, denoted by $\mathbb{G}_2$, is constructed by collecting all pairs scored $\geq$ 3.5. $\mathbb{G}_1$ contains 1095 pairs, and $\mathbb{G}_2$ contains 46 pairs. Since the rest of pairs could be hardly determined as similar or dissimilar even by human judges, then they are not considered in our experiment. The predetermined parameter θ_t takes values 1, 0.95, 0.9, 0.85 and 0.8 for five trials. The stopwords list in use is provided by Onix Text Retrieval Toolkit [19] which contains 571 words.

Method: TopoSem is applied to both $\mathbb{G}_1$ and $\mathbb{G}_2$ to compute a similarity score for each document pair. For each group, the 95% confidential interval of the scores is computed, denoted by $\mathcal{I}_c^1$ and

$\mathcal{J}_c^2$ respectively, where the superscripts are indices. $\mathbb{G}_1$ is set as the positive class (setting $\mathbb{G}_2$ as the positive class is also tested). If a document pair in $\mathbb{G}_1$, denoted by $d_1(\mathrm{x}, \mathrm{y}) \in \mathbb{G}_1$, where 1 is the group index, x, y are document indices, is given a score, denoted by $\alpha(d_1(\mathrm{x}, \mathrm{y}))$, which holds $\alpha(d_1(\mathrm{x}, \mathrm{y})) \leq \mathcal{U}(\mathcal{J}_c^1)$, then $d_1(\mathrm{x}, \mathrm{y})$ is considered as a true positive, where $\mathcal{U}(\cdot)$ takes the supremum of a given interval. If $\alpha(d_1(\mathrm{x}, \mathrm{y})) \geq \mathcal{L}(\mathcal{J}_c^2)$, then $d_1(\mathrm{x}, \mathrm{y})$ is a false negative, where $\mathcal{L}(\cdot)$ takes the infimum of a given interval. Similarly, for a document pair $d_2(\mathrm{a}, \mathrm{b}) \in \mathbb{G}_2$, if $\alpha(d_2(\mathrm{a}, \mathrm{b})) \geq \mathcal{L}(\mathcal{J}_c^2)$, then $d_2(\mathrm{a}, \mathrm{b})$ is considered as a true negative; and if $\alpha(d_2(\mathrm{a}, \mathrm{b})) \leq \mathcal{U}(\mathcal{J}_c^1)$, then $d_2(\mathrm{a}, \mathrm{b})$ is considered as a false positive. Foreach trial (with a specific θ_t), all true positives, true negatives, false positives and false negatives are collected and counted, and then precision, recall and F1 score are calculated.

Table 1. Error rates for TopoSem and control groups of methods

Methods	**$\mathbb{G}_1$ Error Rate**	**$\mathbb{G}_2$ Error Rate**	**Average Error Rate**
TopoSem (θ_t = 1.00)	**2.19 %**	19.56 %	10.88 %
TopoSem (θ_t = 0.95)	2.37 %	19.56 %	10.97 %
TopoSem (θ_t = 0.90)	3.01 %	19.56 %	11.28 %
TopoSem (θ_t = 0.85)	4.29 %	17.39 %	10.84 %
TopoSem (θ_t = 0.80)	6.48 %	21.73 %	14.11 %
Doc2Vec	33.33 %	17.39 %	25.36 %
Text2vec (Jaccard	15.07 %	39.13 %	27.10 %
Text2vec (Cosine	9.50 %	39.13 %	24.32 %
Text2vec (Cosine +	9.50 %	32.61 %	21.01 %
Text2vec (Euclidean +	7.03 %	32.61 %	19.82 %
NLTK	2.28 %	**15.22%**	**8.75 %**

Table 2. Precisions, Recalls and F1 scores with Group 1 as positive and Group 2 as negative.

$\mathbb{G}_1$as positive and $\mathbb{G}_2$ as negative			
Methods	**Precision**	**Recall**	**F1**
TopoSem (θ_t = 1.00)	0.97	0.99	**0.98**
TopoSem (θ_t = 0.95)	0.97	0.99	**0.98**
TopoSem (θ_t = 0.90)	0.96	0.99	0.97
TopoSem (θ_t = 0.85)	0.94	0.99	0.96
TopoSem (θ_t = 0.80)	0.91	0.99	0.95
Doc2Vec	0.60	0.99	0.75
Text2vec (Jaccard	0.84	0.98	0.91
Text2vec (Cosine	0.89	0.98	0.94
Text2vec (Cosine + LSA)	0.89	0.98	0.93
Text2vec (Euclidean +	0.91	0.98	0.94
NLTK	0.96	0.99	**0.98**

Table 3. Precisions, Recalls and F1 scores with Group 2 as positive and Group 1 as negative.

$\mathbb{G}_2$ as positive and $\mathbb{G}_1$ as negative			
Methods	**Precision**	**Recall**	**F1**
TopoSem (θ_t = 1.00)	0.59	0.35	0.44
TopoSem (θ_t = 0.95)	0.59	0.33	0.43
TopoSem (θ_t = 0.90)	0.63	0.31	0.42
TopoSem (θ_t = 0.85)	0.62	0.22	0.32
TopoSem (θ_t = 0.80)	0.50	0.12	0.20
Doc2Vec	0.55	0.03	0.05
Text2vec (Jaccard	0.25	0.04	0.06
Text2vec (Cosine similarity)	0.31	0.07	0.12
Text2vec (Cosine + LSA)	0.44	0.10	0.17
Text2vec (Euclidean + LSA)	0.50	0.16	0.25
NLTK	0.70	0.39	**0.50**

In this experiment, a control group of methods are also tested on the same task. The methods include Doc2vec, Text2vec and NLTK. One of the implementations of Doc2vec (whose name is Gensim [20]) provides a direct interface to compare semantics of two documents. Text2vec provides Jaccard similarity, cosine similarity [22], cosine similarity with TF-IDF, cosine similarity with LSA and Euclidean distance with LSA these methods for comparing document semantics directly. NLTK provides a vector space model based on TF-IDF, then the document similarity can be computed via cosine similarity.

Experimental Results: The experimental results are listed in Table 1, Table 2 and Table 3. Table 1 shows the error rate for each method. In this table, the winner for $\mathbb{G}_1$ is our TopoSem with $\theta_t = 1.00$ while NLTK and our TopoSem with $\theta_t = 0.95, 0.90$ produce similar results. The winner for $\mathbb{G}_2$ is NLTK while our TopoSem with θ_t = 0.85, 1.00, 0.95, 0.90 also produce similar results. In average, the winner is NLTK and our TopoSem with $\theta_t = 0.85$ is the second winner. The difference between the average error rate of NLTK and that of TopoSem with $\theta_t = 0.85$ is 2.09% which is not significant. It can be observed that the error rates on $\mathbb{G}_2$ are higher the error rates on $\mathbb{G}_1$. The reason is that the dataset is skewed so that the misclassified pairs in $\mathbb{G}_2$ impact the error rates on $\mathbb{G}_2$ much more significantly than the misclassified pairs in $\mathbb{G}_1$.

Table 2 shows the precision, recall and F1 score for each method under the case that $\mathbb{G}_1$ is set as positive and $\mathbb{G}_2$ is set as negative. The winner of F1 score is our TopoSem with $\theta_t = 1.00, 0.95$ and NLTK. Table 3 shows the case that $\mathbb{G}_1$ is set as negative and $\mathbb{G}_2$ is set as positive. In the latter case, NLTK is slightly better than our TopoSem but still not significant. Additionally, the reason that the F1 scores in Table 3 are lower than those in Table 2 is again because of the skew in the dataset.

6. Discussion

From table 1, deep learning methods such as Doc2vec and Text2vec have much worse error rate compare with our method and NLTK. Since those two deep learning methods require massive training documents to pretrain the model. If in the scenario that lacks such pretraining dataset, such as Michael D. Lee's dataset we used which only contains 300 background documents. The performance of those methods will hurt. Contrarily, non-training methods such as our topological method and NLTK are capable in any scenario.

Can our method do better? This was the first question we asked ourselves right after the results popped out. Since the performance of our method seems does not significantly better then NLTK and even slightly worse in the case that $\mathbb{G}_1$ is set as negative and $\mathbb{G}_2$ is set as positive. The bottleneck comes from Wordnet and LIN. They are far out of date tools, the number of synsets in Wordnet is not adequate, those out-of-vocabulary words compromised the performance by hindering the formation of simplicial complex (aka, meaningful "holes"). Furthermore, LIN may not be the best option either. WordNet provided LIN as its out-of-box word similarity algorithm.

However, our major goal is to propose a unique novel topological structure that can unify both syntactic and lexical semantics of the document and quantify the semantics without any pretraining procedure. The results proved the validity of our proposal. Moreover, there is a lot of room for improvement.

7. CONCLUSION

In this paper, a novel algorithm for comparing document semantics is proposed. This algorithm is designed based on topological persistence, which is distinguished from most methods for the same task. The experimental results provide strong support to our algorithm showing that it can unify both syntactic and lexical semantics of documents, then produce highly human-consistent results, and also outperform some state-of-the-art methods.

8. FUTURE WORK

Although, TopoSem shows potentials that it is highly consistent with human judgment. The results indicated the performance does not significantly outperform the control methods. There are many aspects that we can do to improve this novel approach. A new version of the algorithm is under development. We are plan to involve parse tree trimming to trim unnecessary nodes in order to reduce the effect of noise homology classes. For the current algorithm, we only use filtration value to weight terms edges formed in step 4. In the next version of the algorithm, we try to not only weigh the terms edges but also parse tree edges then use harmonic mean to combine two types of weights together. We hope this could give TopoSem a more comprehensive similarity function. Furthermore, one of the limitations of TopoSem is complexity, this impedes the application of TopoSem to large datasets. After the algorithm matured, we will focus on the optimization of TopoSem.

REFERENCES

[1] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4), 309-317.

[2] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation.

[3] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[4] Munkres, J. R. (2018). Elements of algebraic topology. CRC Press.

[5] Jurafsky, D., & Martin, J. H. (2014). Speech and language processing. Vol. 3.

[6] Edelsbrunner, H., & Harer, J. (2010). Computational topology: an introduction. American Mathematical Soc..

[7] Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196).

[8] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[9] Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.

[10] Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May). WordNet:: Similarity-Measuring the Relatedness of Concepts. In AAAI (Vol. 4, pp. 25-29).

[11] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[12] Selivanov, D. (2016). text2vec: Modern Text Mining Framework for R: R package version 0.4. 0.

[13] Real, R., & Vargas, J. M. (1996). The probabilistic basis of Jaccard's index of similarity. Systematic biology, 45(3), 380-385.

[14] Dumais, S. T. (2004). Latent semantic analysis. Annual review of information science and technology, 38(1), 188-230.

[15] Zhu, X. (2013, August). Persistent homology: An introduction and a new text representation for natural language processing. In IJCAI (pp. 1953-1959).

[16] Hatcher, A. (2002). Algebraic Topology. Cambridge University Press.

[17] Lin, D. (1998, July). An information-theoretic definition of similarity. In Icml (Vol. 98, No. 1998, pp. 296-304).

[18] Lee, M. D., Pincombe, B., & Welsh, M. (2005). An empirical evaluation of models of text document similarity. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 27, No. 27).

[19] Buckley, C., & Salton, G. (2007) Onix Text Retrieval Toolkit Stopword List2. http://www.lextek.com/manuals/onix/stopwords2.html.

[20] Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

[21] Real, R., & Vargas, J. M. (1996). The probabilistic basis of Jaccard's index of similarity. Systematic biology, 45(3), 380-385.

[22] Karypis, M. S. G., Kumar, V., & Steinbach, M. (2000, May). A comparison of document clustering techniques. In TextMining Workshop at KDD2000 (May 2000).

MODERATION EFFECT OF SOFTWARE ENGINEERS' EMOTIONAL INTELLIGENCE (EQ) BETWEEN THEIR WORK ETHICS AND THEIR WORK PERFORMANCE

Shafia Khatun and Norsaremah Salleh

Department of Computer Science, Kulliyah of Information and
Communication Technology (KICT), International Islamic
University Malaysia (IIUM), Kuala Lumpur, Malaysia

ABSTRACT

In today's world, software is being used in every sector, be it education, healthcare, security, transportation, finance and so on. As software engineers are affecting society greatly, if they do not behave ethically, it could cause widespread damage, such as the Facebook-Cambridge Analytica scandal in 2018. Therefore, investigating the ethics of software engineers and the relationships it has with other interpersonal variables such as work performance is important for understanding what could be done to improve the situation. Software engineers work in rapidly-changing business environments which lead to a lot of stress. Their emotions are important for dealing with this, and can impact their ethical decision-making. In this quantitative study, the researcher aims to investigate whether Emotional Intelligence (EQ) moderates the relationship between work ethics of software engineers and their work performance using hierarchical multiple regression analysis in SPSS. The findings have found that EQ does significantly moderate the relationship between work ethics and work performance. These findings provide valuable information for improving the ethical behaviour of software engineers.

KEYWORDS

Software engineers, emotional intelligence, work ethics, work performance, quantitative study

1. INTRODUCTION

1.1. Ethics in Information and Communication Technology (ICT)

Development of software has become a core element in today's Information and Communication Technology (ICT) based society [1] [2]. Governments, educational institutions, healthcare organizations, national security, transportation, and many other sectors are using systems that involve software. Without the software, these systems would not function. These systems are benefiting society greatly, and in the current global COVID-19 pandemic, their benefits can be felt even more. Many of the services these systems provide can be accessed online from a distance which is crucial during the current pandemic [3]. Therefore, it can be said that the software engineers behind the development of these software are impacting society greatly [1] [2] [4]. As they have the power to affect society, they also have a duty to exercise that power responsibly and ethically or it could cause widespread damage. There are many instances in the past where this has happened.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 153-168, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101412

In March 2018, news all over the world exploded with the Facebook-Cambridge Analytica scandal. Cambridge Analytica, a data analytics (involving software), political, advertising and consulting company based in the UK, was found to have harvested the personal data of millions of people's Facebook profiles without their consent and used it to influence a variety of political campaigns. It resulted in a huge public outcry, a massive fall in Facebook's stock price, Mark Zuckerberg (the founder of Facebook) had to testify in front of the United States' congress, and there were calls for tighter regulation on software companies' use of data [5].

The year before, in 2017, the Volkswagen emissions scandal occurred. Volkswagen was found to have used a defeat device to cheat on emissions tests mandated by the Environmental Protection Agency (EPA) and the California Air Resources Board (CARB) to sell approximately 590,000 diesel vehicles in the US. Misinformation was purposely being generated by software that was developed and implemented by professionals who were knowingly part of this unethical and illegal act [4].

Another example of software misuse is ransomware. According to Lord [6], more than 7,600 ransomware attacks were reported to the Internet Crime Complaint Centre (IC3) between 2005 and March of 2017. In 2015 alone, the IC3 received 2,453 ransomware complaints that cost victims over $1.6 million. According to Lord [6], these figures only represent the attacks reported to the IC3; the actual number of ransomware attacks and costs is much higher. Also, during the current pandemic, countries across the globe have reported an increase in ransomware attacks [3]. Ransomware is a form of malicious software from cryptovirology that threatens to publish the victim's data or perpetually block access to it unless a ransom is paid [7]. This means that people who are able to develop software are intentionally using their skills for unethical purposes. This illegal extortion of money is already a great crime. But when it puts people's lives at risk, the crime becomes even worse. In 2017, in May, a ransomware attack in the UK caused the closure of many of the emergency units in hospitals in the UK, endangering patients [4].

These examples are only a few of the many incidences where software is developed and used for illegal or unethical reasons. Ethics can be simply defined as "a system of moral principles" [8, p. 1]. They influence "how people lead their lives, for life is an unbroken stream of decision-making and ethics are concerned with what is the right moral choice, for individuals and for society" [8, p. 1]. Within an organization, staff behaving ethically would mean that they complete their work responsibly with integrity and honesty, that there work is beneficial and not harmful, and "adhere to policies and rules while working to meet the aims" of the organization [8, p. 1].

Rogerson et al. [4] said that the consideration of ethics in the Information Technology (IT) industry is often neglected until events like those described earlier occur. In the aftermath of these incidences, members of the public reignite the discussion on the ethical standards of IT companies. They call for tighter regulations on them. However, even with them voicing these demands time and time again, not enough is being done. In a keynote speech at ICSE 2015, Grady Booch put forward the notion that "every line of code has a moral and ethical implication." Ferrario et al. [9] said that "anything digital is inevitably affected by values: the organizational values of the project sponsor, the values of the research partners, and the values of each developer and designer" [p. 1]. Thomson and Schmoldt [10] mentioned that, from an ethical perspective, technology improvement should protect human values while improving computer software. However, as said before, the presence of human values in software engineering (SE) is usually neglected or invisible until the widespread consequences of breaching them are felt [4] [9]. In education and research as well, ethics are not given enough emphasis [11]. Logan and Clarkson [12], in 2005, did a review of programs offered by universities in information security and found that there was a pattern of "strong technical content and the absence of an ethics and computer law course" [pp. 157-161]. In 2017, Rogerson et al. [4] said that in reviewing topics covered by several international and national information systems conferences and textbooks, the topic of ethics is totally missing or minimally represented. This is very surprising as, in the IT world, ethics is a very crucial issue. It can be observed that not much has improved between 2005 and 2017. These findings show that there is a dire need to increase awareness, interest and action on the ethical dimension of ICT in education, in industries, and in research.

One way of taking action in research is to investigate the relationship between the ethics of software engineers and other variables, especially interpersonal variables of software engineers. Günsel and Açikgöz [13] said that in the context of organizational behavior (OB), "software development remains a poorly understood process" [p. 14]. Based on what previous studies have mentioned and what the researchers have reviewed in past literature, research on the interpersonal characteristics and competencies of software engineers are scant [14] [15], especially in relation to their ethics. With the recurring incidents of unethical behaviour in the IT/SE industry, educators, practitioners and researchers are becoming more aware that more effort and attention should be given to understanding the software engineers developing the software [14]. If the relationships between the ethics of software engineers and their interpersonal characteristics are investigated, this will help in understanding the dynamics surrounding the ethics of software engineers and what can be done to improve the situation.

Therefore, in this study, the researchers aim to investigate the ethics of software engineers and how it is related with other interpersonal variables as this may give valuable information on what can be done to improve the situation. Two such variables are the emotional intelligence (EQ) of software engineers and their work performance. The rationale for choosing these variables is explained in the next section.

1.2. Emotional Intelligence (EQ) and Work Performance

There is an evident relationship between an employee's individual characteristics, such as his/her ethical beliefs, and unethical behaviour at the workplace [16]. Additionally, there may also be situational characteristics that can cause an employee to behave unethically. Affective Events Theory (AET) which was developed by organizational psychologists, Weiss and Cropanzano [17], explains that emotions experienced by workers at the workplace have an influence on handling workplace situations which can be positive or negative. Software engineers work in rapidly-changing business environments [18] that are demanding and competitive [13] which results in them facing a lot of pressure which leads to stress [15]. Rezvani and Khosravi [15] say that increasing levels of stress affects a software engineer's "ability to self-regulate their feelings and understanding" [p. 139]. As humans, software engineers not only bring their personalities, ethical beliefs, knowledge and skills to their work, they also bring their emotions [19]. Emotional intelligence (EQ) was first described formally by Salovey and Mayer [20]. They defined it as "the ability to monitor one's own and others' feelings and emotions, to discriminate among them and to use this information to guide one's thinking and actions" [p. 189]. Later on, researchers came up with more detailed definitions such as ''the set of abilities that enable a person to generate, recognize, express, understand, and evaluate their own, and others' emotions in order to guide thinking and action that successfully cope with emotional demands and pressures'' [21, p. 72]. The complex nature of a software engineer's job calls for intense social interactions, and such interactions produce many emotions [22]. In stressful situations, a software engineer with a higher EQ is more likely to manage their emotions better, leading to lower levels of stress [15] [23]. Previous studies have found that employees with higher EQs had positive correlations with work performance [15] [24] [25] [26] [27], team performance [13] [28] [29], job satisfaction [30], and successful software projects [13] [31] [32]. In fact, studies have shown that skills associated with EQ are twice as important for career success as intelligence (IQ) or technical skills [33] [34]. Holt and Jones [35] highlighted the economic value of EQ: ''In the age of information and highly specialized work teams, emotional intelligence is becoming a vital skill as people must accomplish their work by collaborating with each other, and their ability to communicate effectively becomes as critical, if not more critical, as technical skills and capabilities" [p. 15].

Furthermore, Conservation of Resources (COR) theory [36], as explained by Rezvani and Khosravi [15], describes that "people aim to keep, safeguard, and construct resources. Resources are defined as objects, conditions, personal abilities, and attributes or energies in which the individual places value [36]. These can either be valued in themselves or facilitate the achievement of specified objectives [36]. A setting that drives an individual to anticipate a possible or actual loss of resources, or the application of resources without the potential to acquire further resources, can therefore generate stress [36] [37]" (p. 140). Adding to what was described earlier, software engineers experience many difficulties in their efforts to successfully complete their projects, especially in complex and broad projects [38], and this increases levels of stress. However, the development of personal competencies such as EQ can serve as a buffer to any loss of resources in such situations [39] [40]. EQ can be an "interpersonal support resource" [15, p. 140] that can be used to reduce the negative influence of stress on software engineers, especially on their ethical decision-making. Self-report EQ measures that are based on Salovey and Mayer's [20] definition of EQ, such as Schutte et al. [41] and Wong and Law [42], assert that "emotional capabilities are acquired skills that can be nurtured and enhanced" [15, p. 140]. Researchers have found that EQ is a skill that can be trained and have conducted studies where they trained students and professionals alike and found the results to be positive; their EQ did increase [43] [44]. Therefore, as EQ is a skill that can be nurtured and enhanced, and according to COR theory, EQ may be an important resource for software engineers to acquire that can assist them in dealing with the various positive and negative situations that occur in their work, especially with regards to ethical decision-making.

Additionally, in relation to ethics, researchers have pointed out that just having a high EQ is not sufficient for employees. Gibbs et al. [45] said that ''without a moral compass to guide people in how to employ their gifts, emotional intelligence can be used for good or evil'' [p. 67]. Also, Maak and Pless [46] argued that ''responsible leaders need both emotional and ethical qualities to guide their action and behaviour in interaction'' [p. 105]. Segon and Booth [47] reviewed several emotional competency inventory frameworks and found that, in terms of their definitional constructs, it is possible for an employee to display high EQ but still be an unethical individual; they demonstrated this through analyzing several high-profile cases of business leaders who were praised by the public but were later found to have behaved unethically. They said that the frameworks did not include an ethical dimension in them. They concluded that ethics competencies must be included in the frameworks to guide EQ. And although studies exploring the relationships between ethics and EQ are scant, the few that are available have shown that the ethics of an employee were most effective when the employee had high EQ and vice versa [48] [49] [50] [51] [52].

Therefore, repeating what was said earlier, many unethical incidents have occurred in the IT industry and keep occurring. These have caused widespread damage and have led to weakening the public's confidence and reducing stakeholders' trust in the integrity of IT companies and their employees. As a result, several calls for reform and closer scrutiny of the ethics of IT companies and their employees are being made by many, including business practitioners and researchers [49]. So, more than ever, investigating the ethics of software engineers, who are key players of the IT industry, and the relationships it has with other variables are of significant importance in understanding what could be done to improve the situation. As discussed earlier, software engineers experience a lot of pressure in their jobs which can lead to high levels of stress. Based on AET [17] and COR theory [36], their emotions are important in dealing with this, and can impact their ethical decision-making [24] [48] [49] [50] [53]. Therefore, EQ is one of the variables that will be investigated in relation to the ethics of software engineers. Studies investigating this relationship among software engineers are scant. And previous researchers of other areas have also recommended this investigation. Holian [54] recommended that future research investigate whether ethical decisions are affected by skills associated with EQ. Also, Mulki et al. [55] proposed that future research address the impact of EQ on an individual's ethical judgment. Bay and McKeage [56] said that ''it may eventually be shown that emotional intelligence is one of the variables that may explain the current gap between ethical understanding and ethical behaviour'' [p. 441].

Therefore, in this study, the researchers will investigate whether EQ moderates the relationship between the work ethics of software engineers and their work performance. The performance of employees is crucial to the overall success of a company [57], and unethical behavior of employees would definitely negatively impact this [4]. The findings of this study are expected to give empirical evidence and valuable information on the need of both work ethics and EQ in software engineers for improving their ethical behaviour in their work.

2. RELATED WORK

2.1. Ethics in ICT

This subsection reviews studies conducted by other researchers that are related to ethics in ICT or SE.

Lurie and Mark [58] did a study in which they proposed "an ethical framework for software engineers that connects software developers' ethical responsibilities directly to their professional standards" [p. 1]. They said that the application of this ethical framework can override the traditional contrast between professional skills and ethical skills which is present in engineering professions. They say that this will improve the engineer's professionalism and ethics. They called it Ethical-Driven Software Development (EDSD). They said that this approach will be more effective than the usual approach in professional ethics that "advocates stand-alone codes of ethics" [p. 1].

Barn [59], proposed that, in light of recent scandals like the Volkswagen-emissions scandal, values should be included in software development as a non-functional requirement. They suggested that "values accompanied by an appropriate framework derived from non-functional requirements" [p. 1] can be used by software engineers as a means for discussion of ethical issues of the design of software. They did a case study on Volkswagen and a qualitative analysis of the views of software engineers from Reddit discussion forums and found that the Volkswagen case study really demonstrated the need of ethics in SE practice, while developers on Reddit were "relatively subdued" [p. 10] in their discussions of ethical issues; they did not give ethics the attention it should be given.

Wilk [60] conducted a study in which they proposed a course for undergraduates, Law for Computer Professionals. They said that it is crucial for students to have a good understanding of "ethics, law, policy, regulation, and responsible software development" [p. 94]. They described the curriculum of the course which consists of "legal aspects, ethical aspects, and professional responsibility" [p. 94]. They also discussed curricular guidelines by the ACM and IEEE. Several related issues and challenges that computer practitioners should address are also mentioned.

A study done by Aydemir and Dalpiaz [1] presented ethics-aware SE in which the expected "ethical values of stakeholders are captured, analysed and reflected in software specifications and in the SE processes" [p. 1]. They argued that SE is a field that has humans at its centre; software is produced by humans for humans. They further stated that other fields that are to do with non-software goods are taking ethical concerns very seriously, like, with regards to how products are produced, "Do subcontractors use child labour? Is fair payment guaranteed?" [p. 1], also, in the field of environment and food, "What is the carbon footprint?... Are there any ingredients that could lead to addiction?" [p. 1]. So, they want similar questions to be asked for software products. They are saying that ethics need to be considered in the decision-making process of every phase of SE. They proposed a framework that can help stakeholders in analysing ethical concerns "in terms of subject (software artefact or SE process), relevant value (diversity, privacy, autonomy etc.), and threatened object (user, developer etc.)" [p. 1]. They also introduce a roadmap that describes the required steps through which SE researchers and practitioners can fully implement ethics-aware SE.

Another study by Karim, Al Ammar and Aziz [61] did a review of work already done in SE related to practicing a code of ethics. They wanted to fill the gap in research in this area regarding the actual use of Software Engineering Code of Ethics (SWECOE) in the practice of SE. Also, they developed a framework as a means for future research and for practicing ethics in the actual phases of SE through the Software Development Life Cycle (SDLC). Their findings show that, although there are quite a lot of research on SWECOE, the reality is that, practically, a code of ethics is "difficult and complicated to implement in the actual work environment" [p. 290]. With their developed framework, they aim to make it easier for software engineers to implement ethics in their work.

Melo and Sousa [62] presented a paper that aimed to examine SE programs. They said that "millennial software engineers must be prepared to make ethical decisions, think critically, and act systematically" [p. 40] in order to produce the best software. They said that this present situation requires constant changes in education and curricula, as wrong decisions in software engineers can lead to major social impacts. So, after examining current SE programs, they proposed a conceptual framework for examining cyberethics education and a number of suggestions on how to incorporate it into SE curricula.

Kumar and Kremer-Herman [11] wrote a paper in which they presented their experiences of incorporating ethics and societal impact in computing through implementing three different courses. They said that, as there is a rising need for there to be more awareness of ethics in the development of ICTs, changes need to occur at the education level for more effective consciousness of ethics in ICT. In the three courses that they implemented, they applied innovative practices such as role playing, case studies, ethical frameworks, ethics scenario development, service learning and so on. In their results, they found that it was not hard to include ethics and societal impact in the three courses. They found that it helped students to consider how their technology would affect society, and felt more ownership of their ethical responsibility. They conclude that their approaches in these three courses are feasible and help in improving ethical awareness among students.

Wood [74] conducted a study in which they explored the relationships between ethics training and decision-making/moral reasoning of IT specialists. They said that following ethical standards is really important when using information systems to prevent negative impacts. So, they wanted to study whether ethics training can improve the ethics of IT specialists. They used two surveys to collect data from a group of IT specialists from different occupations. In their results, among other findings, they found that "there is a positive significant relationship between ethics training and ethical decision making and moral reasoning" [p. 55] and that "there is a significant positive relationship between frequency of training and ethical decision making and moral reasoning" [p. 55]. They concluded that their findings show that ethics training can improve ethical decision-making and moral reasoning of IT specialists. Therefore, ethics training should be implemented in IT/SE industries. This will help in reducing the occurrence of unethical incidents.

Xenos and Velli [63] wrote a paper in which they presented the Ethics Game, a storytelling game in which the players encounter ethical issues that are to do with software engineering. What the players choose will affect how the story will continue. They used this Ethics Game as a tool to mediate learning activity in 144 students in a software engineering course. The findings of their study show that students displayed improvement in their knowledge of software engineering ethics through playing the game. The students found the game to be “a useful educational tool and of high usability” [p. 579]. The researchers conclude that they will improve the game to include other issues of software engineering so that it can become a comprehensive and usable tool through which students can enhance their practical knowledge on ethics in software engineering.

Based on the related work reviewed above, it can be observed that there has been an increase in the amount of research done in the area of ethics in ICT or SE. Many of the studies focused on discussing ethical issues or dilemmas in ICT or SE and have proposed possible solutions. Other studies focused on introducing new ethics courses for universities, introducing new frameworks or codes of ethics, revising existing codes of ethics and so on. However, there is a necessity to investigate the ethics of software engineers themselves and the relationships it has with other interpersonal variables. This is very important for understanding the dynamics that surround the ethics of software engineers and what could be done to minimize the occurrence of unethical incidents in the IT/SE industry.

2.2. Emotional Intelligence (EQ)

This subsection reviews previous studies related to EQ that are relevant to the area of this research.

Chowdhury [51] wrote an article studying the relationships between emotional intelligence (EQ) and consumer ethics. They investigated the effects of different aspects of EQ on consumers’ ethical beliefs. 500 Australian consumers completed an online questionnaire that measured EQ, consumers’ ethical beliefs and personal moral philosophies. Their results showed that the “ability to appraise and express emotions in oneself is directly negatively related” to unethical consumer beliefs, and the “ability to appraise and express emotions in oneself is directly positively related” to ethical/doing-good beliefs, and the “ability to appraise and recognise emotions in others” is also directly positively related to ethical/doing-good beliefs. It can be observed that emotions have a significant effect on consumers’ ethical decision-making. They suggested that public policies be made that allows the EQ of consumers to be developed such as creating social and emotional learning (SEL) programmes in educational institutions as well as EQ training at the workplace.

Yadav, Dubey, and Ali [64] wrote a paper in which they investigated the role EQ and occupational stress play in moral decision-making. They said that previous research has found that stress can affect the quality of decision-making and increase unethical behavior. So, in this study they aim to explore whether EQ can reduce stress and improve ethical behavior. They collected data from 177 marketing executives using questionnaires. In their results, they found that EQ is positively related with moral decision-making, and occupational stress is inversely related to EQ and moral decision-making. They also found that occupational stress mediates the relationship between EQ and moral decision-making. They concluded that, as good EQ significantly affects moral decision-making, organizations in different sectors should administer EQ training to their employees.

Lee, Chan and Lee [65] conducted research in which they investigated the relationship between EQ and job satisfaction among IT professionals. They say that EQ has been studied extensively among other professions such as nurses, teachers, medical professionals and so on, but there are few studies on IT professionals. They further explain that it may seem that EQ is not important to IT professionals as their jobs are technical but on further inspection, it can be observed that the work they do requires a lot of social interaction such as interacting with colleagues, stakeholders, and customers. In order to communicate and complete all of their tasks, their soft skills are important. They said that, “A high level of EQ helps IT professionals to be adept in understanding their job requirements and enables them to work more effectively with people to achieve their work objectives” [p. 140]. Their main motivation of this study is to understand better the turnover of IT professionals. They said that “Turnover is endemic and problematic in the IT industry. It disrupts operations and organizations incur high costs in recruiting and training new employees” [p. 140]. They say that previous research has found that job satisfaction contributes largely to turnover. So, they intend to investigate whether EQ affects job satisfaction. In their results, they found that personal accomplishment mediates the relationship between EQ and job satisfaction. EQ is related to job satisfaction through its indirect relationship with personal accomplishment. With these findings, they suggest that measures of EQ should be used during the staff recruitment process to see who are more likely to experience personal accomplishment. This will then lead to increasing job satisfaction and reducing turnover in IT professionals. They also suggest that EQ training be given in education and in industries.

Rezvani and Khosravi [15] conducted a study in which they investigated the impact of software engineers’ EQ on their stress, trust, and performance. They explained that as software engineers work in demanding environments and face numerous difficulties in completing their projects, this increases levels of stress among them which affects “their ability to self-regulate their feelings and understanding” [p. 139] which then affects their effectiveness. So, they aim to explore whether EQ can reduce stress and instead increase trust among software engineers, which can improve their performance. In their review of the literature, they found that the number of studies focusing on software engineers’ personal skills and competencies were scant. With their study, they aim to improve this lack. In their results, they found that emotionally intelligent software engineers are “more likely to manage the negative influence of stress and are more likely to trust in other team members which result in increased performance” [p. 148]. They suggest that IT/SE organizations should provide training programs that will increase the EQ skills of software engineers. They found in their research that a pre-emptive approach to reducing stress and increasing trust among employees leads to revenue growth and productivity.

From the above studies reviewed, it can be observed that EQ has a significant positive effect on work performance, job satisfaction, and ethical behaviour of people and employees of different fields. Based on the researchers’ review of past studies, EQ of software engineers has not been investigated before in relation to their ethics, and work performance. This research intends to investigate these relationships. It is expected that the findings will give valuable information on the importance and need of EQ and ethics in software engineers for reducing unethical behaviour, and improving work performance, which are important for the success and progress of IT/SE organizations, and their positive contributions to society and the economy.

3. RESEARCH METHODOLOGY

3.1. Research Objective and Research Question

The main objective of this study is to empirically investigate how emotional intelligence (EQ) moderates the relationship between work ethics of software engineers and their work performance.

Figure 1 below models this research objective. Work Ethics is the independent variable (IV), while Work Performance is the dependent variable (DV), and Emotional Intelligence (EQ) is the moderating variable (MV).

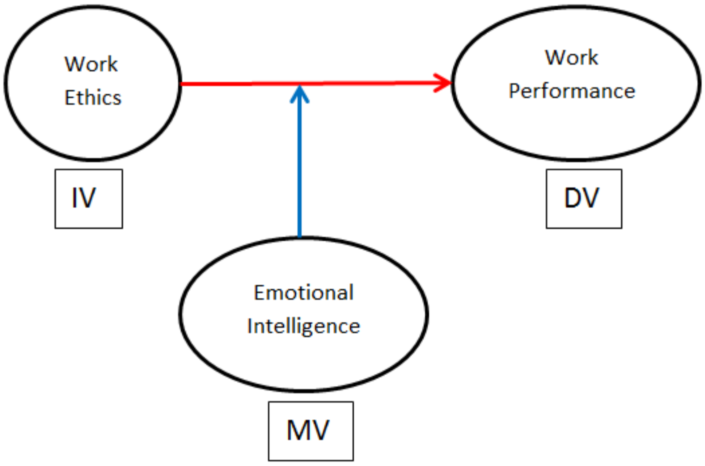


Figure 1. The research model

To address this research objective, our research question is: "*How does EQ moderate the relationship between the work ethics of software engineers and their work performance?*" To answer this research question and fulfill the research objective, hierarchical multiple regression analysis will be used [66]. According to [66], hierarchical multiple regression analysis is frequently used for moderation analysis for these types of research that involve interpersonal or psychological variables.

3.2. Sample and Data Collection

The population of this research is software engineers working in IT/SE companies. The sample was randomly taken from software engineers working in IT/SE companies via LinkedIn. The random sampling method was used in this research. More than 500 software engineers on LinkedIn were selected at random. Each of them was sent a questionnaire containing three instruments to measure three variables: Emotional Intelligence (EQ), Work Ethics, and Work Performance. A total of 170 software engineers responded and completed the questionnaire. The sample size is, therefore, n=170. For the random sampling technique, most statisticians agree that the minimum sample size required in order to get a meaningful result is 100 [67]. Therefore, a sample size of n=170 can be deemed sufficient for this study. The questionnaire was created using Google Forms. The data was collected over a period of 4 months.

3.3. Instruments

The questionnaire sent to each participant contained three instruments to measure the three variables: EQ was measured using the Wong and Law Emotional Intelligence Scale (WLEIS) [68]. It has 16 questions that measure four dimensions: Self-emotions Appraisal, Regulation of Emotions, Use of Emotion, Others-emotion Appraisal; Work Ethics of software engineers will be measured using the Islamic Work Ethics (IWE) [69] instrument. It has 23 questions that measure four dimensions: Effort, Honesty, Teamwork and Accountability (the researchers would like to highlight that these four dimensions are universal work ethics); Work Performance will be measured using the Individual Work Performance [70] instrument. This questionnaire contains 18 questions that measure three dimensions: Task Performance, Contextual Performance and Counterproductive Work Behavior. Each instrument used Likert-scale items: EQ used a 7-point Likert-scale; Work Ethics used a 10-point Likert-scale; and Work Performance used a 5-point Likert-scale.

3.4. Data Analysis

As all three variables in this research use Likert-scale items, they are considered to be ordinal variables. However, as each of the three variables in this research was calculated by summing the score of each question in the instrument together (the level of agreement chosen on the Likert scale for each question; for example: 1=Strongly Disagree, 2=Disagree and so on) to get the total score, the variables can now be treated as interval/continuous variables (as the meaning of addition relies on this property). Researchers have found that ordinal variables with five or more categories can often be used as interval/continuous variables without any harm to the analysis [71] [72] [73] [74]. As all three variables have five or more Likert-scale categories, they will therefore be treated as interval/continuous variables to conduct the data analysis of this research.

After collection of data, data analysis was conducted using IBM SPSS Statistics version 22. In order to investigate how EQ moderates the relationship between Work Ethics and Work Performance, moderation analysis was conducted using hierarchical multiple regression [75]. All the assumptions necessary for conducting hierarchical multiple regression on the data of this research were met [75]. Laerd Statistics [75] was referred to for the analysis.

4. RESULTS

In this section, the results of this study are presented. A total of 170 software engineers responded to the questionnaire. Therefore the sample size is n=170. The demographics of the sample are as follows: **gender:** 110 (64.7%) of them were male, 60 (35.3%) of them were female; **age:** 135 (79.4%) of them are below 30, 29 (17.1%) of them are between 30-35, 3 (1.8%) of them are between 36-40, 2 (1.2%) of them are between 41-45, and 1 (0.6%) of them is above 45; **education:** 11 (6.5%) had a diploma, 2 (1.2%) had an associate degree, 132 (77.6%) had a bachelor’s/undergraduate degree, 23 (13.5%) had a master’s degree, 1 (0.6%) had a doctorate degree/PhD, and 1 (0.6%) did not go to college; **years of experience working in software engineering projects:** 66 (38.8%) had less than 2 years, 57 (33.5%) had between 2-4 years, 28 (16.5%) had 5-8 years, 19 (11.2%) had more than 9 years. ‘Years of experience working in software engineering projects’ is shown below in Figure 2.

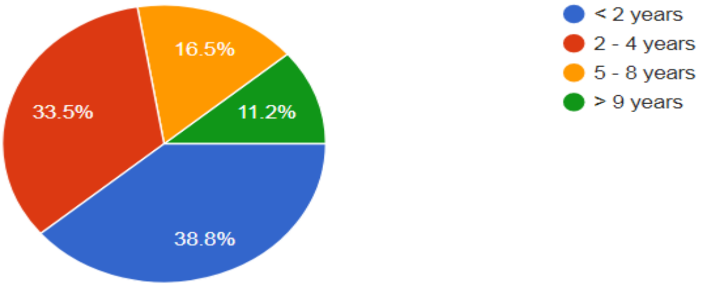
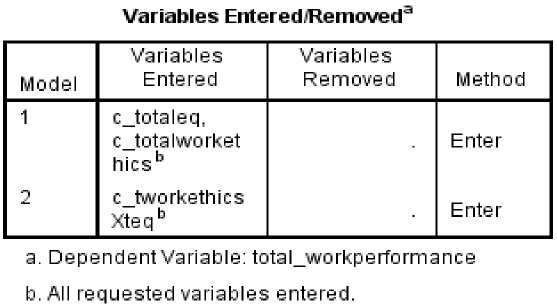


Figure 2. Percentages of years of experience working in software engineering projects of the sample under study

In this study, EQ is the moderating variable (MV), and it is also an independent variable (IV) along with Work Ethics, whereas Work Performance is the dependent variable (DV). In order to avoid multicollinearity problems (i.e. high degree of correlation between the independent variables), the IVs were mean-centred first [75]. Then the centred interaction term, EQ multiplied with Work Ethics (c_tworkethicsXeq), was calculated. And finally, the analysis was run.

There are two models in the analysis. The first model contains the IVs, c_totalworkethics (this is the mean-centred total-work-ethics) and c_totaleq (this is the mean-centred total-EQ), on their own separately. In the second model, the interaction of the IV's is added, c_tworkethicsXeq (total-work-ethics multiplied with total-EQ). Comparing the results of the first model with the results of the second model will reveal whether EQ moderates the relationship between work ethics and work performance or not. Table 1 below shows the two models. And Table 2 below shows the results:

Table 1. Variables included in Model 1 and Model 2



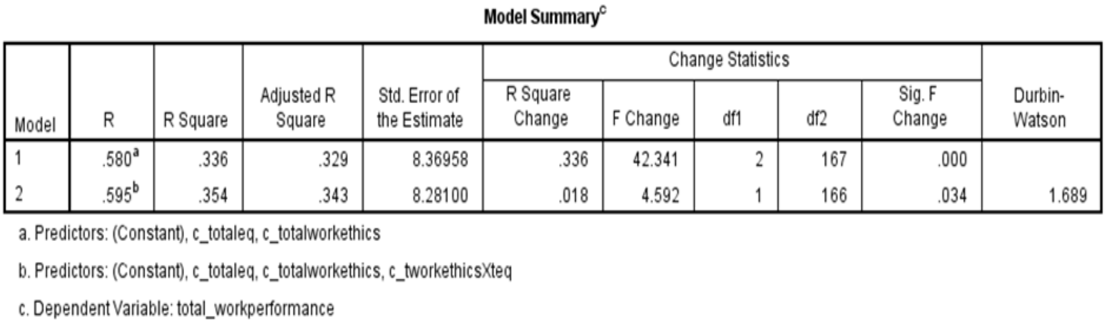
Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	c_totaleq, c_totalworkethics[b]	.	Enter
2	c_tworkethicsXteq[b]	.	Enter

a. Dependent Variable: total_workperformance

b. All requested variables entered.

Table 2. Model Summary for EQ (MV) between Work Ethics (IV) and Work Performance (DV)



Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.580[a]	.336	.329	8.36958	.336	42.341	2	167	.000	
2	.595[b]	.354	.343	8.28100	.018	4.592	1	166	.034	1.689

a. Predictors: (Constant), c_totaleq, c_totalworkethics

b. Predictors: (Constant), c_totaleq, c_totalworkethics, c_tworkethicsXteq

c. Dependent Variable: total_workperformance

A hierarchical multiple regression was run to determine if the addition of the interaction term obtained from a submaximal test improved the prediction of work performance over and above EQ and Work Ethics alone.

Model Fit:

From Table 2 above, it can be observed that Model 2 with the interaction term accounted for significantly more variance in the dependent variable, Work Performance, than just EQ and Work Ethics by themselves. R^2 changed from 0.336 to 0.354, R^2 change = 0.018, p = 0.034 (significant at the $p<0.05$ level). There is significant moderation occurring.

Therefore, the addition of the interaction term to the prediction of Work Performance (Model 2) led to a statistically significant increase in R^2 of 0.018, $F(1, 166) = 4.592$, $p < 0.05$. Therefore, the results show that EQ does significantly moderate/strengthen the relationship between Work Ethics and Work Performance.

5. Discussion and Conclusion

From the results explained earlier, it can be confirmed that EQ significantly strengthens the relationship between work ethics of software engineers and their work performance; for software engineers who had stronger EQs, their work ethics had a stronger positive relationship with their work performance. These findings have important implications for both the IT/SE industry as well as the education sector.

As discussed earlier in the introduction, unethical incidents keep occurring in the IT/SE industry, yet not enough is being done in the IT/SE industry, in education, and in research. This study contributes in the area of research. It has investigated how work ethics of software engineers (who are key players in the IT/SE industry) is related with other interpersonal variables to understand better what could be done to improve the situation. The findings provide empirical evidence that EQ can significantly strengthen the influence of work ethics of software engineers on their work performance. Therefore, the researchers strongly suggest that steps should be taken to incorporate EQ training into IT/SE industries as well as in educational curriculums. Previous studies have found that EQ training did indeed improve the trainees' EQ [43] [44]. At the same time, steps should be taken to teach and train ethics more actively in IT/SE industries and in education. Wood [76] has found that there is a significant positive relationship between "ethics training and, ethical decision making and moral reasoning" [p. 55] among IT specialists. Therefore, training can improve the EQ and ethics of software engineers.

If steps are taken to train EQ and ethics in the IT/SE industry as well as in education, then, based on the findings of this research, it will improve the ethical behavior of software engineers in their work and thus, reduce the occurrence of unethical incidents that cause great harm to society, and instead increase beneficial contributions to society.

6. Contribution and Future Work

This research is expected to increase the number of studies done on the interpersonal characteristics of software engineers and the relationships it has with ethics, and urge other researchers to do further research on this important and urgent area.

The limitation of this research is that all variables were quantitatively measured using self-reported questionnaires. Therefore, there may be bias in the data collected. Future research may try utilizing other types of research such as qualitative studies that utilize interviews, focus groups, observation, case studies and so on as methods of data collection to gather less biased data, which may bring to light additional information on the topic of this research. Longitudinal studies may also be attempted.

The researchers also recommend that further research be done on the relationships between ethics of software engineers/IT professionals and other variables such as team climate, team flexibility and so on as this will continue to increase our understanding of this important area and thus, help in curbing the rampant unethical behavior in the IT/SE world.

ACKNOWLEDGEMENTS

This study is funded by the Ministry of Higher Education Malaysia under the FRGS research grant: FRGS/1/2018/ICT01/UIAM/02/2.

REFERENCES

[1] Aydemir, F. B., & Dalpiaz, F. (2017). A Roadmap for Ethics-Aware Software Engineering. *Conference'17, July 2017, Washington, DC, USA.* ACM.

[2] Jia, J., & Xin, J. (2018). Integration of ethics issues into software engineering management education. *TURC 2018.* ACM.

[3] Radoini, A. (2020, May 11). *Cyber-crime during the COVID-19 Pandemic*. Retrieved June 11, 2020, from UNICRI: http://www.unicri.it/news/article/covid19_cyber_crime

[4] Rogerson, S., Miller, K. W., Winter, J. S., & Larson, D. (2017). Information systems ethics – challenges and opportunities. *Journal of Information, Communication and Ethics in Society*.

[5] *Facebook–Cambridge Analytica data scandal*. (2018, September 27). Retrieved September 27, 2018, from Wikipedia: https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

[6] Lord, N. (2018, April 6). *A History of Ransomware Attacks: The Biggest and Worst Ransomware Attacks of All Time*. Retrieved September 30, 2018, from Digital Guardian: https://digitalguardian.com/blog/history-ransomware-attacks-biggest-and-worst-ransomware-attacks-all-time

[7] *Ransomware*. (2018, September 21). Retrieved September 30, 2018, from Wikipedia: https://en.wikipedia.org/wiki/Ransomware

[8] Belyh, A. (2016, March 19). *Work Ethic Definition & Elements of a Strong Work Ethic*. Retrieved June 12, 2020, from Cleverism: https://www.cleverism.com/work-ethic-definition-elements-strong-work-ethic/

[9] Ferrario, M. A., Simm, W., Whittle, J., Frauenberger, C., Fitzpatrick, G., & Purghathofer, P. (2017, May 06-11). Values in Computing. Denver: ACM.

[10] Thomson, A. J., & Schmoldt, D. L. (2001). Ethics in computer software design and development. Computers and Electronics in Agriculture, 30(1), 85-102.

[11] Kumar, S. & Kremer-Herman, N. (2019). Integrating Ethics Across Computing: An Experience Report of Three Computing Courses Engaging Ethics and Societal Impact through Roleplaying, Case Studies, and Service Learning. *2019 IEEE Frontiers in Education Conference (FIE)*(pp. 1-5), Covington, KY, USA, doi: 10.1109/FIE43999.2019.9028568.

[12] Logan, P. Y., & Clarkson, A. (2005). Teaching students to hack: curriculum issues in information security. SIGCSE Bull, 37(1), 157-161.

[13] Günsel, A., & Açikgöz, A. (2013). The Effects of Team Flexibility and Emotional Intelligence on Software Development Performance. *Group Decis Negot*, 359-377.

[14] Kosti, M. V, Feldt, R., Angelis, L. (2014). Personality, emotional intelligence and work preferences in software engineering: An empirical study. *Information and Software Technology,* 56(8), 973-990. https://doi.org/10.1016/j.infsof.2014.03.004.

[15] Rezvani, A., & Khosravi, P. (2019). Emotional intelligence: The key to mitigating stress and fostering trust among software developers working on information system projects. *International Journal of Information Management*, 139-150.

[16] Robbins, S. P., & Judge, T. A. (2013). *Organizational Behavior*. Prentice Hall.

[17] Weiss, H. M., & Cropanzano, R. (1996). Affective events theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. *Research in Organizational Behavior*, Volume 18, pages 1-74.

[18] Lee, G., Xia, W. (2005). The ability of information systems development project teams to respond to business and technology changes: a study of flexibility measures. *Eur J Inf Syst* 14(1):75–92.

[19] Kelly J. R., Barsade S. G. (2001). Mood and emotions in small groups and work teams. Organ Behav Hum Decis Process 86(1):99–130.

[20] Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. Imagination, Cognition and Personality, 9, 185–211.

[21] Van Rooy, D., & Viswesvaran, D. (2004). Emotional intelligence (A meta-analytic investigation of predictive validity and nomological net). Journal of Vocational Behavior, 65(1), 71–95.

[22] Reus T. H., Liu Y. (2004). Rhyme and reason: emotional capability and the performance of knowledge-intensive work groups. Hum Perf 17(2):245–266.

[23] Gohm, C. (2003). Mood regulation and emotional intelligence: individual differences. Journal of Personality and Social Psychology, 84(3), 594–607.

[24] Goleman, D. (1995). Emotional intelligence: why it can matter more than IQ for character, health and lifelong achievement. New York: Bantam Books.

[25] Goleman, D., Boyatzis, R., & Mckee, A. (2002). Primal leadership. Boston: Harvard Business School Press.

[26] Dulewicz, C., Young, M., & Dulewicz, V. (2005). The relevance of emotional intelligence for effective leadership. Journal of General Management, 30(3), 71–86.

[27] Brackett, M. A., Rivers, S. E., & Salovey, P. (2011). Emotional intelligence: Implications for personal, social, academic, and workplace success. Social and Personality Psychology Compass, 5(1), 88–103.

[28] Druskat, V., & Wolff, S. (2001). Building the emotional intelligence of groups. Harvard Busines Review, 79(3), 81–90.

[29] Jordan, P., & Troth, A. (2004). Managing emotions during team problem solving: emotional intelligence and conflict resolution. Human Performance, 17(2), 195–218.

[30] Sy, T., Tram, S., & O'Hara, L. (2006). Relation of employee and manager emotional intelligence to job satisfaction and performance. Journal of Vocational Behavior, 68(3), 461–471.

[31] Nicholson B., & Sahay S. (2004). Embedded knowledge and offshore software development. InfOrg 14(4):329–365.

[32] Hoegl M., Parboteeah K. P. (2007). Creativity-relevant skills and the task performance of innovation teams: how teamwork maters. J Eng Technol Manag 24:148–166.

[33] Law, K. S., Wong, C. S., & Song, L. J. (2004). The construct and criterion validity of emotional intelligence and its potential utility for management studies. Journal of Applied Psychology, 89(3), 483–496.

[34] Cote, S., & Miners, C. (2006). Emotional intelligence, cognitive intelligence, and job performance. Administrative Science Quarterly, 51(1), 1–28.

[35] Holt, S., & Jones, S. (2005). Emotional intelligence and organizational performance: implications for performance consultants and educators. Performance Improvement, 44(10), 15–48.

[36] Hobfoll, S. E. (2001). The influence of culture, community, and the nested-self in the stress process: Advancing conservation of resources theory. Applied Psychology, 50(3), 337–421.

[37] Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. The American Psychologist, 44(3), 513.

[38] Gupta, M., George, J. F., & Xia, W. (2019). Relationships between IT department culture and agile software development practices: An empirical investigation. International Journal of Information Management, 44, 13–24.

[39] Hobfoll, S. E., & Shirom, A. (2001). Conservation of resources theory: Applications to stress and management in the workplace. In R. T. Golembiewski (Ed.). Handbook of organization behavior (pp. 57–81). (2nd ed.). New York: Dekker.

[40] Holmgreen, L., Tirone, V., Gerhart, J., & Hobfoll, S. E. (2017). Conservation of resources theory. The handbook of stress and health: A guide to research and practice, 443–457.

[41] Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., et al. (1998). Development and validation of a measure of emotional intelligence. Personality and individual differences, 25(2), 167–177.

[42] Wong, C. S., & Law, K. S. (2002). The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. The Leadership Quarterly, 13(3), 243–274.

[43] Grant, A. M. (2007). Enhancing coaching skills and emotional intelligence through training. *INDUSTRIAL AND COMMERCIAL TRAINING*, 257-266.

[44] Herrera, M. A., Salas, A. M., Sarmiento, L. C., Agüero, J. V., Cerdas, M. B., & Jiménez, M. T. (2019). Development of Emotional Intelligence in Computing Students: The "Experiencia 360°" Project. *2019 XLV Latin American Computing Conference (CLEI).* IEEE.

[45] Gibbs, N., Park, A., & Birnbaum, J. (1995). The EQ factor. Time, 146(4), 60–67.

[46] Maak, T., & Pless, M. (2006). Responsible leadership in a stakeholder society– a relational perspective. Journal of Business Ethics, 66(1), 99–115.

[47] Segon, M., & Booth, C. (2014). Virtue: The Missing Ethics Element in Emotional Intelligence. *J Bus Ethics*.

[48] Deshpande, S. P. (2009). A study of ethical decision-making by physicians and nurses in hospitals. Journal of Business Ethics, 90(3), 387–397.

[49] Angelidis, J., & Ibrahim, N. A. (2011). The Impact of Emotional Intelligence on the Ethical Judgment of Managers. *J Bus Ethics*, 111-119.

[50] Fu, W. (2014). The impact of emotional intelligence, organizational commitment, and job satisfaction on ethical behaviour of Chinese employees. Journal of Business Ethics, 122(1), 137–144.

[51] Chowdhury, R. M. (2015). Emotional Intelligence and Consumer Ethics: The Mediating Role of Personal Moral Philosophies. *J Bus Ethics*.

[52] Ashraf, H., Hosseinnia, M., & Domsky, J. G. (2017). EFL teachers' commitment to professional ethics and their emotional intelligence: A relationship study. *Cogent Education*.

[53] Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), Handbook of affective sciences (pp. 852–870). Cary, NC: Oxford University Press.

[54] Holian, R. (2006). Management decision making, ethical issues and emotional intelligence. Management Decision, 44(8), 1122–1138.

[55] Mulki, J., Jaramillo, J., & Locker, W. (2009). Critical role of leadership on ethical climate and salesperson behaviors. Journal of Business Ethics, 86(2), 125–141.

[56] Bay, D., & McKeage, K. (2006). Emotional intelligence in undergraduate accounting students: preliminary assessment. Accounting Education: An International Journal, 15(4), 439–454.

[57] Leonard, K. (2019, March 6). *Importance of Employee Performance in Business Organizations*. Retrieved June 13, 2020, from Chron: https://smallbusiness.chron.com/importance-employee-performance-business-organizations-1967.html#:~:text=One%20of%20the%20most%20important,and%20will%20seek%20help%20elsewhere.

[58] Lurie, Y., & Mark, S. (2015). Professional Ethics of Software Engineers: An Ethical Framework. *Science and Engineering Ethics*.

[59] Barn, B. S. (2016). Do You Own a Volkswagen? Values as Non-Functional Requirements. *HCSE + HESSD 2016*.

[60] Wilk, A. (2016). Cyber Security Education and Law. *2016 IEEE International Conference on Software Science, Technology and Engineering* (pp. 94-103). IEEE.

[61] Karim, N. A., Al Ammar, F., & Aziz, R. (2017). Ethical Software: Integrating Code of Ethics into Software Development Life Cycle. *2017 International Conference on Computer and Applications (ICCA)* (pp. 290-298). IEEE.

[62] Melo, C. d., & Sousa, T. C. (2017). Reflections on Cyberethics Education for Millennial Software Engineers. *2017 IEEE/ACM 1st International Workshop on Software Engineering Curricula for Millennials (SECM)* (pp. 40-46). IEEE.

[63] Xenos M., & Velli V. (2020). A Serious Game for Introducing Software Engineering Ethics to University Students*.* *In: Auer M., Tsiatsos T. (eds) The Challenges of the Digital Transformation in Education. ICL 2018.* Advances in Intelligent Systems and Computing, vol 916. Springer, Cham. https://doi.org/10.1007/978-3-030-11932-4_55.

[64] Yadav, S., Dubey, N., & Ali, A. A. (2015). Emotional Intelligence and Moral Decision Making: Mediating Role of Occupational Stress. *Journal of Contemporary Psychological Research*, 53-59.

[65] Lee, P., Chan, B., & Lee, J. (2017). Emotional Intelligence and Information Technology Professionals. *Proceedings of the 2017 IEEE IEEM* (pp. 140-144). IEEE.

[66] Fairchild, A. J., & McQuillin, S. D. (2010). Evaluating mediation and moderation effects in school psychology: a presentation of methods and review of current practice. *Journal of school psychology*, 48(1), 53–84. https://doi.org/10.1016/j.jsp.2009.09.001

[67] Bullen, P. B. (2020). *How to choose a sample size (for the statistically challenged)*. Retrieved October 10, 2020, from tools4dev: http://www.tools4dev.org/resources/how-to-choose-a-sample-size/

[68] Wong, C. S., & Law, K. S. (2002). The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. *The Leadership Quarterly*, 13(3), 243–274.

[69] Kamaluddin, Norlela, & Ab. Manan, Siti Khadijah (2010). The conceptual framework of Islamic work ethic (IWE*). Malaysian Accounting Review*, 9(2): 57-70.

[70] Koopmans, L. (2014). *Measuring Individual Work Performance.* CPI Koninklijke Wöhrmann, Zutphen.

[71] Johnson, D.R., & Creech, J.C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, 48, 398-407.

[72] Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), pp. 625-632. Retrieved from: https://link.springer.com/article/10.1007%2Fs10459-010-9222-y#citeas.

[73] Sullivan, G. & Artino Jr., A. R. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education*. *5*(4), pp. 541-542.

[74] Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology*, 34, 390-400.

[75] Laerd Statistics (2015). Hierarchical multiple regression using SPSS Statistics. *Statistical tutorials and software guides*. Retrieved from https://statistics.laerd.com/

[76] Wood, K. L. (2019). An Exploration into the Relationships of Ethical Decision Making and Moral Reasoning Among IT Specialists with Ethics Training, Education, and Ethical Leadership. *In Proceedings of the 2019 on Computers and People Research Conference (SIGMIS-CPR '19).* Association for Computing Machinery, New York, NY, USA, 50–56. DOI:https://doi.org/10.1145/3322385.3322398.

AUTHORS

Shafia Khatun completed her primary and secondary education in Malaysia. She went to MAZ International School for her primary education, and International Islamic School Malaysia for her secondary education. She later attended International Islamic University Malaysia (IIUM) to complete her Bachelor's degree in Computer Science. She worked in a software company, Madcat World Sdn. Bhd., for half a year. She is currently doing her Master's degree in Computer Science at IIUM. This research paper presents part of the research conducted for her Master's degree.

Dr. Norsaremah Salleh is an Associate Professor at the Department of Computer Science, International Islamic University Malaysia (IIUM) and the Director of Information Technology Division at IIUM. Her research interests include the areas of empirical software engineering (SE), evidence based SE, human and social aspects of SE and Computer Science/ SE education. She received her PhD in Computer Science from the University of Auckland, New Zealand. Contact her at norsaremah@iium.edu.my

UNIQUE SOFTWARE ENGINEERING TECHNIQUES: PANACEA FOR THREAT COMPLEXITIES IN SECURE MULTIPARTY COMPUTATION (MPC) WITH BIG DATA

Uchechukwu Emejeamara[1], Udochukwu Nwoduh[2] and Andrew Madu[2]

[1]IEEE Computer Society, Connecticut Section, USA
[2]Department of Computer Science, Federal Polytechnic Nekede, Nigeria.

ABSTRACT

Most large corporations with big data have adopted more privacy measures in handling their sensitive/private data and as a result, employing the use of analytic tools to run across multiple sources has become ineffective. Joint computation across multiple parties is allowed through the use of secure multi-party computations (MPC). The practicality of MPC is impaired when dealing with large datasets as more of its algorithms are poorly scaled with data sizes. Despite its limitations, MPC continues to attract increasing attention from industry players who have viewed it as a better approach to exploiting big data. Secure MPC is however, faced with complexities that most times overwhelm its handlers, so the need for special software engineering techniques for resolving these threat complexities. This research presents cryptographic data security measures, garbed circuits protocol, optimizing circuits, and protocol execution techniques as some of the special techniques for resolving threat complexities associated with MPC's. Honest majority, asymmetric trust, covert security, and trading off leakage are some of the experimental outcomes of implementing these special techniques. This paper also reveals that an essential approach in developing suitable mitigation strategies is having knowledge of the adversary type.

KEYWORDS

Cryptographic Data Security, Garbed Circuits, Optimizing Circuits, Protocol Execution, Honest Majority, Asymmetric Trust, Covert Security, Trading Off Leakage.

1. INTRODUCTION

In recent years, the issue of data security has continued to attract global attention. More people are becoming informed of the significance of protecting their privacy. In fact, the cases of Google and Facebook which have been under scrutiny illustrate the trend towards enhancing confidentiality and privacy in society. Software engineering techniques are multiple ways of approaching software development and delivery [8]. A threat, in software engineering, is an application or malicious code that can cause damage to the computer or steal personal data [2]. The need for security measures to safeguard data has been necessitated by the increase in the use of technology in the public and private sectors [7]. Moreover, concerns have been raised by both private and public sectors on the data mined by data mining tools. The application of these data mining tools has conflicted with the privacy policies of individuals.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 169-176, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101413

The garbed circuit is one of the most secure multi-party computation techniques that can be used in securing each party's contribution in instances when two or more parties need to compute a given common result. Trusted execution environments offer data and code-based hardware-implemented seclusion. The seclusion process makes these environments trusted candidates and this makes the secure multi-party tractable. These users can execute their contributions privately and only reveal their output. Overall, unique software engineering techniques offer solutions for the threat complexities in secure multi-party protocol computation with big data.

Private inputs from different parties that do not trust each other have facilitated the development of multi-party computation. Threat complexities propagate due to less on no knowledge concerning the multi-party computations. This paper also has addressed approaches used to mitigate these issues, including adaptive adversaries and static exploration. In multiparty protocol computation, recovering the reuse of a once-corrupted component can act as a solution to adaptive adversaries. Thus, despite some implications, the exploitation of big data justifies the implementation of multi-party protocol computation.

2. APPROACHES TO SECURE COMPUTATION

The three main approaches used in secure computation include homomorphic encryption, secret sharing, and Yao's garbled circuit [17].

2.1. MPC Based On Secret Sharing

Secret sharing enables the computation of information privately. In this case, a secret scheme is relied upon by the involved parties in carrying out the computations [17]. This means that no information relating to the data is revealed as the private data is provided in random values. The advantage of this technique is that it does not require any encryption key and allows information-theoretical security [18]. However, it requires continuous communication among the involved parties.

2.2. MPC Based On Homomorphic Encryption

Homomorphic encryption provides an approach to manipulate data while maintaining confidentiality. Using this approach, parties rely on a homomorphic encryption scheme like the Paillier scheme in the encryption of data [17]. Using the encrypted data, the parties are able to perform computations. Over the last years, the technique has attracted attention owing to its ability to preserve privacy. The technique is pegged on the complexity of the problem at hand [15]. While this feature can be viewed as strength, it can also be regarded as a drawback as it leads to difficulty in dealing with complex problems.

2.3. MPC Based On Yao's Garbled Circuits

Yao's garbled circuit a form of a function that allows communication between two or more parties without infringing on their privacy [16]. When using this approach, one party can encrypt data (input) and carry out computations. The resultant output (computed input) is converted into a circuit and presented in the form of a binary gate. The encrypted input is then sent to the other party in form of a circuit and by evaluating this input; the other party is able to decipher it to generate output – by comparing the bits and combining the results [15]. The approach is regarded as the most efficient since it does not call for continuous communication between the parties involved.

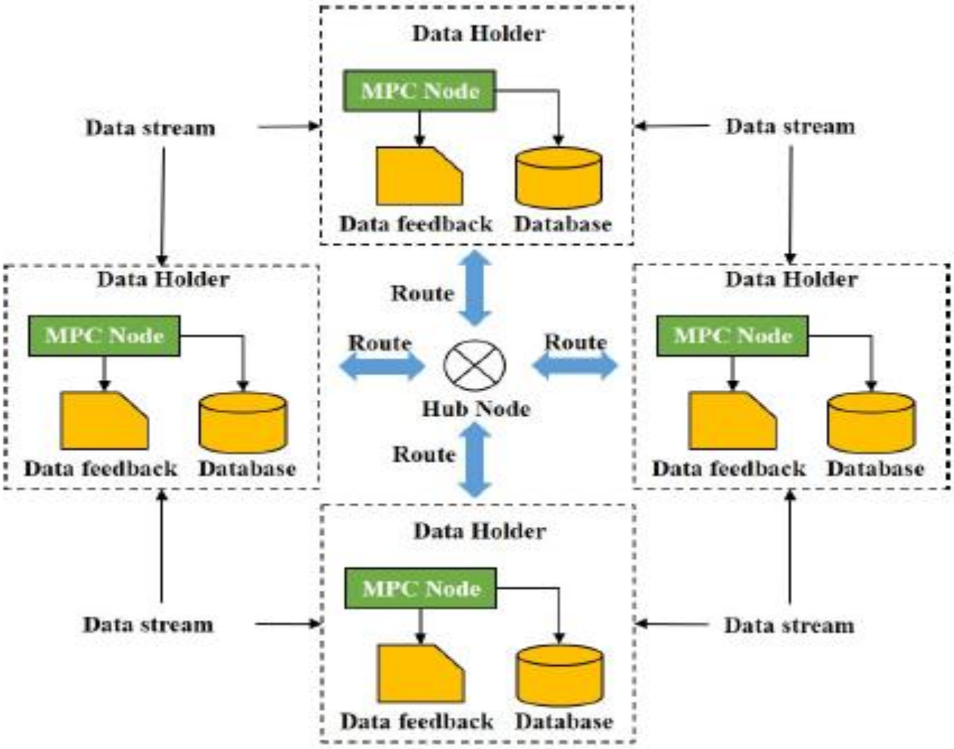


Figure 1. MPC technical framework [16]

Figure 1 above shows the MPC technical framework. The figure illustrates the transmission of data among parties and the specific areas that need to be secured to ascertain the privacy and confidentiality of information. This is where the three approaches of MPC play a key role. A hub node transmits the network and signals control when the MPC computing task is launched. Each of the data holders can commence a collaborative computing task. For secure collaborative computing, addresses are routed through hub nodes and the remaining data holders (of similar data types) are selected for computation. MPC nodes of multiple data holders participate in collaborative computing, query the required data from the databases based on the calculation logic. On the basis of ensuring privacy, all the involved parties get the correct feedback and no data is leaked to any of the participants [16].

3. SECURE MULTIPARTY COMPUTATION (MPC) WITH BIG DATA

In modern business processes, big data analysis has become a significant development that has helped in generating accurate results and accessing private data from different sources. A lot of companies have adopted more privacy measures on their own private data and as a result, employing the use of analytic tools to run across multiple sources has become ineffective. Big data is a field that deals with ways of analyzing complex or too large sets that cannot be handled by traditional data-processing techniques [10]. Joint computation across multiple parties is allowed through the use of secure multi-party computations. This understanding enables the communication of parties without them having to disclose or reveal their private data input.

Implementation of the multi-party protocol computations are only done on larger workflows [9]. There are several challenges that the implementation of the multi-party protocol computation faces. These challenges include poor integration of data processing systems with analytics and multi-party computation [1], requiring significant expert knowledge to be able to run analytics in the multi-party computation framework [6], the incapability of the multi-party computation to support data-parallel process outside the multi-party computations, and poor scaling of frameworks to large data sets. Therefore, the viability of multi-party computation implementation can be established through addressing the challenges, application, and adversaries of the multi-party computations.

Different data owners can be united through secure multi-party in function computation that depends on their data even if they might be having trust issues with each other and this is the primary goal of secure multi-party computation [3]. In multi-party computation, all the participants involved are data input owners. The essential roles of the multi-party computation include the IP (which belongs to the input parties), and they are responsible for sending data to the private computation, and result parties (RP) who receives the result from private computation and the computation parties (CP) whose responsibility is to carry out joint private computations [11]. The most common protocol of the multi-party computation is the lack of a single trust point and it involves many organizations and persons. Due to this, access to encrypted data is prohibited to all the computing parties, and no party can access the data.

There are many cases where MPC can prove worthwhile. However, for the purpose of this paper, two examples are used to show the application of MPC on big data.

3.1. Credit Card Regulation

The financial industry is one of the biggest beneficiaries of big data. Collaboration between the regulators and the industry players can enhance the efficiency of the sector by relying on data sharing and analysis. A government regulator overseeing the consumer credit reporting may wish to estimate the credit score of the consumers based on their geographic region. In this case, the government has social security numbers and the ZIP codes for all the citizens. However, the credit reference organizations have the SSNs of the credit cardholders, credit lines, and credit ratings. As required by law, the government cannot share private information with other parties [17], [15]. Similarly, credit organizations cannot share customer data with external parties. Thus, in this case, MPC is needed to facilitate data analysis and decision making.

3.2. Market Concentration

The law requires the government to regulate the market and avoid monopolies or oligopolies. In most cases, regulators use the Herfindahl-Hirschman Index (HHI) – which is based on the sum of the squared market shares of organizations in the active market. Based on the analysis, a government decides whether scrutiny is required. While public revenue data is easy to obtain, privately-held information is not easily accessed. In this regard, to effectively carry out the analysis, it is paramount to use MPC owing to the presence of private data in various agencies [17], [15]. Moreover, MPC makes it possible to filter and aggregate millions of records that organizations keep confidential.

3.3. Security Guarantees

The use of MPC ascertains the privacy of computed input and intermediate data. The way it works is that MPC does not reveal what is flagged as sensitive data. Moreover, it tends to ensure that the correctness of the output attained is within the standards accepted by all the parties [15].
The output and input processing need to adhere to the set speeds and credibility standards. However, it is required that all the participants must adhere to the regulations – they must be honest in their dealings to ensure that the privacy and confidentiality of the shared data are

maintained at all times. The generally accepted standard is that all the parties must exhibit uttermost honesty in their engagement [15].

4. Managing Threat Complexities in Secure Multiparty Computation (MPC) with Big Data

To enhance security, there are several properties that the multi-party computation employs, and they also help to enhance the efficiency and robustness of the system [4]. In the multi-party system, the most common protocol is that there is no single point of trust. The other most important protocols are n, f, passive security, abort active security and fault tolerance active security. n represents the number of computing parties involved, *f represents* the maximum number of computing parties allowed to regulate and run the protocol intended in which f+1 will be a violation of the system, passive security provides a guarantee to the privacy of source data such as the number of computation parties involved, *abort active security* ensures that corrupt computational parties run the purported protocol and *faulty tolerance active security* has the role of ensuring continuous operation of the system even in instances when the computational parties have ceased to operate correctly [5]. Figure 2, shows the structure of a secure MPC and how the various parties involved compute a function using their inputs, while keeping these inputs private [14].

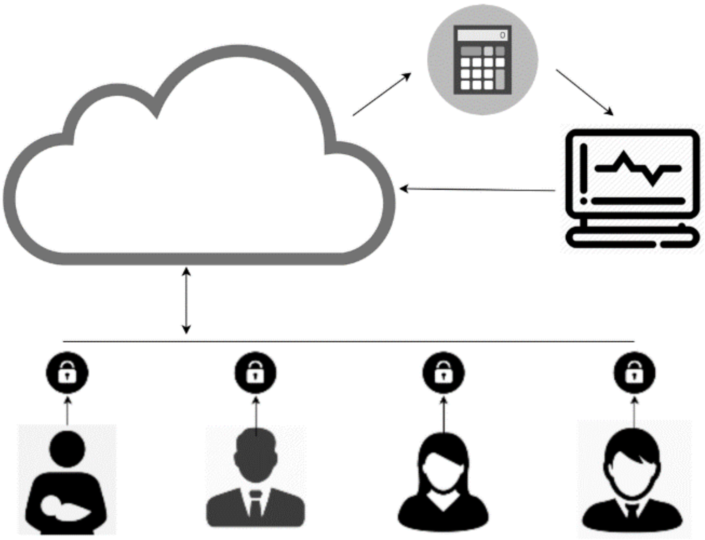


Figure 2. Secure Multiparty Computation [14].

Many trusted entities can be used to create a neural multi-party computation since it does not depend on the organization or individual’s trustworthiness. The other solution that can result in the creation of a more reliable multi-party computation is the combination of entries with conflicting interests and trustworthiness entities. In multi-party computation, the security of the system against threats is a vital consideration. A security attack on multi-party computation can result in many adversaries being attacked since the system deals with big and private data computations. The adversaries that can be attacked can be classified into malicious, covert, and semi-honest since they involve encrypted inputs. Computational security and statistical security are other mitigation techniques that can be used to avoid multi-party computation complexities.

4.1. Cryptographic Data Security

Most large corporations have employed the use of traditional methods of cryptography in their data security. In large corporations, a lot of cryptographic tasks that are diverse in nature and a given number of keys are assigned to individuals to manage them. In terms of control, storage, and usage of these keys, it is carried out under different circumstances in what is referred to as hardware security modules [12]. Companies employ the use of dedicated hardware security modules that provide cryptographic operations across the entire corporation. Exportation or incorporation into the hardware security is the most operations that are done to cryptographic keys. Techniques referred to as key-wrap are used to develop these keys. After the keys have been generated by the hardware security module disintegration to the design is not possible. To maintain the security of the keys, a lock is put on them by the hardware security module and they are safeguarded by key-wrap technique. A call to several cryptographic operations can be made due to the availability of embedded keys on the hardware security module. The hardware security module also backs up the standard cryptographic API. Figure 3 is a cryptographic data security protection platform that features multiple data security products that can be deployed individually or in combination to deliver advanced encryption, tokenization, and centralized key management [13].

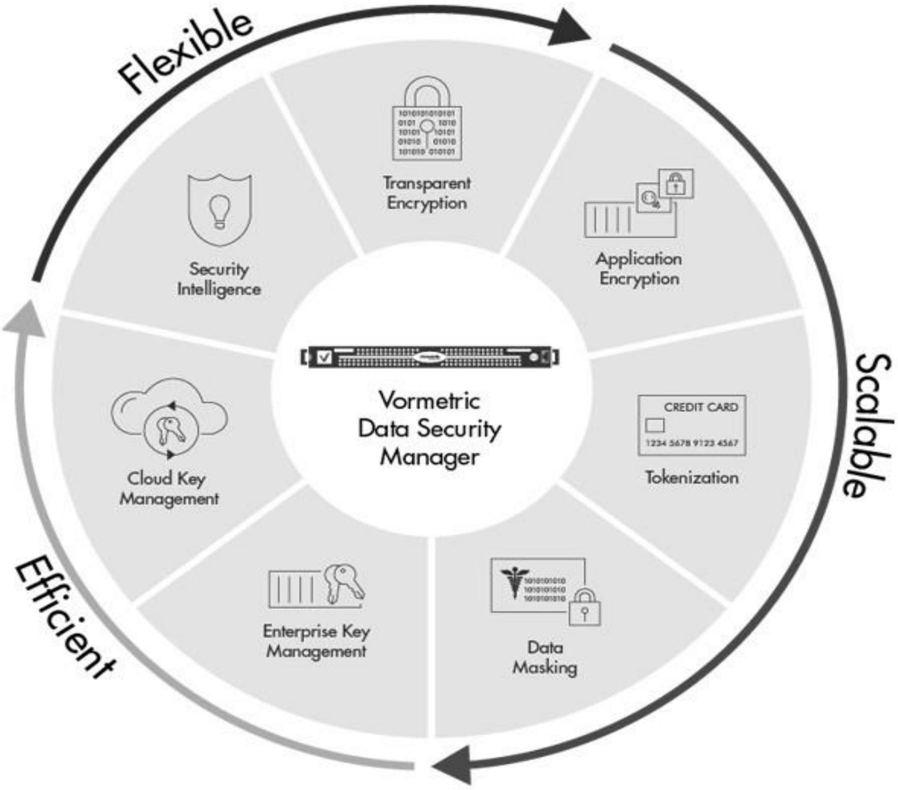


Figure 3. Vormetric Data Security Platform [13].

4.2. Implementing Other Special Software Engineering Technique

Some software engineering techniques are viably utilized in resolving threat complexities in multi-party communications. One of the techniques that can be used is less expensive garbling. Garbed circuits protocol is a special technique used, in which the main cost of executing the circuits is on the computation that is required to evaluate and generate garbed tables and the bandwidth that is required in the transmission of the garbed gates. Traditional garbling methods have been improved in terms of the bandwidth and the computation to evaluate and generate the garbled gates.

Another technique is optimizing circuits which involves reducing the size of the circuit to make a direct impact on the protocol cost. Protocol execution is another implementation technique used which addresses various improvements on the way multi-party communication protocols are executed. In this technique, the scaling issues are addressed and eliminated. Lastly, many programming tools are used in the implementation of these techniques in handling threat complexities in multi-party computations. These tools vary in different ways such as their input languages, the protocol they support, and how they combine input programs into circuits.

4.3. Research Results

Several outcomes exist from the implementation of software techniques in resolving threat complexities multi-party computation. Honest majority is one of the experimental results in which the adversary is likely to corrupt less than two of all parties involved since functions in the system have information-theoretically secure protocols. Asymmetric trust is another experimental outcome that may be central to the standard assumption of all parties being equally distrusting. The other experimental result is covert security which is reasonable in many settings but may be insufficient for some applications. A party can deviate from the protocol and be caught in a fixed probability which is referred to as public verifiable covert. Finally, trading off leakage is seen as an experimental result from the implementation of the multi-party computation. The use of these special software engineering techniques in resolving threat complexities in secure multi-party computation provides very strong security guarantees at a given cost and these security computations may make it difficult for hackers to gain access to a single bit of data.

5. CONCLUSION

In this technology-driven world, big data computations, and business analytics have attracted many adversaries, which tends to access business entities and private data. Due to this trend, developing and implementing special software engineering techniques are very essential in resolving threat complexities encountered in multi-party computation with big data. To ensure high-security levels of multi-party computations, regulations and protocols should be implemented properly. The other essential approach in developing suitable mitigation strategies is having knowledge of the adversary type. This research has clearly shown that some expertise and experience are required to adhere to specific protocols, and by implementing all these techniques, resolving threat complexities in multi-party computations with big data will be feasible and super-efficient.

REFERENCES

[1] Alam, K. S., Xiao, D., Akter, M. P., Zhang, D., Fletcher, J., & Rahman, M. F. (2018). Modified MPC with extended VVs for grid-connected rectifier. *IET Power Electronics*, *11*(12), 1926-1936.

[2] Ansari, M. T. J., Pandey, D., & Alenezi, M. (2018). Store: Security threat-oriented requirements engineering methodology. *Journal of King Saud University-Computer and Information Sciences*.

[3] Archer, D. W., Bogdanov, D., Lindell, Y., Kamm, L., Nielsen, K., Pagter, J. I., ... & Wright, R. N. (2018). From keys to databases—real-world applications of secure multi-party computation. *The Computer Journal*, *61*(12), 1749-1771.

[4] Evans, D., Kolesnikov, V., & Rosulek, M. (2017). A pragmatic introduction to secure multi-party computation. *Foundations and Trends® in Privacy and Security*, *2*(2-3).

[5] Hastings, M., Hemenway, B., Noble, D., & Zdancewic, S. (2019, May). Sok: General purpose compilers for secure multi-party computation. In *2019 IEEE Symposium on Security and Privacy (SP)* (pp. 1220-1237). IEEE.

[6] Houska, B., & Villanueva, M. E. (2019). Robust optimization for MPC. In *Handbook of Model Predictive Control* (pp. 413-443). Birkhäuser, Cham.

[7] Lykou, G., Anagnostopoulou, A., & Gritzalis, D. (2018, June). Implementing cyber-security measures in airports to improve cyber-resilience. In *2018 Global Internet of Things Summit (GIoTS)* (pp. 1-6). IEEE.

[8] Mezhuyev, V., Al-Emran, M., Ismail, M. A., Benedicenti, L., & Chandran, D. A. (2019). The acceptance of search-based software engineering techniques: An empirical evaluation using the technology acceptance model. *IEEE Access*, *7*, 101073-101085.

[9] Montazeri-Gh, M., Rasti, A., Jafari, A., & Ehteshami, M. (2019). Design and implementation of MPC for turbofan engine control system. *Aerospace Science and Technology*, *92,* 99-113.

[10] Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, *30*(4), 431-448.

[11] Tso, R., Alelaiwi, A., Rahman, S. M. M., Wu, M. E., & Hossain, M. S. (2017). Privacy-preserving data communication through secure multi-party computation in healthcare sensor cloud. *Journal of Signal Processing Systems*, *89*(1), 51-59.

[12] Yellepeddy, K. K., Peck, J. T., Hazlewood, K. M., & Morganti, J. A. (2017). *U.S. Patent No. 9,794,063*. Washington, DC: U.S. Patent and Trademark Office.

[13] “Vormetric Data Security Platform,” *Thales*. [Online]. Available: https://cpl.thalesgroup.com/encryption/vormetric-data-security-platform. [Accessed: 03-Jul-2020].

[14] Originally published by Shaan Ray on, “What is Secure Multi Party Computation?,” 09-Jun-2020. [Online]. Available: https://hackernoon.com/what-is-secure-multi-party-computation-232caef900b9.

[15] Volgushev N., Schwarzkopf M. and Getchell B., "Conclave: secure multi-party computation on big data: Extended Technical Report", Conference Paper, 2018. [Accessed 28 January 2020].

[16] Yan S., Liu C., Wang M., Ma P., and Wei K., "Promoting Data Circulation by Secure Multi-party Computation and Blockchain", DEStech Transactions on Computer Science and Engineering, no., 2018. Available: 10.12783/dtcse/ceic2018/24542.

[17] Raeini M. G. and Nojoumian M., "Privacy-Preserving Big Data Analytics: From Theory to Practice", Security, Privacy, and Anonymity in Computation, Communication, and Storage, pp. 45-59, 2019. Available: 10.1007/978-3-030-24900-7_4 [Accessed 28 January 2020].

[18] Ankele R., Kucuk K., Martin A., Simpson A., and Paverd A., "Applying the Trustworthy Remote Entity to Privacy-Preserving Multiparty Computation: Requirements and Criteria for Large-Scale Applications", 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016. Available: 10.1109/uic-atc-scalcom-cbdcom-iop-smartworld.2016.0077 [Accessed 28 January 2020].

NETWORK DEFENSE IN AN END-TO-END PARADIGM

William R. Simpson and Kevin E. Foltz

The Institute for Defense Analyses (IDA), Alexandria, Virginia, USA

ABSTRACT

Network defense implies a comprehensive set of software tools to preclude malicious entities from conducting nefarious activities. For most enterprises at this time, that defense builds upon a clear concept of the fortress approach. Many of the requirements are based on inspection and reporting prior to delivery of the communication to the intended target. These inspections require decryption of packets when encrypted. This decryption implies that the defensive suite has access to the private keys of the servers that are the target of communication. This is in contrast to an end-to-end paradigm where known good entities can communicate directly with each other. In an end-to-end paradigm, maintaining confidentiality through unbroken end-to-end encryption, the private key resides only with the holder-of-key in the communication and on a distributed computation of inspection and reporting. This paper examines a formulation that is pertinent to the Enterprise Level Security (ELS) framework.

KEYWORDS

Appliance, end-to-end security model, ELS, network defenses, web server handlers

1. INTRODUCTION

The Enterprise Level Security (ELS) framework has evolved from a fortress approach, in which the assumption that the threat is stopped at the front door, to a distributed security system that eliminates or mitigates many of the primary vulnerability points inherent with that system, as shown in Figure 1.The basic process of identification involves a two-way contract between two entities that are initiating a communication. Each entity needs to have some assurance that the party they are engaged with is a known entity and, specifically, the one to whom the communication should be allowed. The presentation of claims by each party that can be verified and validated accomplishes this. These claims are often in the form of credentials. [1] provides an extensive description of these processes.

Entities may be active or passive. Passive entities include storage elements, routers, wireless access points, some firewalls, and other entities that do not themselves initiate or respond to web service or web application requests. Active entities are those entities that request or provide services according to ELS. Active entities include users, applications, and services. All active entities have PKI certificates, and their private keys are stored in tamper-proof, threat-mitigating storage. Communication between active entities requires bi-lateral, PKI, end-to-end authentication. A verifiable identity claims-based process provides authentication.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 177-187, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101414

2. LITERATURE REVIEW FOR CURRENT DEFENSE PACKAGES

The elements involved in implementing network and application defense are numerous and complicated. Functionality is provided by a wide ranges of appliances (and by other means).This functionality may be for quality of service to the user or quality of protection to network resources and servers. These appliances are often placed in-line, and some require access to content to provide their service. Figure 2 provides a representation of how these appliances come between the user and the application.

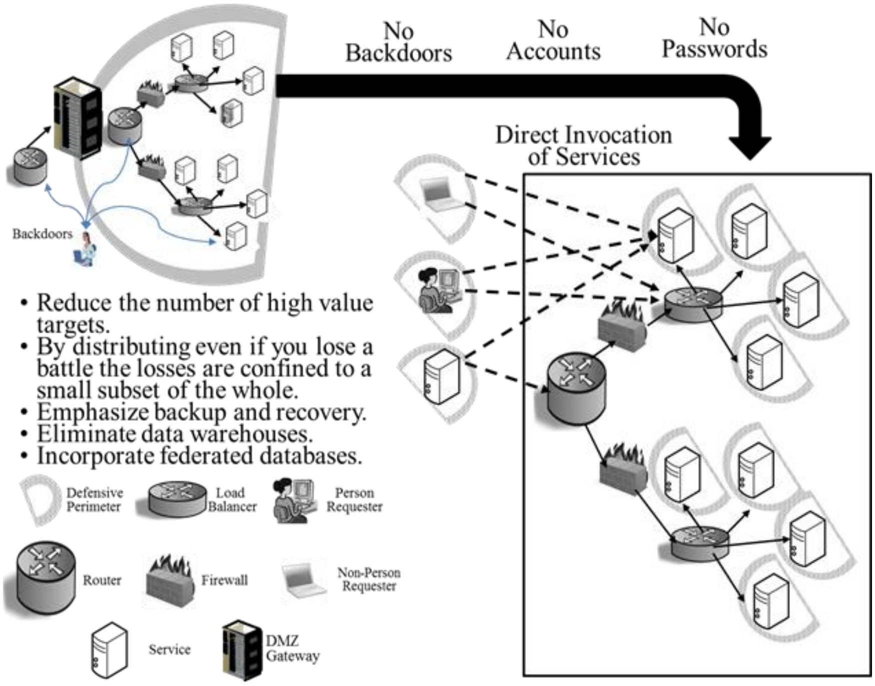

No Backdoors
No Accounts
No Passwords
Direct Invocation of Services
Backdoors
• Reduce the number of high value targets.
• By distributing even if you lose a battle the losses are confined to a small subset of the whole.
• Emphasize backup and recovery.
• Eliminate data warehouses.
• Incorporate federated databases.
Defensive Perimeter
Load Balancer
Person Requester
Router
Firewall
Non-Person Requester
Service
DMZ Gateway

Figure 1 Distributed Security Architecture

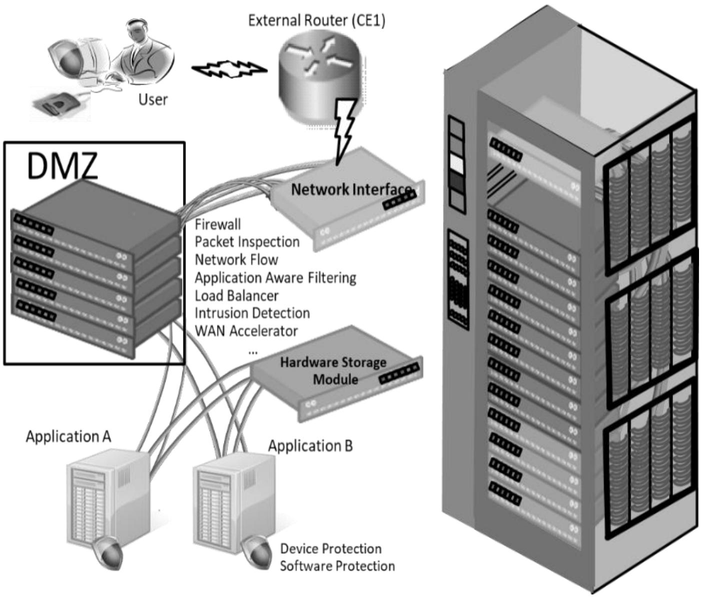


Figure 2 End-Point Access

The number and types of appliances can be quite large. Below is a partial list of functional types as provided in the current literature:

- Header-based scanner/logger [2]
  - Views only unencrypted portion of traffic
  - Synchronous or asynchronous operation
  - Scans for defined behavior, logs traffic
- Content-based scanner/logger [3]
  - Views decrypted transport layer security (TLS) content
  - Synchronous or asynchronous operation
  - Scans for defined behavior, and logs traffic/content
- Header-based firewall [4]
  - Views only unencrypted portion of traffic
  - Synchronous operation
  - Scans for and blocks defined behavior
- Content-based firewall – block only [5]
  - Views decrypted TLS content
  - Synchronous operation
  - Scans for defined behavior and blocks (terminates) connection
- Content-based firewall – modify malicious content [6]
  - Views decrypted TLS content
  - Synchronous operation
  - Scans for defined content, and blocks connection or removes content without blocking the connection
- Web accelerator [7]
  - Views decrypted TLS content
  - Synchronous operation
  - Modifies content for performance
- WAN accelerator [8]
  - Views decrypted TLS content
  - Multi-party system
  - Synchronous operation
  - Modifies content representation between parties, but no end-to-end modification
- Load Balancers [9]
  - Distributes load among destination end-points to improve throughput and reduce latency
  - May decrypt content:
    - May combine encrypted flows through a “secure sockets layer (SSL) accelerator”
    - May distribute content by request to different servers based on load
    - These load balancers are active entities
  - May not decrypt content:
    - Using “sticky” or end-point balances may route all requests from an entity to the same server
    - These load balancers are passive entities

3. Shortcomings of the Current Approaches

Each of the appliances above offers some functionality and increases the threat exposure. None of these are free from vulnerabilities from a security standpoint, and they do increase the threat surface and the vulnerability space. For example, default passwords or other improperly secured access methods allow an attacker access to any data that the appliance can access. For detailed scans, this could include all decrypted network traffic to and from a server. With a large number of independent appliances, this represents a significant security risk. Use of any appliance must be balanced by the increased functionality and the increased vulnerability. The situation is further complicated by vendor offerings of load balancers with firewall capability, "smart" accelerators that scan content, and software-only offerings that will provide most of these functionalities in a modular fashion.

This work is part of a larger body of work termed "Consolidate Enterprise IT Baseline (CEITB)."In this paper, we review the communication models for current network defenses. We then review the inspection processes and its basic architecture. Next, we show how network inspections and reporting are available while maintaining end-to-end communications. Finally, we provide the unique factors that arise with end-to-end approaches and network defenses.

4. THE REAL DE-MILITARIZED ZONE (DMZ)

Figure 3 provides a real-world defense package. Although it may look like a network defense package you have seen, it is not and it is only for illustration purposes. The first thing you see is that it is very complex and has many elements requiring proper configuration to function correctly. In reality, it occupies several racks of equipment. Secondly, the first stop after initial entry from the external router is a load balancer that will decrypt the encrypted packets. This is accomplished by either providing the private keys of all servers or allowing the load balancers (LB1 or LB2) to access the hardware storage module (HSM) of the server as if it were the server. Both break the end-to-end paradigm. Additionally, in most instances, forwarded packets are unencrypted as the appliances are assumed trusted. Each appliance has its own set of vulnerabilities, and this complicates the network defense appreciably.

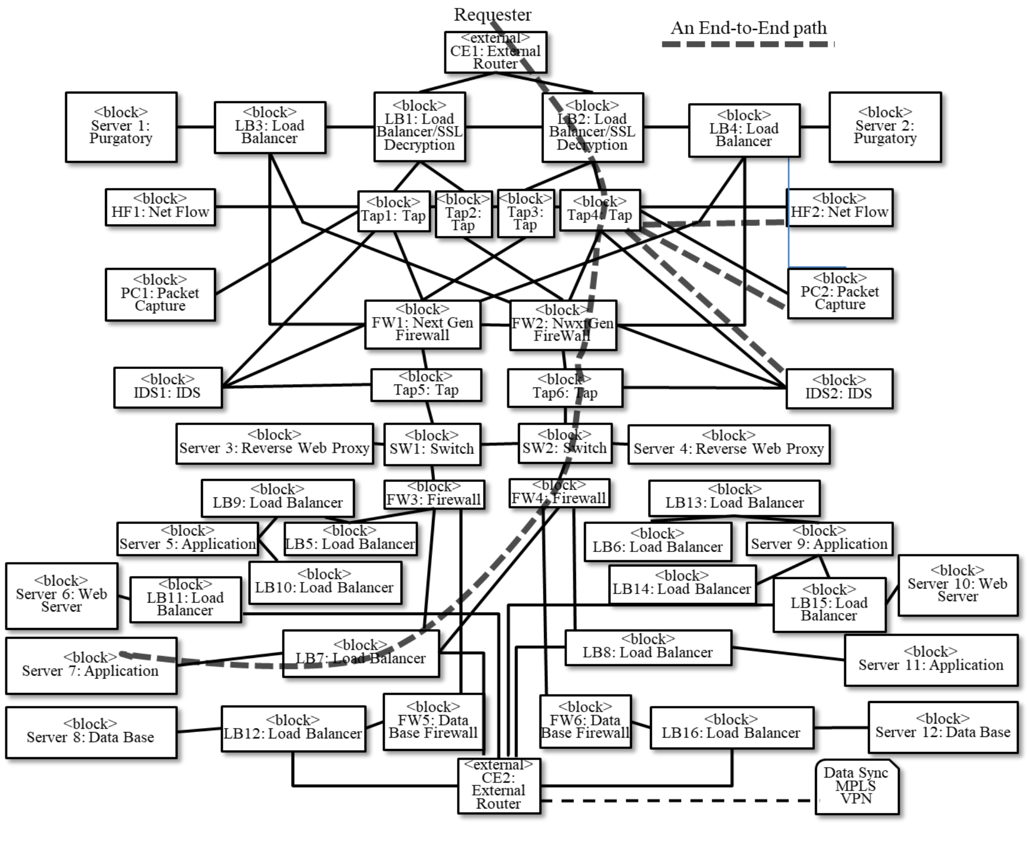

Requester
An End-to-End path
<external> CE1: External Router
<block> Server 1: Purgatory
<block> LB3: Load Balancer
<block> LB1: Load Balancer/SSL Decryption
<block> LB2: Load Balancer/SSL Decryption
<block> LB4: Load Balancer
<block> Server 2: Purgatory
<block> HF1: Net Flow
<block> Tap1: Tap
<block> Tap2: Tap
<block> Tap3: Tap
<block> Tap4: Tap
<block> HF2: Net Flow
<block> PC1: Packet Capture
<block> PC2: Packet Capture
<block> FW1: Next Gen Firewall
<block> FW2: Nwxt Gen FireWall
<block> IDS1: IDS
<block> Tap5: Tap
<block> Tap6: Tap
<block> IDS2: IDS
<block> Server 3: Reverse Web Proxy
<block> SW1: Switch
<block> SW2: Switch
<block> Server 4: Reverse Web Proxy
<block> LB9: Load Balancer
<block> FW3: Firewall
<block> FW4: Firewall
<block> LB13: Load Balancer
<block> Server 5: Application
<block> LB5: Load Balancer
<block> LB6: Load Balancer
<block> Server 9: Application
<block> Server 6: Web Server
<block> LB11: Load Balancer
<block> LB10: Load Balancer
<block> LB14: Load Balancer
<block> LB15: Load Balancer
<block> Server 10: Web Server
<block> Server 7: Application
<block> LB7: Load Balancer
<block> LB8: Load Balancer
<block> Server 11: Application
<block> Server 8: Data Base
<block> LB12: Load Balancer
<block> FW5: Data Base Firewall
<block> FW6: Data Base Firewall
<block> LB16: Load Balancer
<block> Server 12: Data Base
<external> CE2: External Router
Data Sync MPLS VPN

Figure 2 A Real DMZ

5. A NEW APPROACH –CREATING THE PSEUDO-APPLIANCE

The main contribution and unique approach to network defense in a distributed system is in maintaining the inspection process without breaking the end-to-end encryption of communications. The pseudo-appliance captures all of the inspection processes and places them into one software process that resides in the application for processing. This is the first step in realigning the priorities between the current approach and the end-to-end approach, as shown in Figure 4.The path from the user to the application in the top half of the figure shows the processes needed for inspection. Note that the private key for server 7 has been hand passed to the initial load balancer so that the exchange of information is visible. Next, the load balancer decrypts packets for inspection. This includes not only the inspection, but also the necessary reporting.

In the second half of Figure 4, we show the user directly communicating with the load balancer in front of the application (which now contains the inspection process).We have reduced the bandwidth necessary to handle the traffic at the network interface and distributed the computing burden. Tagging the communications between the requester and provider bypasses the DMZ stack. The initial handshake (which is unencrypted) includes the exchange of two white-listed PKI certificates. This exchange in ELS is the bi-lateral authentication of entities and is the initial setup for TLS encryption of all communications. This exchange allows for this tagging. As the decryption now occurs in server 7 prior to inspection, key passing is no longer required, and the end-to-end confidentiality is maintained. Untagged traffic will go through the normal DMZ processing. The reduction in traffic bandwidth at the front door may reduce the need for several of the downstream load balancers. Figure 5 shows the handler makeup in the server.

ELS enhances protection of the application server and provides additional security protections as discussed in the following section.

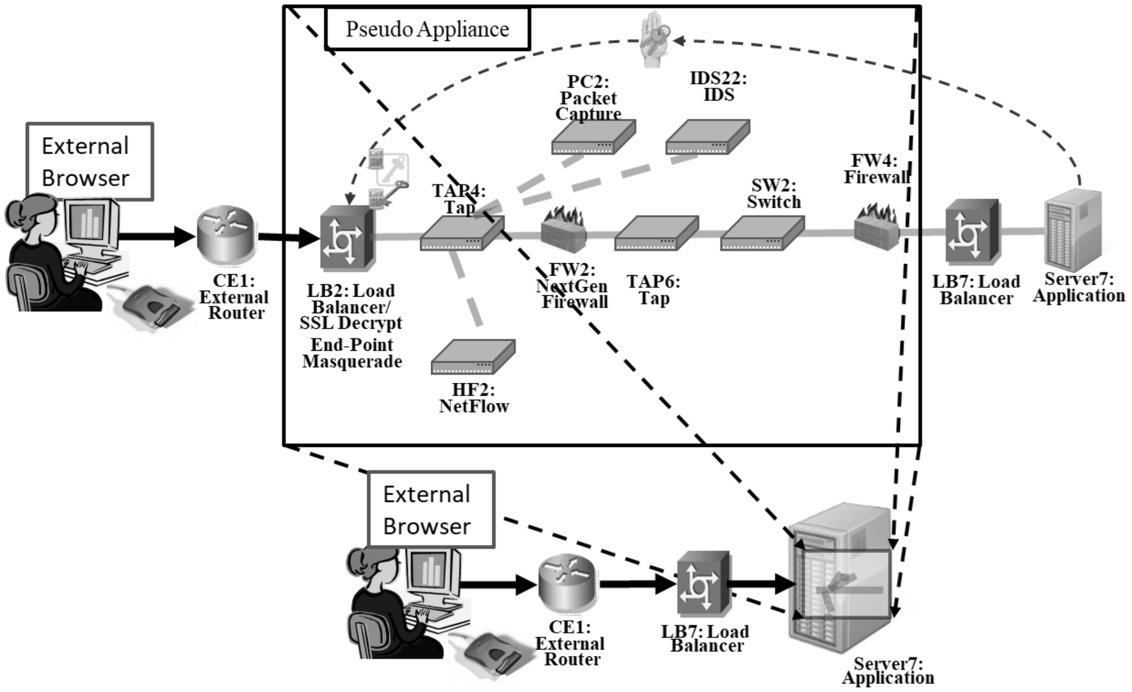


Figure 3 Creating the Pseudo Appliance

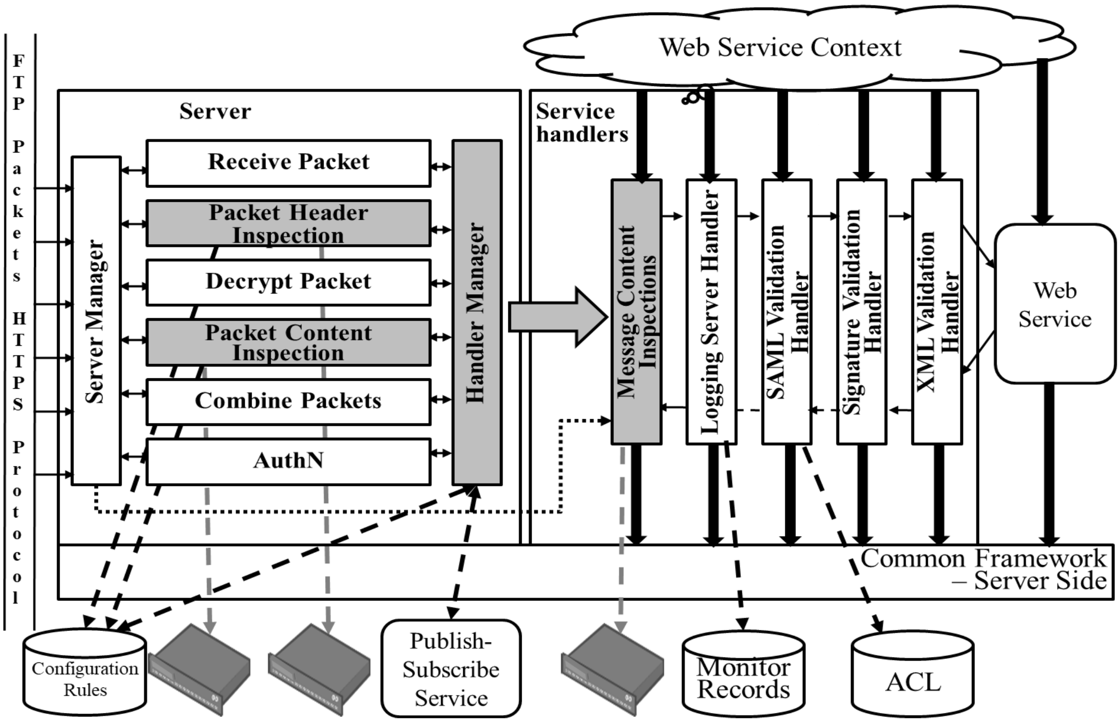


Figure 4 ELS End-point Network Security Functions

6. END-POINT PROTECTION SYSTEMS

The end-point protection system must provide firewall functionality under certain circumstances (as shown in Figure 6) based on end-point, claimed identity, requested action, and other factors.

- Black list – The only functionality enforced is block or drop packets. The black list is centralized, managed, and "pushed" to the protection system (ELS compliant)
- White – Varying degree of firewall enforcement based upon device and criticality. White membership includes The S3ecurity Token Server (STS), for example.
- Gray – Full firewall functionality is enforced. Functionality includes virus scan, malware scan, and other deep packet techniques.

The protection system has the capability to monitor, filter, or shut down traffic to given ports. It scans for malicious code. It examines incoming and outgoing traffic for anomalies or known exploits. It acts in the security context of the end-point for both requester and provider and examines not only the encrypted traffic but also the clear text traffic for malicious behavior or code. This requires access to the unencrypted traffic as well as the encrypted traffic. The protection system provides most but not all of the checks. Figure 6 walks through checks in an ELS enclave provided by the protection system, the server handlers, the service handlers, and the service itself, minimizing the need for in-line appliances.

This capability of the protection system is defined in terms of functional elements, some of which are listed below:

- Maintaining an inventory of assets on all hosts with situational awareness
- Detecting and removing of viruses, Trojans, worms, bots, and root kits in incoming and outgoing email

- Identifying unsafe websites during searches
- Detecting and repairing computer problems
- Enforcing policies on local machine
- Monitoring asset configurations and compare against baseline to detect changes
- Preventing use of unauthorized USB and flash media
- Blocking known and unknown buffer overflow exploits
- Preventing malicious code installation/execution
- Identifying activities that deviate from DoD or organizational policy
- Ensuring firewall functionality
- Monitoring DHCP requests on the network
- Marking any system that does not check in as rogue
- Scanning for compliance with policies
- Identifying host vulnerabilities on the network
- Making data available to the consumer, using ELS security
- Providing situational awareness
- And others as indicated by threat modelling

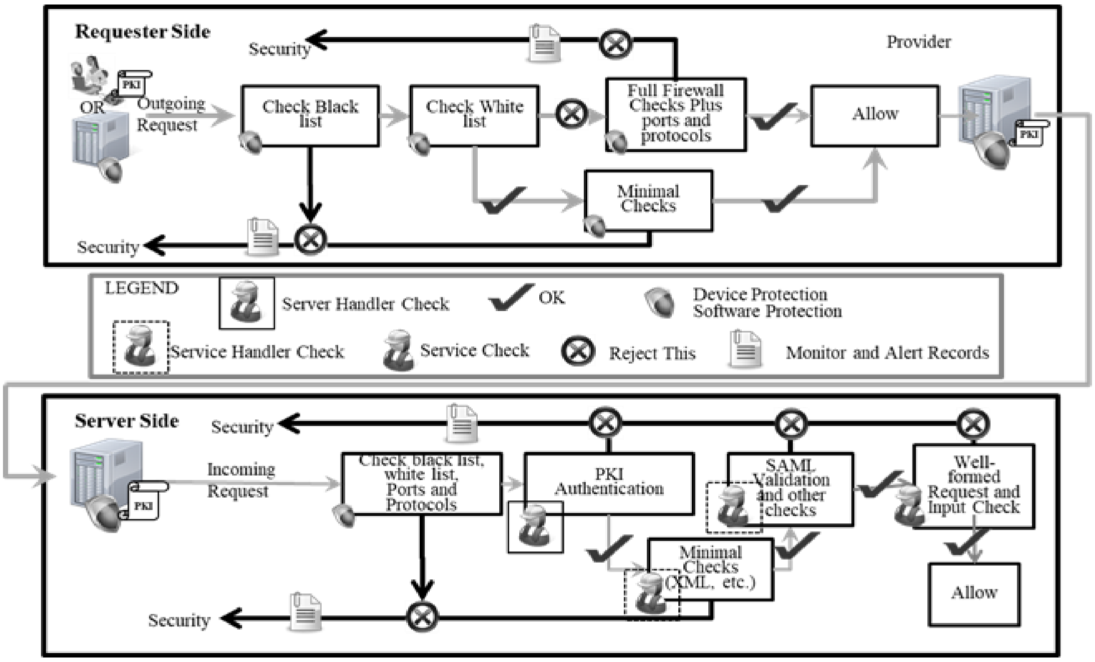


Figure 5 Protection Provided Without In-Line Appliances

The end-point protection system maintains an inventory of what is present (virtual and real) on all devices in the enterprise. Regular updates to this list ensures timely measures can be taken when an incident occurs. The protection system scans applications, configurations, permissions, services, registry entries, and other attributes to ensure that any changes from the baseline configuration have proper authorization. Any unauthorized or questionable differences from an approved baseline are reported to a central monitoring facility.

The protection system detects and removes malicious software from email by extracting, sandboxing and executing attachments to email in the user's security context before the user can do this. The execution is monitored and if malicious the attachment is removed from the email and forwarded to the security team for further analysis. Phishing can overcome people's mistrust of such attachments; this is an important part of device protection.

To prevent web-based attacks, the protection system flags potentially malicious sites to warn users. The protection system uses both heuristics and historical data to determine whether a site is safe or not. As search accesses many new sites, this is the ideal time for performance of such protection functions.

The protection system provides mechanisms to fix problems. Of course, a fully compromised system might be unresponsive to commands to fix certain issues, so this is not always possible. However, for most situations, remotely fixing the problem instead of requiring on-site manual intervention is the best course of action.

The protection system enforces policy on the local machine and enforcement of group policy or other methods for setting policy for compliance. Policies that are not enforced by the device itself must be monitored explicitly by the protection system.

The protection system keeps an accurate record of what the approved baseline configuration is for a given device [10].After a scan of the device, any differences are recorded and made available to the central monitor.

With new threats evolving through non-standard interfaces, such as USB, printers, and other attached devices, the protection system provides a way to manage these interfaces, either by monitoring or filtering traffic on them, disabling them, or using other methods to prevent attacks from these sources.

By closely monitoring code execution, the protection system prevents buffer overflows. Low-level system calls are monitored to track any attempts at writing to unallocated memory spaces, stopping both known and unknown buffer overflows from being exploited. This type of monitoring and prevention requires elevated privilege, as it requires access to system level resources, not just user data.

The protection system stops a user from installing new executable code, unless they are explicitly authorized. This prevents a user from compiling and running code downloaded from, or modified by, a malicious entity. It also provides a generic catch-all for any executables that may have bypassed the email or web monitoring functions. By stopping the user from installing executables, the protection system also stops malicious entities from using hijacked user accounts or sessions to run malicious code.

Enterprise enforcement of rules that govern behavior on their networks and devices is partially achieved by the protection system [11].Although many of these rules will already be handled through group policy or device Security Technical Implementation Guides (STIG), some activity can only be monitored dynamically through the protection system. For example, use of TLS with appropriate version, ciphers, two-way authentication using PKI, and use of appropriate extensions is not typically monitored using existing tools and must be implemented by the protection system.

7. END-POINT PROTECTION IN ELS

In ELS, an agent-type model is preferred. In this model, the packet header filtering and other security functions reside at the web server in the handler chain of the web service. The basic configuration of end-point protection in ELS shown in Figure 6; it provides a complete set of security functions for packet, message, and application layer security tailored for the specific web service being protected. The new functions added in the server are packet header inspection, packet content inspection, message content inspection, and application protection. These functions implement the ports and protocols protection, as well as other security functions normally provided by network devices such as intrusion detection/protection, packet and message content filtering, deep packet inspection, and application/web content filtering such as included in an application firewall.

A service requestor uses HTTPS to establish communication with the server hosting the target web service according to the ELS practice. The packet received by the destination server and the packet header are immediately inspected to perform the ports and protocols blocking, source whitelist/blacklist checking, and other filtering based on only the header, including stateful tracking of client addresses and ports. Until an HTTPS session is established, only packets addressed to the server's IP address and port 443 are allowed. Other ports may be opened as needed as part of the web service following establishment of the HTTPS session.

On the return path, the messages follow a similar process. In effect, the "packet header inspection" module performs the required network-layer filtering and can block traffic based on ports, protocols, and IP address. This makes the personal firewall essentially two-way in its filtering capacity.

In the ELS end-point protection architecture, the end-point protection modules can be configured to communicate with additional security monitoring appliances, such as a NetScout (or other traffic monitoring products), that can compile and track statistics about the security status of the server and the web service. The security appliances should be active entities and communicate with the server via TLS with mutual authentication. If required, the server could send the decrypted message traffic to other security appliances through this interface for additional security functions.

The end-point protection functions are configured through the server configuration management interface, which communicates with the server by TLS with mutual authentication. The ports and protocols, whitelist information, and any software updates are provided through this interface.

It is recommended that the initial configuration of the packet header deny all ports and protocols, both incoming and outgoing (as opposed to the traditional incoming only), and that permissions be configured in as they are identified as needed.

8. HANDLING AND INSPECTION OF TRAFFIC

Handling and inspection is done in software-only modules in the server. The handlers are embedded in the server handler chain at the point when and where the communication is prepared for their use and when and where the functionality has been distributed to packet-header inspection, packet content inspection, and message content inspection. Each of these may perform inspection related to intrusion detection or blacklist blocking, etc.

This is the preferred embodiment for enterprise applications. It moves the inspections to the point of the application itself by inserting handlers within the server and service to do the inspections at the point it makes most sense. The inspections that can be done without decrypting the packets may be done at the front of the web server because they are passive entities. Moving inspections of decrypted traffic inside the server not only preserves the end-to-end paradigm, it encapsulates the security and allows tailoring for the application itself. The encapsulated security with the application is virtualization ready.

9. CONCLUSIONS

We have reviewed the ELS security model and the end-to-end requirement within the enterprise. We have also reviewed the "normal" network defense process, and described the issues that the current network defenses raise and the vulnerabilities that may be introduced. Finally, we have provided an end-to-end approach that allows for both network inspection and reporting and the maintaining of unbroken encryption to the final destination, including enhanced defensive protections afforded by ELS. This approach is based on identifying the instances of official business and deferring the initial inspection until arrival at the target server. For enterprise operations, defining a clear end-to-end approach means a reduced attack space. The approach also reduces bandwidth requirements at the front door of the enterprise and may reduce the need for some load balancing. We have also reviewed the specific requirements for an enterprise level security that is bi-laterally authenticated and encrypted end-to-end. This paper is part of a body of work for high-assurance enterprise computing using web services. Elements of this work are described in [12-22].

REFERENCES

[1] Simpson, William R., CRC Press, "Enterprise Level Security – Securing Information Systems in an Uncertain World",by Auerbach Publications, ISBN 9781498764452, May 2016, 397 pp.

[2] Jack Wallen, "Five free, dead-easy IP traffic monitoring tools," Tech Republic, September 2011, https://www.techrepublic.com/blog/five-apps/five-free-dead-easy-ip-traffic-monitoring-tools/, last accessed 22 November 2019.

[3] Moskovitch R, Elovici Y. "Unknown malicious code detection – practical issues.", In Proceedings of the 7th European Conference on Warfare and Security (ECIW'08), Plymouth, UK, 2008.

[4] A. Begel, S. McCanne and S. L. Graham, BPF+: Exploiting global data-flow optimization in a generalized packet filter architecture, in: *Proc. of ACM SIGCOMM*, Cambridge, MA, USA (1999) pp. 123–134.

[5] M. McDaniel and M.H. Heydari, "Content Based File Type Detection Algorithms," Proceedings of the 36th Annual Hawaii International Conference on System Sciences, IEEE, ISBN: 0-7695-1874-5, DOI: 10.1109/HICSS.2003.1174905, Jan 2003.

[6] Mike Fisk and George Varghese, "Fast Content-Based Packet Handling for Intrusion Detection," Los Alamos National Lab Computing Communications and Networking Division, May 2001, https://apps.dtic.mil/dtic/tr/fulltext/u2/a406413.pdf, last accessed 22 November 2019.

[7] Jian Song and Yanchun Zhang. 2007, "Architecture of a Web Accelerator for Wireless Networks", In Proceedings of the thirtieth Australasian conference on Computer science - Volume 62 (ACSC '07), Gillian Dobbie (Ed.), Vol. 62. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 125-129.

[8] Shin-ichi Kuribayashi, "Improving Quality of Service and Reducing Power Consumption with WAN Accelerator in Cloud Computing Environments," International Journal of Computer Networks & Communications (IJCNC) Vol.5, No.1, January 2013.

[9] Afzal, S., Kavitha, G. "Load balancing in cloud computing – A hierarchical taxonomical classification." J Cloud Comp 8, 22. Decemeber 23, 2019, https://doi.org/10.1186/s13677-019-0146-7

[10] William R. Simpson and Kevin E. Foltz, Lecture Notes in Engineering and Computer Science, Proceedings of the World Congress on Engineering (WCE) 2018, "Enterprise End-point Device Management", pp. 331-336, Imperial College, London, 4-6 July 2018, IBSN: 978-988-14047-9-4, ISSN: 2078-0958.

[11] William R. Simpson and Kevin E. Foltz, Lecture Notes in Engineering and Computer Science, Proceedings World Congress on Engineering and Computer Science(WCECS) 2017, Volume 1, "Enterprise Level Security: Insider Threat Counter-Claims", pp112-117, Berkeley, CA. October 2017.

[12] William R. Simpson and Kevin E. Foltz, Proceedings of the Information Security Solutions Europe (ISSE) 2016, ISBN: 9781541211445, "The Virtual Application Data Center", pp. 43-59,

https://www.amazon.com/isse2016-3-Information-Security-Solutions-Europe/dp/1541211448, Paris, France, November 2016.

[13] William R. Simpson and Kevin E. Foltz,Haeng Kon Kim • Mahyar A. Amouzegar (eds.), Transactions on Engineering Technologies, Special Issue of the World Congress on Engineering 2015, Chapter 15, pp. 205-220, “High Assurance Asynchronous Messaging Methods”, 15 pp., DOI 10.1007/978-981-10-2717-8, Springer Dordrecht 2017.

[14] William R. Simpson and Kevin E. Foltz, Lecture Notes in Engineering and Computer Science, Proceedings of the World Congress on Engineering (WCE) 2017, “Assured Identity for Enterprise Level Security”, pp. 440-445, Imperial College, London, July 2017, IBSN: 978-988-14047-4-9.

[15] William R. Simpson and Kevin E. Foltz, Proceedings of The 21th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI, “Data Mediation with Enterprise Level Security”,WMSCI 2017, Orlando, Florida, 8-11 July 2017, 6 pages.

[16] William R. Simpson and Kevin E. Foltz, Proceedings of the 22nd International Command and Control Research and Technology Symposium (ICCRTS), “Escalation of Access and Privilege with Enterprise Level Security”, ISBN: 978-0-9997246-0-6, Los Angeles, CA. September 2017.

[17] William R. Simpson and Kevin E. Foltz, Sio-Long Ao, et. al. (eds.), IAENG Transactions on Engineering Sciences, Special Issue of the Association of Engineers Conferences 2016, Volume II, pp. 475-488, “Electronic Record Key Management for Digital Rights Management”, 14 pp., World Scientific Publishing, Singapore, ISBN 978-981-3230-76-7, 2018.

[18] William R. Simpson and Kevin E. Foltz, “Secure Identity for Enterprises," IAENG International Journal of Computer Science, vol. 45, no. 1, pp 142-152, ISSN: 1819-656X, February 2018.

[19] William R. Simpson and Kevin E. Foltz, Proceedings of the 8th International Conference on Electronics, Communications and Networks (CECNet 2018), Volume 1, “Cloud Security and Scalability”, pp 27, Bangkok, Thailand, November 2018.

[20] William R. Simpson and Kevin E. Foltz, “Insider Threat Metrics in Enterprise Level security," IAENG International Journal of Computer Science, vol. 45, no. 4, pp 610-622, ISSN: 1819-656X, December 2018.

[21] Simpson W. and Foltz K., Lecture Notes in Engineering and Computer Science, Proceedings World Congress on Engineering and Computer Science 2015, Volume 1, “Maintaining High Assurance in Asynchronous Messaging,” pp. 178–183, Berkeley, CA, October 2015.

[22] William R Simpson, and Kevin E. Foltz, "Mobile Ad-hoc for Enterprise Level Security," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2018, 23-25 October, 2018, San Francisco, USA, pp172-177.

Dr. Simpson has over two decades of experience working to improve systems security. He has degrees in Aeronautical Engineering and Business Administration. He also attended several schools for military and government training. He spent many years as an expert in aeronautics before delving into the field of electronic and system test, and he has spent the last 20years on IT-related themes (mostly security, including processes, damage assessments of cyber intrusions, IT security standards, IT security evaluation, and IT architecture).

Dr. Foltz has over a decade of experience working to improve security in information systems. He has degrees in Mathematics, Computer Science, Electrical Engineering, and Strategic Security Studies. He has presented and published research on different aspects of enterprise security, security modelling, and high assurance systems.

Time Series Classification with Meta Learning

Aman Gupta and Yadul Raghav

Department of Computer Science and Engineering
Indian Institute of Technology (BHU)
Varanasi, India 221–005

Abstract. Meta-Learning, the ability of learning to learn, helps to train a model to learn very quickly on a variety of learning tasks; adapting to any new environment with a minimal number of examples allows us to speed up the performance and training of the model. It solves the traditional machine learning paradigm problem, where it needed a vast dataset to learn any task to train the model from scratch. Much work has already been done on meta-learning in various learning environments, including reinforcement learning, regression task, classification task with image, and other datasets, but it is yet to be explored with the time-series domain. In this work, we aimed to understand the effectiveness of meta-learning algorithms in time series classification task with multivariate time-series datasets. We present the algorithm's performance on the time series archive, where the result shows that using meta-learning algorithms leads to faster convergence with fewer iteration over the non-meta-learning equivalent.

Keywords: Time Series, Classification, Meta Learning, Few Shot Learning, Convolutional Neural Network

1 Introduction

Computers have always been an immense source of fascination for human beings. Be it the computers beginning large computations, forecasting, accounting, it has always fascinated human beings. Furthermore, with the origin of the field of Artificial Intelligence, humans have been eluded to it never than before. The concept of an artificial agent learning to make decisions and perform tasks like humans is very much intriguing in itself. Learning (also referred to as Machine Learning due to the involvement of computers) has been studied in three significant categories Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Besides all these advances in supervised, unsupervised, and reinforcement learning, learning real-life tasks, learning algorithms, and carrying out simple instructions, learning from a few examples, is still regarded as a challenging problem. These are the class of problems that came to highlight in recent years; however, they have been known since the early 70's, known as Meta-Learning. Meta-Learning can be found as early as in the works of Donald B. Maudsley, where he referred to it as "the process by which learners become aware of and increasingly in control of habits of perception,

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 189-201, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101415

inquiry, learning, and growth that they have internalized". John Biggs redefined the same in a significantly more straightforward language as "being aware of and taking control of one's learning". In the context of computer science, meta-learning can be defined as simple as knowledge adaptability. It can be as simple as learning to identify previously seen objects with significantly fewer examples (few-shots learning), learning a task from some demonstrations (learning from demonstrations), or even following simple instructions.

Humans and animals adapt to any environment much faster and more efficiently. Whenever they try to learn any new skills or concepts, they hardly start from scratch. They start from the skills learned in the experience related to the new task, based on the knowledge they already have, that worked well before. For example, A child who has seen a dog, flower, bird only a few times can easily distinguish between them. A person who speaks one language can quickly learn to speak other languages with little practice, or a person who knows how to ride a bike can ride a motorbike with little practice. With every learned skill related to the given task domain, they can learn a new skill with a few examples and training. They simply learn how to learn across different tasks. Machine learning systems have surpassed humans at many tasks; it still requires training with a large number of samples on a specific task, and generally, models are trained from scratch using these samples to reach the same. So, it is not entirely fair for an AI algorithm to compare with humans as humans have prior knowledge and experience in their brain and DNA. Is it possible to imbue an AI system model with similar properties as a human with the ability to learn from experience and knowledge to learn new concepts and skills quickly with a few training examples rather than learning for scratch.

Much of the work has been proposed over the past few years. The earlier works address the problem of meta-learning as Few-Shot learning. The concept behind this is to design a deep neural network that can learn by simulating the datasets with very few instances, just like the babies learn to identify objects by seeing only a picture or two. In [1], the author proposed a method using the convolutional siamese neural network to do a few-shot image classification. Consecutively next year, an embedding method called Full Contextual Embeddings (FCE) [2], which uses bidirectional LSTM to encode the input vector to do the K-shot classification. With the recent success of optimization-based techniques, MAML [3], a model-agnostic algorithm for meta-learning, has been proposed, which has the ability to combined with any other trained gradient descent model, applicable to a variety of machine learning task. They looked at the learning process as maximizing the sensitivity of the loss functions of new tasks with respect to the model parameters such that small changes in the parameters will produce significant improvements on the loss function of any task. Training the model with a small number of gradient steps with a few samples of a task gives good generalization performance on that task.

In this work, we have focused on the optimization-based approach of meta-learning for time series classification tasks. Reptile [4], a meta-learning algorithm, is used as a framework for the experiment combined with a baseline architecture of a convolutional neural network, where the convolutional layer acts as a feature extractor for our classification model. We demonstrate the experiments on the UCR time-series datasets [6]. For the evaluation purpose, we have compared the loss function curve of algorithms showing that the meta-learning approach favors time-series tasks over non-meta-learning counterparts resulting in faster convergence over a sample of tasks with fewer iteration. We successfully demonstrate that reusing knowledge from past tasks and combining them with deep neural networks may provide a better result. To the simplest of our knowledge, this is the first successful attempt in adapting meta-learning for a time series classification task.

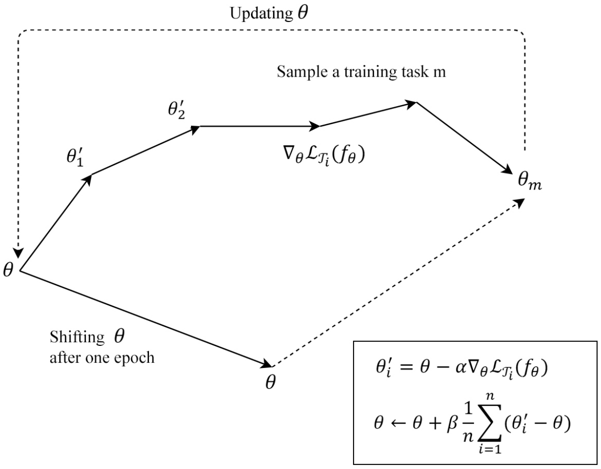

Updating θ
Sample a training task m
θ_2'
θ_1'
$\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
θ_m
θ
Shifting θ
after one epoch
θ
$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
$\theta \leftarrow \theta + \beta \frac{1}{n} \sum_{i=1}^{n} (\theta_i' - \theta)$

Fig. 1: A Meta-learning model optimizes the model parameters θ that can quickly adopt for a new task with parameters θ_i'. The model updates the initialization parameter θ.

Organization. The rest of the paper is organized as follows. Section 2 reviews the existing research work is done in the field of Meta-Learning and Time-Series Classification and builds the base for our main objective. The proposed work to scale the Meta-Learning algorithm to time-series classification tasks and demonstrate its performance on some dataset from the UCR time-series achieve. Section 4 gives the details of the experiment conducted using the explained methods, discusses an experimental study consisting of an evaluation strategy—finally, section 5, where we conclude with our observations in the chapter.

2 Literature Review

2.1 Time Series Classification

TSC has gained much attention in recent years due to its vast application in various domains such as financial services, human activity recognition, healthcare, weather forecasting [11]. Time series is nothing but just a measurement of statistical data taken several discrete times in chronological order. Mathematically, it can be written as,

$$h = f(t) \tag{1}$$

where h is the phenomena (function) at any given time.

Formally, TSC problem can be defined as follows: for a given set of classes Y, a training dataset $T = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ is a collection of pairs (X_i, Y_i) where $X_i = [X_i^1, X_i^2, \cdots, X_i^m]$ consists of M different univariate time series with Y_i as it is their corresponding class label $Y_i \subset Y$. The goal is to learn a classifier or model on a dataset T which, when fed with unseen time-series data points, the classifier is expected to predict its class correctly i.e., finding a function F such that $F(X_j) = Y_j$, and $Y_i \subset Y$.

As the earliest baseline, Lines and Bagnall [14] first demonstrate the effectiveness of distance-based methods such as Euclidean distance or Dynamic time warping (DTW) coupled with the nearest neighbor (NN) classifier to work directly on raw time series data. In the paper [15], the author proposed a feature-based approach extracting a set of features representing the global/local patterns of time series. These sets of features are combined to form Bag-of-words (BoW) or Bag-of-features and feed to the classifier [19] [20]. Several ensemble-based approaches have been explored Elastic Ensemble (PROP) [14], transform-based ensembles (COTE) [21] combining different classifiers over different time series representations. All these approaches need heavy preprocessing; feature extraction has higher time complexity. Earlier, the researchers neglected the fact of pure feature learning. With the success of deep learning models after 2012 [8], researchers began to exploit the idea of feature learning instead of handcrafted features.

2.2 Meta Learning

Meta-Learning is one of the most exciting areas of research in Artificial Intelligence and has been tackled for a long time. It has been tackled through different researchers as meta learner [16], few shots learning [17], meta reinforcement learning [18]. However, the latest mega boom in meta-learning began with the inception of Deep Learning with meta-learning. Much of the work has been proposed over the past few years. The earlier works address the problem of meta-learning as Few-Shot learning. The concept behind this is to design a deep neural network that can learn by simulating the datasets with very few instances, just like the babies learn to identify objects by seeing only a picture or two.

Over the recent year, many of the work has been published related to meta-learning, which classified into three approaches:- metric-based, model-based, and optimization based [10]. In a metric-based approach, the core idea is to learn embedding vectors of input data explicitly and use them to learn the best kernel functions. In the model-based approach, the idea is to design a model that updates its parameters rapidly with a few training steps, learning from the knowledge stored in the memory from the past training to learn a new task. Santoro *et al.* [12] built upon Differentiable Neural Computer and propose a new model, Memory-Augmented Neural Network (MANN). They described a few-shot learning specific data feeding pipeline wherein the answer (output label) of a previous input image is sent concatenated along with the current input. The idea is to encode new information in external memory and using this memory to updates its model parameters rapidly with a few training steps. Thus, the model, with external memory, through its controller had to learn to store the input Santoro *et al.* [12] representation in the external memory, associate the label provided with the current input with the previous input and retrieve the content of the relevant memory locations to produce the answer for the current query. Another approach is optimization-based, the core idea of learning a way to adjust the optimization parameters of the algorithm so that model converges within a small number of optimization steps learning with a few examples. Andrychowicz *et al.* [13] put forward a revolutionary idea of learning the optimizing function instead of hand designing it. The logic being, since we usually are learning a classifier “function” from our training data. Therefore we can also learn an optimizer function that performs the optimization process better. The idea seems simple but is very much revolutionary. This model puts forward meta-learning as well as transfer learning, wherein we can transfer the learned optimizer to other tasks. This meta learner can be scaled to thousands of parameters. The learned LSTM optimizer performed significantly well as compared to the state of the art models such as Stochastic Gradient Descent (SGD), Root Mean Square Propagation (RMSProp), Adam Optimizer (ADAM), etc.

Finn *et al.* [3] introduced the concept of model agnostic meta-learning (MAML). Here, there is a meta learner and a learner. The meta learner trains the learner on a training set that contains a different number of tasks. Through the meta learner, the model will acquire prior experience from training and will learn the common feature representations of the task. Thus, whenever there is a new task, the model will use its prior experience and will be fine-tuned to the new task on a small number of training data. MAML provides a good initialization of model parameters to achieve optimal fast learning.

3 Proposed Work

Frequently, tasks in meta-learning can be expressed as the problem of optimizing an objective function of a model $f(\theta)$ defined across the distribution of task over some domain $\theta \in \Theta$. So, the goal, in this case, is the meta-learning model should be able to find an optimal initialization for parameters of a randomly sampled task where each task is associated with dataset T_i (each dataset is considered as one data sample). Mathematically, it can be expressed as

$$\theta^* = argmin_\theta \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})}[\mathcal{L}_{\mathcal{T}_i} f(\theta)] \tag{2}$$

Our work is built on the recently proposed model Reptile, an optimization-based approach somewhat similar to MAML as they are both trained with gradient descent and model-agnostic, but much simpler in implementation and training. Reptile works repeatedly sampling a task, training on it through multiple gradient descent steps, therefore, shifting the model weights towards the new parameters of the unseen task. The algorithm updates the model into two stages: First, it considers the model as function $f(\theta)$ with parameter θ, then the classifier is trained on a given task T_i, changing the model parameter θ to $\theta_i^{'}$.

$$\theta_i^{'} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i} f(\theta) \tag{3}$$

Generally, a task contains a training set having few or limited examples for each of the classes with one or more test examples (few-shot learning) [7]. The reptile objective is to learn the optimal initialization for the parameters of the neural network model such that the classifier learns faster while optimizing at test time, leads to generalizing the model with fewer examples from the test task. In the final optimization step, it simply updates θ to $(\theta_i^{'} - \theta)$ as a reptile gradient applied with stochastic gradient descent (SGD).

$$\theta = \theta + \beta \frac{1}{n} \sum_{i=1}^{\infty} (\theta_i^{'} - \theta) \tag{4}$$

In this work, we have used the same architecture proposed in the paper [6] as the baseline architecture for our classifier. The author proposed a straightforward deep neural network-based architecture for TSC, which gives remarkable results with 44 UCR time-series datasets. They have tested on three deep neural network architectures, Multilayer Perceptrons (MLP), Fully Convolutional Networks (FCN), Residual Network, to provide a fully comprehensive baseline. For our experiment, we have used only FCN architecture, which can be applied to the dataset without any feature crafting and heavy data preprocessing. An intuition behind using an FCN network is that applying several convolutional layers on time series would be more helpful in learning a discriminative feature for the classification task.

3.1 Network Architecture

Deep convolutional neural network (CNN) has gained much attention in many different domains like regression, classification task, natural language processing (NLP), information retrieval, etc. after the AlexNet [8], won the 2012 ImageNet competition. With the increasing success of this architecture in different research fields, researchers started exploring the success of convolutional network architecture for time-series analysis. As we have seen applying the convolution on the image by sliding the filters in two dimensions (height and width), similarly, we could apply the convolution over time-series by sliding the filter to exhibit in one dimension (time).

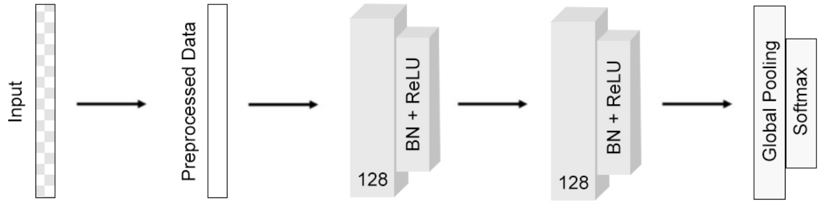


Fig. 2: The baseline architecture of fully convolutional neural network used for training the meta-learning algorithm.

In our problem setting, applying FCN over time-series helps extract multiple discriminative features for the classification task. The softmax layer finally outputs a probability distribution over the class variables in a task. The basic block of FCN mainly consists of three components viz., convolutional layer, followed by batch normalization (BN), and a final non-linearity function Rectified Linear Unit (ReLU). $\otimes$ is the convolution operator, applied with a filter W on a time series data x, and a bias parameter b.

$$\mathsf{C} = \mathsf{ReLU}[\mathsf{BN}(\mathbf{W} \otimes \mathbf{x} + \mathbf{b})] \tag{5}$$

The architecture is built up by stacking three convolutional blocks. The final extracted features from the convolutional block are fed to the global average polling layer reducing the number of weights in a model, thus preventing the risk of overfitting. Batch normalization has been performed after each convolutional layer helping the network to converge quickly and improve generalization. In the end, the softmax layer is applied for classification.

3.2 Preprocessing

Normalization of the dataset has been performed across each channel (attribute); missing values are filled with zeros. Many of the preprocessing steps have been applied for generating the task to fed FCN. To generate the distribution of tasks of the same domain, we hide some of the channels with zeros padding in the dataset and also shuffle the order of the dataset across the channels. We have divided the dataset into the training set and test set such that samples in the train set have classes that are not present in the test set, making a more significant number of task distribution for training the model. The initialization of hyper-parameters, kernel size, and others have been followed the same as given in the paper. *The implementation was done purely in Python and Tensorflow.

3.3 Algorithm

Reptile is the first-order gradient-based meta-learning algorithm. Given the FCN architecture, the adaptation of meta-learning using the Reptile framework can be summarized in algorithm 1. It looks very similar to the ordinary learning tasks, but for each training sample, we create a pseudo-task T_i because the algorithm treats one dataset as one data sample.

Algorithm 1 Reptile with FCN

```
Require: Training Dataset $\mathcal{D} = \{x^{(j)}, y^{(j)}\}$
Require: $\alpha, \beta$ step size hyperparameters
Require: Number of task, examples and batch-size
Denote $p(\mathcal{T})$ as distribution over tasks
Randomly initialize $\theta$
for $iteration = 1, 2, \ldots$ do
    Sample task $\tau_1, \tau_2, \ldots, \tau_n$, batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
    for $i = 1, 2, \ldots, N$ do
        Evaluate $\mathsf{W}_i = \mathsf{SGD}(\mathcal{L}_{\tau_i}, \theta, k)$
        Compute updated parameter with gradient descent using (1)
    end for
    Update $\theta \leftarrow \theta + \beta \frac{1}{n} \sum_{i=1}^{n} (\mathsf{W}_i - \theta)$
end for
```

4 Experimental Study

In this section, we present the information about the datasets we have used, evaluation strategy to compare between meta and non-meta-learning algorithms, and the experimental result analysis and training setups. We have built the architecture by stacking two convolution blocks with the filter sizes $128, 128$ in each block. After each convolutional layer we have applied batch-normalization and ReLu activation layer to improve the generalization capability. Reptile uses 50 examples per task, so we followed the same sample size for the task. We have used Adam optimizer with learning rate 0.01 to optimize the loss.

4.1 Dataset

The performance of our model has been evaluated on the UCR time-series repository, containing more than 40 distinct time-series datasets. The dataset in the repository has been broken into the different domains (Image Outline, Sensor Readings, Motion Capture, Spectrographs, ECG, Electric Devices, and Simulated Data), having different characteristics and varying length. We have used four datasets from the repository and one other dataset (human activity recognition), which we have collected in our lab. All the chosen dataset is multivariate time series, the preprocessing on the dataset as described in Section 3. The details about the dataset used in our experiment are given in Table 1.

Table 1: Time Series Dataset

Name	Dataset				
	Train Size	Test Size	No. of attributes	Length	No. of classes
PenDigits	7494	3498	2	8	10
UWaveGestureLibrary	896	3582	3	315	8
CharacterTrajectories	1422	1432	3	182	20
Motion	390	390	3	300	12
PeekDB	1000	1000	20	60	5

4.2 Results

To evaluated our model, we first trained the meta-learning model on tasks generated from the training dataset. After we got the optimal parameters from training, we trained two models on the task generated from the test dataset. The first model with parameters initialization which we get from meta learning one and other model with randomly initialized parameters. Then we compare the loss function curve of these models while training. The tasks we generate is used for binary classification for each dataset, so that we could generate variety of the tasks from same domain. Here, we calculated the total cross entropy loss to measures the performance of a classification model on test task. Cross-entropy loss decreases as the predicted probability converges to the actual label.

Figure 3. shows the experimental results of our model on the dataset. The graph represents the training loss function curve on the test tasks with the number of epoch while training the model. We observe that the meta-learning model converges more quickly with the fewer number of epoch than the model, which are trained from scratch (non-meta- learning) on every dataset. We believe that the success of metal learning models is because of the optimal initialization of parameters that we obtained from the Reptile algorithm. This experiment confirms that there is a significant difference between the models which are trained with optimal parameter initialization then random initialization. From the experiment, we observe that the meta-learning algorithm performs better with the dataset having more number of attributes (PeekDB) and a larger length of time-series (UWaveGestureLibrary, CharacterTrajectories, Motion). So, using the meta-learning algorithm with time-series tasks can be very useful.

5 Conclusion

In this work, we incorporate a meta-learning algorithm with a fully convolutional network for time-series classification tasks. With this, we establish what different type of learnings are, the emergence of meta-learning, coupling of meta-learning with deep learning, and the state of the art models in meta-learning (MAML, and Reptile). Employing this architecture, we find the optimal parameter initialization for the task. The proposed method is evaluated using a loss function curve, which shows its superiority over the non-meta-learning counterpart. The test suggests the using meta-learning algorithms with time series could be quite effective. Meta-learning helps to learn any new task quickly based on prior experience gained from other similar tasks. It is better to use meta-learning in real-world tasks, as there are plenty of opportunities to learn from prior experience.

6 Acknowledgement

We want to thank Rafael Radrumond, Lucas Brinkmeyer, for helpful discussions and suggestions during the early phases of the research. We would also like to thank the anonymous reviewers for sharing their helpful suggestions and Chaitanya Bhatia and Karan Siwach, for their valuable feedback.

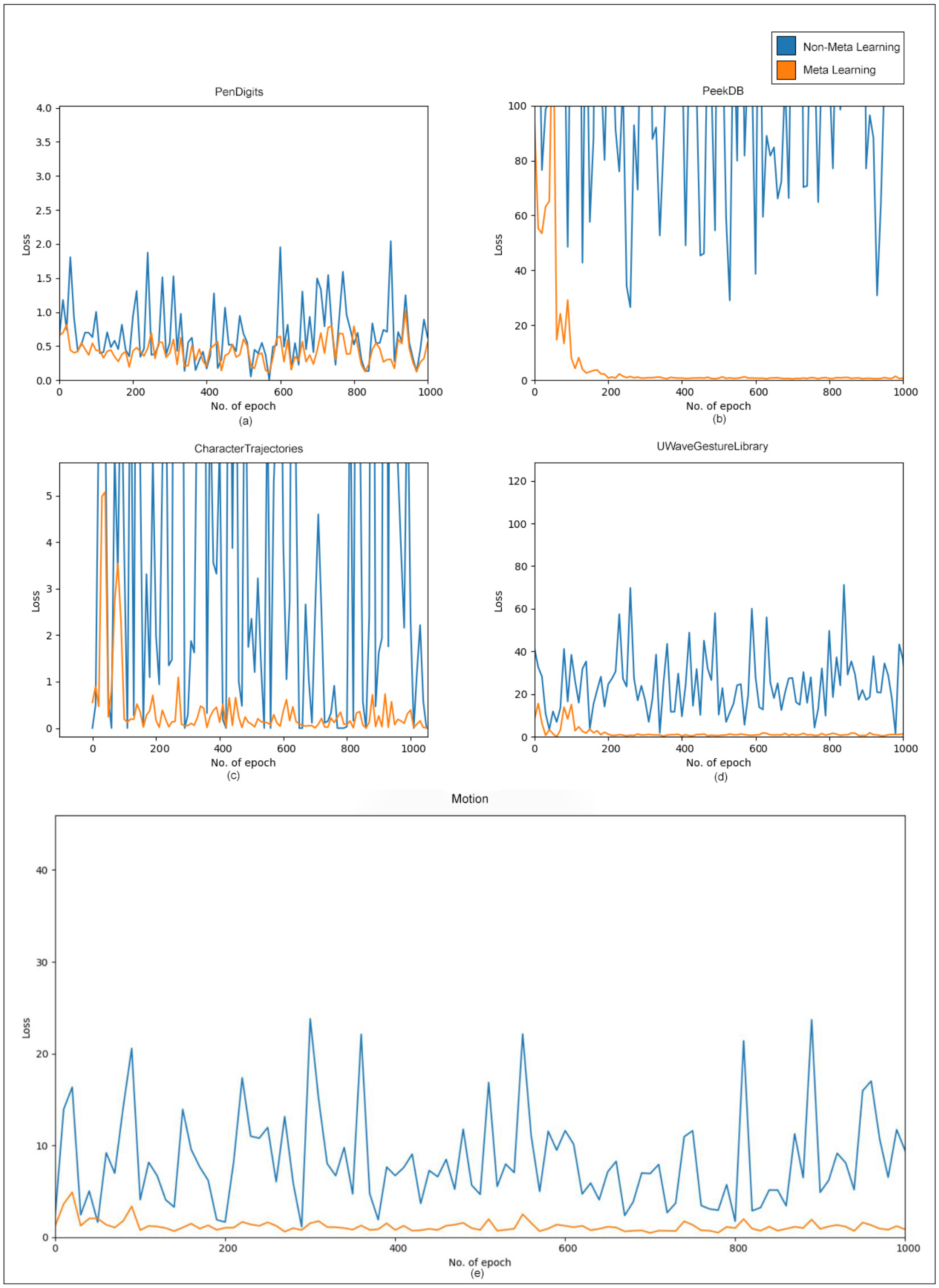


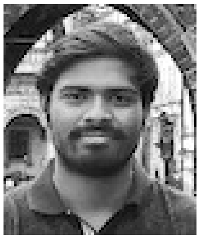
Fig. 3: Demonstrate the effectiveness of Meta-Learning over the Non-Meta-learning model with the help of a loss function curve on a time series dataset.

References

1. Gregory Koch and Richard Zemel and Ruslan Salakhutdinov, “Siamese neural networks for one-shot image recognition,” ICML Deep Learning Workshop, 2015.
2. Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, Daan Wierstra,“Matching networks for one shot learning,” NIPS. 2016.
3. Chelsea Finn, Pieter Abbeel, and Sergey Levine,“Model-agnostic meta-learning for fast adaptation of deep networks,” ICML 2017.
4. Alex Nichol, Joshua Achiam, John Schulman, “On First-Order Meta-Learning Algorithms,” arXiv preprint arXiv:1803.02999 (2018).
5. Chen Y, Keogh E, Hu B, Begum N, Bagnall A, Mueen A, Batista G (2015b) The UCR time series classification archive. www.cs.ucr.edu/ẽamonn/time_series_data/.
6. Zhiguang Wang, Weizhong Yan, Tim Oates, “Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline,” arXiv:1611.06455v4 [cs.LG], 2016.
7. Flood Sung, et al, “Learning to compare: Relation network for few-shot learning,” CVPR. 2018.
8. Krizhevsky A, Sutskever I, Hinton GE, “ImageNet classification with deep convolutional neural networks,” In: Advances in neural information processing systems vol 25, pp 1097–1105, 2012.
9. Po-Sen Huang, Chenglong Wang, Rishabh Singh, Wen-tau Yih, Xiaodong He,“Natural Language to Structured Query Generation via Meta-Learning,” arXiv:1803.02400v4 [cs.CL], 2018.
10. Weng, Lilian, “Meta-Learning: Learning to Learn Fast,” lilianweng.github.io/lil-log, 2018.
11. Ismail Fawaz, H., Forestier, G., Weber, J. et al, “Deep learning for time series classification: a review,” Data Min Knowl Disc 33, 917–963(2019). https://doi.org/10.1007/s10618-019-00619-1.
12. Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra and Timothy P. Lillicrap, “One-shot Learning with Memory-Augmented Neural Networks,” CoRR,abs/1605.06065,2016.http://arxiv.org/abs/1605.06065.
13. Marcin Andrychowicz, Misha Denil, Sergio Gomez Colmenarejo, Matthew W. Hoff-man, David Pfau, Tom Schaul, and Nando de Freitas,“ Learning to learn by gradient descent by gradient descent,” CoRR, abs/1606.04474,2016.http://arxiv.org/abs/1606.04474.
14. Lines J, Bagnall A, “Time series classification with ensembles of elastic distance measures,”2015. Data Min Knowl Discov 29:565–592.
15. J. Lin, E. Keogh, L. Wei, and S. Lonardi, “ “Experiencing sax: a novel symbolic representation of time series,” Data Mining and knowledge discovery, vol. 15, no. 2, pp. 107–144, 2007.
16. Nikhil Mishra, Mostafa Rohaninejad, Xi Chen and Pieter Abbeel, “Meta-Learning with Temporal Convolutions,” CoRR,abs/1707.03141, 2017.http://arxiv.org/abs/1707.03141.
17. Sung, Flood and Yang, Yongxin and Zhang, Li and Xiang, Tao and Torr, Philip H.S. and Hospedales, Timothy M.,“ Learning to Compare: Relation Network for Few-Shot Learning,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June, 2018.
18. Sutton, Richard S and Barto, Andrew G, “Reinforcement learning: An introduction,” MIT press; 2018 Oct 19.
19. M. G. Baydogan, G. Runger, and E. Tuv, “A bag-of-features framework to classify time series,” IEEE transactions on pattern analysis and machine intelligence, vol.35, no.11, pp.2796–2802, 2013.
20. P. Schafer, “The boss is concerned with time series classification in the presence of noise,” Data Mining and Knowledge Discovery, vol.29, no.6, pp.1505–1530, 2015.
21. A. Bagnall, J. Lines, J. Hills, and A. Bostrom,“Time-series classification with cote: the collective of transformation-based ensembles,” IEEE Transactions on Knowledge and Data Engineering, vol.27, no.9, pp.2522–2535, 2015.

Authors

Aman Gupta and **Yadul Raghav**, we received a Bachelor's and Master's degree from the Indian Institute of Technology (BHU), Varanasi, majoring in Computer Science and Engineering. Our interest lies in the field of machine learning, deep learning, and data mining.



Aman Gupta

Yadul Raghav

Deep Learning Roles Based Approach to Link Prediction in Networks

Aman Gupta[⋆] and Yadul Raghav[⋆]

Department of Computer Science and Engineering
Indian Institute of Technology (BHU)
Varanasi, India 221–005

Abstract. The problem of predicting links has gained much attention in recent years due to its vast application in various domains such as sociology, network analysis, information science, etc. Many methods have been proposed for link prediction such as RA, AA, CCLP, etc. These methods required hand-crafted structural features to calculate the similarity scores between a pair of nodes in a network. Some methods use local structural information while others use global information of a graph. These methods do not tell which properties are better than others. With an in-depth analysis of these methods, we understand that one way to overcome this problem is to consider network structure and node attribute information to capture the discriminative features for link prediction tasks. We proposed a deep learning Autoencoder based Link Prediction (ALP) architecture for the latent representation of a graph, unified with non-negative matrix factorization to automatically determine the underlying roles in a network, after that assigning a mixed-membership of these roles to each node in the network. The idea is to transfer these roles as a feature vector for the link prediction task in the network. Further, cosine similarity is applied after getting the required features to compute the pairwise similarity score between the nodes. We present the performance of the algorithm on the real-world datasets, where it gives the competitive result compared to other algorithms.

Keywords: Link Prediction, Deep Learning, Autoencoder, Latent Representation, Non-Negative Matrix Factorization.

1 Introduction

With the surge of the Internet, everyone is almost interconnected via social media platforms (e.g., Facebook, Twitter, Instagram, etc.), professional blogs, and websites. Networks are tools to represent these interconnections in their respective scenarios. For example, an individual profile on Facebook is represented by a node in the network, and relationships between two profiles are represented by the links (edges) between two nodes. Thus, a network can be used to model the communication of a system. Everyday, relationships are changing among individuals, i.e., some new links are formed, and some of them are vanished due to several reasons. This

[⋆] These authors have contributed equally.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 203-222, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101416

behavior makes the scenarios quite complex and dynamic, and dealing with them becomes more challenging. The above scenarios can be modeled using a social or a complex network. Lots of issues exist when dealing with these networks. “Finding missing links or future links in an observed network” is one of the interesting problems which is known as link prediction (LP). Nowell and Kleinberg [1] formally defined the link prediction as follows. Suppose a graph $G_{t_0-t_1}(V, E)$ represents a snapshot of a network during time interval $[t_0, t_1]$ and $E_{t_0-t_1}$, a set of edges present in that snapshot. The task of link prediction is to find set of edges $E_{t_0^{'}-t_1^{'}}$ during the time interval $[t_0^{'}, t_1^{'}]$ where $[t_0, t_1] \leq [t_0^{'}, t_1^{'}]$. Link prediction has been applied in several domain of applications like friendship prediction in Facebook, recommendation system in e-commerce [2], protein-protein interactions (PPI) in bioinformatics [3] etc.

Several authors have presented seminal works on classical methods (indices) of link prediction in networks that are broadly classified into structural similarity-based and maximum likelihood-based methods, which are presented in the reviews [1, 4]. Structural similarity is computed based on the properties of the structure of the networks. These properties are easy to compute, and no need to extract extra information like attribute and other side information. Works on these structural properties are grouped into three categories viz., local, global, and quasi-local. Most classical similarity indexes are based on these three categories. Local indices extract local information (like degree, common neighbors, clustering coefficients, etc.) to compute the similarity between two nodes. Common neighbor [1], Adamic-Adar [5], Resource allocation [6], Preferential attachment [7],etc. are the most well known local indices which are heuristics and used in both supervised and unsupervised settings to show the relevance of other methods. Global indices focus on extracting global properties or information where the whole network is taken into account. These indices are more complex and time-consuming that limits its application for large networks. Most path based indices e.g., Katz [8], shortest path [1], average commute time [9], PageRank [10], Leicht-Holme-Newman Index [11], Random walk with restart [12] etc., comes under global indices. Local similarity indices are simple and efficient in computation, whereas the global are complex and computationally inefficient. A trade-off between them is quasi-local indices, which are as efficient as local indices and not limited to neighborhood information. Sometimes, these methods (indices) take the whole network in computation [13]. Local path index (LP) [14], Local random walk (LRW) [15], Superposed random walk (SRW) [9] etc., are such indices.

Several machine learning [15, 16], and deep learning frameworks [17, 18] also have been explored for the link prediction task. Both supervised and unsupervised models have been used to find missing links in the literature. An unsupervised deep learning model aims at finding hidden structures or patterns in the data and learns suitable representation that is useful input for several tasks. One of the critical issues with these models is the data representation that extracts useful information from the data. Unsupervised models detect and eliminate irrelevant variability present in the input data. Simultaneously, it preserves the information that is useful to several tasks like detection, prediction, visualization, etc. Some of these models are based on reconstructing the input from the suitable representation (or code) with some desired properties like sparsity, low dimensionality, etc. The autoencoder is one of the unsupervised deep learning models using which we predict missing links in the paper. An autoencoder consists of two parts encoder and decoder. The encoder maps the input data to latent representation or code or feature vector, and this code can further be used in downside the prediction task. Autoencoder can be expressed in another way as consisting of one input layer, one or more hidden layers, and an output layer, as shown in Figure 1. A deep autoencoder consists of more hidden layers. Reasons for the use of the deep autoencoder are two folds first, once the training is complete, computing code (latent representation) takes less time; the relevancy of the extracted information can be checked through the decoder by reconstructing the input. In contrast, link prediction task in social networks can also be treated as grouping the nodes based on the similar structural properties and behaviour. Given a network, we have defined the node-roles relationship between them, with the intuition that two nodes belong to the same role, or we could say they are well connected to each other if they have similar structural behavior or function. Node-Roles relationships help in understanding the underlying behavior in a network and also exploring the interesting data analysis tasks such as sense-making, node-similarity, and prediction tasks [63].

In this paper, we transfer the effects of roles in a network to link prediction tasks, where node-roles relationship act as additional features to find the similarity between the nodes in a network. Without any other information except the network structure, the key problem with the link prediction task is identifying and deciding what structural features are needed that can be derived from the network, which will lead to predicting links in the network. Once we get the desired features, the link prediction task can be well-formulated as finding similarity between the nodes based on these extracted features. Given a dataset, we define the link prediction problem as

- finding the features matrix and node-roles relationship (where the features matrix is obtained by using the latent representation of autoencoder architecture, and the node-roles matrix is determined by applying the non-negative matrix factorization on these extracted feature matrix).
- determining the similarity score between the nodes using both feature matrix and node-roles matrix by applying the Cosine similarity function.

We summarize the main contributions of this paper are determining a set of structural features using the autoencoder architecture and transferring the effect of node-roles relationship to perform the link prediction task.

Organization. Section 2 talks on some literature work on link prediction. The proposed work is presented in section 3. Section 4 discusses an experimental study consisting of an evaluation strategy and the results of several methods against real network datasets followed by a statistical test. Finally, section 5 concludes our work.

2 Related Work

Newman presented a paper on link prediction on collaboration networks in Physics and Biology [19]. In such networks, two authors are considered to be connected if they have at least one paper co-authored by them simultaneously. In the empirical study, the author demonstrated that the likelihood of a pair of researchers teaming up increments with the numbers of different colleagues they have in mutual relation, and the likelihood of a specific researcher acquiring new partners increments with the number of his past teammates. The outcomes give experimental proof in favor of formerly guessed mechanisms for clustering and power-law degree distributions in networks. Later, Nowell and Kleinberg [1] proposed a link prediction model explicitly for a social network. Each node in the network corresponds to a person or an entity, and a link between two nodes shows the interaction between them. The learning paradigm in this environment can be used to extract the similarities between two nodes by several similarity metrics. Ranks are assigned to each pair of nodes based on these similarities, then higher ranked node pairs are designated as predicted links. Further, Hasan et al. [15] expanded this work and demonstrated that there is a significant increase in prediction results when additional topological information about the network is available. They considered different similarity measures as features and performed a binary classification task using a supervised learning approach, which is similar to link prediction in their framework.

The graph embedding is considered as a dimensionality reduction technique in which higher D dimensional nodes in the graphs are mapped to a lower d ($d \ll D$) dimensional representation space by preserving the graph properties as much as possible [20]. These graph properties can be node pair similarity, node neighborhood similarity, substructure similarity, etc. Recently, some graph embedding techniques [21]- [25] have been proposed and applied successfully in link prediction and node classification problems. Deep learning models of graph embedding have also recently been introduced that can be classified in with and without random walk strategies [20]. In the first case, the graph is represented as paths sampled from it, which are inputs to an embedding model and the whole graph as input for a later case. The deep learning model is then applied to these sampled paths in the framework and encodes to preserve the graph properties (i.e., path properties here). Lots of seminal works based on the first category (i.e., with random walk strategy) are available in the literature like DeepWalk [24], Node2vec [22], HARP [26]. These models are mainly shallow in nature, moreover deep model are based on without random walk strategies like SDNE [27], DNGR [28], VAGE [29], SEAL [30].

Machine learning and deep learning for link prediction. In the literature, most of the well-known link prediction approaches focus on heuristics, which are domain-specific and ignore the evolutionary behavior of the networks. They mainly work on static networks. From the last decade, several machine learning approaches have been applied to improve link prediction performance. In such approaches, the challenging task is to represent features in a format suitable for the application, which vastly affects the performance results. M. A. Hasan et al. [15] proposed a seminal work on link prediction using supervised learning in which three types of features of graph viz., proximity features, aggregated features, and topological features are employed with several classifiers. Likewise, Doppa et al. [16] put forward works based on a supervised approach on link prediction where k-means classifier employed on feature vectors. Recently, deep learning, a new direction in machine learning have been proposed in the literature. The seminal works based on deep models, for examples, stacked denoising autoencoders (SDAE) [17] and convolutional neural networks (CNN) [18] have shown their great potential of representing and learning features in computer vision and natural language processing. One problem of conventional deep learning models is the independent and identically distribution of the input, which cannot model relational data. To overcome this issue, H. Wang et al. [31] employed a Bayesian deep learning framework that learns relational data (network data) effectively. They jointly model high-dimensional node attributes and link structures in their framework and product of Gaussian as an inference approach. Xiaoyi Li et al. [32] introduced a novel deep learning framework, namely Conditional Temporal Restricted Boltzmann Machine (ctRBM), that captures the evolutionary patterns of networks (i.e., dynamic networks). Their framework is based on the joint inferential effects of seed nodes and their local neighbor's influences. [33] proposes a supervised framework of deep learning where two different architectures show their competitive results with the state-of-the-art.

Graph convolutional neural networks (GCNs) [34] are the recent class of deep network approaches used in network embedding, node classification, and link prediction. The model learns representation from a localized first-order approximation of spectral convolutions. Thomas N. Kipf et al. [29] introduced a framework based on GCN that uses a simple inner product decoder and learns node features of structured graph data. In this paper [65], the author considers the link prediction task as a collaborative filtering problem, where they treated the nodes as items and edges like the rating in the recommendation system and proposed a non-negative matrix factorization approach combined with a bagging technique to predict which nodes are expected to connect. Similarly [64], a unified framework has been proposed for link prediction tasks based on non-negative matrix factorization with coupling multivariate information where they have used both the internal latent feature information and external node attribute information of the network. Different approaches have been used for role extraction in the network. In [63], RolX an unsupervised learning approach has been proposed where it automatically determines the underlying structural roles in a network. It also assigns a mixed-membership of these extracted roles to each node in the network. The author has analyzed different methodologies, research issues, and characteristics that should be considered during the role analysis.

3 Proposed Work

3.1 Network Architecture

The proposed architecture is an unsupervised framework of deep learning model which maps the adjacency matrix of the given graph into the node-features matrix. The architecture does not need any hand-crafted (manual selection) features; rather, it extracts important features automatically. The overall architecture has been divided into two stages. The first stage consists of the autoencoder neural network, which is an unsupervised framework of deep learning that uses backpropagation to update synaptic weights. It mainly consists of two parts encoder and decoder. The encoding layer compresses its input to a lower-dimensional code, known as latent representation. The objective of the decoding layer is to reconstruct the input using this compressed code. Clearly, an autoencoder can be considered as a dimensional reduction technique. The encoder maps the input data to latent representation or feature vector, and this vector can further be used in downside the prediction task. Through this model, we get only the important features that are needed to represent the graph by filtering out the unnecessary details from the graph. We have used this feature vector as a subset of our final node-feature matrix. The second stage uses the given latent representation from the neural network to find out the roles in the graph. Using these roles, we have assigned a mixed-membership of roles to each node in the graph network, giving a node-role matrix, which acts as an additional feature set for our node-feature matrix. In summary, to get the final Node-feature matrix, we have concatenated the feature vector, which we get after the encoder layer from the autoencoder neural network with the node-role matrix. In the proposed architecture, the first hidden layer consists of 16 neurons and the second hidden layer (or latent representation) contains 8 neurons (called latent variables). The learning rate is set to a low value of 0.01 in descent gradient optimization during the learning process. The workflow of our proposed architecture has been shown in Figure 1.

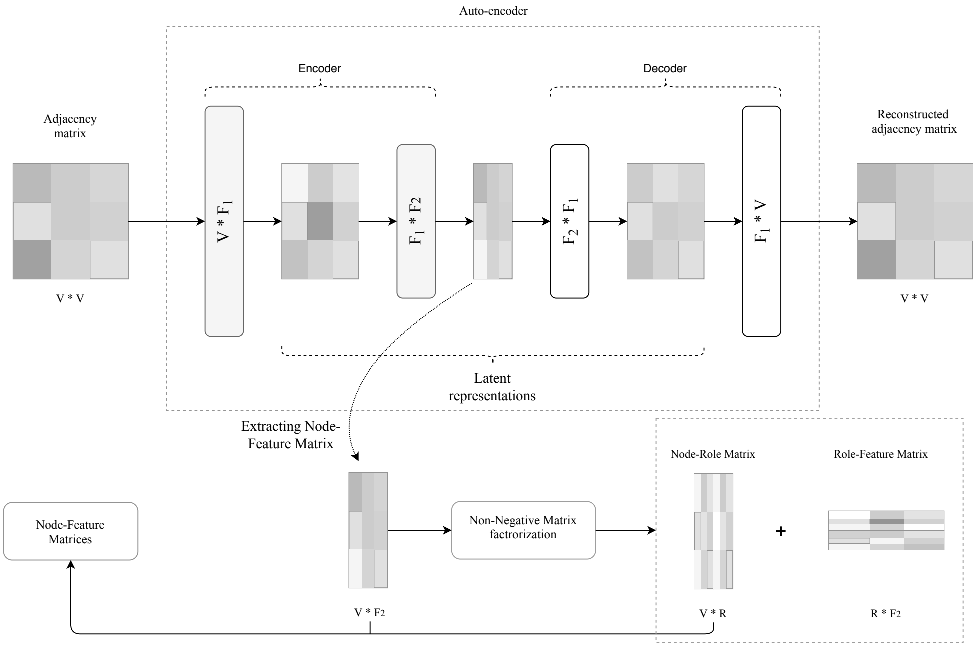


Fig. 1: The Deep Autoencoder Framework.

3.2 Problem characterization for deep autoencoder framework.

Considering a simple undirected network (also applicable to directed and weighted networks), $G(V, E)$, where V is the set of vertices (or nodes) and E is the set of edges. The given graph network can be represented as an adjacency matrix $A \in \mathbb{R}^{|V|\times|V|}$. Inputs to the neural network is a matrix, which is the adjacency matrix A. Aim of the deep autoencoder is to learn low dimensional latent representation $Z \in \mathbb{R}^{|V|\times F_2}$ for the nodes with the constraint of minimization of the reconstruction error (loss). Other half of the architecture aims to find the node-role matrix of dimension $R \in \mathbb{R}^{|V|\times r}$ with the help of feature vector Z. The main focus of the overall architecture is to find the node-feature matrix of dimension $F \in \mathbb{R}^{|V|\times F}$.

3.3 Preprocessing.

Input to the proposed model is a normalized adjacency matrix (A_{Norm}), which is the output of the preprocessing step. Normally, neural architectures use the original adjacency matrix in a layer-wise propagation function that causes a change in the scale of feature vectors. That is, larger degree nodes have more contribution (i.e., feature value), and smaller degree nodes have lower feature values in the feature representation. The different scales of input feature values are problematic in training those networks that use stochastic gradient descent algorithms. To mitigate this problem, the original adjacency matrix is normalized by taking the average of corresponding neighboring nodes features as described in the paper [34]. The symmetric version of this normalization is expressed as follows

$$A_{Norm} = \tilde{D}^{\frac{-1}{2}} \tilde{A} \tilde{D}^{\frac{-1}{2}} \tag{1}$$

where $\tilde{A}$ (i.e. $A + I$,that enforces to include own features also) is the adjacency matrix of the network and $\tilde{D}$ is the node degree matrix of $\tilde{A}$.

3.4 Roles Extraction.

After getting a latent representation of the graph from the neural network, we have a feature matrix $Z \in \mathbb{R}^{|V| \times F_2}$, the next step of our algorithm is to generate a rank r approximation $PQ \approx Z$, where $P \in \mathbb{R}^{|V| \times r}$ represents the node-roles relationship and $Q \in \mathbb{R}^{r \times F_2}$ define how each identified roles contributes to estimated feature values. To do this, we have used Non-negative Matrix Factorization as it aims to find two non-negative factor matrices which simplify the interpretation of the node-roles relationship. Since we have not defined the number of roles required, we decided to use the Minimum Description Length criterion [37], to find the optimal number on roles r as described in the paper that results in the best compression. Mathematically, it can be written as,

$$minimize \|Z - PQ\|_F^2 \, w.r.t. \, P, Q \, s.t. \, P, Q \geq 0 \tag{2}$$

In the last step, we have added both the feature matrix Z and node-roles matrix P such that $F = Z + P$, where $F \in \mathbb{R}^{|V| \times (F_2 + r)}$ is overall features set that is derived from the dataset. Finally, for each node pair, cosine similarity [35] index is used to find similarity between them. Once the score of non-observed links is available in the sorted order, we can compute the Area under the Operating receiver characteristics (AUROC) and average precision to evaluate the accuracy of our approach.

4 Experimental Study

4.1 Evaluation metrics

Consider a simple undirected graph/network $G(V, E)$ where V characterizes a vertex-set and E, the edge-set. Although a simple graph is considered so, parallel edges and self-loops are not permitted. In a simple graph, a universal set U

contains a total $\frac{|V|(|V|-1)}{2}$ edges, where $|V|$ represents the size of the vertex-set in the graph. $(U - E)$ number of edges is termed as the set of nonexistent links, some of which may be missing that may appear in the future. Finding out such missing links is the aim of link prediction [14]. The accuracy of an algorithm can be tested by partitioning the set of observed links E into two sets. E^T, a training set about which we know at all, and a test set (or validation set), E^P in which there are edges which is not present in the training set. Therefore, $E^T \cup E^P = E$ and $E^T \cap E^P = \phi$ with this strategy, it may be possible that some edges may not ever be chosen in the test set or others could be repeated, which results in statistical bias. This problem is overcome by a procedure of sampling known as K-fold cross-validation. To calculate the accuracy of algorithm, generally, two metrics are used: AUROC [40], and precision [41], [42]. Based on the above definitions, the following observations can be made in a graph:
Total possible links in the graph $= U$,
Existent links $= E$,
Non-existent links $= U - E$,
Observed links $= E^T =$ Training set,
Non-Observed links $= U - E^T$,
Missing links $= E^P =$ Test set.

Area under the Operating receiver characteristics (AUROC) Given a ranking of total non-observed links, the term AUROC is estimated as the likelihood that a chosen missing link is given a higher score than a randomly chosen non-existent link. Each time two edges are selected randomly one from each set and compared their scores. Then, AUROC can be calculated using the following expression:

$$AUROC = \frac{n_1 + 0.5 \times n_2}{n} \tag{3}$$

where, n is total independent comparisons, n_1 is number of times the missing link with a higher score,n_2 is number of times they have same score. The standard value of AUROC should be 0.5 which will be possible under an independent and identical distribution. A score greater than 0.5 represents improved accuracy.

Precision Given the ranking of non-observed links, precision can be characterized as the proportion of relevant items to the number of items chosen i.e.,

$$Precision = \frac{L_r}{L} \tag{4}$$

where, L represents predicted links having top scores, and L_r, the number of predicted links which are correct.

4.2 Datasets

The performance of the proposed method has been evaluated on twelve real-world network datasets collected from diverse areas.

- Karate[1] [43]: A friendship network of 34 members of karate club at a US university.
- Dolphins[1] [44]: A social network of dolphins living in Doubtful Sound in New Zealand.
- Lesmiserables[1] [45]: Co appearance network of characters of the novel LesMiserables.
- Adjnoun[1] [46]: Adjacency networkof common adjectives and nouns in the novel David Copperfield by Charles Dickens.
- Football1 [47]: American football games network played between Division IA colleges during regular season Fall 2000.
- Celegansneural[1] [49]: A neural network of C. elegans compiled by D. Watts and S. Strogatz in which each node refers a neuron and, an edges joining two neurons either by a synapse or a gap junction.
- Netscience[1] [46] is a co-authorship network of scientists working on network theory and experiment compiled by Newman in 2006.
- Political bolgs[1] [5] is a directed network of hyperlinks in political blogs of US election 2004.
- Jazz[2] [48] is the collaboration network of jazz musicians.
- Usair97[3] is an airline network of US where a node represents an airport and an edge shows the connectivity between two airports.
- Facebook[4] [50] is social network of user profiles and network data extracted from 10 ego-networks.
- Ca-GrQc[4] is collaboration network from the e-print arXiv of General Relativity and Quantum Cosmology.

Table 1 shows some basic topological properties of the considered networks datasets.$|V|$ and $|E|$ are the total numbers of nodes and edges of the networks respectively.$\langle D \rangle$ represents node pairs average shortest distance, $\langle K \rangle$ the average degree and $\langle C \rangle$ the average clustering coefficient of the network. r and H are the coefficient of assortativity and degree of heterogeneity respectively.

4.3 Baseline methods

- Common Neighbor(CN). In a given network or graph, the size of common neighbors for a given pair of nodes x and y is calculated as the size of intersection of

[1] http://www-personal.umich.edu/ mejn/netdata/
[2] http://konect.uni-koblenz.de/networks/
[3] http://vlado.fmf.uni-lj.si/pub/networks/data/
[4] https://snap.stanford.edu/data/

Table 1: Topological informations of real-world network datasets

Datasets	$\|V\|$	$\|E\|$	$\langle D\rangle$	$\langle K\rangle$	$\langle C\rangle$	r	H
Karate	34	78	2.337	4.588	0.570	-0.475	1.693
Dolphins	62	159	3.302	5.129	0.258	-0.043	1.326
Lesmiserables	77	254	2.606	6.597	0.573	-0.165	1.827
Adjnoun	112	425	2.512	7.589	0.172	-0.129	1.814
Football	115	613	2.486	10.661	0.403	0.162	1.006
Jazz	198	2742	2.235	27.697	0.620	0.020	1.395
Celegansneural	297	2148	2.447	14.456	0.308	-0.163	1.800
Usair97	332	2126	2.738	12.807	0.749	-0.207	3.463
Political blogs	1490	16718	2.738	22.440	0.361	-0.221	3.621
Netscience	1589	2742	5.823	3.451	0.878	0.461	2.010
Facebook	4039	88234	3.693	43.691	0.617	0.063	2.439
Ca-GrQc	5242	14496	6.049	5.531	0.687	0.659	3.051

the two nodes neighborhoods.

$$S(x,y) = |\Gamma(x) \cap \Gamma(y)| \tag{5}$$

where $\Gamma(x)$ and $\Gamma(y)$ are neighbors of the node x and y respectively.The likelihood of the existence of a link between x and y increases with the number of common neighbors between them. In a collaboration network, Newman calculated this quantity and demonstrated that the probability of collaboration between two nodes depends upon the common neighbors of the selected nodes. Kossinets [52] and Neal [53] investigated a large social network and recommended that two students are more likely to be friends who are having numerous common friends. It has been observed that the common neighbor approach performs well on most real-world networks and beats other complex methods.

– Jaccard Coefficient(JC). The Jaccard coefficient is defined as the probability of selection of common neighbors of pairwise vertices from all the neighbors of either vertex.

$$S(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{6}$$

– Resource Allocation (RA). Consider two non-adjacent vertices x and y. Suppose node x sends some resources to y through the common nodes of both x and y then the similarity between the two vertices is computed in terms of resources sent from x to y. This is expressed mathematically as

$$S(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \tag{7}$$

– Preferential Attachment(PA). The idea of preferential attachment is applied to generate a growing scale-free network. The term growing represents the incremental nature of nodes over time in the network. The likelihood incrementing new connection associated with a node x is proportional to k_x, the degree of the node. Preferential attachment score between two nodes x and y can be computed as

$$S(x, y) = k_x * k_y \tag{8}$$

- Node Clustering Coefficient(CCLP). This index is also based on the clustering coefficient property of the network in which the clustering coefficients of all the common neighbors of a seed node pair are computed and summed to find the final similarity score of the pair. Mathematically, this index can be expressed as follows

  $$S(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} C(z) \tag{9}$$

  where

  $$C(z) = \frac{t(z)}{k_z(k_z - 1)} \tag{10}$$

  and k_z is the degree of node z and $t(z)$ is the total triangles passing through the node z.
- CARIndex. CAR-based indices are presented based on the assumption that the link existence between two nodes increases if their common neighbors are members of local community (LCP theory) [56].

  $$S(x, y) = CN(x, y) \times \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{|\gamma(z)|}{2} \tag{11}$$

  where $CN(x, y)$ is common neighbour of (x, y) and $\gamma(z)$ is the subset of neighbors of node z that are also common neighbors of x and y.
- Katz Index. It directly aggregates over all the paths between x and y and dumps exponentially for longer paths to penalize them. It can be expressed mathematically as

  $$S(x, y) = \sum_{l=1}^{\infty} \beta^l |paths_{x,y}^{<l>}| \tag{12}$$

  where, $paths_{x,y}^{<l>}$ is considered as the set of total l length paths between x and y, β is a damping factor that controls the path weights.
- Node2vec(N2V). This is a node embedding technique where it learns a low dimensional continuous representation of nodes in a graph with the objective of preserving the neighborhood structure.

Table 2: Avg. Precision

	CN	JC	RA	PA	CCLP	CAR	Katz	N2V	ALP
Karate	0.148878	0.130261	0.069634	0.134948	0.162239	0.017582	0.260558	0.041504	**0.653197**
Dolphins	0.179974	0.205618	0.050494	0.082726	0.189028	0.007007	0.110008	0.024493	**0.610708**
Lesmiserables	0.249618	0.366998	0.249818	0.085681	0.256686	0.148050	0.085832	0.146678	**0.701639**
Adjnoun	0.079462	0.076121	0.018148	0.067761	0.084578	0.008218	0.105588	0.011761	**0.242644**
Football	0.459779	0.447029	0.111022	0.092703	0.468967	0.061658	0.178761	0.114445	**0.681193**
Jazz	0.326022	0.367671	0.319647	0.106431	0.328317	0.351645	0.263667	0.086058	**0.691399**
Celegansneural	0.118607	0.134162	0.036424	0.048702	0.146307	0.011771	0.051729	0.018931	**0.743431**
Usair97	0.126467	0.271941	0.231041	0.263769	0.141365	0.204706	0.050021	0.029793	**0.370255**
Political blogs	0.079613	**0.146614**	0.061253	0.015059	0.088426	0.063305	0.021715	0.008462	0.106545
Netscience	0.392733	0.433671	0.128983	0.002171	0.430501	0.108762	0.204778	0.081755	**0.587527**
Facebook	0.244687	0.172462	**0.440321**	0.019037	0.238218	0.235341	0.002152	0.125269	0.314704
Ca-GrQc	0.223600	0.063274	0.133345	0.019163	0.253608	0.217245	0.046797	0.048952	**0.381180**

Table 3: AUROC

	CN	JC	RA	PA	CCLP	CAR	Katz	N2V	ALP
Karate	0.693750	0.628375	0.757500	0.760375	0.656438	0.550500	0.611500	0.721125	**0.897395**
Dolphins	0.745418	0.769799	0.822843	0.726425	0.618911	0.357150	0.826348	0.750734	**0.882697**
Lesmiserables	0.889440	0.871515	**0.934390**	0.699676	0.888291	0.695770	0.912502	0.854812	0.906194
Adjnoun	0.665667	0.568315	0.647079	0.735380	0.653959	0.450198	0.658491	0.613725	**0.743431**
Football	0.873762	0.859288	0.854359	0.252409	0.813651	0.585301	0.854977	0.862271	**0.879024**
Jazz	0.948143	0.959044	**0.963302**	0.789540	0.955104	0.931445	0.452756	0.873276	0.909838
Celegansneural	0.815419	0.792798	0.848494	0.735148	**0.872517**	0.450223	0.416356	0.795693	0.720588
Usair97	**0.958332**	0.914826	0.946785	0.905626	0.957295	0.772429	0.50310	0.884882	0.838057
Political blogs	**0.941012**	0.907954	0.939749	0.934223	0.938549	0.739912	0.345143	0.866696	0.781857
Netscience	0.944599	0.953620	0.944769	0.639099	0.897433	0.532846	0.939401	0.892410	**0.966444**
Facebook	0.991824	0.989581	**0.995177**	0.832809	0.992504	0.944786	0.492362	0.991560	0.894836
Ca-GrQc	0.921563	**0.929400**	0.913091	0.741728	0.892601	0.605524	0.718201	0.908955	0.849303

4.4 Experiments Result Analysis

Accuracy Analysis: Table 2 shows the average precision results of the proposed method ALP with the baseline methods. Best accuracy values are shown in boldface against each network. We observe that the proposed method gives the best result on ten out of twelve network datasets, as shown in the table. Our method performs much better on all the datasets except Political blogs and Facebook dataset; RA is the best performer on Facebook, and JC is best on the Political blogs dataset. However, ALP shows the second-best result in both the datasets.

The AUROC results of the proposed and baseline methods are shown in Table 3. ALP performs best on karate, dolphins, Adjnoun, Football, and Netscience networks. RA best performs on Lesmiserables, Jazz, and Facebook networks, while CN is the best performer index on Usair97 and Political blogs networks. Jaccard (JC) shows the best result on Ca-GrQc that are collaboration networks of scientists in computer science and general relativity. CCLP is best on Celegansneural network.

Robustness Analysis: Figure 2. shows the robustness measure of the existing and the proposed ALP method. The figure presents the effects of random noise (i.e., links are randomly added to the network) and random removal links. This concept is well explained in the Zhang, P. et al. [58] work on robustness under noisy environments. The parameter "$Ratio''$ on the X-axis defines the fraction of noisy links that are added or deleted to/from the training data as described in the above work. Positive values of this parameter represent a fraction of added links to the training data, and negative values represent a fraction of deleted links from the training data. Figure 2. shows the dependence of AUROC on these fraction of added and removal links.

ALP shows the best robustness with higher accuracy compared to the baseline methods against both added and deleted links on Karate and Dolphin networks [Fig.2a and 2c]. On Lesmiserables data, ALP shows better AUROC after the CN, JC, and CCLP; however, it shows the least fluctuation (highly robust) in the AUROC values [Fig. 2b]. It is the average performing method on Adjnoun, Jazz, Usair97, and Netscience datasets with AUROC values lower than CN, JC, and CCLP on Jazz Usair97 and Netscience [Figs. 2d, 2f and 2h], moreover it shows better robustness for both added and deleted links as shown in the figures. PA and Node2Vec are the best performing indices on Adjnoun and Netscience data, respectively. ALP shows the comparable result on the Football dataset [Fig. 2e]. One thing to note that the fluctuation of the AUROC values for random deleted links is greater than randomly added links, which is similar to the work [58]. In other words, random links deletion are more vulnerable to link prediction. Due to computational issues, robustness results of the remaining datasets are not shown.

Statistical Test: In this experiment, we conduct a statistical test [59] to show the significant difference between the proposed method (ALP) with the baseline methods. We perform the Friedman test [60], [61], to analyze whether there is a significant difference among multiple methods. It is a non-parametric counterpart of the repeated measures ANOVA. If the test result shows a significant difference, we further applied post hoc analysis to check the degree of rejection of each hypothesis. For the post hoc analysis, several methods are available in the literature, and we applied the post hoc counterpart of the Friedman test known as the Post hoc Friedman Conover method.

The Friedman test results for both average precision (Avg. Precision) and area under the ROC curve (AUROC) are shown in Table 4. The observed test values of the Friedman test for both Avg. Precision and AUROC are 60.222 and 35.956 which are greater than the corresponding χ_2 value (i.e., $\chi_2(c, D_f)$). With the confidence interval α =0.05 and degree of freedom $D_f = 8$, χ_2 value is 15.51, obtained from the χ_2 table available in the literature. This results in the rejection of the null hypothesis, as shown in the last column of the table. This test confirms that there is a significant difference among the methods for both the accuracy measures. The proposed method ALP is considered as the control algorithm in the post hoc analysis.

5 Conclusion

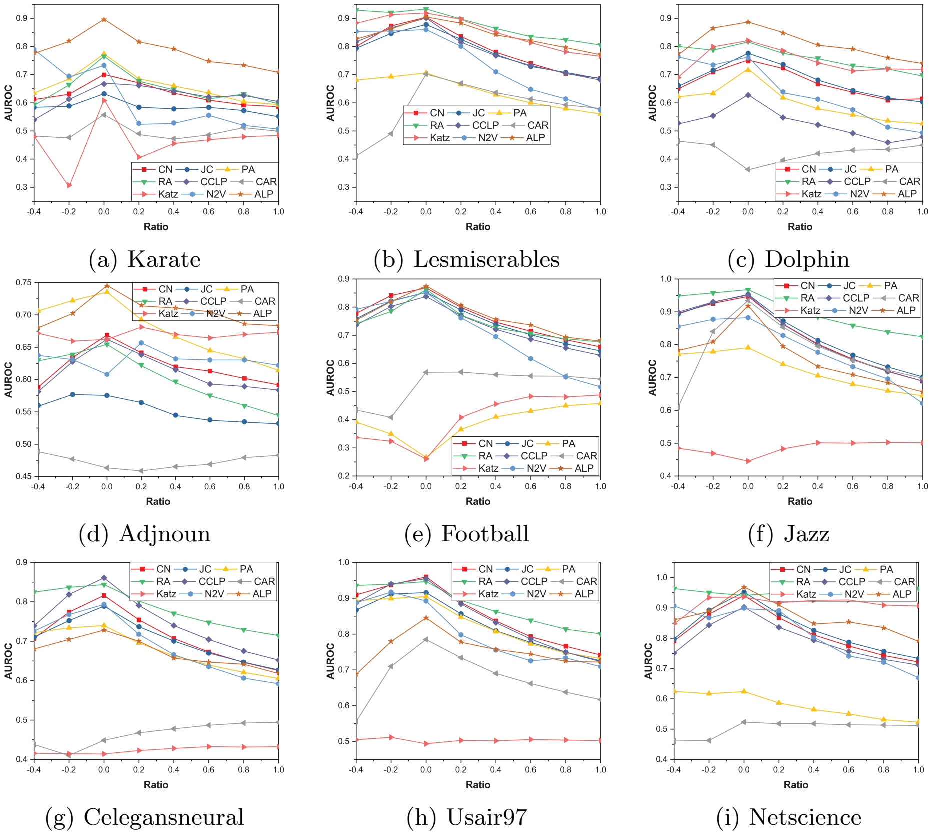
In this work, we incorporate an unsupervised framework of deep learning viz., graph autoencoder for link prediction in networks. Employing this architecture, useful latent representation of nodes is extracted and using these node embeddings, and the similarity matrix is computed for all node pairs. Moreover, missing links are identified based on this similarity matrix. The proposed method (ALP) is evaluated on average precision and AUROC, which show its superiority over the baseline methods. Robustness analysis against the noisy links (added or deleted) is shown, which represents the effectiveness of the proposed method.

6 Acknowledgement

We want to thank Ajay Kumar and Dr.Bhaskar Biswas for helpful discussions and suggestions during the early phases of the research. We would also like to thank the anonymous reviewers for sharing their helpful suggestions and Chaitanya Bhatia and Karan Siwach, for their valuable feedback.

Table 4: The Friedman test on Avg. Precision and area under the ROC Curve

	Dataset	IS-value									Test value	State Result
		CN	JC	RA	PA	CCLP	CAR	Katz	N2V	ALP	F_f	Is $F_f > \chi^2$?
Avg. Precision	Karate	0.148878	0.130261	0.069634	0.134948	0.162239	0.017582	0.260558	0.041504	0.653197	60.222	Null Hypothesis Rejected
	Dolphins	0.179974	0.205618	0.050494	0.082726	0.189028	0.007007	0.110008	0.024493	0.610708		
	Lesmiserables	0.249618	0.366998	0.249818	0.085681	0.256686	0.148050	0.085832	0.146678	0.701639		
	Adjnoun	0.079462	0.076121	0.018148	0.067761	0.084578	0.008218	0.105588	0.011761	0.242644		
	Football	0.459779	0.447029	0.111022	0.092703	0.468967	0.061658	0.178761	0.114445	0.681193		
	Jazz	0.326022	0.367671	0.319647	0.106431	0.328317	0.351645	0.263667	0.086058	0.691399		
	Celegansneural	0.118607	0.134162	0.036424	0.048702	0.146307	0.011771	0.051729	0.018931	0.743431		
	Usair97	0.126467	0.271941	0.231041	0.263769	0.141365	0.204706	0.050021	0.029793	0.370255		
	Political blogs	0.079613	0.146614	0.061253	0.015059	0.088426	0.063305	0.021715	0.008462	0.106545		
	Netscience	0.392733	0.433671	0.128983	0.002171	0.434501	0.108762	0.204778	0.081755	0.587527		
	Facebook	0.244687	0.172462	0.440321	0.019037	0.238218	0.235341	0.002152	0.125269	0.314704		
	Ca-GrQc	0.223600	0.063274	0.133345	0.019163	0.253608	0.217245	0.046797	0.048952	0.381180		
AUROC	Karate	0.693750	0.628375	0.757500	0.760375	0.656438	0.550500	0.611500	0.721125	0.897395	35.956	Null Hypothesis Rejected
	Dolphins	0.745418	0.769799	0.822843	0.726425	0.618911	0.357150	0.826348	0.750734	0.882697		
	Lesmiserables	0.889440	0.871515	0.934390	0.699676	0.888291	0.695770	0.912502	0.854812	0.906194		
	Adjnoun	0.665667	0.568315	0.647079	0.735380	0.653959	0.450198	0.658491	0.613725	0.743431		
	Football	0.873762	0.859288	0.854359	0.252409	0.813651	0.585301	0.854977	0.862271	0.879024		
	Jazz	0.948143	0.959044	0.963302	0.789540	0.955104	0.931445	0.452756	0.873276	0.909838		
	Celegansneural	0.815419	0.792798	0.848494	0.735148	0.872517	0.450223	0.416356	0.795693	0.720588		
	Usair97	0.958332	0.914826	0.946785	0.905626	0.957295	0.772429	0.500310	0.884882	0.838057		
	Political blogs	0.941012	0.907954	0.939749	0.934223	0.938549	0.739912	0.345143	0.866696	0.781857		
	Netscience	0.944599	0.95362	0.944769	0.639099	0.897433	0.532846	0.939401	0.892410	0.966444		
	Facebook	0.991824	0.989581	0.995177	0.832809	0.238218	0.944786	0.492362	0.991560	0.894836		
	Ca-GrQc	0.921563	0.929400	0.913091	0.741728	0.892601	0.605524	0.718201	0.908955	0.849303		



(a) Karate (b) Lesmiserables (c) Dolphin

(d) Adjnoun (e) Football (f) Jazz

(g) Celegansneural (h) Usair97 (i) Netscience

Fig. 2: Robustness

References

1. D. Liben-Nowell, J. Kleinberg, “The Link-prediction Problem for Social Networks,” J. Am. Soc. Inf. Sci. Technol, 2007.
2. Z. Huang, X. Li, H. Chen, “Link prediction approach to collaborative filtering,” in: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL ’05, pp. 141–142.
3. E. Airodi, D. Blei, E. Xing, S. Fienberg, “Mixed membership stochastic block models for relational data, with applications to protein-protein interactions,” Proceedings of International Biometric Society-ENAR Annual Meetings (2006).
4. L. Lu, T. Zhou, “Link prediction in complex networks: A survey,” CoRRabs/1010.0725 (2010).arXiv:1010.0725.
5. L. A. Adamic, N. Glance, “The political blogosphere and the 2004 U.S.election: Divided they blog,” in: Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD ’05, ACM, New York, NY, USA,2005, pp. 36–43.doi:10.1145/1134271.1134277.
6. T. Zhou, L. Lu, Y.-C. Zhang, “Predicting missing links via local information,” European Physical Journal B 71 (2009) 623–630.doi:10.1140/epjb/e2009-00335-8.
7. A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek, “Evolution of the social network of scientific collaborations,” Physica A Statistical Mechanics and its Applications 311 (2002) 590–614.doi:10.1016/S0378-4371(02)00736-7.
8. L. Katz, “A new status index derived from sociometric analysis,” Psychometrika 18 (1) (1953) 39–43.
9. W. Liu, L. L u,“ Link prediction based on local random walk”, EPL(Europhysics Letters) 89 (5) (2010)58007.
10. S. Brin, L. Page, “The anatomy of a large-scale hypertextual web search engine,” in: Proceedings of the Seventh International Conference on World Wide Web 7, WWW7, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, 1998, pp. 107–117.
11. E. A. Leicht, P. Holme, M. E. J. Newman, “Vertex similarity in networks,” Phys. Rev. E 73 (2006) 026120.doi:10.1103/PhysRevE.73.026120
12. H. Tong, C. Faloutsos, J.-Y. Pan, “Fast random walk with restart and its applications,” in: Proceedings of the Sixth International Conference on Data Mining, ICDM ’06, IEEE Computer Society, Washington, DC, USA, 2006,pp. 613–622.doi:10.1109/ICDM.2006.70.
13. V. Mart ınez, F. Berzal, J.-C. Cubero,“ A survey of link prediction in complex networks,” ACM Comput. Surv. 49 (4) (2016) 69:1–69:33.doi:10.1145/3012704.
14. L. Lu, C.-H. Jin, T. Zhou, “Similarity index based on local paths for link prediction of complex networks,” Phys. Rev. E 80 (2009) 046122.doi:10.1103/PhysRevE.80.046122.
15. M. A. Hasan, V. Chaoji, S. Salem, M. Zaki, “Link prediction using supervised learning,” in: In Proc. of SDM 06 workshop on Link Analysis,Counter terrorism and Security, 2006.
16. J. R. Doppa, J. Yu, P. Tadepalli, L. Getoor, “Learning algorithms for link prediction based on chance constraints,” in: Proceedings of the 2010European Conference on Machine Learning and Knowledge Discovery in Databases: Part I, ECML PKDD’10, Springer-Verlag, Berlin, Heidelberg,2010, pp. 344–360.
17. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” J. Mach. Learn. Res. 11 (2010) 3371–3408.
18. A. Krizhevsky, I. Sutskever, G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1,NIPS’12, Curran Associates Inc., USA, 2012, pp. 1097–1105.
19. M. E. J. Newman, “Clustering and preferential attachment in growing networks,” Phys. Rev. E 64 (2001) 025102.doi:10.1103/PhysRevE.64.025102.
20. H. Cai, V. W. Zheng, K. C. Chang,“A comprehensive survey of graph embedding: Problems, techniques, and applications,” IEEE Transactionson Knowledge and Data Engineering 30 (9) (2018) 1616–1637.doi:10.1109/TKDE.2018.2807452.

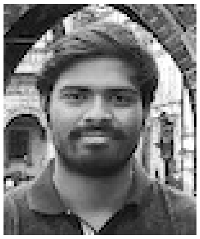
21. M. Belkin, P. Niyogi, “Laplacian eigenmaps and spectral techniques forembedding and clustering,” in: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS’01, MIT Press, Cambridge, MA, USA, 2001, pp. 585–591.
22. A. Grover, J. Leskovec, “Node2vec: Scalable feature learning for networks,” in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, ACM, New York, NY,USA, 2016, pp. 855–864.doi:10.1145/2939672.2939754.
23. S. Mehran Kazemi, D. Poole, “Simple Embedding for Link Prediction in Knowledge Graphs,” ArXiv e-prints (Feb. 2018).arXiv:1802.04868.
24. B. Perozzi, R. Al-Rfou, S. Skiena, “Deepwalk: Online learning of social representations,” in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’14, ACM,New York, NY, USA, 2014, pp. 701–710.doi:10.1145/2623330.2623732.
25. S. T. Roweis, L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” Science 290 (5500) (2000) 2323–2326.
26. H. Chen, B. Perozzi, Y. Hu, S. Skiena, “HARP: hierarchical representation learning for networks,” CoRR abs/1706.07845 (2017).arXiv:1706.07845.
27. D. Wang, P. Cui, W. Zhu, “Structural deep network embedding,” in:Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, ACM, New York, NY,USA, 2016, pp. 1225–1234.doi:10.1145/2939672.2939753.
28. S. Cao, W. Lu, Q. Xu, “Deep neural networks for learning graph representations,” in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI’16, AAAI Press, 2016, pp. 1145–1152.
29. T. N. Kipf, M. Welling, “Variational graph auto-encoders,” CoRRabs/1611.07308 (2016).arXiv:1611.07308.
30. M. Zhang, Y. Chen, “Link prediction based on graph neural networks,” CoRRabs/1802.09691 (2018).arXiv:1802.09691.
31. H. Wang, X. Shi, D. Yeung, “Relational deep learning: A deep latent variable model for link prediction,” in: AAAI, 2017, pp. 2688–2694.
32. X. Li, N. Du, H. Li, K. Li, J. Gao, A. Zhang, “A deep learning approach to link prediction in dynamic networks,” in: Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia,Pennsylvania, USA, April 24-26, 2014, 2014, pp. 289–297.doi:101137/1.9781611973440.33.
33. A. Dadu, A. Kumar, H. K. Shakya, S. K. Arjaria, B. Biswas, “A study of link prediction using deep learning,” in: Advanced Informatics for Computing Research, Springer Singapore, Singapore, 2019, pp. 377–385.
34. T. N. Kipf, M. Welling, “Semi-supervised classification with graph convolutional networks,” CoRR abs/1609.02907 (2016).arXiv:1609.02907.
35. G. Salton, M. J. McGill, “Introduction to Modern Information Retrieval,” McGraw-Hill, Inc., New York, NY, USA, 1986.
36. P. V. Tran, “Learning to make predictions on graphs with autoencoders,” CoRR abs/1802.08352 (2018).arXiv:1802.08352.
37. B. Singh, S. De, Y. Zhang, T. Goldstein, G. Taylor, “Layer-specific adaptive learning rates for deep networks,” CoRR abs/1510.04609 (2015).arXiv:1510.04609.
38. T. Schaul, S. Zhang, Y. LeCun, “No More Pesky Learning Rates,” arXive-prints (2012) arXiv:1206.1106arXiv:1206.1106.
39. J. Shu, Q. Chen, L. Liu, L. Xu, “A link prediction approach based on deep learning for opportunistic sensor network,” International Journal of Distributed Sensor Networks 13 (4) (2017) 1550147717700642.arXiv:https://doi.org/10.1177/1550147717700642.
40. J. A. Hanley, B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (roc) curve,” Radiology 143 (1) (1982)29–36.doi:10.1148/radiology.143.1.7063747.
41. S. Geisser, “Predictive inference: An introduction chapman and hall,” NewYork (1993).

42. J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, “Evaluating collaborative filtering recommender systems,” ACM Trans. Inf. Syst. 22 (1)5–53.
43. W. Zachary, “An information flow model for conflict and fission in small groups,” Journal of Anthropological Research 33 (1977) 452–473.
44. D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten,S. M. Dawson, “The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations,” Behavioral Ecology and Sociobiology 54 (4) (2003) 396–405.doi:10.1007/s00265-003-0651-y.
45. D. E. Knuth, “The Stanford Graph Base: A Platform for Combinatorial Computing,” ACM, New York, NY, USA, 1993.
46. M. E. J. Newman, “Finding community structure in networks using the eigen vectors of matrices,” Phys. Rev. E 74 (2006) 036104.doi:10.1103/PhysRevE.74.036104.
47. M. Girvan, M. E. J. Newman, “Community structure in social and biological networks,” 99 (12) (2002) 7821–7826.doi:10.1073/pnas.122653799.
48. P. Gleiser, L. Danon, “Community Structure in Jazz,” eprint arXiv:cond-mat/0307434.
49. D. J. Watts, S. H. Strogatz, “Collective dynamics of ’small-world’ networks,” Nature 393 (6684) (1998) 440–442.doi:10.1038/30918.
50. J. G. White, E. Southgate, J. N. Thomson, S. Brenner, “The structure of the nervous system of the nematode caenorhabditis elegans,” Philos Trans RSoc Lond B Biol Sci 314 (1165) (1986) 1–340.
51. J. McAuley, J. Leskovec, “Learning to discover social circles in ego networks,” in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS’12, Curran AssociatesInc., USA, 2012, pp. 539–547.
52. G. Kossinets, “Effects of missing data in social networks,” Social Networks28 (3) (2006) 247 – 268.doi:https://doi.org/10.1016/j.socnet.2005.07.002.
53. J. W. Neal, “Kracking the missing data problem: applying krackhardt’s cognitive social structures to school-based social networks,” Sociology of Education 81 (2) (2008) 140–162.
54. P. Jaccard, “Distribution de la flore alpine dans le bassin des dranses et dansquelques r egions voisines”, Bull Soc Vaudoise Sci Nat 37 (1901) 241–272.
55. “Link prediction with node clustering coefficient,” Physica A: Statistical Mechanics and its Applications 452 (2016) 1 – 8.doi:https://doi.org/10.1016/j.physa.2016.01.038.
56. A. Kumar, S. S. Singh, K. Singh, B. Biswas, “Level-2 node clustering coefficient-based link prediction,” Applied Intelligence 49 (7) (2019) 2762–2779.doi:10.1007/s10489-019-01413-8.
57. C. V. Cannistraci, G. Alanis-Lobato, T. Ravasi, “From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks,” Scientific Reports 3 (2013) 1613.doi:10.1038/srep01613.
58. X. Glorot, Y. Bengio, “Understanding the difficulty of training deep feed forward neural networks,” in: Y. W. Teh, M. Titterington (Eds.),Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Vol. 9 of Proceedings of Machine Learning Research, PMLR, Chia Laguna Resort, Sardinia, Italy, 2010, pp. 249–256.
59. P. Zhang, X. Wang, F. Wang, A. Zeng, J. Xiao, “Measuring the robustness of link prediction algorithms under noisy environment,” Scientific Reports6 (2016) 18881.doi:10.1038/srep18881.
60. J. Demˇsar, “Statistical comparisons of classifiers over multiple data sets,” J.Mach. Learn. Res. 7 (2006) 1–30.
61. M. Friedman, “The use of ranks to avoid the assumption of normality implicit in the analysis of variance,” Journal of the American Statistical Association32(200)(1937)675–701.doi:10.1080/01621459.1937.10503522.
62. M. Friedman, “A comparison of alternative tests of significance for the problem of m rankings,” Ann. Math. Statist. 11 (1) (1940) 86–92.doi:10.1214/aoms/1177731944.
63. Henderson, Keith and Gallagher, Brian and Eliassi-Rad, Tina and Tong, Hanghang and Basu, Sugato and Akoglu, Leman and Koutra, Danai and Faloutsos, Christos and Li, Lei, “Rolx: structural role extraction & mining in large graphs,” Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012.

64. Wang W, Tang M, Jiao P., “A unified framework for link prediction based on non-negative matrix factorization with coupling multivariate information,” PLoS One. 2018;13(11):e0208185,2018, doi:10.1371/journal.pone.0208185.
65. Z. Wu and Y. Chen, “Link prediction using matrix factorization with bagging,” 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, 2016, pp. 1-6, doi: 10.1109/ICIS.2016.7550942.

Authors

Aman Gupta and **Yadul Raghav**, we received a Bachelor’s and Master’s degree from the Indian Institute of Technology (BHU), Varanasi, majoring in Computer Science and Engineering. Our interest lies in the field of machine learning, deep learning, and data mining.



Aman Gupta

Yadul Raghav

INVERSE SPACE FILLING CURVE PARTITIONING APPLIED TO WIDE AREA GRAPHS

Cyprien Gottstein[1], Philippe Raipin Parvedy[1], Michel Hurfin[2], Thomas Hassan[1] and Thierry Coupaye[1]

[1]TGI-OLS-DIESE-LCP-DDSD, Orange Labs, Cesson-Sevigné, France
[2]Univ Rennes, INRIA, CNRS, IRISA, 35000RENNES, France

ABSTRACT

The most recent developments in graph partitioning research often consider scale-free graphs. Instead we focus on partitioning geometric graphs using a less usual strategy: Inverse Space-filling Partitioning (ISP). ISP relies on a space filling curve to partition a graph and was previously applied to graphs essentially generated from Meshes. We extend ISP to apply it to a new context where the targets are now Wide Area Graphs. We provide an extended comparison with two state-of-the-art graph partitioning streaming strategies, namely LDG and FENNEL. We also propose customized metrics to better understand and identify the use cases for which the ISP partitioning solution is best suited. Experimentations show that in favourable contexts, edge-cuts can be drastically reduced, going from more 34% using FENNEL to less than 1% using ISP.

KEYWORDS

Graph, Partitioning, Graph partitioning, Geometric partitioning, Spatial, Geography, Geometric, Space Filling Curve, SFC, ISP.

1. INTRODUCTION

Nowadays, graphs are extensively used in the IT industry. As one of the most expressive and widely used data structures, graphs have reached a scale never seen before. Along with the explosion of social networks, new marketing opportunities arose as graph knowledge about people became a very profitable resource. Many features of important services such as Google, Facebook or Amazon rely on graph analysis to provide their users with the content or experience that best satisfies their needs. As performing such analyses requires heavy computation, it is essential for those graphs to be widely distributed across a large set of machines, i.e. partitioned. As K. Andreev and H.Räcke shown, graph partitioning is a NP-Complete problem [1].

Today, a wealth of graph partitioning solutions exists with a large focus on the OnLine Analytic Processing (**OLAP**) context. Choosing the appropriate graph partitioning solution for a given graph has become a difficult task due to the wide variety of solutions. Most recent researches on graph partitioning [2] heavily focuses on scale-free graphs. Scale-free graphs, including social graphs, are graphs whose degree distribution follows a power law. **However, data spatiality tend to be neglected in these works.** This is justified because very few, if not none, scale-free graphs hold precise spatial information onto their nodes. Consequently, geometric graphs, graphs

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 223-241, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101417

whose nodes and edges are associated to spatial elements placed in a plane, and geometric based partitioning strategies have received little interest in the past decade.

Nonetheless, we believe geometric graphs and geometric partitioning will quickly regrow scientific interests with the uprising of the Internet of Things and the appeal of graph technology. Because of new services such as Microsoft Azure Digital Twins and new digital uses developed by companies working on smart-cities or smart-industry, we expect a new family of graphs will rise: geometric graphs based on real-world infrastructures with scales similar to that of social graphs. As these graphs do not exists yet, it is difficult to define them with precise characteristics. Being precise is difficult mainly because these graphs will probably be the resulting aggregate of different types of infrastructures: roads, power or water supply chain, buildings and equipment, each transformed into graphs with their associated properties. In the rest of the paper, we will refer to this family of graphs as Wide Area Graphs (**WAG**).

Some major characteristics will be common to all WAGs. WAG is a sub family of geometric graphs. Recall that geometric graphs are graphs for which vertices or edges are placed on a plane. Therefore, each node of a WAG has an associated location. Also, for a graph to be considered as a WAG it must be composed of millions of nodes and edges and cover an area as large as a country’s surface or even wider. At last, a WAG should be infrastructure related e.g. mixing power grid networks, road networks, buildings, equipment’s or even supply chains. Furthermore, it should have layers associating local objects interacting with each other and connected to a larger network. We do not include anything regarding connectivity and density properties into this high level characterisation of a WAG because these aspects may vary significantly from one WAG to another. However we envision a WAG as a highly polarized version of graphs issued from Finite Element Meshes (**FEM**).

FEM are used to simulate real world experiences as wind or water movement (see Figure 1). FEM graphs are planar graphs with varying density gradually peaking usually within a single continuous area. A planar graph induces a low connectivity ratio, compared to scale-free graphs, and a low edge Euclidean distance as edges only exists between nodes that are close to each other on the plane. Connectivity will not be limited to being planar in a WAG; it may contain nodes with high connectivity leading to crossing edges with higher Euclidean distance. In WAGs, we expect both smooth and harsh density gradation with several peaks scattered randomly across the plane, this would correspond to not-so-well shaped meshes [3]. At last, a WAG may have several nodes located exactly at the same place on the multidimensional plan.

This work addresses the problem of partitioning a WAG. Since WAGs are heavily related to geometry, we believe that geometric partitioning which has already been proposed and applied on graphs (See Section 2) is an appropriate approach for this new challenge. While those solutions are known to be computationally expensive, Pilkington and Baden came up with Inverse Space-Filling curve Partitioning (**ISP**) [4] a much cheaper algorithm, which produces partitions based on a Space-Filling Curve. A space-filling curve is a mapping from a multidimensional plane to a single dimension line. The curve is used with a depth, or zoom, to define its granularity. A space-filling curve is based on recursion and is self-similar through the levels. Those curves are known and used because of their ability to preserve spatial proximity into the single dimension space. ISP is at the core of our work as we propose an extension of the original solution to cope with the new context of WAGs. Our main goal is to determine in which cases such a solution is relevant.

This work has been done in the context of Orange research project's Thing’ In, a platform dedicated to map connected and unconnected objects of the real world into a single graph representing their interactions. Thus, Thing’ In is a spatial graph highly bound to geography and it could be seen as a prototype of a WAG. Furthermore, Thing’ In is a use-case agnostic platform.

Each of its use case will need to store, retrieve and query data in a customised manner and will most likely require optimizations to perform those actions in an acceptable speed. As our objective is to study the WAG family and not only the single specific Thing' In graph, we will also consider synthetic graphs during our experiments.

This paper brings several contributions. First, we provide an update of ISP to show how it behaves when applied to WAGs which are in essence a harder version of FEM graphs. Then, we experimentally demonstrate that edge spatial distance is the decisive factor regarding the performance of ISP. We propose a fine analysis metric and decision matrix that indicates whether to choose ISP or not as a partitioning strategy. At last, we evaluate its benefits and drawbacks through a set of metrics, mainly comparing our results against FENNEL algorithm [5].

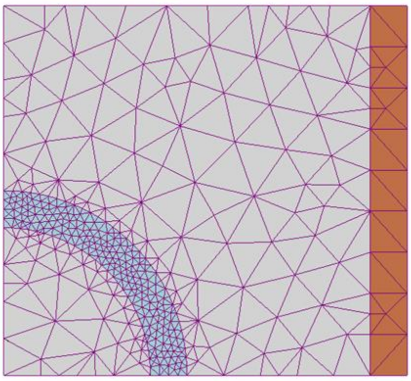


Figure 1. Illustration of a graph issued from a mesh.

2. RELATED WORK

Graph partitioning is a humongous field of research. The graph partitioning, or the balanced k-partitioning, problem is described as such: given a graph G and a number k corresponding to the number of partitions, we seek to cut G into k balanced pieces while minimizing the number of edges cut. **Edge-cut** is the default performance measure to any graph partitioning strategy; it represents the ratio of edges for which both ends are stored on separate partitions. It is important to have a minimum amount of edge-cuts as it dramatically increases edge traversal time: a basic operation for a graph algorithm. Obviously, **Load-Balancing** is also mandatory; a fairly distributed load is needed to provide horizontal scalability.

In this section, we depict some of the state-of-the-art strategies in the field of graph partitioning. We do not aim to cover the whole field of research, but instead we will limit ourselves to the most relevant research related to our target, e.g. partitioning strategies able to partition a WAG. For example, a strategy such as METIS [6], the most popular example of multi-level strategy, is considered to be extremely costly if it is intended to run on WAG scale. While METIS can be applied on any type of graphs, some strategies are specialized for geometric graphs.

A well-known approach for geometric partitioning is Recursive Coordinate Bisection [7] (RCB).The RCB method sorts the graph nodes based on the most expanded dimensional axis and then recursively bisects that axis to distribute the load evenly. This approach then evolved with random sphere [8] and partitioning based on space filling curve: Inverse Space-Filling Partitioning [4] (ISP). Random sphere strategy performs geometric sampling over the mesh to find an efficient circle center point to bisect the mesh with an even load; the algorithm is then executed recursively to reach the given number of partitions. ISP also relies on geometry but with a much cheaper approach based on SFC. Nodes which are close on the multi-dimensional space are mapped close on the single dimension, then they are grouped together and form a partition. We describe how ISP works in Section 3.

Both RCB and Random sphere are geometric strategies which could have been used to partition WAG. But given that RCB is known to yield partitioning of poor quality [9] and random sphere works best on “well-shaped” graphs, which is not what should be expected from WAGs, we will not consider those strategies for our comparisons.

More recent works aim to partition graphs based on geometry [10]–[12]. Akdogan proposes to partition geometric graphs using Voronoi tessellation whereas Volley places data on a 3D sphere representing the earth and groups data spatially close to reduce edge-cuts. These solutions work relatively well but require a lot of computations. Delling et al. strategy is based on the presence of natural cuts within a graph that minimize edge-cuts. Their solution detects such cuts to build partitions. Unfortunately the natural cut hypothesis seems unlikely to hold within a WAG.

Today, graph partitioning is considered, at least from the analytic perspective, to be divided into two categories: offline which requires storing the whole graph in memory to perform global partitioning decision and online for cheaper strategies based on local decision making. The latter is in general based on streaming, well-suited for the heavy scaling that characterizes scale-free graphs or WAGs. Single pass streaming applied to balanced graph partitioning consists of streaming every node of the graph only once and to build the partitions on the fly. For each node streamed, scores are computed assuming the node can be associated with each partition; the node is then attached to the partition with the highest score and may never leave it. Once the streaming is done, so is the partitioning as a whole. When the graph increases, this process can be resumed without the need to restart from scratch.

To the extent of our knowledge, the first study of this strategy has been proposed by Kliot and Stanton [13] in which they analysed multiple score heuristics to handle the placement of nodes inthe partitions. They determined Linear Deterministic Greedy (LDG) as the best heuristic for graph streaming partitioning. Their work have been improved with the upcoming of FENNEL [5] as it offers a generalization framework able to perform just as LDG but with extended possibilities. Even though streaming algorithms are relatively simple in their application they still offer a lot of possibilities by adjusting the formulas used to measure the load of each partition and the possible imbalance factor.

3. ISP EXTENSION

In this section we discuss how we extend ISP previous work. As said before, geometric partitioning within a multi-dimensional space is very expensive. Alternatively, ISP [4] proposes to map the multi-dimensional space on a single dimension thanks to a space filling curve leading to a simplified 1D partitioning of the curve. The curve is divided into cells and their order is defined by the curve algorithm. Each cell is a point on the curve mapped to a bounding box on the multi-dimensional space. A cell is to be assigned to a single partition and has an associated weight defined by the number of nodes it contains. Thus, the weight of a partition is equal to the sum of the weights of its cells.

When ISP was first proposed, graph partitioning streaming strategies have not yet been identified as a particular class of strategies. Therefore, ISP is classified as a geometric partitioning strategy. We believe it is fully streaming compliant and should therefore be considered as such, the only requirement is to stream the nodes in the order defined by the cells of the curve used to perform the partitioning.

In our implementation of ISP, based on the total number of nodes (denoted n) and the number of partitions (denoted k), we assign an expected capacity to each partition. Usually the capacity is the same for all the partitions and is more or less equal to n/k. This is not a mandatory requirement. We define the load of a partition as its weight divided by its capacity. During the partitioning process, a partition may be assigned fewer or more nodes than expected. In any case, the quality of the load balancing can and should be evaluated according to the load ratio of the partitions.

A partition corresponds to a set of contiguous cells of the SFC. Partitions are created sequentially. To create the first partition, we start from the first cell of the SFC and we consume the cells in the order defined by the curve. We greedily assemble cells until the partition is fully loaded (i.e., its load is no longer lower than its expected capacity). Then we move on to the next partition and the next cell. As a result, the curve is segmented into k intervals and each interval is corresponding to a partition. Partitions are continuous segment over the curve which corresponds to continuous area over the multi-dimensional space thanks to the space filling curve properties.

We have respected the original design choice of ISP: even if measuring the load of a partition has seen some evolution like using the number of edges [13] or other custom metric[5], the load metric of a partition is still based on the number of nodes it actually stores. We view the number of nodes as the most suited metric because we are targeting the OnLine Transaction Processing (**OLTP**) context. We also kept the Hilbert curve[14] as it has been proved to be the optimal space filling curve for spatial proximity preservation by Knoll and Weis [15]. At this point, whenever we mention a Space Filling Curve (SFC), we will implicitly assume it is the Hilbert SFC.

Our main contribution is about using ISP [4] to partition WAG which are harder to handle than graphs issued from meshes. Besides the properties we described for WAG, we also mentioned that nodes may overlap in a WAG.

SFCs are frequently used with adaptive refinement, a method which allows the SFC to zoom in and out depending on current needs e.g. local density for graph partitioning. With graphs where nodes never overlap e.g. graph issued from meshes, adaptive refinement guarantees that each cell will store at most one node. In our context, nodes may end up at the exact same position e.g. a building with multiple-storeys stored in a 2D plane in which case adaptive refinement will not be able to split the cell's weight. This is an important issue as instead of assembling cells which have a weight of exactly one node, we may have to cope with cells with high weight which may induce load imbalance.

Figure 2 illustrates this risk. If the cells have different weights (1, 2, 3 or 5 in the example), the consumption of the cells in the order defined by the SFC sometimes leads to achieve exact load balance (at the bottom of the figure) or, on the contrary, leads to exceed this threshold by adding a last cell with too many nodes (at the top of the figure). As such, the existence of super cells (SFC cells with extreme load) may or may not be problematic.

4. EDGE DISTANCE AND DENSITY, ISP DECIDING FACTOR

ISP can potentially be applied to any kind of graph as long as each of its nodes has position on a Euclidean space. We argue in this section that there are precise conditions for ISP to perform well. The decisive factor is the distance covered on average by an edge in proportion to the surface covered on average by a partition, we defined it as: **EDTPS** (Edge Distance To Partition Square Size).

The idea is the following: the more we increase the number of partitions, the more the space covered by each partition decreases as the space is finite and the likelier we are to produce edge-cuts using a geometric partitioning strategy such as ISP. The higher the EDTPS, the closer is an edge from being as large as the average square surface of a partition. Geometric partitioning with low EDTPS should therefore behave well.

Let G(V,E) be a directed geometric graph placed on a 2D plan. The graph G has a surface englobing polygon P composed of points ((x_i, y_j)) so that P= ((x_0, y_0), (x_1, y_1), …(x_{n-1}, y_{n-1})). The points in P are all sorted in order. Let D(e) be the Euclidean distance between the source and destination for each edge e of E. We first define the function used to convert a polygon to a surface area.

$$Area(P) = \frac{1}{2}\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i) \quad \text{Where } x_n = x_0 \text{ and } y_n = y_0$$

Equation 1. Area formula

Then, the EDTPS formula is as follows:

$$EDTPS(G,k) = \frac{\sum_{e \in E} D(e)}{|E|} * \frac{1}{\sqrt{\frac{Area(P)}{k}}}$$

Equation 2. EDTPS Formula

A graph partitioning solution is evaluated over its ability to preserve edge from being cut and how well it maintains load balance. ISP, as mentioned in Section 3, may have trouble handling graph with high local density peak, especially without adaptive refinement. To measure those potential troubles we introduce our second metrics: **CDTPC** (Cell Density To Partition Capacity).

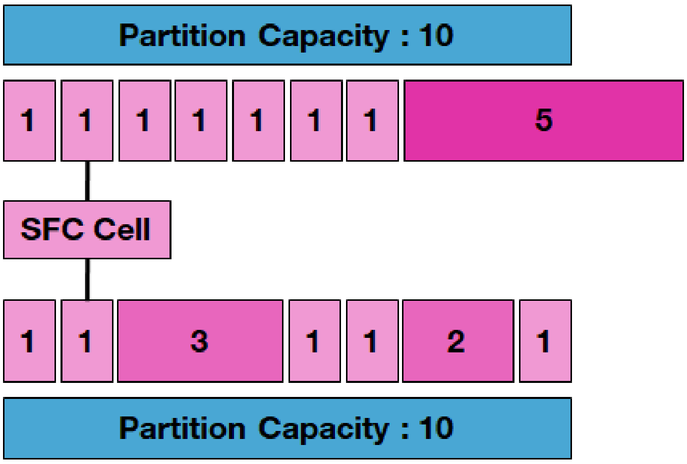


Figure 2. ISP Imbalance scenario

As shown in the second example of Figure 2, a hyper dense SFC cell may not automatically trigger imbalance, it needs to be consumed within a specific interval. This interval is the ratio between the weight of the cell and the space left in the partition. The heavier the cell, the larger the interval is to build imbalance partition. CDTPC is designed to evaluate such interval and detect in advance potential load balance troubles. The higher it is, the larger the interval will be. While a high CDTPC does not necessarily imply an imbalance problem, a small ratio most likely guarantees a very sane balance.

In order to define CDTPC we need first to compute maximum cell density which we will refer as **MCD**. To do so we map every nodes position to the SFC and retain the SFC cell which includes the most nodes. The CDTPC formula is then the following:

$$CDTPC(G,k) = \frac{MCD}{\frac{|N|}{k}}$$

Equation 3. CDTPC Formula

Note that in both formulas we need k as EDTPS and CDTPC depends on the number of partitions. It is normal since the performances of the ISP partitioning algorithm vary depending on the number of partitions.

5. Geometric Graph generator

Faced with the lack of appropriate datasets that may be close to the future WAGs (a geometric graph, covering a large area and with a high number of nodes and edges and due to the difficulty to crawl an equivalent in an open platform, we resorted to use a graph generator. There already exists geometric graph generators [16], [17]. Yet, none can provide a WAG like graph (with the desired density or connectivity characteristics). They also fail to enable multiple edge distances and they do not impose a minimum distance to build an edge between two nodes. To be able to produce synthetic WAGs with varying EDTPS and CDTPC, we need a generator that accepts parameters specifying density and edge Euclidean distance characteristics as an input. For all these reasons we have designed our own geometric graph generator.

Our generator is not suited to produce scale-free graph, it does not include the preferential attachment model [18]. Its goal is to produce graph similar to infrastructure or IoT graphs. The generator is written in Java and integrates its own R-Tree index [19]. The generation process is performed in three steps: process population density through a SFC then create and assign the nodes on the plane and finally build the edges.

It requires the following input parameters: the number of nodes and edges, a surface plane, indications on how the density of nodes may vary in the plane and on the length of the edges. The surface plane needs to be a square composed of GeoJSON coordinates. Density is configured through a set of categories, each holding a surface unit and density factor. The surface unit is used to define how much of the surface of the plane is covered with this category. The density factor defines relatively how dense is a category compared to the others. At last, edge distances are also defined through categories, each with a ratio, and a minimum and maximum possible distance. Each node on the plane is assigned the same number of starting edges, the edge ratio is used to determine how many of those edges belong to which distance categories. At the end, nodes will have the same number of outgoing edges but a different number of incoming edges.

To generate the density distribution of the node population, we first slice the plane into a grid and map each cell to an SFC. The number of columns and rows of the grid is determined through the definition of the SFC level of granularity given by the user. The surface units of each density categories are converted into relative ratio to determine how much of the SFC (and consequently of the underlying plane) will be assigned to a given category. The density factor weighted by the surface units enables us to process how many of the total nodes will fit in a given category and determine its local cell density (the number of nodes will be populated inside a SFC cell mapped to the plane).

In order to produce a representative density of the real world, the whole surface of a given density is segmented into several blocks spread across the SFC. We proceed from highest to lowest density. For each new segment, we start at a free random point on the curve and apply a random Pareto function to determine its length. The denser a category is, the harder it will be to build long continuous segments. At the end, the whole SFC is corresponding to a sequence of multiple segments assigned to a density category. It remains only to populate those segments with nodes and proceed to the edge binding. We provide a visual example at Figure 3.

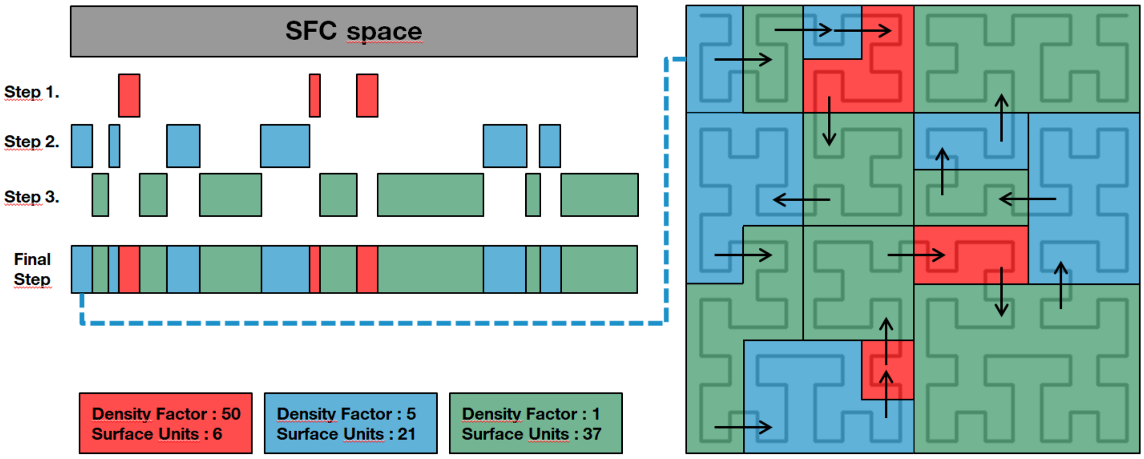


Figure 3. Density distribution of the node population

Populating the nodes only consists in taking a cell from the SFC, checking its density, and producing the corresponding number of nodes. Each node is assigned to GPS coordinates of a random point located within the cell bounding box. Edge binding requires more work. We rely on the embedded R-Tree to query the nodes using geometric squares. For each edge distance category, we form squares with a diameter twice the size of the maximum distance allowed for this edge distance category, thus large enough to contain the longest edges, and slice the whole plane in a grid manner. It enables us to perform edge binding within a limited part of the graph while still satisfying the distance requirements. To prevent any form of advantage which would lead to anomalies, randomized selection is performed each time a target node is needed to build an edge. Also, we perform several passes with offsets into the geometric square query to make sure no connected component is formed because of our approach.

6. EXPERIMENTAL SETUP

6.1. Setup

We evaluated our solution on a high processing computing machine with 92 GB of RAM and 48 CPU cores. The spatial graph generator and the metric processor were written in Java 8, everything related to visualization was done using Python 3.6. We choose to rely on the GraphX library of Spark (https://spark.apache.org/) to perform parts of our metrics; Spark is an expensive overhead in both development and computation but brings high scaling potential. Because Spark is built to work over Hadoop (See https://hadoop.apache.org/) and HDFS, we deployed a standalone HDFS service to handle the data read and produced by the metrics jobs. To evaluate the different partitioning solutions, we measure the amount of edge-cuts produced and the imbalance of the partitions. Regarding load imbalance, we use the **normalized maximum load** metric [5] (defined Section 4.1). We also include the custom heuristics EDTPS and CDTPC defined in section 4.

6.2. Dataset

We used two kinds of graphs in our experimentation: the graph of Thing' In whose characteristics are presented in the following Section 6.2.1 and synthetic WAGs produced using the graph generator already described in Section 5. We consider different WAGs with distinct characteristics as it is impossible to reduce WAGs to a single set of properties. Precisely, we aimed to first produce WAGs close to Thing'In properties and then gradually deviate to obtain a wide range of results for the application of ISP. We also wanted to find an inflexion point where ISP becomes better or worse than concurrent strategies.

6.2.1.Thing'In Graph

The Thing'In graph is an aggregate of multiple open data sources holding representation of physical objects from the real world. It is a generic graph; it holds potentially any kind of data as long as there is the appropriate definition from the semantic web to describe its content. As it is a research project not every node is associated to a GPS coordinates, we removed those to only keep a fully compliant geometric graph. As nodes were excluded, the remaining data, representing about half of the graph (~ 27.2 millions of nodes and ~27.5 millions of edges), are mainly composed by transportation network like trains, buses, roads and subways.

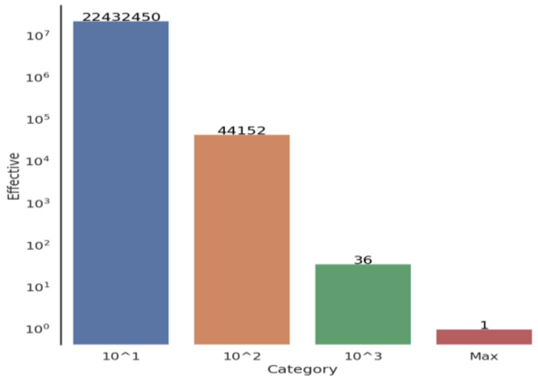


Figure 4. Thing' In Degree Distribution, blue column is for nodes with degree less than 10.

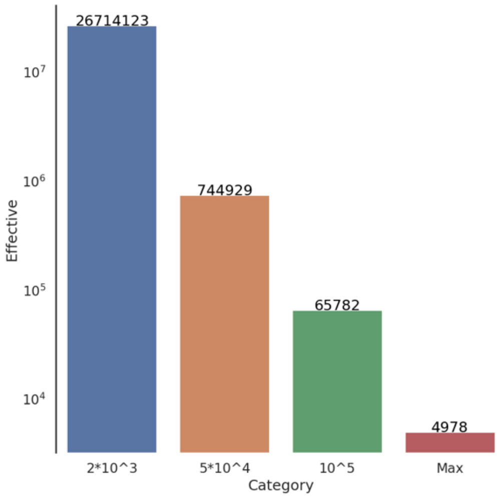


Figure 5. Thing'In edge euclidean distance distribution

As shown in Figure 3 we see that indeed, Thing'In is a spatial graph. There is no super nodes (nodes with hyper connectivity) and the most connected node represent the city of Rennes (It is the city where the Thing'In main developer team is located. They performed many experimentations related to this city and, consequently most edges of the graph are related to this city). The cumulative degree distribution function shows that Thing'In does not follow a power-law distribution. In average, the Euclidean distance covered by the edges (see Figure 4) is very short: 97,03% of the edges are shorter than 2km.

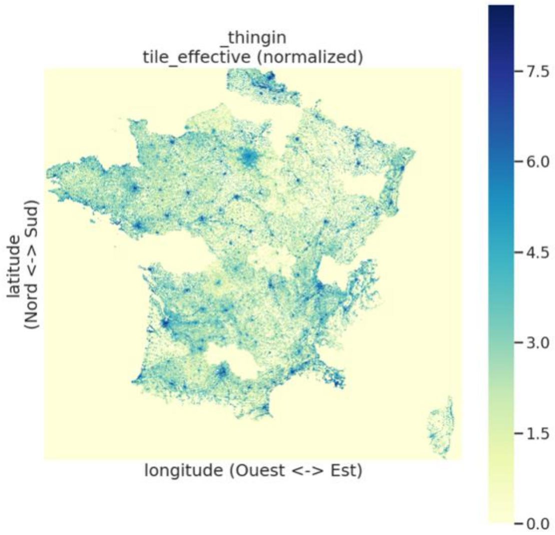


Figure 6. Thing'In node population distribution

Nodes are distributed in a uniform manner across the territory of France (See Figure 6). There are still some regions where no data has been injected, thus with a density of 0, whereas all large cities show progressive density towards the town center. Based on its properties, we consider Thing'In as a WAG. Consequently, it represents an interesting graph to be partitioned with ISP.

6.2.2.Synthetic datasets

We built the synthetic datasets under two assumptions. First, we assume that population density in a WAG should behave the same as human density. Second, we think ISP performances (and in fact any geometric partitioning strategy), applied to a geometric graph, should be directly related to the EDTPS and CDTPC metrics (see Section 4). Regarding density, there will be deserts, a zone with low to no density, and metropolis, a zone with extremely high density.

Table 1. Density configuration

Category	**0**	**0.5**	**5.0**	**20.0**	**50.0**	**500.0**
HD	1000	1000	100	0	10	1
MD	1000	1000	100	0	10	0
LD	1000	1000	100	10	0	0

The density settings can be found in Table 1; each column represents a density category, the header represents the density units and the values are the surface units. The HD configuration is based on a roughly simplified Zipf's (https://en.wikipedia.org/wiki/Zipf\%27s_law) distribution of human population across the Earth. It is composed by 47.37% empty surface (ocean and desert), 47.37% scarce density surface (rural area), 4.73% low density (urban area), 0.47% of medium to high density (large city) and 0.05% of extreme density (metropolis). We declined it with medium (MD) and low (LD) densities to obtain CDTPC and load balance variation. If a row as 0 surface units for a given column, it means it does not use this density category.

The same density and volumetry configuration will behave differently with a larger or tinier plane. A large plane will smooth out the density whereas it will spike even more with a tiny one. We consider graphs with varying density and plane size to see how it impacts the CDTPC metric and ISP's load balance. In particular, hyper dense hot spot may be difficult to manage.

Table 2. Edge distance (km) configuration, each interval respect the following pattern [min_distance : max_distance](%total_edges)

Category	**interval**
TINY-	[0:2[(97%), [2:30](3%)
TINY	[0:30](100%)
SHORT	[0:2[(30%), [2:50[(50%), [50:100](20%)
MEDIUM	[0:20[(25%), [20:50[(25%), [50:100](50%)
MEDIUM+	[0:10[(40%), [10:100[(40%), [100:200](20%)
LONG	[0:160](100%)
LONG+	[100:250](100%)
HUGE	[300:500](100%)

Table 3. Synthetic plane configuration

Type	**Latitude**	**longitude**
FR (Country)	[41.15:50.33]	[-5:9.85]
EU (Continent)	[41.15:65.33]	[-5:40.85]

Concerning ISP performances correlated to EDTPS and CDTPC, various configurations are tested to obtain a wide range of EDTPS values and evaluate its accuracy. We established 8 edge distance configurations described in Table 2. Note that the first distance category together with the medium density specifically aims at producing graphs similar to Thing'In. We limited ourselves to two plane sizes (See Table 3). We seek with this variation to prove that absolute plane size has no impact on ISP performance unlike EDTPS which is a ratio relative to plane size. By default the density is set to HD, the most penalizing load-balance wise. TINY- has been generated only with MD density whereas LONG has been tested with all 3 densities.

Table 4. Average edge distance and EDTPS of synthetics WAGs

Category	**AVG_Dist**	**EDTPS**	
		FR	**EU**
TINY-_MD	1041.49	0.00192	0.00076
TINY	13521.19	0.02123	0.00967
SHORT	25860.80	0.04773	0.01765
MEDIUM	48902.51	0.09026	0.03395
MEDIUM+	55766.50	0.10293	0.03562
LONG	84392.80	0.15608	0.0562
LONG_MD	86869.69	0.16034	0.05508
LONG_LD	89290.85	0.14142	0.06008
LONG+	176261.23	0.32547	0.12684
HUGE	391988.63	0.7235	0.2865

We have performed some early metrics on the generated graphs to check our requirements were respected. We measured edge distance with EDTPS (See Table 4) and cumulative connectivity distribution. As expected, we could not generate graph with scale-free properties: the connectivity of the generated graphs remained very low. Moreover we could not exactly reproduce the properties of Thing'In with our generator (See the line TINY-), the EDTPS should be even lower. On the other hand we managed to approach closely its value in absolute terms and obtained a great set of EDTPS values ranging from 0.1% to 72.35% with 4 partitions. We also included graphs with the same configuration on plane, edge distance and density but with varying volumetry to ensure volumetry is not a performance factor of ISP. Due to space limitations, these additional experiments are not reported here.

7. RESULTS

We seek in this experiment to compare existing state of the art streaming graph partitioning method to ISP. We applied ISP, FENNEL and LDG on each dataset with the same streaming approach. The only difference aside the decision algorithm is the streaming order. In ISP nodes were streamed ordered by the cells of the SFC whereas nodes were streamed randomly for FENNEL and LDG because, given our scale, DFS (Depth First Search) and BFS (Breadth First Search) were too expensive to apply. We first look at how strategies behave in terms of edge-cuts and secondly how they handle balance across their partitions. Note that LDG results are not showed in the tables: LDG is always worse than FENNEL regarding edge-cuts and equivalent in load-balance.

7.1. Edge-cuts

As it was our objective, we managed to reach a breaking point where the performance of FENNEL finally exceeds ISP. It is shown in Table 5 (See LONG+) that on a plane with the size of a country like France, with edge Euclidean distance past 175 kilometres on average and 16 partitions, it becomes better to apply FENNEL rather than ISP.

We expected while building our datasets that the performances of ISP would heavily correlates to EDTPS, it has been confirmed through the results. Each time EDTPS increases so does the number of edge-cuts produced by ISP without exceptions. We also managed to reach a scale where ISP is bound to perform extremely poorly (HUGE), in that case up to **95%** of the edges are cut. On the contrary, when edges are extremely short (e.g. TINY-), ISP outperforms **112** times the results of FENNEL in synthetic graphs and up to **125** times when applied to Thing'In (See Table 5 and 7).

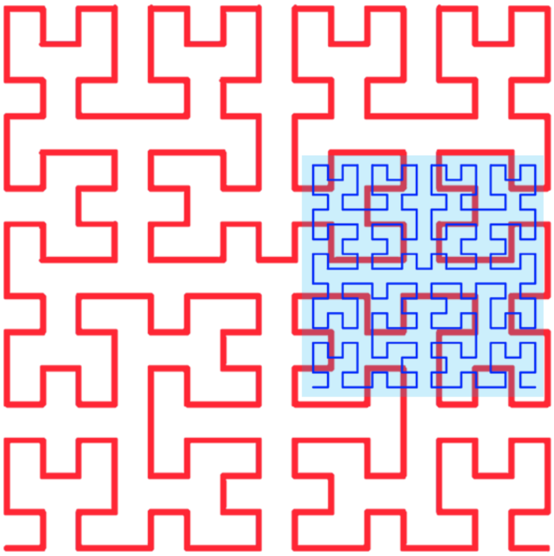


Figure 7. Visual example of generator (Blue SFC) and ISP (Red SFC) mismatch

We used the same space-filling curve based strategy for graph generation and ISP. Yet this has no impact on the results for two reasons. First, both SFCs are not matching. For two curves to match, those needs to use the same algorithm, granularity and map the same multidimensional plan. Although algorithm and granularity used are identical, the multidimensional plan are different. ISP covers the whole world map whereas the generator cover only the bounding boxes given in Table 3, changing both SFC cell bounding box surface and SFC cell ordering (See Figure 7 for a visual example). Second, the generator edge binding step ignores the curve presence, edges are created randomly around a node with respect to edge Euclidean distance configuration.

Interestingly, we see multiple EDTPS and ISP results correlation, it must however be weighted with the amount of borders induced by the partitions which is currently ignored by our metric. Each new partitions produces additional border which may cut edges. As EDTPS does not include those borders in its estimation, we use the following interpretation rule: two graphs with identical EDTPS but different number of partitions should have different behaviour with ISP. The one with less partition should yield better results. We joined a part of the results obtained for the EU plane size in Table 6 and this interpretation rule does hold true.

Table 5. EDTPS and performances of ISP and FENNEL over the generated graphs for FR plane

Category	K	EDTPS	FENNEL	ISP
TINY-_MD	4 8 16	0.00192 0.00272 0.00384	38.232±0.002 45.811±0.002 50.194±0.002	0.34±0.0 0.55±0.0 0.83±0.0
TINY	4 8 16	0.02123 0.03002 0.04245	47.733±0.0 56.241±0.0 61.074±0.0	2.61±0.003 5.1±0.007 8.05±0.006
SHORT	4 8 16	0.04773 0.0675 0.09546	47.872±0.0 56.337±0.0 61.129±0.0	5.89±0.008 9.86±0.017 15.21±0.007
MEDIUM	4 8 16	0.09026 0.12764 0.18052	48.014±0.001 56.545±0.001 61.345±0.001	10.9±0.018 17.9±0.032 27.59±0.021
MEDIUM+	4 8 16	0.10293 0.14556 0.20586	47.908±0.001 56.372±0.001 61.16±0.001	11.35±0.013 19.45±0.014 27.37±0.01
LONG	4 8 16	0.15608 0.22074 0.31217	47.914±0.0 56.379±0.0 61.175±0.0	16.77±0.013 29.39±0.017 41.18±0.006
LONG_MD	4 8 16	0.16034 0.22675 0.32067	47.917±0.0 56.38±0.0 61.171±0.0	18.12±0.026 28.21±0.031 41.06±0.04
LONG_LD	4 8 16	0.14142 0.2 0.28285	47.897±0.0 56.361±0.0 61.159±0.0	15.5±0.05 25.99±0.082 38.21±0.116
LONG+	4 8 16	0.32547 0.46028 0.65093	48.07±0.0 56.564±0.001 61.364±0.001	30.06±0.017 47.71±0.013 67.48±0.004
HUGE	4 8 16	0.7235 1.02318 1.44699	48.056±0.001 56.557±0.002 61.359±0.002	63.88±0.017 86.02±0.007 95.49±0.004

If we compare the results of EU LONG+ and FR MEDIUM with k = 8 and 16 (see Table 5 and 6), they have similar EDTPS but EU LONG+ has consistently better edge-cuts results as it uses fewer partitions to reach the same EDTPS. The same pattern can be detected comparing EU LONG_LD to FR MEDIUM k = 8.It can also be applied to graphs within the same plane: take FR TINY with k = 16 and FR MEDIUM+ k = 4 or EU HUGE k = 4 and EU LONG_LD k = 16. At last, ISP performances are similar when both EDTPS and number of partitions are matching(See FR TINY and EU SHORT). The EDTPS values are interleaving and so do the edge-cuts results with ISP. Finally, it is easy to see EDTPS and edge-cuts correlation (See Figure 8). The only unmatched curve corresponds to HUGE which is perfectly acceptable as there is a ceiling effect for the edge-cuts. EDTPS may go further than 100% but edge-cuts are ultimately limited to 100%.

Table 6. Partial EDTPS and performances of ISP and FENNEL over the generated graphs for EU plane

Category	**K**	**EDTPS**	**FENNEL**	**ISP**
SHORT	4 8 16	0.01765 0.02495 0.03529	47.279±0.001 55.79±0.001 60.601±0.001	2.49±0.003 4.18±0.004 7.24±0.004
LONG	16	0.1124	61.167±0.001	18.36±0.014
LONG_MD	16	0.11017	61.163±0.0	17.81±0.01
LONG_LD	16	0.12016	61.159±0.0	18.65±0.002
LONG+	4 8 16	0.12684 0.17937 0.25367	48.365±0.001 56.899±0.001 61.709±0.001	14.37±0.022 25.49±0.018 36.19±0.007

Table 7. EDTPS and performances of ISP and FENNEL over Thing'In

	K	**EDTPS**	**FENNEL**	**ISP**
Thing'In	4 8 16	0.00006 0.00008 0.00011	6.349 34.636 37.883	0.132 0.277 0.435

While EDTPS seem appropriate to evaluate ISP performance, it ignores other potential factors like connectivity and volumetry, we argue that it does not affect its veracity. We did run tests to check the volumetry impact and we couldn't observe anything significant, the problem is more about connectivity. Connectivity is ignored in both ISP and EDTPS and our dataset does not present much connectivity variation. Still, in essence, as connectivity degree rise for a given node, it has to reach further and further to connect to new nodes as there is a finite number of nodes in its neighbourhood. Consequently, edge distance tends to be longer. But as edge distance is already used in EDTPS, ignoring connectivity should not be a problem.

We expected ISP results to be a bit dissonant; nonetheless we are surprised by FENNEL extreme stability. Again, our most serious guess is based on the fact that all the synthetics graphs share the same low connectivity with absence of super nodes which FENNEL relies on to build its partitions.

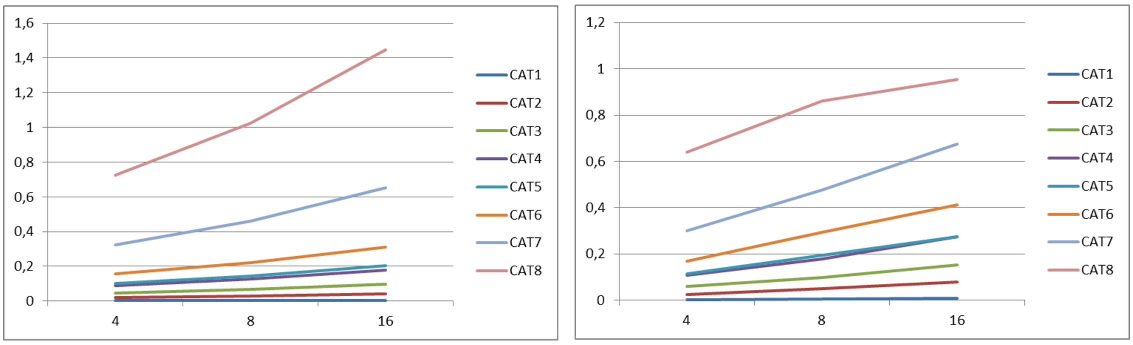


Figure 8. (Left) EDTPS, (Right) Edge-Cuts percentage of graph partitioned with ISP for each category. X-axis is the partition number, bounding box is FR, density information excluded as impact less on EDTPS and edge-cuts.

Contrary to our synthetics graphs, Thing'In possess less edges but its connectivity degree distribution has a better scaling, it could explain why FENNEL yields better results when applied on Thing'In. There remains a wide disparity between 4 and 8 partitions with FENNEL on Thing'In for which we have no appropriate explanation. Overall, we cannot be sure that the results similarities for FENNEL are really due to the connectivity. It is however safe to assume that edge distance has no direct impact on this solution.

7.2. Partition balance

We discuss now about how the different strategies handle balance across the partitions. FENNEL and LDG are pretty much perfect at preserving load balance across the partition, the same cannot be said for ISP looking at Table 8. We didn't put the other graph categories but they all use the same density except for TINY-. The partition imbalance ranges from 2.26% to 11.18% with 16 partitions on FR plane. Obviously, the results regarding imbalance are better for the EU plane. Its worst imbalance case has been recorded at 2.31% which is normal: there is the same number of nodes and density but for a larger plane.

Table 8. CDTPC and load imbalance of ISP and FENNEL over the generated graphs for LONG+

Category	K	CDTPC	FENNEL	ISP
LONG+	4	0.04916	0.0±0.0	0.22±0.002
	8	0.09831	0.0±0.0	2.13±0.013
	16	0.19662	0.0±0.0	7.09±0.022
LONG+_MD	4	0.00836	0.0±0.0	0.16±0.002
	8	0.01673	0.0±0.0	0.6±0.003
	16	0.03345	0.0±0.0	1.96±0.003
LONG+_LD	4	0.00325	0.0±0.0	0.1±0.001
	8	0.0065	0.0±0.0	0.19±0.001
	16	0.01299	0.0±0.0	0.61±0.004

The balance provided using the ISP strategy is very fragile. It hugely depends on the graphs if it is "well behaved" or not that is, if the density does not peak too much relatively to the precision of the SFC. Even though a huge CDTPC ratio will not automatically trigger imbalance problem, it is a good risk indicator as a graph with a very low CDTPC will never encounter balance trouble. With the graphs generated based on the density HD given in Table 1 over a plane of a size similar to France and the Hilbert SFC of zoom 12, a single cell might contain more than 500k of nodes all by itself. This does pose a problem because in some circumstances with a high number of partitions and extremely peaked density a single cell has a load high enough to provoke single-cell partitions.

This problem can be solved using SFC with higher zoom as it splits the load of a single cell into multiple cells. There is no complete guarantee that zoom splits the load but in practice it works. A cell which stores 500k nodes at zoom 12 may be reduced to 64 sub cells at zoom 15 with a load of at most 10k nodes by cell, dropping effectively the CDTPC from 17% to 1.5%. With such a small CDTPC, the risk of imbalance becomes almost zero, and in all fairness, zoom 15 is a rather standard precision level. Unfortunately, we could not afford zoom 15 as our metric system wouldn't have been able to hold the additional load memory wise.

8. CONCLUSION

In this paper we went back through an unusual graph partitioning solution reserved to FEM issued graph and applied it, after little extension, to a new class of graph we defined and called WAG. We proposed additional metrics, along with the solution, to analyse why ISP performs well or not and provided our customized geometric graph generator designed to produce customized WAG. We evaluated our partitioning solution through an extended synthetic dataset and compared it to state of the art graph streaming partitioning solutions: LDG and FENNEL. Overall, we showed that when ISP is applied to WAG, performances mainly depend on the graph edge distance and the distribution of its density across the plane. Although ISP is unsuited for long range distance edges, e.g. social graphs, and loses to streaming strategies, it obtains great result for short to medium edge distance typical of spatial networks and outperforms other solutions. In a future with large scale WAG composed of tiny objects with edge distance under one hundred meters as physical interaction between objects is characterized to such range, ISP could prove to be the best solution for such graphs.

For the upcoming work, we would like to further extend ISP. Rather than the analytical context we are much more interested in the database context and we would like to combine ISP with diffusive load balancing technique to optimize both edge cuts and load balance. Replication within ISP is also an interesting subject we would like to explore. At last we have ideas to enhance our WAG generator, we believe we can mix it with the preferential attachment model to produce WAG with high, spatially logical, connectivity.

REFERENCES

[1] K. Andreev et H. Räcke, « Balanced Graph Partitioning », in Proceedings of the Sixteenth Annual ACM Symposium on Parallelism in Algorithms and Architectures, New York, NY, USA, 2004, p. 120–124, doi: 10.1145/1007912.1007931.

[2] Recent Advances in Graph Partitioning | SpringerLink ». https://link.springer.com/chapter/10.1007/978-3-319-49487-6_4 (consulté le oct. 07, 2020).

[3] F. Payan, C. Roudet, et B. Sauvage, « Semi-Regular Triangle Remeshing: A Comprehensive Study », Comput. Graph. Forum, vol. 34, no 1, p. 86-102, 2015, doi: 10.1111/cgf.12461.

[4] J. R. Pilkington et S. B. Baden, « Partitioning with Spacefilling Curves », 1994.

[5] C. Tsourakakis, C. Gkantsidis, B. Radunovic, et M. Vojnovic, « FENNEL: streaming graph partitioning for massive scale graphs », in Proceedings of the 7th ACM international conference on Web search and data mining - WSDM '14, New York, New York, USA, 2014, p. 333-342, doi: 10.1145/2556195.2556213.

[6] G. Karypis et V. Kumar, « Multilevelk-way Partitioning Scheme for Irregular Graphs », J. Parallel Distrib. Comput., vol. 48, no 1, p. 96-129, janv. 1998, doi: 10.1006/jpdc.1997.1404.

[7] Partitioning of unstructured problems for parallel processing - ScienceDirect ». https://www.sciencedirect.com/science/article/abs/pii/095605219190014V (consulté le janv. 20, 2020).

[8] J. R. Gilbert, G. L. Miller, et S.-Hua. Teng, « Geometric Mesh Partitioning: Implementation and Experiments », SIAM J. Sci. Comput., vol. 19, no 6, p. 2091-2110, nov. 1998, doi: 10.1137/S1064827594275339.

[9] K. Schloegel, G. Karypis, et V. Kumar, « Graph partitioning for high-performance scientific simulations », in Sourcebook of parallel computing, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, p. 491–541.

[10] A. Akdogan, « Partitioning, Indexing and Querying Spatial Data on Cloud », p. 80.

[11] S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, et H. Bhogan, « Volley: Automated Data Placement for Geo-Distributed Cloud Services », p. 16.

[12] D. Delling, A. V. Goldberg, I. Razenshteyn, et R. F. Werneck, « Graph Partitioning with Natural Cuts », in 2011 IEEE International Parallel Distributed Processing Symposium, mai 2011, p. 1135-1146, doi: 10.1109/IPDPS.2011.108.

[13] I. Stanton et G. Kliot, « Streaming graph partitioning for large distributed graphs », in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’12, Beijing, China, 2012, p. 1222, doi: 10.1145/2339530.2339722.

[14] D. Hilbert, « Über die stetige Abbildung einer Linie auf ein Flächenstück », in Dritter Band: Analysis • Grundlagen der Mathematik • Physik Verschiedenes: Nebst Einer Lebensgeschichte, Berlin, Heidelberg: Springer Berlin Heidelberg, 1935, p. 1–2.

[15] M. Knoll et T. Weis, « Optimizing Locality for Self-organizing Context-Based Systems », in Self-Organizing Systems, 2006, p. 62-73.

[16] B. M. Waxman, « Routing of multipoint connections », IEEE J. Sel. Areas Commun., vol. 6, no 9, p. 1617-1622, déc. 1988, doi: 10.1109/49.12889.

[17] M. D. Penrose, « Connectivity of soft random geometric graphs », Ann. Appl. Probab., vol. 26, no 2, p. 986-1028, avr. 2016, doi: 10.1214/15-AAP1110.

[18] R. Albert et A.-L. Barabasi, « Statistical mechanics of complex networks », Rev. Mod. Phys., vol. 74, no 1, p. 47-97, janv. 2002, doi: 10.1103/RevModPhys.74.47.

[19] A. Guttman, « R-trees: a dynamic index structure for spatial searching », in Proceedings of the 1984 ACM SIGMOD international conference on Management of data, New York, NY, USA, juin 1984, p. 47–57, doi: 10.1145/602259.602266.

AUTHORS

Cyprien Gottstein graduated from the University of Rennes 1 with a master degree in Software Engineering in 2016, Rennes, France. He worked from 2016 to 2018 as an engineer at Thales services and started his PhD in large scale graph partitioning for the Internet of Things (IoT) in 2018 at Orange Labs, Rennes, France.

Philippe Raipin Parvedy has received his PhD from the University of Rennes 1, Rennes, France, in 2004. He was a temporary lecturer and research assistant in the university of Rennes 1 from 2004 to 2006 and a post doc researcher in Orange from 2006 to 2007, since 2007, he has been a research engineer in Orange. He is currently the research program manager of the Web of Things platform in Orange. His research interests include dependability, large systems and graph databases.

Michel Hurfin holds a Ph.D. in Computer Science from the University of Rennes (1993). After one year spent at Kansas State University, he is currently conducting his scientific research activities at the INRIA Rennes Bretagne Atlantique research center. He defended his HDR (Habilitation à Diriger des Recherches) in 2004. Since 2012 he is a member of the CIDRE project that aims at addressing security problems. His research interests include distributed systems, software engineering and middleware for distributed operating systems. His recent works focus mainly on dependability and security issues in distributed systems (in particular, intrusion detection mechanisms).

Thomas Hassan is a researcher at Orange Labs, Rennes, France since 2019. He obtained a Thesis in Computer Science from the University of Burgundy, Dijon, France in 2017. His main research areas are Semantic Web, Machine Learning and Big Data. His research was applied to the domains of pervasive computing, Internet of Things, geographic information system and recommender system.

Thierry Coupaye is head of research on Internet of Things (IoT) inside Orange, Orange Expert on Future Network and Orange Open Source Referent. Prior to that, after he completed his PhD in Computer Science in 1996, he had several research and teaching positions at Grenoble University, European Bioinformatics Institute (Cambridge, U.K.) and Dassault Systems. He joined France Telecom in 2000 where he had several research expert, project manager, project and program director positions in the area of distributed systems architecture, autonomic computing, cloud/fog/edge computing and networking. He is the author of more than 75 refereed articles and has participated in multiple program and organisation committees of conferences in these areas. His current research interests include Digital Twins and Edge Intelligence (AI@Edge).

ENERGY AWARE ROUTING WITH COMPUTATIONAL OFFLOADING FOR WIRELESS SENSOR NETWORKS

Adam Barker and Martin Swany

Department of Intelligent Systems Engineering,
Indiana University, Bloomington, Indiana, USA

ABSTRACT

Wireless sensor networks (WSN) are characterized by a network of small, battery powered devices, operating remotely with no pre-existing infrastructure. The unique structure of WSN allow for novel approaches to data reduction and energy preservation. This paper presents a modification to the existing Q-routing protocol by providing an alternate action of performing sensor data reduction in place thereby reducing energy consumption, bandwidth usage, and message transmission time. The algorithm is further modified to include an energy factor which increases the cost of forwarding as energy reserves deplete. This encourages the network to conserve energy in favor of network preservation when energy reserves are low. Our experimental results show that this approach can, in periods of high network traffic, simultaneously reduce bandwidth, conserve energy, and maintain low message transition times.

KEYWORDS

Ad Hoc Network Routing, Q-routing, Wireless Sensor Network, Computational Offloading, Energy Aware.

1. INTRODUCTION

A wireless sensor network (WSN) is characterized by a network of small, low power devices, operating remotely with no pre-existing infrastructure, and little to no human intervention or central management. They typically rely on on-board energy storage such as batteries and can be required to operate for months or years at a time. Whereas many WSN collect sensor information to be relayed to a central location, some instances, particularly those for use in military and first-responder applications operate in a peer-to-peer paradigm in which a sensor node can also be a consumer of sensor information from other nodes. This peer-to-peer concept implies any node can be both a source and destination making routing paths between source and destination dynamic. Because WSN operate without fixed infrastructure, they must handle their own routing and can form mesh networks where sensor information can traverse several hops across other intermediate sensor nodes to reach their destination. The lack of fixed infrastructure also leads to limited access to bandwidth. Typical WSN operate in the tens to hundreds of kilobits per second such as seen in IEEE 802.15.4 or Semtech's LoRa protocol. Every bit being transmitted comes at a cost of not only energy to transmit over the wireless link, but also a cost of time to get critical data where it needs to be. As data sets continue to increase in size, the hundreds of kilobits per second data rates will continue to form a bottleneck. The peer-to-peer pattern coupled with a mesh topology can offer some unique opportunities to help reduce bandwidth consumption, reduce the time data needs to travel in the network, and preserve scarce energy reserves.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 243-259, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101418

Until recently, wireless sensor networks have seen limited use, however the concept of the Internet of Things (IoT) has begun to gain popularity putting WSNs at the center of focus as one of the technologies needed to enable the IoT. The IoT promises networks of millions to billions of small devices connecting everything from wearable technologies to autonomous drones. For this scale of technology to be realized, energy, bandwidth, and data delivery time become critical aspects that need to be properly managed. The unique structure of WSNs allow for novel approaches to data reduction and energy preservation.

This paper presents a concept to assist in management of these factors by re-examining existing routing concepts and improving upon them for the unique use case of peer-to-peer mesh WSNs. This concept is implemented through a series of adaptions to the Q-routing protocol we collectively refer to as energy aware Q-routing with computational offloading (EAQCO.) Typical WSN routing algorithms only route messages based on least-cost routes without considering in-situ computation options to reduce the message data prior to forwarding. The EAQCO concept optimizes trade-offs between energy-expensive message passing and time-critical computational offloading in an effort to deliver usable information to a destination node within a WSN. To the best of our knowledge the EAQCO concept is the only research that manages the trade-offs of delay, energy consumption, and bandwidth in a WSN. The results of our experiments show that EAQCO is a viable concept that can minimize time to deliver processed data while minimizing bandwidth used, particularly in networks with high data traffic.

The remainder of this paper will be divided into sections, starting with section 2, which describes the background of WSN routing research. Section 3 will then present a novel approach to address the major concerns of time, energy, and bandwidth, implemented through EAQCO. Finally, section 4 will present results of experiments performed on a wired mesh network of small IoT type devices implementing the EAQCO algorithm.

2. Background

Wireless sensor network routing has been the subject of much research over the past 30 years. With the implementation of 5G technologies, specifically concepts involving Multi-Access Edge Computing (MEC), and increased interest in the Internet of Things (IoT), research in WSN routing has been on the rise. Research in this area can be grouped into 3 general focus areas: energy efficiency, computational offloading, and traffic/congestion management.

2.1. Energy Efficiency

Because WSN operate from energy storage devices such as batteries, maintaining the longevity of the network is the focus of this area. Notable examples of current and past research focused on energy efficient routing include [1], [2], [3], [4], and [5]. Much of the earlier research, such as [1] and variations such as [2], focus on clustering of nodes and aggregation by a selected (in the case of LEACH, randomly selected) cluster head before forwarding data to a fixed end point. Other variations of LEACH allow nodes to enter a low power sleep mode to further conserve energy when they are not a cluster head or transmitting data, such is in [3]. The concept of clustering is very viable for a densely populated sensor network, but in a sparse, widely distributed, network where nodes may only have 2-3 neighbours by which to route data through, the number of cluster heads may come close to the number of nodes in the network. Additionally, clustering relies on sharing the cost of transmission. If the nodes are heterogeneous, particularly regarding energy storage capacity, equally sharing the cost of being a cluster head may not be the most productive approach.

Other novel approaches such as [4] formulate the energy maximization problem as linear program and use multi-commodity flow algorithms to solve for minimum cost, maximum flow where the cost is measured in remaining energy storage and flows are data transmissions. Maximum flow algorithms are widely available and relatively efficient to compute giving an optimum solution very quickly. However, to formulate the affine function across the entire network requires constant updates of the exact network topology including knowing the energy storage of each node at any given time. For small to medium networks with constant transmissions spanning the entire network this is a viable solution, but again for sparse heterogeneous networks, flooding the entire network with all nodes' energy storage state may require significant additional bandwidth and energy to ensure all nodes receive the most up-to-date topology information.

More recent works have examined more advanced approaches such as the use of genetic algorithms in [5]. Their work is novel, however the requirement for a middle layer, similar to the notion of a cluster head, and the construction of their genetic algorithm, require an existing model of the network. In many cases of WSN deployment the exact structure of the network may not be known and therefore no model can be built *a priori*.

2.2. Computational Offloading

Computational offloading is the process of moving the task of processing raw data to a node or set of nodes physically separated from the sensing node. Much work has been done in this area and has developed into core businesses such as Amazon Web Services and Google's Cloud Services. With the increase in the volume of data, sensor nodes with typically reduced computational capacity need to move local computation to a central location for processing, however larger cloud services require high bandwidth, typically orders of magnitude more than WSNs are able to provide. For IoT devices, the majority of the research has focused on building relatively lightweight algorithms and processes that adapt cloud services for mobile networks outlined in work such as [6] and [7]. With the advent of 5G networks, research into computational offloading has turned toward the MEC. The research concept for MEC typically focuses on a centralized algorithm to manage the decision process between multiple nodes within a network and the edge nodes and servers that can perform the offloading. One recent example, [8], uses a model of the network, the self-reported transmission time, and estimated energy consumption of each node to construct a game theory algorithm that settles into a Nash equilibrium for the optimum strategy for all nodes and all tasks in the network.

Each of these approaches are novel and have merit given the use case presented, however the current research is limited in two ways. Firstly, the offloading algorithms require a network model for which to optimize the decision process against, and secondly the algorithm is centrally planned and managed in which each individual node receives the offloading decision from a central node, or base station. In a purely peer-to-peer *ad hoc* network the network model may not be available and distributed decision making would need to take the place of central management. Additionally, most of the offloading models involve arbitrarily complex functions that would need to be offloaded from a multi-function device such as a wireless handset, however WSNs typically involve a single function, or limited functions, on devices that run limited scope computational algorithms. These devices may not benefit from a complicated model that hands off functions via passing of virtual machine objects as is common research thread for MEC.

2.3. Traffic/Congestion Management

Perhaps the most widely researched area in WSN routing is that of traffic management. Data gathered via WSN is typically time sensitive, in the case of real-time applications measured in the microseconds, and therefore optimizing the fastest route to the destination is of highest concern. Routing in a WSN, like other data networks, is often resolved using shortest path algorithms such as ad hoc on-demand distance vector (AODV), optimized link state routing (OLSR) and dynamic source routing (DSV). Variations of these algorithms exist in research in an effort to optimize energy usage and node mobility which are both key features of WSN. The limitation of these algorithms is they rely on information that is gathered in an instant in time, either as needed, or at some time in the past. Due to the dynamic nature of WSN, these protocols do not allow for prediction or learning of the dynamics of the network.

To address variability in wired networks the authors in [9] developed a reinforcement learning (RL) approach to compare with shortest-path-first algorithms used in most networks. Using their example, a message is required to be sent from source *x* to destination *d* via its neighbour*y*. Their approach was a simple variation of the RL concept called Q-learning using the formula:

$$Q_x(\bar{y}, d) = Q_{\bar{y}}(\bar{z}, d) + q_y \tag{1}$$

where $Q_y(\bar{z}, d)$ is *y's* estimate of the remainder of the message's journey to *d* after it leaves *y* and $q_y$is *x's* estimate to get to *y*. Whenever *x* receives a reply from *y* it includes *y*'s estimates to get to each destination, known as *y*'s Q-table. Using *y*'s Q-table, *x* updates its estimate to *d* using the update rule:

$$Q_x(\bar{y}, d) = Q_x(\bar{y}, d)^{old} + \eta\left(Q_{\bar{y}}(\bar{z}, d) + q_y - Q_x(\bar{y}, d)^{old}\right) \tag{2}$$

Where η is known as the learning rate. Neighbour *y* may have several routes to *d* therefore *x* selects $Q_y(\bar{z}, d)$ using:

$$Q_{\bar{y}}(\bar{z}, d) = \min_{z \in neighbors\, of\, \bar{y}} Q_{\bar{y}}\,(\bar{z}, d) \tag{3}$$

Each node maintains its own Q-table using Equations 1, 2, and 3 therefore Q-routing is a distributed learning process.

Many variations of the Q-routing process have been researched to include [10] which adds a confidence factor and backward exploration to address the probability a node in the ad hoc network may drop out periodically. Additional variations include [11] which incorporates a separate learning phase to develop quality of service metrics and [12] which combines any cast routing with the learning capabilities of Q-routing.

2.4. Other Related Work

Our work retains elements of energy efficiency, computational offloading, and congestion management to build a novel approach to routing based off of the Q-routing algorithm from [9] and is detailed in section 3. Other approaches have examined similar aspects, such as in [13] and [14].

In [13], researchers develop a computational offloading scheme for internet of vehicles (IoV) applications. Their work utilizes a genetic algorithm to optimize the offloading of compute tasks to edge servers. Additionally, they include the option to route offloading tasks through adjacent nodes that may be in range of alternate edge servers. While their approach is novel, their multi-objective optimization approach requires global knowledge of every compute task in the local network. In a peer-to-peer WSN such as we outlined above, global knowledge of required computing tasks is not available to individual nodes without imposing a significant cost of communication and time to gather global data at a designated head node. Therefore, we do not believe their formulation is applicable to our intended use case.

In [14], researchers examine industrial IoT (IIoT) applications of multi-hop computational offloading and develop an algorithm using a game theory approach. Their approach is distributed as each IIoT node makes its decision to offload computation independently of the next node. The decision process works as a game giving each node the option to determine if it is more advantageous to compute locally versus remotely. The game ends when all nodes reach a Nash equilibrium. Their algorithm allows for multi-hop routing between nodes and edge compute nodes making it a routing optimization problem. While their efforts are similar to our use case, their model is limited to specific edge compute nodes and doesn’t allow for transfer of compute capabilities to adjacent nodes. In a peer-to-peer WSN there is no designated edge compute node and any determination of computational offloading benefit is made at each node based on its local ability to perform compute functions.

The next section discusses our approach to the unique problem of peer-to-peer WSN and addresses the issues that research to date has not taken into account.

3. DISCUSSION

In an effort to improve the efficacy of Q-routing specifically in a WSN environment while maintaining awareness of limited energy storage capacity and utilizing the computational power latent within the network itself, we present the concept of energy aware Q-routing with computational offloading (EAQCO.) The EAQCO concept utilizes the simplicity and distributed capabilities of the Q-routing algorithm and adds in additional decision logic that determines if it is more feasible to perform data reduction in-place or forward raw data. The primary metric to determine optimal route selection is time, however the energy-awareness component adds an additional factor that increases the cost as energy reserves become depleted.

3.1. Computational Offloading

The EAQCO concept begins with the basic Q-routing algorithm. For an uninitialized network, a series of ping messages carrying a timestamp are sent from a node to all of the node's neighbours. Each neighbour immediately responds to the ping with time it took to receive and process the ping. These responses form the basis of the node's Q-table. In addition to the time to receive the message, the node's neighbours include in their response, a copy of their own Q-tables from previous iterations of ping messages sent to their neighbours. These neighbour Q-tables contain the best estimates of time to send messages to their neighbours. As the ping messages and responses continue for a few iterations the entire network is mapped out with estimates of time to send data between any two nodes in the network.

When a node receives its neighbour's Q-tables it adds the time it takes to send data to its immediate neighbours and the time recorded in the neighbour's Q-tables to build an estimate of the total time it takes to send a message to each destination covered in the neighbour's Q-table.
The sending of pings is only needed to establish the initial topology of the network and can be completed with a maximum number of pings equal to the longest route in the network. Once the topology is established and messages containing data are sent throughout the network, each node

that receives a data message acknowledges the receipt of the data with a response identical to the receipt of a ping. Therefore, each node's Q-tables are continually updated as long as data is flowing in the network. For sparse communications and/or to verify if various nodes are still reachable, pings can be sent on a periodic basis to ensure the Q-tables remain up to date. The authors in [9] noted that early versions of their Q-routing algorithm settled into an optimal route quickly but were unable to recognize when an alternate route was available. This concept, known as exploitation versus exploration, is a well-researched issue within the field of reinforcement learning. To address the issue a parameter, ε, known as the exploration factor, is included in the Q-learning algorithm. By adding in ε, instead of choosing the action with the highest Q-value, or in the case of Q-routing, the shortest time, there is a probability, ε, that the algorithm will choose a random route to explore. Values for ε are typically 0.1 - 0.5. Other factors such as those proposed in [10] could be added to the Q-routing algorithm as needed to enhance the performance, but they would not disrupt the computational offloading capabilities described below.

So far, this process described above is no different from standard Q-routing, however, EAQCO adds an additional step that is determined in parallel to message routing. When a data message is first made available for forwarding, and every time a message is received by a node, there is a decision process to forward or perform data reduction (computation) in place. If data reduction is selected, the message is placed in the node's computation queue and processed in the order it was received. The cost of computing in place is measured in time by determining the length of the processing queue and is updated using the same update rules for determining new Q-values shown in Equations 1, 2, and 3. This process is built into the Q-routing algorithm process at the point where the optimal route is selected.

With no loss to generality, a simple example is used to explain the decision process. Figure 1 shows an extremely simple network of just 3 nodes.

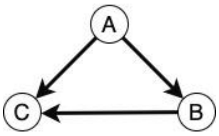


Figure 1. Simplified Example Network

Assuming Node A had a requirement to send data to Node C there are two paths it could take: A to C directly or through B. The Q-routing algorithm, given in [9], would determine the optimal route by choosing the minimum time of the 2 routes:

$$Q_A(c) = min\{ Q_x(\bar{c}, c). Q_x(\bar{b}, c)\} \tag{4}$$

However, with the computational offloading step, a third option exists by including the Q-value for performing data reduction in place. The set of actions that can be performed by A are then:

Table 1. Actions available to A

Action	***Reward***
Fwd to C	$Q_x(\bar{c}, c)$
Fwd to B	$Q_x(b, c)$
Compute Locally	$Q_x(\overline{compute})$

Therefore Equation 4 is now:

$$Q_A(c) = min\{\, Q_x(\bar{c}, c), Q_x(\bar{b}, c), Q_x(\overline{compute})\} \tag{5}$$

If, in this simple case, the data could be reduced and sent to C in less time than it would take to send the raw data to C, A would choose to perform data reduction before forwarding the resultant information on to C. Because of the nature of WSNs, most of the data is redundant and can be repetitive, however without some data reduction step, it cannot be known what is useful data, therefore the default is to forward all data and let the end user determine what to keep. The computational offloading step performs the pre-defined reduction before the data reaches the end destination. The addition of the compute action is also included in B's action set. If A chose to forward to B, B now has two available actions: continue forwarding to C or compute in place, therefore the entire tree of actions for moving data from A to C is expanded with the inclusion of the option to compute locally at every node.

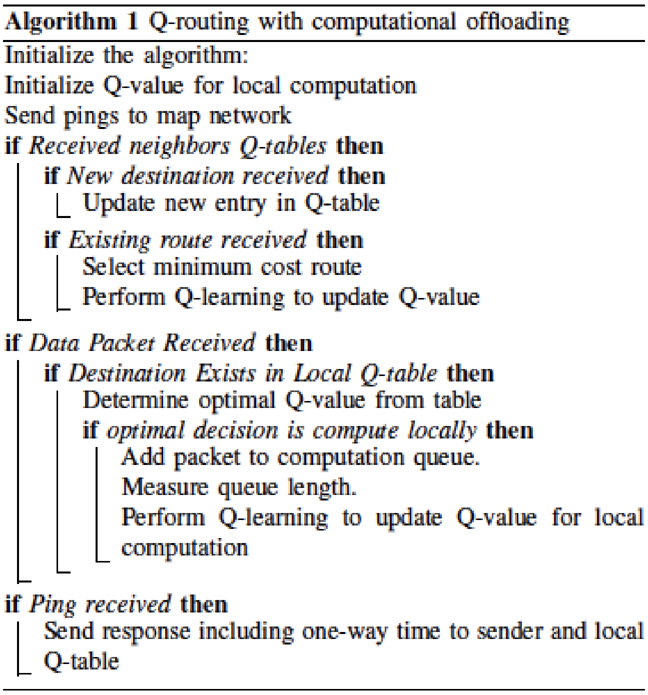
As an example, if a WSN of seismic sensors was monitoring the vibrations at a specified location they might be capturing data several hundred times a second. Once there is enough data it is forwarded on to the destination to determine when there is a significant change. Ultimately what is needed is the time and magnitude of the change point. This could be captured in a tuple of 2 32-bit numbers (time and magnitude), representing a total of 64 bits of information, however, the dataset needed to determine the change point may contain several thousand samples of 64-bit tuples. Therefore, computation can result in a 1000-fold reduction in bandwidth. Because WSN are typically made up of small computationally constrained devices, determining the change point may take a significant portion of the originating node's computational capabilities and could not be performed for every dataset without incurring significant delay. As the datasets become more complex and the computation more intensive, such as the case with depth mapping full motion video or IQ-processing for software defined radios, the cost of computation continues to increase and therefore efficiency benefits from computation being distributed throughout the network.

Because the EAQCO is distributed and the forwarding versus computation decision process is happening at each node, the decision to compute locally is done at each node and with each compute decision, the data set is slowly being reduced focusing on optimizing overall data propagation time. Each decision to compute has the added benefit of reducing the overall bandwidth needed to transmit the streaming data which reduces the overall energy needed to transmit. This detail will be revisited in a follow-on section.

Each iteration of a compute task updates the Q-value for local computations just as each forwarded message updates the Q-value for the time to forward a message to the next neighbour. As the process continues, each node builds its Q-table of forwarding routes and comparative computation costs. This is the Q-table that is returned after each data message is sent or a ping is received. Therefore, each node that receives a neighbour's Q-table is receiving their neighbour’s decision to either forward a message or compute locally. This indicates if a node decides to forward a message to a neighbour because its Q-value is lower than the other options, that node does not know whether their neighbour will forward that message on or perform data reduction in place.

There are limitations to this approach, namely that the data and the resultant computation must be severable. If computation of one subset of data requires the entire dataset, this approach will not be feasible, however most streaming applications such as video, audio, and IQ-data are often severable and can be computed in slices as needed.

High level pseudo code of ping and data sending process with computational offloading is shown in figure 2.



```
Algorithm 1 Q-routing with computational offloading
Initialize the algorithm:
Initialize Q-value for local computation
Send pings to map network
if Received neighbors Q-tables then
    if New destination received then
        Update new entry in Q-table
    if Existing route received then
        Select minimum cost route
        Perform Q-learning to update Q-value
if Data Packet Received then
    if Destination Exists in Local Q-table then
        Determine optimal Q-value from table
        if optimal decision is compute locally then
            Add packet to computation queue.
            Measure queue length.
            Perform Q-learning to update Q-value for local
            computation
if Ping received then
    Send response including one-way time to sender and local
    Q-table
```

Figure 2. Pseudo code of EAQCO process

3.2. Energy Awareness

The timeliness of sending data within a WSN is of utmost importance for time critical applications such as military, public safety, and high-mobility platforms like self-driving automobiles. However, a unique feature of most WSNs is a limited energy storage capacity. They are typically battery operated and at times intended to operate for several months or even years without human intervention. The scarce resource of energy presents another set of concerns for WSN. To extend the battery life of wireless sensors and other energy constrained devices, many studies have researched the energy consumption profiles to quantify the subsystems with the largest energy consumption profile. One example in [15] found that a mobile phone's Wifi or GSM module can consume more than 8 times the energy of the CPU.

Determining the energy used to transmit wireless messages is the subject of much research, however generally, energy is consumed in two parts: energy used to transmit radio frequency (RF) signals, and energy used to receive RF signals. Typical wireless packet-sending protocols start with the transmission of a message, once it is received an acknowledgement is returned to confirm message receipt. If no acknowledgement is received within a predetermined timeframe the message originator retransmits the message. This process continues until the message is acknowledged or a retry threshold is met. Determining the amount of total energy consumed during this transmit/acknowledge process is a stochastic process based on the characteristics of multiple wireless parameters. Researchers in [16] examined this process and developed a simplified model that estimates the expected energy consumed. Using the energy consumption model developed in [16], a formulation was developed to determine energy consumption based on data message size and includes factors for transceiver hardware characteristics and wireless link parameters. The formulation developed is shown in Equation 6 for a data message of size x bits and a data rate of r:

$$E[e_t] \leq \frac{1}{p^2}\left((P_{xmtr}) \times \frac{x_{data\_packet}}{r_{data\_rate}} \right) + \frac{1}{p}\left(\frac{P_{rcvr}}{r_{ACK\_rate}} \times x_{ACK\_packet} \right) \tag{6}$$

where $E[e_t]$ is the expected energy used to transmit a message, p is the message reception probability, which is a function of the characteristics of the wireless link, $P_{xmtr}$is the wireless radio transmit power (including losses due to transmitter inefficiencies), and $P_{rcvr}$is the wireless radio receiver power consumption. The inequality is the upper bound considering message retransmits due to signal loss and applies as long as the number of retransmits is less than ∞.

From this formulation the cost of transmitting per bit over a particular link can be estimated giving a factor for transmitting future messages over a wireless link given the current energy storage of the wireless sensor device. This factor, we designate as the energy factor (*ef*), is expressed as an exponential function, the factors of the exponential are determined by the specific type of wireless link used as expressed in Equation 6, however the variation used in the proceeding experiments assumed a low data rate with a transmission power of milliwatts to one watt such as used in the LoRaWAN™ wireless link [17]. An exponential function was chosen, because we desire the *ef* to have minimal impact when there is minimal energy reserves used, such as below 50% energy capacity level, but asymptotically approach a factor of 10 as energy used is above 50%.

Using this configuration, the energy factor formulation becomes:

$$ef = 0.001 \times 10^{4\times(batt\_used)} \times msn\_remain \tag{7}$$

where *batt_used* is the fraction of battery capacity consumed and can be retrieved from the on-board battery monitoring circuitry. The term *msn_remain* is an added factor to account for a WSN's potential for a predefined operating time, which is often the case in military applications. The *msn_remain* is the fraction of total estimated mission time remaining and devalues the battery capacity factor as the mission time gets closer to its expected completion point. For operations where the WSN is needed to operate indefinitely, this factor can be eliminated.

The energy factor is computed with every update to the Q-routing algorithm and essentially wraps the computed Q-value for a given forwarding route in the exponential function. As the energy factor increases, the overall Q-value for any given forwarding route increases as well, therefore as the energy storage reserves are depleted the advantage of computing in place increases even as the compute queue increases. The overall effect is data flow rate within the network decreases as energy reserves deplete, shifting the priority from optimizing the information delivery time to overall network preservation.

Given Equations 1, 2, 3, 5, and 7, and a requirement to send information from *x* to *d,* the final formulation of the EAQCO algorithm becomes:

$$Q_x(\mathrm{y},d) = Q_x(\mathrm{y},d)^{old} + \eta\left(ef\left(t_{x\to y}\right) + \min_{z\in neighbors\ of\ y} Q_y(\mathrm{z},d) - Q_x(\mathrm{y},d)^{old} \right) \tag{8}$$

$$A_x(d) = min\{\, Q_x(\mathrm{y},d), Q_x(\overline{compute}) \forall y \in Y\} \tag{9}$$

Where $Q_x(\mathrm{y},d)$ is *x*'s value of forwarding a message to *d* through neighbour *y*. Greek letter η is the learning rate, *ef* is the energy factor from Equation 7, and $t_{x\to y}$ is the estimated time to send a message from *x* to *y.* The element $Q_y(\mathrm{z},d)$ is the Q-table forwarded from all of x's neighbours, including any of *y*'s calculation of $Q_y(\overline{compute})$ which is their estimation to compute locally.

Therefore, x selects the min cost action, $A_x(d)$, between either forwarding to any of its neighbours in the set of all neighbours Yor computing locally as shown in Equation 9.

Equations 8 and 9 form the basis for the EAQCO algorithm. The addition of the energy factor in Equation 8, bias the forwarding of messages based on the stored energy reserves of a given node and the inclusion of the action to compute locally to reduce energy consumption and bandwidth usage are the primary contributions of the EAQCO algorithm. Using these Equations, several experiments were run as is described in the following section.

4. EXPERIMENTS

To test the application of EAQCO, a physical network of Beaglebone® Blacks [18] was built, where each node contained at least two network devices and could perform as a router for network traffic while simultaneously performing sensor and compute functions. The network was instantiated using copper Ethernet as the communication link as physical limitations of the lab did not allow for the use of a wireless link. The algorithm was built in Python3 using the Sockets API with a UDP transport protocol for all messages. Sensor data was simulated by randomly generating data sets of 100 16-byte floating point numbers with an associated 16-byte timestamp. The overall message size was 2360-2400 bytes including header information. The configuration of the network is shown in figure 3 with each node corresponding to a single Beaglebone® Black.

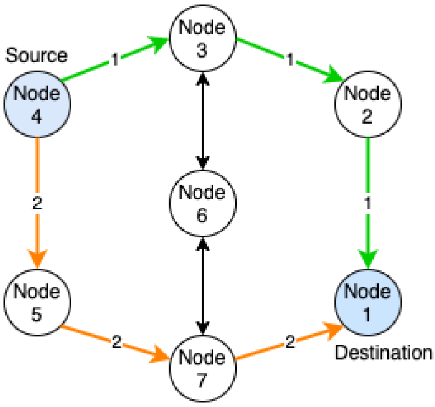


Figure 3. Beaglebone Test Network

4.1. Test Setup

Three scenarios were tested while varying different parameters as shown in the results section:

- Static Routing
- Q-routing with Computational Offloading
- Q-routing with Computational Offloading and Energy Awareness

For the Q-routing cases, each node is running an instance of the EAQCO algorithm and is generating and receiving corresponding ping and data messages, however data was only gathered between Node 4 (source) and Node 1 (destination) to reduce the size of the data set. For the static case, each node routes messages along a predefined static route, however this route could have been developed using any standard *ad hoc* routing algorithm such as AODV or OLSR. For all cases, all nodes, except for source and destination, generate random data destined for any of the randomly selected nodes to ensure the network is sufficiently loaded, commensurate with the specific phase of testing.

In each scenario, Node 4 streams a series of 750 messages containing payloads of 2360 bytes to the destination, Node 1. In the static scenario, the messages follow route 1 as designated in figure 3. In the alternate scenarios, both route 1 and route 2 are utilized and occasionally the algorithm will explore sending data via a path that traverses Node 6. All three scenarios were run multiple times in configurations designated low and high where in the low configuration the intermediate nodes are generating random data messages approximately once per second and in the high configuration messages are generated sequentially as fast as possible. The high rate typically corresponded to an effective data rate of approximately 1.5 Mbps/node. The network loading corresponded to an approximate 10-fold increase in message transmit times between low and high.

Each test scenario was run 10 times and the results were averaged over across the runs. Once a baseline was established for the three scenarios, additional tests were performed varying the learning rate and ε parameters. The results of the tests are shown in the next section.

4.2. Results

Data was collected to look at the primary metric of overall effect on message processing times. Processing time, in the context of the tests, is defined as the time a message takes to transition from source to destination, including time to perform the necessary data reduction/computation. For the static routing case all data reduction/computation takes place at the destination node after the message traverses the predefined route. For the Q-routing cases, computation can take place anywhere in the network as determined by the decision process of the routing algorithm. The results of the 10 trials of each low and high data rates are show in figure 4.

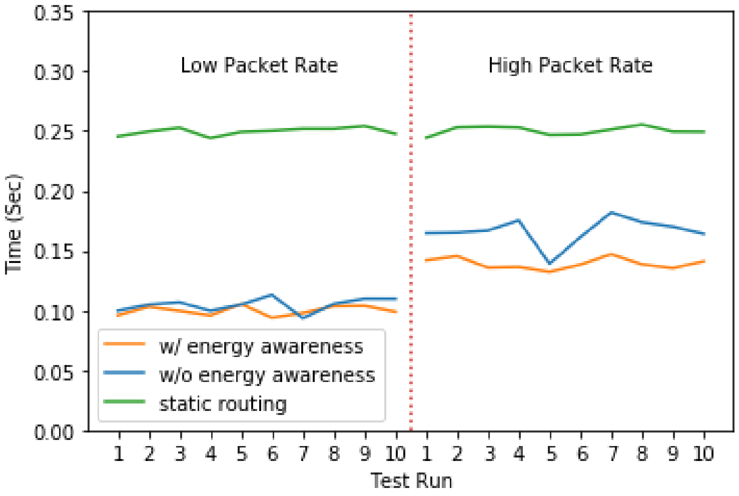


Figure 4. Mean total processing time per message

It can be seen in the results that Q-routing with computational offloading results in a significant reduction in total message processing time. For the low message rate case, energy awareness does not appear to affect the mean time significantly, however in the high message rate case there is a marked difference as the decision process tends to skew towards computation in place rather than forwarding.

In addition to timing, data was examined to determine the overall energy usage of the message stream's traversal from Node 4 to Node 1. The mean energy usage of the 10 trials at each the low and high configurations was determined using the energy model from Equation 6, assuming no retransmits, and hardware profile equivalent to LoRaWAN™ to determine energy usage per byte transmitted. Each trial used a learning rate η of 0.1 and an ε of 0.1. Total energy used was then calculated using:

$$E_{total} = s \times h \times p$$

where *s* is the size of the message in bytes, *h* is the number of hops, and *p* is the total number of messages in each stream. In the static routing case all factors for energy are constant, which results in a linear result across all test runs. For both cases of computational offloading the number of hops varied depending on the optimal decision made at each node. The results are shown in Figure 5.

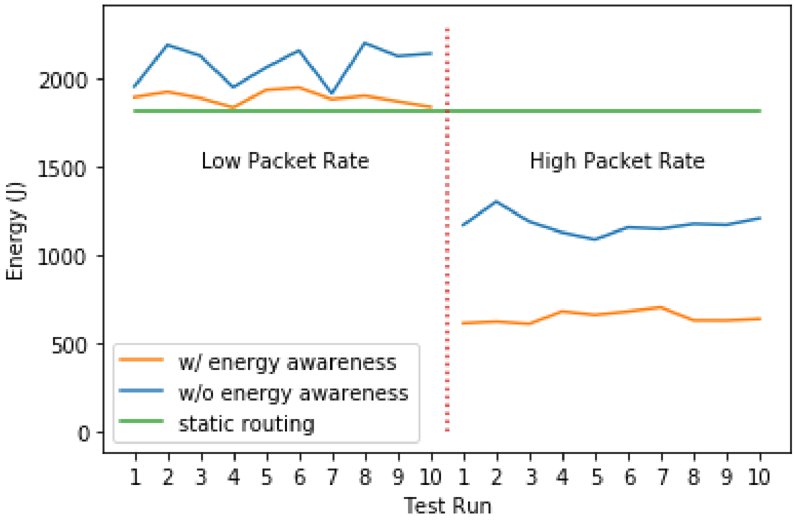


Figure 5. Mean total energy used to transmit data stream

For the low rate, both versions of computational offloading result in higher energy usage compared to the static baseline. There are three factors contributing to this effect. First the Q-routing algorithm performs exploration as noted previously, therefore, randomly, with a probability ε the algorithm may choose a more energy costly route such as traversing the path through Node 6. Secondly the primary factor for route optimization is time rather than number of hops, therefore, due to some congestion at Node 3 the algorithm may find the optimal route is through Node 6, which minimizes time rather than energy. This is highlighted in the difference between the case with energy awareness and the case without. As the energy storage capacity is depleted, the algorithm favours lower energy routes; fewer hops over more hops. Thirdly, in the low data rate case, the cost to compute locally does not compare favourably to the relatively short message transit times therefore the algorithm is forwarding messages more than computing locally. Conversely, examining the high message rate results show the effect of congestion in the network as the algorithm is deciding to compute locally rather than forward at a higher rate. The greater difference between the two cases of with and without energy awareness is a result of the increase in message processing at all nodes in network. The more messages that are processed the more the energy storage reserves are depleted and the more favourable the decision compute becomes. This can be further highlighted in figure 6.

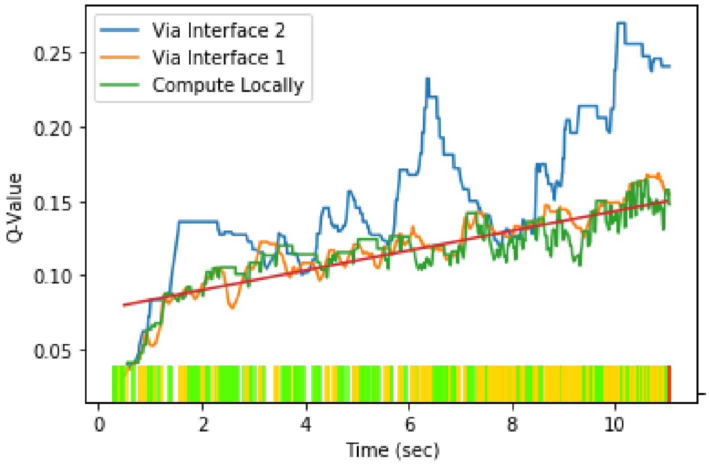


Figure 6. Q-value of action space over time

Figure 6 shows the Q-values over time from Node 4 in a randomly selected run. The red line highlights the upward slope as energy reserves are depleted. The bottom tick marks show action selections where green corresponds to forwarding and yellow corresponds to computing locally. It is also worth noting that the Q-value for computing locally also increases over time. This is due to a higher rate of decisions to compute locally resulting in longer compute queues. The long-term effect of this is the entire action space increases as energy reserves are depleted and, although it becomes more favourable to compute in place, no one action becomes completely dominant.

Once initial results were examined with a set learning rate and set exploration factor, additional experiments were run varying either factor to examine the effects. For this set of experiments, only the Q-routing with computational offloading and energy awareness algorithm was utilized. The experiments were run using the variable data rate loading and divided into sections of 20 seconds each. Table 2 shows the breakdown of the 20 second segments.

Table 2. Variable message rates by segment

Segment	Time Period	Data rate/node
1st	1-20	Pings only
2nd	21-40	5 messages/sec
3rd	41-60	Max Rate
4th	61-80	5 messages/sec
5th	81-100	Max Rate

The results of varying the learning rate between 0.1, 0.5, and 1.0 are shown in Figure 7. The data is taken from one node with only 2 Ethernet interfaces therefore the viable actions space is either forward via one of the two interfaces or compute locally, shown as "Forwarding Via Interface 1", "Forwarding Via Interface 2", and "Compute Locally" respectively.

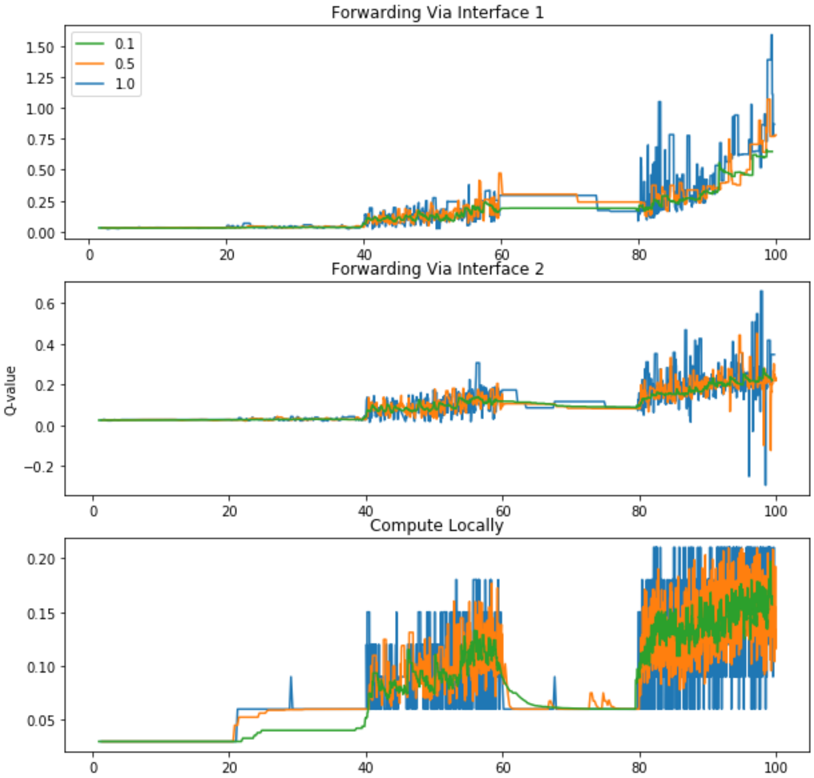


Figure 7. Effect of different learning rates on action space

The results show the increase in variability of the Q-value for higher values of learning rate. Given the Q-value is used to predict the future reward of the a given routing decision, higher variability may not be desirable. For lower values of learning rate, variability is obviously markedly decreased however, the system requires a long time to recover from transition points. A good example of this can be seen at the 60 second mark in the "Compute Locally" section. When the network transitions from high data rate to low data rate the Q-value for computing locally takes approximately 5 seconds to recover. The difference in variability between forwarding and computing is also notable indicating separate learning rates for both processes may be more ideal. From this simple experiment it can be surmised that a learning rate closer to 0.1 is preferable for the forwarding actions while a learning rate closer to 0.5 may be preferable to allow quicker recovery during transition periods. Variable learning rates dependent on network loading could be examined further as an alternate option.

The effect of varying the exploration parameter, ε, was also examined. Epsilon was varied from 0.0, which equates to no exploration and only exploiting the optimal known Q-value, to 0.5 which equates to randomly exploring different actions half of the time. The results are shown in Figure 8.

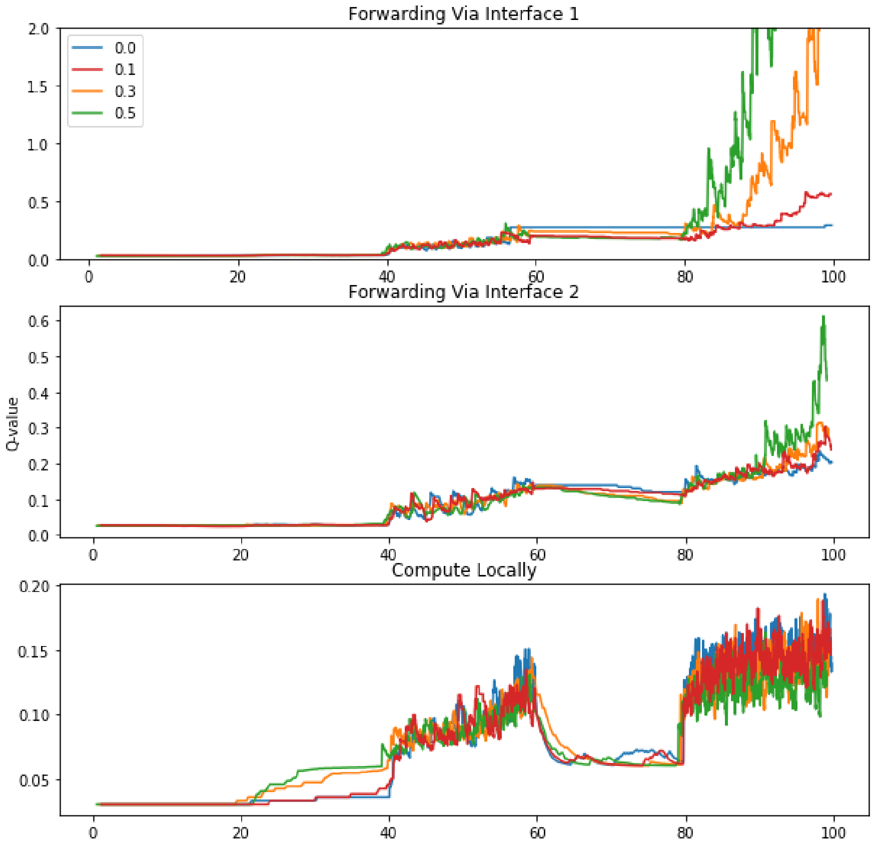


Figure 8. Effect of different ε on action space

The results of varying ε highlight interesting patterns. For the first several sections of the test, the different values have minimal effect on the reward received for a given action, but later in the test as energy reserves begin to deplete there is a large divergence between different values, particularly notable in Interface 1. This is likely due to the cost of forwarding increasing rapidly while energy reserves deplete and randomly selecting forwarding during exploration results in a higher Q-value or lower reward. The option to compute locally seems to benefit from a higher exploration rate. This is likely due, again, to the influence of the energy awareness. As the energy reserves deplete the optimal decision is heavily skewed toward computing locally, however random exploration chooses more messages to be forwarded thereby allowing the compute queue to shrink and lowering the compute locally Q-value. For this particular experiment it appears an ε of 0.1 is ideal as it keeps all options relatively balanced particularly during high data rates and low energy reserves.

5. CONCLUSIONS

This paper has presented a variation of Q-routing algorithm with a focus on usability for WSNs. The addition of the option for the algorithm to determine if computational offloading, instead of forwarding, allows optimization of total processing time while not requiring any one node in a multi-hop network to bear the cost of performing the data computation. The result of this addition allows the EAQCO algorithm to optimize for both time and energy simultaneously, while the energy awareness factor alters the action space to account for depleting energy reserves. The results of the simple experiments performed have shown that EAQCO is a viable option when energy, time, and computational capacity are critical factors as is often the case for a WSN.

Additionally, the simplicity of this algorithm allows it to be integrated into other WSN routing protocols particularly other variations of the Q-routing algorithm, further optimizing the capability of the network.

Based on the results of our experiments we intend to continue to expand the action space of EAQCO algorithm in future research to include store-and-carry-forward options in addition to forwarding and computing locally while simultaneously integrating other quality of service metrics such as variable transmission power and link reliability to expand the possible state space. These additional metrics could further increase the utility of the EAQCO for future deployment in WSN.

REFERENCES

[1] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy efficient communication protocol for wireless microsensor networks,” in Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Jan. 2000, pp. 10 pp. vol.2.

[2] C. Ambekar, D. Mehta, and H. Ashar, “OPEGASIS: Opportunistic Power Efficient Gathering in Sensor Information Systems,” in 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), Jan. 2015, pp. 1–5.

[3] N. M. Shagari, M. Y. I. Idris, R. B. Salleh, I. Ahmedy, G. Murtaza, and H. A. Shehadeh, “Heterogeneous Energy and Traffic Aware Sleep-Awake Cluster-Based Routing Protocol for Wireless Sensor Network,” IEEE Access, vol. 8, pp. 12 232–12 252, 2020.

[4] N. Sadagopan and B. Krishnamachari, “Maximizing Data Extraction in Energy-Limited,” International Journal of Distributed Sensor Networks, vol. 1, Apr. 2004.

[5] L. Kong, J.-S. Pan, V. Sn´aˇsel, P.-W. Tsai, and T.-W. Sung, “An energy aware routing protocol for wireless sensor network based on genetic algorithm,” Telecommunication Systems, vol. 67, no. 3, pp. 451–463, Mar. 2018.

[6] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya, “Cloud-Based Augmentation for Mobile Devices: Motivation, Taxonomies, and Open Challenges,” IEEE Communications Surveys Tutorials, vol. 16, no. 1, pp. 337–368, 2014.

[7] E. Baccarelli, N. Cordeschi, A. Mei, M. Panella, M. Shojafar, and J. Stefa, “Energy-efficient dynamic traffic offloading and reconfiguration of networked data centers for big data stream mobile computing: review, challenges, and a case study,” IEEE Network, vol. 30, no. 2, pp. 54–61, Mar. 2016.

[8] N. Shan, Y. Li, and X. Cui, “A Multilevel Optimization Framework for Computation Offloading in Mobile Edge Computing,” Jun. 2020, iSSN: 1024-123X Library Catalog: www.hindawi.comPublisher: Hindawi Volume: 2020.

[9] M. Littman and J. Boyan, “A distributed reinforcement learning scheme for network routing,” in In Proceedings of the 1993 International Workshop on Applications of Neural Networks to Telecommunications. Erlbaum, 1993, pp. 45–51.

[10] R. Desai and B. P. Patil, “MANET with Q Routing Protocol,” International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS), vol. 3, no. 2, pp. 255–262, Feb. 2013.

[11] T. Hendriks, M. Camelo, and S. Latr´e, “Q2-Routing: A Qos-aware Q-Routing algorithm for Wireless Ad Hoc Networks,” in 2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Oct. 2018, pp. 108–115, ISSN: 2160-4886.

[12] S. Khianjoom and W. Usaha, “Anycast Q-routing in wireless sensor networks for healthcare monitoring,” in 2014 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), May 2014, pp. 1–6.

[13] Xiaolong Xu et al. “Multi-objective computation offloading for Internet of Vehicles in cloud-edge computing,” in Wireless Networks26.3 (Apr. 2020), pp. 1611–1629. ISSN: 1572-8196.doi:10.1007/s11276-019-02127-y.

[14] Z. Hong, W. Chen, H. Huang, S. Guo and Z. Zheng, "Multi-Hop Cooperative Computation Offloading for Industrial IoT–Edge–Cloud Computing Environments," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 12, pp. 2759-2774, 1 Dec. 2019, doi: 10.1109/TPDS.2019.2926979.

[15] A. Carroll and G. Heiser, “An Analysis of Power Consumption in a Smartphone,” p. 14.

[16] J. Vazifehdan, R. V. Prasad, M. Jacobsson, and I. Niemegeers, “An Analytical Energy Consumption Model for Packet Transfer over Wireless Links,” IEEE Communications Letters, vol. 16, no. 1, pp. 30–33, Jan. 2012.

[17] Lora Alliance Technical Committee, “LoraWAN 1.1 Specification,” Oct. 2017. [Online]. Available: https://lora-alliance.org/sites/default/files/2018-04/lorawantm specification -v1.1.pdf

[18] “BeagleBoard.org - black.” https://beagleboard.org/black

FRUSTRATION INTENSITY PREDICTION IN CUSTOMER SUPPORT DIALOG TEXTS

Janis Zuters and Viktorija Leonova

Department of Computer Science, University of Latvia, Riga, Latvia

ABSTRACT

This paper examines the evolution of emotion intensity in dialogs occurring on Twitter between customer support representatives and clients ("users"). We focus on a single emotion type—frustration, modelling the user's level of frustration (on scale of 0 to 4) for each dialog turn and attempting to predict change of intensity from turn to turn, based on the text of turns from both dialog participants. As the modelling data, we used a subset of the Kaggle Customer Support on Twitter dataset annotated with per-turn frustration intensity ratings. For the modelling, we used a machine learning classifier for which dialog turns were represented by specifically selected bags of words. Since in our experimental setup the prediction classes (i.e., ratings) are not independent, to assess the classification quality, we examined different levels of accuracy imprecision tolerance. We showed that for frustration intensity prediction of actual dialog turns we can achieve a level of accuracy significantly higher than a statistical baseline. However we found that, as the intensity of user's frustration tends to be stable across turns of the dialog, customer support turns have only a very limited immediate effect on the customer's level of frustration, so using the additional information from customer support turns doesn't help to predict future frustration level.

KEYWORDS

Neural Networks, Emotion Annotation, Emotion Recognition, Emotion Intensity, Frustration.

1. INTRODUCTION

With the growing popularity of social networks and the exponential increase of user-generated content volume, automated language understanding is becoming ever more relevant. And emotion recognition plays no small part in this understanding. By their nature, humans are emotional beings, and emotions are very important for interpersonal communication. For this reason, many researchers have studied automatic emotion annotation, probably for as long as the machine learning field has existed. Most of these researchers have focused on variants of Ekman's emotion classification schema [1], annotating texts with several basic emotions. However, being interested in a specific task — namely, conversations between customers and customer support representatives — we concentrate on one specific emotion, frustration, and how it changes over the course of a dialog. The reason for this is that the main indicator of success for customer support is customer satisfaction or dissatisfaction, where dissatisfaction is captured by the emotion that we label as frustration.

In this work, we examine two hypotheses:

1. In customer support dialogs, the user's turn-by-turn frustration intensity can be predicted from the text of the user's message, and, in particular, from the presence of keywords – a

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 261-273, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101419

set of words (including also emojis and other non-lexical textual tokens) that correlate with specific frustration intensity levels.

2. In customer support dialogs, the frustration intensity of the user's current turn can be predicted from keywords in the user's previous turn together with keywords (from a different set) in the intervening turn from customer support. This targets the intuition that the manner in which the customer support representative responds to the user's utterances should have some effect on the user's emotional state going forward.

To test these hypotheses, we built a machine learning model and trained it on a dataset annotated specifically for this purpose, running a series of experiments as described in Section 5, Experiments and Results.

This paper is structured as follows: in Section 2 "Background and Related Work" we examine the previous works in the field of the emotion recognition and emotion intensity annotation, including the evolution of emotion in dialogs and available datasets. In Section 3 "Data Selection and Annotation" we explain how we the dataset for training the model was selected and annotated. Section 4 "Frustration Intensity Prediction" explains the concept of frustration used in our research, the definition of frustration intensity and its evolution is given, and the main terms are introduced. Section 5 "Experiments and Results" provides the detailed description of conducted experiments, models constructed, and results achieved. In Section 6 "Discussion" we discuss the results provided in Section 5 and their interpretation. Finally, Section 7 "Conclusions and Future Work" gives a short summary of this work, results achieved and their possible development.

2. BACKGROUND AND RELATED WORK

Virtually since the beginning of Machine Learning (ML) research, there have been attempts to apply ML to emotion annotation, first of speech (as the easier task, since speech signals carry additional, paraverbal information about the speaker's emotional state) and then also of text, as early as in 2005 by Alm et al. [2]. Most such researches used one or another version of Ekman's six emotion model [1]. Examples include Balahur et al., 2013 [3], Kao et al., 2009 [4] and others. With the development of social networks, the focus of work in emotion annotation has shifted toward emotion annotation in messages posted by users in social networks, such as Facebook, e.g. Al-Mahdawi and Teahan, 2019 [5], Weibo, e.g. Lee and Wang, 2015 [6] or Twitter, like Duppada and Hiray, 2017 [7], with Twitter being one of the most fruitful sources due to the open and concise nature of the posts it supports: short texts, sometimes accompanied by a picture or self-annotated with hashtags. Such self-annotations can even be used as the foundation for gold standard corpus labelling, as done by Gonzalez-Ibanez et al. in 2011 [8]. Several emotions have found their way into automated annotation, especially the basic emotions as identified by Ekman (fear, anger, joy, disgust, surprise and sadness), as for example Badaro et al., 2019 [9]. And even such elusive notions as sarcasm and irony have been researched, for example by Reyes et al., 2013 [10]. Frustration, however, has not been widely researched. There have been a few papers focusing on frustration, such as Klein et al., 2006 [11], or Hone, 2002 [12], but not many. Hu et al., 2018 [13] discuss the correlation between the emotional tone of customer support messages and user messages, and the tones they study include frustration among others. We believe that, especially in the field of business communication, automatic frustration recognition targets a relatively unaddressed need.

Whereas much earlier work sought primarily to output binary, categorical labels (predicting the presence or absence of specific emotions), labelling and predicting gradations of emotion intensity is only recently becoming more widespread. Examples include Goel et al., 2017 [14], Bravo-Marquez et al., 2019 [15], and Badaro et al., 2019, analysing emotion intensity in tweets and providing a Weka package for automatically annotating tweets with intensities ratings for anger, fear, joy, and sadness. However, as there has been little work on frustration recognition in general, automatic recognition of frustration intensity mostly remains unaddressed — one exception being the aforementioned Hu et al., 2018, who annotated and modelled intensities for 8 differing emotional "tones" (or language production styles): anxious, frustrated, impolite, passionate, polite, sad, satisfied, and empathetic; our work differs from theirs in that we focus exclusively on frustration, while they explore correlations between the user's vs. the support agent's tone for all pairwise combinations of these 8 tones. Their work and ours also differ in the methods used for selecting keywords associated with a given tone or emotion, and in the architecture and goal of the machine learning models developed. Whereas they train a seq2seq model (sequence-to-sequence, using a recurrent neural network) for generating dialog responses with specified tones, we develop relatively simpler neural models for predicting user frustration gradations given previous user + support agent turns (their analysis of correlations between user vs support agent tones is carried out via linear regression.)

While there are several publicly available dialog datasets, for example Taskmaster-1 [16] or DailyDialog [17], none have directly addressed the modelling of participants' turn-to-turn emotional state dynamics in a goal-oriented context, to the best of our knowledge. With respect to dialog datasets and research on automated dialog agents (or "chatbots") an important distinction is often drawn between goal-oriented dialog agents (where the user is seeking to accomplish some task with assistance from the automated agent) vs. free-chat agents (which attempt to simulate human-style conversations with users, "chatting" with no specific goal other than entertainment, or, possibly, some kind of therapeutic objective). The labelling and structure of the datasets associated with each of these chatbot types are, in general, very different. (Taskmaster and DailyDialog are prototypical examples of datasets for goal-oriented vs free-chat agents, respectively). In one case, the primary focus is on identifying the user's 'intent' (what she is trying to achieve) and shaping further interactions to elicit whatever additional information might be required to complete it. Free-chat agents, on the other hand, are mostly concerned with generating responses that simulate what a human conversational partner might say in the same situation. The free-chat setting is where most previous research on identifying emotions and generating responses with emotionally appropriate language has been done.

Customer support agents can be viewed as a hybrid of goal-oriented and free-chat agents, in that the client usually does have a specific objective (resolving or at least reporting a specific problem), but emotional dynamics are also very important: in the final analysis, the primary objective of the dialog agent can be formulated as an emotional state ("client satisfaction"). Automated goal-oriented dialog agents have been studied in quite a few works, for example Ham et al., 2020 [18], as have affect-driven free-chat dialog agents e.g. Colombo et al., 2019 [19], and Lubis et al., 2018 [20], focusing on providing affect-sensitive responses, but very few works have investigated dialog agents that attempt to address both concerns simultaneously [21], [13].

3. DATA SELECTION AND ANNOTATION

For our research, we reviewed a number of publicly available conversation-based datasets and selected, as a basis for additional annotation, the Kaggle Customer Support on Twitter dataset[1]. In the modern world, customer support via social media is becoming increasingly popular, and it would certainly be valuable to be able to automatically gauge a customer's frustration level,

[1] https://www.kaggle.com/thoughtvector/customer-support-on-twitter

ideally enabling an automated agent to increase customer satisfaction by tailoring dialog responses appropriately. As a first step, we assembled dialogs from the raw tweet data contained in the dataset, by automatically linking the messages by their ids and reply ids, so that they would constitute a complete conversation. After that, these conversations were filtered to exclude examples where more than one user participated in a specific conversation, and so that each of the dialogs contained no less than two user turns, separated by a customer support turn in between. This was done in order to enable modelling of the effect that a customer support turn might have on the emotional state of the specific user.

Having prepared the conversations in this manner, we selected a subset of four hundred of them for annotation. We simply took the first four hundred, as the source data was not organized in any specific way, so choosing in this way provided an essentially random sample while allowing to extend the dataset as needed by simply adding subsequent samples. The conversations were then anonymized and unified by replacing the user Twitter ids and support ids with generic “USER” and “SUPP” labels, respectively. In cases of dialogs containing several sequential user or support messages, they were joined together, so that the sequence of turns was always USER -> SUPP -> USER -> SUPP -> etc. Other sensitive information, such as email addresses, had been already replaced in the Kaggle Customer Support dataset by generic placeholders like “__email__”. Each of the dialogs was assigned a unique id. Files with prepared dialogs were sent to three annotators along with the instruction on their annotation. The annotators were asked to assign a single value to each of the customer turns. The values could be integers from 0 to 4, marking a customer’s frustration level as perceived by the annotator, where zero was to mean that that customer is satisfied or is in a neutral emotional state, while four indicated ultimate frustration. Another allowed value was “n”, which meant that it is not possible to make a conclusion about the customer’s emotional state from the message, e.g. in the case of giving single-word answers or providing purely technical information in response to a question. In addition, there was a possibility to leave the value empty, which meant that the annotator could not interpret the message, for example, in case if the language was not known to him or the text was in some other way not comprehensible.

After the annotated files were received back from the individual annotators, they were combined into a single master file, in which every user turn in every conversation was associated with three assigned values, one from each annotator (note that some of these values could be ‘n’ or blank, as previously described). This file was then further filtered to exclude dialogs that didn’t meet the criteria for dialog length, and keeping just the conversations involving only a single user with one customer support representative — yielding a total of 376 dialogs, with an average dialog length of 5.2 turns.

4. Frustration Intensity Prediction

This section describes the proposed approach for frustration intensity prediction for dialogs, as well as experiments conducted with the purpose to validate the concept.

4.1. Method Overview and Data Preparation

The research focuses on frustration intensity prediction for user-side turns in Twitter-originated customer support dialogs and includes two different tasks:

1. actual frustration intensity prediction – frustration intensity prediction given actual text (of the turn),

2. frustration intensity dynamics prediction – frustration intensity prediction given texts of the previous two turns (user's and support's, respectively).

An annotated dialog d is represented by a sequence of 2-tuples d_1, d_2, ..., d_n, where

- d_i represents one turn - odd turns are user's turns, even turns are support's turns,
- d_i is a tuple containing
    - $d_i(1)$ - text,
    - $d_i(2)$ - frustration intensity, an integer value in the range [0..4].

The aim of the experimentation is to show that the user's frustration intensity can be predicted from keywords found in the text – either of the current turn, or previous ones.

For our experiments, we used a corpus of 376 dialogs having an average dialog length of 5.2 turns (counting both user and support turns, e.g. 3 user turns and 2 intervening support turns). This dataset included 1038 annotated user turns, of which 843 were selected as valid for the modelling process as they were rated with a numeric value by all 3 annotators (we excluded turns that received an 'n' or didn't receive a rating from one or more annotators), as well as 470 valid support turns (support turns occurring between two valid user turns were considered valid for the dynamics prediction modelling task). The distribution of ratings (0, 1, 2, 3, 4) over the 843 valid user turns were (155, 125, 234, 239, 90), respectively; thus, rating values of 2 and 3 are the most common and are almost equally frequent. For the 470 valid support turns, the average frustration intensity change from the previous user turn to the current one was -0.35, so that, in general, frustration intensity is observed to decrease from turn to turn, but only slightly. Over the course of a short dialog, the user's frustration intensity rating is, on average, expected to remain essentially unchanged.

Individual dialog turns for our predictive models were represented using a bag-of-words encoding, using keywords/tokens from selected subsets of the overall vocabulary (encoded as binary vectors, from two separate vocabularies: one for user texts, V_{user}, and another for customer support texts, V_{supp}.

The vocabularies were constructed by selecting from lower-cased tokens occurring at least 3 times in the corresponding valid turns (user's and support's, respectively). This criterion resulted in base vocabulary sets with cardinalities $|V_{user}| = 941$, and $|V_{supp}| = 450$. Only the k 'best' tokens from these vocabularies were used for text encoding – the first k tokens when ranked according to increasing standard deviation of the ratings assigned to turns containing these particular tokens. These thresholds, k_{user} and k_{supp} , served as two of the hyperparameters for our models (with hidden layer size being another), which we evaluated over the ranges: k_{user} in [50 to 700] tokens for user turns, and k_{supp} in [50 to 350] for support turns.

4.2. Quality Measures and Experiment Tasks

For both classification tasks (predicting integer frustration intensity values) we used the following quality measures:

1. absolute accuracy,
2. accuracy with tolerance +/- 1 (so that an "off-by-one" prediction is also considered correct) — as individual intensity grades are not actually independent classes, but form an ordered sequence, this measure seems more adequate (e.g. predicting 2 when the

"correct" rating is annotated as 3, is not equally wrong to predicting 0 in the same situation).

Experiment Task #1: Frustration Intensity Prediction (see Fig. 1).

Task description:

- Given the user's turn text (encoded as described in Section Data-Prep),
- Predict the frustration intensity of this turn.

Baseline:

- For the 'exact' accuracy – predict the most statistically frequent rating, i.e. 3 (as per Section Data-Prep) for all inputs;
- For accuracy with tolerance +/-1, predict frustrationequal to 2 for all inputs as it is the most frequent in this setting.

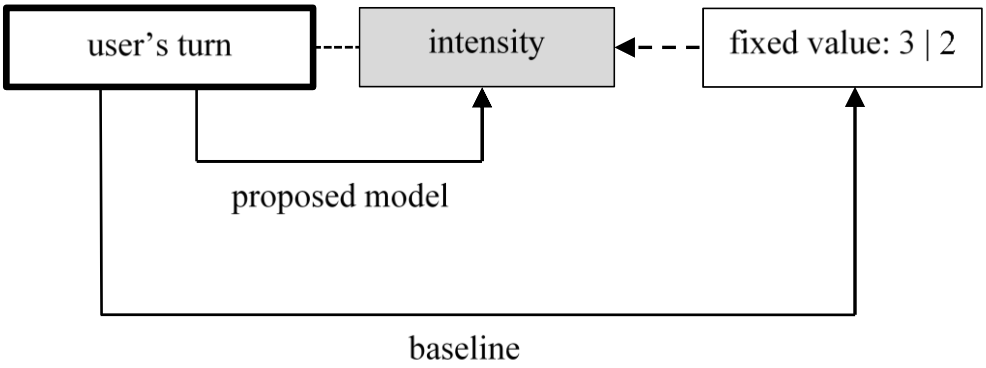


Figure 1. Modelling actual frustration intensity prediction.

Experiment Task #2: Frustration Intensity Dynamics Prediction (see Fig. 2).

Task description:

- Given the user's turn text, and the following support's text (both encoded as per Section Data-Prep);
- Predict the frustration intensity of the next user's turn.

Baseline:

- Predicted frustration of the initial user's turn (obtained the way task #1 is being solved) returned — so that the reference model computes the actual user's frustration intensity and regards that it won't change in the next turn.

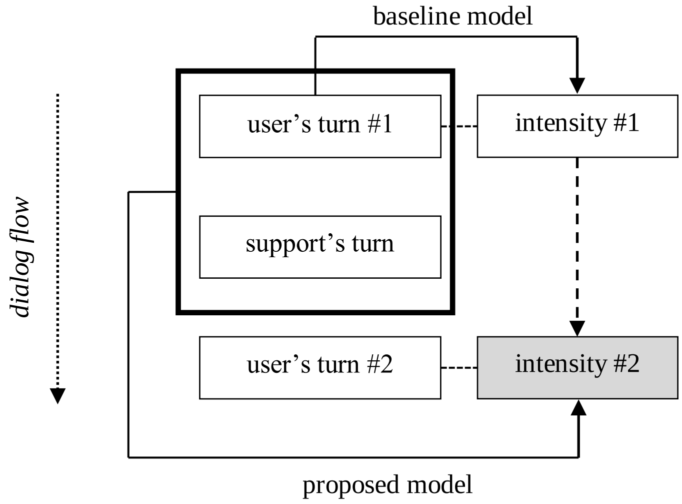


Figure 2. Modelling frustration intensity dynamics.

5. EXPERIMENTS AND RESULTS

5.1. Experiment Configuration and Flow[2]

The experimentation was carried out in two phases:

1. Preparation:
    a. hyperparameter search for neural network models,
    b. selection of the set input configurations to represent dialog data (as shown in column #1 of Tables 1 and 2);
2. Main run – run experiments with selected hyperparameters and with every selected input configuration.

5.1.1. Experiment Preparation

For each of the two experiment tasks, a hyperparameter search covering several thousand experimental runs was conducted, to select a final model consisting of a multi-layered perceptron with one hidden layer with the following final configuration:

1. Input: binary input of several hundreds of values (as per Section "Method Overview and Data Preparation") representing the text of one or two dialog turns — as amounts of keywords per turn:

    a. for experiment task #1 we have selected the following input configurations: 50, 100, 300, 500 (see Table 1) for the number of keywords to represent a user's turn,
    b. for experiment #2 we have selected the following input configurations: 50/50, 200/100, 500/200, 700/350, meaning that to train the baseline model input configurations 50, 200, 500, 700 were used respectively (see Table 2) —the number of keywords to represent the previous user's turn/the number of keywords to represent the following support's turn:

[2] Source code is available at https://github.com/zuters/dfrustration

2. Hidden layer: 64 neurons;
3. Output: categorical of 5 possible values representing frustration intensity (0..4);

The number of epochs for each experiment: 50.

5.1.2. Experiment Main Run

With the obtained model architecture, we have conducted a further series of experiments using a different input—encoding configurations as selected in the preparation phase.

For each input configuration of each of the two experiment tasks, we used Leave-one-out cross-validation to evaluate the model for the average accuracy (see Section "Quality Measures and Experiment Tasks"). Experimentation for a fixed input configuration consisted of the following steps:

- For all annotated n data points in the dialog dataset relevant to the experiment (as for Section "Quality Measures and Experiment Tasks"):
    - Prepare the data for the proposed (target) model:
        - the current data point is reserved for testing:
            - for task #1 – a data point is one user's turn in a dialog to predict the current intensity (as in Fig. 1),
            - for task #2 – a data point is the current user's turn, as well as the following support's turn to predict the next intensity (as in Fig. 2);
        - the rest of n-1 data points go for training;
    - Prepare the data for the baseline:
        - for task #1, a fixed baseline value is used – the most common label in the dataset (Baseline columns in Table 1),
        - for task #2, separate data for the baseline model are prepared (current user's turn only);
    - Train the models for 50 epochs:
        - for task #1, only the target model is trained (as the baseline is fixed),
        - for task #2, the baseline model is also trained;
    - Collect the experiment results:
        - For task #1 – apply the model to the test data and collect accuracy measurements (Result columns in Table 1),
        - For task #2 – apply both models to the test data and collect accuracy measurements:
            - target model accuracy (Result columns in Table 2),
            - baseline accuracy (Baseline columns in Table 2).

Evaluate the input configuration: the final result is the average accuracy of the n models of the input configuration (as obtained using Leave-one-out cross-validation).

5.2. Experiment Results

When running our series of experiments, we found that the results for repeated runs using a given configuration generally varied only within a range of one percent, so here we report all results rounded to whole numbers (see Tables 1 and 2).

Experiment Task #1: Frustration Intensity Prediction.

Table 1. Results for frustration intensity prediction.

Input configuration: user keyword count	Accuracy, %		Accuracy with tolerance 1, %	
	Result	Baseline	Result	Baseline
50	37	28	74	71
100	41		78	
300	41		80	
500	41		80	

Experimental results show that:

- Frustration intensity can be effectively predicted from the presence of selected keywords;
- 100 keywords can be sufficient for predicting the frustration with the 'exact' accuracy (with no tolerance);
- Using more keywords gives better results for accuracy with tolerance.

Experiment Task #2: Frustration Intensity Dynamics Prediction.

Table 2. Results for frustration intensity dynamics prediction.

Input configuration: user keyword count / support keyword count	Accuracy, %		Accuracy with tolerance 1, %	
	Result	Baseline	Result	Baseline
50/50	34	28	58	67
200/100	34	30	62	70
500/200	34	33	68	70
700/350	30	31	65	69

Experimental results show that:

- Frustration intensity can be to some extent predicted from presence of selected keywords in the user's previous turn (baseline model);
- Using additional keywords from the customer support turn doesn't improve the predictions.

6. Discussion

In this work, we have constructed a neural network-based model for predicting user frustration intensity from the text of a user tweet addressed to a customer support. This model takes an encoded representation of the user message as an input and gives an output in the form of an integer rating of frustration intensity on a 5 point scale (0 to 4), achieving a precision of 41%, 14% higher than a baseline which simply assigns the most frequent label to all instances. In addition to exact precision, we also calculate precision with tolerance (allowing a difference of 1 between the actual and predicted rating). Using this "+/-1 accuracy" metric, our model achieves 80%, 9% higher than the baseline (71% using this metric). This allows us to say that to a certain degree frustration intensity can be predicted from the text of a user's message precisely, and in 80% of cases it can be predicted approximately.

In addition, we have constructed another neural network-based model that predicts the user's emotional state dynamics from the contents of the support agent's reply to a preceding message from the user. From encoded representations of the user's message and the support agent's message, it attempts to predict the frustration rating that annotators assigned to the next (user's) turn. As the baseline model we have used the prediction of the frustration intensity for the initial user message, under the assumption that the user’s frustration remains unchanged. The achieved precision was 34%, a very slight (1%) improvement over the baseline. Also, for this scenario, allowing +/- 1 tolerance in the predicted frustration intensity doesn’t improve over the baseline (just using the prediction for the initial message is better), thus implying that knowing the contents of the support agents message provides no additional useful information toward predicting changes in the user’s state of frustration (and which, in general, does not significantly change from one turn to the next).

We have already noted the overall tendency for the user's level of frustration tends to remain mostly unchanged from turn to turn. We hypothesize that this might be at least partially explained by the fact that customer support representatives are already formulating their replies with the goal of trying to reduce, or at least to not increase, the customer's frustration or level of dissatisfaction with their company's products or services (they are, in fact, often trained and explicitly motivated to do so).

Manually examining our data in more detail, we find only 7 examples of dialogues where the user's level of frustration has been labelled as changing for the worse by more than 1 point from one turn to the next (in all such examples the increase is +2 points; there are no examples of a jump of +3 or +4 points). A change in rating for the better is relatively more common: there are 44 examples of turn-to-turn transitions with a -2 delta (where the user's level of frustration has decreased by two points), 13 with -3, and one with -4 (which would mean that the user started out maximally frustrated/dissatisfied but transitioned to being completely satisfied within a single dialog turn). Some examples of such exceptional dialogs can be seen in Appendix 1.

But such outlier transitions are the exception rather than the rule — the overall finding in terms of turn-to-turn dynamics is well illustrated by the relatively strong performance of our baseline model, which simply assumes that the user’s frustration level will remain unchanged from the previous turn.

7. Conclusions and Future Work

In this paper, we have presented a new dataset — a subset of the Kaggle Twitter Customer Support dialogs consisting of close to 400 dialogs and comprising almost 900 individual customer tweets, annotated for frustration intensity on the scale of 0 to 4. We have selected the most popular grade as a baseline and demonstrated that frustration intensity can be predicted based on the contents of an individual tweet with an accuracy significantly higher than the baseline (41% compared to 27%). This result was achieved by constructing a neural network and training a simple classification model. We also examined the effect of customer support turns on the emotional state of the user and found that, typically, the user’s emotional state mostly remains unchanged, with a small decrease of 0.34 points on average from one turn to the next. Currently, in contrast to our generally positive finding for predicting turn-by-turn frustration ratings from text-based features, we conclude that, given the challenges in precise calibration of the user’s frustration level — due at least partially to the subjective and fleeting nature of the emotion itself and the difficulty of estimating it by a third party purely from the text of a conversation, trying to model this dynamic as a function of the emotional valence of the support agent's messages doesn't yield any strong results (at least not using classification models like the neural models we tried).

In the future, we are looking towards possibly adapting and applying this methodology to dialogs in Latvian, Latvian being a low-resource language where practically no work on automatic emotion annotation with machine learning methods has been undertaken, and analysing the effect of another language on the accuracy of automatic annotation of frustration level, and on the feasibility of predicting the dynamics of the user's emotional state.

ACKNOWLEDGEMENTS

The research has been supported by the European Regional Development Fund within the joint project of SIA TILDE and University of Latvia "Multilingual Artificial Intelligence Based Human Computer Interaction" No.1.1.1.1/18/A/148.

REFERENCES

[1] P. Ekman, (1992) "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp.169–200.

[2] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text," *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT 05, 2005,* pp.579–586.

[3] A. Balahur, J. M. Hermida, A. Montoyo, and R. Muñoz, (2013) "Detecting implicit expressions of affect in text using EmotiNet and its extensions," *Data & Knowledge Engineering*, Vol. 88, pp.113–125.

[4] E. C.-C. Kao, C.-C. Liu, T.-H. Yang, C.-T. Hsieh, and V.-W. Soo, (2009) "Towards Text-based Emotion Detection A Survey and Possible Improvements," *2009 International Conference on Information Management and Engineering*, pp.70-74.

[5] A. Al-Mahdawi and W.J. Teahan (2019) "Automatic emotion recognition in English and Arabic text" (*Doctoral dissertation, Bangor University*).

[6] S. Lee and Z. Wang, (2015) "Emotion in Code-switching Texts: Corpus Construction and Analysis," *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pp.91-99.

[7] V. Duppada and S. Hiray, (2017) "Seernet at EmoInt-2017: Tweet Emotion Intensity Estimator," *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp.205-211.

[8] R. González-Ibánez, S. Muresan, and N. Wacholder, (2011). Identifying sarcasm in Twitter: a closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* pp.581-586.

[9] G. Badaro, H. Jundi, H. Hajj, and W. El-Hajj, (2018) "EmoWordNet: Automatic Expansion of Emotion Lexicon Using English WordNet," *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics,* pp.86-93.

[10] A. Reyes, P. Rosso, and T. Veale, (2012) "A multidimensional approach for detecting irony in Twitter," *Language Resources and Evaluation,* Vol. 47, No. 1, pp. 239–268.

[11] J. Klein, Y. Moon and R.W. Picard, R.W., (2002). "This computer responds to user frustration: Theory, design, and results." Interacting with computers, Vol. 14, No. 2, pp.119-140.

[12] K. Hone, (2006). "Empathic agents to reduce user frustration: The effects of varying agent characteristics." *Interacting with computers*, Vol. 18, No. 2, pp.227-245.

[13] T. Hu, A. Xu, Z. Liu, Q. You, Y. Guo, V. Sinha, J. Luo, and R. Akkiraju, (2018) "Touch Your Heart," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18,* pp.1-12.

[14] P. Goel, D. Kulshreshtha, P. Jain and K.K. Shukla, (2017) Prayas at emoint 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis,* pp. 58-65.

[15] F. Bravo-Marquez, E. Frank, B. Pfahringer, and S. M. Mohammad, (2019). Affective tweets: A weka package for analyzing affect in tweets. *Journal of Machine Learning Research*, Vol. 20, No. 92, pp.1–6.

[16] B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, B. Goodrich, D. Duckworth, S. Yavuz, A. Dubey, K.-Y. Kim, and A. Cedilnik, (2019) “Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).*

[17] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, (2017) Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957.*

[18] D. Ham, J. G. Lee, Y. Jang, and K. E. Kim, (2020) End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.583–592.

[19] P. Colombo, W. Witon, A. Modi, J. Kennedy, and M. Kapadia, (2019) “Affect-Driven Dialog Generation,” *Proceedings of the 2019 Conference of the North*, pp.3734–3743

[20] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, (2018). Eliciting positive emotion through affect-sensitive dialogue response generation: *A neural network approach. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp.5293–5300.

[21] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, (2017) “A New Chatbot for Customer Service on Social Media*,” Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 3506–3510.

APPENDIX 1

Turns for the worse: frustration rating increases by +2 points

Major transitions for the worse (+2 delta in frustration rating) seem often to be situations where the customer support representative tells the user to do something that he or she has already tried. This, presumably, is something that the customer support representative had no way of knowing and is certainly not part of the information available in the input data for our model (which sees only the text of the preceding turns of the dialogue). Another pattern we noted is when the user was probably in fact more frustrated than his first question or statement suggests, but expresses his full frustration only in a subsequent turn. Once again, this is not something that the customer support agent (or a machine learning model) is likely to be able to anticipate.

Outlier transitions for the better (-3 or -4 delta in frustration rating), on the other hand, in general seem to be due to something that has happened in the real world (as opposed to in the dialog) to resolve the user's complaint (e.g. the user found a way to resolve it themselves, or the problem got otherwise resolved in the meantime).

TWCS-T1466 (DELTA: +2)

USER:	i have bought dlc diablo 3 necromancer and hav phys disc D3 ROS but when i click necromaner pack on game , dont download https://t.co/JDdn50e0wo	1
SUPP:	Hi there, Have you tried accessing this content from your download queue: https://t.co/lyTEIVBTBn Let us know the results.	
USER:	I have followed your instructions, but not, my ps4 has set up automatic download, and find in library there is no diablo 3 nercromancer my psn id quochuy046, can you help me? Or can I email someone a try for help? live chat on web block me? wtf?	3
SUPP:	Hi there. Please follow the steps in the next link: https://t.co/PR9L0S0kEu Let us know the outcome!	

TWCS-T3988 (DELTA: +2)

USER:	can you please provide me with your complaints procedure?	2
SUPP:	Hi Ellen, we're available on Twitter if you would like to DM us? Alternatively, you can contact our Customer Relations team using the 1/2 following link: https://t.co/SIhdl3TbaN. ^Jane 2/2	
USER:	Hi @2042 I’ve spoken to customer relations 8 times and been lied to each time re price guarantee. Have now raised complaint.	4
SUPP:	I'm sorry you're unhappy with our price guarantee service, Ellen. The team will respond to your complaint in due course. ^Kimbers	

Turns for the better: frustration rating decreases (-3 or -4 points)

TWCS-T1691 (DELTA: -4)

USER:	I legitimately spent an hour trying to deal with USPS cause I had 1 question and they just hung up on me or wasn’t any help, I could haveSaved my fucking time by just checking my mailbox because sure enough I got the UPS letter saying my package was in oh my gOD	4
SUPP:	Is there something that we can assist you with? DM our team ^WS https://t.co/wKJHDXWGRQ	
USER:	Nope, I’ve got my package thanks	0

TWCS-T36 (DELTA: -3)

USER:	somebody from @VerizonSupport please help meeeeee 😩😩😩😩 I'm having the worst luck with your customer service	3
SUPP:	Help has arrived! We are sorry to see that you are having trouble. How can we help? ^HSB	
USER:	I finally got someone that helped me, thanks!	0

AUTHORS

Jānis Zuters received the PhD degree in Computer Science in 2007. His research interests include machine learning, neural networks, and natural language processing. Since 1999, he has been with the University of Latvia, Faculty of Computing (since 20 17 as a Professor).

Viktorija Ļeonova is a PhD student in the University of Latvia, Computer Science Department. Acquired M.Sc. in Computer Science in Open University of Cyprus in 2015.

AN AUTOMATED DATA-DRIVEN PREDICTION OF PRODUCT PRICING BASED ON COVID-19 CASE NUMBER USING DATA MINING AND MACHINE LEARNING

Zhuoyang Han[1], Ang Li[2] and Yu Sun[3]

[1]University of California, Irvine, California, USA
[2]California State University, Long Beach, USA
[3]California State Polytechnic University, Pomona, USA

ABSTRACT

In early 2020, a global outbreak of Corona Disease Virus 2019 (Covid-19) emerged as an acute respiratory infectious Disease with high infectivity and incidence. China imposed a blockade on the worst affected city of Wuhan at the end of January 2020, and over time, covid19 spread rapidly around the world and was designated pandemic by the World Health Organization on March 11. As the epidemic spread, the number of confirmed cases and the number of deaths in countries around the world are changing day by day. Correspondingly, the price of face masks, as important epidemic prevention materials, is also changing with each passing day in international trade. In this project, we used machine learning to solve this problem. The project used python to find algorithms to fit daily confirmed cases in China, daily deaths, daily confirmed cases in the world, and daily deaths in the world, the recorded mask price was used to predict the effect of the number of cases on the mask price. Under such circumstances, the demand for face masks in the international trade market is enormous, and because the epidemic changes from day to day, the prices of face masks fluctuate from day to day and are very unstable. We would like to provide guidance to traders and the general public on the purchase of face masks by forecasting face mask prices.

KEYWORDS

Corona Virus, Machine Learning, Price Prediction, Linear Regression, Poly Regression, Data Cleaning

1. INTRODUCTION

Coronavirus [1] is a kind of RNA virus which exists widely in nature. It has the tropism of gastrointestinal tract, respiratory tract and nervous system. The coronavirus founded in December 2019 is named 2019-nCoV, which causes covid-19. The major media of COVID-19 are direct, aerosol, and contact. Direct transmission refers to the patient sneezes, speaking when the droplets were inhaled close to other people caused by infection; Aerosol transmission refers to the droplets in the air formed Aerosol, was inhaled after the infection; Contact infection is a kind of infection caused by droplets attached to the surface of articles and finally contacting the mucous membrane of eyes, mouth and nose through intermediary articles. Covid-19 is as transmissible as influenza, and because of its initial clinical manifestations of fever, dry cough and weakness, it may lead patients to mistakenly believe that they have the common cold, thus lowering their risk of infection, and delayed the best time for treatment. As a result, the virus continued to spread

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 275-283, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101420

around the world and grow rapidly in March, after an outbreak in China and the closure of the city in January.

Open problem: The price of the mask as an important epidemic prevention material should be related to one or more features. However, the relationship is unknown and different algorithms need to be tried with extensive experiments.

Solution: Machine learning [2] models were used to find the relationship between mask prices and features, which could then be applied to business situations such as buying at a low price or selling at a high price.

We concluded from our predictions that the price [3] of face masks would fall over the next four days. There is a strong correlation between mask prices and China daily cases for some time to come. By comparing the relationships between eight features and mask prices, we find out that the most closely linked function is the exponential function of China daily case. In the future, we believe using this method can effectively predict mask price and provide trade guidance for business and life demand.

The rest of the paper is organized as follows: Section 2 lists the key challenges to be solved in this problem scope. Section 3 details the solution, followed by presenting the experimental results in Section 4. Section 5 analyzes the related work and we conclude the paper in Section 6 with the future work summarized.

2. CHALLENGES

2.1. The most important global data on early infectious diseases are unreliable

Covid-19 is still controlled in China from the end of 2019 to February 2020, and the world has not started to provide large-scale testing of COVID-19 to patients, so early world data are lacking. Even if there were confirmed cases of influenza-like illness caused by coronavirus in various countries, or deaths due to symptoms caused by Covid-19, it would not be included in the statistics. As China is the world's first, worst and fastest outbreak of the disease, the confirmed figures for January and February are based on China.

2.2. China's Data May not be a Perfect Reflection of Global Models

Even if China's data were used as a sample of earlier data, given the varying levels of health system soundness in different countries around the world and the different measures proposed and implemented by health authorities for epidemic prevention, China's infectious diagnosis model does not necessarily fit all countries, especially underdeveloped countries with inadequate health systems and developed countries with slow response to epidemic prevention. Except for a few Asian countries such as China, Japan, South Korea and Singapore, most countries did not take emergency measures during this period, and thus may have contributed to the spread of the virus, so in terms of epidemiology, data for countries that have implemented measures are likely not to be exactly similar to data for countries that have not implemented measures effectively and comprehensively.

2.3. There is not Enough Data on Mask Prices

The data about the price of the mask was taken from amazon.com. I recorded the daily prices from January 21 to March 8 and collated them as data input into the algorithm. However, the data was interrupted on March 8 when Amazon announced the removal of the N95 mask, thus missing the mask prices in March and April. In this project, I used China's data from the early days of the epidemic, when the city of Wuhan was shut down from January to March and the country was on high alert, so the peak time coincided with most of the mask price data, so the lack of data from Amazon's announcement about the removal of the masks does not unduly affect the accuracy of the predictions.

3. SOLUTION

3.1. Overview of the Solution

The CDC first obtained daily figures for confirmed cases and deaths worldwide from January1st through April 11th. The features are then obtained through data cleaning, and a fitting algorithm is used to guess which model fits. The relationship between the number of cases and mask price was predicted by combining the mask price data as dependent variable and independent variable.

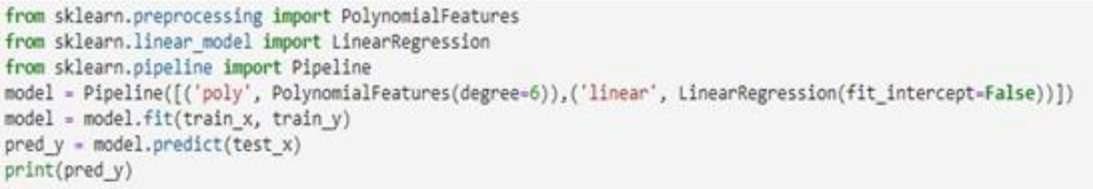
3.2. Machine Learning Model and Feature Selection

The following features have been identified in the data model construction:

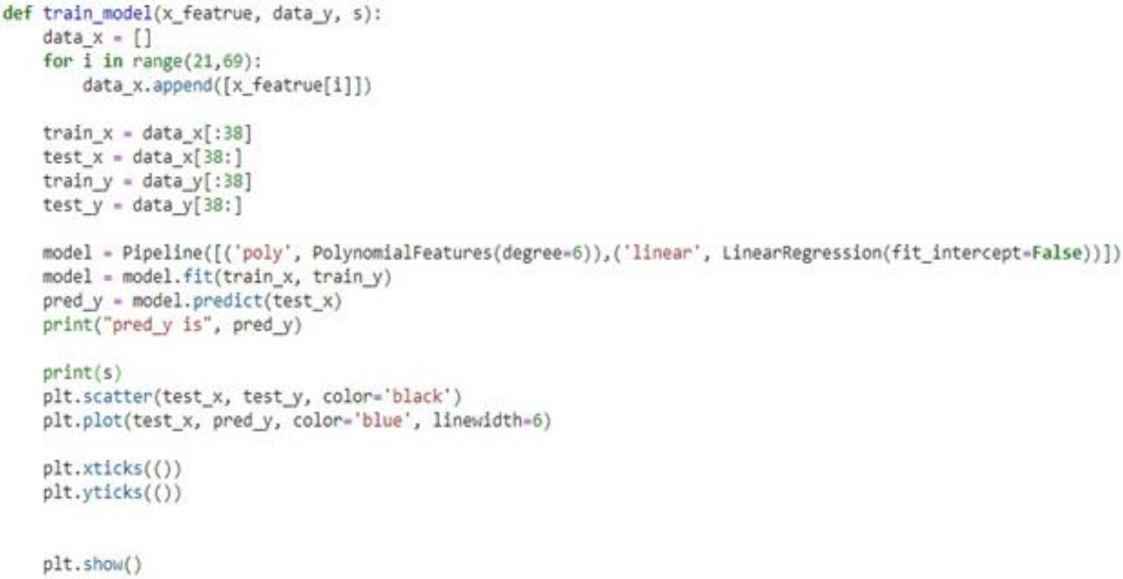
1) China daily case
2) China total case
3) China daily death case
4) China total death case
5) World daily case
6) World total case
7) World daily death case
8) World total death case
9) Mask price/5pcs

3.3. Training and Prediction

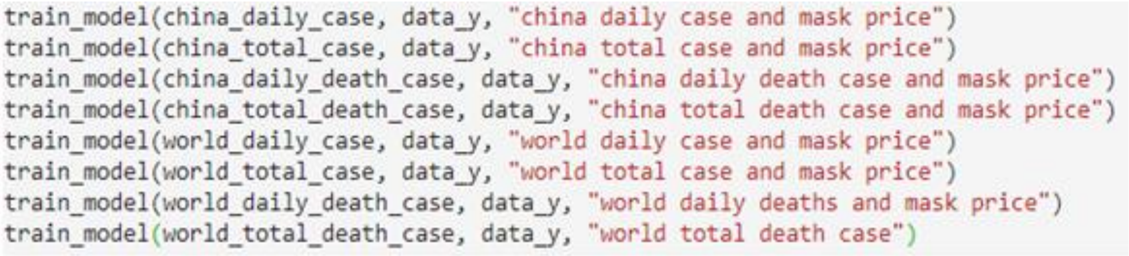
We used the machine learning library scikit-learn [4] to train and predict the model. The models we used in the project are linear regression [5] and polynomial regression. Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. It is helpful to interpret data on a modular level, especially when we want to quantify cases and prices. Polynomial regression provides the best approximation between variables and is compatible for many functions [6].



```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.pipeline import Pipeline
model = Pipeline([('poly', PolynomialFeatures(degree=6)),('linear', LinearRegression(fit_intercept=False))])
model = model.fit(train_x, train_y)
pred_y = model.predict(test_x)
print(pred_y)
```



```
def train_model(x_featrue, data_y, s):
    data_x = []
    for i in range(21,69):
        data_x.append([x_featrue[i]])

    train_x = data_x[:38]
    test_x = data_x[38:]
    train_y = data_y[:38]
    test_y = data_y[38:]

    model = Pipeline([('poly', PolynomialFeatures(degree=6)),('linear', LinearRegression(fit_intercept=False))])
    model = model.fit(train_x, train_y)
    pred_y = model.predict(test_x)
    print("pred_y is", pred_y)

    print(s)
    plt.scatter(test_x, test_y, color='black')
    plt.plot(test_x, pred_y, color='blue', linewidth=6)

    plt.xticks(())
    plt.yticks(())


    plt.show()
```



```
train_model(china_daily_case, data_y, "china daily case and mask price")
train_model(china_total_case, data_y, "china total case and mask price")
train_model(china_daily_death_case, data_y, "china daily death case and mask price")
train_model(china_total_death_case, data_y, "china total death case and mask price")
train_model(world_daily_case, data_y, "world daily case and mask price")
train_model(world_total_case, data_y, "world total case and mask price")
train_model(world_daily_death_case, data_y, "world daily deaths and mask price")
train_model(world_total_death_case, data_y, "world total death case")
```

4. EXPERIMENT RESULTS

4.1. Comparison of Different Features

We first draw the plot of features based on dates, from 2020-01-01 to 2020-04-11.

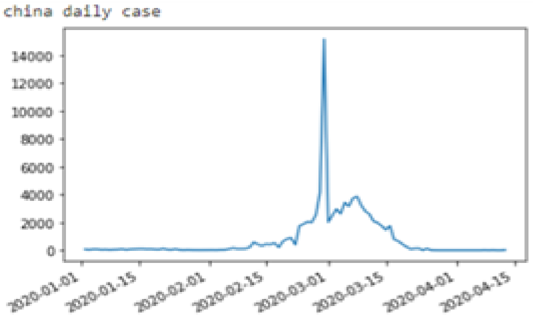


Figure 1. China daily case

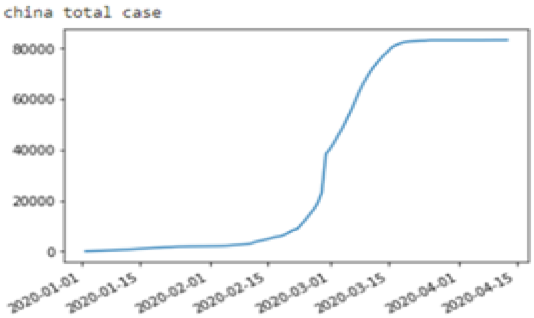


Figure 2. China total case

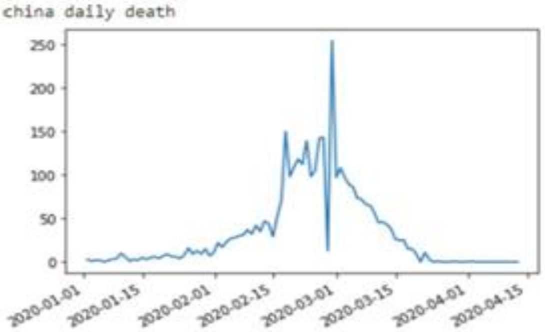


Figure 3. China daily death

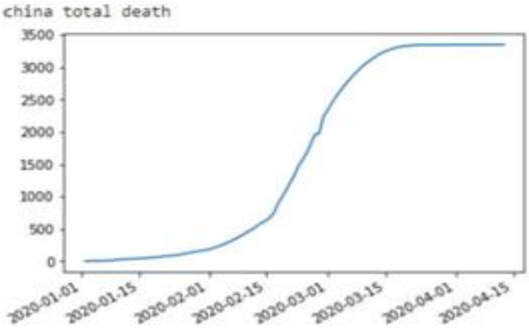


Figure 4. China total death

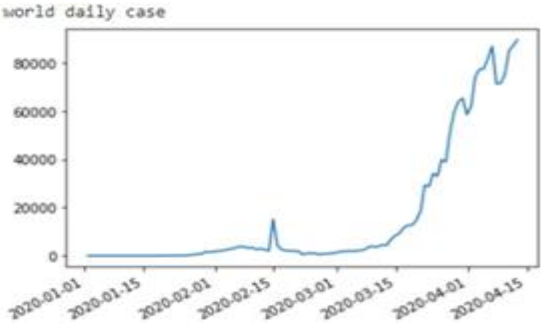


Figure 5. World daily case

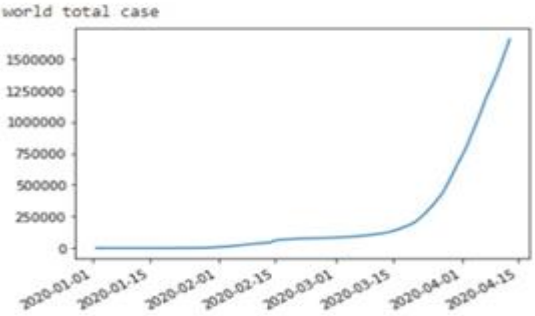


Figure 6. World total case

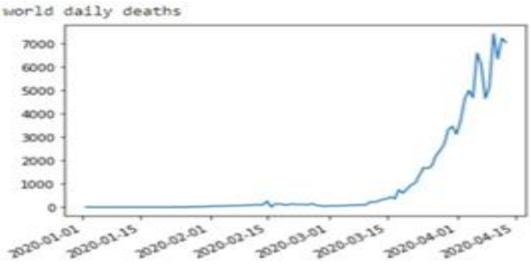


Figure 7. World daily deaths

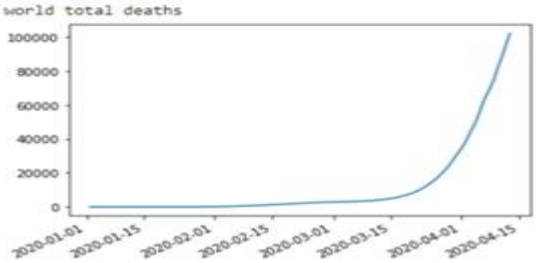


Figure 8. World total deaths

Figure 1, figure 3 show that China had an outbreak of daily confirmed cases and daily deaths in early March, when the figures peaked. China's total number of confirmed cases and deaths (Figure 2,4) and the world's total number of confirmed cases and deaths (Figure 6,8) both experienced a dramatic increase. The curve between the total number of confirmed cases and the number of deaths in China is s-shaped, which means that the epidemic has reached an inflection point, while the curve between the total number of confirmed cases and the number of deaths in the world is exponential[7]. This means that covid-19 is still highly contagious [8] in the world, and the number of patients is growing rapidly every day.

4.2. Comparison of Different Models of Training Features

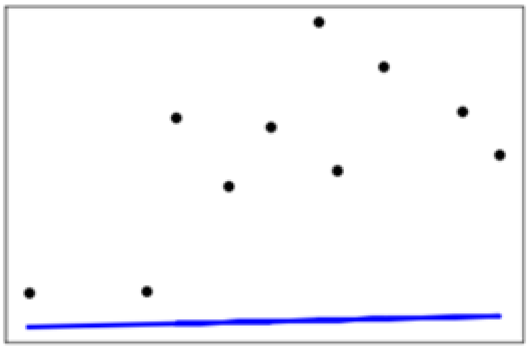


Figure 9.Linear Regression

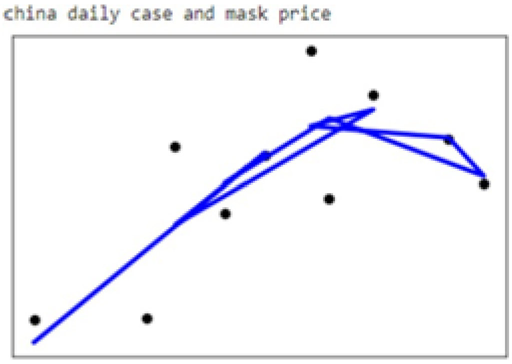


Figure 10. Polynomial regression

As can see from figures 9 and 10, linear regression is not ideal, but polynomial regression fits better.

4.3. Comparison of Prediction of all 8 Features

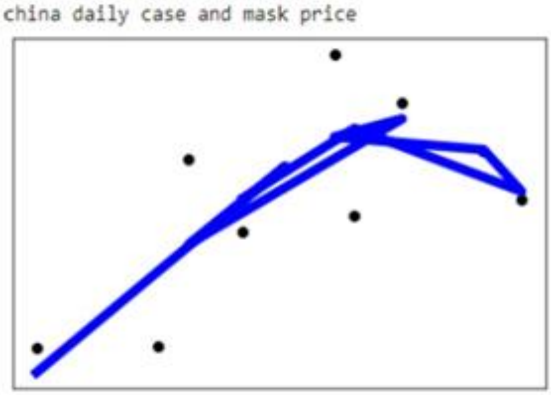


Figure 11. china daily case and mask price

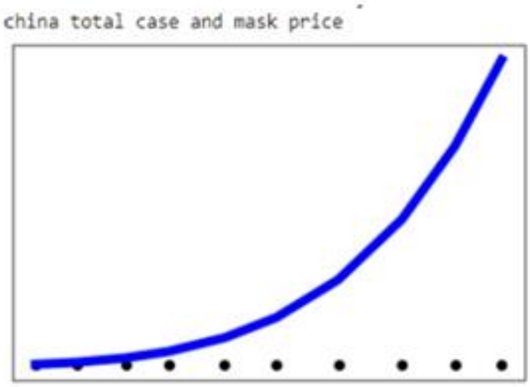


Figure 12. china total case and mask price

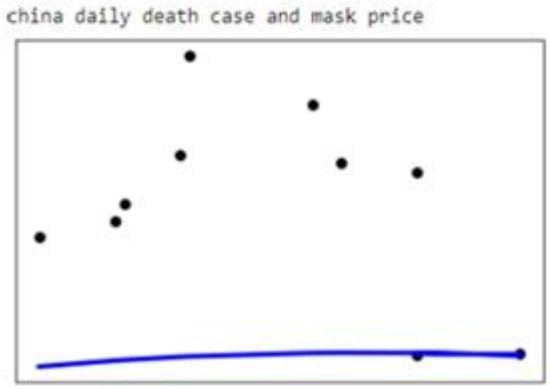


Figure 13. china daily deaths and mask price

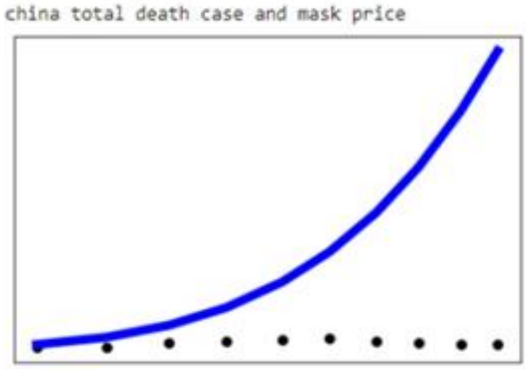


Figure 14. china total deaths and mask price

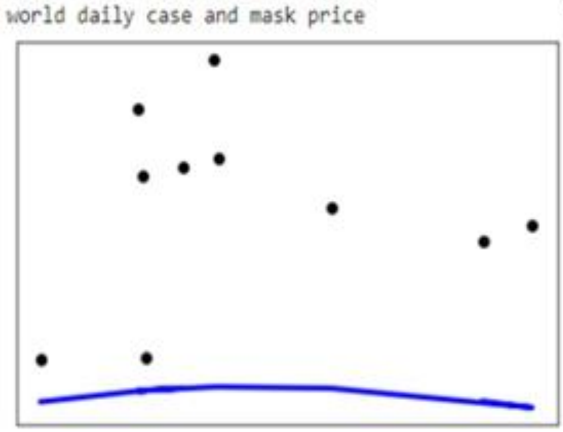


Figure 15. world daily case and mask price

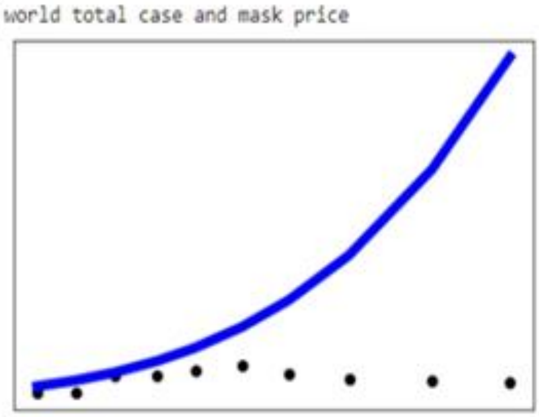


Figure 16. world total case and mask price

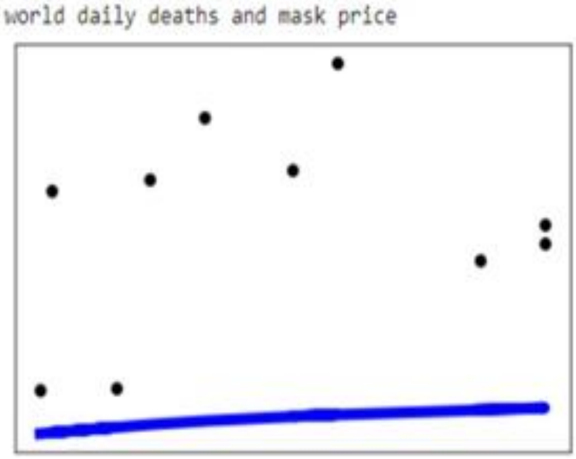


Figure 17. World daily death and mask price

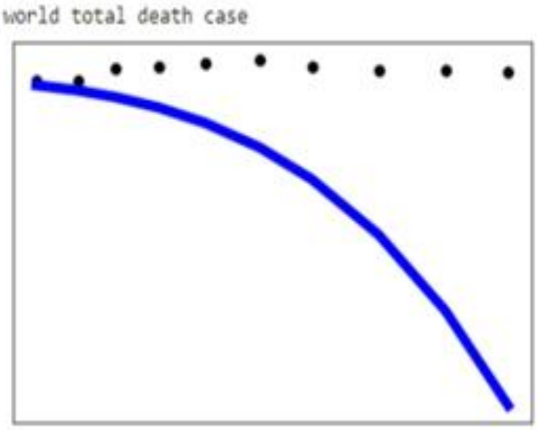


Figure 18. World total death and mask price
When Polynomial degree [9] was 6, Figure 1
was more accurate.

5. RELATED WORK

Liu, Y. et al [10] calculated reproduction number(R0) of the COVID-19 virus to find out that the ability of the virus to spread is higher than WHO expected, which can provide explanation to why the cases increased fast world-widely. Wu, Z. and McGoogan, M. J. [11] analyzed the emergent measures applied in Wuhan, China from January to March. They showed the whole timeline to explain the outbreak in China and the international impact. By comparison, we used machine learning to analyze and present the relationship between Chinese cases and World cases. Grasselli, G. et al [12] used linear and exponential models to estimate Italy's ICU demand [13]. In this project, we used linear and polynomial regression models to predict the world's mask prices. Fanelli, D. et al [14] analyzed the outbreak in China, Italy and France using a simple susceptible-infected-recovered-deaths model to indicate the relationships of situations in three countries. We focused on the infected and deaths number of China as it was the first country to break out, the first to block, and the first to reach the inflection point. Harrell FR Jr. et al [15] introduced the advantages of regression models in making accurate predictions than other methods.

6. CONCLUSIONS

We concluded from our predictions that the price of face masks [16] would fall over the next four days. There is a strong correlation between mask prices and China daily cases for some time to come. By training eight different features, we have eight different data, and know that the price of the most closely linked function is the exponential function.

The Algorithm can effectively predict the trend of mask price fluctuation, so we believe in the future using this method can effectively predict mask price and provide trade guidance for business and life demand.

REFERENCES

[1] Holmes, Kathryn V. "SARS-associated coronavirus." New England Journal of Medicine 348.20 (2003):1948-1951.

[2] Alpaydin, Ethem. Introduction to machine learning. MIT press,2020.

[3] Lee, Jae Won. "Stock price prediction using reinforcement learning." ISIE 2001. 2001 IEEE International Symposium on Industrial Electronics Proceedings (Cat. No. 01TH8570). Vol. 1. IEEE,2001.

[4] Seber, George AF, and Alan J. Lee. Linear regression analysis. Vol. 329. John Wiley & Sons,2012.

[5] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011):2825-2830.

[6] Pant, Ayush. "Introduction to Linear Regression and Polynomial Regression." Medium, Towards Data Science, 16 Jan. 2019, towardsdatascience.com/introduction-to-linear- regression-and-polynomial-regression-f8adc96f31cb.

[7] Maier, Benjamin F., and Dirk Brockmann. "Effective containment explains sub- exponential growth in confirmed cases of recent COVID-19 outbreak in Mainland China." arXiv preprint arXiv:2002.07572(2020).

[8] Dubé, C., et al. "Comparing network analysis measures to determine potential epidemic size of highly contagious exotic diseases in fragmented monthly networks of dairy cattle movements in Ontario, Canada." Transboundary and emerging diseases 55.9-10 (2008): 382- 392.

[9] Homer, Steven. "Minimal degrees for polynomial reducibilities." Journal of the ACM (JACM) 34.2 (1987):480-491.

[10] Liu, Ying, et al. "The reproductive number of COVID-19 is higher compared to SARS coronavirus." Journal of travel medicine(2020).

[11] Wu, Zunyou, and Jennifer M. McGoogan. "Characteristics of and important lessons from the

coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention." Jama(2020).

[12] Grasselli, Giacomo, Antonio Pesenti, and Maurizio Cecconi. "Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response." Jama(2020).

[13] Villari, Paolo, et al. "Unusual genetic heterogeneity of Acinetobacter baumannii isolates in a university hospital in Italy." American journal of infection control 27.3 (1999):247-253.

[14] Fanelli, Duccio, and Francesco Piazza. "Analysis and forecast of COVID-19 spreading in China, Italy and France." Chaos, Solitons & Fractals 134 (2020):109761.

[15] Harrell Jr, Frank E., et al. "Regression models for prognostic prediction: advantages, problems, and suggested solutions." Cancer treatment reports 69.10 (1985):1071-1077.

[16] Moore, Rachael E., et al. "Nasal mask." U.S. Patent Application No.29/166,190.

PHONE CLUSTERING METHODS FOR MULTILINGUAL LANGUAGE IDENTIFICATION

Ronny Mabokela

Technopreneurship Centre, School of Consumer Intelligence and Information Systems, Department of Applied Information Systems, University of Johannesburg, Johannesburg, South Africa

ABSTRACT

This paper proposes phoneme clustering methods for multilingual language identification (LID) on a mixed-language corpus. A one-pass multilingual automated speech recognition (ASR) system converts spoken utterances into occurrences of phone sequences. Hidden Markov models were employed to train multilingual acoustic models that handle multiple languages within an utterance. Two phoneme clustering methods were explored to derive the most appropriate phoneme similarities between the target languages. Ultimately a supervised machine learning technique was employed to learn the language transition of the phonotactic information and engage the support vector machine (SVM) models to classify phoneme occurrences. The system performance was evaluated on mixed-language speech corpus for two South African languages (Sepedi and English) using the phone error rate (PER) and LID classification accuracy separately. We show that multilingual ASR which fed directly to the LID system has a direct impact on LID accuracy. Our proposed system has achieved an acceptable phone recognition and classification accuracy in mixed-language speech and monolingual speech (i.e. either Sepedi or English). Data-driven, and knowledge-driven phoneme clustering methods improve ASR and LID for code-switched speech. The data-driven method obtained the PER of 5.1% and LID classification accuracy of 94.5% when the acoustic models are trained with 64 Gaussian mixtures per state.

Keywords

Code-switching, Phone clustering, Multilingual speech recognition, Mixed-language, Language identification

1. INTRODUCTION

Most multilingual societies are capable of code-switching in their daily conversations. This appears to be an acceptable modern-day style of communication, usually preferred in multilingual societies [1], [2]. Code-switching speech is commonly more spoken than formally written, and a large textual dataset is required to build a suitable language model which is necessary for developing a multilingual ASR system [3], [4]. However, the African reality in many communication episodes is that English is frequently mixed with indigenous under-resourced official languages.

Code-switching speech has a significant impact on existing ASR systems and a large speech corpus is required to develop suitable context-dependent acoustic models [3]. The existing monolingual ASR systems are not accurate enough to handle code-switched speech utterances. Consequently, acoustic, pronunciation and language models need to be redesigned to accommodate foreign or unknown words from different languages [5]. A multilingual ASR

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 285-298, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101421

system employs multilingual language models that allow the exploitation of multilingual pronunciation dictionaries. It is highly plausible to classify code-switched speech itself in the same category as under-resourced languages due to lack of speech technology resources for developing accurate ASR systems [2], [3]. ASR systems deployed in this environment should be able to process multilingual speech that includes such code-switching utterances. The LID system identifies language speech processing applications, such as telephone calls routing to human operators, particularly for handling emergency calls [3], [5]. In this paper, we propose an ASR system that is integrated with an LID system to classify code-switched speech for Sepedi and English. Only the experiments conducted using two official South African languages are reported on.

There are two ASR approaches that are reported to handle code-switched speech [1], [6], [7]. The first approach employs two monolingual ASR systems and an LID module. The LID module extracts the input code-switched utterances and then decides on the identity of each speech segment before passing them into their respective monolingual ASR systems. This approach is very simple because it applies acoustic and language modelling methods which achieve excellent monolingual performance. However, this approach is not preferred by many researchers due to LID error propagation which leads to poor ASR performance. The second approach employs a single-pass multilingual ASR system comprising a multilingual acoustic model of the languages concerned, a multilingual pronunciation dictionary which combines the words from targeted languages, and a multilingual language model that allows mixing of different language units. The approach needs a complete redesign of the acoustic and language models [6]. The major advantage of this approach is that it does not require the use of an LID system and it avoids the errors presented by the LID system.

In this research we propose a multilingual ASR system to perform LID on mixed-language corpora. We investigated whether the second approach can be adopted to achieve suitable multilingual acoustic modelling which can be used to handle Sepedi-English code-switching speech. We present the first study which relied solely on the mixed monolingual speech corpus of Sepedi and English but was evaluated using a code-switched speech corpus. Furthermore, we investigated which phoneme clustering method yields better ASR accuracy. We also examined how a multilingual acoustic model can impact LID classification accuracy in a mixed-language corpus. The novel approach proposed in this study is the first to offer a framework that integrates acoustic features and phonotactic information to achieve the LID system for mixed-language speech. This is a relevant study since it is common in South Africa for more than one language to be spoken in the same region.

2. RELATED WORK

In Singapore, Mandarin and English are often mixed in spoken conversations [1], in Hong Kong code-switching between Cantonese and English takes place on many occasions [8] and in Taiwan, Mandarin-Taiwanese code-switching speech has been reported [9]. Mixed-language speech has also been found to occur in India between Hindi and English [10]. Code-switching is also observed in South Africa, and two South African indigenous languages, Xhosa and Zulu, were studied for LID and multilingual speech recognition. Recently, Modipa et al. [11] reported a context-dependent modelling technique of English vowels in Sepedi code-switched speech where the process of obtaining phone mapping from embedded language to the matrix language was investigated.

There are few reported approaches in code-switched speech. One approach is to integrate multiple cues such as acoustics, prosodics and phonetics to distinguish between languages in a code-switched speech utterance [8]. A language boundary detection (LBD) method is applied to detect multiple languages within an utterance [9]. The second approach, such as the delta-Bayesian information criterion (Delta-BIC) and latent semantics analysis (LSA), has been used to separate English, Mandarin and Taiwanese in code-switched utterances [9]. Lastly, an approach that uses maximum a posteriori-based estimation was used to jointly segment and identify utterances of a mixed language [13]. The above approaches use an LID module that incorporates an LBD module. The LID systems that incorporate an LBD module are usually not preferred due to incorrect assumptions that code-switched speech segments are independent of each other and as a result, errors in the LID module cannot be recovered [1]. Therefore, if the LBD module cannot achieve 100%, it will directly influence the performance of the LID module, thereby limiting the performance of the speech recognition module [1], [10].

On the other hand, a multilingual ASR approach can handle code-switched speech. It comprises a multilingual acoustic model, a multilingual pronunciation dictionary and a multilingual language model that allows the mixing or sharing of models across different language units [1], [10]. A multilingual ASR approach does not need an additional LID module to identify speech segments since language information is incorporated directly into the system [1]. One technique is to use a linguistic knowledge-based method to establish a multilingual phone set mapping or clustering of similar phonetic features that share the training data [7]. Common examples are the International Phonetic Alphabet (IPA), Speech Assessment Methods Phonetic Alphabet (SAMPA) and Wordbet [15]. Another technique is to map language-dependent phones using a data-driven approach such as clustering specific phones according to distance measured between similar acoustic models. Examples of data-driven methods are the Confusion Matrix, Bhattacharyya Distances and Kullback-Leibler Divergent which takes spectral characteristics into consideration [15].

Lyu et al. [16] propose a word-based lexical model LID system which uses the lexicon information to distinguish between code-switching speech within an utterance. A two-stage scheme system is used with a large vocabulary continuous speech recognition (LVCSR) system. Then a trained word-based lexical model is applied to identify languages via recognised word sequences. The approaches such as Parallel Phone Recognition and Language Modelling (P-PRLM) [4, 18] and parallel phoneme recognition vector space modelling (PPR-VSM) [17] are some of the most popular approaches to the LID system. The P-PRLM approach employs multiple phoneme recognisers that tokenise the speech waveform into sequences of phonemes. The resulting sequence of phonemes is then passed to the n-gram language model which determines the most probable language from the target languages [6], [18]. The supervised support vector machine (SVM) model has proven to be the best classifier [18]. A similar approach was used to distinguish between 11 officially spoken languages of South Africa [18]. It was implemented using P-PRLM architecture and techniques such as phoneme frequency filtering - where an SVM-based classifier is used to classify languages at the back end. The SVM classifier was able to achieve an average LID rate of 71.78% on test samples of 3-10 seconds long and an LID rate of 82.39% when clustering similar language families [18]. The diagram below shows the P-PRLM system employed for LID in South African languages. The Figure 1 below shows the PPRLM system employed for LID in South African languages. This work is similar to this current research work, but was employed for single-language identification.

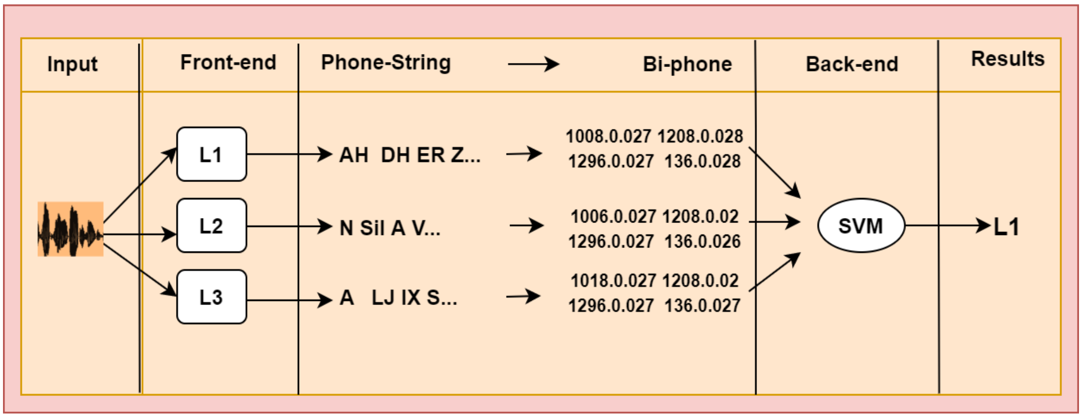


Figure 1. LID system based on P-PRLM scheme (adopted from[18]).

3. Mixed-Language Corpus

The speech corpus in this study contained two monolingual speech sets of data for Sepedi and English. The third speech corpus was Sepedi-English code-switched speech data which was recorded during the speech data collection phase. The speech corpora were divided into training and testing datasets. The amount of mixed-language speech data that was used for the training and testing of the system is summarised in the sections below.

3.1. Training Dataset

The corpus used for training the acoustic model included recorded speech data and the respective transcriptions of locally produced Sepedi developed within the Telkom Centre of Excellence for Speech Technology (TCoE4ST) and freely available LWAZI South African English speech data. The TCoE4ST locally produced Sepedi speech corpus had 3 465 utterances. From the LWAZI English speech corpus, 1 680 recorded speech data and the respective sentences from utterances were selected. The training dataset contained a total of 5 505 speech utterances and was 5.5 hours long.

3.2. Testing Dataset

The speech corpus used for testing the phone recognition contained 660 speech utterances which were not part of the training speech data. Code-switched speech is generally spoken but not formally written. However, it is not easy to find code-switched speech data. It is for this reason that simple finite loop grammar was used to generate 60 artificially code-switched sentences that were syntactically correct. These texts were recorded by 10 speakers to produce 660 utterances, 300 of which were used as a testing set and the remaining 300 were included in the training set. The quality of the utterances was manually improved by removing dysfluencies such as long pauses, laughs and hiccups. The average ratio of code-switched English words within each utterance was not more than 0.5. The testing utterances were 1 hour long.

4. Phoneme Mapping Methods

In this research project, two different phone mapping strategies are proposed to determine the similarity between the target languages. The first mapping strategy is based on an IPA-based scheme which requires linguistic experts, and the other strategy is a data-driven method derived by measuring a confusion matrix.

4.1. Linguistic-Knowledge Phoneme Mapping

The methods which are used to deal with similar phoneme inventories have been studied [1], [8], [9]. A single-pass speech recogniser on two languages with a multilingual acoustic modelling technique for the available speech corpus was proposed in this study. The multilingual acoustic model was developed by mapping the English phonemes to the Sepedi phonemes. This approach was motivated by the occurrence of similar phonemes from the target languages and also aimed to reduce a larger number of phonemes. The criteria that were used to construct the linguistic-knowledge strategy are described below.

Phoneme Mapping Criteria

C-1. If the IPA classification is like a Sepedi phoneme, then the English phoneme is mapped directly to the Sepedi phoneme.
C-2. Each English phoneme is mapped to its closest matching Sepedi phoneme.
C-3. If no closely matching phoneme is found, then the English phoneme that occurs most frequently in the phoneme inventory is extended to the phoneme set.
C-4. If none of the above criteria are applicable, then each phoneme is mapped to the Sepedi phoneme that it is mostly confused with, according to a confusion matrix.

The above criteria resulted in the phoneme mapping list indicated in Table 1 below. Phonemes such as ***/au/*** and ***/e@/***from the LWAZI dictionary were decoupled to a single phoneme ***/a/, /u/***and ***/e/, /@/*** respectively.

Table 1. Examples of the phoneme list achieved with linguistically motivated method.

Phoneme mapping list							
from	**to**	**from**	**to**	**from**	**to**	**from**	**to**
{	**E**	Oi	**O i**	i:	**i**	g	**k_>**
3:	**E**	p	**p_h**	i@	**i @**	@i	**@ i**
a:	**a**	Q	**O**	k	**k_h**	u:	**u**
au	**a i**	r\	**r**	O:	**O**	u@	**u**
ai	**a u**	t	**t_h**	@:	**a**	Z	**d_0Z**
d	**l’**	T	**f**	U	**u**	h_b	**h**
D	**l’**	tS	**tS_h**	v	**B**	**Additions**	
e@	**E @**	@u	**O**	z	**s**	@	**b**

In our case, the diphthongs of the English language were separated into vowels using the traditional IPA-based strategy in the mixed phoneme set. Each phonemic vowel was mapped toits equivalent Sepedi phoneme directly using C-2 criteria.

4.2. Data-driven Phoneme Mapping

The data-driven mapping which is based on the confusion matrix was built by including all the Sepedi and English phonemes [10, 11]. The confusion matrix was generated when the acoustic models of the source language were applied to the speech utterances of the target language. The recognised phoneme sequences of the source phoneme candidates were then mapped to phoneme sequences of the target phoneme candidates as indicated for Table 2. This mapping method consists of the counts of the confusion pairs when aligning the speech recognition output and transcriptions of the speech data. The advantage of this approach is that it is fully data-driven and does not require a linguistic expert, which can be time-consuming. For each phoneme of the

English language, the most often confused phoneme with the Sepedi language was selected for mapping. For each phoneme P_{L1} from the target language $\boldsymbol{P_{L2}}$, the best respective source candidate phoneme was matched. The similarities were measured by selecting the number of phoneme confusions as **C** ($\boldsymbol{P_{L1}}$, $\boldsymbol{P_{L2}}$). The target phoneme was matched as follows:

$$P_{Ln} = MaxC(P_{L1}, P_{L2}) \tag{1}$$

where ***L1*** denotes the target language (Sepedi) and ***L2*** the source language (English). Thus, for each target phoneme, a source candidate phoneme with the highest number of confusions was determined. If the same number of confusions occurred on two or more source candidate phonemes, the decision on the choice of the target phoneme was made by a knowledge expert. The same strategy was followed even when there were no confusions found between target and source candidate phonemes.

Table 2. Examples of the phoneme list achieved with data-driven method.

Phoneme mapping list			
From	**to**	**from**	**to**
{	a	Oi	**E**
3:	E	p	**p_>**
a:	**a**	Q	**O**
au	**i**	r\	**r**
ai	**u**	t	**t_h**
d	**l’**	T	**F**
D	**l’**	tS	**tS_h**
e@	**E**	@u	**O**
G	**G**	u:	**U**
@i	**E**	u@	**O**
i:	**E**	U	**U**
i@	**a**	v	**B**
k	**k_>**	z	**S**
O:	**O**	Z	**d_0Z**
@	**a**	b	**B**
h_b	**h**	@:	**a**

5. Proposed Multilingual ASR-LID System

The multilingual ASR-LID system is targeted to identify only two languages. A multilingual recognition system takes speech waveform and outputs the corresponding phone sequences. This is done when an ASR system estimates the likelihood score of the optimal phone sequences given the acoustic features extracted from the speech utterance waveform. To achieve this, a multilingual acoustic and language model was employed to estimate the likelihood scores for the spoken utterance. The phoneme mapping technique was applied to generate the shared phoneme set for robust multilingual acoustic modelling. Figure 2 shows the proposed multilingual ASR system.

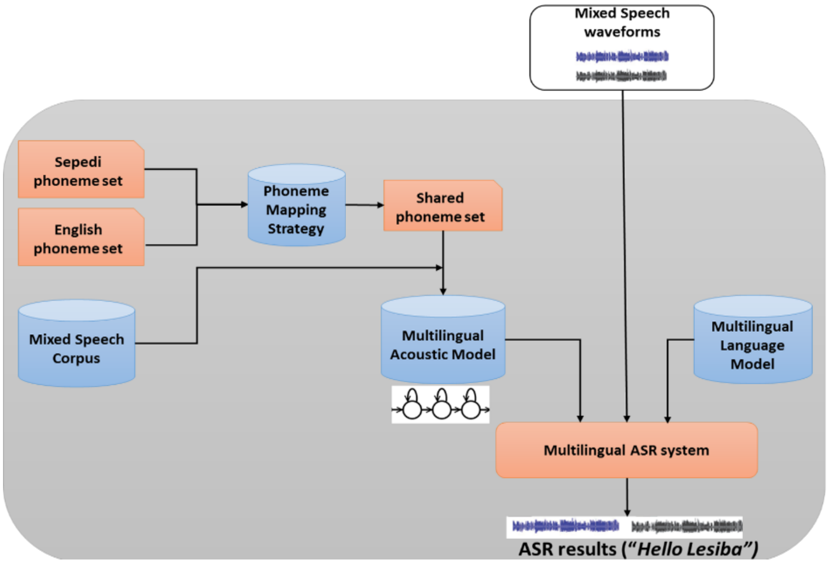


Figure 2. Multilingual automatic speech recognition system.

Multilingual acoustic models are used to perform HMM-based parameter re-estimation. For recognition purposes, the multilingual acoustic features were compared with the HMM-based multilingual acoustic models as well as the language model. The sequences of phone strings were decoded by the Viterbi decoding algorithm, which searches the optimal sequence of the phones using the combined likelihood scores from the multilingual acoustic model and language model.

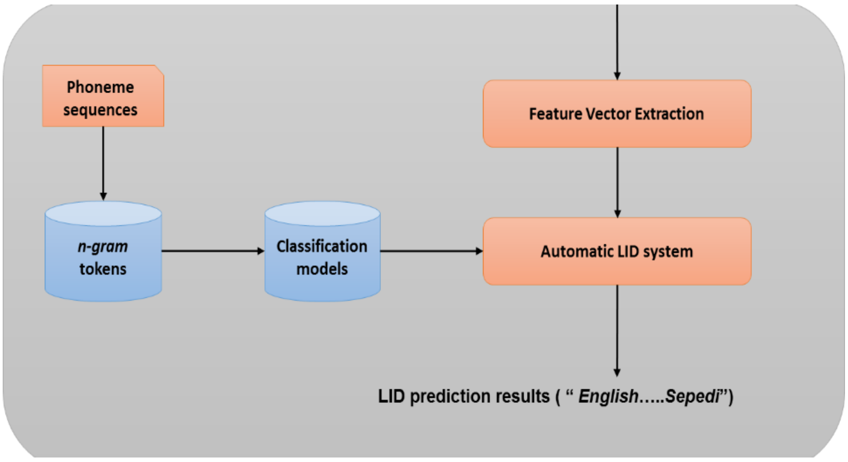

Phoneme sequences
n-gram tokens
Classification models
Feature Vector Extraction
Automatic LID system
LID prediction results (" *English.....Sepedi*")

Figure 3. The language identification system for mixed-language speech.

For each phone sequence generated from the ASR system, the bi-phone occurrences were extracted from the phone sequences and converted into a suitable SVM format with a unique numerical representation. This approach is like vector space modelling [5]. As a final step (see Figure 3), the SVM-based classifier was used to identify only two class feature samples; languages outside the targeted range were not classified. For each phoneme sequence generated from the ASR system, the phoneme occurrences were extracted from the phoneme sequences and converted into a suitable SVM format with a unique numerical representation. The classification model with the highest log-likelihood score was chosen to be the most likely sample for classification. The bi-phone vectors were then used as an input to the SVM-based classifier to build the classification model. The phoneme feature vectors have the following numerical attributes: a label is the class label in a numerical representation, a feature index represents ordered feature indexes - that is, the location of that particular phoneme feature, usually integer representation, and in our case, a feature value represents the frequency count or occurrences of each phoneme feature attribute. The SVM classification model was used to separate vectors in a binary classification and hypothesise the maximum likelihood score of the bi-phone frequencies of each language [18].

6. MODEL DEVELOPMENT AND SYSTEM SETUP

This section describes the experimental setup, the tools used to develop the ASR systems, LID system and the configurations of each system setup. All the multilingual ASR systems were developed, and experiments were performed using the hidden Markov Model Toolkit (HTK) [12]. The experimental results obtained from these ASR and LID systems are later analysed and discussed.

6.1. Baseline Acoustic Models

To build multilingual acoustic models, we applied a Hamming window of 25ms length with an overlapping window frame length of 10ms. Acoustic features were obtained using 39-dimensional static Mel-frequency Cepstral Coefficients (MFCCs) with 13 deltas and 13 acceleration coefficients. The Cepstral Mean and Variance Normalization (CMVN) pre-processing and semi-tied transformations were applied to the hidden Markov Models (HMM). The CMVN was used to overcome the undesired variations across the channels and distortion. Figure 4 shows the configuration file that was used for extracting speech features. The HMM-based ASR system was created with a widely used standard HTK [12]. The acoustic model used a three-state left-to-right HMM. The HMM-based system consisted of the tied-state triphones clustered by a decision tree technique. Each HMM state distribution was modelled by eight Gaussian mixture models (GMM) with a diagonal covariance matrix. Furthermore, the optimal phone insertion penalties and language scaling factors were properly tuned to balance the number of inserted and deleted phones during speech decoding.

```
CEPLIFTER     =   22
ENORMALISE  =   FALSE
NUMCEPS       =   12
NUMCHANS     =   26
PREEMCOEF   =   0.97
SAVECOMPRESSED =   TRUE
SAVEWITHCRC     =   FALSE
SOURCEFORMAT   =   WAVE
TARGETKIND     =    MFCC_0_D_A_Z
TARGETRATE     =    100000.0
USEHAMMING     =    TRUE
WINDOWSIZE     =    250000.0
ZMEANSOURCE    =    TRUE
LOFREQ          =      150
HIFREQ          =      4000
```

Figure 4. The configurations used for mfcc feature extraction

6.2. Language Modelling

To build the multilingual language model, many data texts had to be collected and normalised. The word coverage of the multilingual speech was improved by applying language-aware context-based text normalisation. Thus, for example, digits, temperature, time, currency amount, percentage, etc., were converted to appropriate words. A phone language model was incorporated in the speech recogniser for the purpose of robust speech decoding. The training transcriptions, together with the generated code-switched texts, were formatted into phone transcriptions and were used to develop the multilingual language model. The combined vocabulary that was used to train the language model consisted of over 85 000 unique word tokens for both Sepedi and English. The language model that was used for multilingual ASR experiments was implemented using the Stanford Research Institute language model toolkit [13]. It was trained independently with discount interpolation. The interpolation weights were optimised using the training set perplexity as a performance measure.

6.3. Multilingual Dictionary and Phoneme Set

The experimental bilingual pronunciation dictionary used was achieved by merging several monolingual pronunciation word lexicons without retaining duplicate words. For the primary Sepedi language, we used a limited vocabulary of a freely available Sepedi pronunciation dictionary that was locally produced within the TCoE4ST and LWAZI. For the English language, we used a freely available LWAZI English pronunciation dictionary often used for speech technology research tasks. All the words in the pronunciation dictionary were manually verified and checked for redundant phone representation. There were 7 176 Sepedi and 78 722 English words in the bilingual dictionaries. The dictionaries were further redesigned and rectified where necessary.

The combined bilingual dictionary contained 85 898 unique words. The representation used in the bilingual pronunciation dictionary followed the SAMPA notations based on IPA rules and also taking into consideration the pronunciation rules [8], [14]. Some 67 phones were combined into a mixed phone set, attained by combining Sepedi and English phones directly without silent phonemes. In this case, phones with similar phonetic features were mapped into one best phone candidate representation to lower confusion within the combined phone set. Some of the English vowel phones were left unmapped since they did not match any Sepedi vowel phones.

The combined mixed phone set included all phonemes of Sepedi and English that were used during the training phase without performing phoneme mapping. We used a knowledge-based IPA method to construct linguistically motivated phonetic pairwise mappings. The IPA-based phoneme set, and data-driven phoneme set contained 38 phonemes, excluding the silent phonemes. In this case, to train our multilingual acoustic model that effectively handled Sepedi-English code-switched speech, we adopted the technique used by Biswas et al. [6], Shan et al. [7] and Bhuvanagiri and Kopparapu [8]. Lastly, problematic words of Sepedi or English origin were manually reviewed for correct pronunciation prior to training the HMMs.

6.4. Language Identification Classifier

The SVM-based classifier based on bi-phone frequencies as an output was used to classify only two class feature samples; languages outside the targeted range were not classified. For each phone sequence generated from the phone recognition, each bi-phone occurrence was extracted from the phone sequences and converted into a suitable SVM format with a unique numerical representation called a bi-phone feature vector. Therefore, the phoneme features were calculated for every utterance. The bi-phone frequency vectors were then used as input to the back-end SVM classifier. The SVM classification model was used to separate vectors in a binary classification and hypothesise the maximum likelihood score of the bi-phone frequencies of each language. The bigram phones were trained to create the classification model.

The SVM-based classifier was implemented using LIBSVM library [15]. The SVM training dataset size was 12 147 KB of phone tokens. The training dataset that was used for training the SVM-based classifier was extracted from the phone-based transcriptions. The phone sequences were used to train the SVM classifier, which resulted in support vectors from models. The training process was also aimed at maximising the margin as well as minimising the training errors. The bi-phone vector attributes for both testing and training were scaled in the range of [0, 1]. The benefit of scaling datasets is to speed up the training and classification process in order to obtain the best model performance and to avoid numerical differences that could lead to over-fitting if the training data attributes are in a large range [16]. A grid search was used to estimate the SVM parameters such as ***C***, ***gamma***, **margin error**, **trade-off parameter** and **kernel** width before training the classifier [5], [17]. The Radial Basis Function (RBF) kernel was used for training the classifier. We obtained the optimal parameter for this kernel and applied five-fold cross-validation to the training set and estimated each grid point for the accuracy of the classifier.

7. Results and Discussion

For experimentation, three ASR and LID systems were developed. The baseline ASR system was achieved by directly combining monolingual ASR systems for Sepedi and English into a multilingual ASR system. The baseline (i.e. directly mixed) ASR-LID system was evaluated and compared with the multilingual ASR-LID systems that were developed using the two phoneme mapping techniques. No specific phoneme mapping was performed in the phoneme set. The phoneme set size was large with 67 phonemes. The HMM-based acoustic models trained on these systems contained eight Gaussian mixtures per state.

In this experiment, we modified a mixed recognition system by applying two different phoneme mapping techniques. We trained the front-end acoustic models on both Sepedi and English speech data and performed modelling of code-switching at the pronunciation dictionary level. Linguistically motivated and data-driven phoneme mapping methods were applied to determine the similarity between the phonemes of our target languages. We first applied a linguistically motivated phoneme mapping method using an IPA-based scheme. Our experiments were performed on the same speech data which was used for developing the directly mixed LID system. As a result of phoneme mapping, the number of phonemes in the directly combined phoneme set were reduced. The results obtained from the three systems using mixed-language speech are shown in Table 3.

The experimental results presented in Table 3 show that phone error rate (PER) and LID classification accuracy improved when the phoneme clustering methods were applied. The quality of the correct phoneme recognition output was typically captured by the PER metric formulated as:

$$PER = \frac{(S+I+D)}{N} \tag{2}$$

where (N) is the total number of labels, (D) is the number of phone deletion errors, (S) is the number of phone substitution errors and (I) is the number of phone insertion errors.

In Table 3, the SVM classifier yielded a promising and acceptable LID accuracy rate of 95.2% on directly mixed speech utterances with a total of 3 201 support vectors. In this case, mixed speech utterances included both monolingual and code-switched utterances. The experimental results of the SVM-based LID classifier were also obtained using RBF kernel. The SVM-based classifier was trained using a five-fold cross-validation and RBF kernel which yielded an estimation rate of 99.75% on the trained classification models. Both phoneme mapping approaches achieved a significant improvement over the baseline system results. The data-driven approach outperformed the baseline directly mixed system and the IPA-based system. The IPA-based approach performed better with a PER of 4.5% but LID classification accuracy was about 9% lower. The data-driven approach performed better with a PER of 14.5% as well as a LID classification accuracy of 2.3%. These methods allow sharing of the parameters in the HMM-based acoustic models of the target languages.

The IPA-based LID system was able to achieve a better PER reduction of 4.5%, outperforming the directly combined mixed LID system. The best performance of the PER reduction of 19.2% was achieved by the data-driven LID system. The data-driven LID system achieved a PER difference of 9.4% compared to the IPA-based LID system. We also observed that the ASR system with a larger number of phonemes in the phoneme set performed badly compared to when a phoneme mapping method was engaged to reduce the phoneme set. The amount of quality training speech data can also improve the multilingual ASR system performance significantly.

Table 3. Experimental results showing the PER of the multilingual ASR and LID classification accuracy with 8, 16, 32 and 64 Gaussian mixtures per state.

Number of Gaussian Mixtures	**ASR-LID Systems**	**PER(%)**	**LID Accuracy (%)**
8	Directly mixed	33.2	95.2
	IPA-based	**28.7**	**85.8**
	Data-driven	**19.2**	**87.3**
16	Directly mixed	21.4	83.8
	IPA-based	**17.8**	**89.7**
	Data-driven	**15.6**	**89.5**
32	Directly mixed	12.9	83.8
	IPA-based	**12.9**	**84.7**
	Data-driven	**7.3**	**96.7**
64	Directly mixed	5.4	83.9
	IPA-based	**7.3**	**83.7**
	Data-driven	5.1	**94.5**

Table 3 shows the PER and LID accuracy or rate that was attained on the four LID systems when the HMM-based acoustic models contained 16 Gaussian mixtures per state. The use of context-dependent HMM-based acoustic models with increased Gaussian mixtures per state was adopted during training of the acoustic models as they tend to improve the performance of the phoneme recognition systems. The PER has a direct proportionate relationship to Gaussian mixtures per state. The triphone models were then improved by gradually increasing the number of Gaussian mixtures and performing four iterations of embedded re-estimation after each increase. This procedure was continuously repeated until the acoustic models had 32 Gaussian mixtures per state, after which the phoneme recognition results no longer improved significantly on the test set. The performance of the PER and LID classification accuracy was significantly better on both IPA-based and data-driven LID systems. Both systems were able to outperform the directly mixed LID system only in the LID classification accuracy. A slight difference of 0.12% in LID classification accuracy was observed between the IPA-based and data-driven LID systems. This was due to a larger number of the phoneme occurrences that were observed within the testing set.

Table 3 shows the PER and the LID classification accuracy attained on the four LID systems when the HMM-based acoustic models contained 32 Gaussian mixtures per state. The PER reduction was better, but the LID accuracy was a bit lower than expected for the data-driven LID system. We observed that a better performance of the PER does not necessarily result in a positive bias of the LID accuracy on both directly mixed and IPA-based LID systems, since the phoneme recognition systems are used only to obtain phoneme strings for SVM-based classifier training.

A slight reduction of the PER was achieved with eight Gaussian mixtures, but a better PER reduction was achieved with 32 Gaussian mixtures. This clearly shows that a data-driven LID system achieves a better PER with all Gaussian mixtures per HMM state represented in Table 3. The directly mixed LID and IPA-based LID system nearly achieved the same PER with 32Gaussian mixtures per HMM-based acoustic model state. However, a slight improvement of 0.3% was achieved. The phoneme recognition results in the experiment show that the application of phoneme mapping methods to our targeted languages and the increase of Gaussian mixture per shared HMM-based acoustic model significantly improve the performance of the phoneme recognition and LID system.

We also observed that by applying IPA-based and data-driven phoneme mapping techniques, these could yield extreme results such as increased sentence and phone correctness, phone recognition and LID accuracy of the proposed mixed-language integrated LID system, as well as reduced PER when 32 Gaussian mixtures per HMM state are considered. However, both the directly mixed ASR-LID system and data-driven system show a significant reduction of PER with 64 Gaussian mixtures but no further increase in the LID classification accuracy as indicated in Table 3.

8. CONCLUSIONS

This paper presents an integration of multilingual speech recognition into LID for code-switched speech using phonotactic features as language information. In this research work, two strategies are reported on to perform similar phoneme mapping of the target official languages. We have shown that existing monolingual corpora can handle code-switching utterances. The IPA-based approach is derived from linguistic knowledge, whereas the data-driven approach is based on the confusion matrix. Appropriate phoneme mapping approaches across the target languages offered robust context-dependent multilingual acoustic models which tended to produce acceptable ASR-LID system performance. Our proposed IPA-based and data-driven approaches have shown a significant improvement in both PER and LID classification accuracy. The data-driven method outperforms the IPA-based approach. An acceptable PER was achieved with the data-driven approach when multilingual acoustic models were employed that were trained with 32 Gaussian mixtures per state. Again, the 64 Gaussian mixtures do improve the PER, but has no impact on the performance of LID accuracy. In future, we hope to train the multilingual ASR-LID system with more robust context-dependent code-switched acoustic models for further evaluation and performance analysis. We also aim to collect more code-switched speech for ASR and LID research in future. As part of extending this research work, we aim to investigate more South African languages where speakers use code-switching in their daily conversations.

ACKNOWLEDGEMENTS

We acknowledge the team from Council for Scientific and Industrial Research (CSIR) that developed LWAZI speech data and the University of Limpopo for allowing us to use their Sepedi speech corpus. We also acknowledge Telkom SA for funding this project.

REFERENCES

[1] J. Weiner, N. T. Vu, D. Telaar, F. Metze, T. Schultz, D. C. Lyu, E. Chng, and H. Li, "Integration of language identification into a recognition system for spoken conversations containing code-switches," in Proc. of SLTU, 2012, pp. 1–4.

[2] K. R. Mabokela, "A multilingual ASR of Sepedi-English code-switched speech for automatic language identification," in International Multidisciplinary Information Technology and Engineering Conference (IMITEC), November 2019, pp. 430–437.

[3] K. Mabokela, M. Manamela, and M. Manaileng, "Modeling codeswitching speech on under-resourced languages for language identification," in SLTU 2014, May, pp. 225–230.

[4]` L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," Speech Communication Journal, vol. 56, pp. 85–100, 2014.

[5] H. Li, B. Ma, and K. Lee, "Spoken language recognition: from fundamentals to practice," in Proceedings of the IEEE, May 2014, pp. 1136– 1159.

[6] A. Biswas, F. de Wet, E. van der Westhuizen, E. Ylmaz, and T. Niesler, "Multilingual neural network acoustic modelling for ASR of under-resourced English-isiZulu code-switched speech," in Proc. Interspeech 2018, pp. 2603–2607.

[7] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Investigating end-to-end speech recognition for Mandarin-English codeswitching," in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6056– 6060.

[8] K. Bhuvanagiri and S. K. Kopparapu, "Mixed language speech recognition without explicit identification," In American Journal of Signal Processing, 2012, pp. 92–97.

[9] F. Diehl, "Multilingual and cross-lingual acoustic modelling for automatic speech recognition," PhD dissertation, Universitat Politecnica de Catalunya, Newark, 2007. [Online]. Available: http://mi.eng.cam.ac.uk/ fd257/publications/

[10] T. Modipa, M. Davel, and F. De Wet, "Pronunciation modelling of foreign words for Sepedi ASR," in Proceedings of Pattern Recognition Association of South Africa, 2010, p. 185-189.

[11] T. Modipa and M. H. Davel, "Implications of Sepedi/English code switching for ASR systems," in 24th Annual Symposium of the Pattern Recognition Association of South Africa, 2013, pp. 64–69.

[12] S. J. Young, A. D. K. G. Evermann, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, in Cambridge University (For HTK Version 3.2.1), 2013. [Online]. Available: http://htk.eng.cam.ac.uk

[13] A. Stolcke, "Srilm - an extensible language modelling toolkit," in Proc. ICSLP, 2002, pp. 901–904.

[14] U. T. W. Zhirong, T. Schultz, and A. Waibel, "Towards universal speech recognition," In Proc. ICMI 2002, 2002.

[15] C. C and C. -J. Lin, "Libsvm - a library for support vector machine," 2001. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[16] O.Giwa and M. Davel, "Language identification of individual words with joint sequence models," in Proceedings of Interspeech 2014, 2014, pp. 1400 –1404.

[17] M. Peche, M. Davel, and E. Barnard, "Development of a spoken language identification system for South African languages," in SAIEE Africa Research Journal, vol. 100(4), Dec 2009.

[18] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: atutorial", In IEEE Circuits and Systems Magazine, 11(2), pp.82-108, 2011.

AUTHORS

Mr. Ronny Mabokela holds an MSc. degree in Computer Science from the University of Limpopo. Mr Mabokela has vast industry experience, having worked for Telkom and Vodacom South Africa. He was one of the remarkable team that established the formation of high-speed fibre-based internet. He contributed to the successful development of Vodacom internal systems, API integrations and business process automation. He received numerous awards, including being an exceptional performer in both Telkom & Vodacom and seconds prize award of research excellence at University of Limpopo (2013). He became a session chair and peer reviewer of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC) and International Conference on Computing and Communications Technologies (ICCT) in 2014 and 2015, respectively. He has presented his work on numerous platforms including International Workshop on Spoken Language Technologies (SLTU) in Russia. He has a keen research interest to broadband network services, computational linguistics, natural language processing, speech technologies and machine learning.

RESOLVING CODE SMELLS IN SOFTWARE PRODUCT LINE USING REFACTORING AND REVERSE ENGINEERING

Sami Ouali

College of Applied Sciences, Ibri, Oman

ABSTRACT

Software Product Lines (SPL) are recognized as a successful approach to reuse in software development. Its purpose is to reduce production costs. This approach allows products to be different with respect of particular characteristics and constraints in order to cover different markets. Software Product Line engineering is the production process in product lines. It exploits the commonalities between software products, but also to preserve the ability to vary the functionality between these products. Sometimes, an inappropriate implementation of SPL during this process can conduct to code smells or code anomalies. Code smells are considered as problems in source code which can have an impact on the quality of the derived products of an SPL. The same problem can be present in many derived products from an SPL due to reuse. A possible solution to this problem can be the refactoring which can improve the internal structure of source code without altering external behavior. This paper proposes an approach for building SPL from source code. Its purpose is to reduce code smells in the obtained SPL using refactoring source code. Another part of the approach consists on obtained SPL's design based on reverse engineering.

KEYWORDS

Software Product Line, Code smells, Refactoring, Reverse Engineering.

1. INTRODUCTION

Software Product Line (SPL) is a family of related software systems with common and variable functions whose first goal is reusability [1]. The SPL approach intends at upgrading software productivity and quality by relying on the similarity that exists among software systems, and by managing a family of software systems in a reuse-based way. SPL aims to minimize effort and cost of development and maintenance, to reduce time-to-market and to ameliorate quality of software [2], [3], [4]. Unsuitable development of a SPLs may give rise to bad programming practices, called code anomalies, also referred in the literature as "code smells" [5].

Code smell is often considered as key indicator of something wrong in the system code [5] or undesired code source property. Like all software systems, artifacts of a SPL may contains several code anomalies [6]. Therefore, if these code smells are not systematically removed, the SPL's quality may degrade due to evolution. Code Smells are very-known in classic and single software systems [7]. However, in the context of SPL, Code Smell is a young topic. [8] proposed a specific SPL's smell, called "Variability Smells". [9] discussed two types of bad smells related on SPL: Architectural Bad Smells and Code Bad Smells. [6] and [10] proposed detection strategies for anomalies in SPL.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 299-306, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101422

The main goal of this work is to propose a solution to reduce code smells in SPL. Unsuitable development of a SPLs may give rise to bad practices such as architectural smells and code smells. Our work tries to reduce development problems through the source code analyze of product variants to detect and correct code smells, identify the variability and build the variability model of SPL. Detecting and refactoring code anomalies in source code from the start give us a chance to develop a SPL with a high quality. Thus, the reverse engineering is a preliminary strategy for a clean SPL and to obtain the variability model of SPL.

This paper is organized as follow. Section 2 provides background on code smells, SPL and reverse engineering. Section 3 presents the related work. Section 4 shows the proposed approach. The last section concludes and presents future work.

2. BACKGROUND

2.1. Software Product Lines

The evolution of software development and the growth of product numbers have motivated the emergence of many reuse concepts. Software development communities recognize SPL as a successful approach for reuse [11], [12]. This success results from the reduction of production costs and time to market. SPL is a software development paradigm that share common feature to satisfy the specific needs of particular market segment [13].

Software product line's approach focus on the sharing of a reference architecture between products. These products can differ and the approach allows this variation with respect of particular characteristics and constraints. This difference is the variability present in SPL, which is the ability of a core asset to adapt to usages in the different product contexts that are within the product line scope [14]. Variability must be anticipated and continuously maintained to obtain wished results. The production process of product lines is well known as software product line engineering (SPLE) which tries to maximize the commonalities and reduce the cost of variations [15]. The SPLE process focuses on two levels of engineering [14]: Domain Engineering (DE) and Application Engineering (AE). DE focuses on developing reusable artifacts which are used in AE to construct a specific product. Fig. 1 presents the SPLE process.

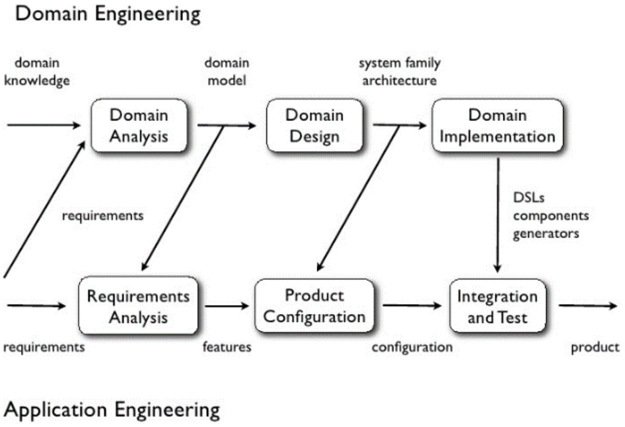


Figure 1. Domain Engineering and Application Engineering [14]

2.2. Code smells

A software system evolves over time. Its evolution is one of the critical phases of the process of its development. Moreover, the software system changes, moreover the structure of the program deteriorates. So, complexity increases until it becomes more profitable to rewrite it from the scratch. Which can involve threats on the software quality.

Software system's bad quality is a key indicator of existing bad programming practices, also known in the literature as source code flaw, code smells or code anomalies [5].

Code smells are usually symptoms of low-level problems such as anti-patterns. They are indicators of something wrong that structures in the source code [5], their presence can affect in maintenance and slow down software development.

In literature, different Code Smells have been defined. For instance, in Fowler's book [5], Beck define a list of 22 code smells, for example "Long Method" is a method that is too long and has too many responsibilities, so it makes code hard to maintain, understand, change, extend, debug and reuse. "Large Class" is a class contains many fields, methods or lines of code, means that a class is trying to do too much. "Duplicated Code" has negative impacts on software development and maintenance. For example, they increase bug occurrences: if an instance of duplicate code is changed in one part of the code for fixing bugs or adding new features, code may require various changes in other parts all over the source code simultaneously; if the correspondents are not changed inadvertently, bugs are newly introduced to them [16].

2.3. Reverse Engineering

Reverse Engineering is the process of analyzing a system. The purpose is to identify system structure, its components and the relationships between them [17].

Reverse Engineering can create representations of the system through transformations between or within different abstraction levels. It can also extract design information from source code [17] and may be used to re-implement the system.

The reverse engineering process can be done through automated analysis or manual annotations. The next steps concern the identification of program structure and the establishment of traceability matrix.

2.4. Refactoring

Refactoring's purpose is to improve the quality of an existing code [5]. This process tries through the software system changing to improve its internal structure without having an impact on the external behavior of the code.

Refactoring can be a solution for code smells. This process takes as input a source code with problems and outputs good ones. The resulting code can be reused. The refactoring allows the code smells identification. Also, it offers the possibility to change the original code containing these code smells by code restructuration to get an output code without code smells.

3. RELATED WORK

Common industrial practices lead to the development of similar software products, then they are usually managed to each other using simple techniques, e.g., copy-paste-modify. This is bad practice leading a low software quality, as we mentioned above the “Duplicated Code” code smell. During the past few years, several studies have investigated two things: how to detect code smells [18], [19], [20], [21], [22], [23] and how to correct [5], [18], [24] them in a single software. To the best of our knowledge we found few studies [6], [8], [9], [10], [25], [26] that can be considered related to our research.

[9] performed a Systematic Literature Review (SLR) to find and classify published work about bad smells in the context of SPL and their respective refactoring methods. They classified 70 different bad smells divided in three groups: (i) Code Smells; that are symptoms of something wrong in the source code, (ii) Architectural Smells; that are an indication of problem in higher levels of abstraction and (iii) hybrid Smells; that are a combination between architectural smell and code smells. [26] proposed a method to derive metric thresholds for software product lines.
The goal is to define thresholds values that each metric can take in order to identify potential problems in the implementation of features. They use 4 software metrics: Lines of Code (LOC) counts the number of uncommented lines of code per class. The value of this metric indicates the size of a class. Coupling between Objects (CBO) counts the number of classes called by a given class. CBO measures the degree of coupling among classes. Weight Method per Class (WMC) counts the number of methods in a class. This metric can be used to estimate the complexity of a class. Number of Constant Refinements (NCR) counts the number of refinements that a constant has. Its value indicates how complex the relationship between a constant and its features is. Their study is based on 33 SPLs which are divided into three benchmarks according to their size in terms of Lines of Code (LOC).

Benchmark 1 includes all 33 SPLs. Benchmark 2 includes 22 SPLs with more than 300 LOC. Finally, Benchmark 3 is composed of 14 SPLs with more than 1,000 LOC. The goal of creating three different benchmarks is to analyze the results with varying levels of thresholds. In term of that they illustrate a detection strategy to detect two types of code smells, God Class and Lazy Class. Figure 2 presents the way to identify God Class and Lazy Class.

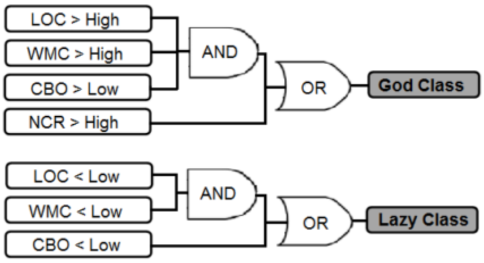


Figure 2. Code Smells identification.

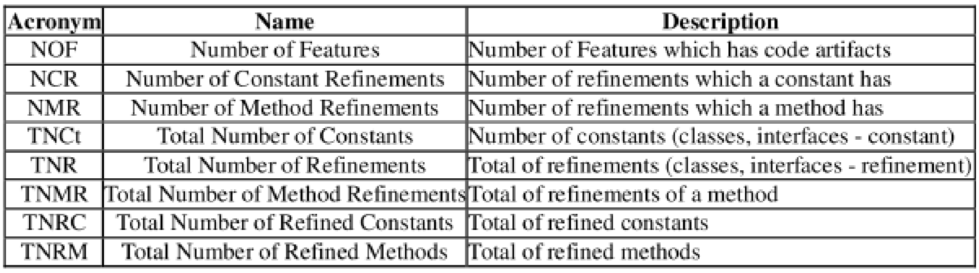
Apel et al. [8] proposed bad smell specific to SPLs called variability smell; that is an indicator of an existing undesired property in all kinds of artifacts in an SPL, such as feature models.

Fernandes and Figueiredo [6] investigated code anomalies in the context of SPLs, they propose new detection strategies for well-known anomalies in SPL such as God Class and God Method, ultimately they propose new anomalies and their detection strategies and they propose supporting tool for the proposed detection.

De Andrade et al. [25] conducted an exploratory study that aims at characterizing architectural smells in the context of software product line.

Abilio et al. [10] proposed means to detect three code smells (God Method, God Class, and Shotgun Surgery) in Feature-Oriented Programming source code, FOP is a specific technique to deal with the modularization of features in SPL. They performed an exploratory study with eight SPLs developed with AHEAD; which is an FOP language, to detect code smells in a SPL by using 16 source code metrics. These metrics corresponds to the detection of three code smells mentioned above. Table 1 presents some of these metrics.

Table 1. Metrics used to detect code smells [10]



Acronym	Name	Description
NOF	Number of Features	Number of Features which has code artifacts
NCR	Number of Constant Refinements	Number of refinements which a constant has
NMR	Number of Method Refinements	Number of refinements which a method has
TNCt	Total Number of Constants	Number of constants (classes, interfaces - constant)
TNR	Total Number of Refinements	Total of refinements (classes, interfaces - refinement)
TNMR	Total Number of Method Refinements	Total of refinements of a method
TNRC	Total Number of Refined Constants	Total of refined constants
TNRM	Total Number of Refined Methods	Total of refined methods

Considering the discussed related work, we propose an approach aiming to develop an SPL with minimal code smells risks.

4. PROPOSED APPROACH

The main goals in our study are to (i) investigate the state of the art on code smells in the context of SPLs as we show above, (ii) propose a solution to decrease code smells in developing software product lines.

Unsuitable development of a SPLs may give rise to bad practices such as architectural smells and code smells that induce maintenance and development costs problems. Therefore, we propose to build an SPL from the scratch using reverse engineering methods, which can help us to detect and correct code smells from the start. Thus, we can guarantee great quality of SPL.

The main challenge in this task is to analyze the source code of product variants in order to (i) detect and correct code smells, (ii) identify the variability among the products, (iii) associate them with features and (iiii) regroup the features into a variability model. The proposed approach is object-oriented language and only uses as input the source code of product variants.

First of all, we use as input source code of product variants then we apply detection strategies for code anomalies as duplicated code, uncovered code by unit tests and too complex code, after that we correct them using an automated bad smell correction technique based on the generation of refactoring concepts. Refactoring is a change made to the internal structure of software to rewrite the code, to “clean it up”, to make it easier to understand and cheaper to modify without changing its observable behavior [27]. In step 2 and after having a clean code, we are interested in the determination of the semantic relations between the names of the classes, the names of the methods and the attributes of all the source codes of the existing products having different terminologies and not necessary having the same meaning. In term of that we are interested in the harmonization of names, and more particularly in unifying fragments of source codes. During unification, we determine the semantic correspondences between the source code elements based on semantic knowledge base YAGO [28].

YAGO is a semantic knowledge base derived from many data sources like Wikipedia, WordNet, WikiData, GeoNames, and other. Aside YAGO, we will base on Machine Learning methods to get better semantic correspondences between source code elements. In fact, Machine Learning algorithms can be helpful in the classification of the features. Machine Learning proved his efficiency in many complex domains like Predictive Analytics [29], image processing [30], and signal processing… At the end of this step, all names with a semantic relationship would be harmonized and can be further analyzed in the next step of identifying commonalities and variability. Thus, we extract features by identification of common block (CB) and variation blocks (VB). CB groups the elements present in all the products while VB groups the elements present in certain products and not all of them. The role of these blocks is to group subsets to implement features. Once the common block and the variation blocks are completed, the extraction of mandatory elements and variation atomic blocks is supported, we associate them to features. Once the common properties and variability of product variants are identified, the feature model(s) will be constructed. Consequently, we can obtain one or more than one SPL. Our approach is presented in Figure3.

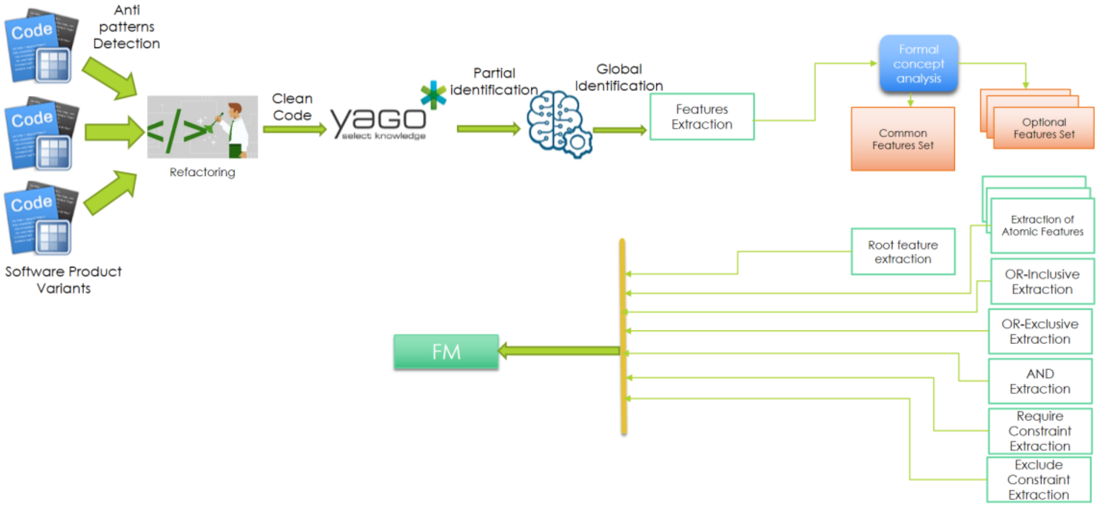


Figure 3. Proposed Approach.

5. Conclusions

Software reuse is an important challenge in software engineering. Software Product Line is one of the technique used to ensure the success of this challenge. The obtained products can contain reused parts or components. These parts can include some problems in their source code more known as Code Smells. These problems can propagate between the different products.

A solution to avoid the Code smells in source code, is refactoring which can improve the internal structure of software system by trying to find the problem and avoid it using some restructuration techniques.

In this paper, we try to present an approach which combines refactoring to eliminate code smells and reverse engineering to propagate modifications to the design level. Our purpose is to obtain a software product line model free from code smells.

Our future works will be the refinement of the different parts of the approach. Also, we will choose the appropriate tools to use in our prototype.

REFERENCES

[1] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC *Transactions on ECE*, Vol. 10, No. 5, pp120-122.

[2] Gizem, Aksahya & Ayese, Ozcan (2009) *Coomunications & Networks*, Network Books, ABC Publishers.

[1] Rincón, L. F. et al., 2014. An ontological rule-based approach for analyzing dead and false optional features in feature models. In Electronic Notes in Theoretical Computer Sciences, Vol. 302, pp 111–132.

[2] Jacobson, I. et al., 1992. Object-oriented software engineering: a use case driven approach. Addison-Wesley, USA.

[3] Xue, Y., 2011. Reengineering legacy software products into software product line based on automatic variability analysis. Proceedings of the 33rd International Conference on Software Engineering, New York, USA, pp. 1114–1117.

[4] Ouali, S. et al., 2012. From Intentions to Software Design using an Intentional Software Product Line Meta-Model. Proceeding of the 8th International Conference on Innovations in Information Technology, Al Ain, UAE.

[5] Fowler, M., 1999. Refactoring: Improving the Design of Existing Code. Addison-Wesley, Boston, MA, USA.

[6] Fernandes, E. and Figueiredo, E., 2017. Detecting Code Anomalies in Software Product Lines". Proceedings of 7th Brazilian Conference on Software: Theory and Practice, Maringa, Brazil, pp. 49–55.

[7] Zhang, M. et al., 2011. Code Bad Smells: A Review of Current Knowledge. In Journal of Software Maintenance and Evolution: Research and Practice, Wiley Online Library, pp. 179-202.

[8] Apel, S. et al., 2013. Feature-Oriented Software Product Lines: Concepts and Implementation. Springer.

[9] Vale, G. et al., 2014. Bad Smells in Software Product Lines: A Systematic Review. Proceedings of the 8th Brazilian Symposium on Software Components, Architectures and Reuse (SBCARS), Brazil, pp. 84-94.

[10] Abilio, R. et al., 2015. Detecting Code Smells in Software Product Lines-An Exploratory Study. Proceeding of the 12th International Conference on Information Technology - New Generations, pp. 433–438.

[11] Ouali, S. et al., 2011. Framework for evolving software product line. In International Journal of Software Engineering & Applications, Vol. 2, No. 2, pp. 34-51.

[12] Weiss, D. M. and LAI, C. T. R., 1999. Software Product-Line Engineering: A Family-Based Software Development Process. Addison-Wesley.

[13] Pohl, K. and Metzger, A., 2006. Software Product Line testing. In Communication of the ACM, pp78-81.

[14] Czarnecki, K. and Eisenecker, W., 2000. Generative Programming: Methods, Tools, and Applications. Addison-Wesley.

[15] Thiel, S. and Hein, A., 2002. Modeling and Using Product Line Variability in Automotive Systems. In IEEE Software Vol.19, No.4, pp. 66-72.

[16] Hotta, K. et al., 2012. An Empirical Study on the Impact of Duplicate Code. Advances in Software Engineering.

[17] Chikofsky, E. J. and Cross, J. H., 1990. Reverse engineering and design recovery: A taxonomy. IN IEEE Software, Vol.7, pp. 13–17.

[18] Moha, N. et al., 2010. DECOR: A Method for the Specification and Detection of Code and Design Smells. Transactions on Software Engineering, Vol. 36, No. 1, pp. 20-36.

[19] Sjoberg, D. et al., 2013. Quantifying the effect of code smells on maintenance effort. Software Engineering IEEE Transactions, Vol. 39, No. 8, pp. 1144–1156.

[20] Van Emden, E. and Moonen, L., 2002. Java quality assurance by detecting code smells. Proceeding Working Conf. Reverse Engineering, IEEE Computer Society Press, pp. 97—107.

[21] Marinescu, C. et al., 2005. Iplasma: An integrated platform for quality assessment of object-oriented design. Proceedings of the 21st IEEE International Conference on Software Maintenance, Budapest, Hungary.

[22] Liu, X. and Zhang, C., 2016. DT: a detection tool to automatically detect code smell in software project. Proceedings of the 4th Int. Conf. Mach. Mater. Inf. Technol. Appl., vol. 71, pp. 681–684.

[23] Fontana, F. et al., 2012. Automatic Detection of Bad Smells in Code: An Experimental Assessment. In Journal of Object Technology.

[24] Campbell, D. and Miller, M., 2008. Designing refactoring tools for developers. Proceedings of the 2nd Workshop on Refactoring Tools, New York, NY, USA.

[25] De Andrade, H. S. et al., 2014. Architectural bad smells in software product lines. Proceedings of the 1st International Conference Dependable Secur. Cloud Comput. Archit., pp. 1–6.

[26] Vale, G. and Figueiredo, E., 2015. A Method to Derive Metric Thresholds for Software Product Lines. Proceedings 29th Brazilian Symposium on Software Engineering, pp. 110–119.

[27] Regulwar, G. B. and Tugnayat, R. M., 2012. Bad Smelling Concept in Software Refactoring. International Proceedings of Economics Development and Research, Vol.45, pp. 56–61.

[28] Rebele, T. et al., 2016. YAGO: a Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. Proceeding of the 15th International Semantic Web Conference, Kobe, Japan.

[29] Demsar, J. et al., 2004. Orange: From experimental machine learning to interactive data mining. Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy.

[30] Ebrahimi, K. S. et al., 2013. Combining modality specific deep neural networks for emotion recognition in video. Proceedings of the 15th ACM on International conference on multimodal interaction, Sydney, Australia.

AUTHORS

Sami Ouali is an assistant professor in the College of Applied Sciences of lbri in Oman. He is a member of the RIADI labs, Tunisia. His research interests lie in the areas of software engineering and software product line.

DATA DRIVEN SOFT SENSOR FOR CONDITION MONITORING OF SAMPLE HANDLING SYSTEM (SHS)

Abhilash Pani, Jinendra Gugaliya and Mekapati Srinivas

Industrial Automation Technology Centre, ABB, Bangalore, India

ABSTRACT

Gas sample is conditioned using sample handling system (SHS) to remove particulate matter and moisture content before sending it through Continuous Emission Monitoring (CEM) devices. The performance of SHS plays a crucial role in reliable operation of CEMs and therefore, sensor-based condition monitoring systems (CMSs) have been developed for SHSs. As sensor failures impact performance of CMSs, a data driven soft-sensor approach is proposed to improve robustness of CMSs in presence of single sensor failure. The proposed approach uses data of available sensors to estimate true value of a faulty sensor which can be further utilized by CMSs. The proposed approach compares multiple methods and uses support vector regression for development of soft sensors. The paper also considers practical challenges in building those models. Further, the proposed approach is tested on industrial data and the results show that the soft sensor values are in close match with the actual ones.

KEYWORDS

Sample Handling System, Soft-Sensor, Variance Inflation Factor (VIF), Local Outlier Factor (LOF), Support Vector Regression.

1. INTRODUCTION

Adverse impacts of rapid industrialization on world's environment are acknowledged worldwide which are mostly irreversible. Hence, various government agencies along with industries have started monitoring emissions to control associated environmental pollution. Continuous emission monitoring (CEM) devices are used across industries for monitoring real-time pollutant content in flue gases [1] [2] . As governments across the globe becomes more vigilant and stringent on emission norms, reliability and availability of CEM systems have become very crucial. Reliability of CEM systems are dependent not only on CEM devices but also on associated sample handling systems (SHSs). As reported in [3] majority of failures in CEM systems are due to issues in SHSs. Therefore, manufactures have started offering sensors for condition monitoring of SHSs, which helps in improving reliability of these systems and hence reliability of overall CEM systems.

Performance of any condition monitoring system depends on sensors and is adversely impacted by sensor failures. Therefore, many studies have discussed methods for sensor fault detection and isolation. In [4] autocorrelation is used to detect sensor failure of pitot static system in airplanes. Spectral clustering technique based faulty sensor detection and deletion from wireless sensor network is proposed in [5] . [6] provides a detailed review on sensor fault detection methods and report that 40% of literature on sensor fault detection are based on either principal component

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 307-320, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101423

analysis (PCA) or artificial neural network (ANN). [7] is one of the earliest and most cited paper on using PCA for sensor fault detection. In [8] feed forward neural network and Locally weighted regression are proposed for sensor fault detection in Predictive Emission Monitoring System.(PEMS) unlike traditional CEMs, pollutant concentration is estimated using process data and parameters instead of measuring them directly. Once faulty sensor is detected and isolated using fault detection technique, data driven soft sensor model is used to estimate the true value of the faulty sensor. These estimates can be used in place of faulty hardware sensor measurement which adds fault resilience characteristics to condition monitoring systems [9] . Recently [10] has proposed a deep learning-based vision sensing applied to printing quality control. Online adaptive ensemble PLS approach is proposed for a chemical process in [11] A Gaussian probabilistic regression approach [12] is proposed to develop soft sensor. A hierarchical clustering method is proposed in [13] .

However, in literature there is not enough material which provides a detailed framework for development of data driven soft sensor for SHS. This is essential as there are practical constraints which are important to be considered during development of soft sensors.

This work focuses on a framework to develop data driven soft sensor for condition monitoring of SHSs. There are few practical assumptions/constraints which are considered during development of the framework. They are mostly related to SHS system and are discussed in the upcoming sections. Rest of the paper is organized as follows. Section 2 introduces Sample Handling System, Section 3 discusses the objective of the work, Section 4 details on the proposed framework, Section 5 provides the results with an use case and Section 6 concludes the study.

2. SAMPLE HANDLING SYSTEM

The main objective of SHSs is to (1) Provide a path for sample collection (2) Transport collected sample without contamination (3) Remove particulate matter and moisture present in the gas sample (4) Maintain desired temperature and regulated flow of gas sample to CEM device.

There are multiple stages to ensure above objectives are achieved. Figure 1 provides a block diagram of SHSs. Each stage in Figure 1 will have one or more components and to monitor functioning of these components there are multiple sensors placed across SHS. The SHS considered in this analysis has 2 temperature sensors, 3 pressure sensors and one flow sensor.

The first stage of SHS consists of sample probe and filter components which are responsible for collecting sample gas and to filter particulate matter present in it. Normally this stage is placed close to exhaust stack and away from remaining stages of SHS. The distance between first stage and remaining stages varies from plant to plant and in some cases can reach values close to 500 meters. The second stage of sample processing consists of temperature treatment which ensures that temperature of collected sample does not drop below certain threshold value to prevent condensation of available moisture. The third stage is responsible for removing moisture by cooling the incoming gas sample. This stage also removes the collected condensate from the process. Fourth stage contains sample pump which is the heart of the SHS. Sample pump produces a pressure difference to ensure enough flow of sample gas to CEM devices. Control valves present in fourth stage regulate the flow as per design specification.

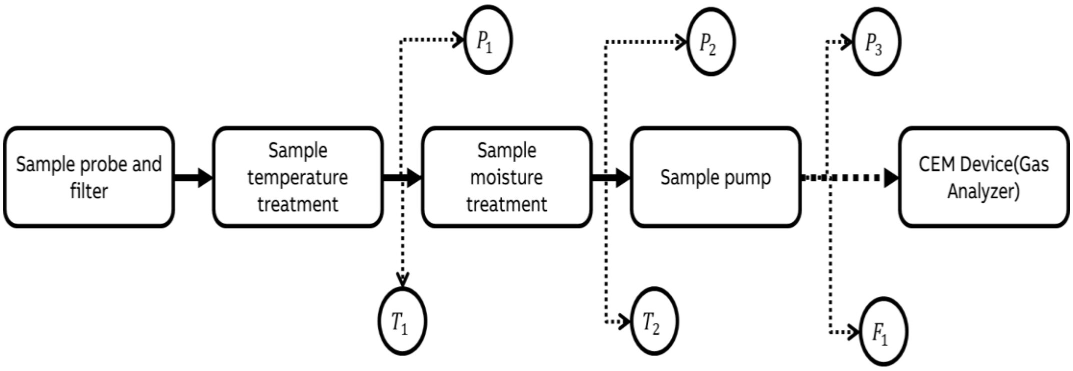


Figure 1: Block Diagram of Sample Handling System

3. OBJECTIVE AND ASSUMPTIONS

The objective of this work is to build a framework/pipeline to develop soft sensors for SHS such that each soft sensor can model measurement of a faulty hardware sensor using measurements from other hardware sensors. In order to do so there are few assumptions which were made considering practical experience.

Assumption 1. The framework was developed for single sensor failure only, which means developed soft sensors will only work if there are single sensor failures. This assumption was considered as simultaneous failure of multiple sensor are rare in field. Secondly the time between two successive hardware sensor failures is large enough to schedule a planned shutdown of process to replace failed sensor.

Assumption 2: Availability of training data will be less (~4000 samples). As developed framework utilizes machine learning approach, historical data from all hardware sensors are required for training. For a new installation getting a large amount of training data is not feasible. Therefore, in this study we have not considered machine learning algorithms which need large volume of data for its training.

Assumption 3: The available training data is collected during normal operation of SHS. For a new installation with all testing done, it is extremely rare to encounter fault/failure and therefore, this assumption should be validated before developing soft sensor.

4. FRAMEWORK

The proposed framework for soft sensor development can be divided into two major modules.

1. Data Pre-processing module
2. Machine learning algorithm evaluation and selection module.

4.1. Data Pre-Processing Module

This is the first module in the framework which ensures quality of data for modeling of soft sensor. This module performs 5 data preprocessing steps, and the details of each step are given in the following.

4.1.1. Missing Value Imputation

Missing observations are common in any data and there are many imputation methods available in literature [14] . Each method has its own advantage and disadvantage and the best imputation method mainly depends on amount of missing data and its type. In this work we have considered time series data from SHS and hence central tendency-based imputation methods (mean/median imputation methods) are not suitable. Imputation by last observation carried forward, imputation by next observation carried backward and imputation by interpolation are three popular methods used for imputation in time series analysis. In this study imputation by interpolation is considered as it considers both previous and next observation value for imputation.

4.1.2. Removal of Off Condition and Outlier

In the training data there are off conditions where SHS is offline. These samples with off condition should be removed before proceeding for further analysis as these samples may impact soft sensor models adversely. The off condition can be checked using sensor measurement. In this study, SHS is considered off-line if all pressure measurements are 0.

Given a sample from an unknown population a point is labeled as outlier if it is located away from majority of the samples. This is known as distance-based outlier definition. According to density-based outlier definition, a point is labeled as outlier if the point is present in a low-density region in multidimensional feature. There are many approaches for outlier detection and removal. In this work we have considered a two-stage outlier removal process in which stage 1 removes distance-based outliers whereas stage 2 removes density-based outliers.

In stage1, quartile-based univariate thresholds for each variable are calculated as in the following

$$\text{Upper Threshold} = Q_3 + 1.5 \times IQR$$
$$\text{Lower Threshold} = Q_1 - 1.5 \times IQR$$

Where $Q_1$and Q_3represents first and third quartiles and $IQR$is the inter quartile range calculated as $IQR = Q_3 - Q_1$.This is a simple distance based univariate approach and is performed first to remove outliers which are far away from majority of population.

In stage 2, density-based local outliers are removed using Local outlier factor (LOF). LOF assigns an outlier score to each sample based on relative density of that sample with respect to its K nearest neighbors. More details on LOF can be found in[15] .

4.1.3. Removal of Features with Multicollinearity

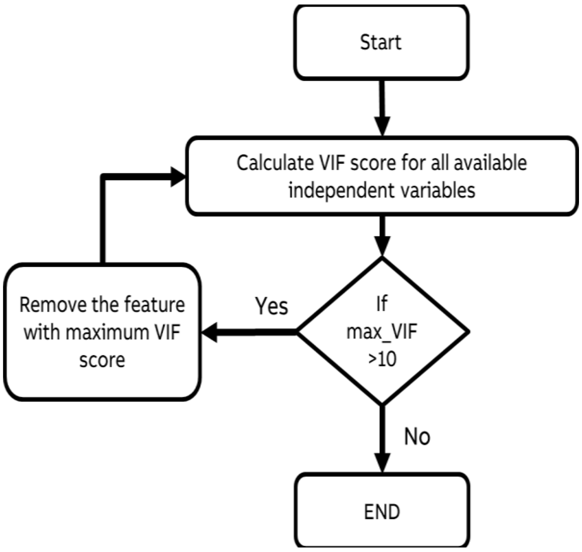
Multicollinearity is a phenomenon in which one or more independent variables can be expressed as a linear combination of other independent variables. In the presence of multi collinearity influence of independent variables on target variable cannot be estimated accurately. Therefore, interpretation of trained models becomes difficult.

Correlation among independent variables can be calculated using Pearson correlation or Spearman correlation coefficient. These correlation coefficients calculate correlation among two independent variables at a time, which is a major limitation. Therefore, in proposed framework, Variable inflation factor (VIF) score is used to identify and remove correlated variables. VIF score for an independent variable is calculated by regressing it against every other independent variable in the model according to the below equation.

$$VIF_i = {}^{1}/_{(1-R_i^2)}$$

Where VIF_i is the VIF score for i^{th} independent variable and R_i^2 is the coefficient of determination (R_squared value) obtained by regressing i^{th} independent variable against every other independent variable. A VIF score of 1 indicates no correlation. A VIF score of more than 10 is considered as extremely correlated and corresponding variable /feature should be dropped. However, dropping all variables with VIF score more than 10 at once is not a good strategy. For example, let's take a regression with 4 independent variables $[X_1, X_2, X_3, X_4]$ and assume that X_1, X_3 and X_4 not correlated. However,X_2 can be expressed as $3X_3 + 4X_4$. By rearranging this equation we can show that there is multicollinearity among $[X_2, X_3, X_4]$. Therefore, for this set up, VIF score for X_1will be minimum and will be higher for $[X_2, X_3, X_4]$. Let's assume that VIF score for X_2, X_3 and X_4 is greater than 10 and if all variables with VIF >10 are dropped at once loss of uncorrelated variables may happen (in this case X_3 and X_4). This is not an efficient way. Therefore in this framework an iterative removal of features based on VIF score is proposed and detailed steps are shown in

Figure 2.

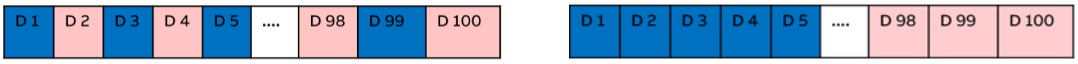

Start
Calculate VIF score for all available independent variables
If max_VIF >10
Yes
Remove the feature with maximum VIF score
No
END

\

Figure 2 Flow chart of Iterative feature removal method

4.1.4. Time based Data Split

It is a standard practice to divide data into three sets namely train, validation and test datasets. Popularly this is done at random with each sample having uniform probability. This is not a suitable strategy for splitting time series data. In case of time series there is an inherent temporal dependency and hence the test accuracy obtained by random splitting will be misleading. There for in this framework, time-based data splitting is used. A visual representation of random and time series splitting is provided in Figure 3.



(b)

(a)

Figure 3: (a) Random splitting and (b) Time-based splitting

4.1.5. Data Normalization

Data normalization is the process of transforming each independent variable such that transformed variables will have 0 mean and 1 standard deviation and normalization step should follow data splitting step. The sequence of last two preprocessing steps is important as normalization of test data should happen based on training data distribution otherwise data leaking issue will arise.

4.2. Machine Learning Algorithm Evaluation and Selection Module

Considering assumption 2 (in Section 3), we have considered 5 algorithms in this framework. Details of these algorithms can be found in [16] and [17] .

1. Linear regression
2. K nearest neighbour
3. Decision Tree
4. Random Forest
5. Support vector regression

There are other machine learning algorithms like ensemble models (GBDT, stacking) and Neural network which are proven accurate for learning complex relation between independent and dependent variable. However, due to their higher flexibility they are prone to overfitting. Considering low volume of training data overfitting issue will worsen. Therefore, these algorithms are not considered in the framework.

In this module hyper parameter tuning for each of the algorithms is done using grid search. In order to evaluate these algorithms, mean square error is considered in this framework. After evaluation, the best model is considered for modeling of soft sensor.

5. RESULTS AND DISCUSSION

In this section results of soft sensor models built using proposed framework are presented. Process data from one of the CEM system installation in India was used for evaluation of the proposed approach. One month of process data with 4,320 data points with 6 features is used for building regression models for soft sensor. The feature values are standardized to 0 mean with 1 standard deviation. In this work temperature measurements are represented as T_1 and T_2. Similarly, pressure measurements are represented as P_1 , P_2 and P_3 and flow measurement is represented as F_1.

After removal of missing values and off conditions from available dataset, two stage outlier removal is performed. As discussed, earlier distance-based outliers are removed in the first stage followed by removal of density-based outliers using LOF scores in second stage. In order to visualize identified outliers, multidimensional feature space is embedded into two-dimensional space using t-distributed Stochastic Neighbour Embedding (t-SNE). In Figure 4, scatter plot of dataset with identified outliers is shown. In this plot inliers/normal data points, outliers identified using IQR method (distance based) and outliers identified by LOF score (density based) are represented with circular, star and square markers respectively. Number of data point belonging to normal, distance based outlier and density based outlier are provided in Table 1. Distance based outliers are located towards outer edges of the scatter plot, whereas density-based outliers are present towards inner side of the plot.

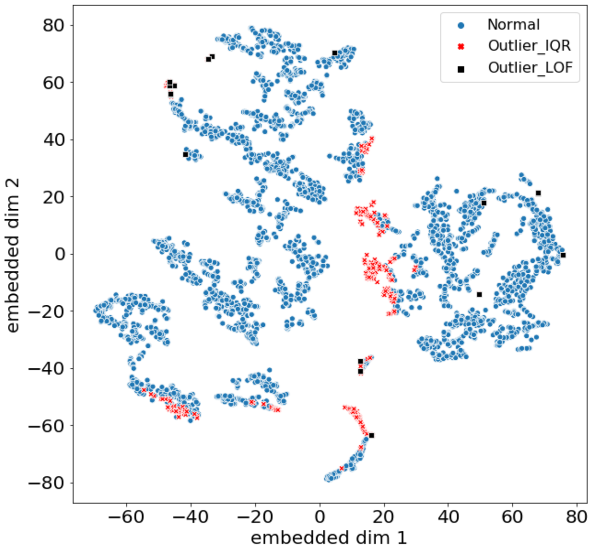


Figure 4: Outliers in the given data set

Table 1: Number of normal data points and outliers

Point type	**Number of samples**
Normal	3813
Distance based Outliers	312
Density based Outliers	21

Next step in data pre-processing is the removal of features with multicollinearity using a recursive method. In oder to showcase results of proposed method let's consider the case of modeling of soft sensor for T_2. where, remaining 5 measurements are considered as predictors and multicolinearity is evaluated for these 5 predictors. Feature-wise log VIF scores are plotted in Figure 5 and absolute VIF scores are provided in Table 2. VIF score of 10 is considered as threshold for feature elimination. From the plot it is evident that P_1 and P_2 have VIF score more than the threshold.

Using recursive feature elemination method first P_1 is removed from predictor list as it has highest VIF score and VIF scores for remaining 4 predictors are evaluated again. VIF scores in second itteration arepresented in Table 2. As evident from this table,after droping P_1, VIF scores for remaining 4 predictors including P_2 are less than 10, which indicates no sever multi colinearity. Therefore, as discussed earlier , droping P_1 and P_2 at once based on threshold is not a good statergy as multicolinearity in P_2 can be eliminated by droping P_1 only.

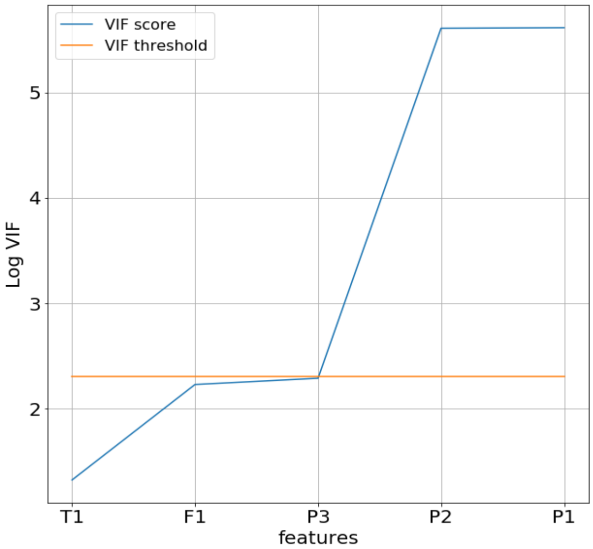


Figure 5: Log VIF Scores of all independent features

Table 2: VIF Scores of Different Features

variables	**VIF score**	
	Iteration 1	**Iteration 2**
T_1	3.75682	1.298446
F_1	9.296621	3.734248
P_3	9.861192	6.246676
P_2	273.2662	7.746631
P_1	274.688	---

To demonstrate impact of features having sever multicollinearity on regression model, modelling of soft sensor for P_3 is considered. Two different models were trained using decision tree algorithm. All 5 features are considered for training of model 1 whereas, feature P_1 is dropped from feature list while training model 2. Model 1 identified $P_1$and P_2 as top two important features for prediction of P_3. However, order of feature importance and their corresponding values change drastically when training datset was changed slightly. This makes interpretation of feature iportance difficult in preesence of multicolinearity . P_2 and F_1 are identified as top two important features by modle 2. Order and value of feature importance are consistant compared to model 1 which makes impterpretation easier. This problem becomes even more significant for systems with large number of features and therefore it is a good practice to remove feature with sever multicolinearity.

After removal of features with multi colinearity first 2,860 (~75%) samples are considered as training data set and remaining as test dataset. These datasets are used to train and evaluate soft sensor models developed using 5 different machine learning algorithms namely linear regression(LR), K nearest neighbour (KNN), support vector regression(SVR) , decision tree(DT) and random forest(RF). Hyper parameter tunning for each algorithm is perfomred by further splitting training data into train and cross validation dataset. Table 3 provides tuned parameter vaues and comparison of aforementioned algorithms is performed using mean square error.

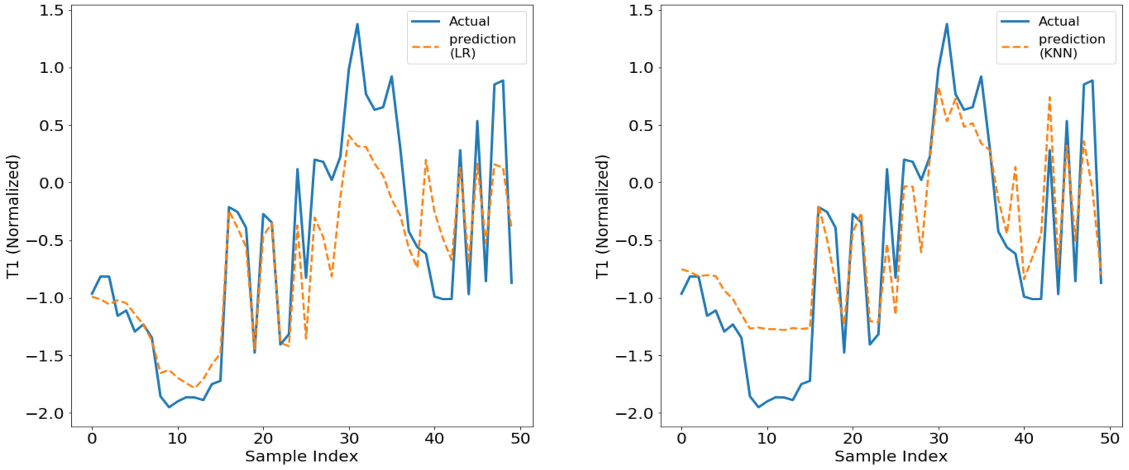
Table 3 Parameter values after hyperparameter tuning

Algorithm	**Parameter values**
Linear regression	L2 Regularization parameter = 0.1
K nearest neighbour	Number of nearest neighbour = 50
Decision tree	Maximun depth = 5
Random forest	Maximum Depth = 50, Number of estimator = 500
Support vector regression	Regularization parameter(C) = 10, Kernel = “RBF”, Kernel coefficient (gamma) = 0.2

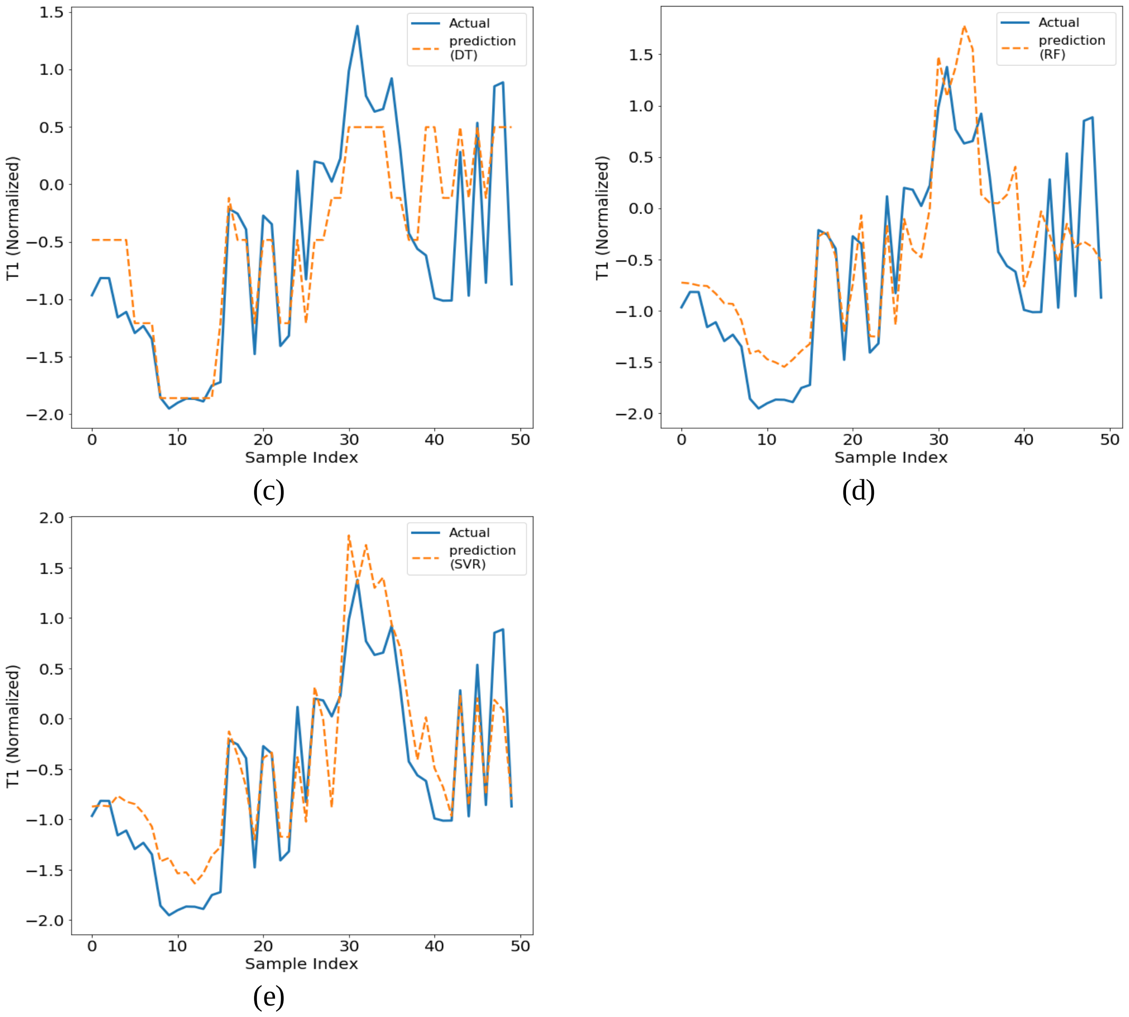
To compare above mentioned algorithms, modeling of soft sensor for T_1 is considered and the calculated mean sqaure error values for both train and test datasets are given in Table 4. From this table it is evident that SVR has minimum test MSE followed by RF. However, the difference in train and test MSE for model obtained by RF algorithm has higher varience issue. This issue can be avoided by increasing number of base estimators in RF given higher volume of training data. Due to the constraint of low volume of training data (assumption 2) ensembling algorithms and neural networks are not used for this application. Plot of actual and predicted values by various algorithms is shown in Figure 6 for visual comparison. The prediction by SVR model is very close to actual values of T_1 followed by that of RF and DT. Predicted values of LR and KNN models could not capture peak patterns in T_1 which are captured by other three algorithms. From the above comparative analysis, SVR with RBF (radial basis function) kernel is selected for modeling of soft sensor in SHSs.

Table 4: MSE values for different ML Algorithms for $\boldsymbol{T_1}$ Soft Sensor Model

Algorithms	**Linear Regression**	**KNN**	**SVR**	**Decision Tree**	**Random forest**
Train MSE	0.403	0.214	0.134	0.310	0.111
Test MSE	0.307	0.306	0.134	0.289	0.236



(a) (b)



(c)

(d)

(e)

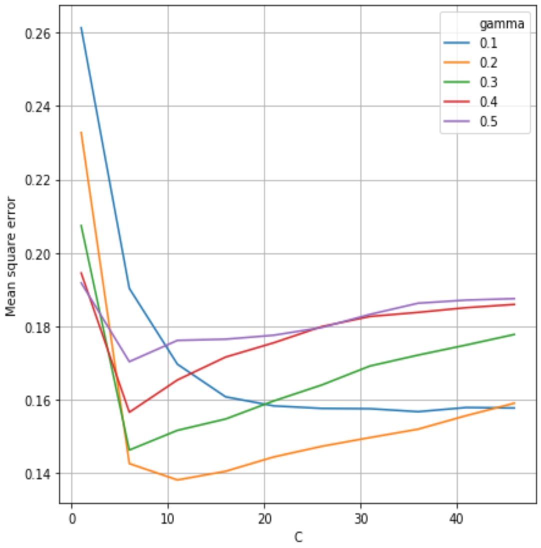
Figure 6: Actual and Predicted Values for $\boldsymbol{T_1}$ Soft Sensor (a) Linear regression (b) K nearest Neighbour (c) Decision Tree (d) Random Forest (e) SVR

Performance of SVR with RBF kernel is superior as it can learn complex nonlinear function by projecting data to higher dimension using Kernel trick. T is advantage becomes more significant when amount of training data is limited. Impact of tuning parameters on performance of SVR is shown in Figure 7. Two parameters of SVR with RBF kernel are considered in this analysis. Parameter C is the regularization parameter of SVR, and strength of the regularization is inversely proportional to C. Parameter gamma represents spread of radial basis function. A range of values for C and gamma are selected and a grid search approach is used for parameter tuning. From Figure 7 it is evident that minimum MSE on cross validation data was obtained for C= 10 and gamma = 0.2. Hence a soft sensor model for T_2 using SVR with RBF kernel, C= 10 and gamma = 0.2 is developed and plot of actual and predicted value by this model is shown in Figure 8. From this figure it can be observed that predicted values could capture overall trend, however model is not able to capture extremely peak patterns in actual data.

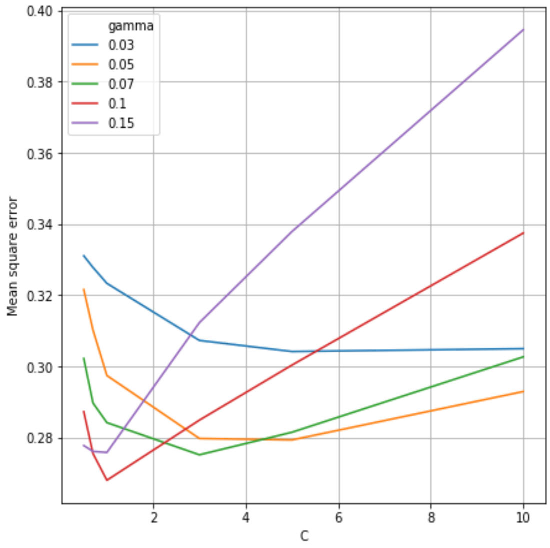
Similar approach is adapted for modeling soft sensors for other measurements/physical sensors and obtained test and train mean square error along with R square values are provided in Table 5.

Table 5: R Square and MSE values for Different Sensors

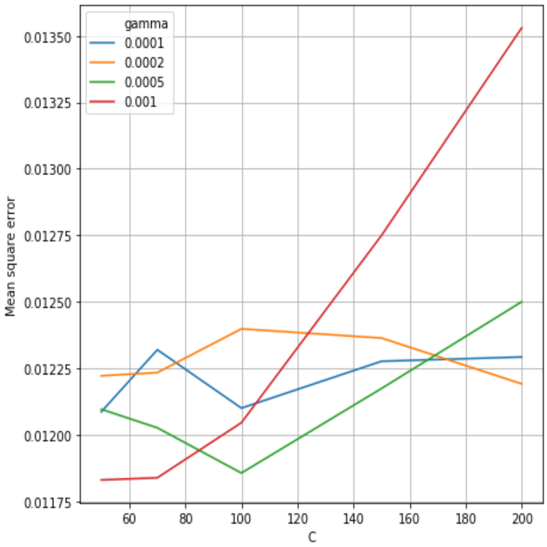
Physical sensor	**Test MSE**	**Train MSE**	**Test R square**	**Train R square**
T_1	0.23	0.13	0.83	0.89
T_2	0.15	0.20	0.78	0.80
P_1	0.011	0.008	0.83	0.93
P_2	0.006	0.007	0.96	0.99
P_3	0.002	0.001	0.85	0.90
F_1	0.04	0.02	0.746	0.83



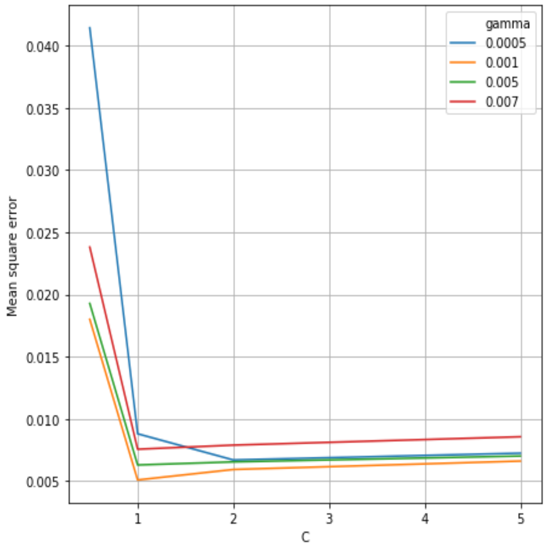
T_1



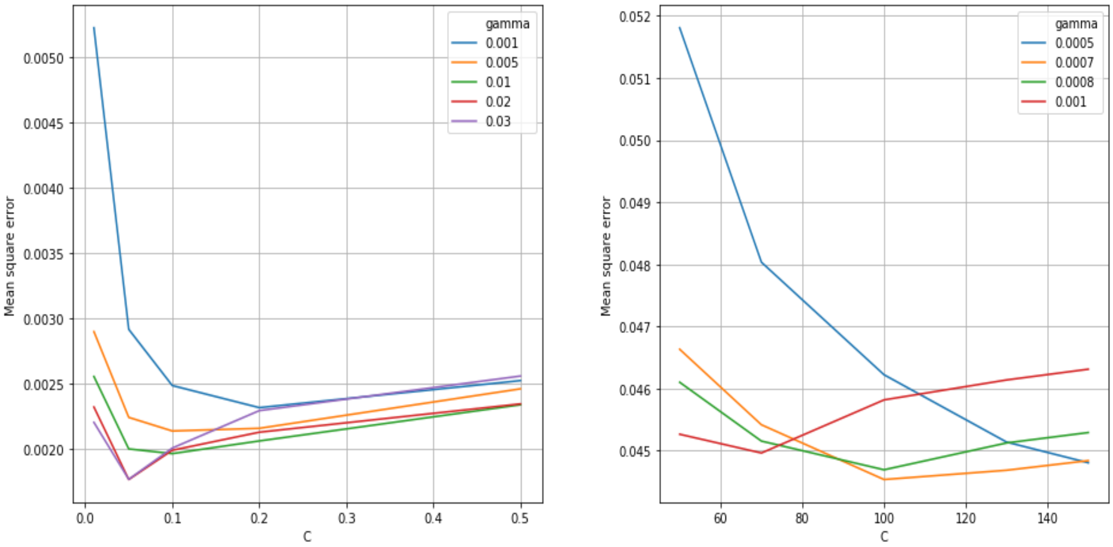
T_2



P_1

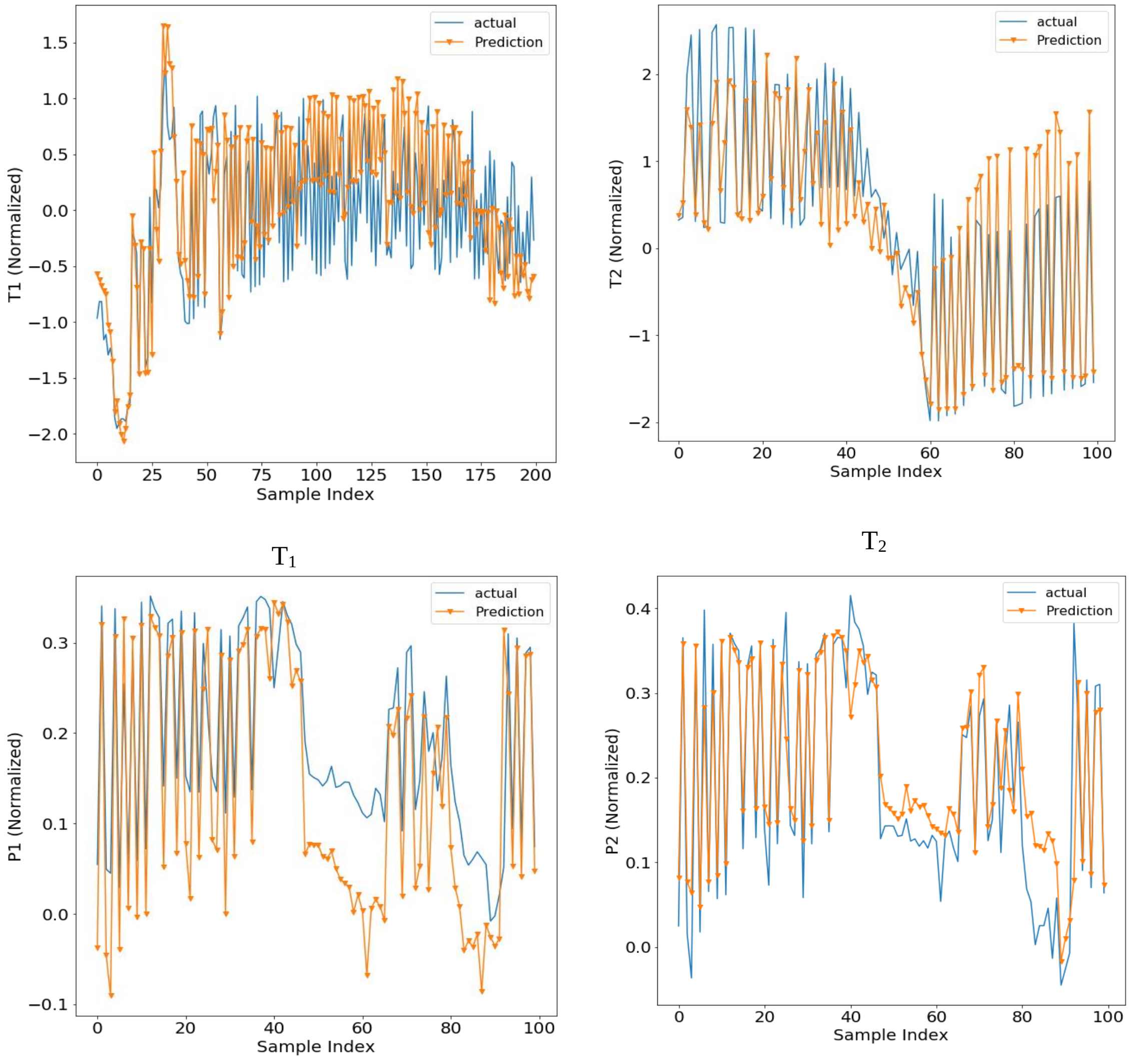


P_2



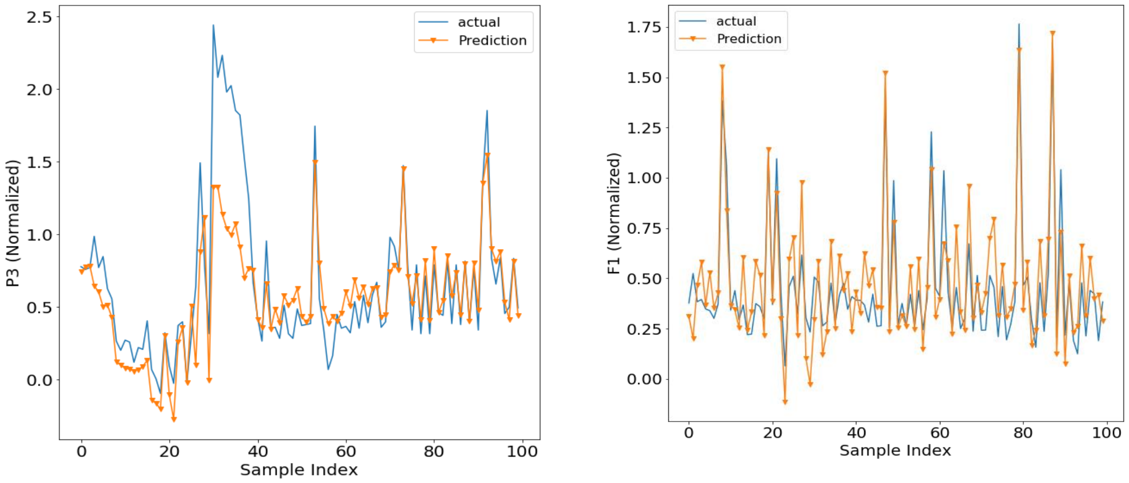
P$_3$ F$_1$

Figure 7: Plot of MSE for various Gamma and C values of SVR for 6 Soft Sensor Models



T$_1$ T$_2$

P$_1$ P$_2$



P$_3$ F$_1$

Figure 8: Actual and Predicted Values for 6 Sensors using SVR

6. CONCLUSIONS

In this study a framework is proposed for developing data driven soft sensor for sample handling system in CEMs. The framework consists of two modules: (i) Data Preprocessing module (ii) Machine learning algorithm evaluation and selection module. In module (ii), 5 machine learning algorithms which includes Linear Regression, KNN, SVR(RBF), Decision Tree and Random Forest are evaluated on industrial data that consists of 6 SHS measurements. From the comparison SVR is found to be better than other methods in predicting the values for all the 6 SHS measurements. The future work would involve exploration of Deep ML approaches and compare their performance against the proposed approach when data availability is not a constraint.

REFERENCES

[1] J. Jahnke, Continuous Emission Monitoring. Wiley, 2000.

[2] E. Arioni, N. Bonavita and M. Paco, “Keeping an eye on emissions,” Hydrocarbon Engineering Magazine, vol. 18, no. 10, pp. 43–49, October 2013.

[3] Y. Yang, X. Zhang, Z. Zhao, G. Wang, Y. He, Y. Wu and J. Li, "Applying Reliability Centered Maintenance (RCM) to Sampling Subsystem in Continuous Emission Monitoring System" IEEE Access, vol. 8, pp. 55054-55062, 2020.

[4] Swischuk, Renee C. and Douglas L. Allaire. “A Machine Learning Approach to Aircraft Sensor Error Detection and Correction.” Journal of Computing and Information Science in Engineering, vol. 19, pp. 1-19, 2019.

[5] A. M. T. Nasser and V. P. Pawar, "Machine learning approach for sensors validation and clustering," 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, pp. 370-375, 2015.

[6] H.Y. Teh, A. W., Kempa-Liehrand K. Wang. “Sensor data quality: a systematic review.” Journal of Big Data, vol. 7, pp. 1-49, 2020.

[7] R. Dunia, S. J. Qin, T. F. Edgar and T.J. McAvoy, “Use of principal component analysis for sensor fault identification” Computers & Chemical Engineering, vol. 20, pp. S713-S718, 1996.

[8] G. Ciarlo, E. Bonica, B. Bosio and N. Bonavita, “Assessment and Testing of Sensor Validation Algorithms for Environmental Monitoring Applications”, Chemical Engineering Transactions, vol. 57, pp. 331-336, Mar. 2017.

[9] D. Angelosante, M. Guerriero, G. Ciarlo and N. Bonavita, "A Sensor Fault-Resilient Framework for Predictive Emission Monitoring Systems," 21st International Conference on Information Fusion (FUSION), Cambridge, pp. 557-564, 2018.

[10] Villalba-Diez, J.; Schmidt, D.; Gevers, R.; Ordieres-Meré, J.; Buchwitz, M.; Wellbrock, W. Deep Learning for Industrial Computer Vision Quality Control in the Printing Industry 4.0. Sensors 2019, 19, 3987.

[11] Cang, W.; Yang, H. Adaptive soft sensor method based on online selective ensemble of partial least squares for quality prediction of chemical process. Asia-Pac. J. Chem. Eng. 2019, 14, 2346.

[12] Xiong, W.; Shi, X. Soft sensor modeling with a selective updating strategy for Gaussian process regression based on probabilistic principle component analysis. J. Frankl. Inst. 2018, 355, 5336–5349.

[13] Yu, W. A mathematical morphology based method for hierarchical clustering analysis of spatial points on street networks. Appl. Soft Comput. 2019, 85, 105785.

[14] W. Young, G. Weckman and W. Holland, “A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits.“ , Theoretical Issues in Ergonomics Science, vol. 12(1), pp. 5–43, 2011.

[15] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jorg Sander, "LOF: Identifying Density-Based Local Outliers", Proc. of the 2000 ACM SIGMOD on Management of Data, pp. 93-104, 2000.

[16] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press, 2014.

[17] H. Drucker, C. Burges,L. Kaufman, A. Smola and V. Vapnik, "Support Vector Regression Machines", Proceedings of the 9th International Conference on Neural Information Processing Systems (NIPS'96). MIT Press, Cambridge, MA, USA, 155–161.

AUTHORS

Abhilash Pani is currently working as scientist ABB Industrial Automation Technology Centre Bangalore and his areas of interests are applied machine learning for industrial analytics and condition monitoring of industrial assets.

Jinendra K Gugaliya is working as Principal Scientist at ABB Industrial Automation Technology Centre Bangalore and his areas of interests are applied machine learning for industrial analytics, reinforcement learning for industrial process controls, model predictive control, and advanced optimization.

Mekapati Srinivas is working as Principal Scientist at ABB Industrial Automation Technology Centre Bangalore and his areas of interests are process modelling, simulation and optimization.

FACE RECOGNITION USING PCA INTEGRATED WITH DELAUNAY TRIANGULATION

Kavan Adeshara and Vinayak Elangovan

Division of Science and Engineering, Penn State Abington, PA, USA

ABSTRACT

Face Recognition is most used for biometric user authentication that identifies a user based on his or her facial features. The system is in high demand, as it is used by many businesses and employed in many devices such as smartphones and surveillance cameras. However, one frequent problem that is still observed in this user-verification method is its accuracy rate. Numerous approaches and algorithms have been experimented to improve the stated flaw of the system. This research develops one such algorithm that utilizes a combination of two different approaches. Using the concepts from Linear Algebra and computational geometry, the research examines the integration of Principal Component Analysis with Delaunay Triangulation; the method triangulates a set of face landmark points and obtains eigenfaces of the provided images. It compares the algorithm with traditional PCA and discusses the inclusion of different face landmark points to deliver an effective recognition rate.

KEYWORDS

Delaunay Triangulation, PCA, Face Recognition

1. INTRODUCTION

1.1. Face Recognition and Principal Component Analysis

Face recognition has always been a topic of keen interest for computer vision researchers. It is extensively studied to advance the system's efficiency and minimize errors. Notably, PCA is one of the prominent approaches applied by the researchers to develop user-authentication system; it allows a user to train programs in distinguishing individual faces by providing a set of training and testing data. Using mathematical functions such as Singular Value Decomposition (SVD), the approach obtains eigenvalues and eigenvectors to project an image onto the eigenspace. The projected images, called Eigenfaces, helps to determine the Euclidean distances between the testing and training images, recognizing the training image with the least distance.

According to studies [1] and [2], employing PCA generates fewer errors when comprehensive training data is provided, such as variations in face illumination, expressions, and angles. However, when insufficient data are trained on, this approach fails to retain an accurate recognition rate. In most applications, all imagery data are set to be in a grayscale format with a preferred dimension. Colored images and images with varying dimensions can often be tedious to process leading to increase in computation time.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 321-331, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101424

1.2. Alternate Variants of Principal Component Analysis

To mitigate the limitation of traditional PCA based face recognition, several alternative versions of PCA have been proposed in the past decade. For instance, Article [3] proposes a modular PCA algorithm that divides an image into smaller sub-images which are least affected by variable changes such as a change in pose and illumination. The algorithm implements PCA on these 2 unaltered sub-images. Using the Yale Face, ORL, and UMIST datasets [4], the study concludes an improved efficacy of their proposed algorithm, unaffected by face pose, lighting, and face tilt.

Article [5] examines the implementation of a composite Kernel PCA algorithm which successfully deals with a large training database by outlining the data into high dimension vector space with non-linear transformation. Using the ORL and FERET [4] face database, the algorithm outputs better processing and recognition rates of above 90% for a large sample size but falls short on effective time consumption.

1.3. Other Notable Techniques

Besides PCA, different unique techniques have been executed for the development of a robust face recognition system. In [6], the implementation of Delaunay Triangulation for face verification is discussed. The approach sets the image in a triangular plane where the triangulation would distribute the entirety of the face into normalized triangles. A squared difference between the training and testing image's triangular area is calculated to determine the closest recognized image in the database. The article experiments the technique with 15 frontal, 256 x 256 images, observing an improved recognition rate with illumination and pose change.

Article [7] proposes a novel technique that uses a Radial basis function neural network to acquire the centers of hidden neuron layers for face recognition. From the experimental results using Yale Face, ORL, AR, and LF dataset [4], it shows a clear advantage of combining the firefly algorithm with the neural network that significantly optimizes the feature selection and speeds up the convergence rate. Article [8] proposes an algorithm that utilizes a combination of three different algorithms: Wavelet Transformation, Local Linear Embedding, and Support Vector Machines. The proposed algorithm breaks the face image into four components using wavelet transformation, then uses local linear embedding to analyze the key features of the four components, after which a weighted fusion is determined to perform face recognition. Lastly, SVM will be utilized to train the eigenvectors of the face data and the face classifier class. The algorithm is tested on three different image dataset and the algorithm yields an improved accuracy rate among all the mentioned datasets. However, one key area of improvement needed is to determine a reasonable weight among different face image weights. In the article [9], Gabor filter is used to obtain Gabor amplitude of face images after which Uniform Local Binary Histogram is obtained. Then, a dictionary is implemented using Fisherion criterion and an image is classified using spare representation coding. Using the Yale B and AR image datasets, the amalgamtions of all these algorithms yielded an improved recognition rate in variable image environments such as change in lighting, illumination etc.

1.4. PCA integrated with different techniques

Some studies combine PCA with different novel techniques to achieve a better computational and accuracy rate. Namely, researchers at [10] and [11] integrates PCA with Delaunay Triangulation, Fisherface algorithm, and Convolutional Neural Network respectively to provide increased accuracy with minimal time needed for computation.

In [10], a face model is derived by implementing a synthesis of PCA, Delaunay Triangulation, and Fisher face algorithm. It procures the shape and the texture of the face by acquiring key information from both the triangulation of face landmarks and the Fisher faces using PCA. It applies a collection of 22 points for face landmarks. Colored images from the AR database are used for testing the proposed algorithm. It yields an average of 95% and above when a large sample is trained for biometric verification.

In [11], a combination of Color 2D-PCA with Convolutional Neural Network (CNN) is developed which integrates feature extraction of CNN and 2D-PCA into one decision-level fusion. The algorithm uses the CNN model for depth features, and Color 2D-PCA for detailed image color and spatial information. The experiments, conducted using LFW and FRGC database, show a reduced training time and increased accuracy of above 90% for various image noise levels.

1.5. Integration of PCA with Delaunay Triangulation

This paper investigates the combination of PCA with Delaunay Triangulation to accomplish an enhancement in the face recognition system. The amalgamation of the two approaches demonstrates a slight improvement in the recognition rate when a certain number of landmarks are employed.

The program acquires a set of training images and converts the given images into a triangular mesh of key face landmark points while also projecting all the images onto the eigenspace. The information is stored in the database by conjoining both the Euclidean distances and the average relative area of the following images, weighing both the values equally. The research draws a comparison between the traditional PCA and three different variations of landmarks in the investigated approach to see which approach and variation produces an improved recognition rate.

The following sections are organized as follows: Delaunay triangulation implementation, Principal Component Analysis, integration of PCA with Delaunay triangulation, experiment and results, and conclusion.

2. DELAUNAY TRIANGULATION IMPLEMENTATION

2.1. Introduction and Application

In computational geometry, Delaunay Triangulation refers to the triangulation of a set of distinct points such that no point in the set is inside the circumcircle of any triangle. It maximizes the minimum of angles of the triangles. The triangulation does not occur when the given points are in the same line. The condition that should be met when performing the triangulation, often referred to as the ‘Delaunay condition’, is that the circumcircle of any triangle should have empty interiors.

Figure 1 describes the steps for conducting Delaunay triangulation where a set of point is first plotted. A triangle is drawn using any of the three points, and a circumcircle of the triangle is drawn to check the Delaunay condition. After the condition is satisfied, a triangle is drawn including the leftover point. Once Delaunay condition is satisfied for the second triangle, this mathematical function provides a final triangulation of four points.

In computer vision, Delaunay Triangulation generates a triangular mesh of key face landmark points such as the eyes, nose, mouth, and the shape of the face. Its application underlies within face swapping, face effects seen in social media app filters, and emotion detection [12] based on the triangulation of different expressions. For face recognition, the approach can provide key information about the change in average relative area of the triangulation when different expressions are posed.

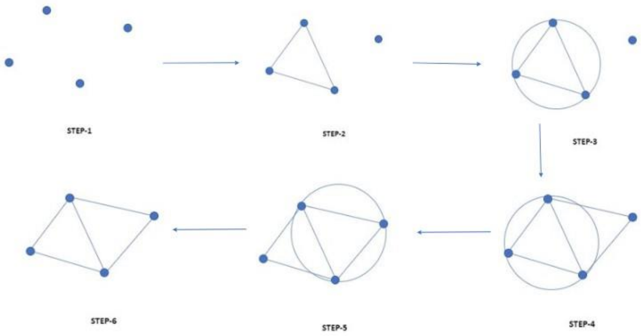


Figure 1: Graphical Representation of the concept

2.2. Developed Step-by-Step Implementation of Delaunay Triangulation

- Establish a training dataset for Delaunay.
- Convert all the images in the training dataset into vector to get pixel value information.
- Divide all the image vectors by 255 to simplify the mathematical calculations.
- Find face landmarks from the vectors. There are various techniques to find face landmark points. Some techniques [13] use neural network that train a classifier by feeding numerous images with manual hand drawn face landmarks on numerous faces whereas some techniques find face landmarks by finding a face that minimizes the deviation with its mean shape [14].
-

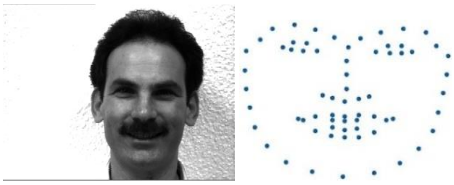


Figure 2: Sample image illustrated with 68 face landmarks [15]

- Perform Delaunay Triangulation on the face landmark points. A plethora of different algorithms exists for the triangulation. In this case, the Sweep hull algorithm is used where a sweep-hull is created in an ordered manner by looping over a set of two-dimensional points, connecting the triangles in the visible area of the convex hull.

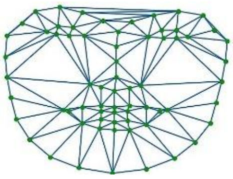


Figure 3: Delaunay Triangulation of Figure 2

- Obtain the set of vertices of all the triangles in the triangulation. This can be done by obtaining the simplices of the triangulation.
- Find edge lengths of all the triangles. The equation used for finding an edge length is as follows:

$$L = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

where L is the edge length

- Obtain areas of all the triangles from the previously calculated edge lengths.

$$S = \frac{L_1 + L_2 + L_3}{2}$$

$$A = \sqrt{S(S - L_1)(S - L_2)(S - L_3)}$$

where S is the semi perimeter, and A is the area

- Calculate the relative area of each triangular area.

$$RA = \frac{A_i}{A_{max}}$$

where RA is the relative area, A_i is each triangular area, and A_{max} is the largest area in the triangulation Ø Obtain the average relative area of all the relative areas

$$RA_{avg} = \frac{RA_1 + RA_2 + RA_3 + \cdots + RA_e}{N}$$

where N is the total number of average relative areas in the triangulation

- Perform Delaunay Triangulation on a test image and acquire its average relative area. Save the average relative area of all the image in the database for error estimation and face comparison.

3. PRINCIPAL COMPONENT ANALYSIS

3.1. Introduction and Application

Principal Component Analysis (PCA) is a dimensionality reduction method that reduces the dimensionality of a large dataset by transforming the given dataset into a smaller one that contains enough information to represent the large dataset. It is used in many fields such as medical and technological field to analyze a large chunk of data by extracting only the vital information from the data and discounting the surplus data.

In face recognition, the method is used to reduce the dimensions of the pixel value of the image in a manner that, after the reduction, the compressed image should have ample information to represent the majority of the unique features of the image.

3.2. Step by Step Implementation of PCA

The following section elaborates the steps followed in developing PCA.

- Establish a training dataset to perform Principal Component Analysis.

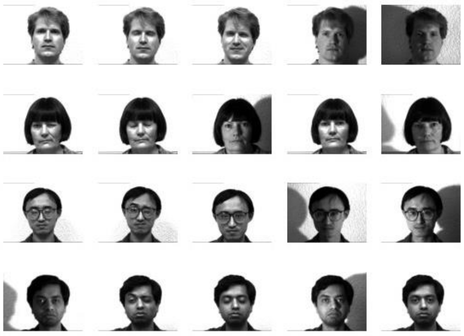


Figure 4: Images taken from Yale Face Dataset

- Convert all the images into vectors.
- Find the mean of all the image vectors.

$M = \frac{\sum_{i=1}^{n} x_i}{n}$ where M is the mean image, x isthe individual image, and n is the total image

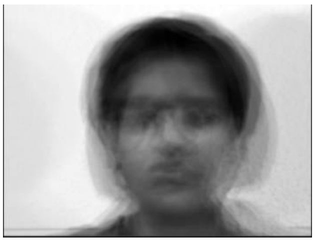


Figure 5: Mean face of all images from Figure 4

- Subtract the mean image vector from each individual image vector in the dataset.

$$I = M - x_i$$

where I is the subtracted image vector

- Perform Singular Value Decomposition to calculate the eigen vectors and eigen values.

$$SVD = U\sum V^T$$

where U and $V^T$are orthogonal matrix of eigenvector, and $\sum$ is the diagonal matrix of eigenvalues

- For k best eigen values, reconstruct the image vectors and project them onto the eigen space. In this research, k is set to 25 for the projection of image vectors onto the eigen space.

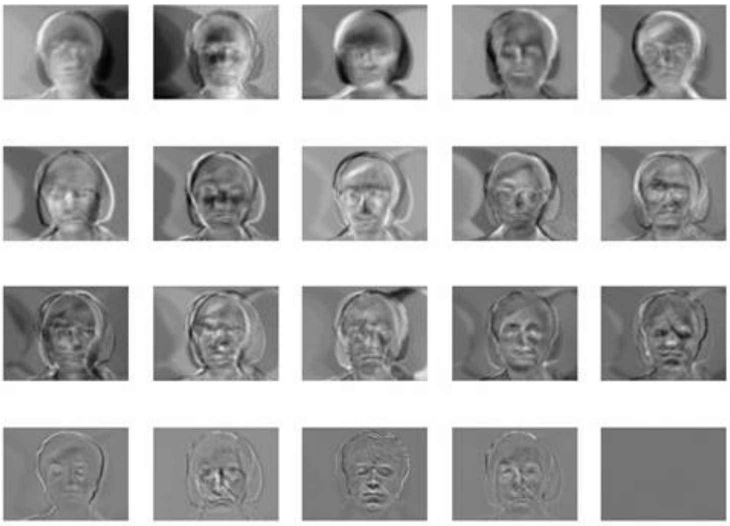


Figure 6: Eigenfaces of all images from Figure 4

- Store the information of the eigenfaces in the training database.
- Establish a testing dataset.
- Perform the same steps by converting all the testing images into image vector.
- Subtract the testing image vectors from the training mean image vector.
- Calculate the eigen values and eigenvectors of the subtracted test image vectors.

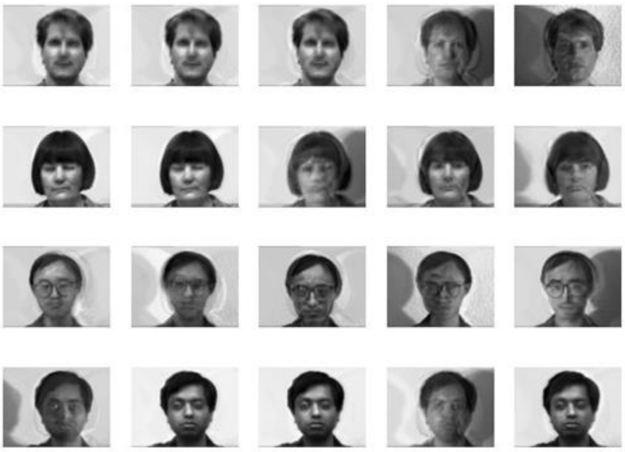


Figure 7: Reconstructed images for k = 9

- Project the images onto the eigenspace to find the closest image associated with the testing image 7
- Calculate the Euclidean distances of the testing image in the eigenspace with all training images and recognize the image with the least distance

4. Integration of PCA With Delaunay Triangulation

The following section discusses the step by step integration of PCA with Delaunay Triangulation.

- Perform Delaunay Triangulation on a set of test images and acquire their average relative areas.
- Calculate the difference between test and training images' average relative area.

$$D = \sqrt{(Tt_{avg} - Tn_{avg})^2}$$

where D is the positive difference, and $\mathrm{Tt_{avg}}$ and $\mathrm{Tn_{avg}}$ ar

e the average relative areas for test and train image

- Add the Euclidian Distance from PCA and difference from Delaunay Triangulation to acquire the image with least result value.

$$RV = ED + \frac{D}{0.001}$$

where RV is the resultant value, ED is the Euclidian distance and D is the difference from the triangulation

5. Experiments and Results

Three experiments were conducted for testing the traditional PCA with DT (Delaunay Triangulation) integrated PCA. For each experiment, different amount of training and testing images are used. Furthermore, different number of landmark points such as 68, 79 and 194 landmarks as shown in Figure 8, are also used for the triangulation to further test the efficiency of each landmark combination to determine which landmarks work best with the discussed integration.

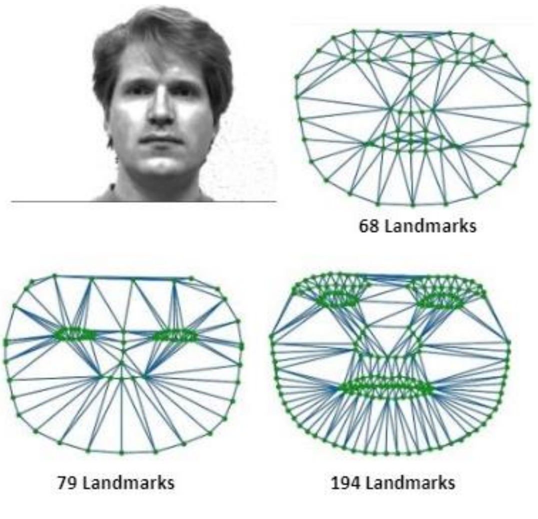


Figure 8: Triangulation of different landmarks

5.1. Images used for the Experiment

For all experiments, 135 frontal, 320 x 243 images from the Yale Face dataset are used. Each image is in grayscale format. The dataset contains 9 total variations per individual and the variations are as follows – sleepy, happy, sad, wink, glasses, surprised, left-light, right-light, and

8 normal. The different conditions in the dataset help assess the accuracy of the two approaches being analyzed.

5.2. Variation in Landmarks

The paper tests three different variation in landmarks: 68, 79, and 194 landmarks [Figure 8]. The 68 landmark contains the basic shape of the eyes, nose, and the mouth. The 79 landmark contains only eyes, nose and overall shape of the face. This landmark point help in differentiating individuals if they wear a mask, but is less helpful when comparing faces without mask because of discounting width of the mouth and the location of eyebrows. The 194 landmarks contain detailed landmark points of the eyes, nose, eyebrows, and lips. This can help differentiate other individuals from the depth in eyebrows, distance of the eyes from the nose and such.

5.3. Experiment 1

In experiment 1, 105 images are trained. 7 variant images of each 15 individuals were considered. The remaining images are used for testing. From the results, the recognition rates are improved for DT integrated PCA using 68, 79, and 194 face landmarks.

5.4. Experiment 2

In experiment 2, 75 images are trained on, from which 5 images are variants. The rest of the images are utilized as test images. The recognition rate for DT-PCA 68-L, DT-PCA 194-L is higher comparatively higher than traditional PCA. DT-PCA 194-L, especially, yields a significant improvement out of all the three different variation of landmarks in the integration.

5.5. Experiment 3

In experiment 3, 45 images are trained on, from which 3 image are variants. 90 images are tested. It can be observed that the accuracy of DT-PCA 79-L is less than the traditional PCA. This is because the landmark does not account for expression change and individuals that closely resemble each other in terms of geometrical configuration. This shows that the recognition rate of this landmark combination tends to decrease as less training dataset is provided and therefore, the combination is inefficient for use in the current integration.

Table 1: Percent accuracy for traditional PCA and variant of PCA with DT

	Traditional PCA	**PCA with Delaunay Triangulation**		
		68-L	**79-L**	**194-L**
Train – 105 **Test – 30**	86.7 %	93.3 %	90.0 %	95.6 %
Train – 75 **Test – 60**	85.0 %	88.3 %	86.7 %	91.6 %
Train – 45 **Test – 90**	82.2 %	87.8 %	81.1 %	90.0 %

On the other hand, DT-PCA 194-L has retained a better percent accuracy compared to the traditional PCA and all the combination of other landmarks. Thus, from the above experiments, the DT integrated PCA efficiently works with 194 face landmark points as it shows considerable improvement in the accuracy rate compared to the traditional PCA.

6. CONCLUSIONS

Principal Component Analysis can help build a robust face recognition system if it is modified or used with another approach. The algorithm has the level of versatility that enables it to be used with different techniques. This research shows one such example of the combination as it discusses the slight improvement in recognition rates when a DT integrated PCA is utilized. Moreover, the results can be improved further if appropriate face landmarks are used in the integration as seen in the results above. It can be used in an attendance marking system which could clock in and clock out employees or mark a student present or absent for a class. For future work, the face landmark detection could be fine-tuned to detect the very intricate edges of the face to give a detailed statistical view of the face and to acquire a more precise triangulated data. However, there is a trade-off in using face recognition of any approach because factors like image clarity and certain face angle are always going to remain an issue. The system cannot efficiently recognize if there is an overhead or side view of the face and there is always a factor of a person undergoing various sort of surgery that could completely change the look of the face., causing the system to not authenticate the face.

ACKNOWLEDGMENTS

This work was funded by the Multi Campus Research Experience for Undergraduates (MCREU) program, Penn State University.

REFERENCES

[1] Pentland, Moghaddam and Starner, "View-based and modular eigenspaces for face recognition," 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 1994, pp. 84-91, doi: 10.1109/CVPR.1994.323814.

[2] A. L. Ramadhani, P. Musa and E. P. Wibowo, "Human face recognition application using pca and eigenface approach," 2017 Second International Conference on Informatics and Computing (ICIC), Jayapura, 2017, pp. 1-5, doi: 10.1109/IAC.2017.8280652.

[3] J. F. Pereira, R. M. Barreto, G. D. C. Cavalcanti and I. R. Tsang, "A robust feature extraction algorithm based on class-Modular Image Principal Component Analysis for face verification," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, 2011, pp. 1469-1472, doi: 10.1109/ICASSP.2011.5946770.

[4] Web.mit.edu. n.d. Face Database Info. available at: http://web.mit.edu/emeyers/www/face_databases.html#umist Accessed on 17 July 2020.

[5] C. Liu, T. Zhang, D. Ding and C. Lv, "Design and application of Compound Kernel-PCA algorithm in face recognition," 2016 35th Chinese Control Conference (CCC), Chengdu, 2016, pp. 4122-4126, doi: 10.1109/ChiCC.2016.7553997.

[6] Chiang, J.Y., & Wang, R.C. (2011). The Application of Delaunay Triangulation to Face Recognition.

[7] Agarwal, V., Bhanot, S. Radial basis function neural network-based face recognition using firefly algorithm. Neural Comput & Applic 30, 2643–2660 (2018). https://doi.org/10.1007/s00521-017-2874-2.

[8] Zhou, Y., Wang, Y. & Wang, X. Face recognition algorithm based on wavelet transform and local linearembedding. Cluster Comput 22, 1529–1540 (2019). https://doi.org/10.1007/s10586- 018-2157-4

[9] Zhenyu Lu, Linghua Zhang, Face recognition algorithm based on discriminative dictionary learning and sparse representation,Neurocomputing,Volume 174, Part B,2016,Pages 749- 755,ISSN 0925-2312,https://doi.org/10.1016/j.neucom.2015.09.091. (http://www.sciencedirect.com/science/article/pii/S0925231215014241) 10

[10] Fagertun, Jens & Gomez, David & Ersbøll, Bjarne & Larsen, Rasmus. (2005). A face recognition algorithm based on multiple individual discriminative models

[11] J. Li et al, "Robust Face Recognition Using the Deep C2D-CNN Model Based on DecisionLevel Fusion," Sensors, vol. 18, (7), pp. 2080, 2018. available: http://ezaccess.libraries.psu.edu/login?url=https://search-proquestcom.ezaccess.libraries.psu.edu/docview/2108746318?accountid=13158. DOI: http://dx.doi.org.ezaccess.libraries.psu.edu/10.3390/s18072080.

[12] Hsu, Fu-Song & Lin, Wei-Yang & Tsai, Tzu-Wei. (2014). Facial expression recognition using bag of distances. Multimedia Tools and Applications. 73. 10.1007/s11042-013-1616-4.

[13] Wenyan Wu, Xingzhe Wu, Yici Cai, Qiang Zhou, Deep coupling neural network for robust facial landmark detection, Computers & Graphics, Volume 82, 2019, Pages 286-294,ISSN 0097- 8493, https://doi.org/10.1016/j.cag.2019.05.031.

[14] P. Nair and A. Cavallaro, "3-D Face Detection, Landmark Localization, and Registration Using a Point Distribution Model," in IEEE Transactions on Multimedia, vol. 11, no. 4, pp. 611-623, June 2009, doi: 10.1109/TMM.2009.2017629.

[15] P. Belhumeur, J. Hespanha, D. Kriegman, ÒEigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection,Ó IEEE Transactions on Pattern Analysis and Machine Intelligence, July 1997, pp. 711-720.

AUTHORS

Kavan Adeshara is a current sophomore at Penn State University, majoring in Computer Science. His research interests include artificial intelligence, computer vision and software development. He has participated in multiple projects including data analysis of global temperature, face recognition, college major recommendation applications etc.

Dr. Vinayak Elangovan is an Assistant Professor of Computer Science at Penn State Abington. His research interest includes computer vision, machine vision, multisensor data fusion and activity sequence analysis with keen interest in software applications development and database management. He has worked on number of funded projects related to Department of Defense and Department of Homeland Security applications.

OBSTACLE AVOIDANCE AND PATH FINDING FOR MOBILE ROBOT NAVIGATION

Poojith Kotikalapudi and Vinayak Elangovan

Division of Science and Engineering, Penn State Abington, PA, USA

ABSTRACT

This paper investigates different methods to detect obstacles ahead of a robot using a camera in the robot, an aerial camera, and an ultrasound sensor. We also explored various efficient path finding methods for the robot to navigate to the target source. Single and multi-iteration angle-based navigation algorithms were developed. The theta-based path finding algorithms were compared with the Dijkstra's Algorithm and their performance were analyzed.

KEYWORDS

Image Processing, Path Finding, Obstacle Avoidance, Machine Learning, Robot Navigation.

1. INTRODUCTION

After a disaster like a hurricane or a tornado has hit and left the area, often the roads are blocked by debris and people can be trapped in the rubble. Additionally, the areas are often littered with dangers that make it unsafe for humans to navigate. Dangers can include live wires in water or unstable foundations. An ideal approach is to use a robot that can autonomously navigate to a target location. A cohesive program is needed with efficient image processing, obstacle avoidance, and path finding algorithms to create a path between the robot and a target.

An autonomous mobile robot is a machine that operates in a partially unknown and unpredictable environment. Mostly, mobile robots are used for surveillance, inspection, and transportation tasks. The robot must operate safely, i.e. it must stay away from hazards such as obstacles or operating conditions dangerous to the robot itself, and it must pose no risk to humans in the vicinity of the robot. For any kind of mobile robots, navigation is a fundamental capability. One of the most challenging tasks for the mobile robot is to understand the information provided through various sensors, which will guide the robot in the environment and reach the destination by avoiding obstacles in the environment. One important task of a mobile robot intelligence program is environment perception and navigation. Environmental perception enables the robot to be aware of its environment.

There are two kinds of navigational environments for the robot to navigate. Completely known Environments and partially known environments. In the completely known environment, the robot knows complete information about all objects in the robot environment before navigation starts. The status of an obstacle is said to be static when its position or orientation are relative to a known and fixed origin that does not change with time. The status of the obstacles in the environment change with time may be in position or orientation or both according to its origin. In this paper, static obstacles are considered.

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 333-344, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101425

Robot path planning is the determination of how the robot navigates in the environment to reach its target. The path planning involves computing the collision-free path between two locations. Path planning takes a significant part of the computation time for many simulations, mainly in high time-dependent environments where most of the agents are moving. Path planning is typically performed on one agent at a time.

In general, there are two path planning techniques namely, global and local path planning. In global path planning the complete information about the environment is known, and obstacles should be static. In this approach, the algorithm generates a complete path from the start point to the destination point before the robot starts its motion. Local path planning is done while the robot is moving. The robot needs to change the path if there is a change in the environment. If there are no obstacles in the environment the path would be the straight line. The robot will head straight until it detects an obstacle. If it detects an obstacle, the robot will use path planning algorithm to find a feasible path to reach the target. In our research, the latter approach is employed.

2. RELATED WORK

The following section highlights some of the relevant search work carried out in robot navigation, path finding and obstacle avoidance.

2.1. Driver Assistance System based on Raspberry Pi [1]

This research paper describes how to navigate a robot when there are boundary lines present that can guide the robot. An edge detection is performed, and Hough line transform is applied to detect the lines on a road. These lines are then used to create a frame of reference for the robot to navigate.

2.2. Intelligent Survelliance Robot with Obstacle Avoidance Capabilities Using Neural Networks [2]

This paper explores how neural networks can be used to detect obstacles ahead and then avoid them appropriately. They used a combination of ultrasound sensors and camera. This method has shown great results to avoid obstacles to the right, left, and in the front of the robot.

2.3. A Comprehensive Study on PathFinding Techniques for Robotics and Video Games [3]

Various techniques of path finding algorithms that are being used within the realm of robotics and video games are investigated. This paper reports that using a traditional grid technique, it takes up considerable memory compared to hierarchical techniques that give more accurate representation near obstacles and less detail to large open fields where a lot of processing power is not needed.

2.4. Robot Navigation Control Based on Monocular Images: An Image Processing Algorithm for Obstacle Avoidance Decisions [4]

This paper highlights the issues of the dynamic environment in robot navigation. The paper details a 2-step approach in which they first use image segmentation to separate different parts of the image. A Balanced Histogram Thresholding was developed to find an optimal thresholding value that divides the image into a foreground and background. Edge detection is used to partition an image on abrupt changes in intensity between pixels. They also discuss how the Hough line transform method is used to highlight boundary lines.

2.5. Online Aerial Terrain Mapping for Ground Robot Navigation [5]

This research combines UAV (Unmanned aerial vehicle) and UGV (Unmanned ground vehicle) in a coherent system with user access from a ground station. The robots use GPS to track positions relative to each other. Processing is split among 4 computers, 2 being on the drone and the robot, and the other two as the base stations. One computer focuses on telemetry and the other as the mission planner and directly controls the drone. The program initiates by receiving images from the drone, creates an orthomoasic map which is stitched from all the images. The UAV creates a terrain map from which a path finding algorithm is used to create a path the UGV follows. The UGV does a lidar scan and creates an obstacle map and then send commands to motors to turn the robot wheels.

2.6. Shortest Path Finding and Tracking System Based on Dijkstra's Algorithm for Mobile Robot [6]

This research paper explores the Dijkstra’s Algorithm specifically for mobile robot navigation. The research used a car-like robot with front-wheel steering. There were three different classifications of the shortest path algorithm which are single-source shortest path algorithm, single destination shortest path algorithm, and all-pairs shortest path algorithm.

2.7. Deep Learning using Rectified Linear Units (ReLU) [7]

This paper explored the performance of Rectified Linear Units with neural networks. They stated that certain neurons do not get activated properly and eventually die. This will affect the training of the algorithm and can often impede it. However, even with that drawback ReLU generally is seen as an effective method to prevent returning a negative value.

2.8. Image-Based Segmentation of Indoor Corridor Floors for a Mobile Robot [8]

A novel approach was proposed by combining three different parts of the image and using those visual cues to isolate the floor. The method proposed is extremely effective with the algorithm detecting 90% of the wall floor boundary. This algorithm is also optimized well for real time mobile robot navigation. A limitation has been stated if the floor is highly textured it can then become very difficult to detect. This can be a potential problem with carpets being potentially hard to detect.

2.9. Monocular Vision for Mobile Robot Localization and Autonomous Navigation [9]

This paper explored a method where a singular camera is used to navigate outdoor environments. This approach was compared to RTK GP. A 3d map was created to help the robot navigate. The paper states there has been difficulty updating the 3d map as the localization algorithm is not capable of a monitoring dynamic environment efficiently.

2.10. Obstacle Avoidance of Mobile Robot Based on HyperOmni Vision [10]

This research paper combines two approaches which are IDWA (Improved Dynamic Window Approach) and artificial potential field to avoid obstacles. The robot uses a concept of attraction and repulsion that would create a repulsive force when near an obstacle. This would cause the robot to move out of the way of the obstacle. The robot also establishes a dynamic window and predicts trajectories using the OmniHyper camera.

One approach for robot navigations is to employ an aerial camera to capture images over the targeted area and create a 2 D map to trace a path between the robot the target. A raspberry pi-controlled robot with a camera can be programmed using image processing techniques to identify obstacles and traverse around them. This research paper follows the above-mentioned approach for autonomous robot navigation. Various algorithms are developed in three main areas namely: robot navigation, path finding, and obstacle avoidance. Robot navigation focuses on interpreting the signals as well as processing the information that the robot will receive. The robot is programmed to detect the lines on a road and then calculates a path to navigate to the target. This was done using Hough line transform method. The next portion of the research paper highlights a different path finding techniques that can be used. A neural network model is developed and trained using YOLO object identification system to detect objects encountered in the navigation path. Data from an ultrasound sensor and imagery data from the robot camera are processed for obstacle detection. Combining the techniques of time of flight principle and ray tracing from the camera improves the efficacy of obstacle avoidance. These methods are built on a framework that uses a client-server protocol that allows communication between the robot and a workstation. The remaining part of this paper is organized as such: robot navigation, path finding, obstacle avoidance, a combination of systems, results, and conclusions followed by references.

3. Robot Navigation

In order for the robot to navigate the map and get to the target, a 10-foot by 10-foot square testbed was created with lines the robot needed to follow. These lines lead the robot to the target. A lane keep assist program was developed to run on the workstation and send signals to the raspberry pi robot.

3.1. Lane Keep Assist Algorithim

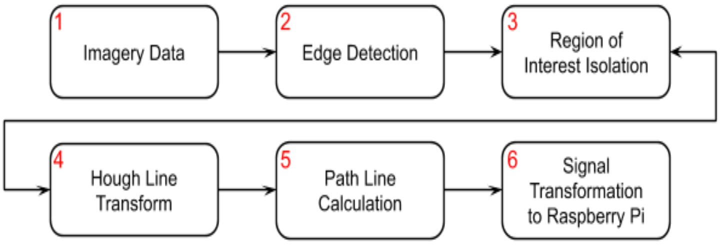


Figure 1: Lane Keep Assist Process

Figure 1 shows the step by step process in lane keep assist program. In step 1, the program received imagery data or a live stream video from the raspberry pi. After reading the image in step 2, the canny edge detection algorithm is applied to highlight the contours of the lines of the road. In step 3, a Region of Interest (RoI) is identified based on where the lines from the floor are in the image, and it was highlighted and isolated. This is done to reduce the processing time and is important especially for the Hough line transform method. RoI is altered based on changes in the environment. Factors to consider while calibrating the RoI is the width of the road, the height of the camera relative to the ground, and the focal length of the camera. In step 4, the Hough line transform method is used to identify the lines on the image. This returned the endpoints for each line on the floor to guide the robot. In step 5, the distance between the endpoints was calculated to get the midline of the road. Using the endpoints of the midline, a theta value was calculated for the robot to make necessary turns following the path. Figure 2 below shows the detection of the lane and path for the robot to navigate.

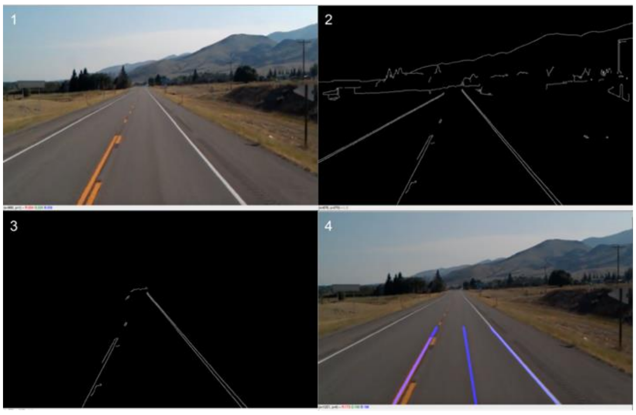


Figure 2: Lane Keep Assist Detections

3.2. Techniques of Angle Based Navigation

Multi iteration-based theta value:

In the first step, data from the previous iteration of the endpoints are saved; this has been done by using a FIFO data structure or a queue. In the second step, after the program passed the data values from the previous two iterations, the top endpoint and the bottom endpoint of the previous iteration and the top endpoint of the current iteration are used to calculate the theta value as shown in Figure 3. Also, note that the top endpoint has an x and y value, however, the program takes the value from the previous iteration for the top endpoint and writes it over the current iteration's y value for the top endpoint. This is done to reduce computing power.

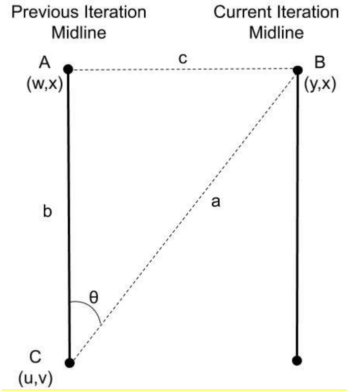


Figure 3: Diagram for Multi-Iteration theta values

The third step calculates the theta value as shown in equation-1 which combines the distance values and the law of cosines.

$$EQ\ 1: \theta = \cos^{-1}(\frac{a^2 + b^2 - c^2}{2ab}) * \frac{180}{2\pi}$$

$$EQ\ 2: a = \sqrt{(y-u)^2 + (x-v)^2}$$
$$EQ\ 3: b = \sqrt{(w-u)^2 + (x-v)^2}$$
$$EQ\ 4: c = w - y$$

In the fourth step, the slope of the line is calculated. If the slope is negative the theta value is negative. If the slope is positive, then the theta value is positive. Finally, the calculated theta value is sent back to the raspberry pi.

Single iteration-based theta value:

The program takes the endpoint values for the midline and creates astraight line from the x value of the bottom end point.

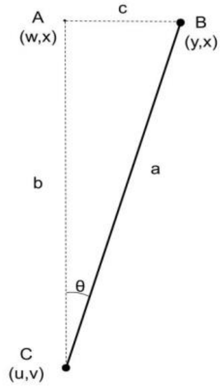


Figure 4: Diagram for Single-Iteration theta values

Equation 1 is used once again however different equations are used to find the variables c and b. The variable a however uses the same equation 2.

$$EQ\ 5: c = w - y$$
$$EQ\ 6: b = w - u$$

The third step is to find the slope needs to be calculated. If the slope is negative, then the theta value is negative. If the slope is positive the theta value is positive. The calculated theta value is sent to the raspberry pi to control the robot.

3.3. Advantages and Disadvantages of each technique

The advantages of the multi iteration-based theta value are that it gives the ability for the robot to correct itself when one or both of the lines are not detectable as it can use the previous values to gauge how far the robot has moved out of the line. The advantage of the single iteration-based theta value is that it is much more efficient and faster at returning a theta, something that is important when a robot is moving and needs to make quick decisions. However, the system was not accurate especially when the camera loses sight of one of the lines on the road. A future test is to use a combination of both ideas to retain the accuracy of the multi iteration-based theta value with the simplicity of the single iteration-based theta value. A proposed system can store previous iterations of midline while processing theta values in real-time by using the single iteration theta method. The two systems can verify each other with a margin for error.

3.4. Steering the robot with theta values

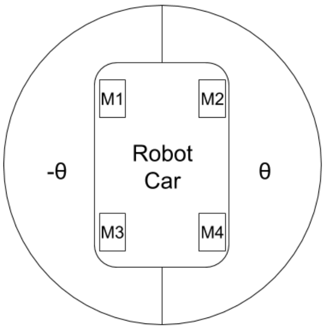


Figure 5: Diagram of the robot car with the right being all possible theta values and the left being all the negative theta values

Before a mathematical model is developed, an experiment was run to find out how much the robot turns in one second. To find the theta value the robot was placed on a line of tape in the testbed. The robot turned for 1 second. and the tape was placed where the robot turned. A protractor was used to measure by what degree the robot has turned. This was done 5 times to find the average of theta values. This returned the result that the robot turned 23 degrees every second. Therefore, a relationship can be stated that for every degree the robot needs to turn for 0.0435 seconds.

$$EQ\ 6: t = \frac{1}{23}|\theta|$$

An absolute value was used to prevent returning a negative time. After finding the seconds needed to turn, the robot then needs to find which way it will turn. A simple if-else statement is implemented where if the theta value was negative it will turn left but if the theta value was positive it will turn right.

4. PATH FINDING

Path finding is the method of formulating a path from one spot on the map to another spot on the map. Path finding will enable the program to determine the best path the robot should take to reach the target location. For the pathfinding algorithm originally, a camera was attached to moving platform on top of the area of interest. This moving platform could move in the x, y, and z-direction simulating the changing conditions of an aerial vehicle. Instead, a smartphone camera took a picture which is then sent by email to the workstation. Then the image is manually uploaded to the program. Numerous methods were used and tested keeping in mind of processing power and speed of the program.

4.1. Path finding algorithim using yolo object identification

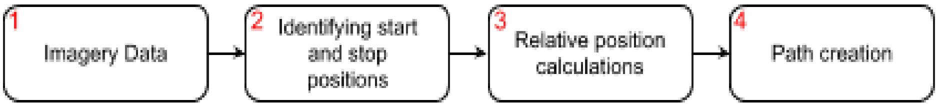

1
Imagery Data
2
Identifying start and stop positions
3
Relative position calculations
4
Path creation

Figure 6: Block Diagram for theta-based path finding algorithm

In step 1, image data from the smartphone is manually inputted into the program. In step 2, to identify the start and stop positions the YOLO object identification was used. To identify the two objects a training program was run to train the machine learning and adjust the weights. The input nodes for the neural network were the pixels of the image. The algorithm learned by changing the weights of the hidden layers to get the desired outcome. Initially, pictures of the robot were taken to be used as the data set for the training algorithm. However, later on, it was found the algorithm was not identifying the robot consistently. The problem was found that there were not enough images of the robot. Instead, a JavaScript program was run that can download images of the raspberry pi. Then using the new dataset, a new set of weights was calculated that can be used to detect the raspberry pi on top of the robot. The same was then done to the target as well. To train the machine learning algorithm, GPU space was rented for free using Google Collab. The algorithm was trained for the maximum amount of time that google provided for free which was 12 hours. In step 3 once the start and end objects were identified the objects needed to be given a coordinate so a path can be formed.

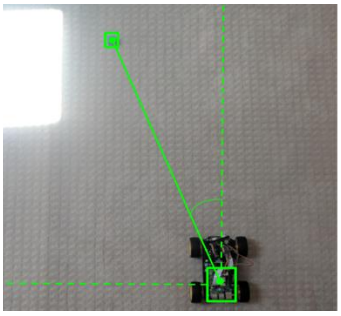


Figure 7: shows the theta angle calculation.

In step 4, once the coordinate values are given the path needs to be created for the robot. This was done by creating a straight line between the robot (start point) and the target (end point). Then using trigonometry, the angle theta was calculated to find how far the robot needs to turn to face the target. Once the robot has faced the target the robot went in a straight line until the robot encountered an obstacle which it will navigate around.

4.2. Advantages and Disadvantages

One of the big advantages of this method is simplicity. This method requires very little processing power and time is and is easy to code as well. The step which requires the most processing is the YOLO object identification. The process takes less time and processing power than alternatives like Dijkstra’s Algorithm. Additionally, with Dijkstra’s Algorithm, the value returned needed to be interpreted differently as it will not return a theta value which meant that an interpreter needed to be made to convert the values to GPIO signals for the robot to understand.

A major disadvantage of the program is its ability to not detect obstacles and create a path around it. This meant that the robot needed to perform the obstacle avoidance and there were no further redundancies. Also, if a hole was present in the ground it would be difficult for the robot to detect it compared to an overhead camera that can see it. Dijkstra’s Algorithm with Yolo object identification, however, would be able to navigate around a hole.

5. OBSTACLE AVOIDANCE

5.1. Navigating with a ultrasonic sensor

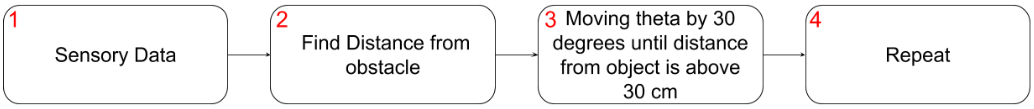

1
Sensory Data
2
Find Distance from obstacle
3 Moving theta by 30 degrees until distance from object is above 30 cm
4
Repeat

Figure 8: Block Diagram for obstacle avoidance program

In step 1, sensory data is received to the robot from the ultrasound sensor in the form of distance in cm from the obstacle. In step 2, the robot receives the distance from the robot and will be inputted into the program. In step 3, if the robot encounters an obstacle, the robot will keep moving forward until the obstacle is less than 30 cm away. The robot will then move 30 degrees to the right and then check the distance again to repeat the process.

If the obstacle is 30 cm away still, it will once again turn right by 30 degrees. If the obstacle is still 30 cm away from the robot after 90 degrees of motion. The robot will then turn left 120 degrees. The robot will continue to do the same to the left as well. This process will be repeated. Once the robot has navigated the obstacle the theta value at which the target is saved, and the robot will once again face the target and move straight.

6. HARDWARE

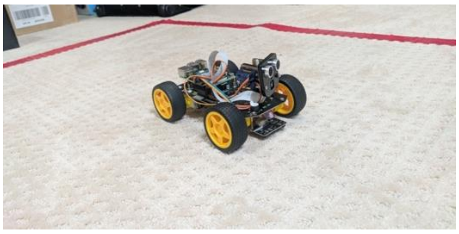


Figure 9: A sample of experimental set up.

The hardware used as an edge device is the raspberry pi 4b+. This particular model was chosen as it is the fastest raspberry pi. This speed helped greatly in making the program run faster. Another reason this was chosen is so the raspberry pi can act as an edge device that will be able to run its own programs. The robot chosen was Freenove 4WD Smart Car Kit for Raspberry Pi. The robot includes LEDs, speaker, lane detector, 4 motors for the wheels, an ultrasound sensor, and a camera. The lane detector was not used as the camera was used instead to detect lanes. A MacBook 15 inch was used as the workstation. The specification of the MacBook is an intel core i9 and 16 GB of ram. The overhead camera used was from a smartphone. To train any neural networks, google colab was used.

7. COMBINATION OF SYSTEMS

The program used yolo object identification to identify the target and the robot. Afterward, the program acquires the coordinate and calculates the theta between the two objects. The information is sent to the raspberry pi.

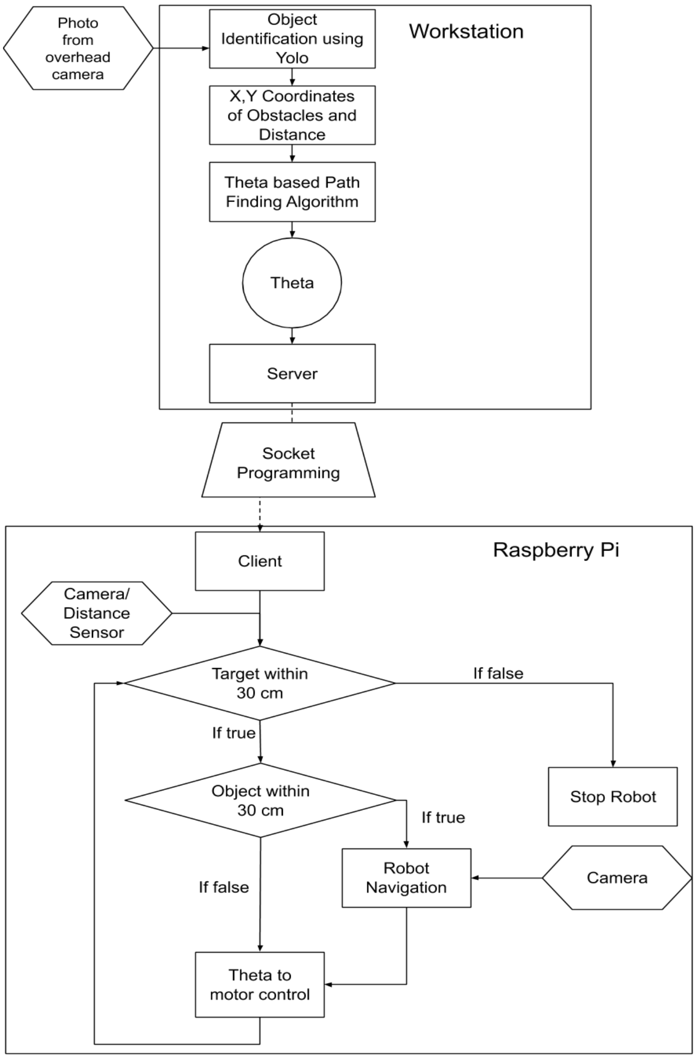


Figure 9: Block Diagram of the entire system

7.1. Socket Programming

The program sends the theta value from a server-class running on the workstation to the client that is running on the raspberry pi. The client first established a connection with the server on a specific port. The client will then listen to any data packages coming from the workstation. Once the client has received a package of data from the server, the program on the raspberry pi will start.

Once the theta value has been received, the robot will turn theta degrees to the left or right towards the target and start heading straight. The robot will check the distance sensor to see if an obstacle is detected within 30 cm of the robot. The camera will then be activated, and a picture is taken and further processed. The program will check for the target using Yolo object identification. If a target has not been detected the program returns a null and will navigate around the robot. The robot will then move towards the target and proceed straight, navigating around the obstacle when needed. Once it reaches the target the robot will first identify it using Yolo object identification. Afterward, the robot will stop and flash its LEDs and make a sound indicating that it has finished and reached the target.

8. Results

Comparisons have been made between Dijkstra's Algorithm and theta-based path finding. Dijkstra's Algorithm on average takes 0.112 seconds while theta-based pathfinding algorithm took on average about $1.249 * 10^{-5}$ seconds. Dijkstra's Algorithm had 108 function calls compared to theta-based path finding had 9 function calls. The program was measured using cProfile which is a library in python. Theta based path finding had a lot less process running which contributed toward the program itself completing much faster.Theta based navigation is good at taking the best of multi computer processing and single computer processing. Since the robot only has to communicate with the workstation once to get the necessary theta value. Once the robot has that information it can navigate through the obstacles. This gives the robot the advantage of being able to utilize the resources of the workstation to find a clear path and also increases the range of the robot while not having to deal with the issue of an unreliable connection to the workstation. However, if there is an uneven surface (example: apit) in the path, the robot would not detect it and would interpret it as an even surface ground. Another disadvantage is that due to the robot performing image processing using the raspberry pi, the reaction time of the robot slows down which can be potentially delay the rescue process in disaster ridden areas.

9. Conclusion

The research paper explored various techniques and how these techniques can be used together to form a coherent system. Lane keep assist and theta-based navigation system was developed and differences between single iteration and multi iteration lane keep assist were explored. It was found that with multi iteration the robot can remember its relative place even if one of the lines is not showing however with a single iteration-based system the program can be executed faster which lead to a faster response from the robot and follows the line more accurately. Path finding was another method investigated. A comparison was made between Dijkstra's Algorithm and a theta-based path finding algorithm where it was found that the theta-based path finding algorithm had a quicker run time which increased responsiveness time.

References

[1] PandharinathPawar, Narayan & Patil, Minakshee. (2014). Driver Assistance System based on Raspberry Pi. International Journal of Computer Applications. 95. 36-39. 10.5120/16682-6794.

[2] Budiharto, Widodo. (2015). Intelligent Surveillance Robot with Obstacle Avoidance Capabilities Using Neural Network. Computational Intelligence and Neuroscience. 2015. 1-5. 10.1155/2015/745823.

[3] Abd Algfoor, Zeyad & Sunar, Mohd Shahrizal & Kolivand, Hoshang. (2015). A Comprehensive Study on Pathfinding Techniques for Robotics and Video Games. International Journal of Computer Games Technology. 2015. 1-11. 10.1155/2015/736138.

[4] Benn, William and Stanislao Lauria. "Robot Navigation Control Based on Monocular Images: An Image Processing Algorithm for Obstacle Avoidance Decisions." (2012).

[5] Peterson, J.; Chaudhry, H.; Abdelatty, K.; Bird, J.; Kochersberger, K. Online Aerial Terrain Mapping for Ground Robot Navigation. *Sensors* 2018, *18*, 630.

[6] WaiWai Maw, WaiPhyo Ei. (2017) Shortest Path Finding and Tracking System Based on Dijkstra's Algorithm for Mobile Robot. International Journal of Science, Engineering and Technology Research (IJSETR) Volume 6, Issue 11, November 2017.

[7] Agarap, Abien Fred. (2018). Deep Learning using Rectified Linear Units (ReLU).

[8] Li, Yinxiao & Birchfield, Stanley. (2010). Image-Based Segmentation of Indoor Corridor Floors for a Mobile Robot. Proceedings of the ... IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE/RSJ International Conference on Intelligent Robots and Systems. 837 - 843. 10.1109/IROS.2010.5652818.

[9] Royer, E., Lhuillier, M., Dhome, M. et al. Monocular Vision for Mobile Robot Localization and Autonomous Navigation. Int J Comput Vision 74, 237–260 (2007).

[10] Shih-An Li, Li-Hsiang Chou, Tsung-Han Chang, Chao-Hsu Yang, and Yu-Cheng Chang, Obstacle Avoidance of Mobile Robot Based on HyperOmni Vision, Sens. Mater., Vol. 31, No. 3, 2019, p. 1021-1036.

AUTHORS

Poojith Kotikalapudi is an Undergraduate student at Penn State Abington. He is planning to major in Computer Science. He has a passion for robotics and coding. His major interests are building and programming robots to address real world problems.

Dr. Vinayak Elangovan is an Assistant Professor of Computer Science at Penn State Abington. His research interest includes computer vision, machine vision, multi-sensor data fusion and activity sequence analysis with keen interest in software applications development and database management.

IMPROVING WORKLOAD ESTIMATION IN INEXPERIENCED TEAMS WITH HYBRID AGILE APPROACH

Fabiano S. Pires, Cícero A. L Pahins and Paulo Fonseca

Sidia R&D Institute, Manaus, AM, Brazil

ABSTRACT

SCRUM framework is an agile technique that is widely used by development teams in order deliver incremental value to customers and dynamically react to project needs. SCRUM framework might be adapted to conform the development team specificities. In the context of an industry project, we have found that an inexperienced development team frequently faced difficulties with estimating the time needed to complete tasks, which led to missed deadlines in most of the projects. Such problem hampers the risk management and degrades the relationship with the customer. Upon closer analysis, it was identified that the main reason to this issue was the team's inability to breaking down a larger task into smaller sub-tasks and associate a realistic workload to each part. Then, based on traditional techniques, a structured approach to workload estimation was introduced in the SCRUM planning meeting to leverage the team's estimation skill. This approach was implemented in two development projects and increased the accuracy in the estimate defined by the team, yielding realistic schedules and increasing technical visibility.

KEYWORDS

Agile, Hybrid Agile Approach, Management, Scrum, Inexperienced Teams, PMBOK, Workload Estimation.

1. INTRODUCTION

Novel business models, strategies, technologies, and global transformations have changed the way companies manage their businesses in an increasingly competitive market. This scenario thrives the companies to adopt agile-oriented attitudes in most aspects of their workflow, e.g., project and team management. The search for efficiency in management activities has assumed a fundamental role to overcome daily faced challenges and strengthen success chances. Nevertheless, off-the-shelf strategies frequently were unable to meet all requirements and internal policies of large companies, leading to novel management models.

In a scenario of sustained unpredictability, where projects suffer from constant scope changes, agile methods tend to reflect stakeholders' requirements along the project course. During the last few years, the Scrum framework becomes standard for most companies in the software industry, despite presenting challenges in the planning phase [1]. Another challenge is its dependency on empirical methods based on developers' experience to estimate project tasks accurately. Deriving and estimation tasks from backlog items are complex activities that are even more challenging to inexperienced teams, typically composed of recently graduated professionals with limited industry and software development experience. According to Cerpa et al. [2], Scrum's premises

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 345-352, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101426

contribute to creating high-risk scenarios on projects developed by inexperienced teams since unrealistic schedules may result in unplanned costs that increase the probability of failures.

Another important factor of concern is the alignment with the client's expectations. Software clients typically expect developers to deliver high-quality projects, on the schedule, and at the lowest cost [3]. Most of Scrum's ceremonies are related to these expectations, mainly those designed to create a well-defined product and sprint backlogs. The product backlog is a document that contains all required features of a project that later was refined and collected into the sprint backlog. Inexperienced teams may have difficulty creating the backlogs since it also depends on the ability acquired on previous projects. A delay, or even poorly documented backlogs, can consequently impact the perceived quality associated with project deliveries and the result.

In order to improve workload estimation in inexperienced teams during the development of complex software projects, we designed and evaluated a hybrid agile approach that combines both (i) Scrum agile-oriented principles and (ii) PMBOK (Project Management Body of Knowledge) planning practices. This combination decreases the risk of missing deadlines due to inaccurate workload estimation. In addition to theoretical guidance in designing our approach, we took advantage of the lessons learned on two real-world projects from a large software company with well-defined policies and high client expectations about the development course deliveries.

Our paper is organized as follows: Section 2 presents the background information about traditional, agile, and hybrid project management approaches and discusses related work. Section 3 provides the context of our proposal. Section 4 describes our hybrid model and how it combines both PMBOK and Scrum aspects to produce a structured workload estimation that can be used by inexperienced developers. Section 5 discusses the lessons learned while implementing our model in real-world projects from a sizeable mobile-related software development company. We conclude the paper with a discussion of results and possible avenues for future work in Section 6.

2. RELATED WORK

In the last two decades, agile methods increased their popularity and were established as primary software development methodologies. A large number of efforts [4-7] were focused on studying their application under different circumstances and evaluate their advantages and shortcomings. Felker et al. [8] evaluated the implementation of the Scrum framework for both UX and software development in an undergraduate team without Scrum's previous experiences. Felker et al. describe multiple challenges the team faced, such as difficulty deciding Sprint length, balancing the amount of new functionality and a bug fix in each sprint, and estimating the time required to complete each task. Most of these problems are due to the team's lack of experience with the Scrum framework. To cope with these issues, the authors suggest allocating more time than the team initially suggested for some tasks and planning each task in a detailed fashion.

An approach to deal with estimating effort during planning sessions is to use computational models that automatically predict the required effort. Bilgaiyan et al. [9] review the most relevant works with this common goal and found that most of them use machine learning techniques, such as neural networks and support vector regression. These models can achieve accurate results. However, they require an extra effort of model training and fine-tuning its hyper-parameters. Besides, these approaches do not help an inexperienced team develop the team's skill in estimating task complexity and effort.

Another way to handle the challenges during Scrum implementation is to adapt and modify the standard process. This approach has the advantage of engaging team members in improving the team's deficiencies. Graphenthin et al. [1] facilitates task breakdown by using a meeting room with four whiteboards. Each board is used to analyze a single aspect, i.e., business, integration, data, and interaction) of a backlog item. Hayata et al.[10] propose a hybrid methodology that uses traditional software life cycle methodologies during the initial (e.g., requirements analysis, documentation) and final (e.g., testing phases of a project)t, whereas using Scrum for the design and implementation phases. These methods have the advantage of developing a team's skills; however, they add overhead to the usual process, e.g., an interaction room, detailed documentation, and still depend on expert's participation for success, e.g., waterfall practitioners.

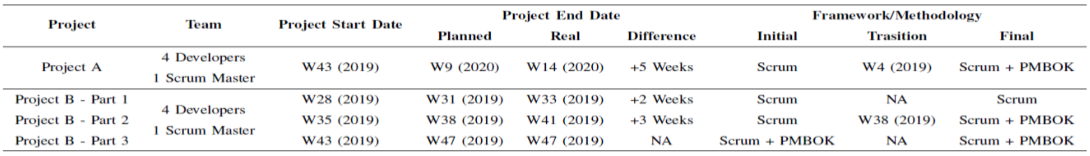
In [12], the authors assess review recommenders ensuring expertise during the review, reducing the core team's review workload, and the turnover risk. For this, they implement a recommender system that combines learning and retention; the aware recommenders effectively reduce the risk of turnover. In [13], the authors propose a hypothetical work commitment model to agile programming improvement groups. Utilizing auxiliary condition demonstrating, we found that agile practices lessen work requests (saw remaining task at hand and job vagueness) and backing position assets (saw importance and occupation self-governance). In [14], it is investigated how to improve the information on the best way to quantify productivity being developed groups where many inconstancies may exist because of the human factor. The primary spotlight is on disclosing the entire cycles and analyzing them as far as proficiency and adequacy. Like this, the authors uncover conceivably shrouded expenses and dangers to make remedial moves practically during the product venture life cycle. This work uses PMBOK's Work Breakdown Structure definition to provide a structured approach for workload estimation that can be used by inexperienced developers.

3. FROM SCRUM TO A HYBRID MODEL: TEAM PROFILE

The design of our hybrid model was based on projects that ran in a large mobile-related company that is mainly responsible for customizing the Android Operating System to Latin American countries with the requirements of its clients, e.g., device manufacturers and mobile carriers.

In this experience report, we focused on two projects that were performed during 2019 and were developed by two different developers' teams, as shown in Table 1. Each team was composed of four developers responsible for carrying out all technical activities, with an average of six months of experience in software development, and a member in the Scrum Master's role. As suggested in the Scrum framework, both teams were multidisciplinary and had all the necessary technical skills to carry out all the projects. Another premise was that both teams worked in a self-organized way: themselves defining, within the organization's context and the Scrum structure, how to perform the tasks and how to manage the progress towards the goals agreed with the Product Owners.

Table 1: Overview of different projects using (i) scrum only, (ii) transitioning from scrum to hybrid model, and (iii) hybrid model



Project	Team	Project Start Date	Project End Date			Framework/Methodology		
			Planned	Real	Difference	Initial	Trasition	Final
Project A	4 Developers 1 Scrum Master	W43 (2019)	W9 (2020)	W14 (2020)	+5 Weeks	Scrum	W4 (2019)	Scrum + PMBOK
Project B - Part 1	4 Developers 1 Scrum Master	W28 (2019)	W31 (2019)	W33 (2019)	+2 Weeks	Scrum	NA	Scrum
Project B - Part 2		W35 (2019)	W38 (2019)	W41 (2019)	+3 Weeks	Scrum	W38 (2019)	Scrum + PMBOK
Project B - Part 3		W43 (2019)	W47 (2019)	W47 (2019)	NA	Scrum + PMBOK	NA	Scrum + PMBOK

The first project, hereafter called Project A, consisted of 3 phases, where each one was associated with a release with well-defined functionalities. In July 2019, at the beginning of Project A, the development team defined an initial Sprint backlog that estimated 2 Sprints (4 weeks) to deliver the first project phase. However, during Sprint 1, unforeseen tasks were added to the planned stories, which increased their complexity. This led to the Sprint failure and increased the number of stories that needed to be completed in Sprint 2 to meet the defined deadline. During the Sprint retrospective, the team could not detect the leading cause of the deadline missing and assumed that increasing working hours would be enough to complete the Sprint 1 remaining tasks and Sprint 2 planned stories. Nevertheless, Sprint 2 occurred, and the did not planed tasks were added to the Sprint, causing another failure. Such behavior led to an increasing of 2 weeks in the first phase and three weeks in the second phase of the project, representing a 50% error in the development team's estimation. In the first two phases, it is possible to observe sufficient technical knowledge to meet the project's development demand. Scrum ceremonies performed according to the recommended guideline, the correct scope definition, task duration, and effort estimates were insufficient.

In this context, we designed a hybrid model to mitigate the workload estimation errors that were typically committed by an inexperienced team of developers. We detail the design of our hybrid model in Section 4. We also considered the lessons learned through the transition of Scrum only to our hybrid project management approach, as illustrated in Table I. Note that, in every transition to the hybrid model, both the scope and schedule of the projects were revisited. The revised scopes consisted of product backlog reviews and WBS (Work Breakdown Structure) creations. In comparison, the revised schedules were concerned about the definition, sequencing, and estimation of the tasks.

4. DESIGNING A HYBRID MODEL

Scrum is a framework based on empirical theories of process control. The knowledge comes from both experience and evidence manners, and their progress is formed on observations of reality, thus generating a solid premise that the team needs to be multidisciplinary and formed by experienced people.

Chow et al. [11] surveyed multiple software projects and concluded that agile techniques are one of the critical factors for project success. Thus, our goal designing a hybrid model was two-fold: i) Minimize the changes to the Scrum guidelines that were necessary to cope with the team's inexperience in order to retain Scrum's advantages and ii) improve the team's estimation skill so that, eventually, the team was capable of returning to Scrum original guidelines.

Therefore, our proposal consists of using concepts from the PMBOK, a globally recognized guide to the best practices in project management, and applying it to the agile-oriented principles and ceremonies of the Scrum Framework to design a hybrid project management model offers a structured workload estimation to inexperienced developers. By combining both PMBOK and Scrum, we aimed to provide a more elaborated project scope by using SCRUM best practices during initial phases, e.g., *Initiation* and *Planning*, and *Scrum ceremonies* during execution, as shown in Figure 1.

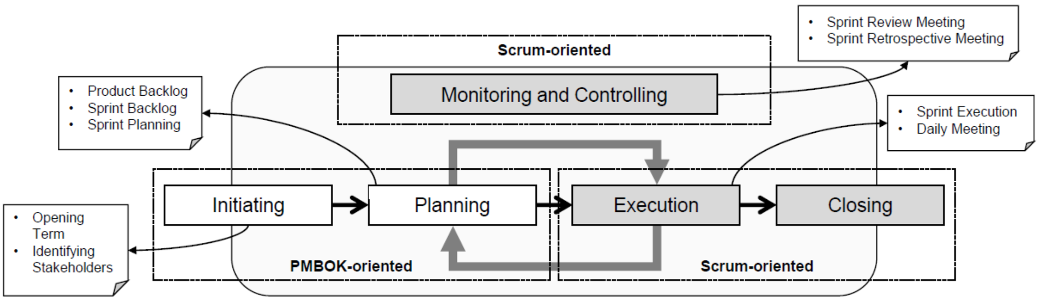


Figure 1: Model Lifecycle with PMBOK and SCRUM.

Our hybrid model combines both (i) PMBOK best practices of project management during initial phases with (ii) Scrum agile-oriented principles and ceremonies of execution to provide a structured approach for workload estimation that can effectively assist teams of inexperienced developers. An essential aspect of our proposal is that it prevents the definition of workload without ensuring the correct task prioritization and knowledge of critical paths that are not under the development team's responsibility, which is difficult to promise on empirical methods of estimates when conducted by inexperienced teams. A concern while designing the hybrid model was to maintain Scrum's premises, such as:

1. not bureaucratizing the process.
2. not excessively and unnecessarily documenting, and
3. not creating or using unnecessary processes to avoid adding slowness to the project's execution is a commonly criticized aspect of traditional methodologies.

In this context, our hybrid model attempts to avoid traditional project management's main disadvantages that are often so bureaucratic that a project's requirements may change even before development begins [12]. To refine our approach, we also considered the lessons learned through the transition of Scrum only to our hybrid project management model, as illustrated in Table 1. Note that, in every transition to the hybrid approach, both the project's scope and schedule were revisited. The revised scopes consisted of product backlog reviews and WBS creations. Simultaneously, the revised schedules were concerned about the definition, sequencing, and estimation of the tasks.

As depicted in Figure 2, our hybrid model implements a set of tools and techniques suggested by PMBOK during the Planning Phase to ensure a more reliable project scope definition by structuring the (i) scoping and (ii) schedule generation processes:

1. **Work Breakdown Structure (WBS) Creation:** developed to establish a common understanding of the project scope by creating a decomposition of the work into easily manageable parts called work packages, estimated with further exactness.
2. **Activities Definition:** creates the activities list (the work packages broken down into the activities needed to produce them) and the list of activity attributes (information related to the activities). These activities can be considered as actions that need to be performed to execute each work package
3. **Activities Sequencing:** Studies of the relationship between project activities determine the logical Sequence that serves as the basis for the project schedule.
4. **Activities Duration Estimation:** estimates the duration of each activity based on a three-point method (PERT), i.e., the best-case, most likely, and worst-case estimates.

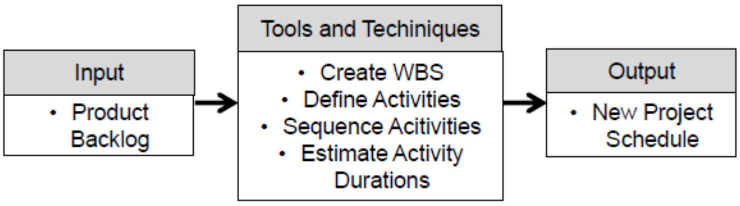


Figure 2: Set of tools and techniques suggests by PMBOK that are implemented in our hybrid model to provide a structured approach for workload estimation.

The designed model provides a structured approach to planning sessions while the remaining Scrum ceremonies (daily meeting, sprint review, and sprint retrospective) are performed without changes. The team does not need to have an extensive knowledge of PMBOK to apply these changes. Besides, given that PMBOK is a widely known process management guide, many companies may use existing collaborator knowledge to leverage such a hybrid model.

5. LESSON LEARNED

We applied the hybrid model in the phase 3 of the Project A. During the planning meeting, the *WBS creation activities definition* helped the team in obtaining a more detailed vision of required tasks, whereas *activities sequencing* helped them visualize inter-dependency among tasks that were not obvious at first sight. The PERT approach to effort estimation also assisted the team by leading them into thinking of all possible scenarios.

Using this model, the team estimated this project phase would require 5 weeks (3 sprints) to be completed. The team was able to meet the planned deadline and no additional tasks were observed during this period. In contrast to the observed in previous phases of the project, we noted that the designed model helped to leverage the team's estimation skill and yield planning sessions with improved accuracy.

The same pattern of task underestimation was observed in another development project, hereafter called Project A. After consecutive sprint failures and unforeseen tasks being added to the sprint, when there were 5 weeks remaining until the next release, it was clear that the project would miss its initial deadline. The hybrid model was applied to generate a new deadline and improve the detailing of product backlog and estimates. During the planning session, the team anticipated unforeseen tasks and interdependence, and estimated that it would require 10 weeks to complete the remaining stories. The team was able to meet the new deadline with only a few tasks being added to the sprints.

By combining (i) the best practices suggested by the PMBOK Guide for task definition and estimation and (ii) the Scrum agile-oriented principles and ceremonies, we notice that our hybrid model was able to effectively provide a more elaborate project scope during the initial phases of *Initiation* and *Planning*.

We also observed that our hybrid model helps to verify that companies with agile-oriented environments can benefit from the use of more traditional concepts brought by the PMBOK Guide without having to necessarily ignore the SCRUM's dynamism, thus being able to add a layer of maturity to the planning stage of the projects by eliminating the necessity to have self-managed and experienced teams.

Another important gain was to provide a *win-win* model for the team, in which the technical members can improve their task estimation skills in a structured approach, with no need of an external tool for effort estimation such as those discussed in Section 2. On the other hand, the *Project Manager* can benefit from mature and precise planning that guides the development team into producing within an agile structure.

6. CONCLUSIONS

This work addressed the challenge of accurately estimate effort during the Scrum planning meeting when there are no experts in the development team. We presented a hybrid project management model that assists inexperienced teams in workload estimation and minimizes the risk of missing deadlines. This model was implemented in two real-world software projects, and it was effective in improving the accuracy of effort estimation. We evaluated that traditional software development models can be useful in tackling Scrum challenges in a structured way and might be reproduced by other teams with similar issues. Besides improved accuracy, we also observed that this hybrid model retains Scrum qualities, such as dynamism and teams' autonomy, whereas developing teams' estimation skills leads to increasingly accurate planning meetings. Finally, another contribution was the understanding that new hybrid approaches to project management can be built and used to meet the most diverse demands, without needing to be limited to a single software development model.

In the future, we intend (i) to study how our model copes with developing a team of inexperienced developers in the long-term; (ii) to combine the PMNOK with other agile methodologies, for example, Extreme Programming-XP; (iii) validate our model with others contexts of applications and companies.

ACKNOWLEDGEMENTS

This work was partially supported by Samsung Eletrônica da Amazonia Ltda, under the auspice of the informatic law no 8.387/91. This work was supported by the Sidia: Institute of Science and Technology.

REFERENCES

[1] S. Grapenthin, S. Poggel, M. Book, and V. Gruhn, “Facilitating task breakdown in sprint planning meeting 2 with an interaction room: An experience report,” in 2014 40th EUROMICRO Conference on Software Engineering and Advanced Applications, Aug 2014, pp. 1–8.

[2] N. Cerpa and J. M. Verner, “Why did your project fail?” Commun. ACM, vol. 52, no. 12, p. 130–134, Dec. 2009. [Online]. Available: https://doi.org/10.1145/1610252.1610286.

[3] A. Tonini, M. Silva, and M. Spinola, “Software expectation management by means of service level agreements in software maintenance,” in 19th International Conference on Production Research. Available: http://www. icpr19. cl/mswl/Papers/252. pdf. Citeseer, 2007.

[4] I. Zada and S. Shahzad, “Issues and implications of Scrum on global software development,” Bahria University Journal of Information & Communication Technologies (BUJICT), vol. 8, no. 1, 2015.

[5] L. Silva, C. Santana, F. Rocha, M. Paschoalino, G. Falconieri, L. Ribeiro, R. Medeiros, S. Soares, and C. Gusm˜ao, “Applying xp to an agile–inexperienced software development team,” in International Conference on Agile Processes and Extreme Programming in Software Engineering. Springer, 2008, pp. 114–126.

[6] E. Hossain, M. A. Babar, and H.-y. Paik, “Using Scrum in global software development: a systematic literature review,” in 2009 Fourth IEEE International Conference on Global Software Engineering. IEEE. 2009, pp. 175–184.

[7] J. L´opez-Mart´ınez, R. Ju´arez-Ram´ırez, C. Huertas, S. Jim´enez, and C. Guerra-Garc´ıa, “Problems in the adoption of agile-Scrum methodologies: A systematic literature review,” in 2016 4th international conference in software engineering research and innovation (conisoft). IEEE, 2016, pp. 141–148.

[8] C. Felker, R. Slamova, and J. Davis, “Integrating ux with Scrum in an undergraduate software development project,” in Proceedings of the 43rd ACM technical symposium on Computer Science Education, 2012, pp.301–306.

[9] S. Bilgaiyan, S. Mishra, and M. Das, “A review of software cost estimation in agile software development using soft computing techniques,” in 2016 2nd International Conference on Computational Intelligence and Networks (CINE), Jan 2016, pp. 112–117.

[10] T. Hayata and J. Han, “A hybrid model for it project with Scrum,” in Proceedings of 2011 IEEE International Conference on Service Operations, Logistics, and Informatics. IEEE, 2011, pp. 285–290.

[11] H. F. Cervone, “Understanding agile project management methods using Scrum,” OCLC Systems & Services: International digital library perspectives, vol. 27, no. 1, pp. 18–22, Feb. 2011. [Online]. Available: https://doi.org/10.1108/10650751111106528.

[12] Ehsan Mirsaeedi and Peter C. Rigby. 2020. Mitigating turnover with code review recommendation: balancing expertise, workload, and knowledge distribution. In <i>Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering</i> (<i>ICSE '20</i>). Association for Computing Machinery, New York, NY, USA, 1183–1195. DOI:https://doi.org/10.1145/3377811.3380335

[13] HUCK-FRIES, Veronika et al. The Role of Work Engagement in Agile Software Development: Investigating Job Demands and Job Resources. In: Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019.

[14] CALDEIRA, João et al. Assessing Software Development Teams' Efficiency using Process Mining. In: 2019 International Conference on Process Mining (ICPM). IEEE, 2019. p. 65-72.

AUTHOR INDEX