`

Computer Science &Information Technology 156

Natural Language Computing

`

David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

- 7th International Conference on Natural Language Computing (NATL 2021), November 27~28, 2021, London, United Kingdom
- 5th International Conference on Networks & Communications (NETWORKS 2021)
- 7th International Conference on Fuzzy Logic Systems (Fuzzy 2021)
- 7th International Conference on Computer Science, Engineering And Applications (CSEA 2021)

Published By

AIRCC Publishing Corporation

`

Volume Editors

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

Dhinaharan Nagamalai (Eds),
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-54-1
DOI: 10.5121/csit.2021.112001 - DOI:10.5121/csit.2021.112013

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

`

Preface

The 7th International Conference on Natural Language Computing (NATL 2021) November 27 ~ 28, 2021, London, United Kingdom, 5th International Conference on Networks & Communications (NETWORKS 2021), 7th International Conference on Fuzzy Logic Systems (Fuzzy 2021) and 7th International Conference on Computer Science, Engineering And Applications (CSEA 2021) was collocated with 7th International Conference on Natural Language Computing (NATL 2021). 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 NATL 2021, NETWORKS 2021, Fuzzy 2021 and CSEA 2021 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, NATL 2021, NETWORKS 2021, Fuzzy 2021 and CSEA 2021 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 NATL 2021, NETWORKS 2021, Fuzzy 2021 and CSEA 2021.

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,
Dhinaharan Nagamalai (Eds)

`

General Chair	Organization
David C. Wyld,	Southeastern Louisiana University, USA
Dhinaharan Nagamalai (Eds)	Wireilla Net Solutions, Australia

Program Committee Members

Abdel-Badeeh M. Salem,	Ain Shams University, Cairo, Egypt
Abdelhadi Assir,	Hassan 1st University, Morocco
Abdessamad Belangour,	University Hassan II Casablanca Morocco
Abdulraqeb Alhammadi,	University of Technology Malaysia (UTM), Malaysia
Abhas Kumar Singh,	Sri Venkateswara College of Engineering, India
Abtoy Anouar,	Abdelmalek Essaâdi University, Morocco
Addisson Salazar,	Universitat Politècnica de València, Spain
Adeyanju Sosimi,	University of Lagos, Nigeria
Adnan Mohammed Hussein,	Northern Technical University, Iraq
Alexander Gelbukh,	Instituto Politécnico Nacional, Mexico
Alireza Valipour Baboli,	University Technical and Vocational Babol, Iran
Amal Zouhri,	Sidi Mohammed Ben Abdellah University, Morocco
Ana Luísa Varani Leal,	University of Macau, China
António Abreu,	ISEL, Portugal
Aravinda c v,	NMAM Institute of Technology, India
Archit Yajnik,	Sikkim Manipal University, India
Arunkumar Thangavelu,	VIT, India
Basant Verma,	G H Raisoni Group of Colleges, India
Benyamin Ahmadnia,	University of California, USA
BeshairAlsiddiq,	Prince Sultan University, Saudi Arabia
Bin Zhao,	Northwestern Polytechnical Unviersity, China
Bipan Hazarika,	Gauhati University, India
BrahimLejdel,	University of El-Oued, Algeria
Callistus Ireneous Nakpih,	C.K. Tedam University of Technology, Ghana
Cheng Siong Chin,	Newcastle University, Singapore
Chuan-Ming Liu,	National Taipei University of Technology, Taiwan
Dadmehr Rahbari,	Tallinn University of Technology, Estonia
Daniel Rosa Canedo,	Federal Institute of Goias, Brazil
Dário Ferreira,	University of Beira Interior, Portugal
Desmond Bala Bisandu,	Cranfield University, UK
Dharmendra Sharma,	University of Canberra, Australia
Diyar Qader Saleem Zeebaree,	Duhok Polytechnic University, Iraq
Domenico Ciuonzo,	University of Naples Federico, Italy
Dongping Tian,	Baoji University of Arts and Sciences, China
Dong-yuan Ge,	Guangxi University of Science and Technology, China
Ederval Pablo Ferreira da Cruz,	Instituto Federal do Espírito Santo, Brazil
El Murabet Amina,	Abdelmalek Essaadi University, Morocco
Elzbieta Macioszek,	Silesian University of Technology, Poland
Emilio Jimenez Macias,	University of La Rioja, Spain
Fatih Korkmaz,	Cankiri Karatekin University, Turkey
Felix J. Garcia Clemente,	University of Murcia, Spain

`

Fernando Zacarias Flores,	Universidad Autonoma de Puebla, Mexico
Francesco Zirilli,	Sapienza Universita Roma, Italy
Froilan D. Mobo,	Philippine Merchant Marine Academy, Philippines
Goutam Sanyal,	National Institute of Technology, India
Grigorios N. Beligiannis,	University of Patras, Greece
Grzegorz Sierpinski,	Silesian University of Technology, Poland
Guilong Liu,	Beijing Language and Culture University, China
Gurpreet Singh,	Punjab Institute of Technology, India
Hamed Taherdoost,	Canada West University, Canada
Hamid Ali Abed AL-Asadi,	Iraq University college, Iraq
Hamidreza Rokhsati,	Sapienza University of Rome, Italy
Hedayat Omidvar,	Research & Technology Dept, Iran
Hilal adnanfadhil,	Al -farabi university college, Iraq
Hiromi Ban,	Nagaoka University of Technology, Japan
Hossein Rajaby Faghihi,	Michigan State University, USA
Ilango Velchamy,	CMR Institute of Technology, India
Israa Shaker Tawfic,	Ministry of Science and Technology Baghdad, Iraq
Jasy Liew Suet Yan,	Universiti Sains Malaysia, Malaysia
Jawad K. Ali,	University of Technology, Iraq
Jesuk Ko,	Universidad Mayor de San Andres (UMSA), Bolivia
Jian Wang,	China University of Petroleum (East China), China
Joey S. Aviles,	Panpacific University North Philippines, Philippines
Jose Alfredo F. Costa,	Federal University, Brazil
K.Suganthi,	Vellore Institute of Technology, India
Kamel Hussein Rahouma,	Minia University, Egypt
Kamel Jemai,	University of Gabes, Tunisia
Kazim Yildiz,	Marmara University, Turkey
Ke-Lin Du,	Concordia University, Canada
Kirtikumar Patel,	I&E Engineer, USA
Klenilmar L. Dias,	Federal Institute of Amapa, Brazil
Kocsis Gergely,	University of Debrecen, Hungary
Laszlo T. Koczy,	Budapest University of Technology, Hungary
Luisa Maria Arvide Cambra,	University of Almeria, Spain
Maad M. Mijwil,	Baghdad College of Economic Sciences University, Iraq
Mahmood Hashemi,	Beijing University of Technology, China
Mai Zaki,	American University of Sharjah, UAE
Manoj Sahni,	Pandit Deendayal Energy University, India
Mansour Y. Bader,	Al-Balqa Applied University, Jordan
Marie-Anne Xu,	Crystal Springs Uplands School, USA
Mario Versaci,	DICEAM - Univ. Mediterranea, Italy
Masoomeh Mirrashid,	Semnan University, Iran
Maumita Bhattacharya,	Charles Sturt University, Australia
Md Azher Uddin,	Ajou University, South Korea
Michail Kalogiannakis,	University of Crete, Greece
Mohamed alielsayedfahim,	benha university, Egypt
Mohamed Anis Bach Tobji,	University of Manouba, Tunisia
Mohamed Arezki Mellal,	M'HamedBougara University, Algeria
Mohamed Fakir,	university Sultan Moulay Slimane, Morocco
Mohamed Skander Daas,	Freres Mentouri Constantine 1 University, Algeria
Mohammad Jafarabad,	Qom university, Iran
Mohammad Mahmiud Abu Omar,	Al-Quds Open University, Palestie
Mohammad Siraj,	King Saud 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)

`

7th International Conference on Natural Language Computing (NATL 2021)

5th International Conference on Networks & Communications (NETWORKS 2021)

7th International Conference on Fuzzy Logic Systems (Fuzzy 2021)

`

7th International Conference on Computer Science, Engineering and Applications (CSEA 2021)

MULTI-LANGUAGE INFORMATION EXTRACTION WITH TEXT PATTERN RECOGNITION

Johannes Lindén, Tingting Zhang, Stefan Forsström and Patrik Österberg

Department of Information System and Technology, Mid.
Sweden University, Sundsvall, Sweden

ABSTRACT

Information extraction is a task that can extract meta-data information from text. The research in this article proposes a new information extraction algorithm called GenerateIE. The proposed algorithm identifies pairs of entities and relations described in a piece of text. The extracted meta-data is useful in many areas, but within this research the focus is to use them in news-media contexts to provide the gist of the written articles for analytics and paraphrasing of news information. GenerateIE algorithm is compared with existing state of the art algorithms with two benefits. Firstly, the GenerateIE provides the co-referenced word as the entity instead of using he, she, it, etc. which is more beneficial for knowledge graphs. Secondly GenerateIE can be applied on multiple languages without changing the algorithm itself apart from the underlying natural language text-parsing. Furthermore, the performance of GenerateIE compared with state-of-the-art algorithms is not significantly better, but it offers competitive results.

KEYWORDS

Information Extraction, IE, Information representation, Knowledge Graph, Natural Language Processing, NLP, Pattern Recognition, Entity Recognition.

1. INTRODUCTION

Modelling data with machine learning algorithms has shown promising results in various areas, such as image processing, robotics and natural language processing. These areas are growing, both in size and number, and machine learning becomes more advanced every day. Especially within news-media companies that tries to reach their customers with functional and promising algorithms. Natural language processing which is a very central part of companies that produce content can combine several models that compute bits and pieces of information about the text into a single model for a particular predictive goal. This research will use previously well explored natural language models to automatically extract information from text. The extracted information can be collected into a database which one can derive knowledge from. The news industries could then use this knowledge to analyse their supply and demand as well as creating summaries of their articles and paraphrase words to make it more understandable by certain target groups. While writing, there are different ways to express the message, while the gist of the text remains the same. Depending on the intended audience, an author can adapt the text with different formulation and terminology to ease the readers understanding of the text. The research problem is about finding this gist from any written text. This paper is trying to achieve this by extracting meta-data from the text. The algorithm developed in this paper is able to identify the gist of the sentence regardless of the grammatical structure and individual expression of each

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 01-17, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112001

author. With the information it would be possible for another algorithm or another person to adapt the text based on who is reading it. Adaptation of texts to different audiences could with this algorithm be dynamically automatized in the future. Therefore, the aim of this research is to extract meta-data in terms of entities and what relations these entities have with each other. This in turn can be captured automatically as the gist of the text. The text entities are words that identify object/things for example "house" or "cat" and relations are binding words between these entities such as "is" or "belongs to".

A knowledge graph one can quickly and flexibly query large data sets of entities together with their relations and sometimes even conclude new relations based on the imported information. The queries can answer what entities are connected and what relation they have with each other. From this information, possible answers can logically be derived by the algorithm. For example, if the goal is to know which colleagues are working with "Johannes Lindén", it is possible to query for entities that have a relation "colleague" with "Johannes Lindén". However, indirect relations and entities can also be found automatically. For instance, based on the sentence "Johannes Lindén works at Mid Sweden University", an indirect entity such as "A colleague" could also have the relation "works at" with the entity "Mid Sweden University". There is a significant number of indirect relations and entities which can be generated and one consequence is that it scales poorly in terms of storage and performance in relational databases. Instead, trained knowledge graph models can be used [1]. The data sets that these knowledge graph models are trained on are often generated from information extraction models. The information extraction algorithm developed in this research is called Generate Information Extraction (GenerateIE).

The goal of this research is to create an algorithm, GenerateIE, that extracts information from plain texts in multiple languages. A second goal is to combine a set of state-of-art algorithms to enhance the information extraction process of the GenerateIE algorithm. To evaluate these goals there are two metrics that will be considered. The first metric will be the accuracy of the number of correctly extracted data-points from the text. The second metric will be the intersection between a set of evaluation algorithms, e.g., a comparison of data-points that were found by one algorithm, but not by other algorithms. The novelty of this research is to use grammatical rules, common to multiple Germanic languages, to deduct which words are entities and relations. Another novelty is to extend the information extraction concept within area of news-media as well as reaching competitive accuracy with other state-of-the-art algorithms.

The remainder of this article is as follows: Section 2 presents relevant related work for this research. Section 3 presents our approach and the proposed model. Section 4 presents our evaluation thereof and the results. Finally, Section 6 presents our conclusions and our future work.

2. RELATED WORK

Information extraction is a difficult task, mostly because of complexity and variations in languages but also because there are few ways to evaluate the output [2, 3]. The output is generally a list of data-points. One data-point consists of three components, subject, relation and object, often referred to as a triple. Today’s algorithms are capable to solve the information extraction task in English with either a greedy result with duplicated triples or missing entire sentences. They are today usually based on a fuzzy dataset, which is used to train a neural network model such as the OpenIE project [4]. A fuzzy dataset contains noise, and it is hard to determine the real accuracy and how well it performs for a certain use-case. The OpenIE algorithm has had several iterations of improvements over the years with contributors from for example Wu and Weld [4] as well as Angeli et al. [5]. Investigations of multi-lingual information extraction have been conducted by Claro et al. [6] which have a similar setup of supervised training. The OpenIE approaches all use a training data set for the information extraction itself and additional training on the dependent components such as dependency parsing and part-of-speech tagging as well.

Gashteovski et.al has developed an information extraction model called MinIE [7]. The MinIE algorithm is a combination of an older algorithm called ClauseIE developed by Del Corro et. al. [8] and aggressive information extraction optimization. The input of MinIE comes from the ClauseIE algorithm that suggests clauses to be considered as informative constituents of the input sentence. MinIE moves the constituents around until a potential relation is found.

There are things that complicate the matter of retrieving a sentence dependency tree and part-of-speech tags even more and that is if a sentence has entities in it that are spanning multiple words such as the cookie brand "Ben and Jerrys" or music band "Rolling Stones". The dependency parsing will not treat these words as entities but rather include them in the dependency tree as if they were separate. These entity problems require a trained named entity recognition model. Extracting information also complicates things when a sentence might refer to previous paragraphs or mentioned entities within such as he, she, it, that, them. A co-reference word model may deal with most of these sentence references which has been investigated by Clark et.al [9]. Since a sentence can be formulated in several different ways and still carry the same information, one consideration would be to try and simplify the sentence before using an information extraction model, a work conducted by Narayan and Gardent where the goal was just this to simplify the sentence [10].

Previous research is struggling to finding a good evaluation method since there are nearly no existing qualified data sets for the task of training such a model. Either a noisy labelled data set is used to estimate the performance of the algorithm, or the existing algorithm is compared with another in different ways. In the handbook of natural language processing, Alexander Clark et al.[11] are writing about different methods of dealing with evaluation when lacking a correct dataset within the NLP field. Alexander’s handbook among other articles summarizes the ways of evaluation into four different methods which are considered in this research [11, 12, 13].

1. Intrinsic evaluation: Manually tag a set of label and compute precision, recall and sometimes F-score.
2. Extrinsic evaluation: Use QA data set and match label with answer and question (for example WikiData→Wikipedia).
3. Laboratory evaluation: Letting people say if the predicted label is correct or incorrect.
4. Real world cases evaluation: Use people with expertise of which label should exist of sentences related to their field of expertise.

Intrinsic and laboratory evaluation requires some form of definition how to label the data points and although this was investigated, a decision was made to postpone this evaluation since the longer the sentences the more complex they became to label. The intrinsic evaluation is more complex and will therefore take longer time to label than the laboratory evaluation and the probability of error is also higher for the intrinsic evaluation. Extrinsic evaluation would work for a customer QA system. However, the news-media business case requires specific types of data which is not available for extrinsic evaluation at the time of writing. The real-world case evaluation re-quires experts within the English language and there are no openly accessible data sets for this purpose.

Google released a data source platform called GDELT [14] that stores billions of news metadata from all over the world, such a system could be used as valuable information to further enhance an IE algorithm. The system has the computer power to store and monitor world news on the internet from certain news sources, new events as well as events reaching as far back in time as 1979.Over 200 million events are recorded from over 240 countries and available for live requests. A similar system for crisis news is the Integrated Crisis Early Warning System (ICEWS). In 2013, a comparison between the GDELT and ICEWS was made that compared the popularity and scale of the two data sources. [15]

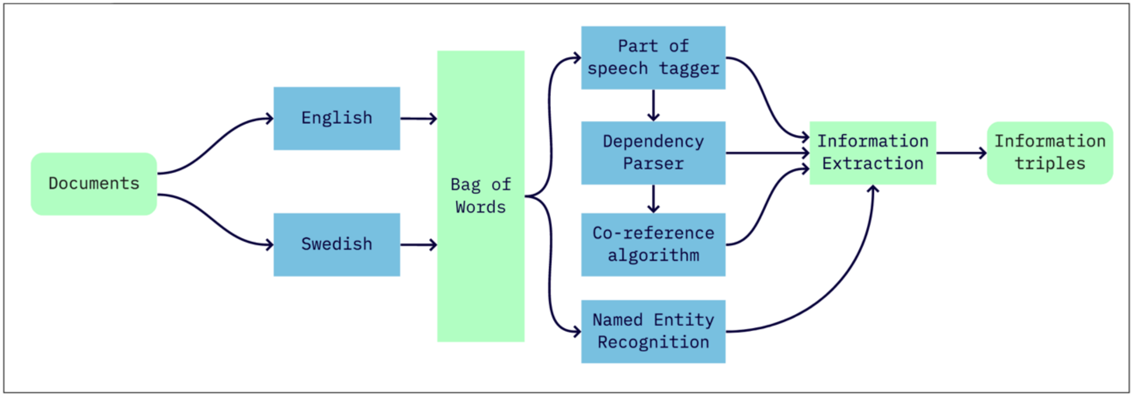


Figure 1. An overview of the proposed GenerateIE model. The input is an arbitrary text and the output is a list of data-points referred to as triples. A triple consists of a subject, relation and object components.

3. METHOD

The method of extracting information meta-data from texts used in this article is to propose a new algorithm called GenerateIE that is composed by several natural language processing models as well as a final algorithm that combines their predictions all together into a list of entity pairs and their relation, called a triple. Figure 1 shows an overview of how these natural language processing models are combined as well as the concept of the algorithm. Furthermore, the method section will explain the data sets, each natural language processing model, the final algorithm and how itis evaluated.

3.1. Dataset

The data set used when training the models of this research is Wikipedia articles selected by random sample with a total of 1000 articles. The data set is a good candidate since information can be extracted for different Germanic languages with texts of the same format and content. While preparing the data set some filtering was done for empty pages with no sentences as well as the format parameters within the articles written in XML syntax. The data set itself does not have any labels thus it is needed to manually evaluate the extracted triples by selecting reviewers. For the models requiring part-of-speech tags and dependencies another data set called Universal Dependencies [16] was used for the English language and a similar derived data set called "Talbanken" was used for the Swedish language [17, 18]. For the named entity recognition model, a pretrained model is fine-tuned with an additional entity type, relation, on WikiData data set. All relations in WikiData are extracted and a string match approach is used from the verbs written in Wikipedia articles.

3.2. Bag of Words

The first step in Figure 1 is a constructed one layered neural network model that will transform text paragraphs from the data set into several word vectors. This neural network is known as a bag of words algorithm. The word vectors are unique vectors that map the paragraph context associated with the word to a high dimensional vector space. A sigmoid activation function makes sure that the elements are bounded and by mapping the activation function according to Equation 1 it is ensured that the bounded values are within the interval -1 to 1. The vectors are constructed such that each element represents neighbouring words in a window. This way it is possible to relate a word by distance from another word [19]. The relation of two words can be obtained by computing the cosine similarity between their vector representations, see Equation 2. The cosine similarity will give a positive value when they are sharing similar contexts, a value close to zero when they have nothing in common, and a negative number when they appear in opposite contexts [19]. The vectors can also use common subtraction and addition operations to retrieve related word vectors, see Equation (3) for an example.

$$activation = \frac{2}{1 + \exp -X} - 1 \tag{1}$$

$$cosine\_similarity = \frac{vector_1 \cdot vector_2}{||vector_1||||vector_2||} \tag{2}$$

$$king - man + woman = queen \tag{3}$$

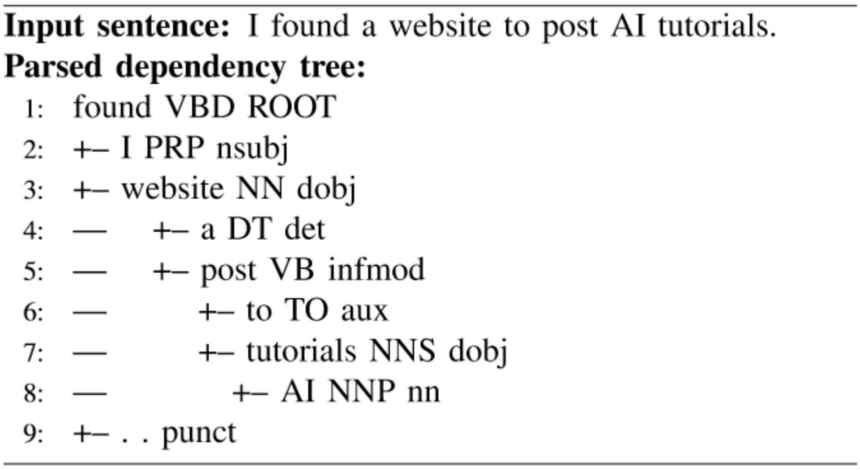
The Bag of Words algorithm takes an unstructured sequence of words forming a text sentence as input. The language of GenerateIE algorithm is for comparison reasons in English but Swedish language has also been evaluated in similar ways. Experiments in previously conducted research show that the continuous bag of words (CBOW) algorithm is reliable choice described by Mikolov et al [19].

3.3. Named Entity Recognition

To identify known entities, such as names of locations, organizations, people and more, a Named Entity Recognition (NER) model is used. Entities are identified by training a model that firstly recognizes in which context an entity usually exists, and secondly recognizes entity variances, i.e., multiple words can become an entity and certain words might be an entity in one context but not in another. A way to avoid that the dependency parser separates the potential entity words, a NER model is used to simplify prediction of the dependency tree and POS parser by replacing multiple word entities with similar single word entities. Different NER models were tried out, among them the model created by Finkel et.al. [20]. Among them, the GenerateIE performed best with a bidirectional BERT model transformer [21] by Google in terms of performance, number of entity types and a state-of-the-art research model. The supported types from the BERT model were Person, Organization, Time, Object, Event and Location. Since Time already is supported by the dependency parser, the entities used in this research were Person, Organization, and Location, Event and Object. Both the Swedish and English pre-trained models where fine-tuned on additional Wikipedia data while an additional entity type, Relation, was added. The Wikipedia dataset did not initially have any entity labels and therefore names of relations from WikiData was used by string matching the existing articles with the relation names.

3.4. Part-of-speech Tagger

The task of a Part-of-speech tagger (POS-tagger) model predicts a part-of-speech tag for each word in the input sentence. The tags could be Nouns, Verbs, Adverbs etc., and there is also a separate tag for special characters. Different languages have different number of tags. In total there are 55 tags in the English language, whereas the Swedish language has 21 tags. There are several algorithms that predict these tags such as SyntaxNet [22], and StanfordNLP [23]. In 2016 Google released a POS-tagger called SyntaxNet [24] with state-of-the-art performance, and one year later they announced an improved version [25]. In the experiments of GenerateIE the SyntaxNet algorithm is used to provide the English part-of-speech. The Swedish part-of-speech tagger is trained on a dataset from the resources mentioned in Nilson et al. [17]. The dataset is originally made by Jan Einarson's project is called Treebank [26, 18].



Input sentence: I found a website to post AI tutorials.
Parsed dependency tree:

```
1: found VBD ROOT
2: +– I PRP nsubj
3: +– website NN dobj
4: —    +– a DT det
5: —    +– post VB infmod
6: —        +– to TO aux
7: —        +– tutorials NNS dobj
8: —            +– AI NNP nn
9: +– . . punct
```

Figure 2. A part-of-speech example sentence parsed by SyntaxNet. The +– characters indicates a child path, followed by the word, part-of-speech tag and relation to patent word.

3.5. Dependency Parser

The task of a dependency parser is to break a sentence down to word dependencies. Each word in the sentence has a dependency to another word except for the root word. The dependencies between each word are named relations, for example "object to", "determinates" etc. Different dependency parsers provide different amount of named relations, e.g., in StanfordNLP [27] there are in total 47 named relations and in MaltParser [28] there are 65 named relations. In the GenerateIE algorithm, the StanfordNLP is used for the English language whereas the MaltParser is used for Swedish. See Figure 2 for an example of such a dependency tree together with the corresponding part-of-speech tags.

3.6. Co-reference Words (Word Linker)

A co-reference algorithm can be used to detect words in a sentence that also refers to other words in the context. For example, the sentence "John Doe went out for a break and he drank a cup of coffee", the co-reference algorithm is to identify that the word "he" refers to "John Doe". By linking these words together, the sentence dependency structure looks different and the extracted triples would make more sense if put into a knowledge graph. The knowledge graph would not contain the words he, she, it, them, they, etc. but instead they would be replaced by the original name of the identified entity. The co-reference models could be trained by using heuristic loss functions or reinforcement learning techniques [9]. The GenerateIE algorithm uses a co-reference model as a word linker to enhance the identification of entities. The mentions of a word will be linked together and replaced in the extracted information to enhance the value of each extracted word entity.

3.7. GenerateIE Algorithm

The GenerateIE algorithm will use the output of all previously mentioned algorithms (e.g., NER, POS-tagger, and dependency parser) and combine them into triples. The output of GenerateIE in Figure 1 is a set of triples T1, T2, ⋯, Tm. GenerateIE takes all the words and converts them into word-vectors using the bag of words algorithm. The triple extraction itself is rule based. That means that each word in a sentence is considered to be an entity and relation candidates. To narrow down the real entities and relations, GenerateIE utilize their part-of-speech tags as well as their word-dependencies. The part-of-speech tagger describes sentence parts as word classes. Dependency parsers work with set of grammatical rules in order to find relations between different word classes. This research uses the grammatical relations discovered by the dependency parser, and the entities found by the part-of-speech tagger, and then connects and reduces them into semantic entity-relation-entity triples, e.g., the elements to find the gist of a text.

Through the part-of-speech-tagger the likeliness of a word being a relation is increased by checking if the word is a verb. Similarly, the likeliness of a word being an entity is increased if the tag is a noun. Furthermore, neighbouring words of nouns are being concatenated depending on the grammatical rules found by the dependency parser. If the noun is bound to any word in the following list, the word would be considered to be part of the entity and concatenated with the noun:

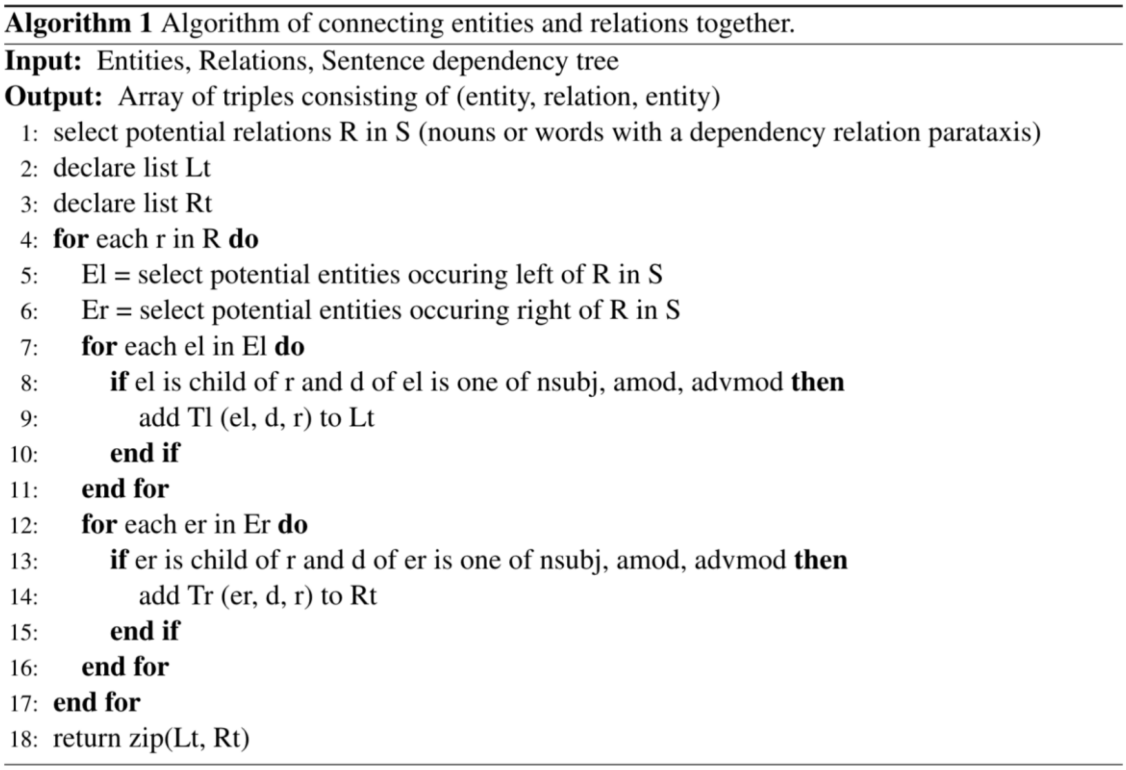
1. The noun is bound to an adjectival modifier (an adjective modifying the noun), for example, “the mother eats **red** meat".
2. The word is a number it is attached to the closest parent in the dependency tree.
3. The noun is bound to a nominal modifier, for example, "**Dr**. Andersson".
4. The noun is bound in a noun phrase as adverbial modifier, for example "I am 100 years **old**".
5. The noun is bound to a compound word, for example, "Let me borrow your **phone** book".

Similarly, the neighbouring words for a relation are being concatenated if the verb is bound to any of the following list, the word would be considered to be part of the relation and concatenated with the verb.

1. The verb is bound to an auxiliary word (a non-main verb), for example, "Meagan might have been lying"
2. The verb is bound to a copula word (link between subject to a subject complement), for example, "The sky is blue"

The found entities and relations are called soft entities and relations since they have yet not been connected into a triple. For entities the likeliness is made certain if the named entity recognition algorithm identifies one of the entities or their co-referenced word as an entity. For the relations the likeliness is made certain if the extended named entity recognition identifies them.

Once the relations and entities are defined, they are connected into triples. The method of doing this is to look at the dependency tree from the dependency parser and see if an entity is linked to another entity going through a relation upwards in the dependency tree towards the root element. Algorithm 1 shows the step-by-step instructions of connecting triples. More formally, parts of the algorithm can be expressed with set theory. For an input sentence Sin=w1, w2, ..., wn where wk is a word in the sequence of words of the sentence and |Sin|=n. The set of all words is defined as W={Sin}. All elements in W will be considered as relations and entities in T= (wk, wl, wm) where T is an ordered triple information in the sentence Sin. The first position of T is the subject, the second position is the relation and the third position is the object constrained to 0<k<l<m<=n. The values of T are further limited to the condition wk ∈{E∩W}, wl∈{R∩W} and wm∈{E∩W} where the entity set E∈{Noun,Conjuction}and relation set R∈{Verb}. The subject and object words, wk and wm, should always be a children of the relation word wl. The child relations are denoted Tl= (wk,dl,wl) and Tr= (wm,dr,wl) for left and right entity respectively, where dl, dr ∈{subject,modification} is the dependency relation between a pair of words in Sin. The retrieval of T from Sin is shown in Algorithm 1. Algorithm 1 can also be used for other Germanic languages like Swedish, given that there are part-of-speech and dependency parser models of the language we want to process. Since the models of OpenIE and MinIE are trained on an English data set, information extraction on Swedish was not possible at this point intime with these algorithms. The dependency parser rules used in this research are formulated for Germanic language structures and therefore a worse accuracy can be expected for other types of languages.



Algorithm 1 Algorithm of connecting entities and relations together.

Input: Entities, Relations, Sentence dependency tree

Output: Array of triples consisting of (entity, relation, entity)

```
select potential relations R in S (nouns or words with a dependency relation parataxis)
declare list Lt
declare list Rt
for each r in R do
    El = select potential entities occuring left of R in S
    Er = select potential entities occuring right of R in S
    for each el in El do
        if el is child of r and d of el is one of nsubj, amod, advmod then
            add Tl (el, d, r) to Lt
        end if
    end for
    for each er in Er do
        if er is child of r and d of er is one of nsubj, amod, advmod then
            add Tr (er, d, r) to Rt
        end if
    end for
end for
return zip(Lt, Rt)
```

There are two special rules about the entities. The first rule is about the named entity recognition algorithm. This will identify multi-word entities from the NER model (for example "Twenty century fox":organization) and make sure they are intact by verifying that an entity does not span multiple entities. After the first rule the second rule is applied. The rule checks whenever a co-referenced word is found that this word will be replaced with the identified entity.

3.8. Evaluation

The GenerateIE algorithm is evaluated using one quantitative and one manual evaluation for com-parison of the algorithms. The quantitative part consists of a comparison of the number of triples found in each algorithm. The reason is to check whether any algorithm is missing any triples that the other algorithms pick up and also to check for errors when an algorithm detects an unreasonable amount of them. The manual evaluation consists of a randomly sampled set of 100 documents were selected reviewers label triples by hand as being correct or incorrect. The labelling is used to compute an accuracy for the selected data set. The order of the triples generated by different algorithms have been randomized, so that reviewers would not recognize which algorithm generated which triples. This would then avoid rating one algorithm more favourably than another. As mentioned in Section 2, there are previous studies on four methods of manual data-labelling to con-sider in our research. A laboratory evaluation (method 3) was chosen over the intrinsic evaluation (method 1). Intrinsic evaluation is a more complex problem since identifying triples without help requires more knowledge than saying weather an already identified triple is correct or incorrect (as done in laboratory evaluation). Since no open accessible data set was found, the real-world cases evaluation (method 4) was not an option. The extrinsic evaluation (method 2) would be possible and should be evaluated, although for news-media the type of triples would need to be verified manually to ensure that the entities and relations actually existed in the text. Furthermore, there are no QA-labels for news-media articles that would be used in an extrinsic evaluation and therefore the laboratory evaluation (method 3) in Section 2 was chosen to evaluate the GenerateIE algorithm.

The label-reviewers, consisting of the members in the research team of five people, went through the extracted triples of all algorithms in a random order and determined from the text context if the data sets were correct or incorrect. The instructions[1] were handed to the label-reviewers of the subjective evaluation before they started labelling the dataset. The instructions consisted of an explanation of how to identify an entity and a relation, how to connect the objects into a triple, and how the labelling tool worked in detail. The explanation of how a triple is tagged is as follows.

Entity:

1. An entity could be a name of a location, person or organization
2. An entity could be I, you, he, she it, they, etc. referring to some definition in 1)
3. An entity can consist of several semantically connected words, for example "green apple" or “Adam Andersson" (note that in a sentence like "I like the green apple", both "green apple “and "apple" are valid entities for the relation "like", whereas "green" is a valid entity for the triple "apple is green")

Relation:

1. A relation consists of at least one verb connected to an entity.
2. A relation that exists along an adverb should be concatenated together to form the relation, relations which do not concatenate the adverb is incorrect (for example "he is running fast through the forest", "running fast" should be the relation)

Triple:

1. A triple can consist of an entity, relation, entity
2. A triple could consist of an entity, relation, adjective describing the entity (for example "The ball is blue")

Once the data set is labelled the algorithms can be compared by accuracy, but since this is a new data set an evaluation baseline is introduced in order to sort out algorithms that perform below this baseline as unsuitable for the information extraction task. A completely random triple picker

[1] Instructions available here:
https://docs.google.com/document/d/1PLUToV2drUlCTIXmLYeIWjpcHzhnLj6krA5MamEB8d0/edit

which considers all words to be entities and relations would have an accuracy of approximately zero. Therefore, the baseline has to be a "reasonable smart" random model that would pair each noun with each verb and consider all combinations of these as triples. Assuming there are one verb and two nouns per sentence the random model would yield an accuracy as shown in Equation4, but since the sentences should strive to be between 16-25 words [29] the accuracy is drastically reduced. Let's say that the baseline picks 3 words randomly from sentences, the accuracy would converge towards one fraction of 16 in best case and one fraction of 25 in worst case, see Equation6. Any algorithm that would have an accuracy above these base lines would be considered a candidate for the triple extraction task.

$$baseline\_accuracy\_3\_words = \frac{1}{5} = 0.2 = 20\% \tag{4}$$

$$baseline\_accuracy\_16\_words = \frac{1}{16} = 0.0625 = 6.25\% \tag{5}$$

$$baseline\_accuracy\_25\_words = \frac{1}{25} = 0.04 = 4\% \tag{6}$$

4. RESULTS

GenerateIE is quantitatively evaluated, comparing the number of extracted information triples, as well as manually evaluated, where a randomly selected sample set of the documents are labelled correctly or incorrectly to estimate the accuracy. The results are presented in the subsequent subsections.

4.1. Quantitative Evaluation

In total, 10000 Wikipedia articles were used to produce the results. Figure 3 shows the number of unique relations recorded for OpenIE, GenerateIE and MinIE after each document. Figure 4shows the matching between each pair of algorithms. It seems like the number of triples is not converging over time, but there are slightly less triples added by each article converging from an exponential increase towards a linear increase of newly found triples. The GenerateIE seem to find less relations than MinIE although finds more triples than MinIE in total as shown in Figure5. Most of the matched triples in Figure 4 are new (the lines are linear) and because multiple algorithms found the same triples, it motivates a higher probability certainty that the triples are correctly extracted triples.

Figure 5 shows a Venn diagram of the extracted triples of the two algorithms. The number of extracted triples of all algorithms are 398,408 triples, whereas 228,245 triples are only extracted by OpenIE, 25,401 triples only by MinIE and 129,966 triples only by GenerateIE. A few thousands of triples are found by all three algorithms. Extracting triples from even less articles was also tested but yielded the same ratios as Figure 5.

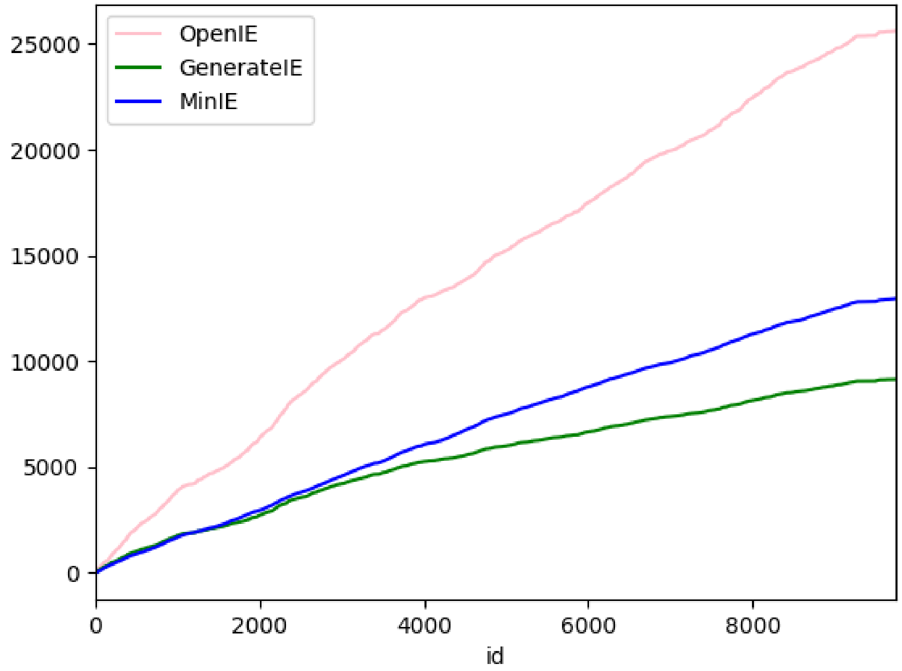


Figure 3. Total amount of unique relations extracted (y-axis) at a given iteration (x-axis)

4.2. Subjective Evaluation

The subjective results were produced by labelling the extracted information as correct or incorrect for each algorithm. In total 100 Wikipedia articles were labelled. Table 1 contains the numbers of triples identified by each algorithm together with how many of them where correct respectively incorrect. From these values, we compute the accuracy and standard deviation in the right-most column. The accuracy is the average number of correctly labelled triples shown in Equation 7. The reviewer's labels, correct and incorrect, could be represented in a stochastic variable X∈{0,1} where x=0 translates to incorrect label and x=1 translates to correct label. Since X could be either one or zero, the standard deviation uses the binomial distribution theorem about estimating the variance σ(X)2 in Equation 8 given ˆp from the previously computed accuracy in Equation 7.

$$\hat{p} = \frac{1}{n}\sum_{n} X \tag{7}$$

$$\sigma^2 = \frac{\hat{p}(1-\hat{p})}{n} \tag{8}$$

Figure 7 shows a bar plot of all extracted triples of the three algorithms and how many of those where correctly extracted. Figure 6 shows the accuracy of each algorithm and all possible combinations of them for the selected dataset. If the word linker is removed from GenerateIE

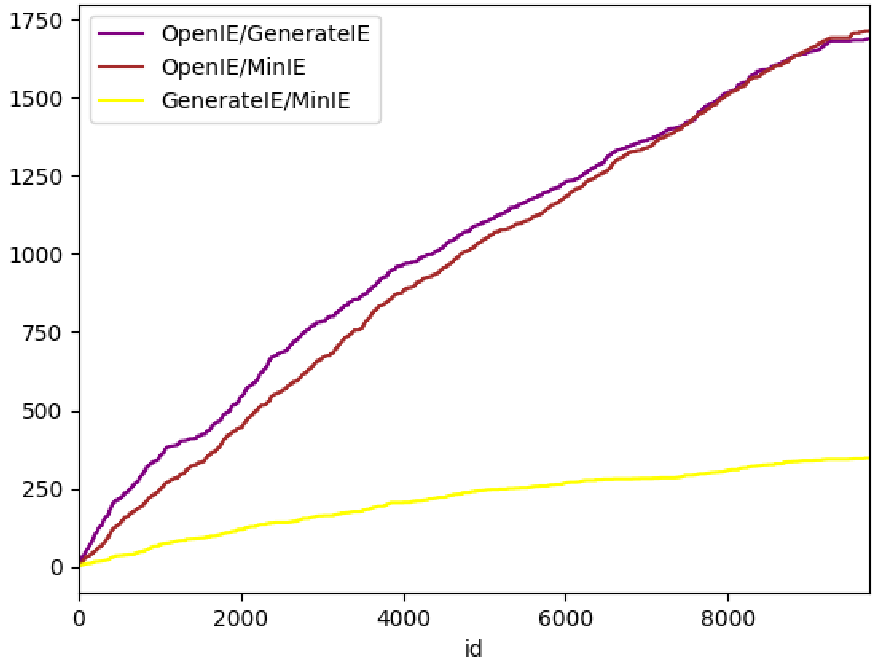


Figure 4. Total amount of unique relations extracted when matching pairs of algorithms (y-axis) at a given iteration (x-axis)

algorithm the accuracy is decreased to 31% while the standard deviation is also decreased to±2.3% as shown in Table 1. The total combination of the common triples of all algorithms are too small to evaluate further although computing the matched triples accuracy between OpenIE and GenerateIE yields83% as shown in Figure 6. The dataset is a subset of the quantitative evaluation with 100 articles, manually tagged triples with correct or incorrect. There are a total of 995 triples extracted by the algorithms, whereas 357 triples are only extracted by GenerateIE, 489 triples are only extracted by OpenIE and 51 only extracted by MinIE. There are a few overlaps where, a pair, or all algorithms have found the same triples. There are 98 correctly extracted triples by GenerateIE, 167 correctly extracted triples by OpenIE and 21 correctly extracted triples by MinIE. The overlap is slightly less when looking at the correctly extracted triples.

Based on this subjective evaluation, it seems like the OpenIE algorithm finds a lot of entities and relations that are not really classified as entities nor triples and it often suggests different mutations of the same entity which yields incorrect result, which may or may not be desired in an application using the algorithm. GenerateIE and MinIE are more conservative in suggesting triples and only suggest one triple combination for each identified subject and object pair. This impression is strengthened by Figure 3, where the line is steeper for OpenIE algorithm. The accuracy seems to be better in the MinIE and OpenIE algorithms compared to GenerateIE, but it is not significant at this point since it falls within the standard deviation interval.

Additionally, of the comparison a Swedish evaluation of the GenerateIE algorithm have been done. It has 79 triples less than the English evaluation of GenerateIE with an accuracy of 41% and a standard deviation of 2.8% as shown in Table 1. The GenerateIE algorithm evaluated on the Swedish dataset does have a high accuracy compared to the English dataset of the same algorithm.

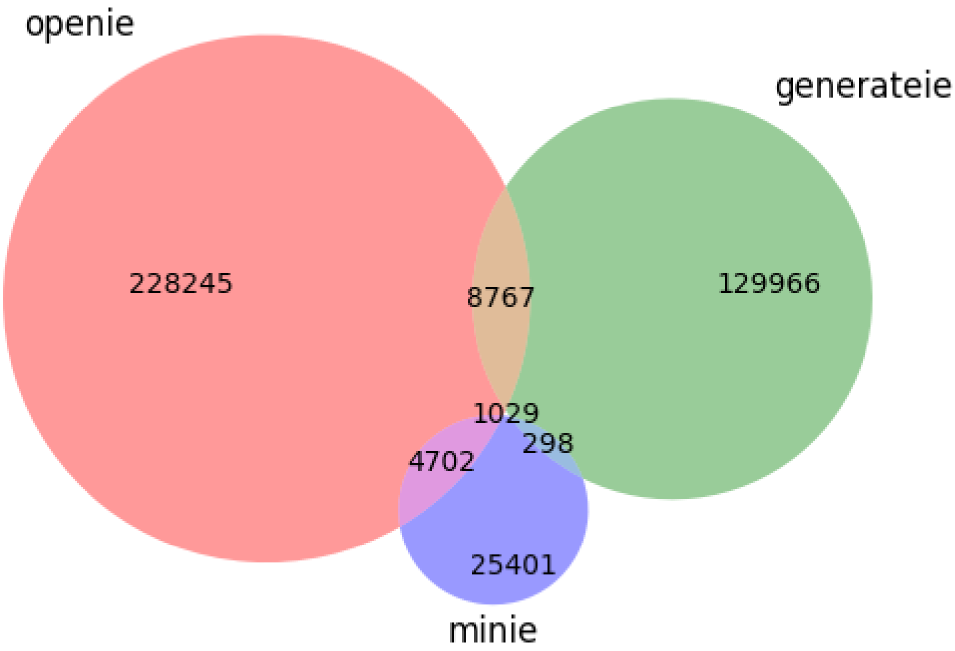


Figure 5. The triples extracted for the compared algorithms on 1000 articles

Table 1: The identified triples of OpenIE and GenerateIE and the result in accuracy and standard deviation when performing 10 cross fold validation

Source	Triples	Error	Correct	Accuracy	Std
All Combined	995	640	355	0.384	±0.016
OpenIE	533	333	200	0.375	±0.021
GenerateIE	394	272	122	0.310	±0.023
GenerateIE+WordLinker	394	272	122	0.378	±0.027
MinIE	68	35	33	0.485	±0.061
Overlap	3	0	3	1.0	0.0
GenerateIE (Swedish)	315	186	129	0.410	±0.028

5. DISCUSSION

After the evaluation there are several aspects that should be highlighted in all the algorithms. The overlapped triples are very few and thus we cannot say very much about a combined accuracy in this article but presumably when all algorithms extract the same triple, it is likely that it is correct. The data-points are relatively few when quantifying the numbers - 100 articles and 995 triples in total - but it is more than enough to compute stable accuracy and standard deviations. There is a slight significance that GenerateIE is performing better with the word linker than without, and that MinIE outperforms all other algorithms with the 68 triples that where been found in the dataset, although that few triples mean that it is missing a lot of triples on almost half of the articles. With the GenerateIE word linker there are no significance between

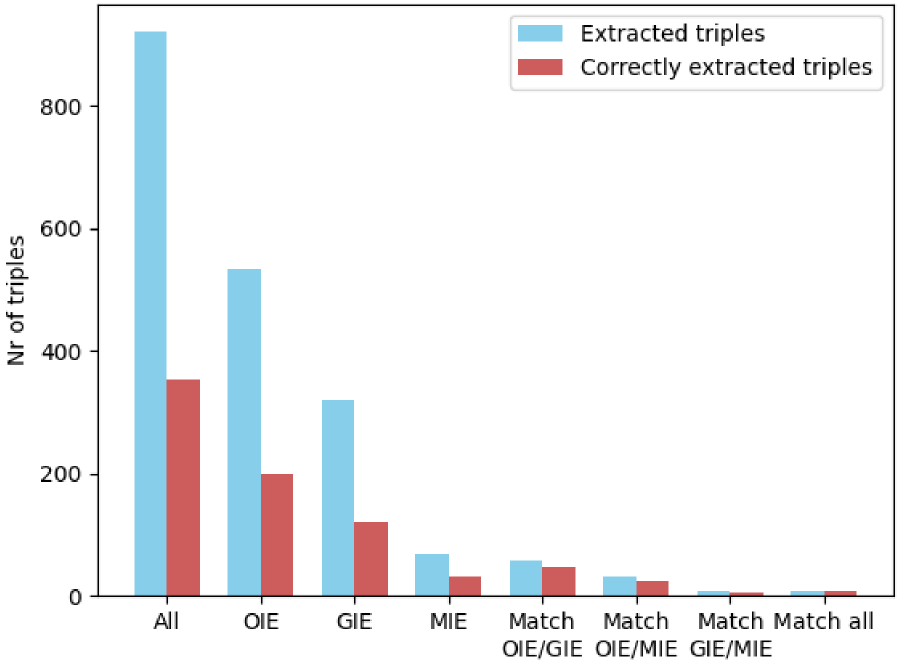


Figure 6. The number of triples found by each algorithm on 100 randomly selected articles from the entire dataset that was extracted and correctly extracted. To make room in the figure the naming was shortened (OIE=OpenIE, GIE=GenerateIE, MIE=MinIE)

OpenIE and GenerateIE. Furthermore, all algorithms are significantly higher accuracy than the baseline discussed in Section 3.8

The number of triples found by OpenIE is two times more than the GenerateIE algorithm in the quantitative evaluation the MinIE is even fewer. Only a small amount of the triples, 0.3%, are extracted by all three algorithms. The overlap is too small to give a just accuracy over all three algorithms, although adding common triples for OpenIE and GenerateIE we get an overlap of about3% and an accuracy of about 83%. The triples not found by any of the algorithms are unknown since we don't have a dataset tagged with triples in our evaluation. The lack of knowledge of total number of triples in a sentence might make the accuracy higher than it actually is e.g., the precision will be lower when including the unknown triples. Another important fact to consider when com-paring these algorithms is that the GenerateIE algorithm can handle multiple languages without any additional tweaks to the main algorithm as long as the part-of-speech and dependency parser exists for the language of the input text, the accuracy seem to be slightly higher of this evaluation even though the standard deviations overlap the OpenIE algorithm and the GenerateIE algorithm with word linker capabilities. It seems like the number of the unique found triples begin to in-crease linear but slightly decreases towards the end for both algorithms for each iteration of new documents we extract information from. The curve could be estimated as an O (n log n) function.

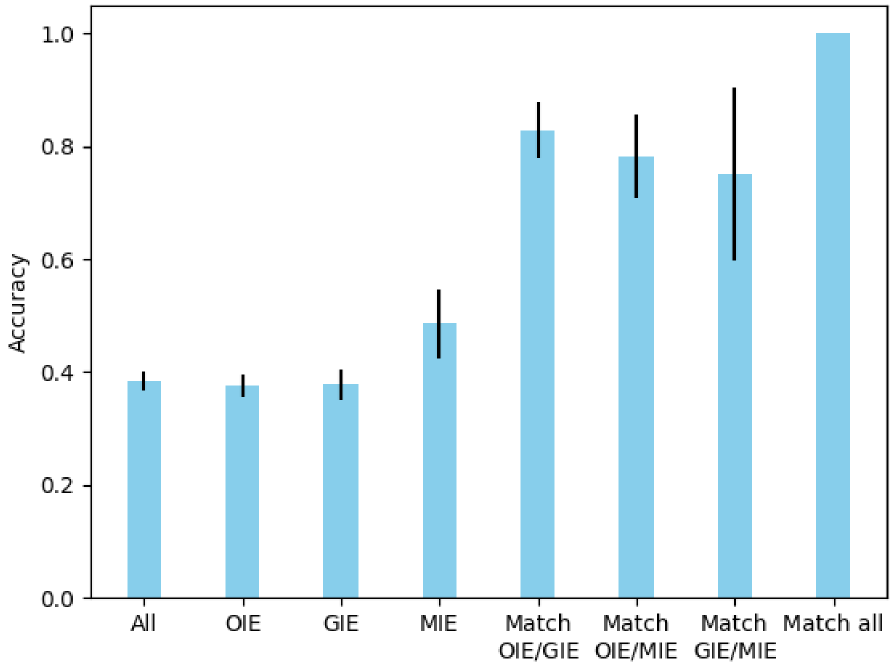


Figure 7. The accuracy of each algorithm on 100 randomly selected articles from the entire dataset. The black lines indicate the standard deviation of each evaluation. To make room in the figure the naming was shortened (OIE=OpenIE, GIE=GenerateIE, MIE=MinIE)

6. Conclusion

The goal of this research was to create the algorithm, GenerateIE, that combines existing algorithms to extract entity-relation-entity triples from plain texts to summarize the gist of it. The extraction can be done with multiple Germanic languages for texts in the news-media domain. A second goal was to combine a set of state-of-art algorithms to enhance the information extraction process. The GenerateIE algorithm can extract triples that will with 36% probability represent a partial gist of a sentence or paragraph. The algorithm has been compared and evaluated with two other algorithms, OpenIE and MinIE, which have the same goal. This evaluation shows that the GenerateIE algorithm performs slightly worse than the others in terms of the number of identified triplets, but in terms of accuracy it performs better than MinIE and worse than OpenIE. Additionally, GenerateIE gives the possibility to transfer the rules to different Germanic languages and thereby also make it possible to use in a multi-language approach. Even though the MinIE and OpenIE algorithms does not support Swedish language the evaluation of the GenerateIE algorithm on Swedish language shows that the accuracy is even higher than its English accuracy on a similar dataset. News-media companies will be able to use this algorithm to further analyse their con-tent. Further also tell what the readers are interested in reading on a much more detailed level than before for both Swedish and English texts. The impact of this work will also affect other domains within NLP such as how one can approach summarizing of text as well as deriving more knowledge from additional languages.

Regarding future work, a continued study is required to tweak the parameters of the algorithm to increase the number of identified triplets and the accuracy. An intrinsic evaluation should be performed to confirm the result in this paper for other languages than English and Swedish. The significance of non-found triples should be further evaluated because the evaluation in this research only reflects the accuracy of the found triples leaving an uncertainty in the false negatives affecting the recall score. It needs further to be determined how this research can be applied to different domains other than the news-media domain.

REFERENCES

[1] X. Han, S. Cao, X. Lv, Y. Lin, Z. Liu, M. Sun, and J. Li, "Openke: An open toolkit forknowledge embedding," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 139–144, 2018.

[2] A. Roy, Y. Park, T. Lee, and S. Pan, "Supervising unsupervised open information extraction models," 2020.

[3] G. Stanovsky, J. Michael, L. Zettlemoyer, and I. Dagan, "Supervised open information ex-traction," vol. 1, 2018.

[4] F. Wu and D. S. Weld, "Open information extraction using Wikipedia," in Proceedings of the48th annual meeting of the association for computational linguistics, pp. 118–127, Association for Computational Linguistics, 2010.

[5] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in 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), pp. 344–354, 2015.

[6] D. B. Claro, M. Souza, C. Castellã Xavier, and L. Oliveira, "Multilingual open information extraction: Challenges and opportunities, "Information, vol. 10, no. 7, p. 228, 2019.

[7] K. Gashteovski, R. Gemulla, and L. del Corro, "MinIE: Minimizing facts in open information extraction," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, (Copenhagen, Denmark), pp. 2630–2640, Association for Computational Linguistics, Sept. 2017.

[8] L. Del Corro and R. Gemulla, "Clausie: Clause-based open information extraction," in Proceedings of the 22nd International Conference on World Wide Web, WWW '13, (New York, NY, USA), p. 355–366, Association for Computing Machinery, 2013.

[9] K. Clark and C. D. Manning, "Deep reinforcement learning for mention-ranking coreference models," in Empirical Methods on Natural Language Processing, 2016.

[10] S. Narayan and C. Gardent, "Unsupervised sentence simplification using deep semantics," arXiv preprint arXiv:1507.08452, 2015.

[11] A. Clark, C. Fox, and S. Lappin, The handbook of computational linguistics and natural language processing. John Wiley & Sons, 2013.

[12] P. Paroubek, S. Chaudiron, and L. Hirschman, "Principles of Evaluation in Natural Language Processing," Traitement Automatique des Langues, vol. 48, pp. 7–31, May 2007.

[13] J. G. Smith and R. Tissing, "Using computational text classification for qualitative research and evaluation in extension., "Journal of Extension, vol. 56, no. 2, p. n2, 2018.

[14] K. Leetaru and P. A. Schrodt, "Gdelt: Global data on events, location, and tone, 1979–2012," in ISA Annual Convention, vol. 2, Citeseer, 2013.

15] M. D. Ward, A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford, "Comparing GDELT and ICEWS event data, "Analysis, vol. 21, pp. 267–297, 2013.

[16] D. Zeman, M. Potthast, M. Straka, M. Popel, T. Dozat, P. Qi, C. Manning, T. Shi, F. G. Wu,X. Chen, Y. Cheng, A. Björkelund, A. Falenska, X. Yu, J. Kuhn, W. Che, J. Guo, Y. Wang,B. Zheng, H. Zhao, Y. Liu, D. Teng, T. Liu, K. Lim, T. Poibeau, M. Sato, H. Manabe, H. Noji,Y. Matsumoto, Ö. Kırnap, B. F. Önder, D. Yuret, J. Straková, C. Vania, X. Zhang, A. Lopez,J. Heinecke, M. Asadullah, J. Kanerva, J. Luotolahti, F. Ginter, Y. Kuan, P. Sofroniev,E. Schill, E. Hinrichs, D. Q. Nguyen, M. Dras, M. Johnson, X. Qian, Y. Liu, D. Vilares,C. Gómez-Rodríguez, L. Aufrant, G. Wisniewski, F. Yvon, S. D. Dumitrescu, T. Boro ̦s,D. Tufi ̦s, A. Das, A. Zaffar, S. Sarkar, H. Wang, H. Zhao, Z. Zhang, R. Hornby, C. Tay-lor, J. Park, M. de Lhoneux, Y. Shao, A. Basirat, E. Kiperwasser, S. Stymne, Y. Gold-berg, J. Nivre, B. K. Akku ̦s, H. Azizoglu, R. Cakici, C. Moor, P. Merlo, J. Henderson,H. Wang, T. Ji, Y. Wu, M. Lan, E. de la Clergerie, B. Sagot, D. Seddah, A. More, R. Tsarfaty,H. Kanayama, M. Muraoka, K. Yoshikawa, M. Garcia, and P. Gamallo, "CoNLL 2017

sharedtask system outputs," 2017. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[17] J. Nilsson and J. Hall, Reconstruction of the Swedish Treebank Talbanken. Matematiska och systemtekniska institutionen, 2005.

[18] J. Einarsson, "Talbankens talspråkskonkordans," Proc. of LREC, 1976.

[19] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents.," in ICML, vol. 14, pp. 1188–1196, 2014.

[20] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in Proceedings of the 43rd annual meeting on association for computational linguistics, pp. 363–370, Association for Computational Linguistics, 2005.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[22] A. Voutilainen, "Part-of-speech tagging, "The Oxford handbook of computational linguistics, pp. 219–232, 2003.

[23] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," arXiv preprint arXiv:1607.01759, 2016.

[24] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally normalized transition-based neural networks," arXiv preprint arXiv:1603.06042,2016.

[25] C. Alberti, D. Andor, I. Bogatyy, M. Collins, D. Gillick, L. Kong, T. Koo, J. Ma, M. Omer-nick, S. Petrov, et al., "Syntaxnet models for the conll 2017 shared task," arXiv preprintarXiv:1703.04929, 2017.

[26] J. Einarsson, "Projektet talbanken. i: C platzack (utg), svenskans beskrivning 8, s76-96,"1974.

[27] S. Schuster and C. D. Manning, "Enhanced english universal dependencies: An improved representation for natural language understanding tasks," in Proceedings of the Tenth Inter-national Conference on Language Resources and Evaluation (LREC'16), pp. 2371–2378,2016.

[28] C. Bosco, S. Montemagni, A. Mazzei, V. Lombardo, A. Lenci, L. Lesmo, G. Attardi, M. Simi, A. Lavelli, J. Hall, et al., "Comparing the influence of different treebank annotationson dependency parsing," in Seventh conference on International Language Resources and Evaluation (LREC'10), pp. 1794–1801, European Language Resources Association (ELRA),2010.

[29] B. B. Kadayat and E. Eika, "Impact of sentence length on the readability of web for screen reader users," in Universal Access in Human-Computer Interaction. Design Approaches and Supporting Technologies (M. Antona and C. Stephanidis, eds.), (Cham), pp. 261–271, Springer International Publishing, 2020.

Intelligent Question Answering Module for Product Manuals

Abinaya Govindan, Gyan Ranjan, and Amit Verma

Neuron7.ai, USA

Abstract. Question Answering (QA) has been a well-researched NLP problem over the past few years. The ability for users to query through information content that is available in a range of formats - organized and unstructured - has become a requirement. This paper proposes to untangle factoid question answering targeting the Hi-Tech domain. This paper addresses issues faced during document question answering, such as document parsing, indexing and retrieval (identifying the relevant documents) as well as machine comprehension (extract spans of correct answers from the context). Our suggested solution provides a comprehensive pipeline comprised of document ingestion modules that handle a wide range of unstructured data across various sections of the document, such as textual, images, and tabular content. Our studies on a variety of "real-world" and domain-specific datasets show how current fine-tuned models are insufficient for this challenging task, and how our proposed pipeline is an effective alternative.

Keywords: machine comprehension, document parser, question answering, information retrieval

1 Introduction

This study examines the challenge of factoid question-answering in a constrained situation, such as the Hi-tech domain, with several product manuals as the data source, where an agent attempts to discover a technical response to a client query by perusing the manuals.With multiple editions of a product, product manuals can be regarded as a constantly evolving source of information. Unlike a knowledge base with a structured source of information like FAQs, which are easier for computers to comprehend and process but may not be exhaustive and require time and manual effort to create and validate, manuals provide a more reliable and up-to-date source that becomes a perfect solution for a long-term and scalable system. These manuals, on the other hand, are intended for humans to decipher rather than machines, making automatic parsing more challenging.

This system intends to reduce the amount of time and effort required to find the most relevant manual that answers the user's question and locate the suitable section with the most appropriate answer by manual intervention.As a result, for any user query, the system returns a set of relevant sections from numerous manuals, rather than just the top one as standard question answering systems do. This decision is based on a business inference that multiple manuals may include information relevant to a user query. A typical case is when a user asks, ***What is the expected***

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 19-32, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112002

time for my battery to be fully charged? and the answer can be found in the manuals for numerous devices, so all of those sections must be recommended to the user. The system also needs to be able to comprehend any additional context provided by the user that aids in narrowing down the manuals - for example, if the user asks ***What is the expected time for device A's battery to be charged?***, the system should recognise that ***device A*** is an additional context and should be able to only look through manuals for ***device A***.

The use of product manuals for question answering involves the inclusion of a document indexer engine in the question answering system, which should be executed at scale because the questions should be addressed in real-time. As a result, the system should be able to retrieve relevant sections of instructions among hundreds of acquired manuals for each user query. Because it can process both textual (paragraphs, summaries, etc.) and non-textual (tables, images, flow diagrams, etc.) information, this system can be extended to any domain as long as manuals, documents, or even books and articles are available.

In traditional question answering scenarios, a small chunk of text can be regarded as an answer to the question asked. We cannot, however, make the same argument for our business use case. If a user inquires about ***What are the steps for me to log in to a device?***, the response cannot be provided in a short segment and must be replied using an entire section titled ***How to use and set up?***. As a result, the system should be able to decide whether the answer should be returned as a short sequence or as a portion of text in real-time.

In this paper, we show how various existing systems try to solve this domain based question answering by comparing their performances on standard business dataset. We also introduce Intelligent Question Answering system which is composed of

- **Document parser**, a transformer-based deep learning model that can parse and manage a wide range of unstructured data, including images, tables, textual content etc.
- **Document indexer**, a module that uses indexed databases to index documents with essential information to keep all different types of data in a single collection, such as images, tables, and so on.
- **Document Retriever**, natural-language-based query processor that handles several business-specific preparation processes, recognizes if the query has any "context," and gets the top relevant chunks of text from the indexed database.
- **Document Reader**, a multilayer transformer-based model which has been fine-tuned for the task of specific domain-based question answering. The document reader also has a classifier that has been trained to decide if the answer should be a small segment or section of text.

In this paper, we study the application of several deep learning models to the question answering task. Our experiments show that the Intelligent Question Answering system outperforms traditional question answering systems on standard business-specific datasets.

1.1 Related work

The first type of question answering that the research teams concentrated on was factoid questions, which are questions for which the answers can be retrieved with certainty from the specified text source.However, in real-world scenarios, there may be a few exceptions to this assumption that can be handled by various classifier modules. Factoid questions such as ***"Where was X born?"***, ***"Which year did Y take place?"*** were the target area for these teams. Now, the focus is on answering complicated problems like ***"How can Y be done?"***, ***"A was moved from B to C and later to D. Where is A now?"***. In some circumstances, these questions require complicated comprehension and inference of contexts, as well as information flow between sentences. Simple comprehension models or named entity models can no longer answer these problems. Starting in 1999, an annual evaluation track of question answering systems has been held at the Text Retrieval Conference (TREC) (Voorhees 2001, 2003b). Following the success of TREC, in 2002 both CLEF and NTCIR workshops started multilingual and cross-lingual QA tracks, focusing on European languages and Asian languages respectively (Magnini et al. 2006; Yutaka Sasaki and Lin 2005 [8]). Other datasets that focused on question answering such as P. Rajpurkar,et al. 2016 [11] and P. Rajpurkar, et al. 2018 [10] were also released. The amount of literature in the general field of QA has grown to the point where there are numerous models with significant performance that reliably address the QA domain. The majority of these models and techniques, on the other hand, concentrate on academic data sources, which have been curated by humans and adhere to excellent grammar and linguistic patterns. In real business world, data is frequently dispersed over pages or portions of pages, causing parsing and further inference of this type of data to perform far worse than in the academic environment. There are also a number of advanced complete pipeline QA systems that leverage either the Web, as does QuASE (Sun et al., 2015) [13], or Wikipedia as a resource, as do Microsoft's AskMSR (Brill et al., 2002) [14], IBM's DeepQA (Ferrucci et al., 2010) [20] and YodaQA (Baudiˇs, 2015; Baudiˇs and ˇ Sediv'y, 2015) [15]. AskMSR is a search-engine-based QA system that focuses on "data redundancy rather than sophisticated language analyses of either questions or potential responses," in other words, it doesn't focus on machine comprehension like we do. Few approaches attempt to address both unstructured and structured information, such as text segments and documents, as well as knowledge bases and databases.One such example is DeepQA. Other systems such as YodaQA which are based after DeepQA combines information extractiom from unstructured sources such as websites, text and Wikipedia. This task is challenging because researchers have to face issues in both scalability and accuracy. In the last few years, rapid progress has been made and the performance of factoid and open-domain QA systems has been improved significantly (Chen et al., 2017;S.Schwager et al., 2019;K. Jiang et al., 2019). Several different approaches were proposed, including twostage ranker-reader systems such as DrQA (Chen et al., 2017) [2], end-to-end transformer based models (S.Schwager et al., 2019) [4] and unified framework based models to solve all text based language problems (Raffel et al., 2020) [7].

2 Our proposal

In the following, we describe our system - Intelligent Question Answering Pipeline which consists of four components: (1) Document Parser module (2) Document Indexer module (3) Document Retriever (4) Document Reader. The whole architecture is depicted in 3.

2.1 Document parser

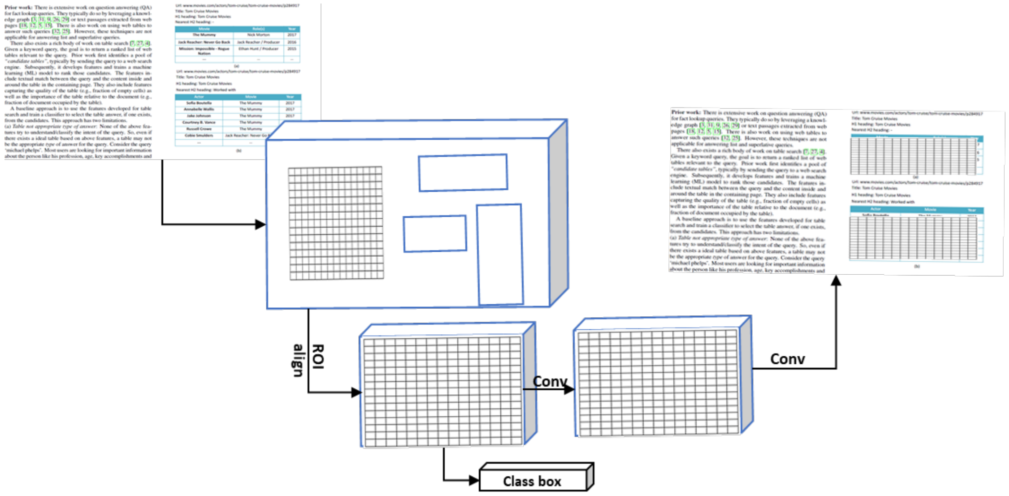


Fig. 1. Mask RCNN framework for instance segmentation

The document parser is the input processing block which is responsible for reading the data and processing it into a format that can be handled by the subsequent modules. The main parts of the document parser are the Mask RCNN based fine tuned instance segmentation model which is fine tuned to identify tables and images with text present in the document. The Mask R-CNN extends Faster R-CNN by adding a branch for predicting segmentation masks on each Region of Interest (RoI), in parallel with the existing branch for classification and bounding box regression . The architecture of a Mask RCNN is as mentioned in 1 and 2

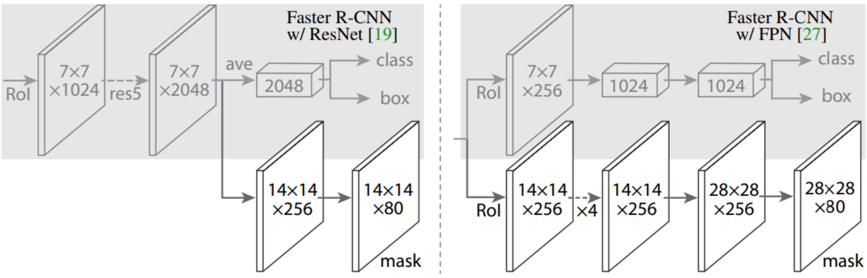


Fig. 2. Head architecture of Faster R-CNN

Mask R-CNN is conceptually simple: Faster R-CNN [17] has two outputs for each candidate object, a class label and a bounding-box offset; to this we add a third branch that outputs the object mask. Mask R-CNN is thus a natural and intuitive idea. But the additional mask output is distinct from the class and box outputs, requiring extraction of much finer spatial layout of an object. Mask RCNN adopts image centric training and hence the images are resized such that their scale is 800 pixels. Formally, during training, a multi-task loss on each sampled RoI is defined as

$$L = L_{cls} + L_{box} + L_{mask}$$

The classification loss $\mathrm{L_{cls}}$ is defined as

$$L_{cls}(p, u) = -logp_u$$

which is the log loss for the true class u and bounding-box loss L_{box} is defined as

$$L_{box}(t^u, v) = \sum_{i\epsilon\{x,y,w,h\}} smooth_{L1}(t_i^u - v_i)$$

where

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases}$$

The mask branch has a Km^2- dimensional output for each RoI, which encodes K binary masks of resolution $m \times m$, one for each of the K classes. To this we apply a per-pixel sigmoid, and define $\mathrm{L_{mask}}$ as the average binary cross-entropy loss. For an RoI associated with ground-truth class k, $\mathrm{L_{mask}}$ is only defined on the k^{th} mask (other mask outputs do not contribute to the loss).

The mask RCNN [18] based object detector is fine tuned as depicted in 1. The following modules form the main stages in the Document Parser :

- The RCNN model has been fine tuned to identify two main objects - tables and images with text caption.
- The Document parser then makes use of the fine tuned model to categorize input document into three classes - tables, images with text caption and paragraph sections
- All the three sections of the document are then stored in appropriate databases based on the detected object.

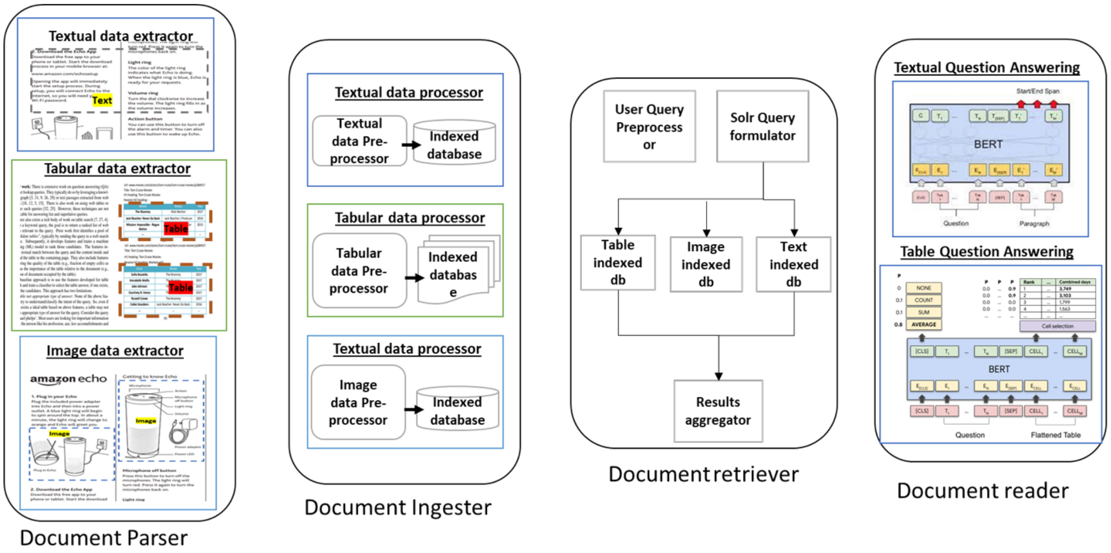


Fig. 3. Detailed overview of Intelligent Question Answering for Product manuals

2.2 Document indexer

The indexer is used to index and store the parsed data into indexed databases and structured tables. During index time, these sections are indexed with necessary indicator/metadata attributes which can later be mapped to the object class such as table, text etc. For the sake of indexed databases, we chose a Lucene based indexer after considering various business parameters such as volume of data, speed of indexing and speed of retrieval during query time. The document parsing and indexing happen at a batch level and triggers have been set to initiate the process if and when new documents get added to the source repository.

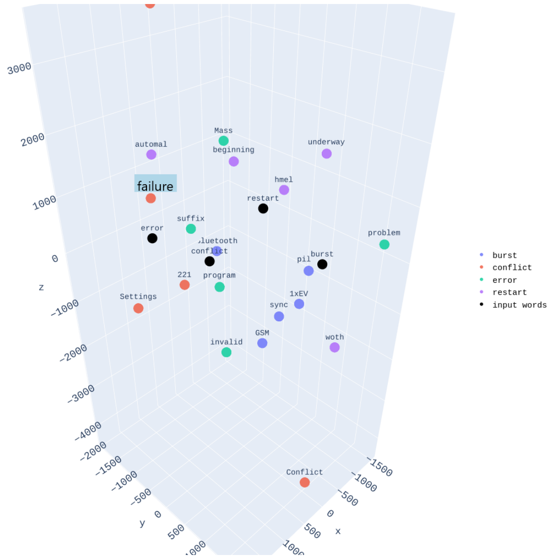


Fig. 4. Fine tuned word embedding vector space

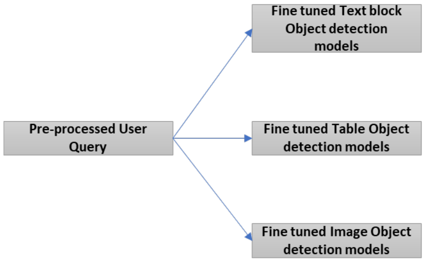

Pre-processed User Query
Fine tuned Text block Object detection models
Fine tuned Table Object detection models
Fine tuned Image Object detection models

Fig. 5. Fine tuned models to parse user query

2.3 Document retriever

Once the documents have been indexed in a batched fashion, the retriever and reader take care of the real time query process. Despite having thousands of documents, which span across hundreds of pages, we have the ability to query through them real-time due to the fast information retrieval that is enabled by the document retriever. The retriever is composed of the following modules -

Contextualised synonyms extractor For the retriever to have semantic abilities during query stage, we leverage GloVe word embeddings [21] which have been fine tuned on our business data. With brevity of this paper in mind, most of the details of these models have not been discussed, aside from the the fact that they attempt to maximize the log probability as a context window scans over the corpus. Training proceeds in an online, stochastic fashion, but the implied global objective function can be written as,

$$J = -\sum_{\substack{i \,\epsilon\, corpus \\ j \,\epsilon\, corpus_i}} log Q_{ij}$$

This results in word embedding that look like 4 in the higher dimensional vector space.

Utilising the fine tuned word embedding, we cluster the words using their embedding to group semantically similar words together and hence leading to a set of contextual synonyms which helps in making the query retrieval semantic and contextualised.

Metadata and Context extractor The next stage of document retrieval is a context extractor which is a named entity model which has been trained on metadata such as product family, product line and model name. This named entity model is trained using a variant of BERT [?] for the task of single sentence tagging. The architecture of this model is depicted in 6. Once the entities has been extracted from the user query, the metadata information is used to further refine the retrieval by extracting information only from documents with corresponding metadata.

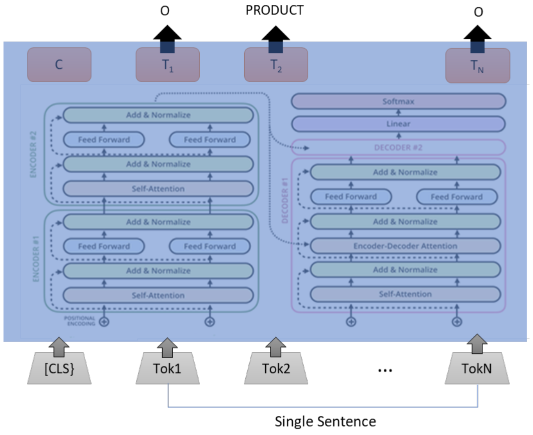


Fig. 6. Architecture for Named entity model fine tuning

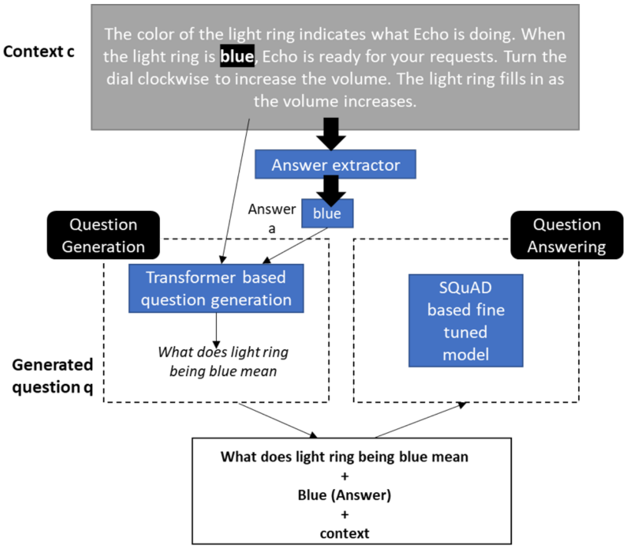

Context c
The color of the light ring indicates what Echo is doing. When the light ring is blue, Echo is ready for your requests. Turn the dial clockwise to increase the volume. The light ring fills in as the volume increases.
Answer extractor
Answer a
blue
Question Generation
Question Answering
Transformer based question generation
What does light ring being blue mean
SQuAD based fine tuned model
Generated question q
What does light ring being blue mean
+
Blue (Answer)
+
context

Fig. 7. Training data preparation for Question Answering

Tf-idf based retriever and collator The pre-processed query is then processed using a Lucene based indexed Query processor which encodes the query into a Lucene index specific format and retrieves n most relevant documents with their IDs, Tf-Idf relevance score and metadata information. The user query is passed in parallel to all the object classes as mentioned in 5.

The collated document results are then passed through our BM25 based similarity scorer that has been used to re-prioritise the retrieved results as follows:

$$fScore(D,Q) = \frac{w1 * (Tf(D,Q) * Idf(D,Q)) + w2 * score(D,Q)}{w1 \ + \ w2}$$

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i,D).(k1+1)}{f(q_i,D) + k1.(1-b+b.\frac{|D|}{avgdl})}$$

where

$$IDF(q_i) = ln(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1)$$

2.4 Document reader

Once the results have been retrieved and aggregated, the document reader is responsible to product the final consumable user results using the following models:

Fine tuned textual and Image question answering model Despite the vast availability of pre-trained question answering models that can answer generic questions, real-life questions often tend to yield less than good performance on these models. For this use case, we created question answering training dataset in an unsupervised fashion (like Cloze translation [16]) which was later than used to fine-tune pre-trained BERT [5] models as depcited in 7.

One difference from standard SQuAD scenario and our business use case is that the questions we might encounter need more descriptive answers. The questions relevant to our business case might be a *What* or *Why* or *How* question such as ***How shall I switch my phone off?*** or ***What is the meaning of error X?***. The first case might need a paragraph as answer whereas the latter needs a short segment of the paragraph to be returned as the answer. A standard Naive Bayes based classifier is used to classify incoming question into one of these classes and decision is taken accordingly if the answer to the question should be a short answer or a long answer.

This model is also used to handle images with text captions. The Image based question answering model includes a OCR parser that extracts textual information from the image segments of the document. We then use the fine tuned question answering model to extract relevant answer segments.

Tabular question answering model The second part of the Data Reader utilises fine-tuned models for table question answering based on the BERT architecture. This approach is based on TableQnA by K. Chakrabarti et al. [2] This module includes identification of the right table from set of tables and using the TableQnA model to identify the cell which can be the answer. The model also has a set of table handling and parsing algorithms which transform various tabular formats to the format desired by TableQnA and a post processor that shall return the answer in a format consumable by the user.

3 Experiments and data

For the scope of comparison, we had considered a comparative analysis between two standard approaches and our pipeline.

3.1 Standard Tf-Idf based retrieval based approach

Information retrieval as a domain has been considered as a search engine based task where retrieval from indexed databases takes us to the final solution. A simple inverted index lookup followed by term vector model scoring performs quite well on this task for many question types. We indexed all the subsections of the manuals into Lucene based indexed database with minimal contextualisation. We then passed the test queries as database queries to the database and extracted the top relevant section of manual as the answer.

3.2 SQuAD based retrieval approaches

The second approach that we wanted to compare was traditional SQuAD based models [4] to see how they fare in real life use cases. The SQuAD model has been always tested on very small text block and hence when we use this for larger pieces of text such as Wikipedia or documents, they often fail in terms of both run time and performance by failing to capture the right section of the manual.

3.3 Comparison of Average run time

The average run time taken by the methods on a test size of 50 user questions are as follows :

Table 1. Average run time (ms)

Tf-Idf based	**SQuAD based**	**Our approach**
100	300000	150

As we can see here, since the problem is to return the response in real-time, the SQuAD approach makes it impractical for the user to wait for nearly 5 minutes for each question for an answer to be produced.

3.4 Comparison of Performance metrics

For this exercise, we decided on metrics that would indicate both lexical and semantic closeness of the predicted answer to the actual answer. The first metric used was Rouge score which was defined as

$$ROUGE_n = \frac{\sum_{S\epsilon\{Refs\}} \sum_{ngram\varepsilon S} count_{match}(ngram)}{\sum_{S\epsilon\{Refs\}} \sum_{ngram\varepsilon S} count(ngram)}$$

where

$$count_{match}(ngram) = n(A \cap B)$$

A token t is considered to be common between A and B if the semantic similarity between t and at least one token in sequence B is greater than a pre-defined threshold. We use $ROUGE_1$ and $ROUGE_2$.

The second metric is the Overlap similarity which aims to capture the closeness between the expected and predicted answer. This is defined as

$$score(S1, S2) = \frac{\sum_{t1 \epsilon S1} \sum_{t2 \epsilon S2} \begin{cases} 1 \ if \ similarity(t1, t2) > t \\ 0 \ otherwise \end{cases}}{\sum_{t1 \epsilon S1} 1}$$

This metric represents the fraction of tokens in expected answer S1 which is semantically similar with S2. Semantic similarity between two vectors A and B is measured as

$$similarity(A, B) = \frac{\sum_{i=1}^{n} A_i . B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

The performance numbers are as below:

Table 2. Performance comparison

Approach	**Metric**	**Performance**
Tf-Idf based approach	***Rouge1 - F score***	0.0838
	Rouge2 - F score	0.0514
	Overlap measure	0.2245
SQuAD based approach	***Rouge1 - F score***	0.0609
	Rouge2 - F score	0.0224
	Overlap measure	0.2041
Intelligent QnA Pipeline	***Rouge1 - F score***	0.2463
	Rouge2 - F score	0.2094
	Overlap measure	0.5918

The numbers that we see here is when only the first prediction was considered. The numbers increase significantly when even top 3 results were considered for the comparison and was accepted for the business case.

3.5 Inference

While the Tf-IDF approach worked well for straightforward cases, the returned answers were too long to be consumed by the end user. In most of these cases, the answers were present only in a small subsection of these manuals and hence going through long passages to locate the answer seems to be impractical in a business scenario. Also, since there were many sections with same keywords, the answer was most often not present in the first returned result and was present somewhere deep down. The SQuAD based results had two main disadvantages - run time and quality of results. The model often failed to return the right answer both for straight forward and complex questions. This proves that even though traditional methods can perform well on a certain scenario, the smaller business specific intricacies that has been fed to our pipeline has proven effective in capturing the right answer in the shortest possible time.

4 Conclusion

In this paper, we proposed a novel pipeline of question answering based on structured and unstructured documents such as manuals, images and product user guides. We have also demonstrated with conclusive evidence, how to overlay bespoke domain knowledge on top of current and traditional systems to deliver contextualised outputs for a certain business domain. Our system has also been applied in a number of business fields, assisting users in quickly identifying appropriate solutions to their problems.

One limitation we discovered is that the existing approach does not allow for the inclusion of human feedback for various sub modules. We intend to do so as part of our pipeline enhancement efforts, as outlined in Future Work.

5 Future Work

There is a simple layer of feedback in the existing system that is utilised to improve the final answer generated by the automated pipeline.

As part of our future work, user signals such as feedback and additional annotated data for new labels will be incorporated. These signals will be plugged into various pipeline submodules, resulting in better performance of individual components. Designing a framework capable of consuming such feedback, as proposed by G. Abinaya et al. citeb22, is one area of improvement. The aforementioned structure will comprise dedicated modules that determine which section of the pipeline the feedback should flow into, the cadence with which the feedback should be reflected in the module, and the weights that should be assigned to feedback based on its eminence, user previliges, among other things.

We prioritised the question answering module, which was trained for the Hi-Tech domain, for the purposes of this paper. The incorporation of domain knowledge and terminologies in the named entity recognizer module will be another area of focus, starting with the generation of domain specific data using unstructured approaches and then building named entity classifiers utilising this data.

References

1. K. Jiang,D. Wu, and H. Jiang, “FreebaseQA: A New Factoid QA Data Set Matching Trivia-Style Question-Answer Pairs with Freebase,”Proceedings of NAACL-HLT 2019 pages 318–323 Minneapolis, Minnesota, June 2 - June 7, 2019.
2. K. Chakrabarti,Z. Chen, S. Shakeri, G. Cao, and S. Chaudhuri “TableQnA: Answering ListIntent Queries With Web Tables,’Proceedings of the VLDB Endowment, Vol. 12, 2020.
3. C. Deng, G. Zeng, Z. Cai, and X. Xiao, “ A Survey of Knowledge Based Question Answering with Deep Learning,” Journal on Artificial Intelligence 2020, No. 4 , pages 157-166
4. S. Schwager and J. Solitario,“Question and Answering on SQuAD 2.0: BERT Is All You Need,” ArXiv e-prints of 2019.
5. J. Devlin , M.W. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” ArXiv e-prints of 2018.
6. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I.Polosukhin “Attention Is All You Need,” 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
7. C. Raffel , N. Shazeer, A. Roberts, K. Lee , S. Narang , M. Matena , Y. Zhou, W. Li and P. J. Liu, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” ArXiv e-prints of 2020
8. Y. Sasaki, Y.Chen, K. Chen and C.J. Lin, “Overview of the NTCIR-5 Cross-Lingual Question Answering Task,” Proceedings of NTCIR-5 Workshop Meeting, 2005, Tokyo, Japan
9. D. A. Chen, J. W. Fisch, and A. Bordes. “ Reading Wikipedia to Answer Open-Domain Questions” ArXiv e-prints, 2017.
10. P. Rajpurkar, R. Jia, P. Liang. “Know What You Don’t Know: Unanswerable Questions for SQuAD.” ArXiv:1806.038221v1, 2018.
11. P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang. “ SQuAD: 100,000+ Questions for Machine Comprehension of Text,” arXiv:1606.05250v3, 2016.
12. W. T. Yih, X. D. He and C. Meek, “Semantic parsing for single-relation question answering,” in Proc. of the 52nd Annual MTG of the Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 643–648, 2014.
13. H. Sun, H. Ma, W. Yih, C. Tsai, J. Liu, and M.i Chang. “ Open domain question answering via semantic enrichment,” Proceedings of the 24th International Conference on World Wide Web. ACM 2015, pages 1045–1055.
14. E. Brill, S. Dumais, and M. Banko. “An analysis of the AskMSR question-answering system,” Empirical Methods in Natural Language Processing (EMNLP). pages 257–264.’
15. P. Baudi˘s and J. ˘ Sediv‘y. “Modeling of the question answering task in the YodaQA system,” International Conference of the Cross- Language Evaluation Forum for European Languages. Springer 2015, pages 222–228.
16. P. Lewis, L.Denoyer and S.Riedel. “Unsupervised Question Answering by Cloze Translation,” ArXiv e-prints of 2019
17. R. Girshick. “Fast R-CNN,” ArXiv e-prints of 2015
18. K. He, G. Gkioxari , P. Doll‘ar and R.Girshick. “Mask R-CNN,” ArXiv e-prints of 2018
19. B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. Rijke ,B. Sacaleanu, D. Santos and R. Sutcliffe “The Multilingual Question Answering Track at CLEF ,” The Multilingual Question Answering Track at CLEF , 2006
20. Ferrucci, David and Brown, Eric and Chu-Carroll, Jennifer and Fan, James and Gondek, David and Kalyanpur, Aditya and Lally, Adam and Murdock, J William and Nyberg, Eric and Prager, John and Schlaefer, Nico and Welty, Christopher “Building Watson: An Overview of the DeepQA Project,”. AI Magazine, 2010. pages 59-79
21. J. Pennington , R. Socher and C. Manning “GloVe: Global Vectors for Word Representation,” Empirical Methods in Natural Language Processing, 2014

22. G. Abinaya, G. Ranjan and P. Aswin Karthik, “Continuous learning mechanism of NLU-ML models boosted by human feedback,” International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pages 1-6

COMMON GROUND, FRAMES AND SLOTS FOR COMPREHENSION IN DIALOGUE SYSTEMS

Philippe Blache and Matthis Houlès

Laboratoire Parole et Langage, CNRS, Aix-en-Provence, France

ABSTRACT

This paper presents a dialogue system for training doctors to break bad news. The originality of this work lies in its knowledge representation. All information known before the dialogue (the universe of discourse, the context, the scenario of the dialogue) as well as the knowledge transferred from the doctor to the patient during the conversation is represented in a shared knowledge structure called common ground, *that constitute the core of the system. The* Natural Language Understanding *and the* Natural Language Generation *modules of the system take advantage on this structure and we present in this paper different original techniques making it possible to implement them efficiently.*

KEYWORDS

Dialogue systems, common ground, natural language understanding.

1. INTRODUCTION

We present in this paper a multimodal dialogue system for training doctors to break bad news. This system consists in asking trainees, following a given scenario, to announce the patient a problem that occurred during a medical act [13]. Doctors are frequently faced with such a situation in real life, and official agencies (e.g. the French "*Haute Autorité de la Santé*") underline the fact that training communication skills for interacting with patients is of deep importance. A typical bad news is a damage associated to the care, consequence of an un-expected event that can be due to a medical complication, a dysfunction or a medical error. Experienced clinicians consider the task of announcing this type of information as difficult, daunting, and stressful. The problem is that training doctors in this perspective remains a complex task: it is organized by hospitals as workshops during which doctors interact with actors playing the role of patient [11]. Such training solution is difficult to implement, expensive and time-consuming. The ACORFORMed project has proposed to develop an immersive platform in virtual reality [13] with an *embodied conversational agent* simulating a patient interacting with the doctor. However, the first version of this platform was equipped with an efficient dialogue system, capable of understanding precisely doctor's productions and generating appropriate reactions. We describe in this paper a system fulfilling these requirements.

The main difficulty in dialogue systems concerns the understanding module and more generally semantic processing. This task still represents a challenge and a research question for open domain dialogues. Fortunately, task-oriented dialogue systems (and even more crucially training-purpose applications) correspond to a very specific situation in which semantics can be controlled precisely ad being restricted to the task itself. In our use case, the system (i.e. virtual patient) has a complete knowledge of the scenario and the context. Moreover, the user (i.e. the doctor to be trained) receives before the interaction a set of information with the patient's medical

DOI: 10.5121/csit.2021.112003

folder, the context description that led to the problem (e.g. surgery, endoscopy, therapy, etc.), the description of the bad news and how (if possible) it should be fixed. He/she also receives several recommendations on the way to deliver the bad news, following official guidelines elaborated by national agencies. The user can freely talk with the virtual patient that generates in response multimodal verbal and non-verbal reactions. In such a context, the dialogue structure is very specific. Technically, this means that the semantic domain is closed and the system has the entire knowledge of the discourse universe (including the specific scenario associated to the task). Beside this characteristic, the system also have information about how the doctor has to announce the new. A last and very important feature of this specific training context is that the doctor remains the main speaker all along the interaction. On its side, the patient (played by the dialogue system) only reacts to the doctor's utterances, without taking the lead of the conversation.

These different characteristics deeply impact on the one hand the behaviour of the agent and on the other hand the technology to be used for the comprehension module. Moreover, as far as agent's behaviour is concerned, the most important feature of the system consists in how it reacts to doctor's utterances in particular by answering questions, producing feedbacks and asking for clarification. In terms of understanding techniques, thanks to the pre-defined knowledge of the discourse universe, the core of the architecture relies on a knowledge structure precisely defined before the conversation.

We propose in this paper a description of the main aspects of the comprehension module of our system. We use machine learning techniques when possible, but the main architecture remains symbolic, in particular because of lack of data in this domain. Moreover, many multimodal behaviours of the virtual agent are directly controlled by rules during the comprehension process. Finally, the system is designed to be used for training: the state of the common ground after the interaction is an important element of evaluation for the trainee. Deep learning approaches would not provide such a facility.

We focus in the remaining of the paper on these two aspects: knowledge representation for understanding and generating patient's behaviours.

2. KNOWLEDGE REPRESENTATION: THE *COMMON GROUND*

As underline above, the context of dialogue systems for training is very specific from many respects. First, the universe of discourse (i.e. the semantic domain)is fully specified both for the knowledge it concerns (in our case the context of the damage, and all the medical aspects) but also in the way the information has to be delivered, according to certain requirements and recommendations [17]. The doctor's discourse is organized around three main phases: *greetings, damage description and remediation*, *closing* [12]. Moreover, and this is of great importance for knowledge representation, both the doctor and the patient have a complete knowledge of the context, the degree of severity, the risk, etc.

In terms of interaction theories [15], information updating consists in building a shared knowledge between the speakers, called "*common ground*" [18], made of what is supposed to be known by both participants. The task consists in adding step by step during the conversation new information in relation to an item of the common ground. In this common knowledge base, many information is also presupposed or can be inferred automatically depending on the instantiated knowledge. What is specific to the common ground is that all participants suppose the others also have access to the same knowledge. In the case of a training dialogue environment, the context, the scenario and the recommendations are already known by the system before the interaction (this fact is hidden to the trainee). The evolution of the information transfer from the doctor to the patient consists in specifying in the knowledge base what has been transferred. Moreover, in the situation of a task-oriented dialogue, the system knows at anytime not only what has been updated, but also what remains to be instantiated.

Formally, the common ground is made of a set of frames in the sense of frame semantics [8], defined as attribute-value matrices gathering different pieces of in-formation, called slots as illustrated in figure 1. A slot value can be atomic (e.g. values of the slots *Name*, *Age*, *Gravity*, etc.) or refer to another frame (e.g. *Person*, *Pathology*). Moreover, each slot can be associated with different control information [2]. First, each slot may be weighted, in a 3-value scale: *mandatory*, *important*, *optional*. Second, an information may depend from another frame or slot value. For example, the doctor cannot describe any remediation before having presented the pathology: in this case, we say that the frame `Remediation` depends on the frame `Pathology` description. Finally, the last important control concerns slots: depending on certain values, particular agent's reactions may be triggered. For example, if the value *high* is instantiated to the slot *Severity*, then an emotional feedback may be generated by the agent. All this information is to be encoded by specific constraints associated to the frame or the slot description, each slot bearing the following features: *type, weight, dependent-values, inference*.

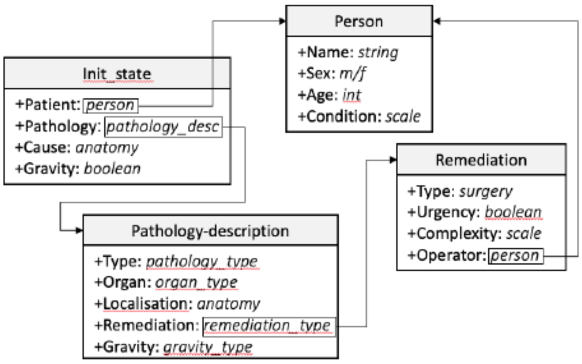


Figure 1: Example of frames and slots

3. Understanding with Common Ground: Frames and Slot Instantiation

As sketched in the previous section, the understanding mechanism of our dialogue system only relies on common ground instantiation. This knowledge structure being based on a frame lattice (each frame being made of slots), the first step of the comprehension module consists then in identifying such frames from the doctor's speech. In this perspective, it is interesting to note the correspondence between dialogue acts [7, 6] and frames required by the common ground representation (for example the different phases of the dialogue). The mechanism for frame identification relies on this correspondence, and we propose to use a dialogue act classification technique for that.

3.1. Dataset

We collected a corpus of training sessions, in French, organized between doctors (the trainees) and patients (played by human actors). The corpus is made of 7 sessions each lasting around 15mns. The audio input (representing 37,000 words) has been transcribed and manually corrected. The corpus has been automatically segmented into inter-pausal units (with pauses higher than 250ms).

Each inter-pausal unit forms an utterance, 1,822 such utterances have been produced throughout the 7 dialogues by the doctors. The corpus has been annotated automatically when possible (tokenization, POS tagging) and manually as for the dialogue acts associate to each utterance (5 annotators among which two experts).

3.2. Step 1: Frame identification

A dialogue in our type of discourse is structured into different phases, on top of classical opening and closing: the description of the patient's initial state (the cause of the hospitalization), the bad news description (typically an incident during a surgery) and the patient's current state. Moreover, the doctor also gives explications, asks questions, reassures the patient and have different types of social interactions. These different actions correspond to different dialogue acts to be annotated: *Opening, Init state, Init remediation, Bad news state, Bad news remediation, Current state, Current remediation, Reassurance, Explication, Social interaction, Discourse, Question, Closing*.

In our approach, each dialogue act corresponds to a frame in the common ground. The first step of the comprehension module is then to identify these frames that can be associated to each doctor's utterance during the dialogue. This problem corresponds to a classification one in which the predictive variables are extracted from the utterance. We propose to implement this classification task on the basis of different linguistic features that already have been shown to be effective [4]:

Classical features: We first use a set of features classically involved in DA identification. It consists in combining TF-IDF principles with word and character n-grams. Applying a principal component analysis, we extracted 4 combinations to be tested:

- f-TFIDF (TFIDF on word n-grams from 1 to 3 words keeping the 250 best, and character n-grams (from 3 to 5 chars, keeping the 250 best)
- s-TFIDF The f-TFIDF features, filtered with a singular value decomposition in order to obtain a better representation density
- w-TFIDF TFIDF only based on the word n-grams, keeping the 500 best
- l-TFIDF TFIDF based on the lemmas n-grams, keeping the 500 best

Morpho-syntactic features: We also involve in the model low-level morpho-syntactic features, based on POS tags: number of discourse markers in the utterance, number of filled-pauses, number of tokens.

Lexical features: A dictionary specific to our domain has been created, containing medical words in which we distinguished pathological terms vs. others. Moreover, we annotated the data with a specific label tagging the medical words depending on they appear for the first time in the dialogue or not (corresponding to the given/new distinction used in discourse analysis).

Context features: As proposed in several works [5, 16], context (i.e. the labels of the preceding dialogue acts) is taken into account. We implemented three different context representations, in a 1 to 5 window: one hot encoding of the preceding DAs, bag-of-words(encoding the number of times the DA appears in the context of the utterance), n-grams of words (up to 0.5% frequency).

Syntactic features: High-level syntactic information can play a role in the characterization of certain classes. In particular, dialogue sequences corresponding to a description or an explanation are usually associated to more complex structures, with more modifiers (adjectives and adverbs) and more complex clauses (subordinates, relatives, prepositional phrases). We propose two features for a simple approximation of these characteristics: the ratio of the number of adjectives and adverbs to the total number of tokens in the utterance $\frac{nbAdj + nbAdv}{\sum tokens}$ and the ratio of the number of conjunctions, pronouns and prepositions to the total number of tokens $\frac{nbConj + nbPrep + nbPro}{\sum tokens}$.

A hierarchical top-down classification, limited to two levels (DA meta-classes and leaf classes) which consists in training a multi-class classifier for each level[10] has been implemented. We keep as first-level classes the meta-classes specified in the ISO 24617-2 scheme: `Opening, Discourse, Inform, Question, Closing`. These classes are not only easier to identify, they also correspond to different agent's reactions (standardized reaction or feedbacks in association with `Opening, Discourse` and `Closing` and appropriate answers (see next section) with `Questions`. The dialogue act `Inform` correspond to the majority class, that we separate into 8 subclasses: `Init_state, Init_remediation, Bad_new_state, Bad_new_remediation,Current_state, Current_remediation, Social_interaction, Explication`.

Different algorithms and feature combinations have been tried for training the classifiers at both levels. The best result for the first level classification has been obtained using a linear regression classifier with an Anova to select the k-best features. As expected, the accuracy is very high, reaching 94% (89% of balanced accuracy). Note that the 1st-level classes are relatively stable and easy to recognize, the context feature did not bring any improvement there.

The second step of the classification consists in applying a new classifier to the sequences labeled `Inform` by the 1st-level classifier. For this step, the best results have been obtained using random forests with the complete set of features and reaching an accuracy of 77.2% (71% balanced accuracy). More details on this stage can be found in [4].

3.3. Step2: Slot filling

The second step in the common ground instantiation concerns slot filling. At each doctor's utterance, we identify a frame (corresponding to the dialogue act)thanks to the classifier. Several approaches have proposed to process at the same time frame classification and slot filling in a unique mechanism [9]. In our case, the dialogue act classifier returns the frame: thanks to the common ground, we know then the set of slots to be instantiated. This task consists in identifying the value and the slot to be instantiated. More precisely, two steps can be specified:

1. Extracting from the utterance the possible values for the different slots
2. Selecting the slot, verifying the type compatibility, instantiating the value

The fact that the list of slots prone to instantiation is very small opens the possibility to adopt a specific mechanism. Instead of trying to identify before-hand the possible slot values from the utterance and then to look for the slot taking into account the compatibility of the type of its value, we propose to implement a reverse mechanism based on semantic similarity. Instead of an abstract type, each slot is associated with a prototypical value. For example, the slot "*specialty*" of the frame “*doctor*” takes as value “*surgeon*”. Then, for each term of the utterance, the semantic distance with the prototypical term of the slots is calculated. Above a given similarity threshold, the slot is then instantiated taking as value the term itself. In our case, the similarity is calculated with the *Gensim* package (*https://radimrehurek.com/gensim/*).

4. QUESTIONS, CLARIFICATION, FEEDBACKS

In a task oriented-dialogue, both the semantic domain and the task to be filled are completely known by the system before the interaction. Moreover, in the case of a medical conversation (typically for breaking bad news), the main speaker remains the user (i.e. the doctor). The main task of the dialogue system understanding module is to update the knowledge transferred by the doctor, in other words to instantiate the common ground. The virtual patient, on its side, has only few information to transfer and its main activity consists in reacting to doctor’s messages and behaviours. This is of deep importance for agent’s naturalness and credibility. Three main reactions have to be implemented in priority: answering doctor’s questions, generating conversational feedbacks, and asking for clarification. This section presents different solutions addressing these specific problems.

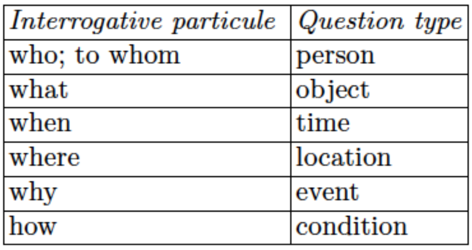
4.1. Answering questions

Questions are identified by the dialogue act classifier. A distinction is done between open-ended questions (wh-questions) and closed-ended questions (yes-no questions).

Yes-no questions: This type of questions focuses on the patient’s condition (“*Are you in pain this morning*?”, “*Did they bring you pain medication*?”), under-standing (“*Do you have any questions?*”), or social aspects (“*Do you want us to call your son?*”). The answers to these questions depend on a scenario or user profile. It does not provide information used by the dialog system. The choice of the type of answer, “*yes*” or “*no*” (or any other positive or negative rephrasing) is left to the system and not based on any particular semantic processing.

Wh-questions: In this case, it is necessary to identify the type of questioning and the informational *focus* of the question. Generally speaking, an open-ended question is made up of an interrogative particle giving the type of question followed by a description of the focus of the question, which is a specific property of the object or event to which the question relates. An open-ended question can therefore be represented by the doublet <*question type; focus property*>. *Question type*: We traditionally distinguish 8 types of open questions, corresponding to different forms of the interrogative particle: *who, what, when, where, why, how, what, to whom*. We propose to associate a generic type for the answer of each of these question types:

Table 1. Interrogative particles and their types.



Interrogative particule	*Question type*
who; to whom	person
what	object
when	time
where	location
why	event
how	condition

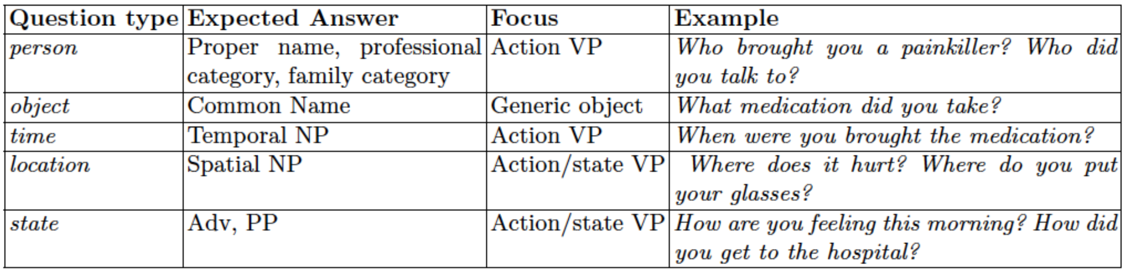
Focus of the question: Depending on the type of question, the associated characteristic is identified, describing a property of the object of the answer. For example, if the type of question is a location, the focus will generally be an action (or possibly a state), represented by a verb phrase (the head of which being an action verb). Table 2 summarizes the prototypical focus for each type of question.

Generating the answer: Knowing the expected answer type and the focus of the question makes it possible to generate straightforwardly an answer, based on generic patterns. All questions refer necessarily to the patient's state or personal data. This information is then already encoded in the common ground, as part of the scenario. Generating the answer consists then in looking into the common ground for a slot value corresponding to the expected answer type, together with the focus as a key for identifying the associated frame.

4.2. Conversational feedbacks

We propose an approach making it possible to generate feedbacks on the basis of different cues that can be identified in real time from doctor's behaviour. In our feedback model, besides low-level classical cues (such as breaks, turn length, POS, etc.), we also integrate higher level semantic or discourse-level cues [3].

Table 2. Expected answers depending on the type of question.



Question type	**Expected Answer**	**Focus**	**Example**
person	Proper name, professional category, family category	Action VP	*Who brought you a painkiller? Who did you talk to?*
object	Common Name	Generic object	*What medication did you take?*
time	Temporal NP	Action VP	*When were you brought the medication?*
location	Spatial NP	Action/state VP	*Where does it hurt? Where do you put your glasses?*
state	Adv, PP	Action/state VP	*How are you feeling this morning? How did you get to the hospital?*

The dialogue systems mainly have two input streams: the audio signal and its transcription. Prosodic features (silent pauses, pitch, IPU duration, etc.) are extracted from the audio stream. Temporal features, also coming from this stream, are kept updated, in particular the duration since the last feedback, the indication of the current state of the production (speech or silent pause), the duration of the speech since the last pause, the duration of the pause, etc. On their side, linguistic features can be acquired from the transcription stream: morphosyntax (POS n-grams), lexicon (some terms can trigger specific feedbacks), but also at a higher level the information structure (the introduction of a new referent) or the discourse organization (transition between phases) can also be associated with specific listener's reactions [1]. Finally, semantics plays a central role in generating feedbacks: many listener's reactions are triggered upon instantiation of the common ground.

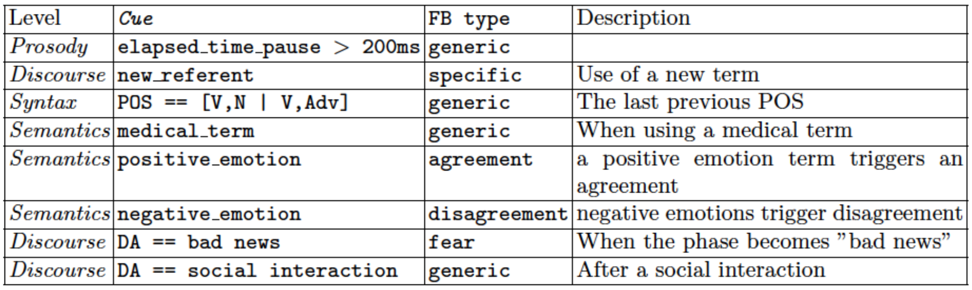
We propose to implement a semantic-based feedback generation by associating CG slots to specific feedbacks (for example, a feed-back expressing fear is triggered instantiation of the slot "urgency"). Note that in the case of a task-oriented dialogue, most feedbacks are triggered by linguistic cues. As a consequence, when the doctor speaks, we first look at linguistic cues whereas during a pause, the feedback generator is mainly based on pause duration. The different cues extracted from the analysis of the input streams serve as input to a feedback type identification function, based on a set of rules, as illustrated in figure 3. Given the feedback type to be generated and the current mode (pause or speech), the last step consists in generating the feedback itself, by choosing among a list of possible candidates. This list is in a probability space which also depends on the current state (e.g. visual or bimodal feedback will be preferred during speech where verbal feedbacks will be favoured during pauses).

4.3. Clarification questions

Clarification questions plays an important role in dialogue not only for the verification of the common ground construction, but also (even to a greater extent) for the naturalness of the virtual agent: such questions show very efficiently that the agent understands and follows the conversation. Different conditions can trigger such questions.

As explained in the description of the common ground structure, frames and slots can be associated with different controls. First, some frames or slots can be instantiated only when other frames or slots are already instantiated. Such values form a pre-requisite and are called *dependent values*. For example a doctor cannot present a diagnostic (i.e. the system cannot instantiate a `diagnostic` frame)before having presented the symptoms (resp. instantiation of a `symptom` frame. In the same way, in the case of our use case, the bad news frame cannot be created before having developed the `init_state` one. Such relations make it possible to implement both sequentiality and semantic dependencies. A dependent value conflict is detected when a slot is about to be instantiated with a dependent value still free.

Table 3. Feedback generation rules



Level	*Cue*	FB type	Description
Prosody	elapsed_time_pause $>$ 200ms	generic	
Discourse	new_referent	specific	Use of a new term
Syntax	POS == [V,N | V,Adv]	generic	The last previous POS
Semantics	medical_term	generic	When using a medical term
Semantics	positive_emotion	agreement	a positive emotion term triggers an agreement
Semantics	negative_emotion	disagreement	negative emotions trigger disagreement
Discourse	DA == bad news	fear	When the phase becomes ”bad news”
Discourse	DA == social interaction	generic	After a social interaction

In the case of a slot dependency, the conflict is directly identified by verifying whether the dependent slot is already instantiated or not. If not, the slot name and its value type are passed to the generation module which select a question pattern filled with this information. As for frames, the dependent value conflict requires a more complex process. The problem consists in identifying whether a frame is instantiated or not: in most of the cases, only part has been informed, the frame remaining incomplete. The problem is then to evaluate whether the frame can be considered as complete or not. As presented with the common ground, we have seen that each slot is associated to a weight (3-value scale). When a frame A is dependent from a frame B, the system verifies whether all mandatory slot values of B are already instantiated1. If not, then a clarification question is generated, using the same generation mechanism as for slots. The second situation triggering a clarification question occurs when no slot can be instantiated in spite of the identification of a frame by the classifier. In this case, none of the terms belonging to the utterance is similar enough (i.e. reaches a sufficient similarity threshold) with one of the prototypical values of the different slots. In this case, we can say that there is a type mismatch between the term used by the doctor for a slot value and the expected value. The clarification question generated is general, simply indicating an incomprehension of the system. The last case processed by our system concerns general questions that can be asked by the system at the end of the interaction or when the doctor asks the agent whether he/she has some questions. The mechanism consists in selecting either a non saturated frame or the frame with the more non-instantiated slots weighted as important. This frame being chosen, the system selects arbitrarily one of the non instantiated important slots and generates a question.

5. GENERATING A MULTIMODAL ANSWER

We briefly sketch in this section the multimodal aspects of the dialogue system. Concerning the input signal, our goal being to develop a generic application, the first constraint concerns the user’s equipment: we want to remain totally free of any instrument or specific sensor, the user freely speaking (or writing) to the agent. Our second goal is to offer the possibility of having a written output (in the situation where the system only generates written sentences), an audio (the system speaks to the user) or a multimodal one. In this last case, the system generates the code for a complete behaviour (speech and gestures) of an embodied conversational agent.

Concerning the input stream, at this stage of development, only the audio modality is processed. Automatic speech recognition is applied to the doctor's production, providing its transcription plus some prosodic information, in particular concerning pauses. The transcription is then segmented into discourse units delimited by discourse markers (*then, because, and, but,* etc.) that indicate approximately a change between different discourse units. Such segmentation makes it possible to identify homogeneous semantic units that can be associated to only one or two frames by the classifier. At this stage, the video signal that could be used for detecting head movement, smiles, etc. is not already processed by the system. As a consequence, we mainly use as input unimodal information based on the transcription only. On the contrary, concerning the output, the system generates multimodal behaviours played by an ECA implemented in the Virtual Interactive Behaviour platform [14]. This platform includes an XML dialect called FML encoding instructions for generating the agent’s verbal and non-verbal behaviour. The mechanism consists then for our system to generate (dynamically when necessary) this FML code for each answer or reaction of the agent.

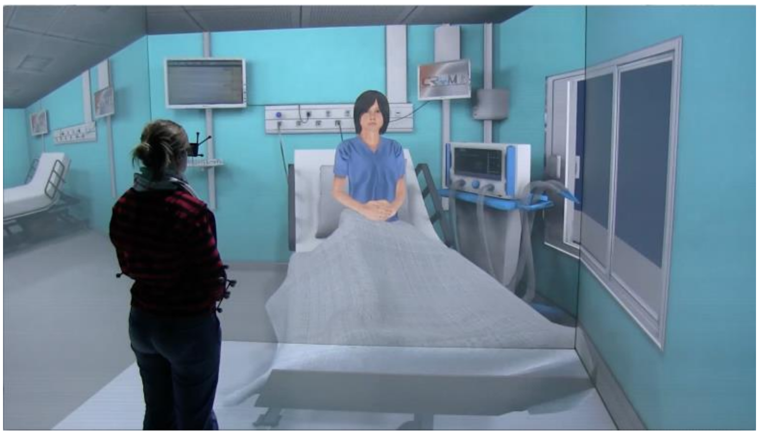


Figure 2. The embodied virtual patient of the ACORFORMed platform

We implemented two different ways for generating such multimodal behaviours depending on whether they include flexible verbal material or not. In the last case (typically feedbacks), the list of possible behaviours is closed and rather small. Moreover, they are very canonical and the verbal material very limited, fixed or even absent. It is then possible (and preferable in terms of efficiency)to create a specific gesture library corresponding to the required feedbacks and then to encode a predefined list of FML files for these prototypical behaviours.

Each feedback may have different realizations that can be chose randomly in order to introduce variability. The generation of flexible behaviours including verbal material requires a three-step mechanism. First, the verbal part is generated following the methods described above. The verbal utterance to produce is then passed to the FML generator. This module generates a first version of the FML code my using the standard VIB function (FMLAnnotator) and then refine this code by completing or adding different instructions (e.g. the prosody).

6. CONCLUSION

Using dialogue techniques for training doctors represent a seminal use case both in the perspective of developing new dialogue techniques, but also in terms of application: training doctor's social skills is known to be of crucial importance in the therapy process. Such dialogue systems require at the same time to be very precise, reactive and to allow doctors to interact freely, with spontaneous speech: this correspond to the most difficult challenges for dialogue technology. We have presented in this paper an approach fulfilling these requirements by taking advantage of the particularities of this type of application. Our approach relies on a precise knowledge representation, the common ground, which constitutes the core of the understanding architecture. The frame-based representation first offer the possibility to use classification techniques identifying directly the frame to be instantiated. Thanks to this first step, we have proposed an original slot filling method, based on the common ground and distributional semantics information. The generation of the agent's reactions and its adaptation to the doctor's speech directly takes advantage of our common ground representation.

ACKNOWLEDGEMENTS

This work has benefited from support from Institut Convergence ILCB (ANR-16-CONV-0002).

REFERENCES

[1] Bertrand, R., Espesser, R. (2017) Co-narration in French conversation storytelling: a quantitative insight. In Journal of Pragmatics111

[2] Blache, P. (2017) Dialogue management in task-oriented dialogue systems. In: International Workshop on Investigating Social Interactions with Artificial Agents

[3] Blache, P., Abderrahmane, M., Rauzy, S., Bertrand, R. (2020) An integrated model for predicting backchannel feedbacks. In: ACM International Conference on IntelligentVirtual Agents

[4] Blache, P., Abderrahmane, M., Rauzy, S., Ochs, M., Oufaida, H. (2020) Two-level classification for dialogue act recognition in task-oriented dialogues. In: COLING'20

[5] Bothe, C., Weber, C., Magg, S., Wermter, S. (2018) A context-based approach for di-alogue act recognition using simple recurrent neural networks. In: LREC 2018

[6] Bunt, H., Fang, A.C., Petukhova, V. (2017) Revisiting the iso standard for dialogue act annotation. In: Proceedings of the 13th Joint ISO-ACL Workshop on InteroperableSemantic Annotation (ISA-13)

[7] Core, M., Allen, J. (1997) Coding dialogs with the damsl annotation scheme. In: Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines. pp. 28–35

[8] Fillmore, C.J., Baker, C. (2009) A frames approach to semantic analysis. In: Heine, B.,Narrog, H. (eds.) The Oxford Handbook of Linguistic Analysis. Oxford UniversityPress

[9] Firdaus, M., Golchha, H., Ekbal, A., Bhattacharyya, P. (2020) A deep multi-task model for dialogue act classification, intent detection and slot fill-ing. Cognitive Computation https://doi.org/10.1007/s12559-020-09718-4,https://doi.org/10.1007/s12559-020-09718-4

[10] Freitas, A., de Carvalho, A. (2008) A tutorial on hierarchical classification with applications in bioinformatics. In Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications pp. 119–145

[11] Granry, J.C., Moll, M. (2012) Rapport de mission, état de l'art en matière de pratiques de simulation dans le domaine de la santé. Tech. rep., HAS 2012

[12] Ochs, M., Blache, P., Montcheuil, G., Pergandi, J.M., Bertrand, R., Saubesty, J.,Francon, D., Mestre, D. (2018) The Acorformed corpus: Investigating multimodality inhuman-human and human-virtual patient interactions. In: CLARIN Annual Conference 2018. p. 16

[13] Ochs, M., Mestre, D., de Montcheuil, G., Pergandi, J.M., Saubesty, J., Lombardo, E., Francon, D., Blache, P. (2018) Training doctors' social skills to break bad news: Evaluation of the impact of virtual environment displays on the sense of presence. In Journal on Multimodal User Interfaces1

[14] Pelachaud, C. (2009) Studies on gesture expressivity for a virtual agent. In Speech Communication51(7), 630–639

[15] Pickering, M., Garrod, S. (2013) An integrated theory of language production and com-prehension. In Behavioral and Brain Sciences36(04), 329–347

[16] Raheja, V., Tetreault, J. (2019) Dialogue act classification with context-aware self-attention. In: NAACL-2019

[17] Schnebelen, C., Pothier, F., Furney, M.: Annonce d'un dommage associé aux soins. (2011) Tech. rep., Haute Autorité de Santé

[18] Stalnaker, R. (2002)Common ground. In Linguistics and Philosophy 25(5), 701–721

AUTHORS

Philippe Blache is senior researcher at the CNRS, France (Laboratoire Parole & Langage, Aix-en-Provence). He is the founder and the former director of the ILCB (Institute of Language, Communication and the Brain)

Matthis Houlès is student at the Engineer school of Nantes

ADVANCED SKILLS MAPPING AND CAREER DEVELOPMENT USING AI

Yew Kee Wong

School of Information Engineering, HuangHuai University, Henan, China

ABSTRACT

Artificial intelligence has been an eye-popping word that is impacting every industry in the world. With the rise of such advanced technology, there will be always a question regarding its impact on our social life, environment and economy thus impacting all efforts exerted towards continuous development. From the definition, the welfare of human beings is the core of continuous development. Continuous development is useful only when ordinary people's lives are improved whether in health, education, employment, environment, equality or justice. Securing decent jobs is a key enabler to promote the components of continuous development, economic growth, social welfare and environmental sustainability. The human resources are the precious resource for nations. The high unemployment and underemployment rates especially in youth is a great threat affecting the continuous economic development of many countries and is influenced by investment in education, and quality of living.

KEYWORDS

Artificial Intelligence, Human Resources, Conceptual Blueprint, Continuous Development, Learning and Employability Blueprint

1. INTRODUCTION

Continuous development is defined as the development that meets the needs of the present without compromising the ability of future generations to meet their own needs [1]. The primary cause of the high unemployment rates is the inefficient education systems that fail to equip young people with the required skills for the labour market. In this research, we propose the use of artificial intelligence to enhance the relationship between education and employment.

Many studies were published on how to improve education curricula to enhance the employability of students; frameworks were designed to facilitate the work of teachers, mentors, career advisers and faculty to guide students through their career exploration and preparation. Numerous papers were published on the impact of artificial intelligence (AI) on education and its impact on employment. However it seems there is a gap in connecting the three important areas of research, 1: education for employment, 2: AI in education and, 3: AI in employment [2]. Further investigations are needed to evaluate and assess how AI can fit in the current learning and employability blueprint and to evaluate what can innovation and entrepreneurship bring to promote better education for employment systems.

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 45-52, 2021. CS & IT - CSCP 2021 DOI:10.5121/csit.2021.112004

2. USING AI TO BUILD A CONCEPTUAL BLUEPRINT

The study is assessing new blueprint for learning and employability and how AI can fit in andfoster the process, so further experiments should be carried out to ensure the effectiveness of theblueprint and the accuracy of results of the AI application on the learning and employabilityprocess [3]. After reviewing literature regarding the impact of AI and its potential on botheducation and employment, as well as reviewing different education for employment blueprints,theories and case studies, this paper attempts to close the gap in the research related to specificscope which is the impact of AI on education for employment [4].

Young people can't find jobs. Yet employers can't find people with required skill set. This mismatch between the supply demand in the labour market might witness a bigger gap in the future with the growth of AI technologies. There are few frameworks for education for employment or in other words "Learning and Employability" [5]. However the existing model didn't address the potential of AI whether in terms of deployment of such technology within the model or in terms of the implications of AI on the learning models or the employment models. So there is a need to find a practical frame for learning and employability that incorporate the advancements of AI to facilitate the university to work transition. This paper seeks to figure out the room for AI potentials through mapping innovative startups that embraced AI capabilities to play a role in the education for employment ecosystem.

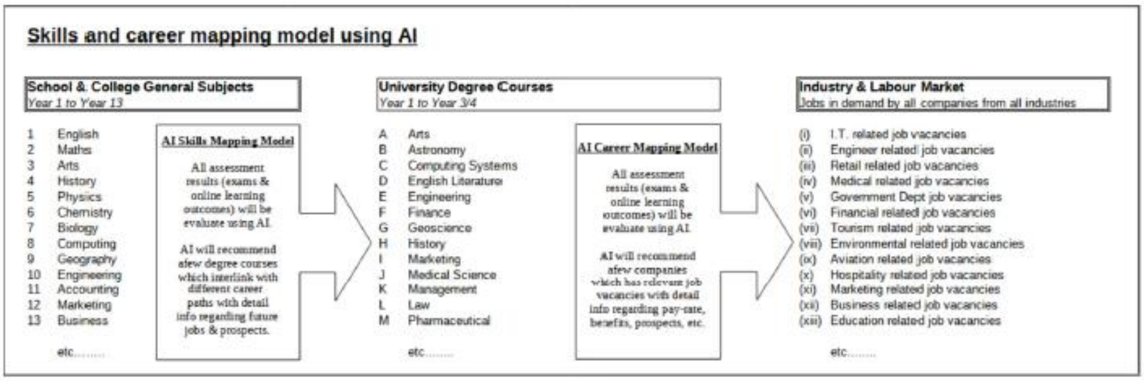
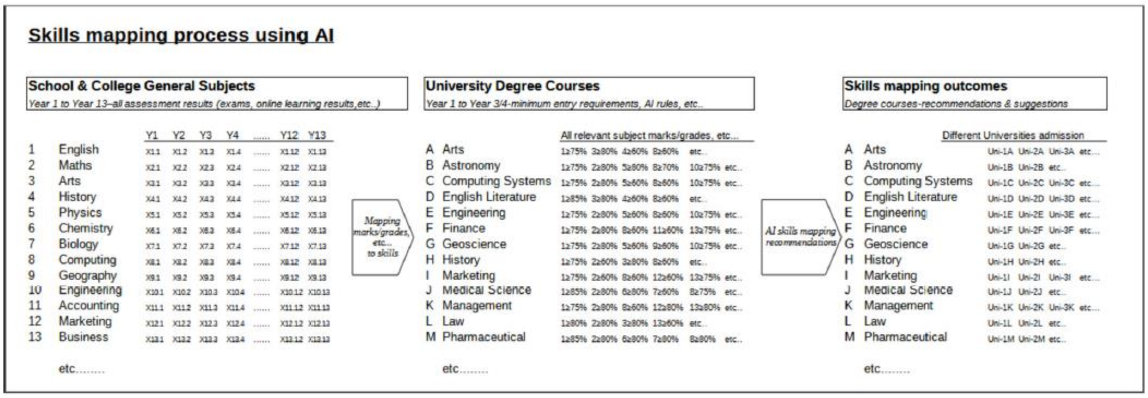

Skills and career mapping model using AI
School & College General Subjects
Year 1 to Year 13
1 English
2 Maths
3 Arts
4 History
5 Physics
6 Chemistry
7 Biology
8 Computing
9 Geography
10 Engineering
11 Accounting
12 Marketing
13 Business
etc........
AI Skills Mapping Model
All assessment results (exams & online learning outcomes) will be evaluate using AI.
AI will recommend afew degree courses which interlink with different career paths with detail info regarding future jobs & prospects.
University Degree Courses
Year 1 to Year 3/4
A Arts
B Astronomy
C Computing Systems
D English Literature
E Engineering
F Finance
G Geoscience
H History
I Marketing
J Medical Science
K Management
L Law
M Pharmaceutical
etc........
AI Career Mapping Model
All assessment results (exams & online learning outcomes) will be evaluate using AI.
AI will recommend afew companies which has relevant job vacancies with detail info regarding pay-rate, benefits, prospects, etc.
Industry & Labour Market
Jobs in demand by all companies from all industries
(i) I.T. related job vacancies
(ii) Engineer related job vacancies
(iii) Retail related job vacancies
(iv) Medical related job vacancies
(v) Government Dept job vacancies
(vi) Financial related job vacancies
(vii) Tourism related job vacancies
(viii) Environmental related job vacancies
(ix) Aviation related job vacancies
(x) Hospitality related job vacancies
(xi) Marketing related job vacancies
(xii) Business related job vacancies
(xiii) Education related job vacancies
etc........

Figure 1. Skills and career mapping model using AI.

2.1. The Use of AI in Skills Mapping Model

In this model, we proposed to use AI to streamline the skills requires by various degrees courses. This process significantly reduce the time for the student to decide on what degree subjects they can register for the university entrance. Furthermore, the model can also assist the student by presenting the detail information regarding the different career paths, current employers that are offering job related vacancies, pay-rate range, related benefits and other prospects [6].

Some students may not have a clear career path after they completed their college study and require further guidance and advise on choosing the appropriate degree course for their future career development [7]. In this evaluation process, the AI will use all the information (i.e. exam grades/marks, understanding level from online learning, etc..) provided by the student, Therefore, AI in this process can only provide advanced in-depth career roadmaps as recommendation for the student. The final decision making still rely on the student.



Skills mapping process using AI

School & College General Subjects

Year 1 to Year 13–all assessment results (exams, online learning results,etc..)

		Y1	Y2	Y3	Y4	......	Y12	Y13
1	English	X1.1	X1.2	X1.3	X1.4	......	X1.12	X1.13
2	Maths	X2.1	X2.2	X2.3	X2.4	......	X2.12	X2.13
3	Arts	X3.1	X3.2	X3.3	X3.4	......	X3.12	X3.13
4	History	X4.1	X4.2	X4.3	X4.4	......	X4.12	X4.13
5	Physics	X5.1	X5.2	X5.3	X5.4	......	X5.12	X5.13
6	Chemistry	X6.1	X6.2	X6.3	X6.4	......	X6.12	X6.13
7	Biology	X7.1	X7.2	X7.3	X7.4	......	X7.12	X7.13
8	Computing	X8.1	X8.2	X8.3	X8.4	......	X8.12	X8.13
9	Geography	X9.1	X9.2	X9.3	X9.4	......	X9.12	X9.13
10	Engineering	X10.1	X10.2	X10.3	X10.4	......	X10.12	X10.13
11	Accounting	X11.1	X11.2	X11.3	X11.4	......	X11.12	X11.13
12	Marketing	X12.1	X12.2	X12.3	X12.4	......	X12.12	X12.13
13	Business	X13.1	X13.2	X13.3	X13.4	......	X13.12	X13.13
	etc........							

University Degree Courses

Year 1 to Year 3/4-minimum entry requirements, AI rules, etc..

		All relevant subject marks/grades, etc...					
A	Arts	1≥75%	3≥80%	4≥60%	8≥60%	etc..	
B	Astronomy	1≥75%	2≥80%	5≥60%	8≥70%	10≥75%	etc..
C	Computing Systems	1≥75%	2≥80%	5≥60%	8≥60%	10≥75%	etc..
D	English Literature	1≥85%	3≥80%	4≥60%	8≥60%	etc..	
E	Engineering	1≥75%	2≥80%	5≥60%	8≥60%	10≥75%	etc..
F	Finance	1≥75%	2≥80%	8≥60%	11≥60%	13≥75%	etc..
G	Geoscience	1≥75%	2≥80%	5≥60%	9≥60%	10≥75%	etc..
H	History	1≥75%	2≥60%	3≥80%	8≥60%	etc..	
I	Marketing	1≥75%	2≥60%	8≥60%	12≥60%	13≥75%	etc..
J	Medical Science	1≥85%	2≥80%	6≥80%	7≥60%	8≥75%	etc..
K	Management	1≥75%	2≥80%	8≥60%	12≥80%	13≥80%	etc..
L	Law	1≥80%	2≥80%	3≥80%	13≥60%	etc..	
M	Pharmaceutical	1≥85%	2≥80%	6≥80%	7≥80%	8≥80%	etc..
	etc........						

Skills mapping outcomes

Degree courses-recommendations & suggestions

		Different Universities admission			
A	Arts	Uni-1A	Uni-2A	Uni-3A	etc...
B	Astronomy	Uni-1B	Uni-2B	etc..	
C	Computing Systems	Uni-1C	Uni-2C	Uni-3C	etc...
D	English Literature	Uni-1D	Uni-2D	Uni-3D	etc...
E	Engineering	Uni-1E	Uni-2E	Uni-3E	etc...
F	Finance	Uni-1F	Uni-2F	Uni-3F	etc...
G	Geoscience	Uni-1G	Uni-2G	etc..	
H	History	Uni-1H	Uni-2H	etc..	
I	Marketing	Uni-1I	Uni-2I	Uni-3I	etc...
J	Medical Science	Uni-1J	Uni-2J	etc..	
K	Management	Uni-1K	Uni-2K	Uni-3K	etc...
L	Law	Uni-1L	Uni-2L	etc..	
M	Pharmaceutical	Uni-1M	Uni-2M	etc..	
	etc........				

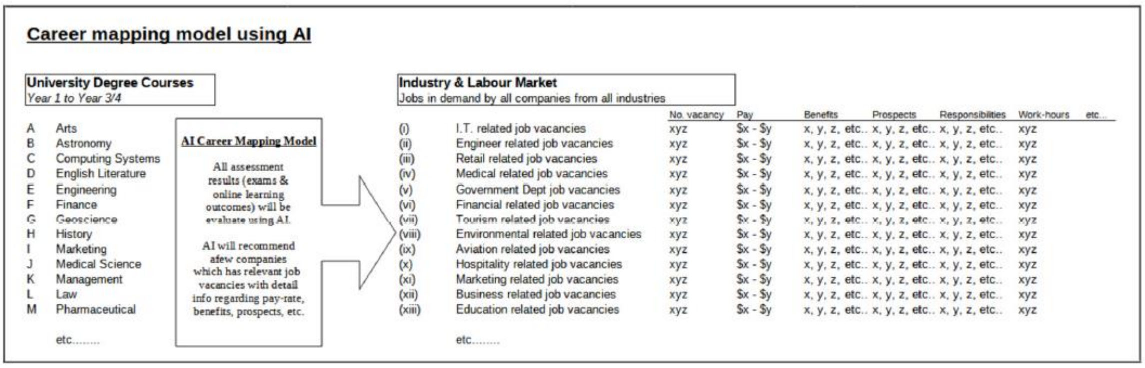
Figure 2. Skills mapping processing using AI

2.2. The Use of AI in Career Mapping Model

With such employment concern, many studies refined such concern and the general consensus now is that AI will generate major transformations in the labour market [8]. According to many researchers, AI will create 2.3 million jobs in 2020, while eliminating 1.8 million and by 2025 AI related job creation will reach two million net-new jobs. Moreover, according to a new report from the World Economic Forum (WEF); 75 million jobs are estimated to be displaced, while 133 million new roles may emerge due to machines and algorithms[9]. The study has argued that this transition to technology should result in favourable unemployment that will allow human labour to better perform activities they were never able to do in their current heavy duty jobs. AI programs will probably be utilized for applications where hiring humans would be too expensive or really dangerous.

AI programs will take over computer tasks allowing humans to dedicate their time to other kinds of tasks including personal services. Service sector companies are optimistic about big data and enthusiastic about AI and robotics deployment as it will have direct impact on productivity improvement that eventually reflects on economic growth. On the other hand, it was realized that AI canpositively impact employment if it is utilized properly within the business model [10]. AI uses in creating effective recruitment systems is seen as an inevitable opportunity to make best use of it. Still this will stay challenging until firms management pay attention to the importance of allocating budgets to finance the required technology for hiring process.

Once the students graduated from the University, they can directly enter the labour market with the help from AI career managing model. The model will provide recommendations and information related to jobs. So that to the graduates can prepare for interviews and other job application process, such as IQ & EQ tests, body check etc. This process can significantly reduce the amount of time graduates need to search for jobs, interviews and other tedious job searching steps and at the same time also reduce the amount of time the employer can recruit the appropriate personnel to fulfil the position and task required in the company, hence can indirectly improve the productivity rate of the company.



Career mapping model using AI

University Degree Courses
Year 1 to Year 3/4

A	Arts
B	Astronomy
C	Computing Systems
D	English Literature
E	Engineering
F	Finance
G	Geoscience
H	History
I	Marketing
J	Medical Science
K	Management
L	Law
M	Pharmaceutical
	etc........

AI Career Mapping Model

All assessment results (exams & online learning outcomes) will be evaluate using AI.

AI will recommend afew companies which has relevant job vacancies with detail info regarding pay-rate, benefits, prospects, etc.

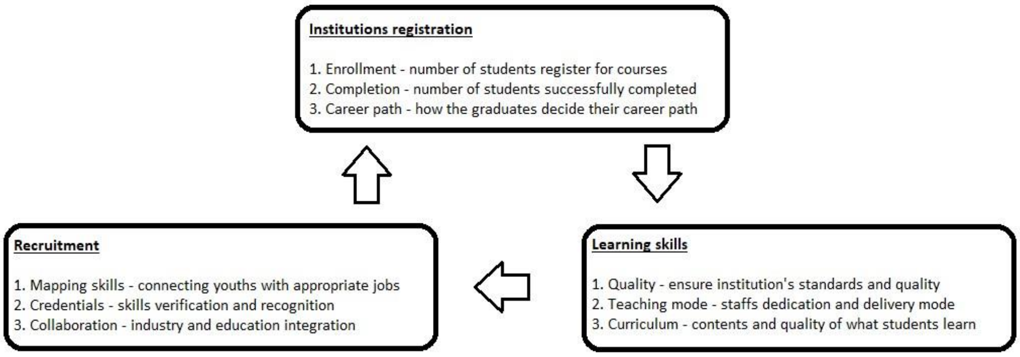
Industry & Labour Market
Jobs in demand by all companies from all industries

		No. vacancy	Pay	Benefits	Prospects	Responsibilities	Work-hours	etc...
(i)	I.T. related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(ii)	Engineer related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(iii)	Retail related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(iv)	Medical related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(v)	Government Dept job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(vi)	Financial related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(vii)	Tourism related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(viii)	Environmental related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(ix)	Aviation related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(x)	Hospitality related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(xi)	Marketing related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(xii)	Business related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
(xiii)	Education related job vacancies	xyz	$x - $y	x, y, z, etc..	x, y, z, etc..	x, y, z, etc..	xyz	
	etc........							

Figure 3. Career mapping processing using AI.

3. LEARNING AND EMPLOYABILITY BLUEPRINT

Aside from the impact of AI in creating new jobs, replacing jobs or even shift in the job and labour market, there are two global employment crises that already exist away from the implications of AI; high levels of youth unemployment and a shortage of talents who possess critical job skills. Mourshed, Farell, & Barton [11] argued that if young people graduating from schools and universities, after exerting lots of efforts, cannot secure decent jobs and observe that sense of respect that comes with such degrees, society may witness outbreaks of anger or even violence. There is an information gap in what works and what does not in preparing young people during their school to employment transition. I summarized this information gap and it clearly shows there is a clear disconnect and misperception about youth job readiness from the point of view of employers vs youth vs educational institutions.



Institutions registration

1. Enrollment - number of students register for courses
2. Completion - number of students successfully completed
3. Career path - how the graduates decide their career path

Learning skills

1. Quality - ensure institution's standards and quality
2. Teaching mode - staffs dedication and delivery mode
3. Curriculum - contents and quality of what students learn

Recruitment

1. Mapping skills - connecting youths with appropriate jobs
2. Credentials - skills verification and recognition
3. Collaboration - industry and education integration

Figure 4. Blueprint for exploring the education to employment system.

3.1. Institutions Registration

A. Information Sharing

All institutions are recommended to develop a comprehensive occupations database and educational/training opportunities and provide information, advice and guidance to help job seekers to make decisions on learning, training and work. The comprehensive occupation database and website allows users to explore different career options including jobs profile, salaries, industry trends and offer webchats with career advisors beside their skills health check assessment that help users to find out what kind of jobs that best suits his/her skills. Users can also find training opportunities.

B. Dealing with Social Perception

It seems that a perception is widespread that getting a decent job with good salary requires being a college graduate. So this puts social pressure on youth to go to college and influence others' choice away from the vocational tracks [12]. Brunello and Rocco [13] argued that youth who graduated from vocational education have a higher likelihood of being not employed and with no education or training within the past 12 months. They also found that vocational education is associated with poorer labour market returns. This as a result impacted on the perception about vocational education.

C. Dealing with Education Affordability

Schultz [14] and Becker [15] introduced individual choice model of human capital investment in which they presented individual's education choice as an investment decision. Individuals sacrifice economically in order to acquire knowledge, referred to as 'human capital', that will enable them to get better rewards in the future. If young people have no access to credit or savings, this may limit their choices and they will not be able to enrol in study.

3.2. Learning Skills

A. Effective Content and Curriculum Design

Mourshed et al. [11] proposed that in order to design relevant curriculum to the employers' requirements, close engagement between, industry leaders and educational providers is needed. Such engagement to succeed, intensive collaboration should exist while defining the core requirements on a very detailed level to ensure that the aspired learning outcomes will be achieved.

B. Effective Delivery Methods

Effective delivery requires still close engagement between employers and educational providers. Mourshed et al. [11] explored two main ways to do so - (1): Classrooms within workplaces. The common model to bring vocational and technical training within the workplaces is through internships or apprenticeships. (2): Workplaces within classrooms. Internships and apprenticeships are types of hands-on learning experiences that are most admired by students, however the number of opportunities are limited to accommodate certain capacities of students.

3.3. Recruitment

A. Assessment for Qualifications and Certifications

Finding a job is a painful process for job seekers. Job seekers strive to market their skills, but can't find a credible way to prove their talents, and Employers can't trust the educational degree as a main reference validating youth skills and knowledge. So both employers and candidates suffer in the hiring and talent acquisition process. One of the well known processes to show one's credentials and prove his skills and knowledge in a credible way is the international professional certifications such as PMP (Project Management Professional) or CPA (Certified Public Accountant) which could be obtained by Individuals after passing standardized tests. Another innovative solution for the assessment and credentials that crossed countries boundaries is the digital badges which introduce much entertainment for online educational activities and experiences.

B. Match Making

Based on their survey that covered more than 100 initiatives in 25 countries, Mourshed et al. [11] observed that there are many cases that educational providers have built strong relationships with employers so that they can hire their graduates immediately after graduation based on the matchmaking and recommendation process that is being done by the educational providers themselves. With current technological advancement, matchmaking could be a game changer in the employment scene.

Flanagan [16] also agreed that Tinder-style matchmaking is beneficial in the job market as well and shed the light on a similar app called "Emjoyment" which allow job seekers to swipe job posts which includes major highlights about the company, location and only one sentence job description and once the job seeker find a good post, he just hits "like". On the other side, employers start to see job seekers who liked their opportunity in a form of cards including resumes main highlights and if the recruiter found an interesting profile, he also hits "like" and at that moment both parties connect together at a push of a button. This kind of matchmaking innovations could decrease the time lost in job applications and finding a good candidate and create direct engagement between employers and job seekers.

4. ARTIFICIAL INTELLIGENCE SYSTEM

The conceptual blueprint using artificial intelligence system include several components which can be integrated as one complete artificial intelligence system [17]. These are the standard components [18]:-

- Reasoning − It is the set of processes that empowers us to provide basis for judgement, making decisions, and prediction.
- Learning − It is the activity of gaining information or skill by studying, practising, being educated, or experiencing something. Learning improves the awareness of the subjects of the study.
- Problem Solving − It is the procedure in which one perceives and tries to arrive at a desired solution from a current situation by taking some path, which is obstructed by known or unknown hurdles.
- Perception − It is the way of acquiring, interpreting, selecting, and organizing sensory information.
- Linguistic Intelligence − It is one's ability to use, comprehend, talk, and compose the verbal and written language. It is significant in interpersonal communication.

The potential of online learning system include 4 factors of accessibility, flexibility, interactivity, and collaboration of online learning afforded by the technology. In terms of the challenges to online learning, 6 are identified: defining online learning; proposing a new legacy of epistemology-social constructivism for all; quality assurance and standards; commitment versus innovation; copyright and intellectual property; and personal learning in social constructivism.

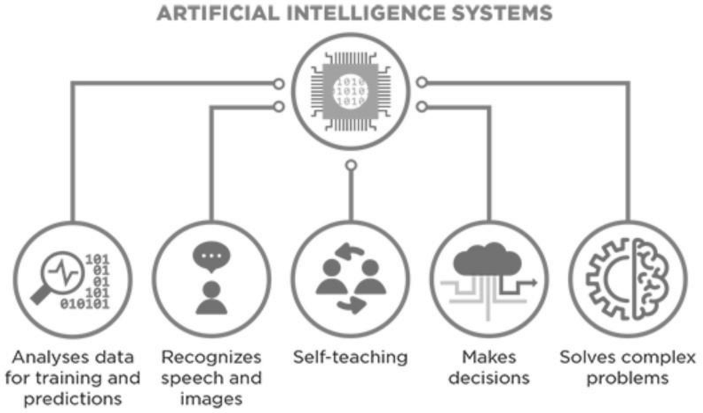

ARTIFICIAL INTELLIGENCE SYSTEMS
Analyses data for training and predictions
Recognizes speech and images
Self-teaching
Makes decisions
Solves complex problems

Figure 5. The artificial intelligence online learning system components.

5. CONCLUSIONS

This research was proposed by understanding the inter-relation between education and employment, what blueprints and systems that worked, and how AI can impact in the education for employment process whether by introducing new innovations that foster students learning process and placement in the job market or by harming the process and introducing unintentional bias, privacy breach, escalating power consumption and replacing human in workplaces. This paper is assessing new blueprints for learning and employability and how AI can fit in and foster the process, so further studies and experiments should be carried out to ensure the effectiveness of the blueprint and the accuracy of results of the AI application on the learning and employability process.

REFERENCES

[1] World Commission on Environment and Development (1987). *Our Common Future.* Oslo.

[2] Tuomi, I. (2018). *The Impact of Artificial Intelligence on Learning, Teaching, and Education.* Publications Office of the European Union.

[3] Popenici, S., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*.

[4] Bayne, S. (2015). Teacherbot: interventions in automated teaching. *Teaching in Higher Education*, 455-467.

[5] Schmidt, A. (2017). *How AI Impacts Education*. Retrieved February 2019, from Forbes:https://www.forbes.com/sites/theyec/2017/12/27/how-ai-impactseducation/#22edd83f792e.

[6] Hawksworth, J. (2018). *AI and robots could create as many jobs as they displace*. Retrieved 2019, from World Economic Forum: https://www.weforum.org/agenda/2018/09/ai-and-robots-could-create-as-many-jobs-as-theydisplace/.

[7] Andrews, W. (2018). *Craft an Artificial Intelligence Strategy: A Gartner Trend Insight Report.* Gartner, Inc.

[8] Khakurel, J., Penzenstadler, B., Porras, J., Knutas, A., & Zhang, W. (2018). The Rise of Artificial Intelligence under the Lens of Sustainability. *Technologies* , 6(4), 100.

[9] *The Future of Jobs* (2018). Centre for the New Economy and Society.

[10] Martens, B., &Tolan, S. (2018). *Will this time be different? A review of the literature on the Impact of Artificial Intelligence on Employment, Incomes and Growth.* Digital Economy Working Paper 2018-08; JRC Technical Reports.

[11] Mourshed, M., Farrell, D., & Barton, D. (2019). *Education to employment: Designing a system that works.* Retrieved April 1, 2019, from McKinsey Center for Government: https://www.mckinsey.com/industries/social-sector/our-insights/education-to-employmentdesigning-a-system-that-works.

[12] Boyer, R. H., Peterson, N. D., Arora, P., & Caldwell, K., (2016). Five Approaches to Social Sustainability and an Integrated Way Forward. *Sustainability, 8(9), MDPI AG*.

[13] Brunello, G., & Rocco, L. (2017). The effects of vocational education on adult skills, employment and wages: What can we learn from PIAAC? *Springer Link*, 8-315.

[14] Schultz, T. W. (1961). Investment in Human Capital. *The American Economic Review* , 1-17.

[15] Becker, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. *Journal of Political Economy* , 9-49.

[16] Flanagan, J. (2014). *Tinder-style matchmaking helps you bag your next job*. Retrieved from New Scientists: https://www.newscientist.com/article/dn25172-tinder-style-matchmakinghelps-you-bag-your-next-job/.

[17] Gus Bekdash, (2019). Using Human History, Psychology, and Biology to Make AI Safe for Humans, Chapman & Hall/CRC.

[18] The Student Circles.com, Artificial Intelligence Study Notes https://www.thestudentcircle.com/quickguide.php?url=artificial-intelligence

AUTHOR

Prof. Yew Kee Wong (Eric) is a Professor of Artificial Intelligence (AI) & Advanced Learning Technology at the HuangHuai University in Henan, China. He obtained his BSc (Hons) undergraduate degree in Computing Systems and a Ph.D. in AI from The Nottingham Trent University in Nottingham, U.K. He was the Senior Programme Director at The University of Hong Kong (HKU) from 2001 to 2016. Prior to joining the education sector, he has worked in international technology companies, HewlettPackard (HP) and Unisys as an AI consultant. His research interests include AI, online learning, big data analytics, machine learning, Internet of Things (IOT) and blockchain technology.

WARRANT GENERATION THROUGH DEEP LEARNING

Fatima T. Alkhawaldeh, Tommy Yuan and Dimitar Kazakov

Department of Computer Science, University of York,
Deramore Lane, Heslington, York, YO10 5GH. UK

ABSTRACT

The warrant element of the Toulmin model is critical for fact-checking and assessing the strength of an argument. As implicit information, warrants justify the arguments and explain why the evidence supports the claim. Despite the critical role warrants play in facilitating argument comprehension, the fact that most works aim to select the best warrant from existing structured data and labelled data is scarce presents a fact-checking challenge, particularly when the evidence is insufficient, or the conclusion is not inferred or generated well based on the evidence. Additionally, deep learning methods for false information detection face a significant bottleneck due to their training requirement of a large amount of labelled data. Manually annotating data, on the other hand, is a time-consuming and laborious process. Thus, we examine the extent to which warrants can be retrieved or reconfigured using unstructured data obtained from their premises.

KEYWORDS

Toulmin model, warrant, fact-checking, and deep learning.

1. INTRODUCTION

The Toulmin model components are necessary for fact-checking as Alkhawaldeh et al. demonstrated in [1]. Argument mining is automatic recognition and extraction of the structure of inference and reasoning expressed in natural language arguments [2]. Habernal & Gurevych [3] identify argument mining as a method for analysing people's argumentation from the computational linguistics point of view and discuss the existing argumentation theories, and they develop a system based on the Toulmin model. Toulmin's arguments should be interpreted as a guideline for concentrating on the most pertinent statements and reasons for supporting or opposing the claim. It is composed of six argument components, as defined in [4]:

- **Claim:** The statement that is being argued to be true. For instance, that cat is most probably friendly.
- **Qualifiers:** Generally, occasionally, in most cases, frequently, few, many, it is possible, perhaps, rarely, in some cases, are all words and phrases that limit claims and are critical for determining the truthfulness of arguments. For instance, students who study more often earn more than students who study less.
- **Data:** Actual data has been gathered to substantiate the perspective (claim). It contains persuasion declarations that add clarity to the claim and demonstrate its truthfulness, such as proof, reasons, opinions, examples, and facts. Data provides evidence to substantiate the perspective (claim). It contains persuasion declarations that add clarity to the claim and demonstrate its truthfulness, such as proof, reasons, opinions, examples, and facts. On the

DOI:10.5121/csit.2021.112005

basis of the data, for example, the following questions could be addressed: "What evidence do you have? "How did you find out? It appears to be raining, for example, the data is that the ground is wet.

- **Warrants**: Reasons why it is critical to make decisions as a supporter or opponent [5]. The warrant will address the following question: "How did you arrive at this claim based on the evidence presented, the logical connection between the data, and how did you resolve this claim."?
- **Backing**: Justification for the warrant as a more specific illustration to substantiate the warrant.
- **Rebuttals/Counterarguments**: Demonstrate an opposing viewpoint as exceptions to the claim and consider other conflicting points of view. For example, social media platforms can communicate with multiple faces using a necessary face for social needs.

An example Toulmin argument is as follows:

- Claim: You should use social media.
- Data: You have been having more trouble with socialising lately, and over 70% of people over age 65 have social difficulty. So, social media is a good chance for elders.
- Warrant: Many social media users say it helps them to be social better.
- Backing: 80% of social media users report a better socially and comfortable lifestyle.
- Rebuttals: 60% of old social media users suffer from a lonely feeling.
- Qualifiers: In most cases, 62% of social media users are well known in the community.

Despite the fact that utilising a warrant can aid in the performance of fact-checking tasks [1], to our knowledge, no previous work has proposed that a claim be connected to a piece of evidence via automated warrant creation rather than manual annotation. Additionally, no experiment was conducted using a labelled dataset, but rather through the use of case studies [6]. Unlike previous approaches that relied on structured annotated warrants [7] or manually generated warrants for emerging claims based on certain linguistic rules [6] that require a higher level of language comprehension and complex reasoning, our work is based on the automated generation of warrants for claims.

This paper examines how to train models to generate warrants data, to address the critical issue of a lack of labelled data for emergent rumours. The works makes the following major contributions:

- In this paper, we have examined the extent to which warrants can be retrieved or reconfigured using unstructured data obtained from their premises.
- To our knowledge, this is the first time that this novel integration of reinforcement learning, and a generative adversarial network has been used to solve the warrant generation problem and alleviate the scarce of labelled data. For this, we have proposed various Deep Learning models for Toulmin Argument warrant generation in this paper.
- We have demonstrated the performance of each of these models and the benefit of combining them with a reinforcement learning agent to improve generation and inference accuracy.
- The results confirm that combining our model with auxiliary data such as the topic and sentiment is necessary to obtain a more robust model. Incorporating a reinforcement learning agent enables the generator to receive rapid and robust training for decoding sequential text successfully.

The remainder of the paper is organised as follows. In Section 2, we discuss the utility of warrants and related work with warrant generation. Then, with an emphasis on the news domain, we propose a novel approach to warrant exploration in Section 3, where we address warrant information filtering and the Generation Model. Section 4 discusses the experiments and the findings, while Section 5 draws the conclusion of the work.

2. RELATED WORK

Toulmin's model of argument has been examined in different of fields, including law and computer science. For instance, in multi-agent systems, multiple agents collaborate to make decisions and inferences to accomplish specific goals [8], where it can be used to generate argument, as in Gabriel et al.'s Belief–Desire–Intention software model (BDI) agents based on Toulmin's models [8], [9]. The warrant is an implicit (or major) premise in the Toulmin argument model that explains how a conclusion (or claim) is deduced from the given premises (or evidence) [10]. According to Hashimoto et al. [11], a warrant is a fictitious logical inference assertion that links the claim and the evidence.

A few works have studied and analysed the task of generating the connection between the claim and the data. In our work, this is referred to as the warrant; in other works, it is referred to as the enthymeme [12] or implicit premise [13], which is typically the warrant (or major premise). Reisert et al. [6] assume that the data are accurate: If the data are accurate, the argument is true. The authors develop a model to generate Toulmin's argument using NLP techniques and some linguistic rules. They demonstrate that argument generation requires a greater understanding of language and complex reasoning and that their system requires significant development to perform argument generation. Boltuzic and Najder[14] investigate how to identify such implicit knowledge by analysing a large amount of text data from a variety of sources. In Habernal&Gurevych's work [14], the warrant is implicit because it is obvious from the statement's meaning, but Rajendran et al. indicated that if it is explicitly required, the argument synthesis method should be used [15]. Rajendran et al. [15] propose a method for creating a premise similar to a warrant in online review opinions that connects an aspect-related opinion to an overall opinion. However, their work's annotated dataset was insufficiently large to be useful for deep learning models. Singh et al. [7] manually generate a warrant in response to a claim and supporting evidence. In Horne & Adali's work [16], workers are asked to think and write what they believe is necessary to explain why the provided evidence supports the provided claim.

3. WARRANT GENERATION MODELS

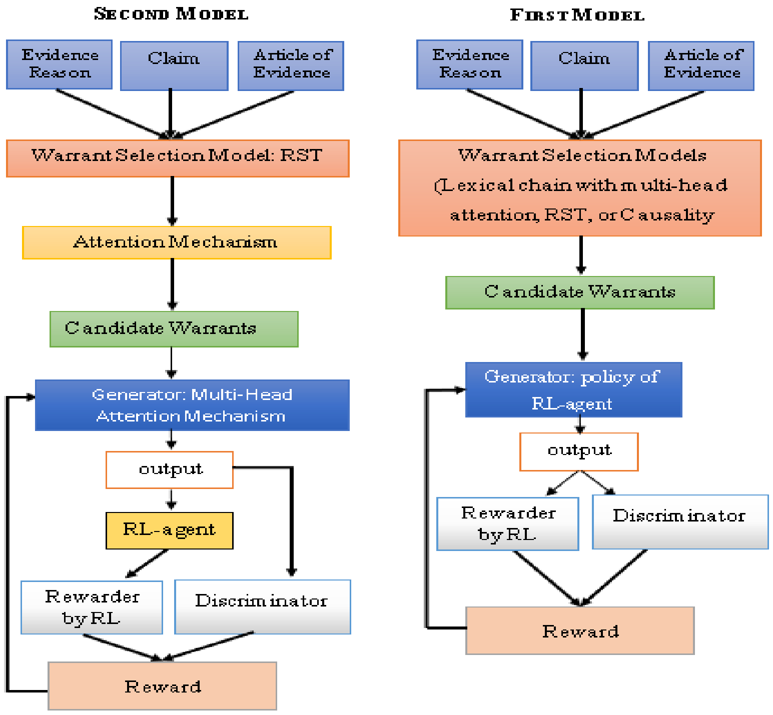


Figure 1. Warrant generation models

Figure 1 shows the overall framework of our warrant generation models. We develop two generator models; the first model trains a reinforcement learning agent to act as a generator, while the second model employs a reinforcement learning agent to enhance different generator via multi-head attention. The purpose of implementing these models is to determine which strategy produces the most promising results: using the RL agent as a generator or as a generator enhancer. The first model has two stages: the initial stage selects warrant-relevant fragments using various methods such as RST and causality, and the second stage selects warrant-relevant words to generate warrants via reinforcement learning agents. While the second model relies on RST and a deep learning mechanism to select candidate warrant relevant fragments, this model utilises a Multi-Head Attention Mechanism enhanced by reinforcement learning to generate warrants.

3.1. model 1 for warrant generation using an RL generator

3.1.1. The Initial Stage: Models for Identifying Warrant-relevant Fragments

The first stage in our warrant generation process is to select (retrieve) information that is pertinent to a claim and unstructured evidence. Increasing the efficiency of false information detection requires developing the ability to recognize the connection between an argument and a piece of evidence. Multiple warrants have been selected from an existing, organized corpus of arguments using developed methods. Our proposed models include a Lexical Chain with Multi-Head Attention, an RST-based algorithm, and a Causality-based selection method, all of which are aimed at capturing more compelling reasoning warrants. Table 1 illustrates an example of the most pertinent information contained in a warrant in light of a claim and evidence which are bold.

Table 1. An illustration of the task of locating the most pertinent information to the claim and supporting evidence from the ARC [17].

Claim: "Greece will destroy the Euro Zone"
Evidence Reason:"Greece cannot support its own economy and is bringing the Euro down"
Article of Evidence: "The euro zone is now furiously bracing itself for the likely collapse of the Greek government. Faced with the prospect of Greece voting for a fully-fledged default and euro exit rather than last week's debt deal, the remaining euro zone members must themselves choose: stick even more closely together or be pulled apart. They will stick together – and survive. **However, the euro zone's survival has very little to do with Greece**. The Greek economy is too small to cause any noticeable impact on the euro zone and even the widespread and substantial financial contagion of a default can be absorbed. Last week's debt deal may not appeal to Greece, but the beefed-up bailout fund is capable of taking care of the immediate consequences of a Greek default. Containment has been addressed and would focus on supporting other indebted states. ***The euro zone's survival has little to do with Greece except to persuade other members to redouble their efforts and stick with the euro. The key reason for Greece continuing to play an important role in deliberations over the euro zone's future is that it highlights the question mark over member states' abilities to resolve the deep-rooted problems of poorly performing economies. The influence that Greece can still wield is a demonstration effect: If Greece leaves, will the result be disastrous or could the economy be galvanized into a better performance, as those who favor exit appear to believe?*** "

3.1.1.1. Lexical Chain with Multi-Head Attention

Inspire by the data retrieval, question answering, and response selection models, a claim is viewed as a query and evidence as an appropriate document from which the candidate's responses should be selected. The lengthy text (as evidenced by ARCC) data will be condensed for warrant selection using the lexical chain model to retain the most informative words that are also the quietest to draw attention to the claim outputs (or a query).

We begin by detecting salient portions of text using Word Sense Disambiguation (WSD) and then extracting the lexical chains described in Al-Khawaldeh & Samawi [18]. In contrast to Al-Khawaldeh & Samawi [18], the proposed model attempts to select sequences from each cluster associated with the claim instead of selecting the sequences that are significant to different topics as in Al-Khawaldeh & Samawi [18]. For example, as in table 1, suppose we have "Greece will destroy the Eurozone," as the evidence reason ". To obtain the correct sense of the term ("zone") ("its senses must be extracted at three levels). Thus, the first level extracts all possible senses for the "zone," the second level extracts the senses for these senses, and so on for the third level. The sense of a word refers to how its meaning is detonated when it is used in a specific context.
The developed WSD algorithm consists of five steps as in Al-Khawaldeh & Samawi [18]:

1) Extract all the possible interpretations (senses) of each word in a sentence of evidence. Extract the three levels of senses for each sense, the first level is the senses of a word; the second level is the senses for each sense in the first level and so on,
2) Each word's senses are compared to the senses of all other words in the text and then establish connections between the related senses, a connection is established when there is a semantic relationship between the current word's senses and any other word's senses.
3) Calculate the strength of the connections.
4) Summing all the strengths of the connections.
5) Select the highest summation sense.

By empirically, the semantic relations and their associated weights are as follows:

- Repetition relation (same occurrences of the word), weight=1.
- Synonym relation (weight=1). In the example above, the word "zone " has a synonym semantic relation with the sense ("area")
- Hypernym and Hyponym relation (weight=0.5): Y is a hypernym of X if X is a (kind of) Y; X is a hyponym of Y if X is a (kind of) Y e.g., X=" zone", Y=" ground"
- Holonym and Meronym relation (weight=0.5): holonymy relation is (the whole of), and meronymy relation is (part of). Y is a holonym of X if Y is a whole of X; X is a meronym of Y if X is a part of Y. X= "state", Y="zone".
- Gloss relation (definition and/or example sentences for a synset), (weight=0.5): consider the word=" zone", gloss=" area having a particular characteristic".

Each sense has several weighted connections to other words' related senses. The weighted connections between the senses are added together. Lexical cohesion is used to differentiate between significant and unimportant sentences in a text. The text is segmented by lexical cohesion. Each segment consists of a series of sentences devoted to a single subject. Each word is assigned the correct sense after the proposed WSD algorithm is applied to the text above. Lexical chains (LCi) are formed by connecting the words' senses (meanings). If these senses have semantic relationships, then the words are related.

LC1:{money, account, transfer, cash, withdraw, bank}
LC2:{area, ground, region, segment, sector}

To begin, we use a Bi-RNNc to model the embeddings of claim words cl and chain words c, where $h^{c}_{i,1}$ denotes the hidden state of the t-th word in the i-th chain and $h^{cl}_{i,1}$ denotes the hidden state of the t-th word in the i-th claim. Following that, we perform an average-pooling operation on these hidden states, eq. 1, to generate a vector representation of the i-th chain, eq. 2

$$a_{vi} = avg\left(\left\{h^{c}_{i,1}, h^{c}_{i,2}, \dots, h^{c}_{i,T^{c}_{i}}\right\}\right) \qquad (1)$$

$$m_i = tanh\left(W_{cl} \cdot \left[a_{vi}; h^{cl}_{j,T^{cl}_{j}}\right] + b_{cl}\right) \qquad (2)$$

Mi can be thought of as a salience score for the i-th chain in the context of the claim representation, $h^{cl}_{j,T^{cl}_{j}}$.. The highest sigmoid output indicates the chain's importance in relation to the claim; thus, the selected segment of evidence should be chosen based on this critical chain, which allows for the omission of irrelevant text. To model the relevancy of the segment of text towards the strongest chain, we first calculate the word alignment of the segment towards the chain. We use the embeddings of words in chain and segment to calculate the semantic alignment score as shown in equations 3 and 4:

$$\mathrm{score}_{\mathrm{i,j,n}} = \mathrm{e}(\mathrm{A}^{\mathrm{c}}_{\mathrm{i}})^{\mathrm{T}}\, \mathrm{e}\left(\mathrm{A}^{\mathrm{s}}_{\mathrm{j,n}}\right) \qquad (3)$$

$$maximum_{i,j} = max\left(\left\{score_{i,j,1}, \dots, score_{i,j,T^{c}_{j}}\right\}\right) \qquad (4)$$

Where $e(A^{s}_{i})^{T}$ is word embedding in the segment, and $e\left(A^{c}_{j,n}\right)$ is word embedding in the chain, $score_{i,j,n}$ is the attention wei for the i-th chain word with the j-th segment word, s is a segment, c is the chain, n is the segment number, i is the index word of the segment, and j is the index word in the chain.

The alignment score, maximum-i,j, is the weight assigned to the jth chain in relation to the ith segment word. We take the highest attention weights from all scores and represent them as candidates' parts, retaining only the relevant parts.

After selecting the most informative text from the evidence and obtaining reduced text, we will use multi-head attention to construct deep contextual representations for tokens located in different representation subspaces at different positions while preserving their syntactic form.
This model's general framework is divided into four steps.

- Apply word embedding for each word in the text.
- Use a BiLSTM and CNN to obtain the vector representation of the text.
- A multi-head attention mechanism that can capture relevant information from different subspaces
- Use SoftMax layer for text classification to select the candidates warrant.

The Elmo word embedding model will represent each word in each sentence as a deep contextual deep word representation. Elmo is a sophisticated, contextualised word representation that extracts the word's complex syntactic and semantic features [19]. On a variety of natural language processing tasks, including query answering and textual entailment, Elmo outperforms previous word embeddings such as word2vec and GloVe [20]. By reading each sentence in two directions: from beginning to end (forward) and from end to start (revers), we extract the most critical information and obtain contextual information about the current word using a CNN and a Bi-LSTM. The final encoded representation combines the Bidirectional hidden state representation and the Bidirectional hidden state representation

Multi-head attention layer for claim-evidence text: A specific section of the text is critical in identifying the candidate warrant in a given claim-evidence. Numerous heads of attention assign each word the appropriate weight to represent the text's general semantics based on various factors. This work makes use of self-awareness to capture the relationship between the claim-evidence pair and the warrant.

In contrast to multi-head -attention from the literature, which typically considers V=K=Q and is derived from the same source, we define Q as each word in a candidate warrant is required to perform an attention calculation using all other claim words as key-words, where the warrant is a candidate sentence from the article. The attention layer receives three input texts: a claim text as a key, a candidate warrants as a query, and an evidence text as a value. Each of them contains a word vector containing all of the words in the input text.Multi-head refers to paying attention not only to the individual words in the sentence but also to the individual segments of the words. The vectors of words are divided into a fixed number of chunks (h, number of heads), and then multi-head attention is applied to the corresponding chunks, resulting in an h context vector for each word. The final values vector is created by concatenating all of that h to generate an encoded representation for each word in the input sequence (representation vectors) and add the word's attention score taking a walk through the primary steps of the example from the ARC [17]:

- Candidate warrants from an article (query Q): “money will not be saved all the way around”
- Claim (keys K): “Privatization is a bad deal for cities and states.”
- Evidence (values V): “The only interest of the private sector is the bottom-line profits.”
  - The query is the input word vector for the Candidates warrantstoken, e.g., “money”.

- The keys are the input word vectors for all of the claim's tokens [Privatization, is, a bad, deal, for, cities, and, states]
- The query's word vector is then DotProducted with the word vectors of each key, yielding n numbers, i.e., "weights." Following that, the weights are scaled.
- The weights are then subjected to a ‘SoftMax' operation, which normalises all weights to values between 0 and 1.
- Finally, the input word vectors, e.g., values, are summed in a "weighted average of the value vectors " using the previously normalised weights. It generates a single output word vector representation of the Candidates warrantsword, as in equations 5 and 6:

$$\mathrm{Attention}\ (\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \qquad (5)$$

$$\mathrm{head_i} = \mathrm{Attention}\left(\mathrm{QW_i^Q}\,,\mathrm{KW_i^K}\,,\mathrm{VW_i^V}\right) \qquad (6)$$

- All word vectors are getting similarly; the attention mechanism is applied to all word vectors. **single output word vector representation** of “Privatization” is finally obtained and so on for all words, resulting in o **output word vector representation,** as shown in equation 7:

$$O = MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^o \qquad (7)$$

O is the output of multiple attention functions used in multi-head attention capturing explicit and implicit patterns. It converts Q, K, and V subspaces to C subspaces using various learnable linear projections. To capture various contexts, information from various representation subspaces at various positions can be prioritised. Each head generates an attention distribution to its subspace to represent the final state when all attention heads are considered. The independent operation's result is then spliced into a linear transformation. To obtain the multi-head attention result M, as in [21]. We construct an auxiliary feature vector from the topic T and sentiment vectors S; the concatenated features are TS. Assuming that those features are consistent across inputs, we combine them with the output of multiheaded attention O to create a new representation, Onew=O+TS; all words vectors are concatenated as S= Onew1, Onew2…. Onew n. Then, using a SoftMax layer as an activation function, classification is performed. Thus, the probability of current candidates warrants Y, as shown in equation 8:

$$Y = softmax(W * S + b) \qquad (8)$$

3.1.1.2. RST-based Algorithm

Due to the causal and semantic relationship between claim, evidence, and warrants, we were inspired by RST's discourse analysis, which identifies a rhetorical relationship between two text spans, nucleus and satellite, where the nucleus contains more informative text than the satellite, which contains additional information. Given that warrant provides reasoning for a claim in the form of cause, purpose, motivation, and circumstance, in our model, the nucleus (span) of the RST relation is matched against the claim and the relationship (primarily implicit or explicit causal) with the satellite; the best candidate warrant is determined by the most pertinent RST relation between the claim and the warrant span discourse units.

RST can be used to describe the relationships between text's internal components. RST relations divide the text into rhetorically related segments that may be further divided, resulting in a hierarchical rhetorical structure. Each segment corresponds to a nucleus or satellite. It demonstrates that coherence relations can have a beneficial effect on both the claim and the justification. For instance, the nucleus contains an idea that the author regards as the nucleus.

We will use RST to conduct discourse analysis, which identifies rhetorical relationships between two text spans: nucleus and satellite, with the nucleus containing more informative text than the satellite, which contains additional information. Several RST relationships could help to explore warrant information from text: as in table 2. Because a warrant justifies a claim, it serves as the cause, purpose, motivation, and circumstance. The nucleus (span) of the RST relation is matched against the claim and the relationship (primarily implicit or explicit causal) with the satellite in our model; the best candidate warrant is determined by the most relevant RST relation between the claim and the warrant span -discourse units. Heilman & Sagae's work will be used to implement RST [22]. An example of a nucleus or satellite, where the claim “I believe the weather is cold and wet” is the nucleus and the supplementary text “since the temperature has decreased by 15 degrees Celsius” is a satellite, connected with the *explanation* rhetorical relation. In this example, the satellite clause explains the nucleus, as in argumentation model such as Toulmin model, the warrant is supplementary for main information, claim, so our work considers warrant is satellite and claim are the nuclei.

Based on this complementary relationship between satellites and nuclei, we argue that certain words in certain nucleus-satellite relationships may be more significant than others, e.g., they indicate the clause has a warrant. Thus, we argue that a satellite should be considered when determining a warrant in a case where the satellite is linked to the claim's nucleus. On the other hand, we argue that the nucleus does not contribute to the satellite's understanding. Thus, words contained within a satellite differ from those contained within a nucleus, as in figure 2:

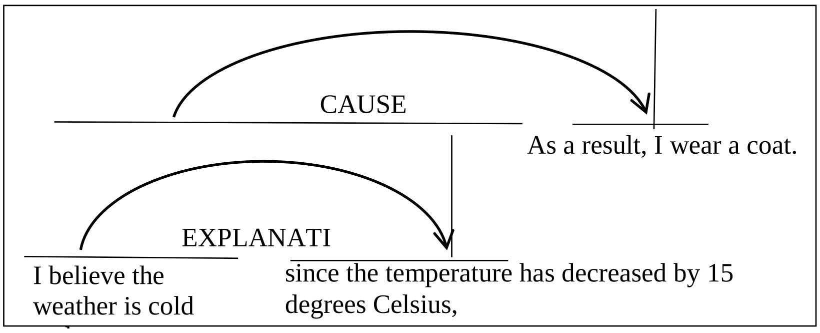

CAUSE
As a result, I wear a coat.
EXPLANATI
I believe the weather is cold
since the temperature has decreased by 15 degrees Celsius,

Figure 2. The relation between a nucleus and a satellite, an example of nucleus or satellite, with RST relation

The RST-based algorithm to select a warrant for a claim is as follows:

1. Input: evidence text, claim query, query expansion
2. Result: warrant
3. Begin
4. Segment texts to clauses based on cure phrases (connectors words)
5. Find rhetorical relations between the clauses to build all RS-trees for evidence text
6. Check the rhetorical relations between the segments: nucleus and satellite, e.g., explanation, interpretation, result or justification.
7. If a segment is a nucleus and is relevant to claim query or query expansion, then the satellite is a warrant and vice versa.
8. Save as candidate part of the warrant and continue to the next candidate warrant
9. End

Table 2. Organization of the Relation Definitions [23]

Circumstance	Antithesis and Concession	Enablement	Otherwise
Summary	Antithesis	Motivation	Interpretation and Evaluation
Elaboration	Concession	Evidence and justify	Interpretation
Background	Condition and otherwise	Evidence	Evaluation
Enablement and Motivation	Condition	Justify	Restatement and Summary
Relations of Cause	Restatement	Purpose	sequence

3.1.1.3. Causality-based Selection

The causal relationships provide knowledge that allows for the interpretation of the evidence-based claim. As the warrant explains how the data leads to the claim, it is necessary to recognise causalities expressed explicitly in answer phrases such as "because" and to use those recognised causalities as a guide for locating proper answers. Causalities expressed in one text may be expressed with explicit cues in other texts. in the form of texts expressing causal relationships (e.g., "[Tsunami occurred] effect as a result of [a sudden displacement of sea water] cause"). If we can identify causal relations in which the effect part corresponds to a target why-question, the cause parts may contain useful information for generating appropriate compact answers, such as important keywords to include in the compact answers. We retrieve causal relation expressions that are relevant to claim C, such as effect and cause relevant statements, given a target claim C.. Thus, we automatically extract causal relations relevant to a target why-question from the web, such as "[Microsoft's machine translation has made significant progress in recent years] effect since [it began using deep learning] cause":

Because the warrant has a casualty and a reason, we used a why–how to approach in our work. A contrast relationship implies adversarial justification (rebuttal). The event causes demonstrate what occurs (effect) in a claim and a warrant. Table 3 illustrates several of these relationships and the position of claim and warrant and evidence. The presence of causality is checked in a sentence, where causality refers to the relationship between cause and effect in a sequence of events. Oh, et al. [24] suggested Causality-attention: A convolutional neural network with multiple columns for why-QA.

Table 3. Examples of Causality Relations

Claim ***as a result of***warrant and evidence	***seeing that*** warrant and evidence, the claim	warrant and evidence ***this led to claim***
because of warrant and evidence, the claim	*Claim* ***So*** warrant and evidence	***this cause*** warrant and evidence, claim
warrant and evidence ***Consequently*** claim	the claim***as a consequence of***warrant and evidence,	***in order to***warrant and evidence, the claim
due to warrant and evidence, the claim	warrant and evidence ***the reason, claim***	warrant and evidence, the warrant and evidence ***resulting in*** the claim
due to the fact warrant and evidence, the claim	warrant and evidence ***therefore claim***	warrant and evidence ***Thereby claim***
on account of warrant and evidence, the claim	***for this reason,*** warrant and evidence, the claim	warrant and evidence ***Similarly claim***

The claim expansion process in our work is inspired by (question query Q) [25]–[27], which employs a word embedding to expand the query (in our work, claim) and wordnet expansion [28]. The model checks for hypernyms, such as food, and hyponyms, such as fruit, in addition to meronyms and holonyms; a branch is a meronym (part meronym) of a tree, whereas heartwood is a meronym (substance meronym) of a tree, and the forest is a holonym (member holonym) of a tree. If the evidence text has causality with the claim or is highly semantically related to the claim (more connected to the claim), those texts will receive additional scores as part of the candidate's warrant.

Along with the most closely related parts by wordnet relation, two types of attention mechanisms will be used to score the candidates' warrants: similarity-attention [29] and causality-attention [24]. The similarity-attention mechanism calculates the cosine similarity between the embeddings of claim and evidence text to generate an attention feature vector for evidence words. In contrast, causality attention focuses on evidence words causally related to claim words and is used to generate causal embeddings focusing on causal relations to generate a causality attention feature vector. When confronted with passages containing possible causes/reasons for a given claim, causality attention can be focused on words and their contexts. The matrix of causality-attention features is constructed using scores indicating the degree to which two words are causally related (one in a claim and another in a warrant passage).

3.1.2. The Second Stage of Warrant Generation: RL for Identifying Warrant-Relevant Words

Candidate warrant selection techniques will be analysed to ascertain the warrant's scope (to retrieve the warrant). We propose to collect significant, warrant-relevant words from a lengthy fragment using reinforcement learning RL (through actions). RL shows a promising result in different method [30]–[32] where the model acquires knowledge through interaction with its environment and is rewarded for completing tasks. In [33], text generation is formulated as the sequential decision-making problem

Due to the discrete nature of the data and no gradient can be obtained, we use RL to guide our sequential decision policy network's training and use lexical in nature measures for evaluation a reward function, for example, rouge or BLEU. We hypothesise that a sequential decision policy network can aid in the detection of warrants. A delayed reward is used to direct the policy's learning process based on the interaction of predicted and actual warrants. As illustrated below, we briefly discuss state, action and policy, motivation, and objective function.

Given a candidate warrant's word sequence$\mathrm{w_i}, 1, \dots, \mathrm{w_i}, \mathrm{k_i}$ the policy network πl attempts to select the warrant-relevant word $\mathrm{w_i}, \mathrm{j}$ and eliminate irrelevant ones. The policy network employs a stochastic policy to check the probability of an action at each state, and it learns through delayed reinforcement after the sequence of actions is completed. We construct the policy πl for selecting words over a word sequence using the Bi-GRU model. We use Bi-GRU because it has fewer parameters than LSTM and thus performs more quickly with efficiency [34].

State (st): given the claim, evidence and candidate warrant as input, the policy aimed to decide the warrant relevant words as delete, keep or generate. Afterword embeddings $e_i$is performed, we use Bi-GRU to get the vector representation of candidate warrant $h_s^{(1)} + h_s^{(1)} + h_s^{(2)} + \cdots + h_s^{(n)}$. Following the acquisition of claim and evidence hidden state representations, we then pool the vectors on an average basis $claim^{(l)}$ and $evidence^{(l)}$ through equations 9-13:

$$\vec{h}_i^{(1)}, \overleftarrow{h}_i^{(1)} = bGRU\left(e_i, \vec{h}_{i-1}^{(1)}, \overleftarrow{h}_{i+1}^{(1)}\right) \qquad (9)$$

$$h_i^{(1)} = W_1\left[\vec{h}_i^{(1)}, \overleftarrow{h}_i^{(1)}\right] \qquad (10)$$

$$claim^{(l)} = \frac{1}{N-1}\sum_j h_j \qquad (11)$$

$$evidence^{(l)} = \frac{1}{m-1}\sum_j h_j \qquad (12)$$

$$st = h_s^{(n)} + claim^{(l)}evidence^{(l)} \qquad (13)$$

To produce a vector representation for both, claim and evidence, we use average-pooling operation over hidden states as shown in equations 14 and 15.

Action: A stochastic policy uses state information for deciding to select the current word or not. We adopt a logistic function (conditional probability) to decide whether this word is relevant for a warrant or not, as in equation 14.

$$action = sigmoid(W * st + b) \qquad (14)$$

Reward-1: We employ attention mechanisms at each stage of text representation, the actual warrant and predicted warrant. By assigning weights to encoding vectors, it is possible to highlight specific parts of the input that are more important for detecting warrants, candidate warrant CW, and actual warrant AW similarity, as in equation 15-20.

$$u_{ij} = tanh\left(W_w \cdot \left[h_{ij}; CW\right] + b_w\right) \qquad (15)$$

$$a_{ij} = softmax\left(u_{ij}\right) = \frac{exp(u_{ij})}{\sum_{t=1}^{N} exp(u_{it})} \qquad (16)$$

$$CW = \sum_{i=1}^{N_i} a_{ij} \cdot h_{ij} \qquad (17)$$

$$u_{ij} = tanh\left(W_w \cdot \left[h_{ij}; AW\right] + b_w\right) \qquad (18)$$

$$a_{ij} = softmax\left(u_{ij}\right) = \frac{exp(u_{ij})}{\sum_{t=1}^{N} exp(u_{it})} \qquad (19)$$

$$AW = \sum_{i=1}^{N_i} a_{ij} \cdot h_{ij} \qquad (20)$$

Finally, reward guides the policy regarding the selection of warrant-relevant words within a warrant sequence. We use the connection of vectors and the SoftMax function to combine the predicted warrant CW a representation and the actual warrant AW representation for similarity classification, as in equation 21:

$$Y = SoftMax(W[CW \oplus AW] + b) \qquad (21)$$

Semantic coherence Reward 2: the generated warrant to check if it is grammatical and coherent as in equation 22:

$$r_{SC} = \frac{1}{N_y} log\, P_{seq2seq}\,(y|x_i) + \frac{1}{N_{x_1}} log\, P_{backward-seq2seq}\,(x_i|y) \qquad (22)$$

Pseq2seq denotes the likelihood of the seq2seq model (the probability of generating the predicted warrant given the previous warrant). Pbackward seq2seq denotes the backward probability of actual warrant given the current generated warrant.

In previous work [35], we trained separate models (single agents) to locate the warrant given a claim and evidence. The first model employs Lexical Chains, as proposed by Al-Khawaldeh and Samawi [18], which aid in extracting the most informative words and thus reducing the text's size. After obtaining the summarised text, the claim's related fragments and evidence are captured using the multi-head attention model. The second model employs a Rhetorical Structure Theory-based algorithm to segment each text into two spans, nucleus and satellite, with a higher probability of being nucleus. Finally, the causality model: because the warrant possesses a causal and rational nature, the causality relations denote the text fragments that contain one of the following relations: justification, interpretation, or confirmation. These are more extraction-oriented models than generation-oriented models. As a result, our model attempts to generate warrants by combining multi-head attention theory and rhetorical structure theory.

3.2. model 2 for warrant generation using a Multi-Head Attention Mechanism generator enhanced by RL

In model 1, we conduct experiments with a reinforcement learning agent as the generator, whereas in model 2, we use reinforcement learning as an enhancer for the generator to determine which is more effective. We develop justifications for why an argument is persuasive, discovering that adding word embedding features improves performance. Given a claim c = c_1; c_2; …; $c_k$containing k words, and an evidence d = d_1; d_2; … d_n consisting of n words, the objective is to generate a warrant for the context y = y_1; y_2; …$_{ym}$ containing m words. The objective is to find an output Y* that maximizes the probability p(Y| c ; d), Y is the warrant, and c and d are claim and evidence, respectively.

The RST based algorithm is used to locate a warrant for the claim, as in section 3.1.1.2.
We take each word as input to get the claim embedding vectors as in equation 23.

$$e_c = \{e_c^1, e_c^2, e_c^3 \dots e_c^n\} \qquad (23)$$

Similarly, the candidate warrant is also embedded as vectors as in equation 24.

$$e_w = \{e_w^1, e_w^2, e_w^3 \dots e_w^m\} \qquad (24)$$

Then we apply cosine similarity to compute the final score as the relevance of a claim to a warrant to detect the candidates' warrants: score (claim, candidates warrant) = cosine similarity (e_c,e_w). The highest score means that it is more likely that the warrant is plausible. The model adopts BiGRU to represent both claim rc and candidate warrant rw because it operates well in learning long term dependencies and is fast in training.

To reduce the spatial size of the representation and retain essential features, we adopt mean pooling to calculate the claim $mcl^{(cl)}$, evidence $mev^{(ev)}$ and warrant $m^{(w)}$ pooling vectors through the equations 25-27:

$$mcl^{(cl)} = \frac{1}{N-1}\sum_i \quad h_{claim}^{(i)} \qquad (25)$$
$$mev^{(ev)} = \frac{1}{M-1}\sum_i \quad h_{evidence}^{(i)} \qquad (26)$$
$$m^{(w)} = \frac{1}{K-1}\sum_i \quad h_{warrant}^{(i)} \qquad (27)$$

We define the attentive representation of claim, evidence, and warrant in relation to one another, i.e., the attentive representation of the effect phrase concerning the cause phrase, to consider the score and impact of each of them on the other, as follows:

The claim representation with its candidates' warrants clw_t as the equations 28-30:

$$a_{t,i}^{cl} = v_{cl} \cdot \ tanh \left(W_1 m^{(warrant)} + U_{cl} h_i^{claim} \right) \qquad (28)$$

$$\alpha_{t,i}^{cl} = \frac{exp\left(a_{t,i}^{cl}\right)}{\sum_{i=1}^{|cl|} \quad exp\left(a_{t,i}^{cl}\right)} \qquad (29)$$

$$clw_t = \sum_{i=1}^{|cl|} \quad \alpha_{t,i}^{cl} h_i^{claim} \qquad (30)$$

The candidates warrant representation with their claim wcl_t as in equations 31-33:

$$a_{t,i}^{w} = v_{w} \cdot \ tanh \left(W_2 m^{(claim)} + U_{w} h_i^{warrant} \right) \qquad (31)$$

$$\alpha_{t,i}^{w} = \frac{exp\left(a_{t,i}^{w}\right)}{\sum_{i=1}^{|w|} \quad exp\left(a_{t,i}^{w}\right)} \qquad (32)$$

$$wcl_t = \sum_{i=1}^{|w|} \quad \alpha_{t,i}^{w} h_i^{warrant} \qquad (33)$$

The evidence representation with its candidates' warrantsevw_a^t as in equations 34-36:

$$a_{t,j}^{ev} = v_{ev} \cdot tanh\left(W_{ev} m^{(warrant)} + U_d h_j^{evidence} \right) \qquad (34)$$

$$a_{t,j}^{ev} = \frac{exp\left(a_{t,j}^{ev}\right)}{\sum_{j=1}^{|ev|} \quad exp\left(a_{t,j}^{ev}\right)} \qquad (35)$$

$$evw_a^t = \sum_i \quad a_{t,j}^{ev} h_{a,i}^{(evidence)} \qquad (36)$$

The candidates warrant representation with its evidence wev_a^t as in equations 37-39:

$$a_{t,j}^{w} = v_{w} \cdot tanh\left(W_{w} m^{(evidence)} + U_d h_j^{warrant} \right) \qquad (37)$$

$$a_{t,j}^{w} = \frac{exp\left(a_{t,j}^{w}\right)}{\sum_{j=1}^{|w|} \quad exp\left(a_{t,j}^{w}\right)} \qquad (38)$$

$$wev_a^t = \sum_i \quad a_{t,j}^{ev} h_{a,i}^{(warrant)} \qquad (39)$$

Finally, we combine all these representations for *causal/noncausal in* equation 40:

$$Y = softmaxY(clw_t + wcl_t + evw_a^t + wev_a^t) \qquad (40)$$

Causal/noncausal Y means the candidates warrant either plausible or not.

Multi-Head Attention Mechanism with Multiple Heads: This model employs the transformer network [21], which is based primarily on deep learning and dot products and is composed of fully connected layers from both the encoder and decoder. It replaced recurrence or convolution with the multi-head -attention transformer's encoder, composed of six identical layers, each of which is composed of two sub-layers: a multi-head -attention mechanism and a position-wise fully connected feed-forward network [36]. A residual connection and layer normalisation are used to generate outputs from two sublayers. The transformer Decoder is also composed of a stack of identical layers to the encoder, except that it includes a third sublayer that implements a multi-head attention mechanism over the encoder's output, as illustrated in figure 3.

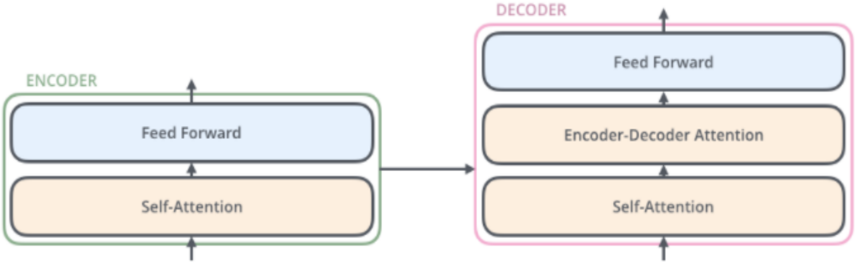


Figure 3. Transformer Encoder-Decoder Architecture [36]

To capture the relationship between words in various positions, it computes the relevance of a set of values (information) using the same attention mechanism. In practice, the attention function is computed concurrently on a set of queries. It computes the attention function for a matrix Query, Keys, and Values that contains a collection of queries, keys, and values. Each head corresponds to a layer of attention [36]. The encoder converts a sequence of discrete representations in the form X = (x1;...xh) to a sequence of continuous representations in the form z = (z1; ... zh). In our work, X refers to the claim, evidence, and the average embedding of selected warrants used to generate warrants. the decoder then generates an output sequence consisting of one element at a time (y1;...yh).For the multi-head attention mechanism, h = 8, implying the use of eight parallel attention layers. To ensure the model's sequence, positional encoding is added to the input embeddings at the end of the encoder and decoder stacks. It can use embedded vectors to represent the relative positions of each sentence's words and then combine them with the sentence embeddings, as in equations 41 and 42:

$$Z_i = head_i = \ attention\big(QW_i^Q\ ,KW_i^K\ ,VW_i^V\big) \qquad (41)$$

$$Z = MultiHead(Q, K, V) = Concat(Z_1, \dots, Z_h)W^o \qquad (42)$$

Our model takes as input a claim concatenated with candidate relevant warrants. After applying word embeddings, W-emb., to input words, we use The BiGRU to capture semantic information about past and future words. BiGRU utilises a forward and backward LSTM as encoder hidden layers to determine the hidden state of the time step t ht. Then, as in Vaswani et al. [21], we use residual connection around the output of the Bi-GRU layer to stabilise the model's training, followed by layer normalisation, as equation 43:

$$h_t^* = LayerNorm(W_{emb} + h_t) \qquad (43)$$

Final encoder layer output H is the output of the add and Norm layer, equation 44.

$$H = (h_1^*, h_2^*, \dots, h_i^*, \dots, h_n^*) \qquad (44)$$

We compute a representation of the sequence using multi-head attention, which is an attention mechanism associated with the various positions of a single sequence. The attention distribution at is calculated as follows: Output H is Query vectors, keys vectors K2, and values vector Ve. The encoder's attention module is largely based on Vaswani et al.’s multi-head attention [21], as in equations 45-47:

$$e^t = \frac{Q^e K^{eT}}{\sqrt{D}} \qquad (45)$$

$$a^t = softmax(e^t) \qquad (46)$$

$$Attention(Q, K, V) = a^t V^e \qquad (47)$$

The multi-head attention adjusts the Q, K and V matrix dimensions by h different linear layers to h queries, keys, and dimension values. The linear transformation parameters W of Q, K and V, are different each time based on the learnable parameter's matrix for the heads. Then, h parallel heads are used to concentrate on distinct semantic spaces. The result of the independent operation is spliced into a linear transformation to obtain the result ce of multi-head attention, as in equations 48 and 49:

$$head_i = attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \qquad (48)$$

$$ce = MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^o \qquad (49)$$

Then decoder d generates word by word based on:

- The encoder e attention context ce is the output of multi-head soft-attention of sequence words input.
- The recurrent attention context, c_t^{ed}, it is based on each hidden state s_t of the decoder as query and hidden state output of the encoder as keys -values vectors of multi-head -attention.
- The decoder attention context c_t^d, Where multi-head-attention of all the predicted tokens is used.
- The decoder hidden state s_t. (equation 50) and the vocabulary probabilities (equation 51)

$$s_t = GRU\left(s_{t-1}, Y_{t-1}, c_{t-1}^{ed}\right) \qquad (50)$$

$$Pv = softmax\left(W'\left(W\left[c_t^e, c_t^{ed}, c_t^d, s_t\right] + b\right) + b^{'}\right) \qquad (51)$$

c_t^{ed}t is the output of multi-head soft attention. The decoder has an embedding layer, a unidirectional GRU and a SoftMax layer. We use the hidden states of the decoder layer and the final encoder layer output H for obtaining the attention context c_t^{ed}. Besides feeding the attention context to all decoder GRU layers, we also feed it to SoftMax. This is important for both the quality of our model and the stability of the training process.

An encoder-decoder LSTM or GRU network is used to automatically approximate internal states and formulate potential actions for the reinforcement learning agents Sarsa or DDQN. The RL agents take the decoder output at time t as input and estimate each action's advantage values that learn to select an action (e.g., a word) from a list of possible actions to improve the current warrant sequence. For Sarsa, because it is learning an action-value function rather than a state-value function, it differs from Q-learning in that it does not require using the maximum reward for the next state. However, Deep Q-Networks is Q-learning with a deep neural network function that employs an epsilon-greedy policy to select actions for the Q-network approximator. Each decoding iteration will modify the current SARSA or DDQN by predicting which actions should be taken to accumulate a larger long-term reward.

4. Experiments and Results

4.1. Implementation Details

We implement our model using Keras and a pre-trained 300-dimensional Glove word Embedding [37]. The encoder employs 300-dimensional hidden states, while the decoder employs 300-dimensional hidden states. We use the Adam optimizer [38], with both the encoder and decoder set to a maximum of 50 tokens and the batch size set to 32. The hyperparameter values used in a model have a significant impact on its performance. We will discuss how to tune hyperparameters to achieve a more robust and generalised mode. We create our implementation of an algorithm by determining the optimal hyperparameter values for a given task and dataset. We divide the available data into training and testing subsets, then repetition of optimization loop until a condition is met and finally, we compare all metric values enables you to select the hyperparameter set that produces the optimal metric value.

4.2. Dataset

We conduct experiments using data from the ARC [17] repository, which annotated in such away serve our work. Habernal et al. [17] developed the ARCCto discover warrants. It contains 188 debate topics for the argument reasoning comprehension task as in the following example [17]:

> *"Reason: Cooperating with Russia on terrorism ignores Russia's overall objectives.*
> *Claim: Russia cannot be a partner.*
> *AW adversarial warrant: Russia has the same objectives of the US.*
> *W warrant: Russia has the opposite objectives of the US."*

We evaluate our models with the metrics used in Park et al.'s model [39] for the quality BLEU-1/2 and Embedding Average/Greedy/Extreme and the diversity Dist-1/2 and Dist-1/2-within of the generated sentential arguments for each. The two metrics, quality and diversity of generated text, are widely used in Park et al.'s [39] text generation task model and will be used in our evaluation. Given that evidence used to substantiate a claim may cover a variety of aspects of an argumentative topic, the diversity and quality of generated text should be evaluated to determine the breadth and variety of word usage in writing, as well as the vocabulary richness and n-gram precision desired in conversational topics.

The model developed by Park et al. [39] illustrates the evaluation results for each model in terms of generating quality using BLEU and word embedding-based metrics. As we can see, our model outperforms the competition on nearly all metrics. Park et al.'s model [39] demonstrated that our model could generate diverse and multiple arguments to examine various aspects of a given claim. employing the PERSPECTRUM. Park et al. [2019] generate claims in response to a given claim, utilising a diversity penalty to encourage the presentation of diverse perspectives. It utilises a Seq2Seq framework and introduces latent mechanisms on the assumption that each latent mechanism can be associated with a single perspective.

- BLEU-1/2: measures N-gram precision of the generated text to multiple target arguments references [40]
- Embedding Average/Greedy/Extreme: measures the semantic similarity between hypothesis and references, using a semantic representation by word embedding [41]
- Dist-1/2: computes the percentage of unique unigrams/bigrams within a sentence to measure the diversity among multiple generated texts [41]
- Dist-1/2-within [39], propose a simple metric to calculate the sum of the numbers of unique N-grams for each result that does not occur in other results) / (The sum of all generated numbers of unigrams/bigrams).

For implicit reasoning, current approaches either locate multiple warrants from an existing structured corpus of arguments via similarity search [7], [35] or incorporate them to improve the performance of evidence detection [7]. While Singh et al. [7] commissioned two annotators to assess the quality of warrants located from the ARCC (ARC Corpus) dataset in relation to various datasets. The proposed method is based on a publicly available dataset ARCC, which stands for

Argument Reasoning Comprehension Corpus from News Comments [42], which was built for the 2018 SemEval task [43] by Habernal et al. [17].

4.3. Analysis and Performance Comparison

To evaluate the quality of our warrant generator and the score of their quality, we use automatic evaluation methods, same to Park et al.'s model [39] evaluation metric, as in table 4 and table 5 shows the results on the diversity. We conduct ablation experiments to demonstrate the effectiveness of reinforcement learning and its associated benefits in terms of generating more enhanced warrants.

Table 4. Automatic evaluation results on warrant generation quality in

Method	**BLUE-1**	**BLUE-2**	**Embedding Average**	**Embedding Greedy**	**Embedding extreme**
Lexical Chain with Multi-Head Attention (without RL-agent)	0.2019	0.0897	0.7107	0.3989	0.2374
Lexical Chain with Multi-Head Attention controlled by RL-agent (SARAS)	0.2974	0.1084	0.7885	0.5265	0.2944
A multi-column convolutional neural network for why-QA (without RL-agent)	0.2717	0.0807	0.6921	0.5282	0.2404
A multi-column convolutional neural network for why-QA Controlled by RL-agent (SARSA)	0.3205	0.1175	0.7744	0.5817	0.2978
RST (without RL-agent)	0.2153	0.0884	0.6408	0.5578	0.3432
RST controlled by RL-agent (DDQN)	0.3381	0.1192	0.7822	0.6168	0.3828
RST-Multi-head attention generator (without RL-agent)	0.3427	0.1069	0.7439	0.5997	0.3834
RST-Multi-head attention generator controlled by RL-agent (DDQN)	0.3749	0.1205	0.7943	0.6227	0.4436

Novel hybrid models for warrant generation are proposed in our work, which combines natural language processing, deep learning, and reinforcement learning techniques. Each model is constructed using a new framework that includes a locator and a generator. To generate warrants, the generator is initially trained using sequence-to-sequence learning. The selector, which is used to identify warrants relevant fragments, is then trained in a variety of environments using supervised or reinforcement learning techniques. The goal of reinforcement learning is to find the best reward function for the expert policy. Finally, the generator is fine-tuned further through reinforcement learning to produce more accurate warrants with a well-trained locator. High prediction success rates have been achieved thanks to the diversity of approaches used in the proposed models.

Table 5. Automatic evaluation results on the diversity of warrant generation of our proposed model.

Method	**Dist-1**	**Dist-2**	**Dist-1-within**	**Dist-2-within**
Lexical Chain with Multi-Head Attention (without RL-agent)	0.0816	0.0955	0.1993	0.2153
Lexical Chain with Multi-Head Attention controlled by RL-agent (SARAS)	0.1266	0.1225	0.2454	0.2881
A multi-column convolutional neural network for why-QA (without RL-agent)	0.1182	0.2265	0.3103	0.3244
A multi-column convolutional neural network for why-QA Controlled by RL-agent (SARSA)	0.1382	0.2963	0.3422	0.3818
RST (without RL-agent)	0.0927	0.2791	0.2695	0.3364
RST controlled by RL-agent (DDQN)	0.1423	0.3210	0.3612	0.4147
RST-Multi-head attention generator (without RL-agent)	0.1102	0.2983	0.3274	0.3908
RST-Multi-head attention generator controlled by RL-agent (DDQN)	0.1528	0.3291	0.3710	0.5007

By experimenting with different SARSA and DDQN for each model, we discovered that they make little difference. This means that they reward similarly to the generator, resulting in very similar results when changing the RL-agent, for example, from SARSA to DDQN and vice versa. We use reinforcement learning in our models to generate more interesting and coherent warrants focusing on the context of claim and evidence reason. The experiments in Tables 4 and 5 demonstrate that automated diversity and quality metrics produce scores that are significantly higher than the baseline (without Reinforcement Learning). The effectiveness of reinforcement learning, which involves the agent performing an action and being rewarded, is demonstrated by the promising outcomes obtained as a result of the reward used to guide the generator. The best performance is obtained when the RST-based algorithm is combined with multi-head attention for warrant generation enhanced by RL-agent.

According to Al-Khawaldeh et al. [35], the RST-based algorithm for filtering a warrant for a claim trained using DDQN has the highest f-score because it assists in detecting the relationship between clauses. This model can benefit from text organisation by dividing it into sub-clauses, either as a nucleus or a satellite, after the semantic structure is parsed using RST. Since RST is useful for determining the structure and relationship of arguments, this model's performance is enhanced. The more fundamental relationships are interpretation, justification, confirmation, illustration, result, explanation, evidence, foundation, and condition.

Causal relationships between two events establish common causes that support the initial event, assisting in causal inference. Given that a warrant justifies the claim based on the evidence, it improves the model's ability to capture the text fragment that supports the evidence. As a result, we investigated that using a multi-column convolutional neural network for the why-QA model proposed by Oh et al. [24], dealing with warrant generation as Why-question answering (why-QA) that retrieves the warrant as to the answer to a relevant document (evidence) and automatically recognises causalities is extremely practical. It ranks second among our proposed models for detecting casualties.

Along with the primary role of the lexical chain, we use the strongest chain as an auxiliary input to select significant sentences. Extracting the highest score (sequence of related words) as an auxiliary input to the model enables the model to pay more attention to the most informative words in the evidence while preserving the main content. In other words, the most robust chain reflects the evidence's central theme. They are extracting the chains of evidence articles to summarise and reduce the data. For Multi-Head Attention CNN-Bi-LSTM, individual attention heads capture more linguistically interpretable representations: syntactic and semantic relations that the encoder finally concatenates to attend to data from distinct representation subspaces. Local and global features are detected using the CNN-Bi-LSTM combination. Al-Khawaldeh et al. [35] used an RST-based algorithm to filter warrants for the claim and DDQN agent-controlled multiheaded attention to generate higher-quality warrants and eliminate irrelevant information.

The RST-based algorithm combined with multi-head attention provides the best performance for warrant generation. The primary objective of our work in utilising Rhetorical Structure Theory RST is to return the appropriate warrant from the retrieved evidence in light of the claim. The input that justifies detection is the claim's "bag of words" and relevant evidence. The RST-based method improves the warrant filtering's f-score measure by nearly 3% and 4%, respectively, compared to Multi-Head Hierarchical Attention CNN-Bi-LSTM combined with the most robust chain evidence and causality attention. In this work, we begin by filtration warrants using an RST-based method and then use Multi-Head Hierarchical Attention as a generator controlled by DDQN. In comparison to the other three models, the fourth model produces the highest-quality warrants based on diversity and quality metrics in addition to the f-score measure.

To determine the warrant associated with a particular claim and evidence, it is necessary to determine the context of that claim within the evidence. The RST connection is used to denote which sections of the text contain the warrant (that could be implicit or explicit). A critical property of an RST analysis in RST combined with the Multi-Head -Attention Mechanism model is that RST parses unstructured text into clauses with rhetorical relations, nucleus or satellite, as in the example below. The warrant is connected to the claim in this example via an explanation relation (As a result) in figure 2.

To filter warrant using RST, we must first identify text units (spans) within the evidence and then determine their relationships (rhetorical relations that hold between them). Certain rhetorical relations contain cues that connect these spans; for example, the relation result contains a "so," the relation evidence connects the claim with the candidate warrant as a cause-effect relationship, the nucleus is the claim, and information aimed at increasing belief in the claim is considered a warrant in our work. DDQN requires both encoder and decoder to have an informative representation of internal states in the form of hidden vectors. The DDQN learns how to determine which action (e.g., word) to choose from a list to modify the current decoded sequence in the long run. It approximates the Q-value function by updating its Q-values through actions and rewards, selecting the action with the highest Q-value in the outputs.

5. CONCLUSION

We propose various Deep Learning models for Toulmin Argument warrant generation in this paper. We demonstrated the performance of each of these models and the benefit of combining them with a reinforcement learning agent to improve generation and inference accuracy. Our investigations confirm that it is necessary to combine our model with auxiliary data such as the topic and sentiment. Incorporating a reinforcement learning agent enables the generator to receive rapid and robust training for decoding sequential text successfully. We generate warrants using RST and a multihued attention mechanism and obtain the best results on the ARC dataset [17]. We will devote additional attention to the remaining Toulmin Arguments for future works: supporting evidence, modifiers, and rebuttals.

REFERENCES

[1] F. T. AlKhawaldeh, Tommy Yuan, D. Kazakov, F. T. Al-Khawaldeh, T. Yuan, and D. Kazakov, "A Novel Model for Enhancing Fact-Checking," in *Proceedings of the 2021 Computing Conference*, 2021, vol. 284, pp. 661–677, doi: 10.1007/978-3-030-80126-7_47.

[2] J. Lawrence and C. Reed, "Argument mining: A survey," *Comput. Linguist.*, vol. 45, no. 4, pp. 765–818, 2019, doi: 10.1162/COLIa00364.

[3] I. Habernal and I. Gurevych, "Argumentation Mining in User-Generated Web Discourse," *Comput. Linguist.*, vol. 43, no. 1, pp. 125–179, 2017, doi: 10.1162/COLI a 00276.

[4] S. E. Toulmin, *The uses of argument (Updated edition, first published in 1958)*. Cambridge University Press2003 ,.

[5] K. S. Hasan and V. Ng, "Why are you taking this stance? Identifying and classifying reasons in ideological debates," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014, pp. 751–762, doi: 10.3115/v1/d14-1083.

[6] P. Reisert, N. Inoue, N. Okazaki, and K. Inui, "A Computational Approach for Generating Toulmin Model Argumentation," in *Proceedings of the 2nd Workshop on Argumentation Mining*, 2015, vol. June, pp. 45–55, doi: 10.3115/v1/w15-0507.

[7] K. Singh, E. Simpson, P. Reisert, I. Gurevych, and K. Inui, "Ranking Warrants with Pairwise Preference Learning," in *Proceedings of the 26th Annual Meeting of the Natural Language Processing Society (March 2020)*, 2020, no. C, pp. 776–779, [Online]. Available: https://www.anlp.jp/proceedings/annual_meeting/2020/pdf_dir/P3-34.pdf.

[8] V. D. O. Gabriel, D. F. Adamatti, A. R. Panisson, R. H. Bordini, and C. Z. Billa, "Argumentation-based reasoning in BDI agents using Toulmin's model," in *Proceedings of the Brazilian Conference on Intelligent Systems, BRACIS 2018*, 2018, pp. 378–383, doi: 10.1109/BRACIS.2018.00072.

[9] V. de O. Gabriel, A. R. Panisson, R. H. Bordini, D. F. Adamatti, and C. Z. Billa, "Reasoning in BDI agents using Toulmin's argumentation model," *Theor. Comput. Sci.*, vol. 805, no. January, pp. 76–91, 2020, doi: 10.1016/j.tcs.2019.10.026.

[10] D. Walton, C. Reed, and F. Macagno, *Argumentation schemes*, no. August. Cambridge University Press, 2008.

[11] C. Hashimoto, K. Torisawat, S. De Saeger, J. H. Oh, and J. Kazama, "Excitatory or Inhibitory: A New Semantic Orientation Extracts Contradiction and Causality from the Web," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 619–630, [Online]. Available: https://www.aclweb.org/anthology/D12-1057.

[12] V. Simaki, C. Paradis, and A. Kerren, "A Two-step Procedure to Identify Lexical Elements of Stance Constructions in Discourse from Political Blogs," *Corpora*, vol. 14, no. 3, pp. 379–405, 2019, doi: 10.3366/cor.2019.0179.

[13] I. Habernal and I. Gurevych, "Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse," in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, no. September, pp. 2127–2137, doi: 10.18653/v1/d15-1255.

[14] F. Boltuzic and J. Šnajder, "Fill the Gap! Analyzing Implicit Premises between Claims from Online Debates," in *Proceedings of the 3rd Workshop on Argument Mining*, 2016, no. August, pp. 124–133, doi: 10.18653/v1/w16-2815.

[15] P. Rajendran, D. Bollegala, and S. Parsons, "Contextual Stance Classification of Opinions: A Step towards Enthymeme Reconstruction in Online Reviews," in *Proceedings of the Third Workshop on Argument Mining (ArgMining2016),* 2016, no. August, pp. 31–39, doi: 10.18653/v1/w16-2804.

[16] B. D. Horne and S. Adali, "This Just In-Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News," in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, 2017, vol. 11, no. 1, pp. 40–49, doi: 10.18653/v1/w18-5507.

[17] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein, "The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicitwarrants," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2018, vol. 1, pp. 1930–1940, doi: 10.18653/v1/n18-1175.

[18] F. T. Al-Khawaldeh and V. W. Samawi, "Lexical Cohesion and Entailment based Segmentation for

Arabic Text Summarization (LCEAS),” *World Comput. Sci. Inf. Technol. J.*, vol. 5, no. 3, pp. 51–60, 2015, [Online]. Available: http://oaji.net/articles/2015/567-1425407917.pdf.

[19] S. Li, Z. Zhao, T. Liu, R. Hu, and X. Du, “Initializing Convolutional Filters with Semantic Features for Text Classification,” in *Proceedings ofthe 2017 Conference on Empirical Methods in Natural Language Processing EMNLP*, 2017, pp. 1884–1889, doi: 10.18653/v1/d17-1201.

[20] M. Peters *et al.*, “Deep Contextualized Word Representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long Papers)*, 2018, vol. 1, pp. 2227–2237, doi: 10.18653/v1/N18-1202.

[21] A. Vaswani *et al.*, “Attention Is All You Need,” in *Proceeding of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 5998–6008.

[22] M. Heilman and K. Sagae, “Fast Rhetorical Structure Theory Discourse Parsing,” *CoRR, abs/1505.02425*, pp. 1–6, 2015, [Online]. Available: http://arxiv.org/abs/1505.02425.

[23] W. C. Mann and S. A. Thompson, “Rhetorical Structure Theory: Toward a functional theory of text organization,” *Text & Talk*, vol. 8, no. 3. pp. 243–281, 1988, doi: 10.1515/text.1.1988.8.3.243.

[24] J.-H. Oh, K. Torisawa, C. Kruengkrai, R. Iida, and J. Kloetzer, “Multi-Column Convolutional Neural Networks with Causality-Attention for Why-Question Answering,” in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 415–424, doi: 10.1145/3018661.3018737.

[25] A. Imani, A. Vakili, A. Montazer, and A. Shakery, “Deep Neural Networks for Query Expansion Using Word Embeddings,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, ECIR 2019., vol. 11438 LNCS, H. D. Azzopardi L., Stein B., Fuhr N., Mayr P., Hauff C., Ed. Springer, Cham, 2019, pp. 203–210.

[26] M. Esposito, E. Damiano, A. Minutolo, G. De Pietro, and H. Fujita, “Hybrid Query ExpansionUsing Lexical Resources and Word Embeddings for Sentence Retrieval in Question Answering,” *Inf. Sci. (Ny).*, vol. 514, pp. 88–105, 2020, doi: 10.1016/j.ins.2019.12.002.

[27] N. Yusuf, M. A. M. Yunus, N. Wahid, N. Wahid, N. M. Nawi, and N. A. Samsudin, “Enhancing Query Expansion Method Using Word Embedding,” in *Proceeding of the 9th IEEE International Conference on System Engineering and Technology*, 2019, vol. 6, pp. 21–24, doi: 10.1109/ICSEngT.2019.8906317.

[28] H. K. Azad and A. Deepak, “A New Approach for Query Expansion Using Wikipedia and WordNet,” *Inf. Sci. (Ny).*, vol. 492, pp. 147–163, 2019, doi: 10.1016/j.ins.2019.04.019.

[29] C. dos Santos, M. Tan, B. Xiang, and B. Zhou, “Attentive Pooling Networks,” *arXiv Prepr. arXiv1602.03609*, pp. 1–10, 2016, [Online]. Available: http://arxiv.org/abs/1602.03609.

[30] K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari, “RL-CycleGan: Reinforcement Learning Aware simulation-to-real,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11154–11163, doi: 10.1109/CVPR42600.2020.01117.

[31] B. Zhao, H. Li, X. Lu, and X. Li, “Reconstructive Sequence-Graph Network for Video Summarization,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, vol. 8828, no. c, pp. 1–10, doi: 10.1109/TPAMI.2021.3072117.

[32] A. Ayoub, Z. Jia, C. Szepesv, M. Wang, and L. F. Yang, “Model-Based Reinforcement Learning with Value-Targeted Regression,” in *Proceedings of the 37th International Conference on Machine Learning*, 2020, vol. PMLR 119, pp. 463–474, [Online]. Available: http://proceedings.mlr.press/v119/ayoub20a.html.

[33] P. Bachman and D. Precup, “Data Generation as Sequential Decision Making,” *Adv. Neural Inf. Process. Syst. 28*, vol. 2015-Janua, pp. 3249–3257, 2015.

[34] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[35] F. T. Al-Khawaldeh, T. Yuan, and D. Kazakov, “RL-GAN BASED TOULMIN ARGUMENT,” *JASC J. Appl. Sci. Comput.*, vol. VII, no. III, pp. 106–120, 2020.

[36] J. Alammar, “The Illustrated Transformer,” 2018. http://jalammar.github.io/illustrated-transformer/ (accessed Nov. 16, 2020).

[37] Jeffrey Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, no. October, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[38] D. P. Kingma and J. L. Ba, “ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION,” in *3rd*

International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015, pp. 1–15.

[39] C. Park, W. Yang, and J. Park, "Generating Sentential Arguments from Diverse Perspectives on Controversial Topic," in *Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 2019, vol. November 4, pp. 56–65, doi: 10.18653/v1/d19-5007.

[40] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, no. JULY, pp. 311–318, doi: 10.1002/andp.19223712302.

[41] C. W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation," in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016, pp. 2122–2132, doi: 10.18653/v1/d16-1230.

[42] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward Controlled Generation of Text," in *34th International Conference on Machine Learning, ICML 2017*, 2017, vol. 4, no. PMLR 70, pp. 2503–2513.

[43] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein, "SemEval-2018 Task 12: The Argument Reasoning Comprehension Task," in *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, 2018, vol. June, pp. 763–772, doi: 10.18653/v1/s18-1121.

AN EVALUATION OF STATE-OF-THE-ART APPROACHES TO RELATION EXTRACTION FOR USAGE ON DOMAIN-SPECIFIC CORPORA

Christoph Brandl[1], Jens Albrecht [1] and Renato Budinich[2]

[1]Nuremberg Institute of Technology Georg Simon Ohm,
Department of Computer Science, Keßlerplatz 12, 90489 Nuremberg, Germany
[2]Fraunhofer Supply Chain Services, Research Group Future Engineering,
Nordostpark 93, 90411 Nuremberg, Germany

ABSTRACT

The task of relation extraction aims at classifying the semantic relations between entities in a text. When coupled with named-entity recognition these can be used as the building blocks for an information extraction procedure that results in the construction of a Knowledge Graph. While many NLP libraries support named-entity recognition, there is no off-the-shelf solution for relation extraction.

In this paper, we evaluate and compare several state-of-the-art approaches on a subset of the FewRel data set as well as a manually annotated corpus. The custom corpus contains six relations from the area of market research and is available for public use. Our approach provides guidance for the selection of models and training data for relation extraction in real-world projects.

KEYWORDS

Relation Extraction, Knowledge Graph, Market Research.

1. INTRODUCTION

1.1. Motivation

Many businesses today are building knowledge graphs to model complex networks of entities and their relationships. Hereby, implementations using graph databases are more flexible than SQL databases and offer unique possibilities like path-based queries and employing network analysis tools for data exploration.

Specifically, we are interested in the automatic creation of a Knowledge Graph from text sources, such as news or Wikipedia articles. The required information extraction process usually involves at least two steps: named-entity recognition (NER) and relation extraction (RE). Relevant entities and the relation types are usually defined by the application domain. Several NLP libraries today support NER with state-of-the-art transformer models (https://spacy.io/usage/facts-figures#benchmarks). RE methods, in contrast, still lack a uniform interface, requiring the user to prepare multiple variants of the training pipeline depending on the chosen model architectures. In addition, the different RE approaches are designed for specific data formats, making a direct evaluation and comparison inconvenient in a real-world scenario.

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 77-89, 2021. CS & IT - CSCP 2021 DOI:10.5121/csit.2021.112006

In this paper, we investigate the suitability of certain state-of-the-art models for relation extraction in the domain of market analysis. Here, the entities represent objects, such as companies, products or technologies. Typical relation types are *manufactures*, *operates* and *operates sth in* (see Figure 1). Our research is part of a project on the detection of market trends in temporal knowledge graphs created from news articles. The work was part of the Future Engineering project at TH Nuremberg and Fraunhofer SCS [1, 2]. The broad focus of this project is the detection of market trends by various means including the analysis of temporal changes in knowledge graphs generated from domain specific news articles.

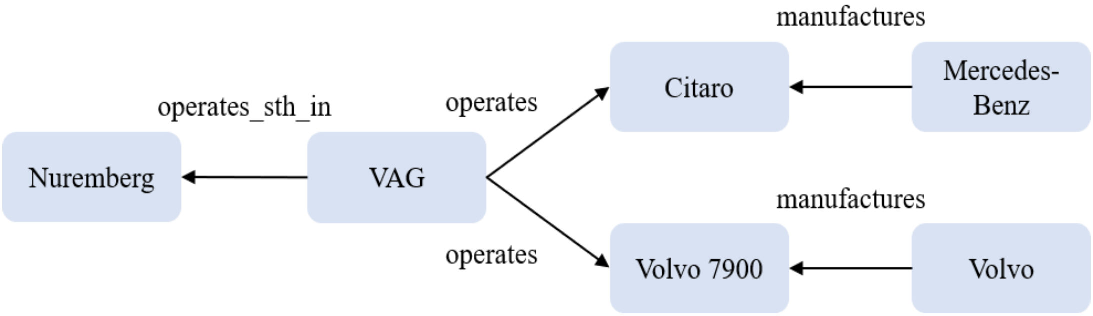


Figure 1. A simple knowledge graph

1.2. Research Questions and Contribution

In the last years, several new training data sets and model architectures have been published for RE (see Section 2). The motivating questions for this work are: Which model should be used for a specific application? Can the performance on a general, non-domain specific data set be used as a reasonable indicator to select the model that will perform best on the domain-specific data of the application?

Among the various available data sets for RE, we chose the FewRel data set published in 2018 [3] since it covers the broadest number of use cases (see Section 3.3). For the evaluation, we selected a subset of six FewRel relations relevant to our domain. In addition, we created a custom training data set with six different and more specific relation types. Both training data sets also contain samples that should be categorized into neither of these relation types ("none of the above").

Thus, the contribution of this work is as follows:

- We compare the performance of several state-of-the-art model architectures to relation extraction on a subset of the FewRel data set and a manually labelled set of custom training data. Both data sets contain six relations relevant to trend analysis.
- We analyze and discuss the difference in performance when using the FewRel data versus the domain-specific training data.
- We propose an interface to streamline the usage of the relation extraction approaches with the Inferencer class.
- We provide a new training data set for relation extraction on company news data for public use.

2. RELATION EXTRACTION

2.1. State-of-the-Art Models for Relation Extraction

The basis for many of the approaches presented in Natural Language Processing in recent years is the BERT model [4], which is based on the Transformer architecture [5].

It provides state-of-the-art results in a variety of different NLP tasks thanks to an effective internal representation of language. Furthermore, it offers the possibility to fine-tune the pre-trained language model for specific tasks, including RE.

Thus, a lot of proposed models within RE utilize adaptations of the BERT model. In order to find the most suitable approach for the use case of trend analysis and the generation of a knowledge graph from text data, we examined five state-of-the-art RE approaches, four of which are based on BERT models and one utilizing a LSTM network structure. However, a prerequisite to all the examined approaches is the identification of named entities in NER, which is usually provided in the training data set.

The selection of the examined approaches is based on two different factors. First, the performances of the approaches in common RE task leader boards were considered (http://nlpprogress.com/english/relationship_extraction/). Further, we paid attention to the availability of implementations of the proposed approaches so that they could be quickly adapted and trained for our use case. The only approach examined that is not based on BERT is the bidirectional Entity-Aware Attention LSTM [6]. Lee et al. are using a bidirectional long short-term memory network that uses both, an attention mechanism and latent entity typing for the classification of relations. This approach makes it possible to use different word embeddings, such as Glove [7] or ELMo [8] whilst using a less complex network structure compared to the BERT model.

The Enhanced Language Representation with Informative Entities (ERNIE) approach [9] tries to leverage additional information about the entities through linked open data resources for the classification process. ERNIE utilizes previously trained TransE embeddings [10] as representation of the contained entities in combination with a relation extraction specific encoder component as well as a new goal for the pre-training phase of the BERT model.

In contrast, R-BERT [11] concentrates on the extraction of entity information contained in the input sentences. Therefore, it only uses the output vectors of the entities together with the `[CLS]` output vector of the standard BERT model for the classification of relations, providing low complexity in the classification process.

Matching the Blanks (MTB) [12] is a basic method for learning relation representations from non-annotated text data during the pre-training phase of the BERT architecture. This leads to high flexibility in the application of this method, since it is still a standard BERT model that can be used arbitrarily. For the relation specific optimization of the BERT model, Soares et al. define a new pre-training goal while replacing some of the entities in the pre-training data with `[BLANK]` tokens in order to force the model to learn semantic relations between general entities.

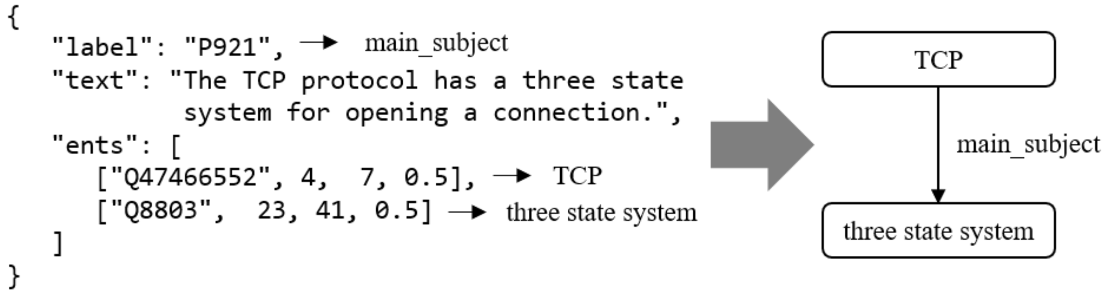
Lastly, BERT Pair [13] is the only approach in this evaluation defining the relation extraction task as an n-way-k-shot scenario. The approach uses a support set for the classification of an input sequence, which contains k examples for each of the n relation types. The authors focus on addressing the "none of the above" issue (see section 2.2) within the field of relation extraction.

Whilst classifying sentences, BERT Pair builds pairs of the input sentence with each of the instances in the support set to identify the most similar support set example.

2.2. Data Sets for Relation Extraction

A wide variety of data sets is available for training and benchmarking of RE approaches providing different tasks and application scenarios. Examples include the TACRED data set [14], the New York Times corpus [15] or the SemEval 2010 Task 8 data set [16]. Those data sets often contain very general relation types such as "Cause-Effect" or "Entity-Origin". These general relations offer a high coverage of sentences, but they do not capture the specific relations in a business domain like market trend analysis. Therefore, these data sets cannot be used in such application scenarios.

A data set with more suitable relation types for trend analysis is the FewRel data set [3]. Proposed in 2018, it provides 100 relation types with a wide thematic spread from different domains, including categories like "owned by", "operating system" and "member of political party". For each of the relations, the data set contains around 700 examples. Every example consists of a sentence, two entities and a relation label (see Figure 2). The entities as well as the relation labels are linked to Wikidata identifiers making it easy to connect them to other linked open data resources.



```
{
   "label": "P921", → main_subject
   "text": "The TCP protocol has a three state
            system for opening a connection.",
   "ents": [
      ["Q47466552", 4,  7, 0.5], → TCP
      ["Q8803",  23, 41, 0.5] → three state system
   ]
}
```

Figure 2. Example sentences from FewRel

As an extension to the FewRel task, Gao et al. propose FewRel 2.0 [17], which does not add new data but addresses the problem of "none of the above" recognition. It describes the case that a sentence does not belong to any of the predefined relation types. Therefore, they propose to classify such sentences into an additional category "NOTA". In previous scenarios, it was assumed that each of the instances to be classified can be assigned to one of the predefined relations. In practical use cases, however, this assumption usually does not hold: instances that do not contain one of the predefined relations or do not contain any relation at all form a significant portion of the sentences. Thus, Gao et al. propose to use only a subset of the relations contained in the FewRel data set and build an artificial "NOTA" class out of the remaining classes.

Due to its specific relation classes as well as the "NOTA" identification task the FewRel data set provides a good starting point for a comparison of RE approaches in a custom application scenario. In addition, the data set allows the creation of a sufficiently sized training data set to ensure to ensure meaningful results.

3. MODEL COMPARISON WITH FEWREL-DATA

3.1. Data Selection

To create a useful subset for our project, business stakeholders were asked to identify the most relevant out of the 100 FewRelrelation types for our scenario of market trend analysis. As a result, a subset consisting of the six relation types listed in Table 1 was selected.

Table 1. Relevant Relation Classes from FewRel Data Set

Relation Class	**Description**
taxon rank	level in a taxonomic hierarchy
movement	literary, artistic, scientific or philosophical movement associated with this person or work
follows	immediately prior item in a series of which the subject is a part
instance of	that class of which this subject is a particular example and member
notable work	notable scientific, artistic or literary work, or other work of significance among subject's works
main subject	primary topic of a work

Thus, our training data set consists of all training samples from these six categories. In addition, we included a random selection of sentences from the remaining FewRel classes and re-assigned them to the category "none of the above" (NOTA). This creates a class with a wide spread of example sentences from different areas of the relation spectrum.

For the generation of the NOTA class, the remaining relation classes are partitioned into training, test and validation data sets, ensuring that the validation and test data sets do not contain any sentences from classes contained in the actual training data set. Subsequently, this newly generated class can be treated as an additional class in the classification scenario.

A train-test-validate split was performed, resulting in 200 samples for each category in the training and test data set and 100 sentences per class in the validation data, following a similar approach to Zhang et al. [9]. Thereby, the equal distribution of examples per class in the FewRel data set was also adopted for the selection of our subset.

Due to the use of the few-shot scenario in BERT Pair, this approach requires a reorganized training data set, which is, however, identical to the training data with respect to the contained sentences.

The comparison of the different relation extraction approaches with this reduced FewRel data set can provide first insights on the performance in our application domain.

3.2. Unified Evaluation Process

Despite the identical initial model and the same objective, the ways in which the approaches are applied differ significantly from one another. For example, the input and output sequences differ from approach to approach due to differences in the adaption to the BERT model. For uniform use and comparison of the approaches, we thus propose the Inferencer interface, that encapsulates each RE approach and provides a uniform interface for the usage of the models. The implementation is open source and provided on Github (https://github.com/th-nuernberg/fe-relation-extraction-natl21).

The functionality is shown in Figure 3. Each of the models can be trained with its individual training routine, but all Inferencer classes implement the same method for relation inference. Hence, it is possible to apply the same evaluation routine to all the approaches, avoiding discrepancies in the evaluation procedures of different machine learning frameworks, which could distort the results of the evaluation.

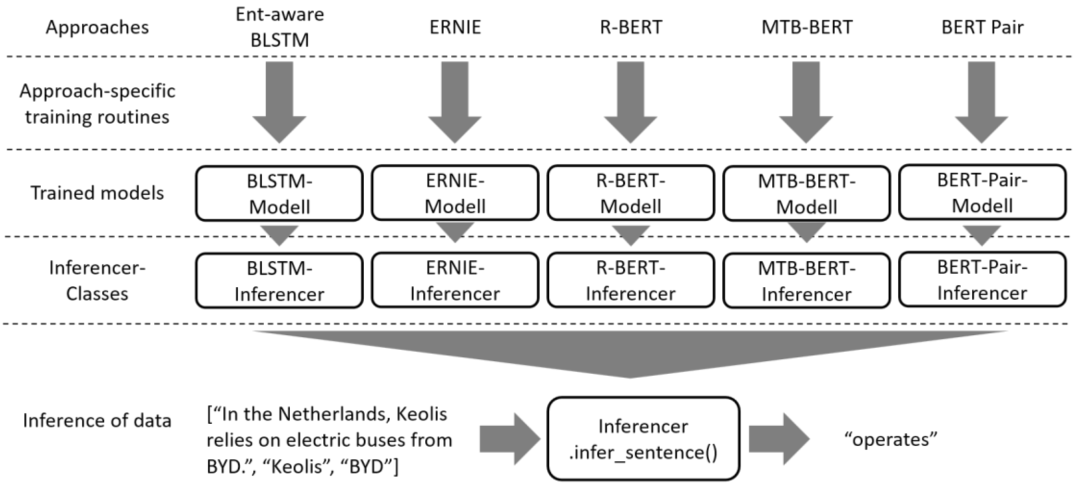


Figure 3. Process of Evaluation

3.3. Training and Evaluation

The hyperparameters used to train the different models were adopted from the original publications [5, 6, 9, 11, 12, 17]. No further hyperparameter tuning was performed. All relation extraction approaches were trained with the same training data. Accuracy, precision, recall and F1-score were used to evaluate and compare the different RE approaches. Table 2 shows the results of the evaluation process with FewRel data. As the training data were equally distributed over the classes, micro and macro average of these metrics are identical.

Table 2. Results of Evaluation with FewRel Data

	R-BERT			**MTB**			**Pair**			**BLSTM**			**ERNIE**		
	p	r	f1	p	r	f1	p	r	f1	p	r	f1	p	r	f1
taxon rank	0.99	1.00	1.00	0.98	1.00	0.99	0.99	1.00	1.00	0.96	1.00	0.98	0.99	1.00	1.00
movement	0.89	1.00	0.94	0.80	0.94	0.86	0.98	1.00	0.99	0.79	0.91	0.85	0.95	0.94	0.94
follows	0.93	0.90	0.91	0.86	0.91	0.88	0.99	0.69	0.81	0.69	0.79	0.74	0.83	0.90	0.86
instance of	0.80	0.89	0.84	0.77	0.77	0.77	0.92	0.57	0.70	0.77	0.68	0.72	0.87	0.94	0.90
notable work	0.97	0.94	0.95	0.89	0.93	0.91	0.94	0.66	0.78	0.81	0.83	0.82	1.00	0.03	0.06
main subject	0.97	0.90	0.93	0.95	0.80	0.87	0.85	0.52	0.65	0.84	0.70	0.77	0.65	0.87	0.74
NOTA	0.79	0.70	0.74	0.66	0.56	0.61	0.41	0.97	0.58	0.60	0.56	0.58	0.51	0.75	0.61
Average	**0.91**	**0.90**	**0.90**	0.84	0.84	0.84	0.87	0.77	0.79	0.78	0.78	0.78	0.83	0.78	0.73
Accuracy	**0.90**			0.84			0.77			0.78			0.78		

All approaches show strong results. The best approach is R-BERT with an accuracy of 0.90 and an F1 score of 0.90. In terms of F1 score, the ERNIE model is the weakest with 0.73. It can also be seen that the ERNIE and BERT Pair model each have higher precision than accuracy values.

The precision is of great importance for the use case of generating a knowledge graph from text data, as only correct relations should be included. But R-BERT outperforms these models even in terms of precision in most but not all of the classes.

In addition, general tendencies and behaviour of all RE approaches can be identified. First, it is clearly visible that all models were able to classify completely or almost completely the classes "taxon rank" and "movement" correctly. These two categories are very different from each other as well as from all other relations present in the data set, which explains the observed behaviour. Furthermore, by comparing the detailed results of all approaches, it can be seen that the category "instance of" is often among the most misclassified ones. Examples of this class are frequently classified as "NOTA" instances. This accumulation can be explained by the high diversity in the category "instance of", which leads to confusion within the classification.

The results gathered give insights about the behaviour of the approaches in a real-world scenario with fewer, domain specific relations than the original FewRel task. R-BERT turned out to be the most suitable approach for the subset of the FewRel data, since it provides the best results in all metrics. However, BERT Pair also proves to be suitable for the use case of generating a knowledge graph because of its strong precision value. The results of Matching the Blanks, the BLSTM and ERNIE are significantly worse and therefore not suitable in such a scenario. Note that these results are not comparable to ones listed on the FewRel leader board (https://thunlp.github.io/fewrel.html) as we only used a subset of the relations.

4. COMPARISON WITH MANUALLY LABELLED DATA

Using an existing data set, such as FewRel, restricts an application to predefined relation types. With an ad-hoc data set, however, it is possible to define custom relations which more precisely match the requirements of the application domain. For our scenario, the analysis of trends in the market of electric buses, we manually created a custom data set with the relation classes shown in Table 3. This data set is available on Github (https://github.com/th-nuernberg/fe-relation-extraction-natl21).

Table 3. Defined Relations for FE Data Set

Relation Class	**Description**
orders	Order process of products
orders something from	Order process with a specific company
operates	Operation or use of a product
operates something in	Location of operation of a product
manufactures	Manufacturing of products
uses/employs	Application of a technology

The data set is based on articles extracted from electrive.com, a news provider targeting decision-makers, manufacturers and service providers in the e-mobility sector (https://www.electrive.com/faq-electrive).

The search on electrive.com was restricted to articles in the "Fleets" section that primarily addresses the purchase and use of electric buses. The data set contains 2,269 articles from the period November 2013 to July 2020 which were extracted by the news crawler `news-please`[18]. This package enables the automated extraction of information such as the publication date, the title, the text, or the language of the article. To annotate these texts for relation extraction, the articles were split into single sentences.

The definition of the relations in Table 3 is based on application requirements and the analysis of information available in the articles. As these are news reports from the field of electric buses, much of the information contained relates to the ordering, use and manufacturing of e-buses. The relations "orders", "orders something from", "operates", "operates something in", "manufactures" and "uses/employs" represent these kinds of information in the classification scenario. All other contained relations are not relevant and therefore annotated as "NOTA" instances. This should provide the opportunity to learn the distinction between relevant and irrelevant relations during training.

For the annotation of the data, we used the tool INCEpTION[19]. It allows the definition of individual layers that capture different information in the annotation process. All contained named entities as well as the relation between all entity pairs were labelled this way. To keep the adaptations in the training routines of the RE approaches as low as possible, the annotated data was converted to match the FewRel data format.

In contrast to the FewRel data set where each sentence appears only once with exactly one combination of two entities, the annotation procedure described makes it possible for the same sentence to appear multiple times with different entity pairs in the data set (see Figure 4). This allows the generation of multiple training examples from a single sentence. Furthermore, this behaviour more accurately represents the use case of extracting information of all entities contained in the sentence and their relations, which is an advantage for the training and later use of the relation extraction approaches.

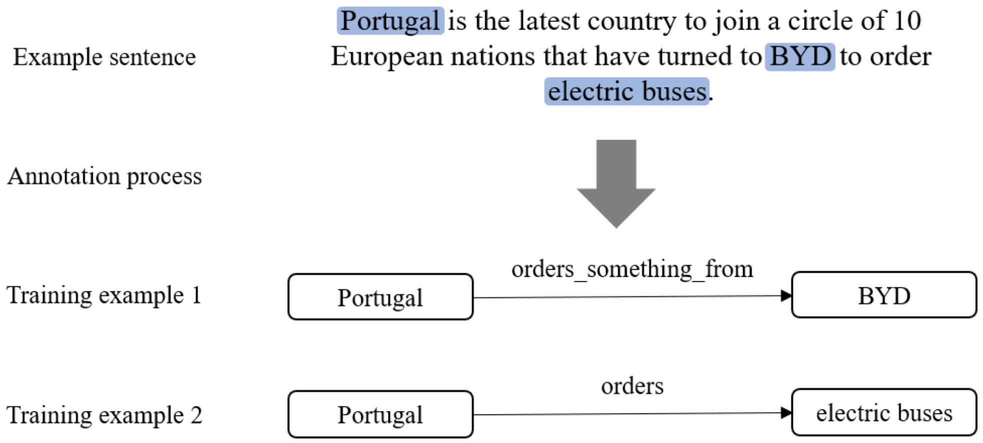


Figure 4. Generation of Multiple Training Examples from Sentence

In total, the data set consists of 1780 examples from 707 different sentences; see Table 4. The data set is divided into training, test and validation data. The training data comprise 1068 examples (60\%), the test and validation data set contain 356 sentences each (20\%).

Thus, the training data set reaches approximately the size of the FewRel training data set, which includes 1400 sentences. The distribution of the relation classes was preserved during the split.

Table 4. Number and Distribution of Examples in our Data Set

Relation Class	**Validation/Test set**	**Training set**	**Overall**
manufactures	79	238	396
operates	47	142	236
operates sth in	40	120	200
orders	69	207	345
orders sth from	31	95	157
uses/employs	32	96	160
Total	**356**	**1068**	**1780**

All approaches are trained with identical data and then evaluated with a likewise identical data set using the same metrics as in Section 3.3. See Table 5 for the results.

Table 5. Results of Evaluation with Future Engineering Data

	R-BERT			**MTB**			**Pair**			**BLSTM**			**ERNIE***		
	p	r	f1	p	r	f1	p	r	f1	p	r	f1	p	r	f1
NOTA	0.78	0.69	0.73	0.62	0.43	0.51	0.34	0.64	0.44	0.56	0.60	0.58	0.36	0.07	0.12
manufactures	0.92	0.89	0.90	0.85	0.59	0.70	0.89	0.73	0.81	0.89	085	0.87	0.23	0.73	0.35
operates	0.80	0.91	0.85	0.75	0.77	0.76	0.67	0.70	0.69	0.74	0.79	0.76	0.00	0.00	0.00
operates sth in	0.74	0.88	0.80	0.63	0.85	0.72	0.66	0.72	0.69	0.62	0.62	0.62	0.00	0.00	0.00
orders	0.93	0.90	0.91	0.70	0.90	0.78	0.94	0.49	0.65	0.88	0.87	0.88	0.55	0.23	0.33
uses/employs	0.71	0.69	0.70	0.71	0.75	0.73	0.68	0.66	0.67	0.79	0.69	0.73	0.40	0.78	0.53
orders sth from	0.90	0.87	0.89	0.69	0.81	0.75	0.86	0.61	0.72	0.85	0.90	0.88	0.00	0.00	0.00
Average	0.83	0.83	0.83	0.71	0.73	0.71	0.72	0.65	0.67	0.76	0.76	0.76	0.22	0.26	0.19
Accuracy	**0.82**			0.71			0.65			0.77			0.29		

* Because of missing information in the training process the results for ERNIE are not valid; see text

Surprisingly, all metrics of all examined approaches have dropped compared to the evaluation with FewRel data. One possible reason might be the increased difficulty resulting from the occurrence of multiple relations in a single sentence. Another point which may explain the decrease of the metrics is the similarity of the relations among each other, as they all target information from a similar context. Whilst R-BERT can almost perfectly classify the relations of the FewRel data set (Figure 5), it has difficulties with more similar relation classes, such as "operates" and "uses/employs" in our data set, which aim for overlapping expressive wordings within the sentences (Figure 6). Figure 5 and Figure 6 also illustrate the problems of the R-BERT approach in identifying instances of the artificial "NOTA" category. This can be explained by the high heterogeneity in the respective relation categories of the two datasets. Looking at the confusion matrix in Figure 6, it can be seen that "uses/employs" instances are often assigned to the category "NOTA", while examples of the classes "NOTA" and "operates sth in" often interchange. A closer look at individual records reveals that many of these misclassified records cannot be unambiguously assigned to one relation, thus explaining many of the uncertainties of R-BERT. The same findings can be observed in all examined approaches.

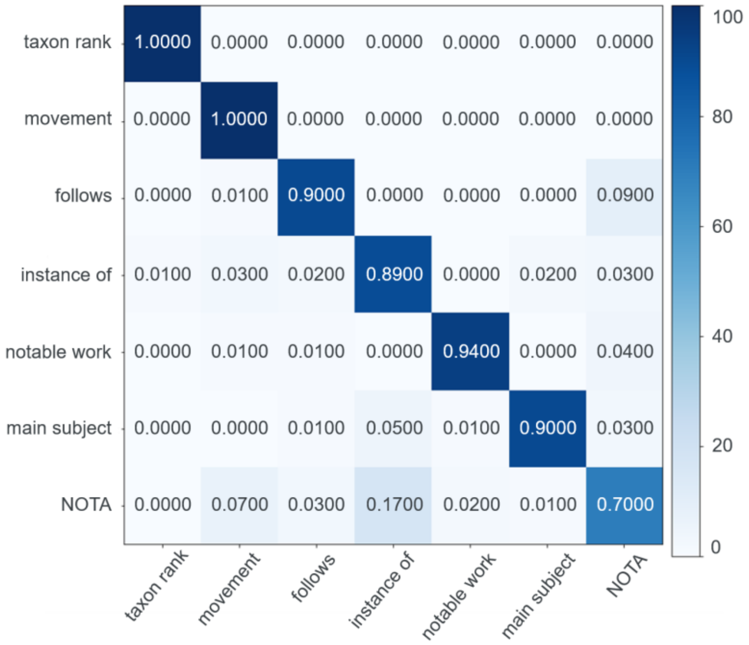


Figure 5. Confusion matrix for R-BERT on FewRel

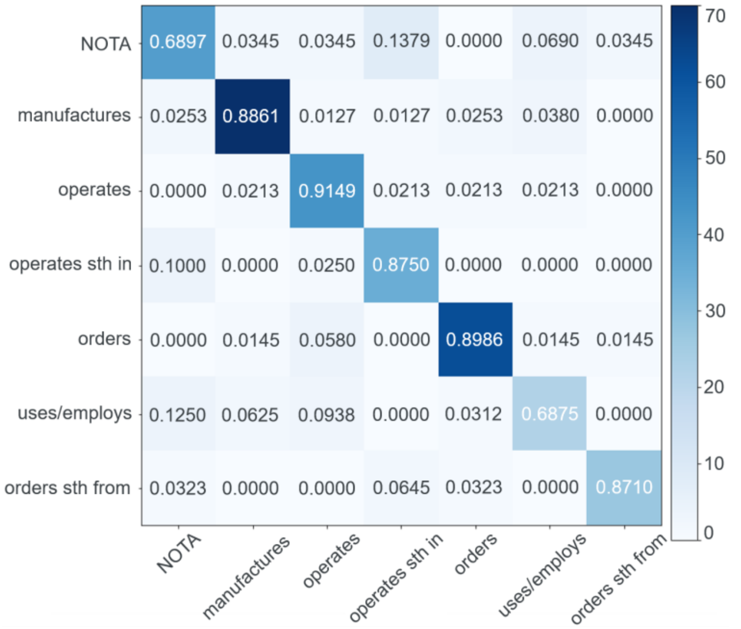


Figure 6. Confusion matrix for R-BERT on FE

The best approach is again R-BERT with an accuracy of 0.82 and an F1 score of 0.83. As before, ERNIE is the weakest model with an F1 score of 0.19. Using our data set, the BERT pair approach again shows a significantly higher precision compared to its F1 score.

As in the evaluation with FewRel data, the identification of "NOTA" instances is still difficult for all models despite the non-artificially generated category. This can be attributed to the fact that even in the new data set there is a high heterogeneity in the class "NOTA". No specific words or phrases exist to identify a relationship as "NOTA" which makes it hard for any model to learn such a class.

Again, R-BERT can be clearly identified as the best performing RE approach. Furthermore, BERT Pair also shows suitable behaviour with our data due to its high precision values. The entity-aware BLSTM model also shows good results with the data. Matching the Blanks, on the other hand, reveals once more weaknesses in identifying "NOTA" instances and seems less suitable for the specific use case.

Regarding the ERNIE model, there is a simple explanation for its very low metrics. The ERNIE model is expecting the entities to be linked to Wikidata identifiers to use previously learned entity knowledge embeddings for the classification. This link is provided within the FewRel data but not with the newly created data set. Consequently, no meaningful optimization of the model during the training process can take place. Therefore, a valid evaluation of the ERNIE model was not possible as it requires additional data, which cannot be provided with our data set.

5. CONCLUSION

Even though relation extraction is an essential task in building a knowledge base from text, there are no standard solutions or easy-to-use recipes available for industrial use cases. System engineers have to experiment with different modelling approaches and create custom training data to create sufficiently performing models. Our work can serve as a guideline and starting point for such an evaluation. The provided open-source implementation of the test, including a common API to all the evaluated models, minimizes the effort to get started.

In our evaluation R-BERT turned out to be the best performing model, showing robust results with the FewRel as well as with our own data set. Therefore, it can be concluded, that the use of the entity vectors in combination with the classification sequence of the BERT model as utilized by R-BERT represents the most promising approach in the experiments performed. In order to find the most suitable RE approach for a real-world scenario with a small set of specific relations and a fixed domain it can be a valid first step to use a subset of an available RE data set (e. g. FewRel) and select relations fitting to the scenario. Nevertheless, it is generally unavoidable to define specific relations and create a custom data set to extract the truly relevant relations for a business use case. In this case we would advise, in order to obtain a better confusion matrix, to carefully design the relations in order to avoid whenever possible any semantic overlap between them. In addition, it should be mentioned that the presented results provide only limited insight into the extraction of a larger number of relations from texts with the considered approaches and are thus not comparable to the tasks of common leaderboards, such as FewRel.

Since the task of relation extraction is a very active field of research, new approaches are constantly being proposed. Interesting recent alternatives are for example RECON [20], utilizing a KG whilst classifing relations and WDec [21], which in contrast to the examined approaches tries to jointly extract entities and relations from texts.

Future research in our group will also include investigation on a completely different approach to RE based on extractive question answering models (e.g. [22], [23]) trained on the SQuAD data set [24] In fact, one can easily reformulate any relation as a parametrized question, whose exact formulation depends on the named entities in the considered sentence (e.g. "Who ordered something from BYD?" for the sentence in
Figure 4). In a scenario where the required relations to extract often vary, this approach seems very appealing because it does not require any fine-tuning.

REFERENCES

[1] J. Albrecht, A. Belger, R. Blum, and R. Zimmermann, “Business Analytics on Knowledge Graphs for Market Trend Analysis,” in *Proceedings of LWDA 2019 (CEUR Workshop Proceedings 2454)*, Berlin, Germany, 2019. Available: http://ceur-ws.org/Vol-2454

[2] A. Belger, R. Budinich, R. Blum, M. Zablocki, and R. Zimmermann, “Market and Technology Monitoring driven by Knowledge Graphs,” Fraunhofer SCS; Technische Hochschule Nürnberg Georg Simon Ohm, 2020.

[3] X. Han *et al.*, “FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation,” 2018. Available: https://arxiv.org/pdf/1810.10147

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018. Available: https://arxiv.org/pdf/1810.04805

[5] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 2017, pp. 5998–6008. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need

[6] J. Lee, S. Seo, and Y. S. Choi, “Semantic Relation Classification via Bidirectional LSTM Networks with Entity-aware Attention using Latent Entity Typing,” Jan. 2019. Available: https://arxiv.org/pdf/1901.08163

[7] J. Pennington, R. Socher, and C. Manning, “Glove: Global Vectors for Word Representation,”*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014, doi: 10.3115/v1/D14-1162.

[8] M. E. Peters *et al.*, “Deep contextualized word representations,” Feb. 2018. Available: https://arxiv.org/pdf/1802.05365

[9] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, “ERNIE: Enhanced Language Representation with Informative Entities,” 2019. Available: https://arxiv.org/pdf/1905.07129

[10] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, “Translating Embeddings for Modeling Multi-relational Data,” in*Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013: Proceedings of a meeting held December 5-8*, Lake Tahoe, Nevada, United States, 2013, pp. 2787–2795. Available: http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data

[11] S. Wu and Y. He, “Enriching Pre-trained Language Model with Entity Information for Relation Classification,” May. 2019. Available: https://arxiv.org/pdf/1905.08284

[12] L. B. Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski, “Matching the Blanks: Distributional Similarity for Relation Learning,” 2019. Available: https://arxiv.org/pdf/1906.03158

[13] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Tylor, “Industry-scale Knowledge Graphs: Lessons and Challenges,”*ACM Queue*, vol. 17, no. 2, pp. 1–28, 2019, doi: 10.1145/3329781.3332266.

[14] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning, “Position-aware Attention and Supervised Data Improve Slot Filling,” in*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2017.

[15] S. Riedel, L. Yao, and A. McCallum, “Modeling Relations and Their Mentions without Labeled Text,” in*Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2010, pp. 148–163. Available: https://link.springer.com/chapter/10.1007/978-3-642-15939-8_10

[16] I. Hendrickx *et al.*, “SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals,” *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 33–38, 2010.

[17] T. Gao *et al.*, “FewRel 2.0: Towards More Challenging Few-Shot Relation Classification,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 6249–6254.

[18] F. Hamborg, N. Meuschke, C. Breitinger, and B. Gipp, “news-please: A Generic News Crawler and Extractor,” in*Proceedings of the 15th International Symposium of Information Science*, 2017. Available: https://www.researchgate.net/publication/314072045_news-please_A_Generic_News_Crawler_and_Extractor

[19] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych, “The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation,”*Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pp. 5–9, 2018.

[20] A. Bastos et al., “RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network,” Sep. 2020. [Online]. Available: https://arxiv.org/pdf/2009.08694

[21] T. Nayak and H. T. Ng, “Effective Modeling of Encoder-Decoder Architecture for Joint Entity and Relation Extraction,” Nov. 2019. [Online]. Available: https://arxiv.org/pdf/1911.09886

[22] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 2019. Available: https://arxiv.org/pdf/1907.11692

[23] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,” Sep. 2019. Available: https://arxiv.org/pdf/1909.11942

[24] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, “SQuAD: 100, 000+ Questions for Machine Comprehension of Text,”*CoRR*, abs/1606.05250, 2016.

AUTHORS

Christoph Brandl is a Software Engineer at GfK. He studied Information Systems and Management at the Nuremberg Institute of Technology, where he graduated in January 2021 with a Master of Science. His focus lies on processing natural language texts with machine learning approaches.

Jens Albrecht is a full-time professor in the Computer Science Department at the Nuremberg Institute of Technology. He works and does research in the areas of data management, data analytics and machine learning with a focus on text.

Renato Budinich is a Research Assistant at Fraunhofer IIS in the Future Engineering group, which works on researching and developing data insights for Businesses. His focus lies on applying natural language processing on social media texts and other online sources.

WHAT MAKES US CURIOUS? ANALYSIS OF A CORPUS OF OPEN-DOMAIN QUESTIONS

Zhaozhen Xu[1], Amelia Howarth[2], Nicole Briggs[2]and Nello Cristianini[1]

[1]Intelligent Systems Laboratory, University of Bristol, Bristol, UK
[2]We the Curious, Bristol, UK

ABSTRACT

Every day people ask short questions through smart devices or online forums to seek answers to all kinds of queries. With the increasing number of questions collected it becomes difficult to provide answers to each of them, which is one of the reasons behind the growing interest in automated question answering. Some questions are similar to existing ones that have already been answered, while others could be answered by an external knowledge source such as Wikipedia. An important question is what can be revealed by analysing a large set of questions. In 2017, "We the Curious" science centre in Bristol started a project to capture the curiosity of Bristolians: the project collected more than 10,000 questions on various topics. As no rules were given during collection, the questions are truly open-domain, and ranged across a variety of topics. One important aim for the science centre was to understand what concerns its visitors had beyond science, particularly on societal and cultural issues. We addressed this question by developing an Artificial Intelligence tool that can be used to perform various processing tasks: detection of equivalence between questions; detection of topic and type; and answering of the question. As we focused on the creation of a "generalist" tool, we trained it with labelled data from different datasets. We called the resulting model QBERT. This paper describes what information we extracted from the automated analysis of the WTC corpus of open-domain questions.

KEYWORDS

Deep Learning, Natural Language Processing, Question Answering, BERT.

1. INTRODUCTION

In 2017 "Project What If" was started at the "We the Curious" (WTC) science-centre of Bristol (UK), with the stated intention of being the first exhibition all about "the curiosity of a city". Its aim was no less than capturing the curiosity of Bristolians by collecting all their questions. It was focused on the questions "of real people", and through these is aimed at understanding what Bristolians were curious about.In other words, it was not so much about the answers to individual questions, as it was about understanding a Community from the questions it asks.

Despite the clear identity of WTC as a science-centre, the organisers of this project were trying to gauge a broader set of concerns, about culture and society, in a time of rapid change. A collection of the spontaneous questions of thousands of people was expected to tell us a lot about the people who asked them.

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 91-105, 2021. CS & IT - CSCP 2021 DOI:10.5121/csit.2021.112007

Over the following three years, the project gathered over 10,000 questions, both in their "museum" venue and in initiatives around the city. That list taken together contained many questions, worries, doubts, and ambitions of thousands of citizens.

At the end, just one final question remained: what did that vast corpus contain? This is not something that can be answered by a single person reading the questions, but also not by a simple statistical analysis. What is needed is intelligent software capable of understanding the questions, their type and topic. As a further level of ambition, we asked: would the AI system be able to answer some of these questions?

We report here on the first content analysis of that set of questions, which was performed with Artificial Intelligence tools specifically created for that task. The AI algorithm was based on Deep Learning technologies and was tasked to solve three main problems: detecting topic of questions; detecting equivalent questions with similar meaning but different wording; locating potential answers to these same questions in Wikipedia.

The field of automated Question Answering (QA) is a new but fast-growing branch of AI, driven by commercial systems such as Alexa and Siri. According to a US report on smart speaker consumer adoption, 84.0% of their users had tried to ask a question through the speaker, 66.0% and 36.9% did so on a monthly and daily basisrespectively [1]. But at the core of any method for QA (as well as other question processing tasks) there is the challenge of representing a short sentence in a way that reflects its meaning. For this purpose, we made use of a deep-learning technique known as "BERT"[2] which will be described below.

We discover that over half of these questions were about the topic of Science and Mathematics, and over a quarter were of the type WHY and HOW. These are known as factual questions and can sometimes be answered by automated systems, perhaps on the basis of Wikipedia. But what was more interesting was the large number of non-factual questions. For example, the counterfactual ones of the type IF which could not be handled in this way.

The QA task itself can be described as an open-domain open-book question answering task, in that no limitation is posed a priori on the topic of the question, and the answering system is allowed to "look at the book" in order to answer.

The main contributions of this article are: a general-purpose method to represent short questions, that is useful for a range of different tasks; and a statistical overview of the contents of the WTC corpus, enabled by that method.

The article describes the WTC corpus in Section 2, the algorithm in Section 3, the content-analysis of the corpus in Section 4, and the discussion of results in Section 5.

2. WTC CORPUS OVERVIEW

We will call our dataset of open-domain questions "the WTC corpus", this section describes its origin and main features.

The dataset is originated from a project run in Bristol (UK) by "We The Curious" (henceforth WTC), an educational charity and science centre.

Between January 2017 and October 2019, WTC collected over 10,000 open-domain questions from a diversity of sources: onsite (at the WTC venue in Bristol), offsite, and online. Offsite question gathering ensured questions were received from various Bristol postcodes (BS1 – BS16), and general submissions were received from all remaining postcodes. These refer to questions collected in the venue, written on cards by visitors, later stored and entered manually into the question database.

Based on this initiative, WTC created a digital database of questions, which is in WTC's possession. WTC is responsible and accountable for protecting the personal data of individuals submitting this information alongside their questions. All personal data is held by WTC in compliance with GDPR protocol and personal data is not shared with other parties, including the analysis team of this project. For the purpose of the present study a smaller dataset was generated, by removing all the personal data that was associated to the questions, and only this was shared with the analysts (ZX and NC).

Manual Curation of the WTC Corpus. The raw corpus also included repeated questions, various types and topics, and other non-question sentences. Questions were first moderated manually by WTC staff. The questions in the database were also screened for any possible identifying information or potentially offensive or inappropriate language or content. These were removed from the database. After moderation, the resulting dataset contained 10,073 questions.

This second, anonymised and moderated, textual dataset is what we will call the WTC-corpus in this paper.

Automated Pre-processing: Some simple pre-processing was performed before content analysis, such as removing exactly identical questions and questions shorter than three words. After these steps, the filtered WTC dataset contained 8.600 questions, using 5,732 words. The length of questions is between 3 and 55 words, with an average of 7.15 words. 87.96% of the questions are within 10 words.

The word cloud in figure 1 shows that the questions cover universe and space, human body, energy and climate change, animals and plants, chemistry and materials, the future and some other topics outside of the typical science categories listed before. We also identified 5,022 "equivalent" question pairs, as will be described in section 4.2.

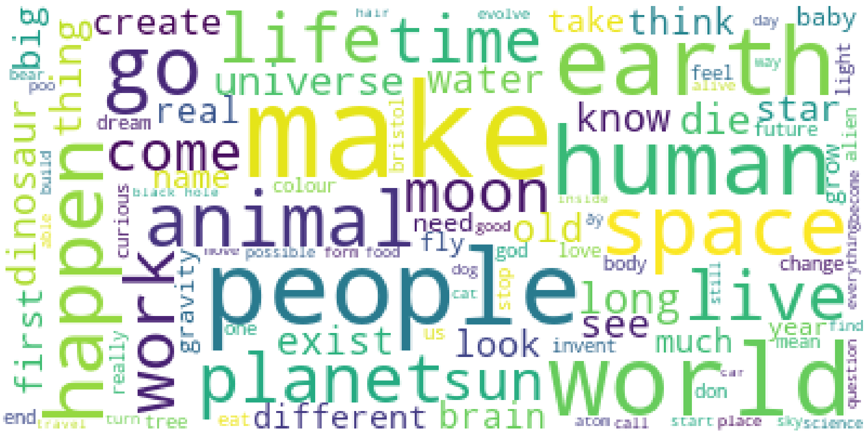


Figure 1. The word cloud is generated from the curated and filtered WTC corpus. The words were lemmatised before generating the graph. The size of the word is proportional to its frequency in the corpus.

3. QBERT: A MULTI-TASK QUESTION-PROCESSING VERSION OF BERT

In order to process the questions in our corpus we embedded them into a vector space of 768 dimensions, using a model based on the BERT method.

In particular, we fine-tuned a BERT based model (sentence-BERT, or S-BERT [3]) by using a diverse set of question-related datasets, which will be described below. For convenience in our experimental comparisons, we named this refined BERT model “QBERT” to indicate that it was specifically fine-tuned to handle questions.

BERT is a standard method for the representation of sentences, based on the technology of Transformers (more specifically, multiple stacked encoders) as described in [2].It adds a classification token [CLS] at the beginning of the input sequence and encodes the input sequence by assembling its token embeddings, segment embeddings, and position embeddings. The bidirectional transformer encoder [4] is then trained with two unsupervised tasks: masked language model and next sentence prediction. The encoder’s output can be used for downstream tasks like classification, question answering, and sentence tagging.

Fine-tuning is an important step in using BERT, as it adapts the model to the specific class of sub-tasks at hand. To train a model which can understand the content of the corpus and find possible answers to the questions, we used three kinds of tasks to fine-tune the standard BERT model: Question-Equivalence (QE), Question-Answering (QA), and Question-Topic (QT), defined as follows.

- QE BERT is given two questions and is required to decide whether they are equivalent.
- QA BERT is given a question and a set of candidate answers and is required to decide which of them is the correct answer.
- QT BERT is given a set of questions and topics and is required to determine the topic of each question.

We tuned the parameters of a pre-trained BERT on each of these tasks, usingthree different datasets that will be described below. We also measured its performance on each of these tasks separately. Our focus was not on achieving record performance on any of these tasks, but rather on creating a model that can perform well on each of them: rather than three specialists, we wanted a generalist model.

3.1. The Method

In order to train the model, we reduced the three NLP tasks described above to a series of standard classification tasks: QE determines if a pair of questions are equivalent or not; QA determines if a candidate answer is appropriate for a given question; and QT categorises the text by topic.

Recent success of pre-training language models proved that training and fine-tuning a single model could increase performance in different tasks [2], [5], [6]. Our approach is based on S-BERT [3], a modified BERT that captures sentence similarity and provides an embedding for a given sentence. Comparing to the original BERT that uses the [CLS] token as sentence embedding, S-BERT applies a pooling method to compute the mean of all output vectors from BERT. In addition, S-BERT concatenates the sentence embeddings with the element-wise difference of the sentence pairs during training so that semantically similar sentences are close to each other in vector space. Another advantage of S-BERT is that it is more time-efficient in finding the most similar sentence while combining with Faiss [7]. As it is shown in figure 2, while training on classification task, S-BERT calculates $softmax(W_t \cdot (u, v, |u - v|))$to predict the label for the sentence pairs. W_t is a trainable weight, and $|u - v|$ is the element-wise difference of the embeddings. The BERTs in figure 2 share the same parameters during training. Through inference, S-BERT generates embeddings with the trained model. The distance between the embeddings can be measured with cosine distance.

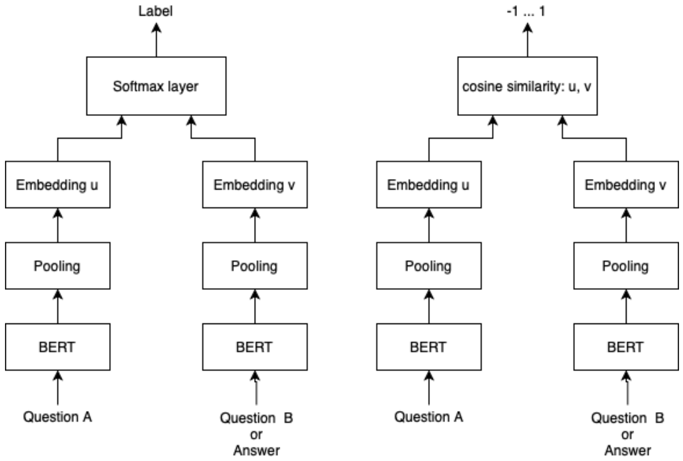


Figure 2. S-BERT architecture. Left: Training on classification task, Right: Inference by giving a cosine similarity between sentences. All the BERTs share the same parameters.

3.2. Training Datasets

One of the main challenges for this research is the lack of labelled data. It is expensive to create reliable labels for each task. Thus, the model is trained and fine-tuned on some other existing datasets then transferred to analyse WTC.

Quora Question Pair (QQP) [8] is a question pair identification competition first released on Quora in 2017. It contains 404,290 pairs of different questions from Quora with annotations. After embedded with S-BERT, there are 537,931 unique questions in the dataset. QQP competition intends to classify if the questions are duplicated, which is ideal for fine-tuning our model. By training on QQP, we figure out the distance threshold that can be transferred to

WikiQA [9] is a question-answering dataset that extracts the questions from real-world query logs on Bing. All the questions in WikiQA are factual queries that start with Wh-word and have at least 5 users click on a Wikipedia page after searching. The answers are consist of candidate sentences from Wikipedia and human labelled as a correct answer or not. The dataset includes 3,047 questions and 26,154 sentences; 1,239 of the questions contain a correct answer from Wikipedia. Training with WikiQA enables us to evaluate our question answering system on open-domain.

Yahoo! Answer [10] is a corpus generated from Yahoo! Research Alliance Webscope program. The corpus contains 1,460,000 samples in 10 different topics. Each sample includes the topic, question title, question content, and the best answer. During training, Yahoo! Answer was separated into two datasets, Yahoo Topic (YT) and Yahoo Question-Answering (YQA). YT contains all the questions and categories used for QT training. YQA is made up of questions and the corresponding answers that are less than 35 words. There are 754,566 question-answer pairs in YQA.

3.3. Training and Fine-tuning

Following S-BERT, we trained the model with classification tasks. The basic S-BERT was only trained on natural language inference dataset and semantic textual similarity dataset that contain sentence pairs with labels such as SNLI [11], NLI [12], and STS [13] dataset. Thus, it has poor performance in detecting similar question pairs. Following the architecture of S-BERT in figure 2, we trained the sentence embedding model to learn the similarity between questions with data from QQP. The length was limited to 35 tokens for each input sequence because 99.93% of the questions are shorter than 35 words in the WTC corpus. The sequences with more than 35 words were truncated after the limited length.

We have used the classification technique described in S-BERT [3] for QE and QA classification. For QQP and WikiQAdatasets, the model took questions pairs or questions answer pairs as the input and produces the label in terms of *1* or *0*. *1* represents that the input sequences are similar or related. Each input sequence was tokenised and embedded by BERT-base then produce an embedding with 768 dimensions. BERT-base used in this model is a smaller BERT version containing 12 layers and 110M parameters. All the weights in BERT were updated during training. Comparing to softmax loss in S-BERT, the contrastive loss is more capable of mapping the similar vector in high dimensional space into nearby points in a lower dimension [14]. Hence, we minimised the online contrastive loss and optimised it by Adam optimiser with a learning rate of *2e-5*. The contrastive loss combines loss from both positive samples and negative samples with a margin of 0.5. The margin ensures that negative samples have a more significant distance than the margin value. YQA was introduced as a supplement dataset for QA task because WikiQA did not have enough data considering the size of the model. Since YQA only contains corresponding question-answer pairs, multiple negatives ranking loss that requires only positive labels is applied instead of online contrastive loss.

In QE and QA, instead of classifying if the sequences are related, it is more important that the system can retrieve all the related sequences for given questions. The problem is how to quantify `related' with embeddings.A cosine similarity threshold was introduced in this model. First, all the sequences in the training set were embedded with the fine-tuned model. The sequence pairs were classified as positive if they have higher similarity than the threshold. We used 2 different strategies to decide the threshold for QE and QA. For identifying similar question pairs, the similarity threshold with the best accuracy in the QQP was found to quantify any questions pairs during training. On the other hand, for question-answer pairs, we leveraged the threshold with the best precision in WikiQA instead. While retrieving answers from the knowledge base, there are usually millions of candidates and we wanted the answer to be as reliable as possible.With both sequence embeddings and the threshold observed above, the model is capable to classify and search all the related sequences in the corpus by calculating the cosine similarity between sequences.

Contrary to QE and QA, QT took one question as input and predicted the question topic with the embedding.We have used the similar classification technique described in S-BERT[3], but only one BERT was needed in the network. An additional softmax layer was applied after BERT to map the embedding into probability for each topic. We fine-tuned the trained BERT and the softmax layer with extra data in YT.

QBERT was trained for 10 epochs, respectively for each dataset, with the training data divided by the data provider for each task. Except for YT, all the data was applied during training and trained for 5 epochs. For QE and QA, the network was trained with a batch size of 150. Moreover, for QT, the batch size was 350. The model was evaluated by accuracy for all classifications. In answer retrieval, we evaluated accuracy, precision and recall for the first answer.

In this experiment, the networks were trained with 1 GeForce GTX TITAN X GPU. It took 7 hours, 0.5 hours, 9 hours, 16.5 hours to train on QQP, WikiQA, YQA, and YT, respectively.

3.4. Performance of QBERT

The results for models fine-tuned with different datasets are illustrated in table 1. In QE and QA tasks, the embedding network was trained with pairs of sentences, so with more semantic textual similarity datasets, they both achieve better performance. However, for QT, it was trained with a different structure fine-tuned based on QE and QA results with the same YT dataset. QT task does not require the embedding model to capture the sentence similarity. Therefore, pre-training with more semantic textual similarity datasets does not significantly affect the result of QT. Furthermore, the fine-tuned QT model performs worse on other tasks because itonly trained with one BERT and limited the performance on capture sentence similarity.

While pre-training the sentence embedding model with all three datasets for QE, we observed that the order in which the training data are presented has a great effect on the final result. As shown in table, with the same datasets, the model trained with QQP as the last outperforms the model trained QQP first around 10.16%, from 80.13% to 90.29%. Besides, the cosine similarity threshold increases from 0.825 to 0.875 means that the questions with similar meanings are closer to each other in vector space.

Possible remedies to this effect will be the object of a separate study, as they are not relevant to the problem we are addressing in this paper.

In QA, we fine-tuned S-BERT on the classification task and evaluated it on both classification and retrieval tasks. Similar to QE, the best threshold contains more datasets from different tasks. In the classification task, the original S-BERT trained on STS+NLI outperforms other models. However, this is due to the bias of WikiQA dataset. 94.89% of the question-answer pairs in the training set of WikiQA are labelled as 0, and 95.24% in the test. S-BERT does not manage to identify the correct answer, and it only uses a large threshold to ensure that all the question-answer pairs are categorised as negative. WikiQA only contains question-answer pairs in the dataset. In order to evaluate the performance on retrieval task, a knowledge base that includes all the candidate sentences in WikiQA dataset was generated. And during evaluation, we leveraged only questions with a correct answer in the answer base. As shown in Table 2,in QA retrieval,training with extra YQA data dramatically increase the accuracy of the WikiQAtest set.

Table 1. Performance of QT, QE, and QA classification tasks. The models are trained with different datasets. The model's name indicates the training sequence of each dataset.

	QT		QE			QA - Classification		
	Accuracy		Threshold	Accuracy		Threshold	Accuracy	
	Train	Test		Train	Test		Train	Test
STS+NLI	85.71%	72.44%	0.800	74.39%	74.77%	0.910	94.89%	95.24%
QQP	82.86%	72.32%	0.850	90.48%	89.59%	-	-	-
QQP+WikiQA	86.21%	72.32%	0.825	91.42%	85.44%	0.713	99.99%	94.70%
QQP+WikiQA +YQA	86.57%	72.51%	0.825	82.08%	80.13%	0.755	95.16%	94.74%
YQA+WikiQA+QQP	86.78%	72.14%	0.875	99.20%	90.29%	0.797	96.34%	93.92%
QQP+YQA +WikiQA	86.58%	72.37%	0.850	78.27%	76.49%	0.756	99.97%	95.18%

Table 2. Performance of QA Retrieval tasks. The models are evaluated on the WIKIQA dataset. The model's name indicates the training sequence of each dataset.

	QA - Retrieval					
	Accuracy@1	**Precision@1**	**Recall@1**	**Accuracy@1**	**Precision@1**	**Recall@1**
	Train			**Test**		
STS+NLI	21.64%	21.64%	19.74%	29.88%	29.88%	27.21%
QQP	-	-	-	-	-	-
QQP+WikiQA	82.06%	82.06%	78.53%	28.63%	28.63%	27.07%
QQP+WikiQA+YQA	53.81%	53.81%	50.53%	46.47%	46.47%	43.36%
YQA+WikiQA+QQP	60.92%	60.92%	57.48%	37.76%	37.76%	34.85%
QQP+YQA+WikiQA	85.17%	85.17%	81.28%	46.06%	46.06%	42.36%

According to QE and QA performance in Table 1 and Table 2, we noticed that the best models for each specific task might have poorer accuracy on another. Besides, both of the models with the best performance are over-fitted on the training set. Therefore, we used the model following the training sequence of QQP, WikiQA, YQA, which has a more balanced performance on all three tasks, as our multi-tasking generalist QBERT. We applied this generalist QBERT to our task of the content analysis of the WTC corpus.

4. CONTENT ANALYSIS OF WTC CORPUS

In the WTC corpus, we have questions with various content from different people. There are factual questions like *"How are mirrors made?"* and *"Who built the internet/electricity?"*; and counterfactual questions such as *"How long will the earth and humans last if we carry on damaging it and nothing changes?"* and *"Would a car weigh more if there was a flying pigeoninside of it?"*. They cover multiple topics and have overlap in the content. By using QBERT trained and fine-tuned with the public datasets YT, YQA, WIKIQA, and QQP, we also analysed our WTC corpus in the three different tasks of QT, QE, and QA.

4.1. QT

To classify the questions, we first identified each question's type and topic. The type of the question was categorised into the following categories: WHAT, WHO, HOW, WHEN, WHERE, WHY, WHICH, and IF which will be further discussed below. For this first classification, we just used simple keyword-matching. On the other hand, the topic of the question was classified by the trained network in section 3.3. The topics included: Business & Finance, Computers and Internet, Education & Reference, Family & Relationships, Health, Politics & Government, Science & Mathematics, Society & Culture, Sports.

It is important to notice that there are many non-scientific questions in this list, which was part of the initial intent of the overall project: to assess the scope and breadth of the curiosity of an entire community.

Notice also that we had 9 types and 10 topics, and therefore 90 Question "Themes" to which we could allocate the over-8000 distinct questions that have survived the various stages of filtering.
Besides the 7 WH-questions, and HOW, we have also defined a further class of questions that we call IF questions. The aim was to find a simple way to approximate the counterfactual questions

of the type “what if”, which however are difficult to capture exactly by keyword matching, but can well be approximated in this context by checking for the use of the word “if”.

A counterfactual (CF) question is defined as a question of the type: “what would happen if X was true”. The understanding is that X is not a true fact, but the asker of the question is considering the possible consequences of X being true. This kind of question takes its name from being “counter to the facts”, is often used in defining the notion of Causality (e.g.in [15]), and indicates a mental process directed at understanding the mechanism behind observations.

Given a question, we assigned it to the type of the first keyword from our list that was found in it, with one notable exception described below. For example, the question *“Why do we get butterflies when we like someone?”* was categorised as WHY. However, questions that contain the keyword “if”, such as “what if” and “How ... if ...” were classified as IF questions. The category OTHER includes yes/no questions or sequences that do not fit into other categories.

Then we applied the topic classification model to identify 10 topics in WTC. Table 3 shows the frequency distribution of the questions across types and topics. The most “asked” topics in the corpus are science & mathematics and society & culture, which make up 66.35% of the corpus. Moreover, half of the questions are HOW and WHY questions.

4.2. QE

Although the corpus was pre-processed to remove identical questions, there are still many “equivalent” questions left in the corpus. For the purpose of this study, we consider two questions as equivalent if they have the same answer. For instance, questions *“What is our purpose?”*and *“What is the aim of our life?”*have different wordings but they have the same answer. Finding equivalent questions in the corpus helps us to further understand any patterns, such as clusters, in the set.

Table 3. Contingency table for topics the types in WTC

	how	**what**	**when**	**where**	**which**	**who**	**why**	**if**	**other**	**Percentage（%）**
Business & Finance	121	100	16	18	0	26	191	30	136	7.42
Computers & Internet	34	9	3	2	0	3	18	5	34	1.26
Education & Reference	132	81	8	11	2	50	84	16	68	5.26
Entertainment& Music	55	56	10	10	0	12	80	39	108	4.30
Family & Relationships	44	32	8	8	0	1	95	14	68	3.14
Health	159	66	18	10	0	6	299	34	84	7.86
Politics & Government	23	18	7	2	0	5	57	22	51	2.15
Science & Mathematics	1355	646	88	99	15	58	1107	392	918	54.40
Society & Culture	142	159	23	21	0	52	286	108	237	11.95
Sports	47	14	5	0	0	15	48	7	59	2.27
Percentage（%）	24.56	13.73	2.16	2.10	0.20	2.65	26.34	7.76	20.50	

To better understand the performance of finding equivalent questions in WTC, we randomly sampled 1,000 questions and generated a list of candidate question pairs. The questions were embedded by the S-BERT that trained only with the NLI dataset. The top 10 questions with the largest cosine similarity were selected as similar question pair candidates for each question. For duplicate question pairs such as [Q1, Q2] and [Q2, Q1], we only kept one of them for annotation. The question pairs were labelled with 0 or 1, where 0 represents different, and 1 for similar. Theauthor labels 5,022 pairs of candidate questions, 728 pairs are similar, and 4,294 pairs are different. Table 4 provides some examples of the data we label.

When the cosine similarity threshold is 0.825, the model obtains 90.80% accuracy on the sampling data. QBERT obtains a better accuracy on WTC with QQP, which has 80.13% accuracy on the test set. This proves that QBERT can be applied to a corpus of unseen questions.

Moreover, we identified clusters by applying a "graph community detection" method [16], [17]in order to group similar questions. To cluster the questions, a graph is built with question nodes using the cosine distance matrix. An edge is added to the nodes if the distance between a pair of questions is smaller than *1-cosine_similarity*. There are 6,060 communities found in the WTC corpus, which represents 6,060 different questions in the corpus. Of these, 5,398 questions do not have any similar question in the corpus.

Table 4. Examples from labelled WTC question pairs

qid1	Question1	qid2	Question2	Label
2	Who is the richest?	6992	Why can't I be rich?	0
33	How do you make glass?	2001	How is glass made	1
50	When did the humans come alive?	7257	When did humans first exist?	1

4.3. QA

We are also interested in whether WTC questions can be answered (with high confidence) by the QBERT model. The model aims to retrieve a sentence as the answer from an unstructured knowledge base.

For WTC, we used Wikipedia summary [18]as a knowledge source during inference. The corpus includes the title and the first paragraph as the summary for each Wikipedia article extracted in September 2017. The raw texts of the Wikipedia have 116M sentences initially. Of these, 22M are in the summaries. After we embedded with QBERT, 21M sentences have different embeddings. The summary of Wikipedia provides the article's primary information. In the meanwhile, the summary reduces about 80% of the sentences from the original Wikipedia.

In order to retrieve the answer from Wikipedia, we calculated the average distance of correct answers in the QA retrieval task described in Section 3.4. The best-scored sentence from Wikipedia Summary with a higher similarity than 0.688 was considered as the answer for given question.

In order to embed all the sentences of Wikipedia summary with SBERT, we used 7 GeForce GTX TITAN X GPU and took 1.5 hours to encode all the sentences. Due to the scale of the dataset, we located answers to questions by using the method of approximate nearest neighbour. The index was trained and built using the inverted file with exact post-verification for 4 hours [7].

After building the index, 3 minutes were needed to search the set of answers for all the 8,600 WTC questions with one GPU, an average of 0.02s per question.

The percentage of questions in different types and topics that can be answered with high confidence are shown in Table 5. There are 24.69% of the questions in WTC that can be answered by QBERT.

Table 5. Number of questions in WTC can be answered with high confidence over the number of the groups.

	how	what	when	where	which	who	why	if	other	All
Business & Finance	32/ 121	44/ 100	4/ 16	5/ 18	0/ 0	12/ 26	42/ 191	2/ 30	34/ 136	175/ 638
Computers & Internet	6/ 34	2/ 9	0/ 3	0/ 2	0/ 0	0/ 3	1/ 18	0/ 5	4/ 34	13/ 108
Education & Reference	42/ 132	36/ 81	2/ 8	1/ 11	2/ 2	14/ 50	14/ 84	1/ 16	14/ 68	126/ 452
Entertainment & Music	9/ 55	17/ 56	2/ 10	4/ 10	0/ 0	4/ 12	7/ 80	1/ 39	22/ 108	66/ 370
Family & Relationships	7/ 44	9/ 32	0/ 8	3/ 8	0/ 0	0/ 1	16/ 95	2/ 14	13/ 68	50/ 270
Health	15/ 159	12/ 66	4/ 18	0/ 10	0/ 0	0/ 6	61/ 299	3/ 34	15/ 84	110/ 676
Politics & Government	3/ 23	6/ 18	2/ 7	1/ 2	0/ 0	1/ 5	12/ 57	1/ 22	6/ 51	32/ 185
Science & Mathematics	416/ 1355	240/ 646	21/ 88	27/ 99	6/ 15	16/ 58	339/ 1107	51/ 392	186/ 918	1302/ 4678
Society & Culture	21/ 142	57/ 159	4/ 23	4/ 21	0/ 0	9/ 52	49/ 286	13/ 108	51/ 237	208/ 1028
Sports	11/ 47	5/ 14	2/ 5	0/ 0	0/ 0	4/ 15	4/ 48	0/ 7	15/ 59	41/ 195
All	562/ 2112	428/ 1181	41/ 186	45/ 181	8/ 17	60/ 228	545/ 2265	74/ 667	360/ 1763	2123/ 8600

5. Discussion of Results

"Project What If" was launched in 2017 across Bristol and involved thousands of people. Its aim was to focus on the questions that were most asked by ordinary Bristolians, rather than on the answers, to see what they said about the local Community.

The automated analysis of that corpus, enabled by QBERT in Section 4, revealed that more than half of the questions are in the domain of Science & Mathematics (54.10%), followed by Society & Culture (12.57%), and then by Health (7.70%). The most frequently asked type of question is of the type WHY (26.34%) followed by HOW (24.56%).

By navigating the IF question, we observed that most of the questions are counterfactual such as *"What if we never went to sleep?"*, *"If you could hear in space, how loud would the Sun be?"*. However, the corpus also contains a number of factual questions, as would be expected in a science-centre setting. For example, *"I'd like to know if atoms are made up of other atoms."*.

Furthermore, we calculated the *P(type, topic)* and *P(type)*P(topic)* to understand the associations between type and topic. The question-type WHO is strongly associated with Education & Reference and with Sport because the *P(Who, Sport)* is 3 times larger than the probability of *P(Who)*P(Sport)*. The topic Education & References strongly associates with types: how, what, when, who. The type IF associates strongly with Politics & Government, but not with Education & Reference.

One limitation of pre-training with Yahoo! Answer dataset is that it only takes the top 10 topics from Yahoo! Answer regardless of all other possible topics. During labelling the topic for WTC corpus, we noticed that many questions did not belong to any of the groups in YT. QBERT can be improved with more question topics or labelling the unknown topics.

After applying QBERT, we found answers for 2,123 questions in the WTC corpus. WH questions, such as WHICH, WHAT, HOW, WHERE, and WHO are more likely to be answered by Wikipedia Summary comparing to IF questions and yes/no questions. Due to the QBERT mechanism, the answer is supposed to be one sentence from Wikipedia. In this case, factoid questions which can be answered with fact expressed in a short sentence are more likely to be answered. Furthermore, non-factoid questions, like some of the WHY or IF questions, that require more explanation in the answer are harder to find one sentence answer from a knowledge source. More than 50% of the Education & Reference, Science & Mathematics questions, and Business & Finance can be answered with confidence by Wikipedia Summary. However, QBERT can only answer around 12% and 17% of the questions in Computers & Internet and Politics & Government, respectively.

Here are some examples of the answers giving by QBERT. The questions are from the WTC corpus, and the answers are from the Wikipedia Summary.

- Q1: *How old is the oldest tree in the world?*
  A1: *A scientific investigation in 1965 of the tree's rings indicated that the tree has an estimated age of 1450-1900 years, and may well be the oldest living oak in northern Europe. (Score: 0.817)*
- Q2: *In the future, will robots gain conciousness?*
  A2: *Throughout history, it has been frequently assumed that robots will one day be able to mimic human behavior and manage tasks in a human-like fashion. (Score: 0.775)*
- Q3: *How do birds lay their eggs?*
  A3: *They lay their eggs into the wet dangling roots of plants. (Score: 0.788)*

From the question answering pair found by QBERT, we notice that for questions lacking in attributes such as Q1 in the example, the retrieved answer is adapted to a specific attribute rather than a general situation as human understanding. Another barrier in QBERT is that the answer sometimes is not included in one single sentence. For example, in A3, the original summary is *"Zygonyx is a genus of dragonflies ... They lay their eggs into the wet dangling roots of plants."*. However, the QBERT is only able to retrieve the most relevant sentence. In this case, QBERT fails to capture the entity, which is more important in the question.

We also observe that some answers with high confidence (over 0.85 cosine similarity) are similar questions that QBERT found in Wikipedia. For example, QBERT is tricked by a sentence in Wikipedia *"Why Does the Sun Shine?"* and considers it the answer to the question *"Why is the sun bright?"*.

6. RELATED WORK

Question answering is always a challenging research task in NLP. Meanwhile, question pre-processing like topic classification and similar question classification is critical to a large scale question answering system.

The pre-trained language models earn state-of-the-art results in many NLP tasks. The model we used, BERT [2], and its modified models have leading performance in classification [19], [20] and question answering[21]–[23].

For question answering, we identified our research as open-domain and open book answer retrieving. The system was designed to infer the correct answer from knowledge sources like Wikipedia in a concise sentence. Comparing to answering question using a knowledge base, open-domain QA is more challenging in using large-scale knowledge sources and machine comprehension. Previous research [21]–[25] leveraged a retriever-reader or retriever-generator that retrieved the relevant passage from the knowledge source and extracted an answer span from the passage. The passage can be a document, paragraph, sentence or fixed-length segment. However, this two stages system is computationally expensive. Inspired by DenSPI [26], QBERT encodes all the sentences in the knowledge base and searches the most relevant sentence with the query. In addition, we perform approximate nearest neighbour search to reduce the searching time.

Some researchers use Glove word embedding [27] or BERT [CLS] token as sentence embedding to encode the questions and the knowledge base. Instead, we trained S-BERT [3], a sentence embedding network that fine-tuned BERT with similar sentences, to retrieve answers. Because S-BERT outperforms Glove and BERT [CLS] in textual similarity tasks. Besides, it reduces the complexity of embedding sentences with BERT.

7. CONCLUSIONS

What did the corpus of questions collected by WTC reveal about the Community that generated it? This was the last question that remained unanswered, and we hope that our AI analysis can provide a first insight.

QBERT, a new generalist model for question-content analysis was applied to the WTC corpus. In the results, we see that the contributors to “Project What If” were very interested in Science, Society, and Health; and asked many questions of the WHY and HOW type. This is not surprising within the setting of WTC as a science institution. But the next finding revealed a lot more: questions of the type IF tend to relate to Politics & Government topics and not with Education & Reference topics. Are Bristolians exploring alternative ways to be a Community?

Curiosity about Society & Culture is also very revealing. This is an emerging theme in the sector of science centres, where there is an ongoing discussion about expanding from Science Centres to Science & Cultural Centres. More generally, there is a movement in the sector currently to explore society and culture alongside traditional science such as Biology, Engineering, Chemistry etc. This seems to be reflected in the kind of questions Bristolians have been asking.

In our modern world, access to information is easy and comprehensible through digital and online channels. Science centres have therefore been challenged to adapt to this changing environment, incorporating social sciences and perspectives from different cultures and presenting a space for exploration of ideas rather than just answers. This is why We The Curious has based “Project What If” on questions, exploration and curiosity, rather than just education and knowledge sharing. These findings support the rationale and aims of these changes.

With QBERT, more than 50% of the questions from Science & Mathematics and Education & Reference topics have been answered. Moreover, the QA system can answer a high percentage of WH questions except for WHY. QBERT is also computational efficient during retrieval answer from Wikipedia. It takes 0.02s per question. Although QBERT managed to answer 43.8% of the questions, there are still some limitations with the QA model. We can further explore the QA system in the future to overcome these deficiencies. One of the possible ways is that we can encode the paragraph instead of sentences for question answering.

Note: The anonymised and moderated dataset is available from We The Curious on reasonable request for research purposes. Please contact information@wethecurious.org.

ACKNOWLEDGEMENTS

We acknowledge Anna Starkey for her central role in "Project What If" and being part of the team who initiated this research, and David May for bringing us all together and providing his insight along the way.

REFERENCES

[1] B. Kinsella and A. Mutchler, "Smart speaker consumer adoption report 2019," 2019.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Nov. 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[4] A. Vaswani et al., "Attention Is All You Need," 2017.

[5] A. Radford and T. Salimans, "Improving Language Understanding by Generative Pre-Training," OpenAI, pp. 1–12, 2018, [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[6] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. v Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," 2019. [Online]. Available: https://github.com/zihangdai/xlnet

[7] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, Feb. 2017, [Online]. Available: http://arxiv.org/abs/1702.08734

[8] K. Csernai, "Quora Question Pairs," 2017. https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs (accessed Jul. 04, 2021).

[9] Y. Yang, W.-T. Yih, and C. Meek, "WIKIQA: A Challenge Dataset for Open-Domain Question Answering," in 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2013–2018. [Online]. Available: http://aka.ms/WikiQA.

[10] X. Zhang, J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for TextClassification," in 28th International Conference on Neural Inforamtion Processing Systems, Feb. 2015, pp. 649–657. [Online]. Available: http://arxiv.org/abs/1502.01710

[11] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in 2015 Conference on Empirical Methods in Natural Language Processing, Aug. 2015, pp. 632–642. [Online]. Available: http://arxiv.org/abs/1508.05326

[12] A. Williams, N. Nangia, and S. R. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," Apr. 2017, [Online]. Available: http://arxiv.org/abs/1704.05426

[13] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation," Jul. 2017, doi: 10.18653/v1/S17-2001.

[14] R. Hadsell, S. Chopra, and Y. Lecun, "Dimensionality Reduction by Learning an Invariant Mapping," 2006. [Online]. Available: http://www.cs.nyu.edu/~yann

[15] J. Pearl, "Causal and Counterfactual Inference," 2019.

[16] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," Phys. Rev. E, vol. 70, no. 6, 2004.

[17] A. A. Hagberg hagberg, lanlgov -Los, D. A. Schult, and P. J. Swart swart, "Exploring Network Structure, Dynamics, and Function using NetworkX," in 7th Python in Science Conference, 2008, pp. 11–16. [Online]. Available: http://conference.scipy.org/proceedings/SciPy2008/paper_2

[18] T. Scheepers, “Improving the Compositionality of Word Embeddings,” 2017.

[19] C. H. McCreery, N. Katariya, A. Kannan, M. Chablani, and X. Amatriain, “Effective Transfer Learning for Identifying Similar Questions: Matching User Questions to COVID-19 FAQs,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2020, pp. 3458–3465. doi: 10.1145/3394486.3412861.

[20] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to Fine-Tune BERT for Text Classification?,” in China National Conference on Chinese Computational Linguistics, May 2019, pp. 196–206. [Online]. Available: http://arxiv.org/abs/1905.05583

[21] K. Lee, M.-W. Chang, and K. Toutanova, “Latent Retrieval for Weakly Supervised Open Domain Question Answering,” in 57th Annual Meeting of the Association for Computational Linguistics, Jun. 2019, pp. 6086–6096. [Online]. Available: http://arxiv.org/abs/1906.00300

[22] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, “REALM: Retrieval-Augmented Language Model Pre-Training,” Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.08909

[23] V. Karpukhinet al., “Dense Passage Retrieval for Open-Domain Question Answering,” Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.04906

[24] D. Chen, A. Fisch, J. Weston, and A. Bordes, “Reading Wikipedia to Answer Open-Domain Questions,” Mar. 2017. [Online]. Available: http://arxiv.org/abs/1704.00051

[25] W. Yang et al., “End-to-End Open-Domain Question Answering with BERTserini,” Feb. 2019. doi: 10.18653/v1/N19-4013.

[26] M. Seo, J. Lee, T. Kwiatkowski, A. P. Parikh, A. Farhadi, and H. Hajishirzi, “Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index,” in 57th Annual Meeting of the Association for Computational Linguistics, Jun. 2019, pp. 4430–4441. [Online]. Available: http://arxiv.org/abs/1906.05807

[27] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation,” in 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

CONVOLUTIONAL NEURAL NETWORK FOR OFFLINE SIGNATURE VERIFICATION VIA MULTIPLE CLASSIFIERS

Fadi Mohammad Alsuhimat and Fatma Susilawati Mohamad

Faculty of Informatics and Computing,
Universiti Sultan ZainalAbidin, Terengganu, Malaysia

ABSTRACT

The signature process is one of the most significant processes used by organizations to preserve the security of information and protect it from unwanted penetration or access. As organizations and individuals move into the digital environment, there is an essential need for a computerized system able to distinguish between genuine and forged signatures in order to protect people's authorization and decide what permissions they have. In this paper, we used Pre-Trained CNN for extracts features from genuine and forged signatures, and three widely used classification algorithms, SVM (Support Vector Machine), NB (Naive Bayes) and KNN (k-nearest neighbors), these algorithms are compared to calculate the run time, classification error, classification loss and accuracy for test-set consist of signature images (genuine and forgery). Three classifiers have been applied using (UTSig) dataset; where run time, classification error, classification loss and accuracy were calculated for each classifier in the verification phase, the results showed that the SVM and KNN got the best accuracy (76.21), while the SVM got the best run time (0.13) result among other classifiers, therefore the SVM classifier got the best result among the other classifiers in terms of our measures.

KEYWORDS

CNN, Signature verification, SVM, KNN, NB.

1. INTRODUCTION

A handwritten signature considered as a personal skill which consists a group of symbol and characters written in a specific language, the signature is one of the operations that use to provide persons with authentication to perform many transactions, such as banking transactions and classes attendance, where the signature can ensure the permitted validity of persons and classify the forged signature from the genuine signature [1].

A signature is sketched out as an extraordinarily composed drawing that an individual composes on any record as a sign of character. A person employments it on a normal wish to sign a check, a legitimate instrument, contract, etc. The matter emerges when once some person tries to duplicate its [2].

Signature verification may be a complex design recognizable proof with inadequacy as no two veritable signatures of a person can be absolutely comparative. In case inadvertently it is winning at that point it'll do genuine damage to an individual. One of the ways is to utilize the biometric features of each person [3].

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 107-116, 2021. CS & IT - CSCP 2021 DOI:10.5121/csit.2021.112008

Nowadays signature discovery and other biometric features are playing a fundamental part in nearly all the field, where mystery and security are the most concerns for all people and nations. Moreover, utilizing signature verification can offer assistance to decide the personality of people and their authorization to do a particular work [2].

A signature recognition system could be a way to confirm the signature in order to distinguish any imitation, sometime recently getting the ultimate result from verification stage, the recognition prepare comprises of a set of stages, incorporate normalization, features extraction, and classification, these three phases are exceptionally imperative to confirm signature since the transcribed signature can shift each time depending on the conduct and position of the person. [3]. Figure 1 shows different types of signatures for the same person.

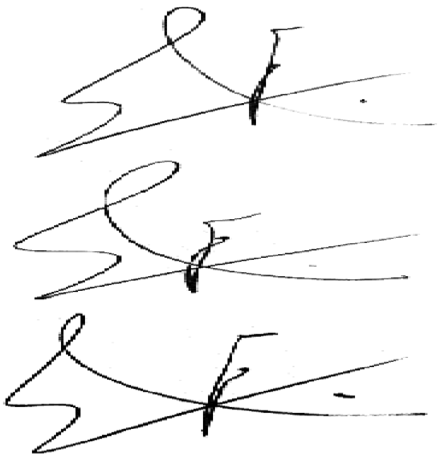


Figure 1. Example of different patterns of signature

The second stage in signature recognition system is features extraction stage, this phase considers a significant phase in signature recognition system because the whole system depends on it in order to verify individuals signatures, where this phase responsible about detecting and determine a group of features in each signature, including number of pixels, width, corner, and length [4].

The features extraction stage depends on detect image highlights with incredible precision through minimizing the measurements of the first picture at that point extricate a group of covered up characteristics within the picture, in arrange to encourage the method of separation between unique and fake marks.

The third stage in the signature recognition system is the classification stage, and this stage is the signature verification stage, in which it is determined whether the signature is false or real in it, through comparing the signature features stored in the database with anyone who wants to verify his/her signature [5].

The classification phase aims at identifying the genuine signature by comparing the enrolled and authenticated signature features. The decision-maker then chooses if the signature should be accepted or denied based on the threshold [6].

Furthermore, the signature is a character trait of individuals used in biometrics systems to verify individuals' identities, as the usage of biometric characteristics in the field of security grows, the signature appears as a biometric feature that provides a secure way of delegating individuals and verifying their identification in legal documents. Furthermore, when compared to other biometric traits like (hand geometry, iris scan, or DNA), the signature has a high level of acceptance by individuals. All these reasons have led to an increase in the proliferation of signature recognition systems and the need for further developments on these systems.

In this paper, our objective is to study the features extraction phase and classification phase for signature images. Therefore, in this research Pre-trained Convolutional Neural Network was used for features extraction phase, then signature image features are classify using (support vector machine (SVM), naive Bayes (NB) and k-nearest neighbor (KNN)), with UTSig dataset [7]. This dataset has (115) classes containing: (27) genuine signatures; (3) opposite-hand signed samples, (36) simple forgeries and (6) skill forgeries; we selected (2475) images as a training group to train the classification algorithms.

2. OVERVIEW OF METHODS

In this section, the features extraction technique and classification algorithms that are used for signature classification and comparison process are described briefly. The suggested signature classification algorithm consists of feature normalization, feature extraction and classification.

2.1. Features Extraction Phase

In this research, a deep learning method was used for offline signature verification. A Convolutional Neural Network (CNN) ad hoc model was used as a deep learning method. A Convolutional Neural Network was firstly proposed by LeCun et al [8] as a method for image processing, where it has consisted of two essential features including spatial pooling and spatially shared weights.

In 1998, they [9] enhanced the CNNs as LeNet-5 which is a pioneering 7-level convolutional network in order to classify digits. At the present time, CNNs considered the most widely utilized deep learning architecture in feature learning, through many successful applications in various areas like autonomous vehicles[10], character recognition [11], video processing [12], medical image processing and object recognition [13].

Figure (2) shows basic structure of CNN.

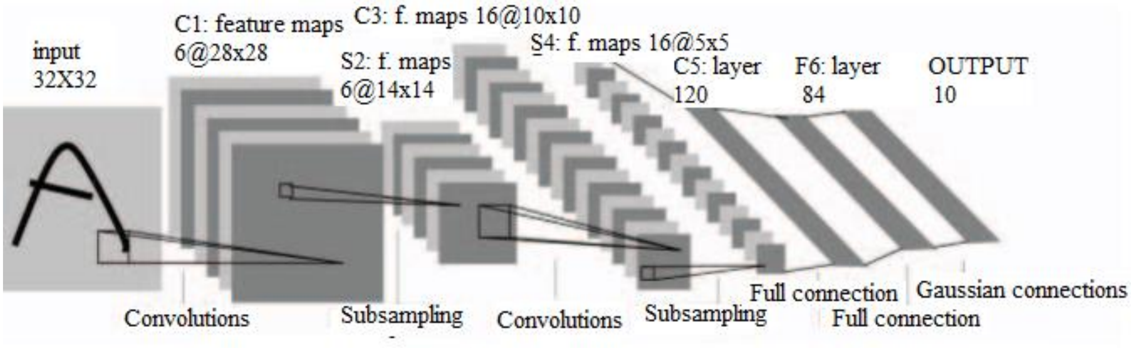


Figure 2. CNN structure

As shown in Figure (2), a CNN has three primary layers: a convolutional layer, a subsampling layer (pooling layer), and a fully-connected layer, that was taken from the study of LeCun et al

[8]. CNN points to define the unique features of pictures utilizing convolutional operations and pooling operations. The features gotten within the first layers identify as edges or colour data, whereas within the final layers they portray parts of shapes and objects [9].

In the convolution layer, the convolution operation is implemented by shifting the filter data matrix on the input data matrix and adding a bias to the multiplication of these matrixes. Basic convolution process represents in Figure. 3, Basic formulation of the convolution operation has been given in equation (1). In the equation, pixels of the output image, pixels of the input image, pixels of the filter (kernel) and bias term were represented by y, x, w and b respectively.

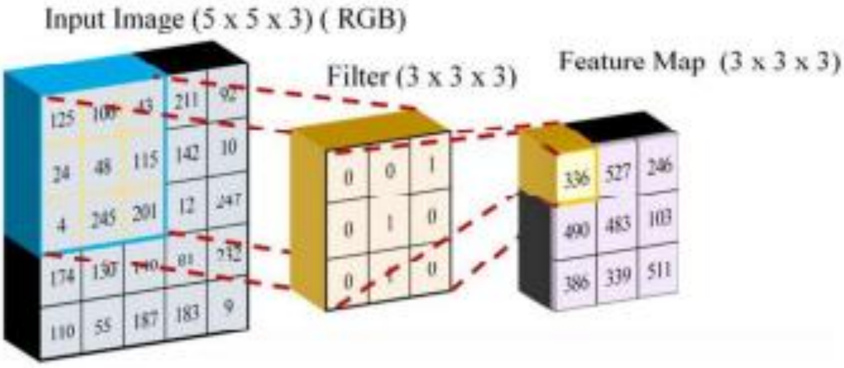


Figure 3. Basic convolution operation

$$y_n = \sum_{n=1}^{9} (x_n . w_n + b_0) \qquad (1)$$

Another tool using by CNNs is called pooling, the pooling tool [58] is utilized to spatially down-sample the activation of the previous layer by propagating the maximum activation of the previous neuron groups. The most objective of the pooling layers is diminishing the computational complexity of the model by continuously diminishing the dimensionality of the representation [9]. If preferred, a rectified linear unit (*ReLU*) activation function can be utilized at the conclusion of each layer for normalization. The main operation of (*ReLU*) was depicted in equation (2).

$$\text{ReLU(x)} = f(x) = \begin{cases} 0 & if\ x < 0 \\ x & if\ x \geq 0 \end{cases} \qquad (2)$$

Fully Connected Layers (FC), which are the primary building components of classical neural networks, are the final layer in CNN. Fully Connected layers are shaped by the association of neurons to each neuron within the following layer. It is at that point normalized to a probability dispersion employing a Soft-Max layer. Moreover, it points to require the high-level sifted pictures and interpret them into votes. These votes are communicated as weights, or association qualities, between each esteem and each category [9], [11].

2.2. Signature Classification

In this paper, we used various algorithms for classification: KNN, SVM, and SVR.

K-nearest Neighbor (KNN): This is a procedure of gathering parameters based on closest tests of the range of inner features [14]. KNN is one of the popular and clear classification calculations.

Learning approach as it joined sparing characteristic vectors and marks of the learning pictures, inner gathering operations.

This unmarked position may be really assigned the title for its *k* closest neighbor's. Regularly, this thing will be categorized based on the marks of its *k* closest neighbor's by utilizing overwhelming portion surveying. On *k*=1, those parameters are categorized based on the power of the parameter closest to it. If there is a need for only two segments, then k should make an odd number. *K* may be an odd number when showing up multiclass arrangement. This stage used the famous distance equation, Euclidean distance, as a related point separation capacity for KNN after changing each image to a vector from claiming fixed-length for true numbers:

$$d(x,y) = (\sum_{i=1}^{m}((x_i - y_i)^2))^{1/2} \qquad (3)$$

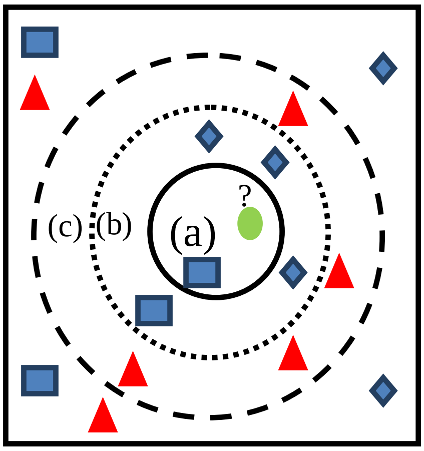


Figure 4. KNN Classification

Support Vector Machine (SVM): This is prepared to assess signature among specific signature qualities [15]. Through applying a classification algorithm to particular features for signature images, during the training procedure, we trained a signature classifier, used every last one of the preparation data. An outline of signature prediction utilized SVM algorithm indicated in Fig. 5 to classify the input signature image with training procedures. The inputs xi is the characteristic vectors. To configure the SVM parameters, we used Gaussian kernel *K*:

$$f(x) = \sum_{i=1}^{N_s} a_i y_i K(s_i, x) + b \qquad (4)$$

$$K(x_i, x_j) = e^{\frac{1}{2\sigma^2}|x_i - x_j|^2}$$

Naive Bayes: Naive Bayes learning refers to the construction of a Bayesian probabilistic model that assigns a posterior class probability to an instance*: P(Y = yj |X = xi)*. The simple naive Bayes classifier uses these probabilities to assign an instance to a class. Applying Bayes' theorem (Eq. 7) [16], and simplifying the notation a little, as shown in equation 5.

$$P(y_i|x_i) = \frac{P(x_i|y_i)P(y_i)}{P(x_i)} \qquad (5)$$

3. EXPERIMENTAL RESULT

This section shows the results of our classifiers, through three mean sections, section (3.1) describes the database which was used. Section (3.2) shows the receiver operating characteristic (ROC) and run-time for each classifier, while section (3.3) indicates the performance of each classifier by calculating (accuracy, classification error, classification loss, and run-time).

3.1. Database

The process of comparing three algorithms implemented on a set of signature images from the (UTSig) dataset. As illustrated in Figure (5), this dataset has "(115) classes containing: (27) authentic signatures; (3) opposite-hand forgeries, (36) easy forgeries, and (6) skill forgeries." Each lesson is assigned to a single actual person. UTSig contains (8280) photos taken from undergraduate and graduate students at the University of Tehran and Sharif University of Technology, where signatures images were scanned at 600 dpi and saved as 8-bit Tiff files" [7, p1].

In this paper, a total of (2475) signature images were chosen to train the set, and (660) signature images were chosen to test our classification algorithms.

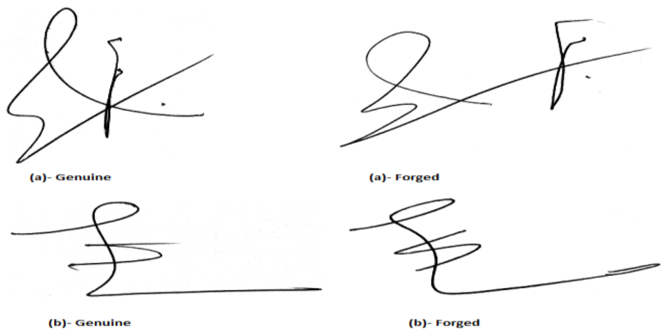


Figure 5. Forger and Genuine signature examples from UTSig dataset.

3.2. Experimental setup

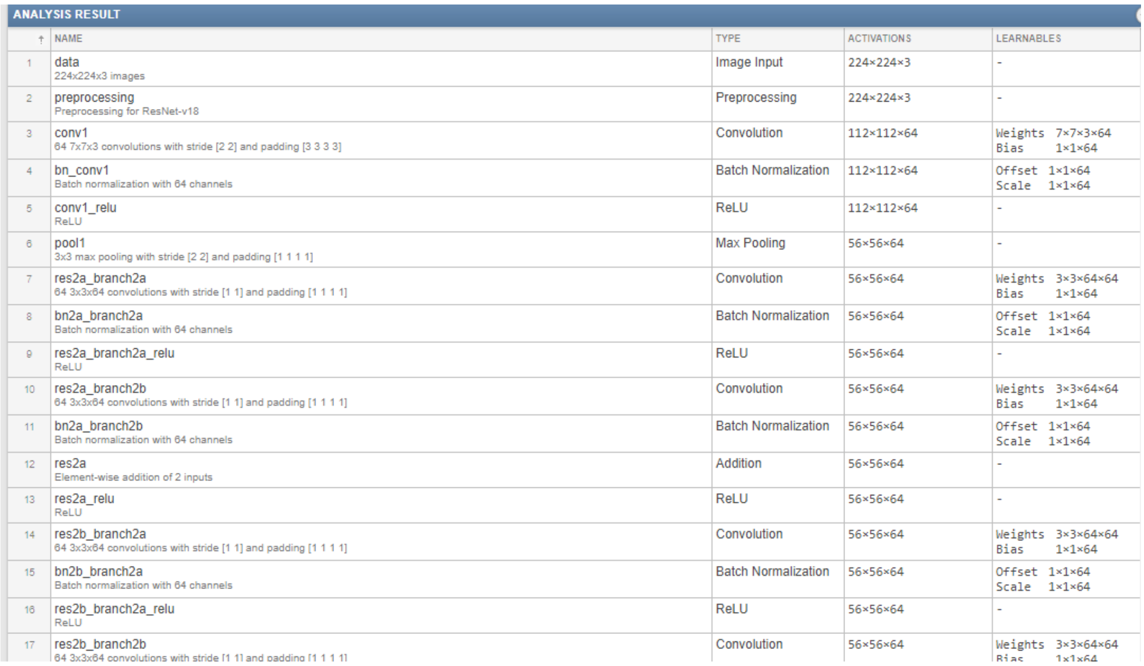
Features were extracted from a pre-trained CNN and then classified in original-forgeries through three classifiers, SVM, KNN and NB. In the first model, CNN was trained via a set of signatures for (75) persons, where each person has 33 signatures which include 27 genuine and 6 forgeries were used, the pre-trained CNN used AlexNet for features extraction process, where AlexNet uses layers property that comprises of 25 layers. There are 8 layers for learnable weights, 5 convolutional layers and 3 fully connected layers. Fig. 5 shows the details of all the layers of AlexNet.

Table I shows the experimental results using (ROC) by calculating the area under the curve for the estimated values of *X* and *Y*. Also, calculate the run-time for each classifier. We discovered that KNN performed better than other classifier algorithms, which include SVM and NB according to ROC values, where the NB classifier run-time was better than other classifier algorithms.

Table 1. Run-Time and AUC values for each classifier

Method	Run-Time	AUC
SVM	70.1	0.998
KNN	1.89	0.999
NB	1.52	0.782

Figure. 6 showed the ROC values for each classifier, where KNN produces better ROC values for higher thresholds, SVM is also got good ROD values and almost equal to KNN values. While the ROC curve for naive Bayes is often lowers than the other two ROC curves, this suggests that the other two classifier algorithms perform better in-sample.



ANALYSIS RESULT

↑	NAME	TYPE	ACTIVATIONS	LEARNABLES
1	data 224x224x3 images	Image Input	224×224×3	-
2	preprocessing Preprocessing for ResNet-v18	Preprocessing	224×224×3	-
3	conv1 64 7x7x3 convolutions with stride [2 2] and padding [3 3 3 3]	Convolution	112×112×64	Weights 7×7×3×64 Bias 1×1×64
4	bn_conv1 Batch normalization with 64 channels	Batch Normalization	112×112×64	Offset 1×1×64 Scale 1×1×64
5	conv1_relu ReLU	ReLU	112×112×64	-
6	pool1 3x3 max pooling with stride [2 2] and padding [1 1 1 1]	Max Pooling	56×56×64	-
7	res2a_branch2a 64 3x3x64 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	56×56×64	Weights 3×3×64×64 Bias 1×1×64
8	bn2a_branch2a Batch normalization with 64 channels	Batch Normalization	56×56×64	Offset 1×1×64 Scale 1×1×64
9	res2a_branch2a_relu ReLU	ReLU	56×56×64	-
10	res2a_branch2b 64 3x3x64 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	56×56×64	Weights 3×3×64×64 Bias 1×1×64
11	bn2a_branch2b Batch normalization with 64 channels	Batch Normalization	56×56×64	Offset 1×1×64 Scale 1×1×64
12	res2a Element-wise addition of 2 inputs	Addition	56×56×64	-
13	res2a_relu ReLU	ReLU	56×56×64	-
14	res2b_branch2a 64 3x3x64 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	56×56×64	Weights 3×3×64×64 Bias 1×1×64
15	bn2b_branch2a Batch normalization with 64 channels	Batch Normalization	56×56×64	Offset 1×1×64 Scale 1×1×64
16	res2b_branch2a_relu ReLU	ReLU	56×56×64	-
17	res2b_branch2b 64 3x3x64 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	56×56×64	Weights 3×3×64×64 Bias 1×1×64

Figure 5. shows the details of all the layers of AlexNet

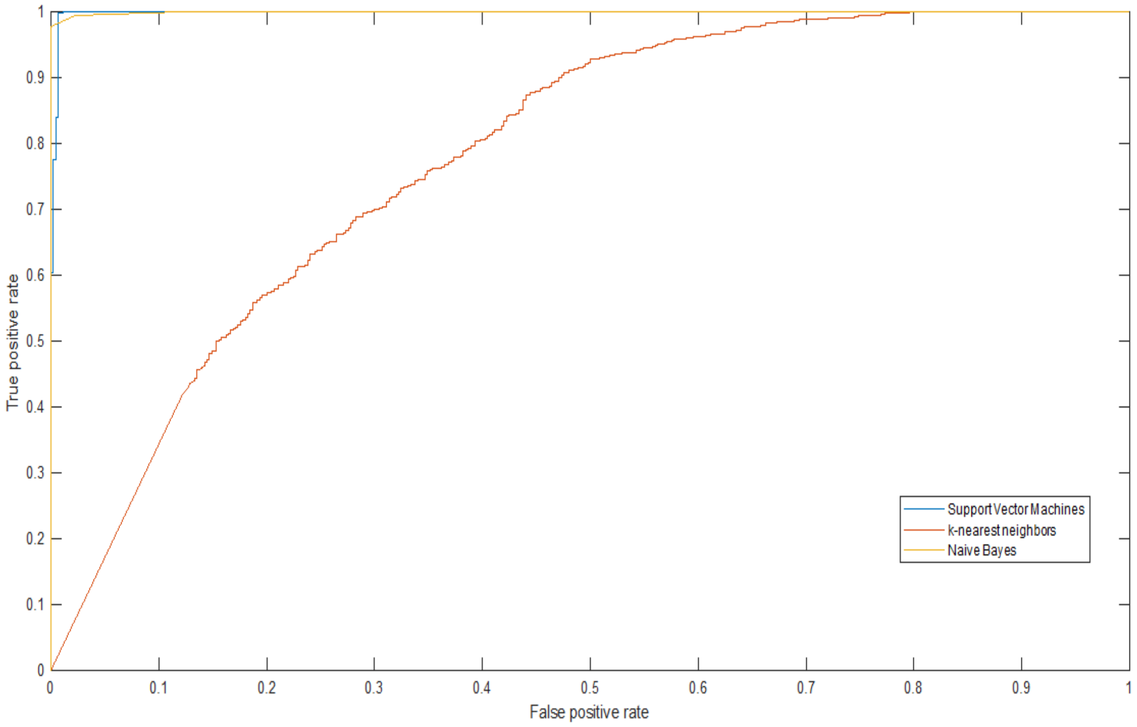


Figure 6. Receiver operating characteristic (ROC) curve

3.3. Efficiency

The efficiency was taken regarding run time, classification error, classification loss and accuracy measurements for each classifier on 660 images sequentially.

Table 2. shows all measures for each classification algorithms

Measures	Methods		
	SVM	KNN	NB
Run-Time	0.13	0.77	0.18
Classification Error	0.24	0.24	0.29
Classification Loss	0.01	0.01	0.26
Accuracy	76.21	76.21	71.36

Data in the above table showed that, for the run time we can note that the best run time was for SVM classifier. Following by NB classifier, and finally KNN classifier, while for classification error we note that, SVM and KNN misclassifies approximately (24%) of the test sample, while NB misclassifies approximately (29%) of the test sample. Besides that, classification loss values indicated that, SVM and KNN classifiers have better value (0.01) than NB classifier (0.26), finally the accuracy value for both classifier SVM and KNN achieved (76.21) which better than the accuracy value for NB classifier.

4. Conclusion and Future Work

In this research, the SVM, KNN, and NB classification algorithms were compared on a set of signature images from the (UTISG) dataset to assess performance by calculating the run time, classification error, classification loss, and accuracy metrics for each algorithm. The three methods described here are popular classification algorithms, with computing complexity and accuracy being the most important factors in selecting a better classification technique.

The comparison process is done between the train set consist of (2475) signature images through pre-trained CNN for features extraction, then the result trained using three classifiers SVM, KNN and NB. After that the run time, classification error, classification loss and accuracy measurements calculated for each algorithm in order to find the best classification algorithm. The experimental results showed that, the best run time was for SVM classifier, following by NB classifier, and finally KNN classifier, while for classification error SVM and KNN got same misclassifies approximately and better than NB misclassifies approximately. In addition, SVM and KNN classifiers have same classification loss values and better than NB classifier, finally the accuracy value for both classifier SVM and KNN was same and better than the accuracy value for NB classifier.

For future work other classification algorithms will be test with the same and different dataset, also using full deep learning system for both phases (extract features and classification) will help in build an accurate signature verification system.

References

[1] F. Alsuhimat, F. S. Mohamad, and M. Iqtait, “Detection and Extraction Features for Signatures Images via Different Techniques,” IOP Conf. Series: Journal of Physics: Conf. Series 1179 (2019) 012087.

[2] F. S. Mohamad, F. Alsuhimat, M. Mohamed, M. Mohamad, and A. Jamal, “Detection and Feature Extraction for Images Signatures,” International Journal of Engineering & Technology, vol. 7, no. 3, pp. 44-48, 2018.

[3] J. Poddar, V. Parikh, S. K. Bharti, “Offline signature Recognition and Forgery Detection using Deep Learning,” The 3rd International Conference on Emerging Data and Industry 4.0 (ED140), April 6-9, Warsw, Poland, 2020.

[4] K. Daqrouq, H. Sweidan, A. Balamesh, and M. Ajour, "Off-Line Handwritten Signature Recognition by Wavelet Entropy and Neural Network", Entropy., vol. 19, no. 6, pp. 1.20, 2017.

[5] T. Jahan., S. Anwar, and A. Al-Mamun, "A Study on Preprocessing and Feature Extraction in offline Handwritten Signatures", Global Journal of Computer Science and Technology: F Graphics & Vision., vol. 15, no. 2, pp. 1.7, 2015.

[6] S. Gunjal, B. Dange, and A. Brahmane, "Offline Signature Verification using Feature Point Extraction", International Journal of Computer Applications., vol. 141, no. 14, pp. 6.12, 2016.

[7] Soleimani, K. Fouladi, and B. Araabi, "UTSig: A Persian offline signature dataset", IET Biometrics., vol. 6, no. 1, pp. 1.8, 2016.

[8] S. Singh, M. Gogate, and S. Jagdale, “Signature Verification Using LDP & LBP with SVM Classifiers,” International Journal of Scientific Engineering and Science, vol. 1, no. 11, pp. 95-98, 2017.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ Images classification with deep convolutional neural networks,” In Advance in neural information processing systems, pp. 1097-1105, 2012.

[10] Y. LeCun et al.,”Handwritten digit recognition with a back-propagation network,” in Advances in neural information processing systems, pp. 396-404, 1990.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffiner, ‘”Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[12] M. F. Yahya and M. R. Arshad, “Detection of markers using deep learning for docking of autonomous underwater vehicle,” in 2017 IEEE 2nd International Conference on Automatic Control

and Intelligent Systems (I2CACIS), pp. 179 184, 2017.
[13] C. Boufenar, A. Kerboua, and M. Batouche, “Investigation on deep learning for off-line handwritten Arabic character recognition,” Cogn. Syst. Res., 2017.

[14] R. Olmos, S. Tabik, and F. Herrera, “Automatic handgun detection alarm in videos using deep learning,” Neurocomputing, vol. 275, pp. 66-72, 2018.

[15] V. D. Nguyen, H. Van Nguyen, D. T. Tran, S. J. Lee, and J. W. Jeon, “Learning Framework for Robust Obstacle Detection, Recognition, and Tracking,” IEEE Trans IntellTranspSyst, vol. 18, no. 6, pp. 1633-1646, 2017.

[16] T. Cover, and P. Hart, “Nearest neighbor pattern classification,” IEEE transactions on information theory, vol. 13, no. 1, pp. 21-27, 1967.

[17] C. J. Burges, “ A tutorial on support vector machines for pattern recognition,” Data mining and knowledge discovery, vol. 2, no. 2, pp. 121-167, 1998.

[18] D. Berrar, “Bayes theorem and naive Bayes classifier,” Encyclopedia of Bioinformatics and Computational Biology, vol. 1, pp. 403-412, 2018.

AUTHORS

Received the B.S. degree in computer information system from Alhussien Bin Talal University, Ma’an, Jordan, in 2007, the M.S. degree in computer science from Utara University Malaysia (UUM), Kedah, Malaysia, and now Ph.D. student in pattern recognition, deep learning at University Sultan ZainalAbidin, Kuala Terengganu (UNISZA), interesting in machine and deep learning, data science and artificial intelligence.

B.Sc degree in information system management from Oklahoma, USA, master degree in computer science from University Kebangsaan Malaysia, and Ph.D in computer science from University Teknologi Malaysia, now work as Associate Professor at Faculty of Informatics and Computing, University Sultan ZainalAbidin, Kuala Terengganu, Malaysia. Current research on Statistical and Biometric Pattern Recognition.

LOW-COMPLEXITY RECEIVER FOR MASSIVE MIMO-GFDM COMMUNICATIONS

Feng-Cheng Tsai, Fang-Biau Ueng and Ding-Ching Lin

Department of Electrical Engineering,
National Chung Hsing University, Taiwan

ABSTRACT

OFDM has two disadvantages. The first is high peak-to-average power ratio (PAPR), and the second is high out-of-band (OOB) radiated power. In the future communication applications, the diversified scenarios such as Internet of Things, inter-machine communication and telemedicine make the fourth-generation mobile communication no longer applicable. The generalized frequency division multiplexing (GFDM) has a pulse-shaping filter, which has less out-of-band radiated power and peak-to-average power ratio and fewer cyclic prefixes (CP) than OFDM. In order to meet high- data-transmission rate, it is an inevitable trend to install massive multi-input multi-output (massive MIMO) antennas. As the number of antennas increases, so does its complexity. This paper employs time reversal (TR) technology to reduce the computational complexity. Although the number of base station (BS) antennas has increased to eliminate interference, there is still residual interference. In order to eliminate the interference one step further, we deploy a zero forcing equalization (ZF equalization) after the time reversal combination.

KEYWORDS

5G, GFDM, MIMO.

1. INTRODUCTION

The fifth generation (5G) of mobile communications is already developed [1][2], and standards have also been formulated at international conferences. In order to be applied to the Internet of Things (IOT) and Wireless Regional Area Network (WRAN), the 5G system uses high-level technology like massive MIMO [3], beamforming [4] and millimeter wave communication, so that 5G has the advantages of big data transmission rate, low time delay, low power consumption and so on. The development of device-to-device proximity service and machine type communication (MTC) [5] makes OFDM face the challenges of future 5G application scenarios. MTC requires very low power consumption, which makes OFDM's orthogonal subcarriers unbearable; tactile Internet requires short-burst data with low time delay, but OFDM adds long overhead in front of each OFDM symbol. The length of the cyclic prefix (CP) exhibits disappointing spectral efficiency in the performance of the spectrum. The more mobile devices that can connect to the Internet, it means that more spectrum is needed, but our spectrum resources are getting less and less. Because OFDM uses rectangular pulse filters for transmission, its sidelobes are large, resulting in OFDM modulation systems that are sensitive to frequency deviations, high OOB radiated power [6] and high PAPR [7] and other shortcomings, so the application in wireless communication technology is severely restricted.

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 117-126, 2021. CS & IT - CSCP 2021 DOI:10.5121/csit.2021.112009

At present, the known literature proposes several multi-carrier technologies as candidates for 5G communication standards, such as filter bank multicarrier (FBMC) [8], universal filtered multicarrier (UFMC) [9]. The sub-carriers of the FBMC system are all individually pulsed filters, and the sub-carriers have a narrower bandwidth, so the transmission filter has a longer impulse response length. Usually, the length of the filter is four times the signal to reduce OOB emission, and good spectrum efficiency is obtained, but the effect of low time delay cannot be achieved. The UFMC system filters their respective sub-carriers through their respective filters to reduce OOB emissions. Because the bandwidth of filter covers many sub-carriers and the impulse response is very short, high spectral efficiency (SE) can be achieved in transmission. UFMC does not use CP, and the symbol time misalignment in a short period of time is more sensitive than OFDM. The technology used in this paper is GFDM [10]. This technology has a flexible multi-carrier modulation scheme and can be adjusted according to the application of different scenarios [11]. GFDM is a multicarrier modulation technology that uses non-rectangular pulse filters. It uses cyclic convolution to realize the DFT filter bank structure in the frequency domain. GFDM uses less CP [12], which improves the spectrum efficiency to a certain extent. Because GFDM uses non-rectangular filters, it can avoid the problems faced by rectangular pulse filters, namely high PAPR and OOB. There is a special feature in the 5G applications, that is massive MIMO system. However, as the number of antennas increases, the anti-interference ability will saturate at a certain deterministic SINR value, so that the interference will no longer decrease. The computational complexity will also increase as the antenna increases, and its performance will be relatively poor. Therefore, in the 5G large-scale multiple-input multiple-output system, due to the increase in the number of antennas and the increase in the modulation order, designing a technology with computational efficiency is a key challenge. Massive MIMO [13][14] system is an extension of the concept of MIMO system. Usually the number of base station antennas is about 100 or more than 100 [15]. Because each channel is independent, the channels of different users will gradually show orthogonality due to the increase in the number of antennas. This method can eliminate multiple user interference (MUI) [16], and then increase the system capacity. Since the OFDM [17] system faces the problems of high OOB and high PAPR, GFDM can change its sub-carrier waveform according to the selection of different filters and rolling factors, to reduce the OOB radiation power and the PAPR [16]. The multipath effect [18] will make the receiving antennas receive the same signal copy generated by multiple paths. When the multipath delay time is too long, it will cause Inter Symbol Interference (ISI) [19]. In this paper, since we use a large-scale MIMO system, the increase in antennas leads to an increase in computational complexity, so we propose a time reversal [20] method to reduce computational complexity. Time Reversal technology is a basic physical phenomenon that uses the inevitable but abundant multipath radio propagation environment to produce space-time resonance effects, the so-called focusing effect [21]. When the bandwidth is larger, the time resolution is better, and therefore, more multipaths can be displayed. Time reversal technology can use a single antenna to achieve a large-scale MIMO-like effect [22]. By using a large number of virtual antennas, a single-antenna time reversal system can achieve excellent focusing effects in the time and space domains, thereby obtaining the promising performance of massive MIMO systems. In addition, since the time reversal system uses the environment as a virtual antenna array and computing resources, its implementation complexity is much lower. Unlike conventional technologies that use multipath propagation environment, if time reversal technology can use a large enough bandwidth, it does not need to deploy complex receivers or a large number of antennas to take full advantage of multipath propagation [23].

2. SYSTEM MODEL

We assume that there are K users on the transmitting end and M base station antennas. The transmitter uses the GFDM system, which can use less cyclic prefix (CP) than that of the OFDM system, and has a lower OOB and PAPR compared with the OFDM system. The data of the k-th user can be expressed as (1),

$$u_k = [u_k(0), \ldots, u_k(P-1)]^T \mathrm{k} \in \{1,2, \ldots K\} \tag{1}$$

Among them, u represents the data stream, and the subscript k represents the k-th user, which contains sampling elements $\mathrm{p} \in \{0,1, \ldots, P-1\}$. Assume that there are K users, each user uses a single-antenna transmission system, and then through symbol mapping according to (1), and then GFDM modulation to GFDM symbol s_k, as in (2),

$$s_k = [s_k(0),\ldots,s_k(P-1)]\mathrm{k} \in \{1,2, \ldots K\} \tag{2}$$

Next, a cyclic prefix (CP) is added, and the length of the cyclic prefix is usually set to be greater than the length of the multipath channel delay to avoid inter-symbol interference (ISI) caused by multipath. Compared with OFDM, the cyclic prefix length used by GFDM is short to achieve the effect of preventing inter-symbol interference. s_k represents the GFDM symbol obtained by the k-th user's transmission signal through the GFDM modulator, which contains sampling elements $\mathrm{p} \in \{0,1,2, \ldots, P-1\}$.$u_k$ will first enter S/P, pass through the pulse, then pass through the subcarrier filter and finally perform frequency domain offset to obtain s_k. X represents sub-carrier, Y represents sub-symbol, g[n] is the subcarrier filter, e^0 represents the (frequency domain) offset. The mathematical formula of the GFDM symbol generated is expressed as follows,

$$s_k[p] = \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} g_{x,y}[p] u_k[p] \tag{3}$$

Among them, u_k is the data symbol, x is the sub-carrier index, y is the sub-symbol index, and p is the sampling index. Here the impulse filter $g_{x,y}[p]$ used by the kth user is a prototype filter after a time and frequency shift version, and the filter used in this paper is a raised cosine filter (RC filter). Among them, the mathematical expression of the raised cosine filter is (4),

$$g_{x,y}[p] = g[(p-mX) \; mod \; P] \cdot e^{-j2\pi \frac{x}{X} p} \tag{4}$$

After (3), the transmission symbols are collected and can be expressed as a matrix. We will write all $g_{x,y}[p]$ as a matrix after collecting and sorting.

3. The Proposed Detector

In this paper, we consider a large-scale multi-input multi-output system with multiple users. We consider that there are K users and M BS antenna arrays, and each user is with single antenna. The signal received by the m-th antenna is expressed as follows,

$$r_m = \sum_{k=0}^{K-1} s_k * h_{m,k} + v_m \tag{9}$$

where $s_k = [s_k(0),\ldots,s_k(P-1)]$ represents the transmitted signal of the k-th user, r_m represents the received signal of the m-th antenna, and v_m represents the complex additive white Gaussian noise (AWGN). The sequence $h_{m,k} = [h_{m,k}(0), \ldots, h_{m,k}(L-1)]$ represents the channel impulse response (CIR) from the k-th user to the m-th BS receiving antenna, where we assume that the channel is a perfect known channel, the multipath channel tap length is L, and the channels from the user to the BS antenna are independent of each other. Then we write the signal received at the m-th BS antenna as a matrix form after the signal passes through the GFDM modulator, which is expressed as follows,

$$\begin{aligned}\bar{r}_m^i &= \sum_{\mathrm{k}=0}^{\mathrm{K}-1}\left(\mathrm{BH}_{\mathrm{m,k}}^{(\mathrm{i,i}-1)} s_k^{i-1} + \mathrm{BH}_{\mathrm{m,k}}^{(\mathrm{i,i})} s_k^i\right) + \mathrm{B}v_m^i \\ &= \sum_{\mathrm{k}=0}^{\mathrm{K}-1}\left(\mathrm{BH}_{\mathrm{m,k}}^{(\mathrm{i,i}-1)}\mathrm{A}u_k^{i-1} + \mathrm{BH}_{\mathrm{m,k}}^{(\mathrm{i,i})}\mathrm{A}u_k^i\right) + \mathrm{B}v_m^i \\ &= \sum_{k=0}^{K-1}\left(\bar{H}_{m,k}^{(i,i-1)} u_k^{i-1} + \bar{H}_{m,k}^{(i,i)} u_k^i\right) + \bar{v}_m^i \quad (10)\end{aligned}$$

where $\mathrm{H}_{\mathrm{m,k}}^{(\mathrm{i,i}-1)}$ and $\mathrm{H}_{\mathrm{m,k}}^{(\mathrm{i,i})}$ are $N\times N$ convolution matrices, s_k^{i-1} and s_k^i respectively represent the tail of symbol time $i-1$ and the head of symbol time i. Matrix $\bar{H}_{m,k}^{(i,i-1)} \triangleq \mathrm{BH}_{\mathrm{m,k}}^{(\mathrm{i,i}-1)}\mathrm{A}and\bar{H}_{m,k}^{(i,i)} \triangleq \mathrm{BH}_{\mathrm{m,k}}^{(\mathrm{i,i})}\mathrm{A}$ are the inter-symbol interference matrix and the inter-carrier interference matrix, respectively, where $B = A^{-1}$. We assume that W_p is a combination matrix of $M\times K$, and the number of subcarriers is $p = 0, \dots P-1$. Generally, there are three common traditional linear combination methods, which are maximum ratio combining (MRC) [24][25], zero-forcing (ZF)[26], and minimum mean square error detection (MMSE).

- Maximum Ratio Combination (MRC):

$$W_p = \bar{\mathrm{H}}_p D_p^{-1} (11)$$

- Zero Forcing (ZF):

$$W_p = \bar{\mathrm{H}}_p\left(\bar{H}_p^H \bar{\mathrm{H}}_p\right)^{-1} (12)$$

- Minimum mean square error (MMSE):

$$W_p = \bar{\mathrm{H}}_p\left(\bar{H}_p^H \bar{\mathrm{H}}_p + \sigma^2 I_K\right)^{-1} \quad (13)$$

Here, in order to find various interference items in the structure of the large-scale antenna, this paper considers $W_p = \frac{1}{M}\,\bar{\mathrm{H}}_P$. Bring the received signal (10) into equation (11), and after the output of the combiner, the vector of the detection signal is obtained as follows,

$$\widehat{u^i}(p) = W_p^H \bar{r}^i(p) \qquad (14)$$

where $\bar{r}^i(p) = [\bar{r}_0^i(p), \dots, \bar{r}_{M-1}^i(p)]^T$ is the received signal vector of $M\times 1$, $\hat{u}^i(p) = \left[\hat{u}_0^i(p), \dots, \hat{u}_{K-1}^i(p)\right]^T$ is the $K\times 1$ detection signal vector. According to (10) and (11), the detection signal $\hat{u}^i(p)$ can be expressed as follows,

$$\begin{aligned}\hat{u}^i(p) = H_{kk,pp}^{(i,i)} u_k^i(p) + \sum_{\substack{q=0 \\ q\neq p}}^{N-1} H_{kk,pq}^{(i,i)} u_k^i(q) + \sum_{q}^{N-1} H_{kk,pq}^{(i,i-1)} u_k^{i-1}(q) \\ + \sum_{\substack{j=0 \\ j\neq k}}^{K-1}\sum_{q=0}^{N-1}\left(H_{kj,pq}^{(i,i-1)} u_k^{i-1}(q) + H_{kj,pq}^{(i,i)} u_k^i(q)\right) + \bar{v}_k^i(p) (15)\end{aligned}$$

which includes the inter-symbol interference coefficient, inter-carrier interference coefficient, multi-user interference coefficient and expected coefficient. Because we use a large-scale multiple-input multiple-output system, we can calculate the above coefficients in the form of the law of large numbers. In a large-scale multi-input multi-output system, we can calculate the convergence value by the law of large numbers, so here we propose the multi-user coefficient and explain how to use the law of large numbers to calculate the convergence value. According to (15), the multi-user term can be expanded into the following formula:

$$H_{kj,pq}^{(i,i)} = \frac{\bar{h}_k^H(p)}{M}\left[\left[\mathrm{H}_{0,j}^{(i,i)}\right]_{pq}, \dots, \left[\mathrm{H}_{M-1,j}^{(i,i)}\right]_{pq}\right]^T \quad (16)$$

$$H_{kj,pq}^{(i,i-1)} = \frac{\bar{h}_k^H(p)}{M}\left[\left[\mathrm{H}_{0,j}^{(i,i-1)}\right]_{pq}, \dots, \left[\mathrm{H}_{M-1,j}^{(i,i-1)}\right]_{pq}\right]^T \quad (17)$$

where $\bar{h}_k(p) = \left[\bar{h}_{0,k}(p), \dots, \bar{h}_{M-1,k}(p)\right]$ is an $M \times 1$ vector, $[\bar{\mathrm{H}}_P]_{m,k} \triangleq \bar{h}_{m,k}(p)$. Before that, let's review the definition of probability. Let $\mathrm{a}=[\mathrm{a}_1, \dots, \mathrm{a}_n]^T$ and $\mathrm{b}=[\mathrm{b}_1, \dots, \mathrm{b}_n]^T$ be two random vectors, and have mutually independent and identically distributed elements. We assume that the i-th element of a and b has $\mathbb{E}\{a_i * b_i\} = C_{ab}, \ \ i = 1, \dots, n$. Then according to the law of large numbers, when n approaches infinity, the sampling average $\frac{1}{n}a^H b$ will converge to a distribution average C_{ab}, that is to say $\frac{1}{n}a^H b \to C_{ab}$, $as \ \ n \to \infty$. When M approaches infinity, (16) and (17) use the form of the law of large numbers into (18) and (19),

$$H_{kj,pq}^{(i,i)} \to \mathbb{E}\left\{\bar{h}_{m,k}^*(p)\left[\bar{\mathrm{H}}_{m,j}^{(i,i)}\right]_{pq}\right\} (18)$$

$$H_{kj,pq}^{(i,i-1)} \to \mathbb{E}\left\{\bar{h}_{m,k}^*(p)\left[\bar{\mathrm{H}}_{m,j}^{(i,i-1)}\right]_{pq}\right\} (19)$$

Since k ≠ j, $\bar{h}_{m,k}(p)$ will be the same as $\left[\bar{\mathrm{H}}_{m,j}^{(i,i)}\right]_{pq}$ and $\left[\bar{\mathrm{H}}_{m,j}^{(i,i-1)}\right]_{pq}$ presents an irrelevant state, so when M approaches infinity, the multi-user coefficients $H_{kj,pq}^{(i,i)}$and $H_{kj,pq}^{(i,i-1)}$ will approach zero. The coefficients of other terms can be proved as follows after derivation,

$$\left[\bar{H}_{m,k}^{(i,i-1)}\right]_{qp} = \frac{1}{N}\sum\nolimits_{n=0}^{N-1}\sum\nolimits_{l=0}^{L-1} h_{m,k}(l) e^{j\frac{2\pi}{N}(nq-lq-np)} \varpi(n-l+N) \ (20)$$

$$\left[\bar{H}_{m,k}^{(i,i)}\right]_{qp} = \frac{1}{N}\sum\nolimits_{n=0}^{N-1}\sum\nolimits_{l=0}^{L-1} h_{m,k}(l) e^{j\frac{2\pi}{N}(nq-lq-np)} \varpi(n-l)(21)$$

$\varpi(n)$ is the window function, and the square function is considered here, which is expressed as follows,

$$\varpi(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ & 0, \ \text{other} \end{cases} \quad (22)$$

When p ≠ q, we can get the following calculation process:

$$H_{kk,pp}^{(i,i)} \to \mathrm{E}\left\{\bar{h}_{m,k}^*\left[\bar{H}_{m,k}^{(i,i)}\right]_{pp}\right\} = \frac{1}{N}\sum\nolimits_{l=0}^{L-1}(N-l)\rho(l) = 1 - \frac{\tau}{N} \quad (23)$$

$$H_{kk,qp}^{(i,i)} \to \mathrm{E}\left\{\bar{h}_{m,k}^{*}\left[\bar{H}_{m,k}^{(i,i)}\right]_{qp}\right\}$$
$$= \frac{1}{N}\mathrm{E}\left\{\sum\nolimits_{n}^{N-1}\sum\nolimits_{l=0}^{L-1}\sum\nolimits_{l'}^{L-1} h_{m,k}^{*}(l)h_{m,k}^{*}(l') \times e^{j\frac{2\pi}{N}(qn-lp+l'p-pn)}\varpi(n-l)\right\}$$
$$= \frac{1-\bar{\rho}(q-p)}{N(1-e^{j\frac{2\pi(q-p)}{N}})} \qquad (24)$$

The normalized channel power delay profile is considered in this paper, that is, $\sum_{l=0}^{L-1}\rho(l) = 1$. $\bar{\rho}(q) \triangleq \sum_{l=0}^{L-1}\rho(l)e^{-j\frac{2\pi lq}{N}}$, $\tau \triangleq \sum_{l=0}^{L-1}\rho(l)l$. The method of deriving the values of other ISI coefficients is the same as the above derivation method, where their values are $H_{kj,pq}^{(i,i-1)} \to \frac{\bar{\rho}(q-p)-1}{N(1-e^{j\frac{2\pi(q-p)}{N}})}andH_{kj,pp}^{(i,i-1)} \to \frac{\tau}{N}$.

Continuing the above concept, we can get the following equation by putting the received signal into the time-reversed channel impulse response,

$$r_k{}^{TR} = \frac{1}{\sqrt{M}}\sum_{m=0}^{M-1} r_m * \beta_{m,k} \quad (25)$$

Among them, $r_k{}^{TR}$represents the received signal of the impulse response of k-th users through the time reversal channel, $\beta_{m,k} = \left[h_{m,k}^{*}(-L+1), \dots, h_{m,k}^{*}(0)\right]$ means that on the k-th user, the channel impulse response between the received signal and the corresponding base station antenna is time-reversed to take the conjugate form. Expand the above formula to obtain the following formula,

$$r_k{}^{TR} = \sum_{j=0}^{K-1} s_j * c_{kj} + v_k{}^{TR} \qquad (26)$$

wheres_j represents the j-th GFDM transmission signal, $v_k{}^{TR}$ represents the k-th AWGN, and c_{kj} is the equivalent impulse response of the original channel and the corresponding time reversal conjugate channel,

$$c_{kj} \triangleq \tfrac{1}{\sqrt{M}}\textstyle\sum_{m=0}^{M-1} h_{m,j} * \beta_{m,k} \qquad (27)$$

$$v_k{}^{TR} \triangleq \tfrac{1}{\sqrt{M}}\textstyle\sum_{m=0}^{M-1} h_{m,k}^{*} * v_m (28)$$

when k ≠ j, c_{kj} presents the crosstalk channel impulse response between terminal k and terminal j; when k = j, $c_{kj} = c_{kk}$ is the time reversal equivalent impulse response of the corresponding channel for user k.

Let $r_k^{i^{TR}} = \left[r_k^{i^{TR}}(iP), \dots, r_k^{i^{TR}}(iP+P-1)\right]^T$ be a P×1 vector, $r_k^{i^{TR}}$ contains the time-reversal received signal of the i-th time portion, and then expressed in matrix form, the following formula can be obtained,

$$r_k^{i^{TR}} = \sum_{j=0}^{K-1} \left(C_{kj}^{(i,i-1)} s_j^{i-1} + C_{kj}^{(i,i)} s_j^{i} + C_{kj}^{(i,i+1)} s_j^{i+1} \right) + v_k^{i^{TR}} (29)$$

where$C_{kj}^{(i,i-1)}$, $C_{kj}^{(i,i)}$and $C_{kj}^{(i,i+1)}$ are the $P\times P$ convolution matrices, which can be expressed as follows,

$$\mathrm{C}_{\mathrm{kj}}^{(\mathrm{i,i-1})} = \mathcal{T}_{\mathrm{P\times P}}\left(\left[\mathrm{c_{kj}}(1), \ldots, \mathrm{c_{kj}}(L-1), 0_{1\times 2\mathrm{P-L}}\right]^{\mathrm{T}}\right)(30)$$

$$\mathrm{C}_{\mathrm{kj}}^{(\mathrm{i,i})} = \mathcal{T}_{\mathrm{P\times P}}\left(\left[0_{1\times \mathrm{P-L}}, \mathrm{c_{kj}}, 0_{1\times \mathrm{P-L}}\right]^{\mathrm{T}}\right)(31)$$

$$\mathrm{C}_{\mathrm{kj}}^{(\mathrm{i,i+1})} = \mathcal{T}_{\mathrm{P\times P}}\left(\left[0_{1\times 2\mathrm{P-L}}, \mathrm{c_{kj}}(1-\mathrm{L}), \ldots, \mathrm{c_{kj}}(-1)\right]^{\mathrm{T}}\right)(32)$$

Among them, they are formed by the way of the Teplitz matrix, which respectively represent the inter-symbol interference matrix and the inter-carrier interference matrix. $\mathrm{c_{kj}} \triangleq \left[\mathrm{c_{kj}}(1-\mathrm{L}), \ldots \mathrm{c_{kj}}(\mathrm{L}-1)\right]^{\mathrm{T}}$ contains the sampling elements of the time reversal channel impulse response $\mathrm{c_{kj}}$. According to the previous concept, it can be written in another form:

$$\bar{\mathrm{r}}_k^{\mathrm{i^{TR}}} = \sum_{\mathrm{j=0}}^{\mathrm{K-1}} \left(\bar{\mathrm{C}}_{\mathrm{kj}}^{(\mathrm{i,i-1})} \mathrm{u_j^{i-1}} + \bar{\mathrm{C}}_{\mathrm{kj}}^{(\mathrm{i,i})} \mathrm{u_j^{i}} + \bar{\mathrm{C}}_{\mathrm{kj}}^{(\mathrm{i,i+1})} \mathrm{u_j^{i+1}} \right) + \bar{\mathrm{v}}_{\mathrm{k}}^{\mathrm{i^{TR}}} (33)$$

4. SIMULATION RESULTS

This section compares the performances of large-scale MIMO GFDM system using the traditional equalizers in the previous paper and the TR-ZF method proposed in this paper. We compare the rate performance of each method. The following are the system parameters used in the simulation of this paper.

.

Table 1. Simulation parameters

Modulation Format	**4QAM,16QAM**
Users(K)	10, 35
Number of Receive Antennas(M)	100, 200
Subcarrier(X)	128
Sub-symbol(Y)	5
Pulse Shaping Filter (g)	RC filter
Roll-Off Factor (a)	0.1
GFDM Demodulator	ZF
Channel Delay(L)	20, 40
Channel	Rayleigh Fading channel

Figure 1 compares the traditional equalizer, the reference paper ZF-FFT, and the TRZF multi-user GFDM system we proposed. The number of users is 10, the number of base station antennas is 100, the channel delay length is 20, and the cyclic prefix length is 20. It can be seen from the figure that when the CP is sufficient, the error rate of the proposed scheme is better than that of the existing ZF and ZF-FFT. Figure 2 compares the traditional equalizer, the reference paper ZF-FFT, and the TRZF multi-user GFDM system we proposed. The number of users is 10, the number of base station antennas is 100, the channel delay length is 40, and the cyclic prefix length is 20. From the figure, it is found that when the channel delay length is greater than the cyclic prefix length, the traditional ZF and ZF-FFT have poor performance due to increased interference caused by the multipath effect. Figure 3 compares the traditional equalizer, the reference paper ZF-FFT and the TRZF multi-user GFDM system we proposed. 16-QAM is employed, the number of users is 10, the number of base station antennas is 100, the channel delay length is 20, and the cyclic prefix length is 20. When the CP is sufficient, the error rate performance of the proposed scheme is better than that of the traditional ZF and ZF-FFT.

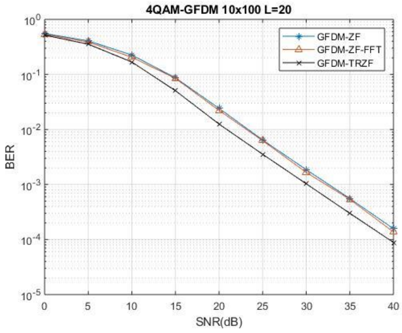


Figure 1.

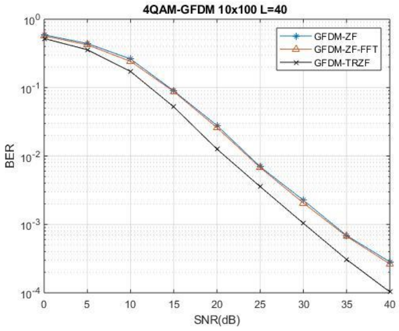


Figure 2.

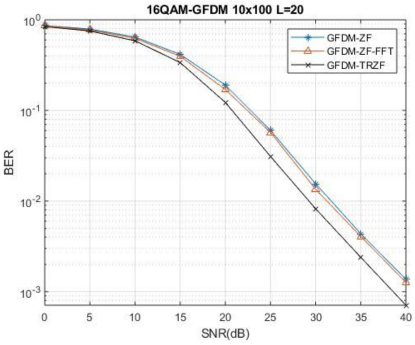


Figure 3.

5. Conclusions

In a large-scale multiple-input multiple-output system, as the number of antennas increase, channel capacity can be increased and irrelevant noise and interference can be eliminated.

However, an unlimited number of base station antennas will increase computational complexity and SNR can no longer be improved. In this paper, we employ time reversal technology to reduce complexity and improve SNR. When the channel delay length is greater than the cyclic prefix length, compared with the traditional MIMO-GFDM system, the proposed TRZF will not increase the error rate and also reduces the computational complexity.

REFERENCES

[1] G. Wunder et al., "5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications," in IEEE Communications Magazine, vol. 52, no. 2, pp. 97-105, February 2014.

[2] G. Fettweis and S. Alamouti, "5G: Personal mobile internet beyond what cellular did to telephony," in IEEE Communications Magazine, vol. 52, no. 2, pp. 140-145, February 2014.

[3] H. Q. Ngo, E. G. Larsson and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems", *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436-1449, Apr. 2013.

[4] U. Hamid, R. A. Qamar and K. Waqas, "Performance comparison of time-domain and frequency-domain beamforming techniques for sensor array processing," Proceedings of IBCAST, pp. 379-385, 2014.

[5] Y. Medjahdi et al., "On the Road to 5G: Comparative Study of Physical Layer in MTC Context," in IEEE Access, vol. 5, pp. 26556-26581, 2017.

[6] J. Van De Beek and F. Berggren, "Out-of-Band Power Suppression in OFDM," in IEEE Communications Letters, vol. 12, no. 9, pp. 609-611, September 2008.

[7] Y. Xiao, Q. Liang, X. He and S. Li, "On the joint reduction of peak-to-average power ratio and out-of-band power in OFDM systems," in Proc.WCSP, Suzhou, 2010, pp. 1-4.

[8] A. Jayapalan, P. Savarinathan, P. Mahalingam, A. Dev N.S. and A. Natraj S., "Analysis of Filter Bank Multi carrier Modulation," in Proc. ICCCI, Coimbatore, 2020, pp. 1-4.

[9] H. Wang and Y. Huang, "Performance evaluation of the universal filtered multi-carrier communications under various multipath fading propagation conditions," in Proc. IEEE iCAST, Taichung, 2017, pp. 466-469.

[10] G. Fettweis, M. Krondorf and S. Bittner, "GFDM- Generalized Frequency Division Multiplexing," in Proc. IEEE VTC Spring, Barcelona, 2009, pp. 1-4.

[11] N. Michailow, I. Gaspar, S. Krone, M. Lentmaier and G. Fettweis, "Generalized frequency division multiplexing: Analysis of an alternative multi-carrier technique for next generation cellular systems," in Proc. ISWCS, 2012.

[12] B. Farhang-Boroujeny and H. Moradi, "Derivation of GFDM based on OFDM principles," in Proc. IEEE ICC, London, 2015, pp. 2680-2685.

[13] A. J. PAULRAJ, D. A. GORE, R. U. NABAR and H. BOLCSKEI, "An overview of MIMO communications - a key to gigabit wireless," in Proceedings of the IEEE, vol. 92, no. 2, pp. 198-218, Feb. 2004.

[14] Wei Hong et al., "Development of the MIMO system for future mobile communications," IEEE/ACES International Conference on Wireless Communications and Applied Computational Electromagnetics, HI, 2005, pp. 634-637.

[15] K. Yamazaki et al., "Field Experimental DL MU-MIMO Evaluations of Low-SHF-Band C-RAN Massive MIMO System with over 100 Antenna Elements for 5G," in Proc. IEEE VTC-Fall, Chicago, IL, USA, 2018, pp. 1-5.

[16] D. A. Feryando, T. Suryani and Endroyono, "Performance analysis of regularized channel inversion precoding in multiuser MIMO-GFDM downlink systems," in Proc. IEEE APWiMob, Bandung, 2017, pp. 101-105.

[17] Y. S. Cho, J. Kim, W. Y. Yang, C. G. Kang, "Introduction to OFDM," in MIMO-OFDM Wireless Communications with MATLAB® , IEEE, 2010, pp.111-151.

[18] M. Cheffena, "Industrial indoor multipath propagation-A physical-statistical approach," in Proc. IEEEPIMRC, Washington, DC, 2014, pp. 68-72.

[19] H. Suganuma, S. Saito, T. Maruko and F. Maehara, "Inter-symbol interference suppression scheme employing periodic signals in coded network MIMO-OFDM systems," in Proc. IEEE Radio and Wireless Symposium (RWS), Phoenix, AZ, 2017, pp. 42-44.

[20] Y. Jin, J. M. F. Moura, N. O'Donoughue and J. Harley, "Single antenna time reversal detection of moving target," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 3558-3561.

[21] G. F. Edelmann, H. C. Song, S. Kim, W. S. Hodgkiss, W. A. Kuperman and T. Akal, "Underwater acoustic communications using time reversal", *IEEE J. Ocean. Eng.*, vol. 30, no. 4, pp. 852-864, Oct. 2005.

[22] Y. Han, Y. Chen, B. Wang and K. J. R. Liu, "Time-reversal massive multipath effect: A single-antenna 'massive MIMO' solution", *IEEE Trans. Commun.*, vol. 64, no. 8, pp. 3382-3394, Aug. 2016.

[23] Y. Chen, B. Wang, Y. Han, H. Lai, Z. Safar and K. J. R. Liu, "Why Time Reversal for Future 5G Wireless? [Perspectives]," in IEEE Signal Processing Magazine, vol. 33, no. 2, pp. 17-26, March 2016.

[24] Y. A. Al-Zahrani, N. K. Al-Mutairi, Y. Al-Hodhayf and A. Al-Shahrani, "Performance of antenna selection for maximum ratio combining MIMO system," 2011 IEEE 3rd International Conference on Communication Software and Networks, Xi'an, 2011, pp. 642-645.

[25] H. Lee, J. An, J. Seo and J. Chung, "Multipath Selection Method for Maximum Ratio Combining in Underwater Acoustic Channels," 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN), Prague, 2018, pp. 519-521.

[26] H. Wang and R. Song, "Low Complexity ZF Receiver Design for Multi-User GFDMA Uplink Systems," in IEEE Access, vol. 6, pp. 28661-28667, 2018.

[27] W. D. Dias, L. L. Mendes and J. J. P. C. Rodrigues, "Low Complexity GFDM Receiver for Frequency-Selective Channels," in IEEE Communications Letters, vol. 23, no. 7, pp. 1166-1169, July 2019..

THERMAL ENTROPY BASED HESITANT FUZZY LINGUISTIC TERM SET ANALYSIS IN ENERGY EFFICIENT OPPORTUNISTIC CLUSTERING

Junaid Anees and Hao-Chun Zhang

School of Energy Science & Engineering,
Harbin Institute of Technology, Harbin, China

ABSTRACT

Limited energy resources and sensor nodes' adaptability with the surrounding environment play a significant role in the sustainable Wireless Sensor Networks. This paper proposes a novel, dynamic, self-organizing opportunistic clustering using Hesitant Fuzzy Linguistic Term Analysis- based Multi-Criteria Decision Modeling methodology in order to overcome the CH decision making problems and network lifetime bottlenecks. The asynchronous sleep/awake cycle strategy could be exploited to make an opportunistic connection between sensor nodes using opportunistic connection random graph. Every node in the network observe the node gain degree, energy welfare, relative thermal entropy, link connectivity, expected optimal hop, link quality factor etc. to form the criteria for Hesitant Fuzzy Linguistic Term Set. It makes the node to evaluate its current state and make the decision about the required action ('CH', 'CM' or 'relay'). The simulation results reveal that our proposed scheme leads to an improvement in network lifetime, packet delivery ratio and overall energy consumption against existing benchmarks.

KEYWORDS

Graph Theory, Wireless Sensor Networks, Hesitant Fuzzy Linguistic Term Set, Opportunistic Routing and RF Energy Transfer.

1. INTRODUCTION

Wireless Sensor Networks (WSNs) should operate for a long time for a specific application like Smart Home, environmental monitoring, disaster management, forest fires, precision agriculture, surveillance system traffic monitoring etc. Power-constrained WSNs have to adjust their sleep/awake cycle according to the application requirements in order to maximize the network lifetime and overall energy consumption [1]. Sectional failure and thermal exposure can cause significant damage to sensor nodes. Moreover, different units of a sensor node behave differently when exposed in sunlight for long period of time for example; the performance of a typical transceiver is degraded with the increase in temperature. The purpose of deploying WSNs is to achieve a shared goal through sensor collaboration and data aggregation. In order to allocate the resources to sensor nodes effectively, topology architecture is needed in which sensors are organized in clusters [1]. The multi-hop routing in this clustering topology can result in the decrease of overall energy consumption and interference among sensor nodes due to specific timeslots allocation [2]. In addition to it, it could also effectively optimize the data redundancy by significantly reducing the collected data size using data aggregation techniques at Cluster Head (CH) level [1-2].

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 127-142, 2021. CS & IT - CSCP 2021 DOI:10.5121/csit.2021.112010

Researchers have proposed different node scheduling techniques to save battery power of sensor nodes i.e. synchronous and asynchronous sleep/awake scheduling. Asynchronous sleep/awake scheduling is designed to prolong the network lifetime and improve energy utilization by creating an opportunistic node connection between sensor nodes in the network [3-6]. A very popular technique to ensure sustainable operation of sensor nodes is Opportunistic Routing (OR) which is a paradigm in WSNs that benefit from broadcasting characteristics of a wireless medium by selecting multiple sensor nodes as candidate forwarders. This set of sensor nodes is called a Candidate Set (CS). The performance of OR significantly depends on several key factors, such as the OR metric, the candidate selection algorithm, and the candidate coordination method. Based on the asynchronous sleep/awake scheduling in OR, a node can sense, process, and transmit/receive during its active state [3-4]. Conversely, a node enters its sleeping state for an interval predefined or calculated according to contemporary environmental conditions. Researchers have also worked on temperature adaptive sleep/awake scheduling techniques [7-8].

The information entropy utilizes probability distribution function (pdf) to statistically measure the degree of uncertainty [9]. The entropy H(X) of a random variable $X = \{x1, x2, \dots xn\}$ having probability distribution as p(X) can be given as $H(x) = \sum_{x \in X} p(x) log_2 p(x)$ for 0 ≤ H(X) ≤ 1 [9-10]. It should be kept in focus that CH or BS should not be hesitant or irresolute about any of their decisions regarding cluster formation and data fusion. Keeping in view OR and temperature adaptive sleep/awake scheduling, we have selected multiple parameters including time frequency parameter, node's gain energy, relative thermal entropy, expected optimal hops, link quality factor in terms of signal-to-noise ratio, as our attributes of hesitant fuzzy linguistic term set. These attributes are used to assess the role of nodes and self adaptively make the appropriate decision in a round of operation. No concept of duty cycle is used in our proposed scheme. Furthermore, our proposed scheme FLOC uses this information in a Multi-Attribute Decision Modelling (MADM) framework to efficiently utilize our hesitant fuzzy linguistic term set to incorporate a qualitative assessment of the parameters by a node and help the node observe a situation adaptive role transition. The rest of the paper is organized as follows: Section 2 contains the discussion on some related works. System modelling is presented in Section 3. Our proposed scheme FLOC is presented in Section 4. HFLTS analysis is provided in Section 5. Section 6 presents the simulation framework and performance evaluation of the proposed technique. Finally, section 7 concludes the paper with some targeted future works.

2. RELATED WORK

Various researchers have focused on proposing different routing protocols for WSNs based on different parameters such as end–end delay, packet delivery ratio, network lifetime, overall energy consumption, control packet overhead, and sink node mobility, etc. Ogundile et al. [1] presented a detailed survey for energy efficient and energy balanced routing protocols for WSNs including the taxonomy of cluster-based routing protocols for WSNs. Routing protocols in WSNs can be segmented into two main categories, i.e., hierarchical and non-hierarchical routing protocols. Non-hierarchical routing protocols are designed in accordance with overhearing, flooding, and sink node position advertisement, whereas hierarchical routing protocols are designed on the basis of grid, tree, cluster and area [1-3][12]. Different hierarchical routing protocols have their own merits and demerits, but as far as cluster-based hierarchical routing protocols are concerned, researchers have been challenged with a task of achieving an optimal balance between end–end delay and energy consumption [9-13].

Yang et al. [5] introduced the utilization of sleep/awake cycle of sensor nodes to prolong the network lifetime. The sleep/awake cycle can be segmented into two categories—synchronous and asynchronous sleep/awake cycle. In this paper, our focus is only towards asynchronous sleep/awake scheduling. Depending on the network connectivity requirements in terms of traffic coverage, Mukherjee et al. [15] proposed an asynchronous sleep/awake scheduling technique with a minimum number of sensor nodes to achieve the required network coverage. As a result of asynchronous sleep/awake scheduling, opportunistic node connections are established between sensor nodes and their neighbours, which brings the need of Opportunistic Connection Random Graph (OCRG) theory to properly model the opportunistic node connections by forming a spanning tree. Anees et al. [6] proposed an energy-efficient multi-disjoint path opportunistic node connection routing protocol for smart grids (SGs) neighbourhood area networks (NAN). Anees et al. [13] also proposed a delay aware energy-efficient opportunistic node selection in restricted routing for delay sensitive applications. In this protocol, the information related to updated position of sink is advertised by multiple ring nodes and data is forwarded to mobile sink using ring nodes having maximum residual energy.

In a few asynchronous sleep/awake scheduling techniques, the sensor nodes are found to remain in active listening mode for a long amount of time, resulting in unnecessary consumption of energy. A popular WSN MAC protocol, Sensor Medium Access Control (SMAC), has been proposed by Ye et al. [16]. SMAC protocol lets the node listen for a fixed interval of time and turn their radio off (sleep state) for a fixed duration. Barkley-MAC (BMAC) [17] provides an adaptive preamble sampling technique to effectively reduce the duty cycle and idle listening by the sensor nodes. They are required to wake up periodically to check for ongoing communications. Shah et al. [18] devised a guaranteed lifetime protocol in which the sink node assigns sleep/awake periods for other nodes depending on residual energy, sleep duration, and coverage by the nodes. A mathematical model for temperature adaptive sleep/awake strategy is developed by Bachir et al. [8] with three proposed algorithms i.e. Stop Operate (SO), a Power control (PC), and Stop-Operate-Power-Control (SOPC). The sensor nodes running any of the algorithms are supposed to observe the contemporary state based on a pre-calculated relationship between node-density and temperature. Thermal entropy of the sensor nodes has been explored in the intelligent sleep-scheduling technique iSleep [19]. Reinforcement Learning based sleep-scheduling algorithm RL-Sleep has been proposed in [7] in which the authors have used a temperature model and Q-learning technique to switch the sleep/awake states adaptively, depending on the environmental situation.

It has been revealed through a detailed literature review that most of the clustering schemes consider energy efficiency, traffic distribution, or coverage-efficiency as the prime criteria for state-scheduling and decision modelling of sensor nodes instead of relative thermal entropy, temperature adaptability or hesitant fuzziness used for nodes' role transition etc. A few entropy based clustering schemes have been proposed in which entropy weight coefficient method is adopted for decision making in cluster-based hierarchical routing protocol [20-21].Multi-Criteria Decision Analysis (MCDA) and Multi-Attribute Sensors Decision Modelling (MADM) using entropy weight coefficients are also types of entropy weight-based multi-criteria decision routing [21]. Anees et al. [9] proposed hesitant fuzzy entropy based opportunistic clustering and data fusion algorithm for heterogeneous WSNs. In this algorithm, the local sensory data is gathered from sensor nodes by utilizing hesitant fuzzy entropy based multi-attribute decision modelling for cluster head election procedure. Zhai et al. [10] developed Hesitant Language Preference Relationships (HLPR) to improve the credibility of WSNs by fusing uncertain information and putting forward exact opinion about different WSN schemes.

Varshney [22] proposed an emerging concept of simultaneous wireless information and power transfer (SWIPT) in which both energy and data are transferred over RF links simultaneously. Guo et al. [23] utilized the concept the SWIPT to extend the network lifetime of a clustered WSN by wirelessly charging the relay nodes which are responsible to share data with BS. Zhou et al. [24] proposed dynamic power splitting (DPS) to adjust the power ratio of information encoding and energy harvesting in EHWSNs. Anees et al. [25] proposed harvested energy scavenging and transfer capabilities in opportunistic ring routing in which a distinguishing approach of hybrid (ring + cluster) topology is used in a virtual ring structure and then a two-tier routing topology is used in the virtual ring as an overlay by grouping nodes into clusters. Overall, to the best of our knowledge, there is no published literature which focuses on thermal entropy based HFLTS analysis for energy efficient opportunistic clustering. In this paper, we have considered a set of attributes that regulate the nodes' decisions about its role transition conducive to the current situation in a cluster and provided a detailed solution for optimally handling problems in energy efficient opportunistic clustering using relative thermal entropy based HFLTS analysis.

3. SYSTEM MODELING

3.1. Network model

A MxM network area denoted as A is considered for FLOC in which N sensor nodes are deployed randomly and independently. We have assumed that sensor nodes follow a uniform distribution. The node-density of the network is denoted as $\lambda_0 = \frac{N}{A}$. All sensor nodes use short radio range (RS) for sensing and transmission purposes whereas sink node can use RS for transmission & reception and long radio range (RL) for data collection tasks using a tag message. However, all sensor nodes can exploit the power control function and communicate with different neighbouring nodes within various power levels. A probe message is shared by each sensor node to acquire the neighbour information as discussed in [6]. Each sensor node is equipped with a power splitting radio, which is composed of a signal processing unit to transfer energy to or from neighbours using RF link. Moreover, it is also assumed that every sensor node is aware of its position using the energy-efficient localization method [26-29].

Each sensor node is characterized by a set of k attributes named as $C = \{c_1, c_2, \dots .c_k\}$ and a set of weights $w_t = \{w_{t1}, w_{t2}, \dots .w_{tp}\}$ is assigned by sensor node to the p criteria of C. Furthermore, the sensor node undergoes y states i.e. $ST = \{ST_1, ST_2, \dots .ST_y\}$, where$ST_1$ represents the favorable state (attribute values are above threshold) and ST_y represents the stressed state (attribute values below threshold. Depending on multiple parameters, the sensor node decides about the most suitable action against the contemporary state i.e. $AC = \{CH, CM, Relay\}$. Hesitant Fuzzy Sets have been used in our proposed scheme which enables the sensor node to decide about the optimal action after assessing the respective conditions.

3.2. Energy Model

The energy consumption model [6] for radio energy dissipation during transmission and reception is considered in which the energy required to transmit l bits of data over distance d can be given in (1) as:

$$E_{Tx}(V_i, V_j) = \begin{cases} E_{elec}l + \varepsilon_{fs} l d_{V_iV_j}^2 & d < d_0 \\ E_{elec}l + \varepsilon_{mp} l d_{V_iV_j}^4 & d \geq d_0 \end{cases} \quad (1)$$

where E_{elec} is the energy spent by transmitter on running the radio electronics, ε_{fs} is the free space energy dissipated by power amplifier depending on the Euclidean distance $d_{V_iV_j}$ between the transmitter and receiver, ε_{mp} is the muti-path fading factor for energy dissipated by power amplifier depending on Euclidean distance $d_{V_iV_j}$between transmitter and receiver. The threshold distance $\mathrm{d_o}$is given as $\mathrm{d_o} = \sqrt{\varepsilon_{\mathrm{fs}}/\varepsilon_{\mathrm{mp}}}$. Likewise, the energy required to receive l bits of data over distance d is given in (2) as:

$$E_{Rx} = E_{elec}l \qquad (2)$$

The energy used for sensing l bits of data in the virtual ring at the beginning of each round can be given as $\mathrm{E_{sense}} = \mathrm{E_{elec}l}$. Accordingly, the total energy consumed by cluster member (CM) can be computed in (3) as:

$$E_{CM} = E_{sense} + E_{Tx} = E_{elec}l + E_{elec}l + \varepsilon_{fs}ld^2_{V_iV_j}(3)$$

Each CH is responsible for data gathering, aggregating the received data and then relaying that data towards sink, so the total energy consumed by a CH can be computed in (4) - (5) as

$$E_{CH} = E_{sense} + \left(\frac{N}{N_C} - 1\right)E_{Rx} + \left(\frac{N}{N_C}\right)lE_A + \left(\frac{N}{r}\right)E_{Tx} \qquad (4)$$

$$E_{CH} = E_{elec}l + \left(\frac{N}{N_C} - 1\right)E_{elec}l + \left(\frac{N}{N_C}\right)l\frac{E_{elec}}{R_{CC}} + \left(\frac{N}{r}\right)E_{elec}l + \left(\frac{N}{r}\right)\varepsilon_{mp}ld^4_{V_iV_j}(5)$$

where $\mathrm{N_C}$ represents the number of clusters in the network, $\frac{\mathrm{N}}{\mathrm{N_C}}$is the number of working sensor nodes per cluster in which we have 1 CH and $\frac{\mathrm{N}}{\mathrm{N_C}} - 1$ CMs. $\mathrm{E_A}$ signifies the data aggregating energy at CH level, r represents the compression ratio and $\mathrm{R_{CC}}$ symbolizes the communication to computation ratio.

4. PROPOSED SCHEME FLOC

4.1. Ambient temperature and Relative thermal entropy

In this section, the proposed scheme FLOC is discussed in detail. As the hesitant fuzzy linguistic term set analysis is based on MADM, we need to consider several parameters like ambient temperature, asynchronous sleep/awake cycle, relative thermal entropy, gain degree, expected optimal hops and link quality factor as attributes of hesitant fuzzy set. Keeping in view the diurnal temperature variation caused by solar radiation, the sensor nodes placed under direct sunlight absorb higher heat energy than the sensor nodes in shadow. According to temperature model in [7], the temperature of a sensor node i after solar heat absorption for amount of time Δt can be represented in (6) as,

$$T^i_{t+\Delta t} = max\left\{T^i_t + \frac{(S_{SUN}(t)\alpha(t) - \eta T^4)}{c_p\theta}Area_{sen}\Delta t, T^i_t\right\} \qquad (6)$$

where T^i_t is the temperature of a node i at time t, $S_{SUN}(t)$ denotes the amount of radiation by the sun at that time, $\alpha(t)$ is the temporal variation of sun exposure, $Area_{sen}$ is the exposed area through which the sensor node absorbs solar heat, η is Boltzman constant, θ represents the mass of the sensor node, c_p represents the specific heat and $T^i_{t+\Delta t}$ symbolizes the ambient temperature. The change in temperature of a sensor node can be extracted from equation (7) i.e.

$$\Delta T_i = \frac{(S_{SUN}(t)\alpha(t) - \eta T^4)}{c_p\theta}Area_{sen}\Delta t \qquad (7)$$

The resultant temperature T_i of a sensor node i is given as $T_i = T^i_{t+\Delta t} = T + \varDelta T_i$. Foregoing in view, the solar radiation pattern for a day can be represented as $S_{SUN}(t) = S^{max}_{SUN} exp^{\frac{-(t-\rho)^2}{2\sigma^2}}$, $0 \leq t \leq 2\rho$ where $\mathrm{S}^{\mathrm{max}}_{\mathrm{SUN}}$ is the peak value of the solar radiation during the day [7]. It has been found that the identical sensor nodes behave differently to the temperature variations due to solar radiation exposure, traffic flow and relative position of the sensor node [8]. Let $\mathrm{T_i}$ be the temperature of the ith node and T_H be the highest temperature for which the ith node becomes non-operational. The probability of failure of a sensor node due to temperature increase can be represented as $p_i = \frac{T_i}{T_H}$, where T_i can be acquired from equation (6)& (7) and T_H symbolizes the highest temperature the sensor node can withstand. Here we have assumed that T_H is the same for all sensor nodes in the network. The cumulative effect of failure likelihood leads to network instability; therefore we need a probability distribution function (PDF) to measure the degree of uncertainty in the sensor network. It should be kept in focus that any sensor node due to failure likelihood should not be hesitant or irresolute about any of the operating mode. This hesitancy or irresolution resembles entropy.

The entropy H(X) of a random variable $X = \{x_1, x_2, \ldots x_n\}$ having probability distribution as p(X) can be given as $H(X) = -\sum_{x \in X} p(x) log_2\, p(x)$ for $0 \leq \mathrm{H(X)} \leq 1$ [9]. Similarly, the Shannon's entropy at ith node can be defined as $H(p_i) = -p_i log_2 p_i$. The relative contribution of a sensor node towards the probable instability of the network can be estimated using relative thermal entropy by calculating the entropy in neighborhood i.e.

$$H^{therm}_{rel} = \frac{H(p_i)}{\sum_{j \in nbr_i} H(p_j)} \qquad (8)$$

Where H^{therm}_{rel} indicates the relative thermal entropy and $\mathrm{nbr_i}$ represents the neighborhood dataset.

4.2. Energy transfer and Asynchronous sleep/awake cycle

The amount of energy a node i could acquire from its neighbouring sensor node j through RF transfer based on sensor node's ability to control their power level, can be defined in equation (9) and (10) as:

$$E_{trans(V_j,V_i)} = \eta_1 \mu P_j |h_{V_i,V_j}|^2 = \eta_1 \mu P_j |\beta_1 d^{-\alpha_1}_{(V_i,V_j)}|^2 \qquad (9)$$

$$\varGamma_{V_i} = \sum_{j=1}^{k} E_{trans(V_j,V_i)} = \sum_{j=1}^{k} \eta_1 \mu P_j |\beta_1 d^{-\alpha_1}_{(V_i,V_j)}|^2 \qquad (10)$$

where $E_{trans(V_j,V_i)}$ is the amount of energy node j can transfer to its neighbor$i$, η_1 is the energy conversion efficiency 0 <η_1< 1, μ is the energy and data splitting ratio 0 <μ< 1, P_j is the signal power received from node j, h_{V_i,V_j} is the channel gain, β_1 is a constant which depends on the environment's radio propagation properties, α_1 is the path loss exponent, and $\varGamma_{V_i}$is the node V_i's gain degree. Foregoing in view, the total available energy at node i can be computed as:

$$E_{T(V_i)} = \varGamma_{V_i} + E_{Bat(V_i)} \qquad (11)$$

where $E_{Bat(V_i)}$ is the remaining battery energy of node i in (11). The amount of energy shared by a node with its neighbours depends on activities such as sensing, relaying, sleep/awake schedule etc. In contrast to conventional routing algorithms in WSNs, our proposed scheme can serve both data and energy in its routing topology. We used opportunistic connection random graph (OCRG) in FLOC to model the opportunistic node connections between sensor nodes. Let $G(S_{SN}, O_C, L)$ be the graph representing OCRG in which S_{SN} represents the set of nodes in the network, O_C represents set of opportunistic connections existing between any two adjacent neighbours and L represents the link connectivity of any two adjacent nodes in S_{SN}. As we know that if any sensor node works for longer time, it is highly likely that sensor node will be able to communicate with neighbours due to higher status transition frequencies, thus it will contribute in improving the link connectivity. The link connectivity also depends on the data routing cost $DR_C\colon O_C \rightarrow R$ such that $DR_C(i,j)$ is the cost associated with link (i,j). Our routing metric can be defined as: $Min \sum_{i=1}^{n-1} DR_{C(V_i,V_{i+1})}$, $DR_{C(V_i,V_{i+1})} \geq 0$. The data routing cost can be computed using $E_{C(V_i,V_j)}$, $E_{T(V_i)}$ and $E_{T(V_j)}$ and is given in (12).

$$DR_{C(V_i,V_j)} = \frac{E_{C(V_i,V_j)}}{(E_{T(V_i)} + E_{T(V_j)})} \qquad (12)$$

where $E_{C(V_i,V_j)}$ is the transmission energy consumed over link (i,j), $E_{T(V_i)}$ is the available total energy (including battery and gained energy through RF transfer) of node i, $E_{T(V_j)}$ is the available energy (including battery and gained energy through RF transfer) of node j. It is pertinent to mention that energy is also transferred along with the data in the routing process to compensate for the transmission energy consumed over each link and more energy is conserved than consumed as the sensor nodes are using strong signals for transmission purposes.

As far as asynchronous sleep/awake cycle is concerned, we have proposed the concept of sleep/awake cycle schedule (W_v/S_v) and status transition frequencies (F_{ST}) to investigate the opportunistic node connection between sensor nodes in each data collection period. We calculated the time-frequency parameter $TF_{V_iV_j}$ based on working time W_{V_i} and W_{V_j} of sensor nodes V_i and V_j, data collection duration T_{CP}, status transition frequencies F_{ST_i} and F_{ST_j} in equation (13) and (14) as,

$$TF_{V_iV_j} = \left(\frac{F_{STV_i}}{F_{ST_{max}}} \times \frac{W_{V_i}}{T_{CP}}\right)\left(\frac{F_{STV_j}}{F_{ST_{max}}} \times \frac{W_{V_j}}{T_{CP}}\right) \qquad (13)$$

$$TF_{V_iV_{SINK}} = \left(\frac{F_{STV_i}}{F_{ST_{max}}} \times \frac{W_{V_i}}{T_{CP}}\right)\left(W_{V_{SINK}}\right) \qquad (14)$$

Using the time-frequency parameter $TF_{V_iV_j}$ and data routing cost DR_C, our link connectivity $L_{V_iV_j}$ can be computed in (15) as,

$$L_{V_iV_j} = \alpha_2 DR_C(i,j) + (1-\alpha_2) TF_{V_iV_j} \qquad (15)$$

$$L_{V_iV_j} = \alpha_2 \left(\frac{E_{C(V_i,V_j)}}{(E_{T(V_i)} + E_{T(V_j)})}\right) + (1-\alpha_2)\left(\frac{F_{STV_i}}{F_{ST_{max}}} \times \frac{W_{V_i}}{T_{CP}}\right)\left(\frac{F_{STV_j}}{F_{ST_{max}}} \times \frac{W_{V_j}}{T_{CP}}\right) \qquad (16)$$

Where α_2 is the appropriate weight assigned to data routing cost and time-frequency parameter in (16).

5. HESITANT FUZZY LINGUISTIC TERM SET (HFLTS) ANALYSIS

We know that the nature of our problem is subjective and uncertain, that’s why we need a fuzzy computation based technique like HFLTS. A generalization of the basic fuzzy set which deals with the uncertainty starting from the hesitation in the assignment of membership degrees of an element is known as Hesitant Fuzzy Set (HFS) [7-9]. We start the HFLTS analysis with a set of inputs containing total number of nodes, sink, neighbor information, context free grammar, transformation function, set of alternatives, set of criteria and weight assignment. In FLOC, a node can attain two states based on the node’s gain energy and energy welfare. If the node’s gain degree is greater than the threshold and the normalized energy welfare is greater than the half of the maximum value of energy welfare, then the state evaluation of a node can be considered as 'optimistic'. Likewise, if the node’s gain degree is less than the threshold and the normalized value of energy welfare is less than the half of the maximum value of energy welfare, then the state evaluation of a node can be considered as 'pessimistic'. After evaluating the state and acquiring the neighbourhood information, the node calculates the relative thermal entropy with reference to neighborhood. The next step is to store different attributes like ambient temperature, relative thermal entropy, node gain degree, link connectivity, EOH [6] and link quality factor in an array and perform the data standardization by normalizing different attributes to obtain the fractional representation of attributes within [0 1] before defining the criteria. The set of required actions of a sensor node is known as alternatives which can be denoted as {‘CH’, ’CM’, ’Relay’} and the suitable action chosen by the sensor node i from alternatives is based on the criteria defined in (17) i.e.,

$$Criteria = \left\{\begin{matrix} Node\,gain\,degree \\ Energy\,Welfare \\ Relative\,thermal\,entropy \\ Link\,Conn \\ Expected\,Optimal\,Hop \\ Link\,quality\,factor \end{matrix}\right\} = \{E_{T(V_i)}, EW, H_{rel}^{therm}, L_{V_iV_j}, EOH, LQR\} \qquad (17)$$

And the corresponding weights assigned to the members defining the criteria will be $w_T = \{w_1, w_2, \dots w_{|Criteria|}\}$, $|Criteria|$ is the cardinality of Criteria. Now we assume our linguistic term set as,

$$S = \left\{\begin{matrix} s_1{:}Extremely\ low(el) \\ s_2{:}very\ low(vl) \\ s_3{:}low(l) \\ s_4{:}medium(m) \\ s_5{:}high(h) \\ s_6{:}very\ high(vh) \\ s_7{:}perfect(p) \end{matrix}\right\} \qquad (18)$$

The normalized attribute values after data standardization in the hesitant fuzzy set are converted to linguistic term set S using triangular membership function as depicted from Fig 1. A context-free grammar G_{CF} [7] has been used to produce linguistic terms for the alternatives against different values of the criteria. These HFS membership values are then transformed into HFLTS using a transformation function E_{GCF} as shown in equation (19) and (20).

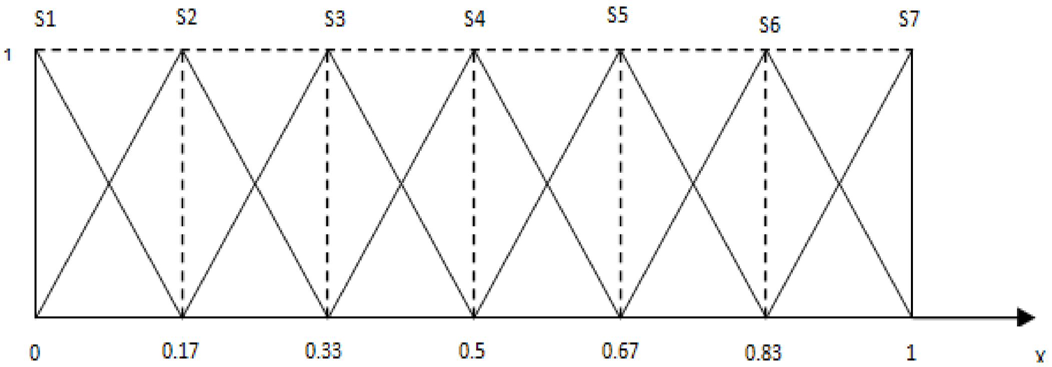


Figure1. Linguistic term set conversion using triangular membership function

$$H = \left[\begin{array}{ccccccc} & c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \\ x_1 & greater\ than\ h & greater\ than\ h & lower\ than\ m & greater\ than\ h & lower\ than\ m & between\ h\ and\ p \\ x_2 & between\ l\ and\ h & between\ v_l \&\ h & between\ l\ and\ v_h & greater\ than\ m & between\ l\ and\ v_h & between\ m\ and\ v_h \\ x_3 & greater\ than\ l & between\ v_l \&\ h & lower\ than\ h & greater\ than\ m & lower\ than\ m & between\ h\ and\ p \end{array}\right]$$

(19)

where, $c_1 = Node\ gain\ degree$, $c_2 = Energy\ Welfare$, $c_3 = Relative\ thermal\ entropy$, $c_4 = Link\ Connectivity$, $c_5 = Expected\ Optimal\ Hop$, $c_6 = Link\ quality\ factor$, $x1$ = CH state, $x2$ = CM state, $x3$ = Relay state. Subsequently, the decision matrix D_M is then converted to HFLTS by using a transformation function E_{GCF}.

$$H = \left[\begin{array}{ccccccc} & c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \\ x_1 & \{v_h, p\} & \{v_h, p\} & \{el, v_l, l\} & \{v_h, p\} & \{el, v_l, l\} & \{h, v_h, p\} \\ x_2 & \{l, m, h\} & \{v_l, l, m\ h\} & \{\ l, m, h, v_h\} & \{h, v_h, p\} & \{l, m, h, v_h\} & \{m, h, v_h\} \\ x_3 & \{m, h, v_h, p\} & \{v_l, l, m\ h\} & \{el, v_l, l, m\} & \{h, v_h, p\} & \{el, v_l, l\} & \{h, v_h, p\} \end{array}\right] \quad (20)$$

The decision matrix D_M includes the members h_{ij} where $i \in \{x_1, x_2, x_3\}$ and$j \in \{c_1, c_2, c_3, c_4, c_5, c_6\}$. According to the definition of hesitant fuzzy linguistic term set, we can easily calculate the envelope of its members h_{ij} using upper bound and lower bound rules. Accordingly, the new decision matrix Y containing the envelopes of H is given in equation (21) as,

$$Y = \left[\begin{array}{ccccccc} & c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \\ x_1 & \{v_h, p\} & \{v_h, p\} & \{el, l\} & \{v_h, p\} & \{el, l\} & \{h, p\} \\ x_2 & \{l, h\} & \{v_l, h\} & \{\ l, v_h\} & \{h, p\} & \{l, v_h\} & \{m, v_h\} \\ x_3 & \{m, p\} & \{v_l, h\} & \{el, m\} & \{h, p\} & \{el, l\} & \{h, p\} \end{array}\right] \quad (21)$$

Now we utilize the node gain degree and energy welfare to classify the status of a node as ‘Optimistic’ and ‘Pessimistic’. The ‘RetainFunc’ will be called if the status of a node is evaluated as ‘Optimistic’ and ‘ChangeFunc’ will be called if the status of a node is evaluated as ‘Pessimistic’ depending upon the status evaluation criteria already discussed.

```
Function: RetainFunc (Alternatives, Criteria, Y)
For i = 1 to |Alternatives| of Y,
    For j = 1 to |Criteria| of Y
        Get 1-cut hesitant fuzzy set $H_S^j(xi)_{\alpha=1}$ for $H_S^j(xi)$,
        where, $H_S^j(xi)_{\alpha=1} = [\{H_{S^-}^j(xi)_{\alpha=1}, H_{S^+}^j(xi)_{\alpha=1}\}]$
    End
End
For e = 1 to |Alternatives| of Y
    For f = 1 to |Criteria| of Y
        Get the intervals $I_{max}(xe)$ foreach alternative $x_i$ with respect to each criterion $f$;
        where, $I_{max}(xe) = [Max(H_{S^-}^f(xi)_{\alpha=1}), Max(H_{S^+}^f(xi)_{\alpha=1})]\ f \in Criteria = [u_{e1}^{max}, u_{e2}^{max}]$
        $Rank_{e1}^{opti} = \max(1 - max(\frac{1-u_{i1}^{max}}{u_{i2}^{max}-u_{i1}^{max}+1}, 0), 0)$ (From equation (26))
    End
End
Return $max(Rank_{e1}^{opti})$
```

```
Function: ChangeFunc (Alternatives,Criteria, Y) Algorithm:
For i = 1 to |Alternatives|of Y
    For j = 1 to|Criteria| of Y
        Get 1-cut hesitant fuzzy set $H_S^j(xi)\alpha=1$ for $H_S^j(xi)$,
        where, $H_S^j(xi)\alpha=1 = [\{H_{S^-}^j(xi)\alpha=1, H_{S^+}^j(xi)\alpha=1\}]$
    End
End
For e = 1 to|Alternatives| of Y
    For f = 1 to |Criteria| of Y
        Get the intervals $Imin(xe)$ for each alternative $xi$ for each criterion f;
        where $Imin(xe) = [Max(H_{S^-}^f(xi)\alpha=1), Max(H_{S^+}^f(xi)\alpha=1)]\ f \in Criteria = [u_{e1}^{max}, u_{e2}^{max}]$
        $Rank_{e1}^{pessi} = \max(1 - max(\frac{1-u_{i1}^{max}}{u_{i2}^{max}-u_{i1}^{max}+1}, 0), 0)$ (From equation (24))
    End
End
Return $max(Rank_{e1}^{pessi})$
```

The ‘RetainFunc’ and ‘ChangeFunc’ functions applies the 1-cut HFLTS to fuzzy sets in Y to generate the envelope for each criteria against every alternative and calculates the probabilistic ranking of the alternatives based on the interval calculated from the envelopes. For instance, if the probabilistic ranking of alternatives is $[x_1 > x_3 > x_2]$, it indicates that the corresponding sensor node in its current state will probably retain its state and perform the role of a CH instead of CM or Relay node. But if the probabilistic ranking of alternatives is $[x_2 > x_3 > x_1]$ or $[x_3 > x_2 > x_1]$, it indicates that the sensor node’s preferred action will be to change its role as CM or relay instead of CH.

6. RESULTS

6.1. Simulation Environment

We have evaluated the performance of FLOC in MATLAB 2019b and OMNET++ using cross platform library (MEX-API). This Application Programming Interface (API) can provide the user an easy bidirectional connection interface between MATLAB and OMNET++. Nodes are arranged in random topology. We have utilized low rate, low cost, short range, flexible and low power consumption standard IEEE 802.15.4 for our PHY and MAC layer. The performance metrics like active node ratio, average energy consumption and packet delivery ratio are analyzed against parametric benchmarks viz. node density and temperature variation. The performance of

FLOC is compared with three different approaches i.e. 1) SOPC [8], 2) BMAC [17], 3) RL-Sleep [7]. Stop-Operate Power-Control (SOPC) is a temperature-aware asynchronous sleep-scheduling algorithm in which energy, link connectivity and network coverage are preserved by putting a few sensor nodes in hibernation mode and controlling the rest of the sensor nodes' transmission power. The communication range and number of active nodes are adjusted to maintain the critical density for consistent connectivity in the network. Berkley-MAC (BMAC) is a low-traffic, low-power-consuming MAC protocol based on adaptive preamble sampling for duty cycling to preserve energy, provide effective collision avoidance and high channel utilization. RL-Sleep is an asynchronous reinforcement learning based procedure based on the adaptive state transition determined by sensor nodes. The state transition is based on temperature sensing and collecting information from the neighbourhood. The effect of various parameters on the performance of FLOC with other existing benchmarks is provided in this section.

Table1. Simulation parameters

Parameter	Value
Deployment area	500 m X 500 m
N	60-90
T_H	80°C
S_{SUN}^{max}	1
Maximum Temperature	80°C
$Area_{sen}$	20cm^2
d_0	20m
R	200m
ε_{fs}	50nJ/bit/m^2
ε_{mp}	10pJ/bit/m^2
E_{elec}	50nJ/bit
Initial Energy for nodes	5J(for neighbors of sink node) 3J (For other nodes)
n_p	2
c_p	0.5
$mass$	50g
r	0.25
Number of packets	1024
Length of packet	8000bits

6.1.1. Active node ratio

Figure (2) depicts the active node ratio comparison of FLOC with SOPC, BMAC and RL-Sleep. It is evident from the figure that the ratio of average number of active nodes to total number of sensor nodes in the network is higher for FLOC than in any other benchmark. Furthermore, the active node ratio for all approaches is optimum for N=80. We also evaluated the performance of FLOC against SOPC, BMAC and RL-Sleep for varying diurnal temperature. Figure (3) shows the comparison of active node ratio of FLOC and other benchmarks for diurnal temperature variations. It has been observed that FLOC outperforms all three approaches in terms of active node ratio. The number of active sensor nodes in the network varies inversely with the diurnal temperature.

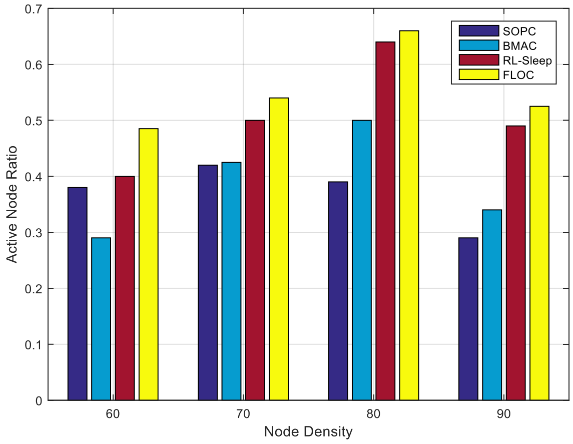


Figure 2. Active node ratio against node density

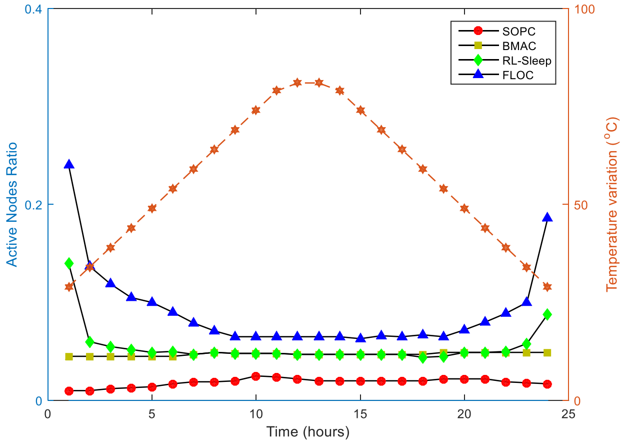


Figure 3. Active node ratio for diurnal temperature variations

6.1.2. Average energy consumption

Figure (4) shows the performance comparison of FLOC with SOPC, BMAC and RL-Sleep in terms of average energy consumption. BMAC outperforms all other algorithms due to its adaptive preamble strategy and short duty cycle which play a significant role in preserving energy. The adaptive adjustment of temperature with respect to communication range leverages higher energy consumption for SOPC. FLOC performs better than SOPC and RL-Sleep but exhibits higher amount of energy consumption against BMAC due to packet broadcasting in the neighborhood. Figure (5) depicts the average energy consumption of FLOC against other approaches for diurnal temperature variation. FLOC and BMAC exhibit almost similar profile for average energy consumed whereas SOPC and RL-Sleep consumed higher amount of energy for N=80.

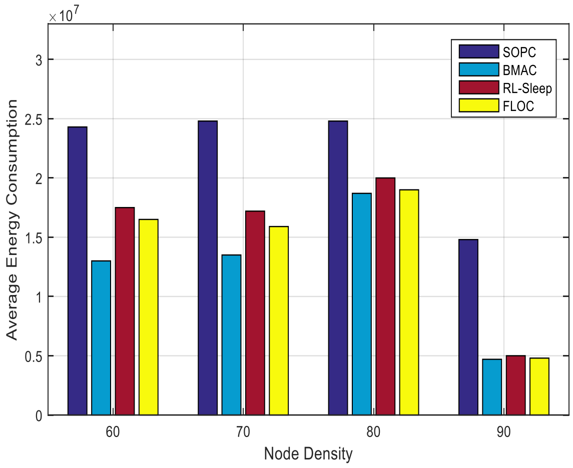


Figure 4. Average energy consumption against node density

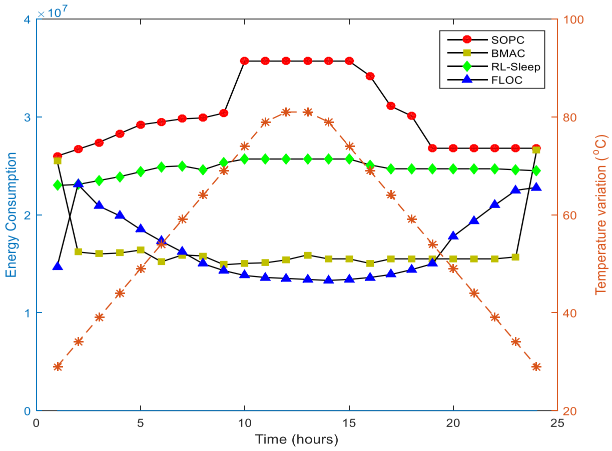


Figure 5. Average energy consumption for diurnal temperature variations

6.1.3. Packet Delivery Ratio (PDR)

Figure (6) depicts the comparison of FLOC with existing benchmarks in terms of PDR. FLOC outperforms other approaches in case of PD. Due to its opportunistic and environment adaptive sleep scheduling strategy, the additional power loss in FLOC can be compensated due to control packet overhead. BMAC shows the worst performance against existing benchmarks. Figure (7) shows the PDR of FLOC with other approaches for diurnal temperature variations. FLOC leverages a better packet delivery ratio in comparison to other techniques. It is pertinent to mention that PDR of FLOC decreases with the increase in diurnal temperature.

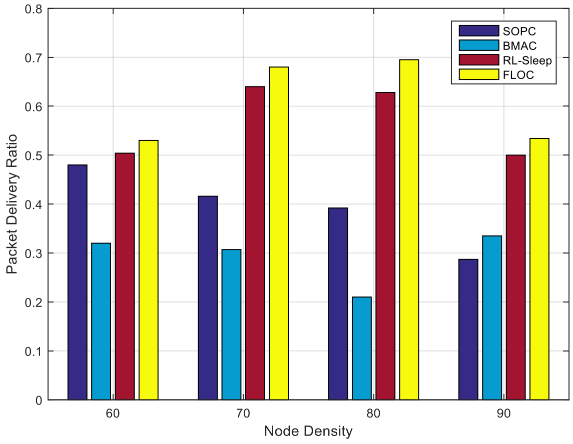


Figure 6. Packet delivery ratio against node density

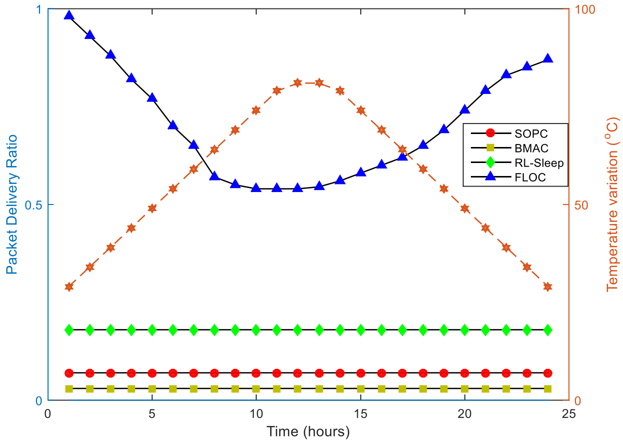


Figure 7. Packet delivery ratio for diurnal temperature variations

7. Conclusions

In this paper, a novel, distributed, FLOC algorithm is proposed based on the hesitant fuzzy linguistic term set (HFLTS) analysis in order to resolve the CH decision making problems and network lifetime bottlenecks using a dynamic network architecture involving opportunistic clustering. The attributes such as energy transfer based opportunistic routing, energy welfare, relative thermal entropy; expected optimal hops and link quality factor are utilized to form the criteria for Hesitant Fuzzy Linguistic Term Set and make a decision about the contemporary role of the node based on its current state. The effectiveness of FLOC is confirmed after carefully analyzing and evaluating its performance against several existing benchmarks. The simulation results have clearly shown that employing FLOC algorithm results in the improvement of active node ratio, average energy consumption and packet delivery ratio. The possible future work would be to perform the hesitant fuzzy linguistic term set analysis for harvested energy scavenging and transfer capabilities in opportunistic clustering.

ACKNOWLEDGEMENTS

The authors would like to thank all the reviewers for their insightful comments and deliberate suggestions for improving the quality of this paper. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1901900, and in part by the National Natural Science Foundation of China (NSFC) under Grant 51776050 and Grant 51536001.

REFERENCES

[1]. O. Ogundile, A. Alfa, “A Survey on an Energy-Efficient and Energy-Balanced Routing Protocol for Wireless Sensor Networks,” Sensors, vol. 17, no. 1084, 2017.

[2]. M. S. Manshahia, “Wireless Sensor Networks: A Survey,” *IJSER*, vol. 74, p. 710–716, 2016.

[3]. A. Boukerche and A. Darehshoorzadeh, ‘‘Opportunistic routing in wireless networks: Models, algorithms, and classifications,’’ ACM Comput. Surv., vol. 47, no. 2, pp. 1–36, Jan. 2015, doi: 10.1145/2635675.

[4]. J. Luo, J. Hu, D. Wu, and R. Li, ‘‘Opportunistic routing algorithm for relay node selection in wireless sensor networks,’’ IEEE Trans. Ind. Informat., vol. 11, no. 1, pp. 112–121, Feb. 2015, doi: 10.1109/TII.2014.2374071.

[5]. G. Yang, Z. Peng, and X. He, ‘‘Data collection based on opportunistic node connections in wireless sensor networks,’’ Sensors, vol. 18, no. 11, p. 3697, Oct. 2018, doi: 10.3390/s18113697.

[6]. Anees, Zhang, Baig, and Lougou, ‘‘Energy-efficient multi-disjoint path opportunistic node connection routing protocol in wireless sensor networks for smart grids,’’ Sensors, vol. 19, no. 17, p. 3789, Sep. 2019, doi: 10.3390/s19173789.

[7]. Banerjee, Partha Sarathi, Satyendra Nath Mandal, Debashis De, and BiswajitMaiti. "RL-sleep: Temperature adaptive sleep scheduling using reinforcement learning for sustainable connectivity in wireless sensor networks." Sustainable Computing: Informatics and Systems, vol. 26, no. 100380, 2020.

[8]. A. Bachir, W. Bechkit, Y. Challal and A. Bouabdallah, "Joint Connectivity-Coverage Temperature-Aware Algorithms for Wireless Sensor Networks," in IEEE Transactions on Parallel and Distributed Systems, vol. 26, no. 7, pp. 1923-1936, 1 July 2015, doi: 10.1109/TPDS.2014.2331063.

[9]. J. Anees, H.-C. Zhang, S. Baig, B. GueneLougou, and T. G. Robert Bona, ‘‘Hesitant fuzzy entropy-based opportunistic clustering and data fusion algorithm for heterogeneous wireless sensor networks,’’ Sensors, vol. 20, no. 3, p. 913, Feb. 2020, doi: 10.3390/s20030913.

[10]. W. Zhai, “Performance Evaluation of Wireless Sensor Networks Based on Hesitant Fuzzy Linguistic Preference Relations,” *Int. J. Online Biomed. Eng.* vol. 14, p. 233–240,2018.

[11]. R. M. Rodriguez, L. Martinez and F. Herrera, "Hesitant Fuzzy Linguistic Term Sets for Decision Making," in *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 1, pp. 109-119, Feb. 2012, doi: 10.1109/TFUZZ.2011.2170076.

[12]. Mo, Xiaoyi, Hua Zhao, and ZeshuiXu, "Feature-based hesitant fuzzy aggregation method for satisfaction with life scale," in Applied Soft Computing, vol. 94, no. 106493,2020.

[13]. J. Anees, H.-C. Zhang, B. G. Lougou, S. Baig, and Y. G. Dessie, ‘‘Delay aware energy-efficient opportunistic node selection in restricted routing,’’ Comput. Netw., vol. 181, Nov. 2020, Art. no. 107536, doi: 10.1016/j.comnet.2020.107536.

[14]. Y. Gao, W. Dong, L. Deng, C. Chen and J. Bu, "COPE: Improving Energy Efficiency With Coded Preambles in Low-Power Sensor Networks," in *IEEE Transactions on Industrial Informatics*, vol. 11, no. 6, pp. 1621-1630, Dec. 2015, doi: 10.1109/TII.2015.2412093.

[15]. M. Mukherjee, L. Shu, L. Hu, G. P. Hancke, C. Zhu, “Sleep Scheduling in Industrial Wireless Sensor Networks for Toxic Gas Monitoring,” *IEEE Wirel. Commun.* vol. 99, p. 2–8, 2017.

[16]. Ye, Wei, John Heidemann, and Deborah Estrin, "Medium access control with coordinated adaptive sleeping for wireless sensor networks,"*IEEE/ACM Transactions on Networking*, vol. 12, no.3, p. 493-506, 2004.

[17]. Polastre, Joseph, Jason Hill, and David Culler, "Versatile low power media access for wireless sensor networks,"*Proceedings of the 2nd international conference on Embedded networked sensor systems*, 2004.

[18]. Shah, Babar, et al. "Guaranteed Lifetime Protocol for IoT based Wireless Sensor Networks with Multiple Constraints." Ad Hoc Networks, no. 102158, 2020.

[19]. P. S. Banerjee, S. N. Mandal, D. De et al, "iSleep: thermal entropy aware intelligent sleep scheduling algorithm for wireless sensor network," *MicrosystTechnol*, vol. 26, p. 2305–2323, 2020. DOI: 10.1007/s00542-019-04706-7.

[20]. P. Musílek, P. Krömer and T. Barton, "E-BACH: Entropy-Based Clustering Hierarchy for Wireless Sensor Networks," 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015, pp. 231-232, doi: 10.1109/WI-IAT.2015.88.

[21]. W. Liang, M. Goh, Y. M. Wang, "Multi-attribute group decision making method based on prospect theory under hesitant probabilistic fuzzy environment," in Computers & Industrial Engineering, vol. 149, no. 106804,2020.

[22]. L. R. Varshney, ''Transporting information and energy simultaneously,'' in *Proc. IEEE Int. Symp. Inf. Theory*, Toronto, ON, Canada, Jul. 2008, pp. 1612–1616, doi: 10.1109/ISIT.2008.4595260.

[23]. S. Guo, F. Wang, Y. Yang and B. Xiao, "Energy-Efficient Cooperative Tfor Simultaneous Wireless Information and Power Transfer in Clustered Wireless Sensor Networks," in *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 4405-4417, Nov. 2015, doi: 10.1109/TCOMM.2015.2478782.

[24]. X. Zhou, R. Zhang, and C. K. Ho, ''Wireless information and power transfer: Architecture design and rate-energy tradeoff,'' *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4754–4767, Nov. 2013, doi: 10.1109/TCOMM.2013.13.120855.

[25]. J. Anees, H. -C. Zhang, B. G. Lougou, S. Baig, Y. G. Dessie and Y. Li, "Harvested Energy Scavenging and Transfer capabilities in Opportunistic Ring Routing," in IEEE Access, vol. 9, pp. 75801-75825, 2021, doi: 10.1109/ACCESS.2021.3082296.

[26]. G. Han, J. Jiang, C. Zhang, T. Q. Duong, M. Guizani, and G. K. Karagiannidis, ''A survey on mobile anchor node assisted localization in wireless sensor networks,'' IEEE Commun. Surveys Tuts., vol. 18, no. 3, pp. 2220–2243, 3rd Quart., 2016, doi: 10.1109/comst.2016.2544751.

[27]. L. Karim, N. Nasser, and T. El Salti, ''RELMA: A range free localization approach using mobile anchor node for wireless sensor networks,'' in Proc. IEEE Global Telecommun. Conf. GLOBECOM, Miami, FL, USA, Dec. 2010, pp. 1–5, doi: 10.1109/glocom.2010.5683802.

[28]. A. Gopakumar and L. Jacob, ''Localization in wireless sensor networks using particle swarm optimization,'' in Proc. IET Conf. Wireless, Mobile Multimedia Netw., Beijing, China, 2008, pp. 227–230, doi: 10.1049/cp: 20080185.

[29]. X. Li, J. Yang, A. Nayak, and I. Stojmenovic, ''Localized geographic routing to a mobile sink with guaranteed delivery in sensor networks,'' IEEE J. Sel. Areas Commun., vol. 30, no. 9, pp. 1719–1729, Oct. 2012, doi: 10.1109/jsac.2012.121016.

AUTHORS

Junaid Anees received B.S. degree from Institute of Space Technology, Islamabad, Pakistan in 2010 and M.S. degree in Electrical Engineering from COMSATS University Islamabad, Pakistan in 2015. Currently, he is a PhD scholar in School of Energy Science & Engineering at Harbin Institute of Technology, China. He holds Senior Manager Position in Ground Segment Network Operations in Public Sector Organization in Pakistan. His research interest includes Energy harvesting Wireless Sensor Networks, Opportunistic Routing, Smart Grids, and Distributed Computing. He is also interested in Cognitive Radio Sensor Networks and Mobile Networking.

Prof. & Dr. Hao-Chun Zhang is currently the Head of the Department of Nuclear Science and Engineering and executive professor of HIT-CORYS Nuclear System Simulation International Joint Research Center (Sino-France). With BE in 1999, ME in 2001 and PhD in 2007 from Harbin Institute of Technology (HIT), Dr. Zhang joined HIT in September 2004. Dr. Zhang has about 200 research publications in peer reviewed journals and conferences, 5 books, and 2 translations of foreign books.

K12 SENIOR HIGH SCHOOL STUDENTS ACADEMIC PERFORMANCE MONITORING SYSTEM FOR PRIVATE INSTITUTIONS WITH DECISION SUPPORT SYSTEM

Dr. Winston G. Domingo, Dr. Erwin N. Lardizabal
and Sheena Marie V. Toledo

College of Information Technology and Engineering, Cagayan Valley Computer and Information Technology College, Philippines

ABSTRACT

The K to 12 Basic Education program uses standards and a competency-based grading system. These are found in the curriculum guides. All grades will be based on the weighted raw score of the learners' summative assessments. Senior High School Students have been graded on three categories the written work, performance tasks, and quarterly assessments. Technology plays a substantial role in helping teachers in the progress, communication, application, and grading of assessment tasks. Thus, this study aims to produce a feasible computerized grading system that will address these issues and problems encountered by the teachers in recording and monitoring grades. The developed K12 Senior High School Students Academic Performance Monitoring System for Private Institutions with Decision Support System was compliant with ISO 25010 quality standards as assessed by SHS Principal, SHS Faculty/ Teachers, and IT Experts. The developed system followed the policy and guidelines set by the department of education in the grading system. The decision support system of the developed system helped the senior high school principal and teachers in monitoring the grades and performance of the students in every subject. Monitoring the performance of the students academically and non-academically, and classifying the students who have at risk in their academic performance.

KEYWORDS

K to 12 Basic Education, competency-based grading system, Decision Support System, Senior High School Students, Academic Performance Monitoring System, ISO 25010 Quality Standards.

1. INTRODUCTION

Information technologies have affected every aspect of human activity and have a potential role to play in the field of education and training, especially, in distance education to transform it into an innovative form of experience. The need for new technologies in the teaching-learning process grows stronger and faster. Technology becomes a time of knowledge providing the complete and unmatched possibility for discovery, exchange of information, communication, and exploration to strengthen the teaching and student learning process. These can help the teachers and students have up-to-date information and knowledge.

Report grades represent teachers' student evaluationsof students' performance. Educators must ensure that grading and reporting always meet the criteria for validity and reliability. And

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 143-157, 2021. CS & IT - CSCP 2021 DOI:10.5121/csit.2021.112011

because of their primary communication purpose, teachers must also ensure that grading and reporting are correct, accurate, and fair. [1].

The K to 12 Basic Education program uses standards and a competency-based grading system. These are found in the curriculum guides. All grades will be based on the weighted raw score of the learners' summative assessments. The minimum grade required to pass a specific subject is 60, which is transmuted to 75 on the report card. The lowest mark that can appear on the report card is 60 for Quarterly Grades and Final Grades. Learners are graded on written work, performance tasks, and quarterly assessments every quarter. These three are given a specific percentage that varies according to the nature of the learning. [2]. Technology plays a substantial role in helping teachers in the development, communication, implementation, and grading of assessment tasks. [3]

Senior High School teachers feel that the time they need to take in the recording of class records. Computing for the grades of their student. With the help of computer technology, schools are taking advantage of a variety of grading systems. However, a greater majority, especially small schools, government schools, and schools in remote areas, still utilize the manual method of recording and computing for the grades of the students.

The researchers want to develop a computerized grading system to lessen the workload of teachers. The common problems encountered in manual recording, accuracy in computations of grades, synchronization of records. As the teacher's workload increases with growing amounts of grades and student lists that need to be attended to, it becomes tedious on the part of the teacher to capably manage them in time for file submission and reporting to higher education authorities. SHS Principal was not able to monitor the updates of class records in every teacher. Thus, this study aims to implement a workable computerized grading system that will address these issues.

1.1. Research Paradigm

This part of the study is about the research paradigm. The proposed study bore three major components: Input, Process, and Output.

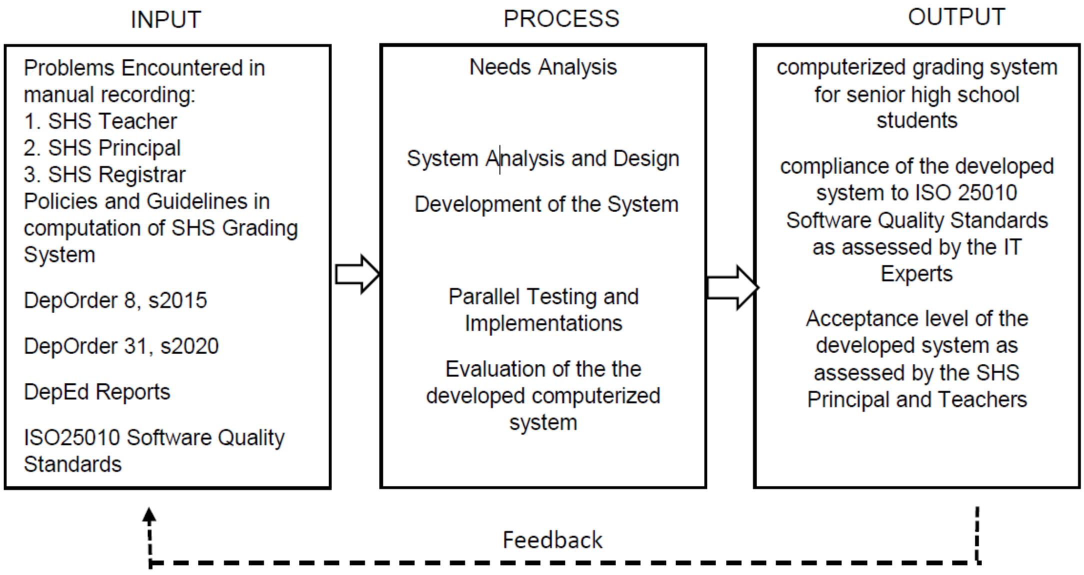

INPUT
PROCESS
OUTPUT
Problems Encountered in manual recording:
1. SHS Teacher
2. SHS Principal
3. SHS Registrar
Policies and Guidelines in computation of SHS Grading System
DepOrder 8, s2015
DepOrder 31, s2020
DepEd Reports
ISO25010 Software Quality Standards
Needs Analysis
System Analysis and Design
Development of the System
Parallel Testing and Implementations
Evaluation of the the developed computerized system
computerized grading system for senior high school students
compliance of the developed system to ISO 25010 Software Quality Standards as assessed by the IT Experts
Acceptance level of the developed system as assessed by the SHS Principal and Teachers
Feedback

Figure 1. Research Paradigm

Fig. 1 Presents the paradigm of the study. Input includes the problem encountered in manual recording, policies, and guidelines in the computation of SHS Grading System, DepEd Order No. 8, s. 2015 and DepEd Reports. The process includes the need analysis and system design and development of System, Testing, and Evaluation. The direct target of this study is to develop K12 Senior High School Students Academic Performance Monitoring System for Private Institutions with Decision Support System, evaluate the level of compliance of the developed system to ISO 25010 Software Quality Standards as assessed by the IT expert, and assess the acceptance level of the system as assessed by the principal and SHS faculty.

1.2. Statement of the Problem

1. What are the problems encountered in the manual grading system in terms of?
    a. Computation of Grades and
    b. Monitoring of Grades
2. What computerized grading system with a decision support system can be developed for the Senior High School?
3. What is the level of compliance of the developed computerized system to ISO 25010 Software Quality Standards as assessed by the IT Expert in terms of:
    - ➢ Functional sustainability;
    - ➢ Performance efficiency;
    - ➢ Compatibility;
    - ➢ Usability;
    - ➢ Reliability;
    - ➢ Maintainability;
    - ➢ Portability and
    - ➢ Security.
4. What is the extent acceptance level of the developed system as assessed by the principal and senior high school teachers in terms of:
    - ➢ Functional sustainability;
    - ➢ Performance efficiency;
    - ➢ Compatibility;
    - ➢ Reliability;
    - ➢ Maintainability;
    - ➢ Portability and
    - ➢ Security.

2. METHODS

2.1. Research Design

This study used a descriptive research design and system development methods. The descriptive method was used to determine the present status and condition of the Senior high school grading system to describe and understand the present environment. Environment analysis and need analysis were done on the adopted grading system of senior high school in this study. The existing senior high grading system policies and practices were analyzed to determine areas of computerization that can be performed for the development of the system, the Software Development Life Cycle (SDLC) methodology was used. This is to ensure that the phases in system development are done in the software building process. The Agile methodology of SDLC was adapted from the business understanding and requirements elicitation phase to testing the developed computerized grading system for senior high school students.

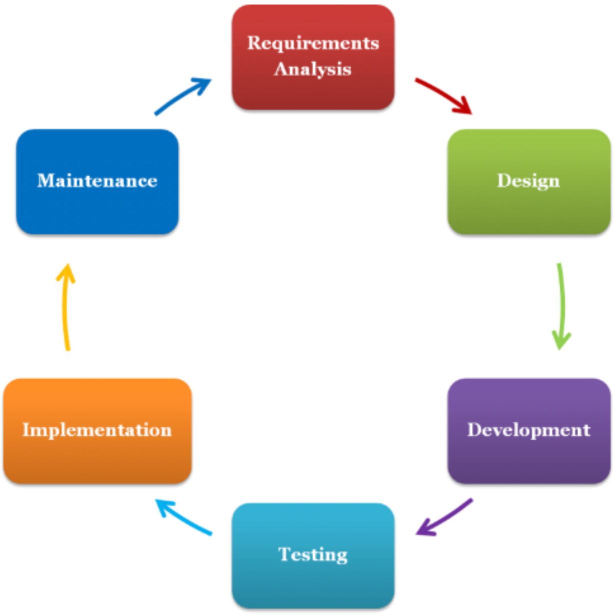


Figure 2. Agile Iterative Model

Fig. 2 Agile Iterative Model was adopted to guide the development of the computerized grading system for senior high [4]. Every iteration in system development involves the following process:

Requirement Analysis. In this procedure, the researcher accompanied a series of interviews with the Senior High School Principal and Teachers who typically administered and monitored the whole actions of the grading system. All the gathered data and information was studied by the researchers to come up with appropriate inputs in designing and developing the computerized grading system for senior high school students.

Design. The researcher chose the appropriate programming software, database, and hardware with which the developed system could be compatible. The researcher constantly coordinated with the users and top management on the features that are suitable for their needs.

Development. The activities involved here are the designing and coding of the user interface. During the development, there were a series of laboratory testing that was conducted in the different modules of the system. Compatibility testing was done and constant coordination with the users was made to align the users' specifications with the developed system.

Testing. In this procedure, the parallel testing of the developed system was done. The researcher collected comments from the testing teams which served as the basis for the modification and redesign of the system.

Implementation. The researcher executed the system in the Department of Senior High School at CVCITC, Santiago City. The system was installed and used. During the implementation phase, a series of training was made to the Principal and Teachers. Calibration and alignment of expectations of the users with the developed system were done.

Maintenance. In this process, the monitoring of the implementation and documentation of the use of the system was done. The problems and challenges encountered by the users were closely recorded and reported. The errors and bugs encountered by the users including suggestions on better features were documented and fixed.

2.2. System Architecture

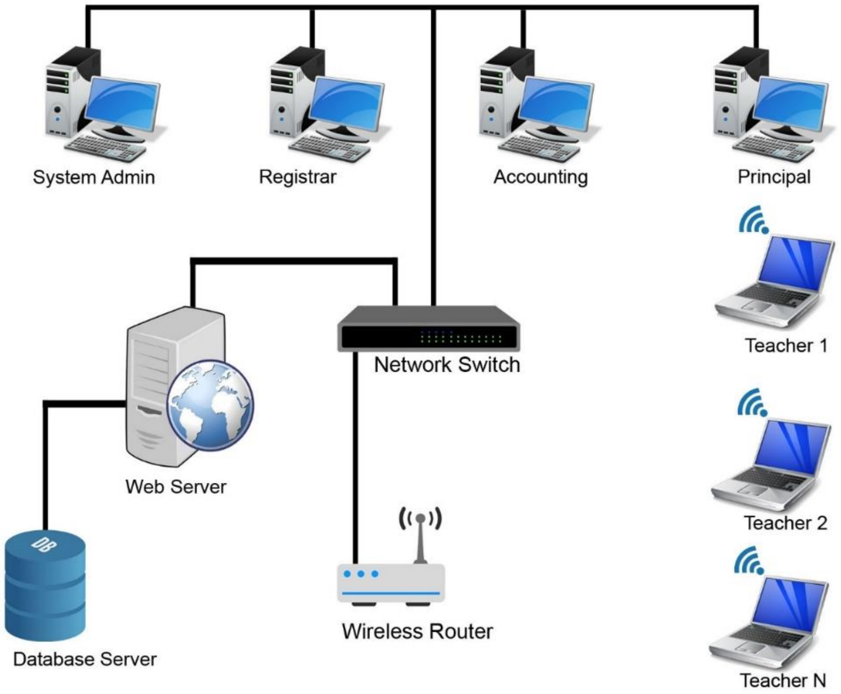


Figure 3. System Architecture

Fig. 3, shows the computerized grading system architecture. The system was designed with a centralized web-based system and database server. The data inputs from the system users were processed on the webserver in it will be stored in the central database server. The system admin is to monitor the overall performance of the system. Registrar is for inputting the student's information during the enrollment and for the subjects enrolled by the teachers. The accounting office is for monitoring the account of each student. Principal monitors the class records of each teacher. Monitoring the permanent records of senior high school students. To check the officially enrolled students for the current term. Teachers were the primary users of the system. They were the ones who input grades into the system. The teachers can check the officially enrolled students through their accounts and subjects. The system can be accessed through the local network wired or wireless.

2.3. Hierarchical Input Process Output

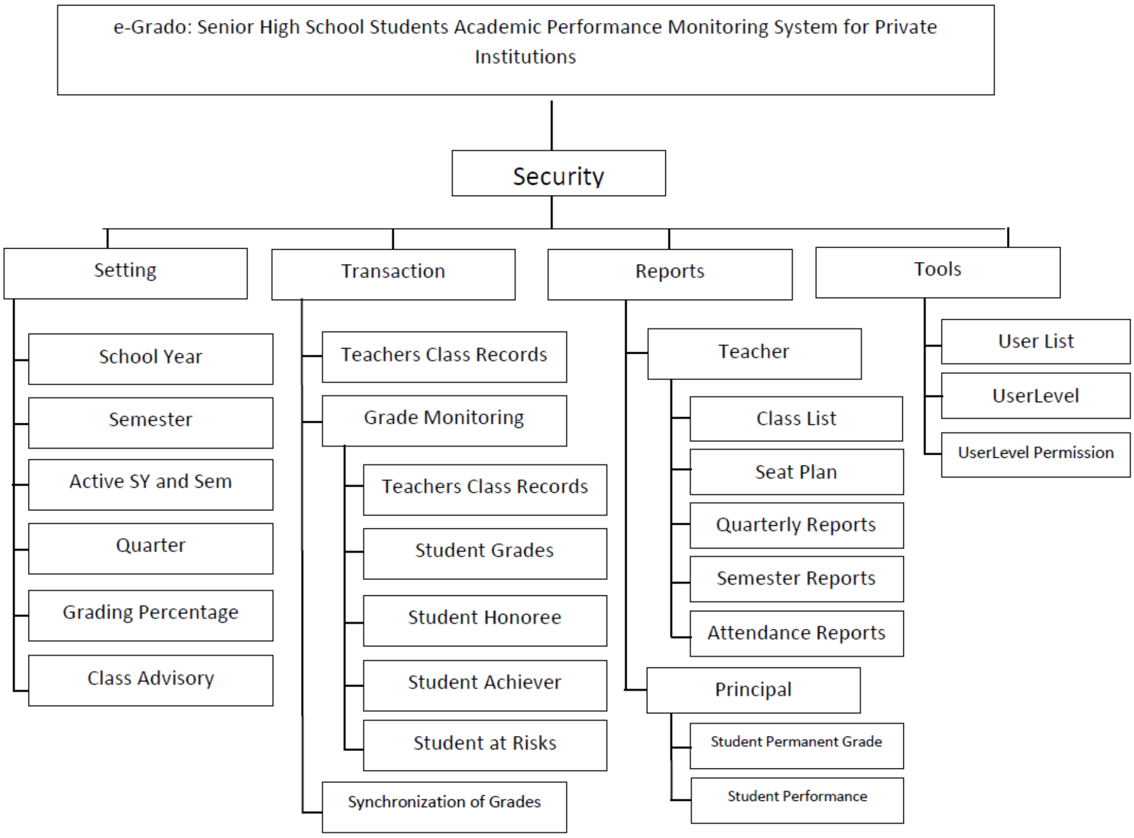


Figure 4. Hierarchical Input Process Output

Figure 4 illustrates the Hierarchical Input Process Output of the system: This figure shows how the system works and the module and sub-modules of each process. It represents the overall design of the system being implemented and the requirements needed.

It is supported by the study of Farahat Ahmed (2015) Hierarchical Input Process Output of the system is a technique and tool for planning and/or documenting a computer program. The HIPO model contains a hierarchy chart that graphically represents the program's control structure and a set of IPO (Input-Process-Output) charts that define the inputs to, the outputs from, and the functions accomplished by each module on the hierarchy chart [6].

2.4. Respondents

There were 16 respondents of the study selected using purposive sampling to determine the practices and policies of the senior high school department. They provided inputs on the Users' specifications such as their needs and challenges. They were the ones directly involved in the operations of the senior high school grading system and the best personnel to get the needed inputs for consideration in the design process of the developed system.

Table 2. Respondents of the Study

Nature of Work	No. of Respondents
Principal	1
SHS Teachers	10
IT Experts	5
Total	**16**

The senior high school principal provided the top management perspective on how the senior high school grading system. How the computerized grading system willhelp in monitoring grades and preparation of reports using computerization? The senior high school teachers are considered as the main users of the system. They will be the ones to use the system, by recording the class records in the system. They also identified the reports that they needed from the developed system as part of the semester and annual reports. As users, they have expressed their report requirements and helped in the evaluation of the developed system. IT Experts will evaluate the performance of the system in terms of Functional sustainability; Performance efficiency; Compatibility; Usability; Reliability and Security.

2.5. Instrument

This study made use of a focus group discussion, observation checklist, interview guide, and documentary analysis.

2.6. Data Gathering Procedure

The researcher secured approval from School Administration and Senior High School Department, CVCITC Santiago City. The study also underwent an ethics review to ensure that there would be no violation of the Privacy Act. The researcher gathered data through a series of interviews. Focus Group Discussion (FGD) was also conducted with the Principal and Senior High School Teachers. The results were the basis of the researchers in the design and development of the system. The researchers conducted form and report evaluation as part of the data gathering procedure to have a deeper understanding of the current grading system. The developed system was tested and used by the users(Teachers and Principals) of the system and they were also involved in the evaluation of the interface of the system. Their recommendations were considered in the development of computerizing the grading system for senior high school.

2.7. Statistical Treatment of Data

Weighted mean was used as the statistical tool. In the evaluation of the developed system, five IT experts were topped, 1 principal and 10 SHS Faculty. The ISO 25010 Software Quality Standards was used as an instrument for assessing the developed system. The results gathered were analyzed employing the 4-point Likert.(4-Highly Accepted, 3-Accepted, 2-Not Accepted, 1-Highly Not Accepted)

Table 3. Likert Scale with Numerical Interpretation

Weight (Likert Scale)	Weighted Mean	Description
4	3.30-4.00	The measure described in the item is Highly Accepted.
3	3.30-4.29	The measure described in the item is Accepted.
2	2.30-3.29	The measure described in the item is Not Accepted.
1	1.00-2.30	The measure described in the item is Highly Not Accepted.

3. RESULTS

1. Issues and Problems Encountered in The Manual Grading System

1.1. Computation of Grades

By using the excel file provided by the DepEd in computing the grades of the students, here were the problems encountered by the teachers:

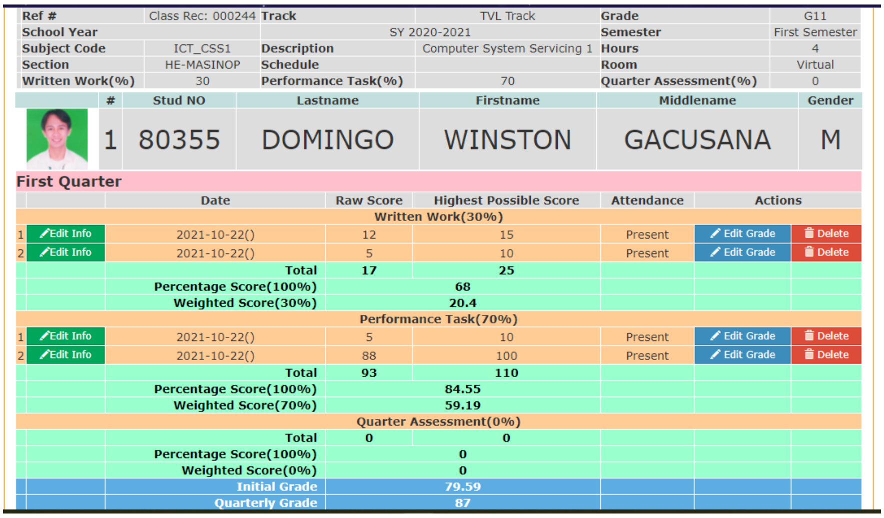
- occasional grades are not accepted in cells even if it is valid;
- cells in the spreadsheets are not automatically computing;
- there are circumstances that fields won't accept input, though it's a valid score;
- some grades were not accurate since some cells are not functioning;
- not easily detect if you inputted wrong values, most especially if you are preparing composite grades, report cards, and permanent record
- occasionally the columns in inputting scores are locked, it is not accessible for editing, especially when it was for completion and
- intrinsic clerical errors in the forms provided by the DEPED to teachers who may not have the technical knowledge to fix or even identify them.

1.2. Monitoring of Grades

The following were the problems encountered in the manual grading system in terms of monitoring of grades:

- not easy to determine failing students;
- class adviser hard to monitor of grades from other subjects;
- there is no indirect monitoring of teachers progress in recording students' grades;
- data inconsistency to the documents that teachers are submitting;
- the preparation of Student Composite Grades (computation of all grades from different courses/subject teachers), it needs more time and effort in completing the report;
- Time-consuming in evaluating students;
- No alternative backup copy of grades; and
- A printed copy is submitted to the office and the digital copy is not shared

2. The developed Senior High School Students Academic Performance Monitoring System for Private Institutions



Ref #	Class Rec: 000244	Track	TVL Track	Grade	G11
School Year			SY 2020-2021	Semester	First Semester
Subject Code	ICT_CSS1	Description	Computer System Servicing 1	Hours	4
Section	HE-MASINOP	Schedule		Room	Virtual
Written Work(%)	30	Performance Task(%)	70	Quarter Assessment(%)	0

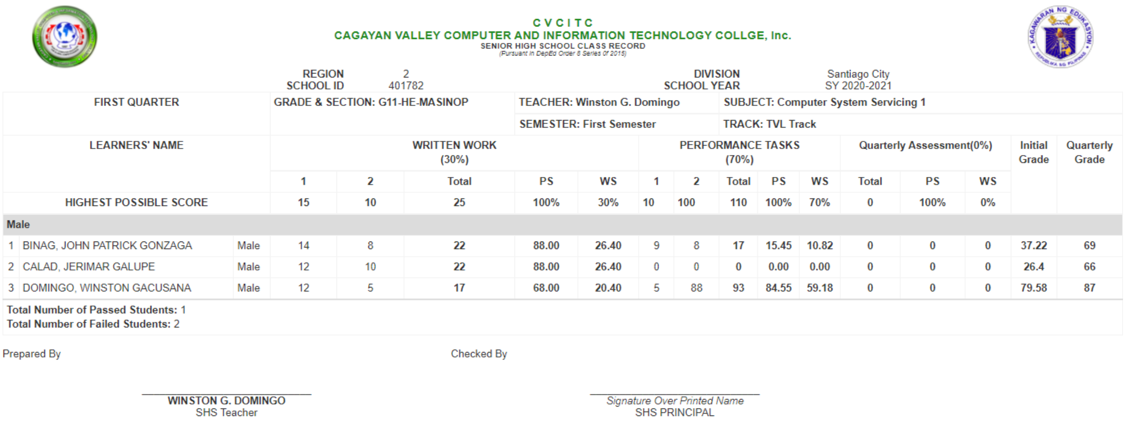
	#	Stud NO	Lastname	Firstname	Middlename	Gender
	1	80355	DOMINGO	WINSTON	GACUSANA	M

First Quarter

		Date	Raw Score	Highest Possible Score	Attendance	Actions	
			Written Work(30%)				
1	Edit Info	2021-10-22()	12	15	Present	Edit Grade	Delete
2	Edit Info	2021-10-22()	5	10	Present	Edit Grade	Delete
		Total	17	25			
		Percentage Score(100%)	68				
		Weighted Score(30%)	20.4				
			Performance Task(70%)				
1	Edit Info	2021-10-22()	5	10	Present	Edit Grade	Delete
2	Edit Info	2021-10-22()	88	100	Present	Edit Grade	Delete
		Total	93	110			
		Percentage Score(100%)	84.55				
		Weighted Score(70%)	59.19				
			Quarter Assessment(0%)				
		Total	0	0			
		Percentage Score(100%)	0				
		Weighted Score(0%)	0				
		Initial Grade	79.59				
		Quarterly Grade	87				

Figure 5. Teacher's Class Records

In fig. 5 teachers' class records display the records in the written work, performance task, and Quarter Assessment. The percentage and weight of each component depend on the track of the programs as indicated in the Department of Education Order No. 8, s. 2015 Table No 5: Weigh of the component for SHS, Page 11[2] and under Department of Education Order No. 31, s. 2020 Grading and Promotion, Table 2: Weight Distribution of the summative assessment components for senior high school [5] The teacher has the privilege to edit a particular record or delete it in case of typographical errors. Computation of Initial Grade and Quarterly Grade based from DepEd Order No 8 Series of 2015, Policy and Guidelines on Classroom Assessment for the K to 12 Basic Education Program, Table 5. Weight of the Components for SHS and Table 7. Steps for Computing of Grades, and DepEd Order No 031 S. 2020, Interim Guidelines for Assessment and Grading in light of the basic education learning continuity plan, Grading and Promotion [5]. The quarterly grade was based on Appendix B. Transmutation Table under DepEd Order No Series of 2015, Policy and Guidelines on Classroom Assessment for the K to 12 Basic Education Program [2]. This module of the system solved the problems and issues encountered by the teacher, the teacher has the privileges to modify the grades, delete grades, and update the grades of the students. In the case of typo errors in the grades, the system has the features to check the score of the students and it will highlight the records which have errors to notify the teacher.



C V C I T C
CAGAYAN VALLEY COMPUTER AND INFORMATION TECHNOLOGY COLLGE, Inc.
SENIOR HIGH SCHOOL CLASS RECORD
(Pursuant in DepEd Order 8 Series Of 2015)

REGION 2
SCHOOL ID 401782
DIVISION Santiago City
SCHOOL YEAR SY 2020-2021

FIRST QUARTER
GRADE & SECTION: G11-HE-MASINOP
TEACHER: Winston G. Domingo
SUBJECT: Computer System Servicing 1
SEMESTER: First Semester
TRACK: TVL Track

LEARNERS' NAME			WRITTEN WORK (30%)					PERFORMANCE TASKS (70%)					Quarterly Assessment(0%)			Initial Grade	Quarterly Grade
			1	2	Total	PS	WS	1	2	Total	PS	WS	Total	PS	WS		
HIGHEST POSSIBLE SCORE			15	10	25	100%	30%	10	100	110	100%	70%	0	100%	0%		
Male																	
1	BINAG, JOHN PATRICK GONZAGA	Male	14	8	22	88.00	26.40	9	8	17	15.45	10.82	0	0	0	37.22	69
2	CALAD, JERIMAR GALUPE	Male	12	10	22	88.00	26.40	0	0	0	0.00	0.00	0	0	0	26.4	66
3	DOMINGO, WINSTON GACUSANA	Male	12	5	17	68.00	20.40	5	88	93	84.55	59.18	0	0	0	79.58	87

Total Number of Passed Students: 1
Total Number of Failed Students: 2

Prepared By

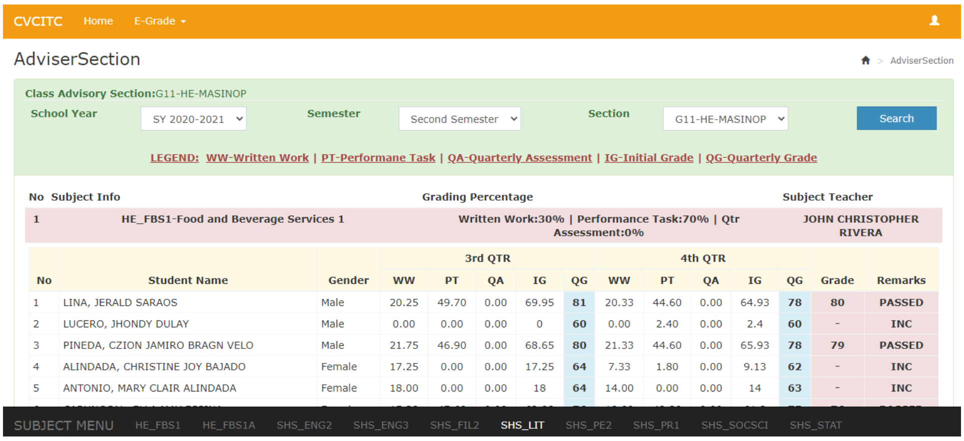
Checked By

WINSTON G. DOMINGO
SHS Teacher

Signature Over Printed Name
SHS PRINCIPAL

Figure 6. Teachers Quarterly reports

Fig. 6 Teachers Quarterly Reports, shows the result of student scores and grades in every quarter. Shows the performance progress of the student in the written task, performance tasks, and quarterly assessment. This report was submitted at the end of the quarter as part of teachers' reports. The format of this report is from table 5. Sample class records page 12 of DepEd Order No 8, series of 2015, Policy and Guidelines on Classroom Assessment for the K to 12 Basic Education Program [5]. The system will count the number of passed and failed students.



Class Advisory Section:G11-HE-MASINOP
School Year: SY 2020-2021
Semester: Second Semester
Section: G11-HE-MASINOP

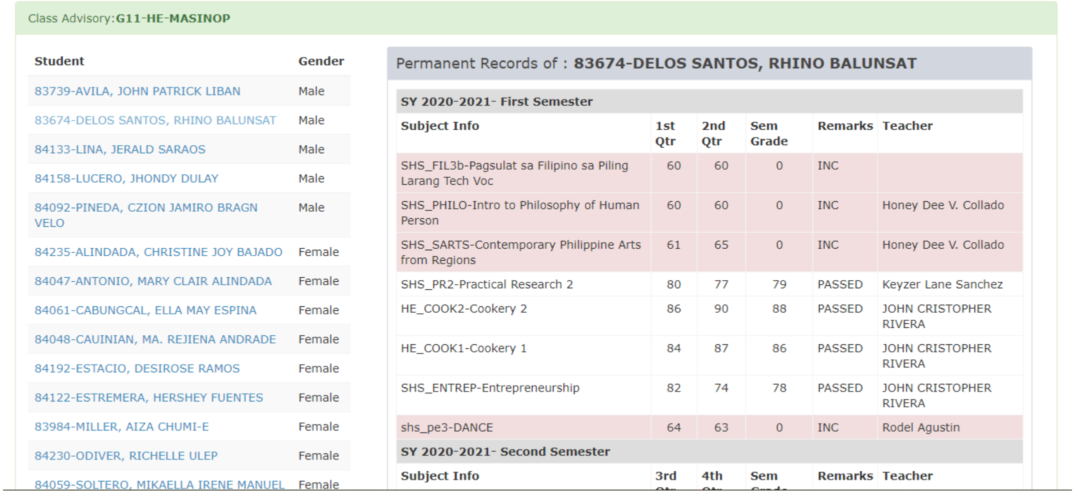
LEGEND: WW-Written Work | PT-Performane Task | QA-Quarterly Assessment | IG-Initial Grade | QG-Quarterly Grade

No	Subject Info	Grading Percentage	Subject Teacher
1	HE_FBS1-Food and Beverage Services 1	Written Work:30% | Performance Task:70% | Qtr Assessment:0%	JOHN CHRISTOPHER RIVERA

			3rd QTR					4th QTR						
No	Student Name	Gender	WW	PT	QA	IG	QG	WW	PT	QA	IG	QG	Grade	Remarks
1	LINA, JERALD SARAOS	Male	20.25	49.70	0.00	69.95	81	20.33	44.60	0.00	64.93	78	80	PASSED
2	LUCERO, JHONDY DULAY	Male	0.00	0.00	0.00	0	60	0.00	2.40	0.00	2.4	60	-	INC
3	PINEDA, CZION JAMIRO BRAGN VELO	Male	21.75	46.90	0.00	68.65	80	21.33	44.60	0.00	65.93	78	79	PASSED
4	ALINDADA, CHRISTINE JOY BAJADO	Female	17.25	0.00	0.00	17.25	64	7.33	1.80	0.00	9.13	62	-	INC
5	ANTONIO, MARY CLAIR ALINDADA	Female	18.00	0.00	0.00	18	64	14.00	0.00	0.00	14	63	-	INC

Figure 7. Adviser Section

Fig. 7 Advisers Section, the principal has to set the class advisory of the teachers. Only teachers with class advisory have access to these features. The adviser can check and monitor the real-time performance of its students under his/her advisory. Subjects under the particular sections will be listed below. The system will provide a summary of class records per subject. The computed Written Work, Performance Task, and Quarterly Assessment per quarter will be displayed. This feature of the computerized system address the a) the class adviser hard to monitor grades from other subjects; b) there is no indirect monitoring of teachers progress in recording students grades; c) data inconsistency to the documents that teachers are submitting; d) the preparation of Composite Grade (computation of all grades from different courses/subject teachers), you will need more time and effort in completing the report; and e) Time-consuming in evaluating students.Also, this feature of the system will be part of the decision support system.



Class Advisory:**G11-HE-MASINOP**

Student	Gender
83739-AVILA, JOHN PATRICK LIBAN	Male
83674-DELOS SANTOS, RHINO BALUNSAT	Male
84133-LINA, JERALD SARAOS	Male
84158-LUCERO, JHONDY DULAY	Male
84092-PINEDA, CZION JAMIRO BRAGN VELO	Male
84235-ALINDADA, CHRISTINE JOY BAJADO	Female
84047-ANTONIO, MARY CLAIR ALINDADA	Female
84061-CABUNGCAL, ELLA MAY ESPINA	Female
84048-CAUINIAN, MA. REJIENA ANDRADE	Female
84192-ESTACIO, DESIROSE RAMOS	Female
84122-ESTREMERA, HERSHEY FUENTES	Female
83984-MILLER, AIZA CHUMI-E	Female
84230-ODIVER, RICHELLE ULEP	Female
84059-SOLTERO, MIKAELLA IRENE MANUEL	Female

Permanent Records of : **83674-DELOS SANTOS, RHINO BALUNSAT**

SY 2020-2021- First Semester

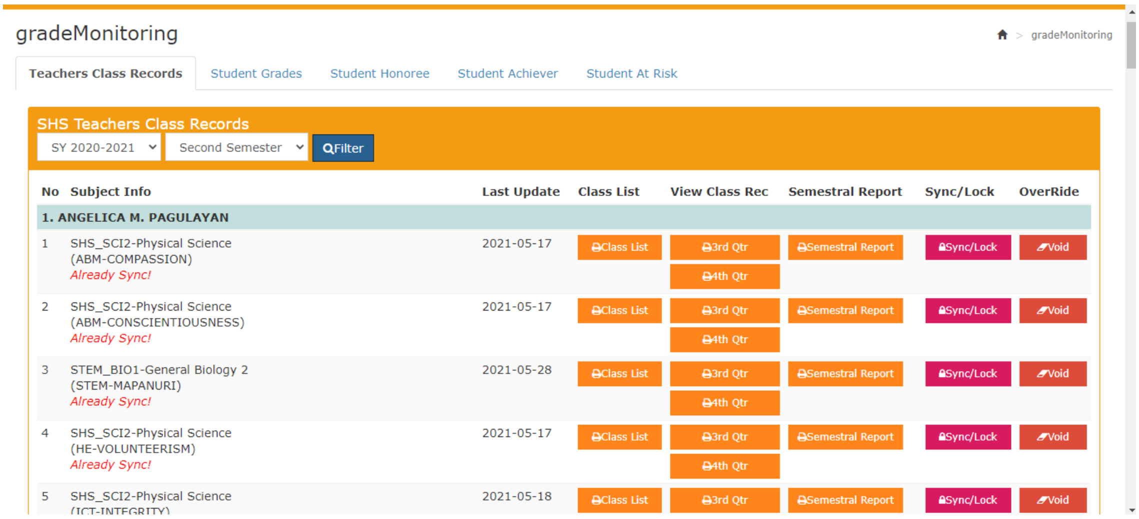
Subject Info	1st Qtr	2nd Qtr	Sem Grade	Remarks	Teacher
SHS_FIL3b-Pagsulat sa Filipino sa Piling Larang Tech Voc	60	60	0	INC	
SHS_PHILO-Intro to Philosophy of Human Person	60	60	0	INC	Honey Dee V. Collado
SHS_SARTS-Contemporary Philippine Arts from Regions	61	65	0	INC	Honey Dee V. Collado
SHS_PR2-Practical Research 2	80	77	79	PASSED	Keyzer Lane Sanchez
HE_COOK2-Cookery 2	86	90	88	PASSED	JOHN CRISTOPHER RIVERA
HE_COOK1-Cookery 1	84	87	86	PASSED	JOHN CRISTOPHER RIVERA
SHS_ENTREP-Entrepreneurship	82	74	78	PASSED	JOHN CRISTOPHER RIVERA
shs_pe3-DANCE	64	63	0	INC	Rodel Agustin

SY 2020-2021- Second Semester

Subject Info	3rd	4th	Sem	Remarks	Teacher

Figure 8. Class Advisory Permanent Grade

Fig. 8 Class Advisory Permanent Grades, this feature of the system was only given to the class advisers. These will generate the permanent records of the per-student under his/her class advisory. This will help the teachers in class cards preparations every end of the semester. The system highlighted the grades with INC or Incomplete Remarks. This shows that the student needs to comply. This will answer the problems and issues encountered by the teachers in preparation of Composite Grade (computation of all grades from different courses/subject teachers), you will need more time and effort in completing the report, this will lessen their time and effort, to make the subject and grades report accurate on time.



gradeMonitoring

SHS Teachers Class Records

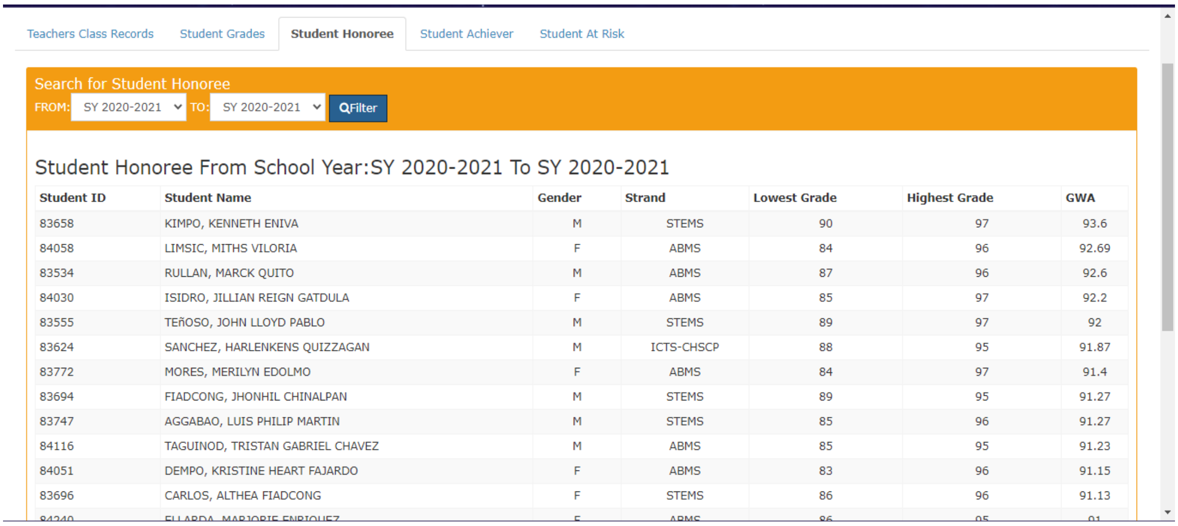
SY 2020-2021 · Second Semester · Filter

No	Subject Info	Last Update	Class List	View Class Rec	Semestral Report	Sync/Lock	OverRide
1. ANGELICA M. PAGULAYAN							
1	SHS_SCI2-Physical Science (ABM-COMPASSION) *Already Sync!*	2021-05-17	Class List	3rd Qtr 4th Qtr	Semestral Report	Sync/Lock	Void
2	SHS_SCI2-Physical Science (ABM-CONSCIENTIOUSNESS) *Already Sync!*	2021-05-17	Class List	3rd Qtr 4th Qtr	Semestral Report	Sync/Lock	Void
3	STEM_BIO1-General Biology 2 (STEM-MAPANURI) *Already Sync!*	2021-05-28	Class List	3rd Qtr 4th Qtr	Semestral Report	Sync/Lock	Void
4	SHS_SCI2-Physical Science (HE-VOLUNTEERISM) *Already Sync!*	2021-05-17	Class List	3rd Qtr 4th Qtr	Semestral Report	Sync/Lock	Void
5	SHS_SCI2-Physical Science (ICT-INTEGRITY)	2021-05-18	Class List	3rd Qtr	Semestral Report	Sync/Lock	Void

Figure 9. Class and Grade Monitoring

Fig 9. Class and Grade Monitoring, these features of the system will produce access to the records of each teacher on their class records. Only the principal or the authorized user has the right to access these modules. This module can display a class list per subject of the teacher, class records for a particular quarter, semesterly report, synching of class records of the students to their permanent records then void/cancel the synched records in case some corrections need to be checked. This void/cancel will be only authorized and approved by the principal. The system will also display when the was the last date of recording of records. This will solve the concerns and

problems of the principal in no indirect monitoring of teachers' progress in recording students' grades, the principal can now check anytime the records of each teacher, and no need to print a hard copy.



Search for Student Honoree

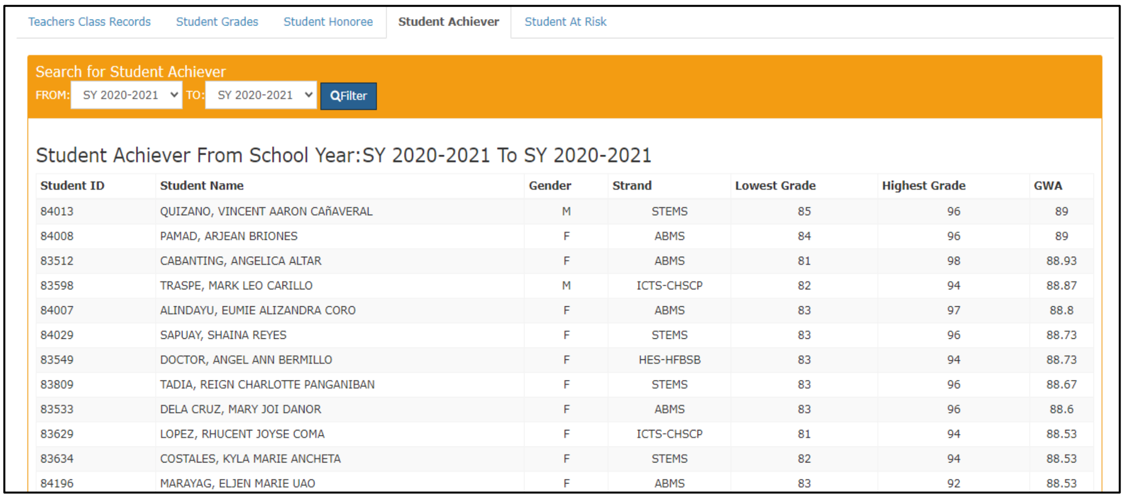
FROM: SY 2020-2021 TO: SY 2020-2021

Student Honoree From School Year:SY 2020-2021 To SY 2020-2021

Student ID	Student Name	Gender	Strand	Lowest Grade	Highest Grade	GWA
83658	KIMPO, KENNETH ENIVA	M	STEMS	90	97	93.6
84058	LIMSIC, MITHS VILORIA	F	ABMS	84	96	92.69
83534	RULLAN, MARCK QUITO	M	ABMS	87	96	92.6
84030	ISIDRO, JILLIAN REIGN GATDULA	F	ABMS	85	97	92.2
83555	TEñOSO, JOHN LLOYD PABLO	M	STEMS	89	97	92
83624	SANCHEZ, HARLENKENS QUIZZAGAN	M	ICTS-CHSCP	88	95	91.87
83772	MORES, MERILYN EDOLMO	F	ABMS	84	97	91.4
83694	FIADCONG, JHONHIL CHINALPAN	M	STEMS	89	95	91.27
83747	AGGABAO, LUIS PHILIP MARTIN	M	STEMS	85	96	91.27
84116	TAGUINOD, TRISTAN GABRIEL CHAVEZ	M	ABMS	85	95	91.23
84051	DEMPO, KRISTINE HEART FAJARDO	F	ABMS	83	96	91.15
83696	CARLOS, ALTHEA FIADCONG	F	STEMS	86	96	91.13
84240	ELLARDA, MARJORIE ENRIQUEZ	F	ABMS	86	95	91

Figure 10. List Student Honoree

Fig. 10. List of Student Honoree, the system will generate a list of possible honoree students. This report will be the basis of the senior high school department to determine the students with academic awards of "with highest honors", "with high honors" and "with honors" during the deliberation of awards, following the criteria with the Academic Excellence Award under DepEd Order 36, series of 2016, Policy Guidelines on Awards and Recognition For The K To 12 Basic Education Program [7].



Search for Student Achiever

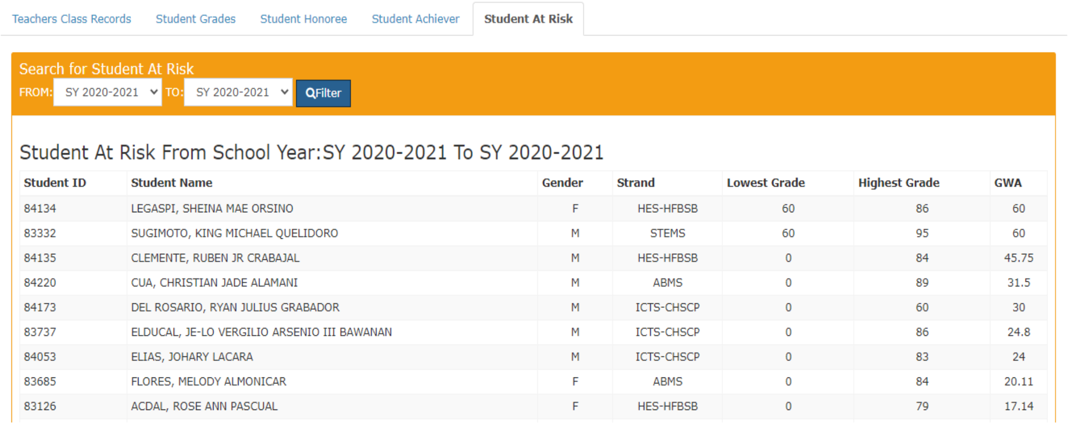
FROM: SY 2020-2021 TO: SY 2020-2021

Student Achiever From School Year:SY 2020-2021 To SY 2020-2021

Student ID	Student Name	Gender	Strand	Lowest Grade	Highest Grade	GWA
84013	QUIZANO, VINCENT AARON CAñAVERAL	M	STEMS	85	96	89
84008	PAMAD, ARJEAN BRIONES	F	ABMS	84	96	89
83512	CABANTING, ANGELICA ALTAR	F	ABMS	81	98	88.93
83598	TRASPE, MARK LEO CARILLO	M	ICTS-CHSCP	82	94	88.87
84007	ALINDAYU, EUMIE ALIZANDRA CORO	F	ABMS	83	97	88.8
84029	SAPUAY, SHAINA REYES	F	STEMS	83	96	88.73
83549	DOCTOR, ANGEL ANN BERMILLO	F	HES-HFBSB	83	94	88.73
83809	TADIA, REIGN CHARLOTTE PANGANIBAN	F	STEMS	83	96	88.67
83533	DELA CRUZ, MARY JOI DANOR	F	ABMS	83	96	88.6
83629	LOPEZ, RHUCENT JOYSE COMA	F	ICTS-CHSCP	81	94	88.53
83634	COSTALES, KYLA MARIE ANCHETA	F	STEMS	82	94	88.53
84196	MARAYAG, ELJEN MARIE UAO	F	ABMS	83	92	88.53

Figure 11. List of Student Achiever

Fig. 11 List of Student Achiever, the system will generate a list of student achievers which the grades of these students did not meet the criteria of the criteria with the Academic Excellence Award under DepEd Order 36, series of 2016, Policy Guidelines on Awards and Recognition For The K To 12 Basic Education Program [7]. These reports will also be used in the preparation of special awards and to be used during the deliberation of awards.



Search for Student At Risk

FROM: SY 2020-2021 TO: SY 2020-2021 Filter

Student At Risk From School Year:SY 2020-2021 To SY 2020-2021

Student ID	Student Name	Gender	Strand	Lowest Grade	Highest Grade	GWA
84134	LEGASPI, SHEINA MAE ORSINO	F	HES-HFBSB	60	86	60
83332	SUGIMOTO, KING MICHAEL QUELIDORO	M	STEMS	60	95	60
84135	CLEMENTE, RUBEN JR CRABAJAL	M	HES-HFBSB	0	84	45.75
84220	CUA, CHRISTIAN JADE ALAMANI	M	ABMS	0	89	31.5
84173	DEL ROSARIO, RYAN JULIUS GRABADOR	M	ICTS-CHSCP	0	60	30
83737	ELDUCAL, JE-LO VERGILIO ARSENIO III BAWANAN	M	ICTS-CHSCP	0	86	24.8
84053	ELIAS, JOHARY LACARA	M	ICTS-CHSCP	0	83	24
83685	FLORES, MELODY ALMONICAR	F	ABMS	0	84	20.11
83126	ACDAL, ROSE ANN PASCUAL	F	HES-HFBSB	0	79	17.14

Figure 12. List of Students at Risk

Fig. 12 List of Students at Risk, the system will generate a list of students with failed grades. These reports will be used also to determine the list of students that need to act with regards to their performance and grades. It will be used also during the deliberation of listing of candidates in graduation.

3. The level of compliance of the developed system to ISO 25010 Software Quality Standards as assessed by the IT Expert

Table 4: Level of compliance of the developed system to ISO 25010 Software Quality Standards as assessed by the IT Expert

ISO 25010 Software Quality Standards.	MEAN	Descriptive Rating
1) Functional Suitability	3.87	Compliant and Highly Accepted
2) Performance Efficiency	3.60	Compliant and Highly Accepted
3) Compatibility	4.00	Compliant and Highly Accepted
4) Usability	3.57	Compliant and Highly Accepted
5) Reliability	3.50	Compliant and Highly Accepted
6) Security	3.68	Compliant and Highly Accepted
7) Maintainability	3.72	Compliant and Highly Accepted
8) Portability	3.80	Compliant and Highly Accepted
GRAND MEAN	3.72	Compliant and Highly Accepted

Table 4 presents the result of the level of compliance of the developed system to ISO 25010 Software Quality Standards as assessed by the IT Expert that obtained the Grand mean of 3.72 with the descriptive rating of compliant and highly accepted. The indicator of ISO 25010 Software Quality Standards such as functional sustainability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability got the descriptive rating of compliant and highly accepted. Therefore, it can be inferred that the developed application was highly approved and accepted by the IT experts.

4. The extent of acceptance level of the developed system as assessed by the principal and senior high school teachers

Table 5: The extent of acceptance level of the developed system as assessed by the principal and senior high school teachers

ISO 25010 Software Quality Standards.	MEAN	Descriptive Rating
1) Functional Suitability	3.76	Compliant and Highly Accepted
2) Performance Efficiency	3.70	Compliant and Highly Accepted
3) Compatibility	3.50	Compliant and Highly Accepted
4) Usability	3.65	Compliant and Highly Accepted
5) Reliability	3.66	Compliant and Highly Accepted
6) Security	3.60	Compliant and Highly Accepted
7) Maintainability	3.56	Compliant and Highly Accepted
8) Portability	3.64	Compliant and Highly Accepted
GRAND MEAN	3.63	Compliant and Highly Accepted

Table 5 presents the result to the extent of acceptance level of the developed system to ISO 25010 Software Quality Standards as assessed by the principal and senior high school teachers that obtained the Grand mean of 3.63 with the descriptive rating of compliant and highly accepted. The indicator of ISO 25010 Software Quality Standards such as functional sustainability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability got the descriptive rating of compliant and highly accepted. Therefore, it can be inferred that the developed application was highly approved and accepted by the principal and senior high school teachers. Thus, it results in the full implementation of the developed system to be used by the SHS faculty and SHS Principal. Comply with all the requirements, policies, and guidelines of Department of Education Order No. 8, s. 2015, Policy on Classroom Assessment for the K to 12 Basic Education Program [5]

4. CONCLUSION

From the above findings, the researcher concluded that the existing manual system/by using Excel for the grading system of the SHS Department can be improved through the adoption of the developed system. The developed K12 Senior High School Students Academic Performance Monitoring System for Private Institutions with Decision Support System was compliant with ISO 25010 quality standards as assessed by SHS Principal, SHS Faculty/ Teachers, and IT Experts. The developed system followed the policy and guidelines set by the department of education in the grading system. The decision support system of the developed system helped the senior high school principal and teachers in monitoring the grades and performance of the students in every subject. To determine the performing students academically and non-academically, to identify the students who have at risk in their academic performance.

And from the findings and conclusions in this study, the researchers recommend the following;

1. The senior high school department may consider using the developed system in inputting of grades;
2. The school may consider acquiring hardware and better equipment capabilities that are necessary to improve the usability and functionality of the developed system;
3. Future researchers and system developers may consider the development of, report for student report card (FORM 138), Transcript of Records (FORM 137), improving

the decision support system features to data analytics, improving the interface design to be responsive in mobile devices to be integrated into the K12 Senior High School Students Academic Performance Monitoring System for Private Institutions with Decision Support System

ACKNOWLEDGMENT

The researcher would like to extend their warmest gratitude to those people who extend their hand in making this research possible; CVCITC Administration, SHS Department, Office of the SHS Principal, SHS Faculty and Teachers, MIS Office, families, and friends.

REFERENCES

[1] Muñoz, M. A., &Guskey, T. R. (2015). Standards-based grading and reporting will improve education. Phi Delta Kappan, 96(7), 64-68.

[2] Department of Education Order No. 8, s. 2015, Policy on Classroom Assessment for the K to 12 Basic Education Program, *Retrieved from http://www.deped.gov.ph/wp-content/uploads/2015/04/DO_s2015_08.pdf*

[3] Department of Education Order No. 31, s. 2020, The Interim Guidelines for Assessment and Grading in light of the Basic Education Learning Continuity Plan, Retrieved from *https://www.deped.gov.ph/wp-content/uploads/2020/10/DO_s2020_031.pdf*

[4[HarinathMallepally (2016), Agile Iterative Model, Retrieved from *http://www.agilemethod.net/p/waterfall.html*

[5] ISO25010 Software Quality Standards, Retrieved from *https://iso25000.com/index.php/en/iso-25000-standards/iso-25010*

[6] Farahat Ahmed (2015), HIPO (hierarchy plus input-process-output), Retrieved from*http://www.engineering-bachelors-degree.com/business-information-management/uncategorized/hipo-hierarchy-plus-input-process-output/*

[7] Department of Education Order No. 36, s. 2016, Policy Guidelines on Awards and Recognition for the K to 12 Basic Education Program, Retrieved from *https://www.deped.gov.ph/2016/06/07/do-36-s-2016-policy-guidelines-on-awards-and-recognition-for-the-k-to12-basic-education-program-2/*

AUTHORS

Dr. WINSTON G. DOMINGO
Dean of the College of Information Technology and Engineering of Cagayan Valley Computer and Information Technology College.

Dr. ERWIN N. LARDIZABAL
BSIT Program Chairperson of Cagayan Valley Computer and Information Technology College.

SHEENA MARIE V. TOLEDO, MSIT
Faculty of College of Information Technology and Engineering, Cagayan Valley Computer and Information Technology College.

ASSESSMENT OF WATER AVAILABILITIES IN THE TANCÍTARO AREA THROUGH THE FUZZY WILLINGNESS TO PAY

José M. Brotons[1], Gerardo Ruiz Sevilla[2], Ruben Chavez[3]

[1]Department of Economic and Financial Studies,
Miguel Hernández University, Elche, 03202 Alicante, Spain
[2]Escuela Nacional de Estudios Superiores. Universidad Nacional
Autónoma de México, Campus Morelia, México
[3]Facultad de Químico Farmacobiología, Universidad Michoacana de San
Nicolás de Hidalgo, Morelia, Michoacán. México

ABSTRACT

The Tancítaro peak is located in the State of Michoacán in Mexico. The current situation of unsustainable consumption of water resources can lead the region to a critical situation if adequate measures are not taken. An improvement in water management involving paying for the use of these resources could improve the situation. This work aims to propose a model allowing obtaining an equilibrium price of the use of water in the Tancítaro area. For this, experts will be consulted among the users of the water and experts among those who currently the right to use it, that is, inhabitant of the reserve area. The use of the Fuzzy logic will allow them to express their willingness to pay and collect data, not in a dichotomous way, but by grading their opinions. The use of Ordered Weighted Average (OWA) will allow the aggregation of these opinions bearing in mind different degrees of optimism or pessimism. The results obtained show an equilibrium price of \$ 0.49 m^{-3}. It should be noted that these are preliminary results and the main objective of the work is the presentation of a methodological proposal.

KEYWORDS

OWA, Water demand function, Water supply function, Willingness to accept, Willingness to pay.

1. INTRODUCTION

The management of Ecosystem Water Services, from an economic and environmental perspective, allows creating the context to generate the necessary conditions, based on organizational policy, aimed at achieving sustainable and comprehensive development [1, 2]. In the west central region of Michoacán, Mexico is located the Tancítaro peak. Due to the economic growth and the development, effective management of the environmental services is required as well as rational use of them. It is predicted that by 2030 several large hydrological regions will be found in a critical condition [3]. In Mexico there is a severe crisis caused by deficient water management, aggravated by both, high rates of deforestation and the loss of the Ecosystem Water Services (representing a country's forests and jungles) [4, 5, 6]

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy - 2021
pp. 159-172, 2021. CS & IT - CSCP 2021 DOI:10.5121/csit.2021.112012

The economic valuation of water resources plays an important role in two aspects: demand management and distribution for its different uses. Optimized management of water resources requires decisions based on economic efficiency, social equality and, above all, ecological sustainability. The values of water resources depend on the quality, location, reliability of access, and availability among others [6].

The state of Michoacán stands out for its fruit production, mainly Hass Avocado (Persea americana). Back in the eighties, the total percentage occupied by fruit trees was only 42%, representing 21,241 ha, and by 2009 the percentage increased to 55% (103,602 ha). The state contributes 10% to the national agricultural Gross Domestic Product (GDP) and agriculture represents 7% of the total state's GDP, establishing itself as the main economic activity in some regions and municipalities [7]. Currently, in Michoacán, there is a planted area of 169,939 ha, from which 64,808 hectares are irrigated and 105.13 hectares are rainfed. The total production is 548,150 tons per season [8] and since 2018, the great economic growth has generated a positive impact on the regional economy, increasing the producers' income, as well as direct and indirect employment [9]. According to De la Tejera et al. [10] more than 47 thousand direct and 187 thousand indirect jobs have been created since then. To sum up, this activity generates annually around $ 30,265,787.40 [11].

These orchards consume about 1,800 l/plant/month, consequently, a hectare of avocado containing 156 trees can consume up to 5.2 times more water than the same area of a natural forest with a density of 677 species per ha. The growth of orchards and their economic benefits forces the change from forest to agricultural land and the intensive use of agrochemicals [12].

The region of Tancítaro peak, with an elevation of 3,800 m., is one of the most important hydrological regions in the state due to the production of avocado whose main destination is exportation. The municipality of Tancítaro is part of this avocado strip [13]. The avocado is the source of the development for approximately 39,783 inhabitants, distributed in 81 towns and communities. This region is one of the most important areas of the country for its production [14]. Here, about 30 million m^3 of water are reported annually, thus benefiting the agricultural activities and domestic use of the inhabitants [13]. The overexploitation and devastation of the forests have provoked the reduction of water availability for agricultural uses. It is expected that the water valuation improves the use efficiency of the water [13, 14, 15].

From an economic logic, the resources' exploitation implies the scarcer the resources, the higher the price would have to be paid for their use. Then, the objective will consist on assessing the economic contribution to irrigation in agricultural systems through the payable provision for the obtained benefits [16,17]. Several methodologies have been used for the valuation of environmental goods, such as the willingness to pay (WTP) or to accept, contingent valuation, travel costs methodology or hedonic prices, among others. In general, these methodologies are based on the user's opinions, in which it is no possible to introduce the subjectivity. Several studies have addressed the willingness to pay for water, such as [18] concerning the Savegre River in Costa Rica, where the cost per opportunity methodology was applied [19] focusing on the Yamuna River, New Delhi. In Mexico, Soto [20] used the contingent valuation method to estimate the benefits of the comprehensive project for the sanitation of Alto Atoyac in Puebla. Sanchez [21], in the Apatlaco River calculated the WTP to improve the water quality of the Apatlaco river basin, or Rodríguez and García [22] studied in the Guayalejo Basin in the south of the state of Tamaulipas.

We are aware of the difficulty that entails making this type of assessment. On many occasions, the responses portray a wish rather than an opinion. In other words, a water buyer tends to indicate a low price when interviewed to avoid to pay a higher real price in the future. For these reasons, we believe that the introduction of subjectivity will make it possible to express opinions to a better way. As a result, the use of Fuzzy Logic is proposed for a better treatment of the subjectivity. Furthermore, the paper will introduce a methodological proposal for the quantification equilibrium price of the water employing Fuzzy Logic, particularly, in the aggregation of subjective information. The use of fuzzy logic introduces a better treatment of the expert's opinions allowing to graduate in them. However, so far, it has not allowed the graduation of the respondent optimism or pessimism degree. A very common aggregation method is the ordered weighted averaging (OWA) operator introduced by Yager [23]. The OWA operator and its extensions have been used in a wide range of applications [24-29].

In this work, given the increasingly pressing water shortage in the Tancítaro area, we propose to make an approximation to the price that could be applied if the public administration makes the necessary improvements to ensure availability for farmers, in the future. For this purpose, experts representing the stakeholders have expressed its opinions through linguistic labels in an artificial market created to determine the equilibrium price. The use of fuzzy logic allows a better treatment of the information provided by the experts. Finally, the use of OWAs and the confidence assigned to each expert allows a graduation of the results according to different degrees of optimism or pessimism.

2. Material and Methods

Next, we will proceed to the estimation of the water demand and supply curves, whose intersection will allow obtaining the equilibrium point.

2.1. Water demand function

In order to estimate the supply curve, a group of J experts has been selected and asked about their willingness to pay a series of prices for water $P=\{P_1,P_2,...,P_P\}$ to ensure water availability in the future. The expert set are administrator of hydraulic resources (CONAGUA) and Organismo Operador de Aguas (OOAPAS) municipal of Tancítaro, Michoacán. In the same way, we have tested the willingness to accept for the people who had the availability of the water in the reservation area. Prices have been presented in ascending way, so that $P_i < P_{i'}, i < i'$. If experts agree to pay for the use of the water, they will be asked if they are willing to pay P_1 \$ m^{-3}. If they are not, the final price would be 0 \$ m^{-3}, and if they are willing to pay P_1, they would be asked for his willingness to pay for a price P_2. If the answer is negative, the maximum price would be P_1 and if it is positive, they would be asked for the next price and so on. Given the subjectivity in each answer, it is accepted that the respondent does not respond with a dichotomous answer (yes / no), but rather that they do so according to linguistic labels such as totally disagree, strongly disagree, disagree, etc. (Table 1). Each of the elements of the table will be assigned a membership function (from zero to one, according to it).

Table 1. Values assigned to the linguistic labels

Linguistic label	**μ_j**
1 Totally disagree	0.00
2: Strongly disagree	0.20
3: Disagree	0.40
4: Neutral	0.60
5: True	0.80
6: Very true	1.00

From this information, it is possible to obtain the water demand function. For this purpose, three different assumptions have been considered

1. Using average means.
2. Assigning different degrees of optimism and pessimism, based on the opinions provided by experts through OWAs.
3. According to the confidence degree generated by each expert.

2.1.1. Using average means

All experts are equally important. In this way, the price that expert j (WTP_j) would be willing to pay for water can be obtained as:

$$WTP_j^1 = \sum_{i=1}^{P} \Delta P_i \cdot \mu_{ij} \tag{1}$$

Being μ_{ij} the membership function assigned by expert j to price i and $\Delta P_i = P_i - P_{i-1}$, that is, the increase that occurs in each new price provided to the expert to express his willingness to pay, over the previous one. In this way, a series of prices has been obtained representing the price that each expert would be willing to pay $WTP^1 = \{WTP_1^1, WTP_2^1, ..., WTP_p^1\}$.

In this way, it is already possible to obtain the water demand function since the price offered by each expert is available. The curve will be obtained:

- The abscissa axis $P = \{P_1, P_2, ..., P_P\}$ will be made up of the prices initially provided to the experts.
- The ordinate axis, the membership function of each P_i, $\mu^1(P_i)$, is obtained by the quotient between the number of experts n_i who were not willing to pay a price WTP_j equal to or lower than P_i and the total number of experts who answered (J).

$$\mu^1(P_i) = \frac{n_i}{J} \tag{2}$$

2.1.2. Assigning different degrees of optimism and pessimism, based on the opinions provided by experts through OWAs.

An ordered weighted average (OWA) is defined as a mapping of dimension n, $F : R^n \to R$ that has an associated weighting vector W of dimension n, $W^T = [w_1, w_2, ... w_n]$, such that $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j = 1$, with

$$f(a_1, a_2, ... a_n) = \sum_{j=1}^{n} w_j \cdot b_j \tag{3}$$

Where b_j is the jth largest of the a_i.

The essence of OWA [23] is the rearrangement of the elements or arguments, causing aggregation in the a_j not associated with a weighting w_j but with the placement order instead.

The weights of expression (3) has been obtained by ordering the prices obtained from the opinions given by the experts (WTP_j) and assigning 1 to the highest, 2 to the second, etc. The weight assigned to position j is, $2j/\left(\left(J+1\right)J\right)$, that is, the quotation between the digit assigned to position the WTP price of expert J in descending order over the total sum of digits (sum of the numbers 1, 2,…, J). This value will be weighted by α factor representing the degree of optimism or pessimism. The highest positive α values correspond to a greater optimism degree, and the lower α values (even negatives) represent a higher pessimism degree.

$$\omega_j^p = \left[\frac{2\cdot j}{\left(J+1\right)\cdot J}\right]^{\alpha} \tag{4}$$

To obtain a weights sum equal to 1, the previous weights will be normalized, dividing them by the total sum of weights.

$$\omega_j = \frac{\omega_j^p}{\sum_{j=1}^{J}\omega_j} \tag{5}$$

In this way, it is possible to obtain the water demand function since the price offered by each expert is available. The curve will be obtained as:

- The abscissa axis, $P=\{P_1,P_2,...,P_P\}$ will be formed by the initial prices
- The ordinate axis, the membership function of each WTP_i, $\mu^2(P_i)$ is obtained as the sum of the weights assigned to each of the experts who provided WTP_j lower or equal to P_i.

$$\mu^2(P_i)=\sum_{j=1}^{J}\omega_j \,/\, WTP_j \le P_i \tag{6}$$

2.1.3. According to the confidence degree of each expert.

In this case, each price is weighted according to the importance assigned to each expert. Each of them was assigned a previous probability ρ_j^* (from 0 to 1) depending on the credibility that they generate. Finally, these probabilities are normalized dividing each of them by the total sum of probabilities.

$$\rho_j = \frac{\rho_j^*}{\sum_{j=1}^{J}\rho_j^*} \tag{7}$$

In this way, it is possible to obtain the water demand function since the price offered by each expert is available. The curve will be obtained:

- The abscissa axis $P=\{P_1,P_2,...,P_P\}$ will be formed by the initial prices
- The ordinate axis indicates the membership function of each WTP_i, $\mu^3(P_i)$, the sum of the probabilities assigned to each expert who obtained a WTP_j lower or equal to each of the prices indicated on the abscissa axis $P=\{P_1,P_2,...,P_P\}$.

$$\mu^3(P_i) = \sum_{j=1}^{J} \rho_j \,/\, WTP_j \leq P_i \tag{8}$$

2.1.4. Water demand function

It will be obtained aggregating the membership functions obtained for each of the three previous methods

$$\mu_i^S = \alpha\mu^1(P_i) + \beta\mu^2(P_i) + \gamma\mu^3(P_i), \text{ with } \alpha,\beta,\gamma \geq 0 \text{ and } \alpha+\beta+\gamma = 1 \tag{9}$$

Where α is the importance assigned to the supply curve obtained considering all the experts the same importance, β considering the optimistic or pessimistic attitude of the demand curve and γ the importance assigned to the demand functions based on the probabilities assigned to each expert.

2.2. Water supply function

The Pico de Tancítaro is made up of 16 hydrological basins together representing 678.1 km^2. They are not large bodies of water, rather, they are low flow runoff between 100-200m^3 s^{-1}, underground hydrography and permeability is medium. So users take advantage of the water through retention or deep excavation. Thus, the study of water demand in the avocado belt focuses on users of the Upper Basin and users of the Lower Middle Basin. As a result, we have proceeded in a similar way to obtain the supply function. Three alternatives will also be used, that is, considering that all experts have the same importance, using OWAs to assign different degrees of optimism and pessimism, and depending on the degree of confidence generated by each expert.

On this occasion, they ask about the price that they will be willing to receive for the resource they have, so they will begin by asking for the higher prices. In this case, the increase indicated in the expression (1) refers to the price reduction provided to the experts in each phase.

2.3. Equilibrium price

The equilibrium point is defined by the intersection of both curves. The equilibrium point will be given by a price p_i and a membership function. The former price p_0 will be the maximum value that a farmer will be willing to pay to obtain water and the minimum that the owner of the resources (inhabitant of the protected areas) will be willing to receive for sharing water resources.

3. RESULTS

3.1. Water demand function

Table 2 shows the responses of the consulted experts. The first column indicates the expert number, the second the degree of confidence of each expert. The following columns indicate the willingness to pay for each of the offered prices. Thus, expert 1, who deserves a degree of confidence of 0.7 indicates that he or she is willing to pay 0.5 and 0.30 \$ m^{-3} of water. However, as the price increases, his willingness to pay decreases. For 0.45 \$ m^{-3} the expert expresses the opinion with (0.8) and for 0.6 with the expression disagrees (0.4), and for the rest of the prices it indicates that he or she is totally disagreeing (0). A similar methodology has been used in other works such as Brotons & Sansalvador [30]. The willingness to pay of the first expert is obtained as:

$$WTP_1 = 0.15 \cdot 1 + 0.15 \cdot 1 + 0.15 \cdot 0.8 + 0.15 \cdot 0.4 + 0.15 \cdot 0 + 0.15 \cdot 0 + 0.15 \cdot 0 = 0.48$$

That is, multiplying the successive increases in each of the prices by the valuation made by the expert for each price. Last column orders the experts by their willingness to pay in descending order.

Table 2. Willingness to pay

		Willingness to pay								
Expert	**Confidence**	**0.15**	**0.3**	**0.45**	**0.6**	**0.75**	**0.9**	**1.05**	**WTP**	**Order**
1	0.70	1.0	1.0	0.8	0.4	0.0	0.0	0.0	0.48	6
2	0.60	1.0	1.0	1.0	1.0	0.8	0.4	0.0	0.78	3
3	0.10	1.0	1.0	1.0	1.0	0.8	0.6	0.2	0.84	2
4	1.00	1.0	0.6	0.2	0.0	0.0	0.0	0.0	0.27	8
5	0.20	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.09	10
6	0.90	1.0	1.0	1.0	0.4	0.2	0.0	0.0	0.54	5
7	1.00	1.0	1.0	0.8	0.2	0.0	0.0	0.0	0.45	7
8	1.00	1.0	1.0	1.0	0.8	0.6	0.2	0.0	0.69	4
9	0.30	1.0	1.0	1.0	1.0	1.0	1.0	0.7	1.005	1
10	0.70	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.12	9

Next, the weights of each expert are shown in Table 3 in three different ways. The first one was obtained assigning equal importance to each expert (average), the second using OWAs: the weight of each expert was obtained according to the opinion expressed in the previous table (WTP), where the WTP were obtained according to expressions (4) and (5), and considering $\alpha = 0.8$. In this case, experts who expressed higher prices were overweighed and the experts who express lower prices were underweighted. OWAs had also been used to estimate unknown values, for example in Sansalvador & Brotons [31]. In the final way, the weight of each expert has been assigned according to the allocated probability and has been obtained according to expression (7).

Table 3. Weights for willingness to pay

Expert	**Average**	**OWA**	**Probability**
1	0.100	0.095	0.108
2	0.100	0.138	0.092
3	0.100	0.152	0.015
4	0.100	0.063	0.154
5	0.100	0.026	0.031
6	0.100	0.110	0.138
7	0.100	0.080	0.154
8	0.100	0.124	0.154
9	0.100	0.165	0.046
10	0.100	0.046	0.108

Membership functions are shown in Table 4. For each WTP, the sum of the weights assigned to each expert who has express that the price to be paid was equal to or lower to the one shown in the first column Table 4. The added value has been obtained by assigning 0.2 to the “average”, 0.4 to the OWA, and 0.4 to the corresponding probability. Las column of Table 4 shows the demand function. Similar weightings have been used in works such as Sansalvador & Brotons

[31] where a new method for the economic evaluation of the ISO 9001 certification was developed.

Table 4. Membership functions for willingness to pay

WTP	Answers	µ average	µ OWA	µ probability	µ weighted average
0.00	10	1.00	1.00	1.00	1.00
0.15	8	0.80	0.93	0.86	0.88
0.30	7	0.70	0.86	0.71	0.77
0.45	7	0.70	0.86	0.71	0.77
0.60	4	0.40	0.58	0.31	0.44
0.75	3	0.30	0.46	0.15	0.30
0.90	1	0.10	0.17	0.05	0.10
1.05	0	0.00	0.00	0.00	0.00

3.2. Water supply function

Similarly, Table 5 shows the willingness to accept the prices for sharing their available water resources.

Table 5. Willingness to accept

		Willingness to accept								
Expert	Confidence	0.15	0.3	0.45	0.6	0.75	0.9	1.05	WTP	order
1	0.8	0.0	0.0	0.4	0.8	1.0	1.0	1.0	0.42	6
2	0.6	0.0	0.0	0.0	0.0	0.4	0.8	1.0	0.72	2
3	0.0	0.0	0.0	0.0	0.0	0.4	0.6	0.8	0.78	1
4	1.0	0.0	0.6	0.8	1.0	1.0	1.0	1.0	0.24	8
5	0.0	0.0	0.2	0.4	0.6	0.8	1.0	1.0	0.45	4
6	0.2	0.0	0.0	0.0	0.2	0.4	0.8	1.0	0.69	3
7	0.3	0.0	0.0	0.2	1.0	1.0	1.0	1.0	0.42	5
8	0.1	1.0	1.0	1.0	0.8	0.6	0.2	0.0	0.36	7
9	0.3	0.7	1.0	1.0	1.0	1.0	1.0	1.0	0.05	10
10	0.7	0.2	1.0	1.0	1.0	1.0	1.0	1.0	0.12	9

The weights allocated to each expert are shown in Table 6

Table 6. Expert Weights for willingness to accept

Expert	Average	OWA	Probability
1	0.100	0.095	0.200
2	0.100	0.152	0.150
3	0.100	0.165	0.000
4	0.100	0.063	0.250
5	0.100	0.124	0.000
6	0.100	0.138	0.050
7	0.100	0.110	0.075
8	0.100	0.080	0.025
9	0.100	0.026	0.075
10	0.100	0.046	0.175

Finally, assigning weights (0.2, 0.4, 0.4) to (α,β,δ), that is, to the memberships obtained by each of the three methodologies (average, OWA and probability), it is possible to obtain the results shown in the last column of Table 7 (μ weighted average), representing the water supply curve.

Table 7. Membership functions for willingness to accept

WTP	Answers	μ average	μ OWA	μ probability	μ weighted average
0.00	0	0.00	0.00	0.00	0.00
0.15	2	0.20	0.07	0.25	0.17
0.30	3	0.30	0.14	0.50	0.31
0.45	7	0.70	0.54	0.80	0.68
0.60	7	0.70	0.54	0.80	0.68
0.75	9	0.90	0.83	1.00	0.91
0.90	10	1.00	1.00	1.00	1.00
1.05	10	1.00	1.00	1.00	1.00

3.3. Equilibrium Price

Figure 1 shows the equilibrium price of the water as a result of the intersection of the previously calculated demand and supply functions. This intersection allows obtaining an equilibrium price of 0.49 \$ m^{-3}, with a membership function is 0.68. The shape of these supply and demand curves depends on the attitude towards risk of the experts consulted [32]. It should also be noted that a greater membership function of the price obtained indicates weaker preference uncertainty.

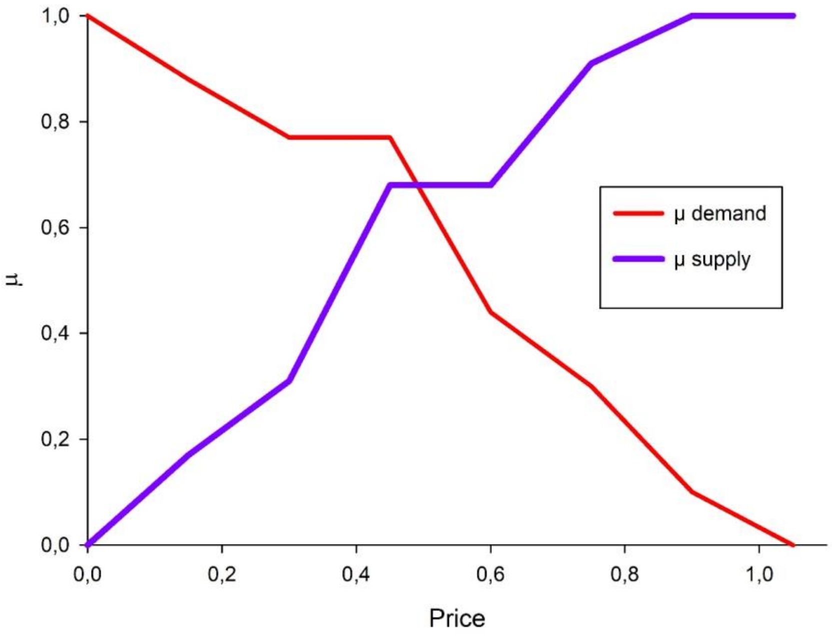


Figure 1. Equilibrium price

The obtained results are in line with those obtained by Rodríguez and García [22], who analyzed the water services payment in sugar cane in the Guayalejo Basin in the state of Tamaulipas, concluding that the price of water could be \$ 0.39 m^{-3}. On the other hand, Chávez and Mancilla

[33] proposed a water rate applied to water users in the Pixquiac River, in Veracruz, Mexico, for which the opportunity cost method was used to assign value to the forest, obtaining a price of $ 0.473 m^{-3} (see Figure 2). Other works show different willingness to pay, mainly due to the peculiarities of each area. For example, Barrantes [34] in the Savegre river in Costa Rica applied the cost per opportunity methodology and obtained a value of US $ 0.0010 m^{-3}. In Mexico, Rodríguez and García [22] studied in the Guayalejo Basin in the south of the state of Tamaulipas, how they have been benefited from the water coming from the "Heaven Biosphere Reserve, obtaining $ 0.39 m^{-3}. Finally, Chávez and Mancilla [33] proposed a water tariff applied to water users in the Pixquiac river, in Veracruz obtaining a value of $ 0.473 m^{-3}.

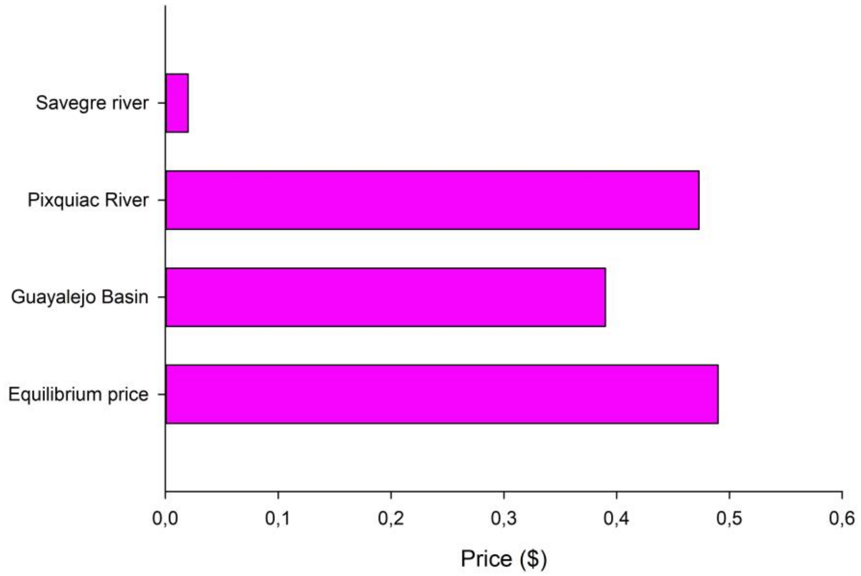


Figure 2. Comparison prices

It should be noted that the application of this methodology has allowed

- The determination of an equilibrium price by creating an artificial market
- The experts to grade their opinions regarding each price with the inclusion of linguistic labels.
- The introduction of OWAs allows graduation the final result according to different degrees of optimism or pessimism in the model
- The introduction of probabilities to each of the experts can improve the quality of the estimation because not all experts deserve the same trust, either because of their knowledge, or because of their interest in obtaining results that are positive for them.

We are aware that it is only an approximation, mainly because on numerous occasions, experts indicate their willingness to pay, something quite different is that if they had to pay, they would actually do it. Therefore, the introduction of fuzzy numbers and OWA extensions should improve the accuracy of the estimation.

4. CONCLUSIONS

The main objective of this work has been to determine an equilibrium price for water in the Tancítaro area. For this purpose, the water users in the lower middle basin area (denser avocado fringe) have expressed the maximum price they would be willing to pay to ensure a continuous supply of water. In the same way, the inhabitants of the protected area (high basin) have expressed their opinion about the minimum price required by them to share their water resources.

The importance assigned to each expert has been considered in two ways, assigning a probability to each expert according the confidence degree in each one as well as just considering their provided values and aggregating them according several degrees of optimism or pessimism.

The fuzzy logic has been introduced in the way the experts express their opinions, using linguist labels. This methodology increases the flexibility of the model since it allows the experts not only to answer in a dichotomous way (yes or no), but also to graduate their opinions.

The intersection of the demand and supply functions allows obtaining the equilibrium price. We are aware that it is a preliminary work and final values may vary significantly compared to those offered in this work, but this paper aims to offer a new methodology applicable cases in which there is no market and it is necessary to create an artificial one to obtain the equilibrium price.

We want to point out that this is a preliminary work and we have used only probability OWAs, OWAs and means, but we are working in the application of some OWAS extension to the water demand and supply, such as induced OWAs. Anyway, the use of intuitionistic fuzzy numbers as well as hesitant fuzzy numbers will improve the quality of our research.

REFERENCES

[1] X. Labandeira, C. J. León & M. X. Vázquez. Economía Ambiental Pearson Educación, S. A., Madrid, 2007 isbn 10:84-205-3651-2

[2] M. Lavado, L. Palma, & M. Cárcamo. Transferencia Tecnológica, Servicios Ecosistémicos y CAPR: Mecanismos de vinculación integral para los diversos actores que conviven en una cuenca: Caso Innova Cuencas APR. Red ProAgua CYTED, Chile, 2013

[3] D. Herrador & L. Dimas. “Aportes y limitaciones de la valoración económica en la implementación de esquemas de pago por servicios ambientales,” 2002.

[4] R. H. Manson, 2004. Los servicios hidrológicos y la conservación de los bosques de México. Madera y Bosques, vol 10, n. 1, 2004, 3-20. Recuperado de http://www.redalyc.org/articulo.oa?id=61710101

[5] M. A. Almendarez-Hernández; L. A. Jaramillo-Mosqueira; G. Avilés Polanco; L. F. Beltrán-Morales; V. Hernández-Trejo & A. Ortega-Rubio. Economic valuation of water in a natural protected area of an emerging economy: recommendations for el Vizcaino Biosphere reserve, Mexico Interciencia, vol. 38 n. 4, Venezuela, 2013, pp. 245-252. Available: https//www.redalyc.org/articulo.oa?id=33926985005

[6] UNESCO, “Valoración económica de los recursos hídricos”. Programa Mundial de Evaluación de los Recursos Hídricos (WWAP), 2017. http://www.unesco.org/new/es/naturalsciences/environment/water/wwap/facts-and-figures/valuing-water/

[7] C. Ortíz-Paniagua, J.C. L. Navarro-Chávez and T. Cortez Ma. “Acercamiento a las metodologías de valoración económica de uso directo extractivo en el contexto de los ecosistemas y elementos para la gestión del desarrollo sustentable”. *Revista Nicolaita de Estudios Económicos*, vol. IV, no. 1 enero – junio, 2009. Pp. 57-84.

[8] SIAP. Servicio de Información Agroalimentaria y Pesquera Avance de Siembras y Cosechas. Resumen por cultivo y entidad, 2020. Recuperado de http://infosiap.siap.gob.mx:8080/agricola_siap_gobmx/ResumenDelegacion.

[9] T. L. Villanueva & J. A. Zepeda-Anaya (2018). "La Producción de Aguacate en el Estado de Michoacán y sus efectos en los índices de pobreza, el cambio del uso de suelo y la migración". *Revista Mexicana Sobre Desarrollo Local*, vol. 0, no 2. ISSN: 2395-863

[10] B. De la Tejera-Hernández; A. Santos; H. Santamaría; T. Gómez & C. Olivares. "El oro verde en Michoacán: ¿un crecimiento sin fronteras? Acercamiento a la problemática y retos del sector aguacatero para el Estado y la sociedad" *Economía y Sociedad*, vol. XVII, no. 29, julio-diciembre, 2013, pp. 15-40. UMSNH, Morelia, Michoacán.

[11] Y. Raya-Montaño, P. Apáez-Barrios; S. Aguirre Paleo, M. Vargas Sandoval; R. Paz Da Silva & M. Lara-Chávez. Identificación de hongos micorrizógenos arbusculares en huertos de aguacate de Uruapan, Michoacán. Revista Mexicana De Ciencias Agrícolas, vol. 23, 2019, 267-276. https://doi.org/https://doi.org/10.29312/remexca.v0i23.2026

[12] Gómez-Tagle (2018) Hydrological impact of the green gold (avocado culture) in central Mexico; rainfall partition and water use comparison with native forests. DOI: 10.13140/RG.2.2.18644.65921 Conference: Joint Conference on Forests and Water. Disponible en: https://www.researchgate.net/publication/329060308_Hydrological_impact_of_green_gold_avocado_culture_in_central_Mexico_rainfall_partition_and_water_use_comparison_with_native_forests

[12] J.J. A. Fuentes-Junco. Análisis morfométrico de cuencas: caso de estudio del parque nacional pico de Tancítaro. Instituto Nacional de Ecología. Dirección General de Investigación de Ordenamiento Ecológico y Conservación de Ecosistemas, 2004.

[13] INEGI. Censo de Población y Vivienda 2010, México: INEGI.

[14] L. Escobar-Jaramillo & A. Gómez-Olaya. "El valor económico del agua para riego un estudio de valoración contingente", *Ingeniería de Recursos Naturales y del Ambiente*, vol 6, 2007, 16-32.

[15] D. Martínez & H. Padgett. "Aguacate: "oro verde" de los templarios", *Periodismo digital*, 2013, Recuperado de https://www.sinembargo.mx/11-10-2013/780868

[16] J. Pérez Roas. "Valoración económica del agua". Centro Interamericano de Desarrollo e Investigación Ambiental y Territorial. CIDIAT. Universidad de los Andes, 2002, Mérida, Venezuela.

[17] C. Montes. Del desarrollo sostenible a los servicios de los ecosistemas. Ecosistemas, vol. 16, no. 3, 1-3. septiembre 2007. Recuperado de http://www.revistaecosistemas.net/index.php/ecosistemas/article/viewFile/87/84

[18] G. Barrantes-Moreno, "Valoración económica de la oferta de agua como un servicio ambiental estratégico", *Ecological Studies*, vol.185, 2006 M.Kappelle (Ed.). Ecology and Conservation of Neotropical Montane Oak Forests. Springer--Verlag Berlin Heidelberg.

[19] Nallathiga y Paravasthu "Economic value of conserving river water quality. Results from a contingent valuation survey in Yamuna River basin, India.". 2010.

[20] Soto G. "Estudio para estimar los beneficios ecológicos del proyecto integral para el saneamiento del Alto Atoyac, Puebla.", 2009.

[21] Sánchez. A et al "Cálculo de Disposición de Pago por Mejora en el Recurso Hídrico en la Cuenca del Río Apatlaco, Morelos, México. Usando el método de valoración contingente" 2010. Tesis. Universidad Pontifica católica de Chile.

[22] Rodríguez R. H., García, G, N., Cantero, M. D., Carreón, P. A., y Del C. Andrade L. E. "Pago por servicios hidrológicos ambientales en la cuenca del Río Guayalejo, Tamaulipas, México". 2012. Papeles de Geografía, (55-56), 167–178.

[23] R.R. Yager "On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision-making", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 18, no. 1, 1988, pp. 183-190.

[24] Liu, X.W. "A general model of parameterized OWA aggregation with given orness level", *Int. J. Approx. Reason*, Vol. 48, No. 2, 2008, pp. 598-627.

[25] Z. Gong, X. Xu, J. Forrest & Y. Yang, "An optimization model of the acceptable consensus and its economic significance", *Kybernetes*, Vol. 45, No. 1, 2016, pp. 181-206.

[26] J.M. Merigó & Casanovas, M. "Decision making with distance measures and induced aggregation operators", *Computers & Industrial Engineering*, vol. 60, no. 1, 2011, pp. 66-76.

[27] J.M. Merigó, A. M. Gil-Lafuente & R.R. Yager, "An overview of fuzzy research with bibliometric indicators", *Applied Soft Computing*, vol. 27, 2015a, pp. 420–433.

[28] J.M. Merigó, D. Palacios-Marqués & B. Ribeiro-Navarrete, "Aggregation systems for sales forecasting", *Journal of Business Research*, Vol. 68, No. 11, 2015b, pp. 2299-2304.

[29] R.R. Yager, "Fuzzy logic in the formulation of decision functions from linguistic specifications", *Kybernetes*, vol. 25, no. 4, 1996, pp. 119 – 130.

[30] J.M. Brotons, M.E. & Sansalvador. "Proposal for a Fuzzy Model to Assess Cost Overrun in Healthcare Due to Delays in Treatment". *Mathematics*, vol. 9, no. 408, 2021. https://doi.org/10.3390/math9040408.

[31] M.E. Sansalvador & J.M. Brotons. "The Application of OWAs in Expertise Processes: The Development of a Model for the Quantification of Hidden Quality Costs". *Economic Computation and Economic Cybernetics Studies and Research*, vol 51, no. 3, 2017 73-90.

[32] G. Cornelis van Kooten & E. Krcmar. "Fuzzy Logic and Non-market Valuation: A Comparison of Methods". *IIFET 2000 Proceedings*, 2000.

[33] M. Chávez Cortés & K. E. Mancilla-Hernandez. "Charging Scheme for Hydrological Services Provided by the Upper Pixquiac River Watershed". *Tecnologia y ciencias del agua*, vol 5, no 5:155-170, September 2014

[34] C. A. Barrantes & E. R. Flores. Estimating the willingness to pay for the conservation of andean rangelands. *Ecol. apl.* vol.12 no.2 Lima ago./dic. 2013

AUTHORS

José M. Brotons Martínez

JM Brotons is Economist (1991, Alicante University) and PhD in Business Administration (2003). From 1999 he is a member of the Economic and Financial Studies Department of Miguel Hernández University in Elche (Spain). He has an 7 h-index (Web of Science), with 29 documents listed in Web of Science (7 in Q1 and 11 in Q2) and 118 citations (accessed on 18/11/2020). In the period 2010-2020, J.M. Brotons has contributed in 47 scientific publications (29 of them published in indexed journals in the Journal Citations Report), 29 national and international congress communications, 11 books; he has participated in 25 projects and 2 contracts; 6 end-of-degree projects. He has obtained two recognised six-year research periods by CNAI and four five-years teaching periods. He is expert in financial valuation and economic analysis applied to the agriculture. In 2003 he defended the thesis "Financing of the Wastewater Treatment Company Sector. An Operational Approach". One of his interest research is the economic analysis of the agriculture production. Another interest research is the company financial valuation as a result of the use of new methodologies such as the innovations in fuzzy logic and the analysis of the quality system in the enterprises.

Gerardo Ruíz Sevilla

Ruíz-Sevilla is a Biologist graduated (1995) from the Michoacan University of San Nicolás de Hidalgo, later he obtained the degree of Master of Science (2002) from the Faculty of Biology of the same University and recently obtained the degree of Doctor (2021) for the Institute for Economic and Business Research with the research line of Water Ecosystem Services and Geographic Information Systems. He obtained a Diploma in Geographic Information Systems and Applications from the Institute for Geo-Information Science and Earth Observation (ITC) of the Netherlands and the Institute of Geography of the UNAM (2003). He is currently a teacher at the National School of Higher Studies of the National Autonomous University of Mexico Morelia campus and has participated as an associate researcher in the project: Two-dimensional flow dynamics and sortive properties of the RAMSAR - Pátzcuaro wetland, with (2012), Processing of the database and cartographic images of project No. 14: "Aquaculture Charter" of the program: Defense for the conservation of aquatic resources of the Fisheries Commission of the State of Michoacán. Morelia, Michoacán (2008), Associate Researcher in the project: Management of a wetland for the conservation of the lake coast in the south of Lake Pátzcuaro, (2010.), Associate Researcher in the project: The Zirahuén Lake Basin: Evaluation integral of its natural resources (2005), Associate Researcher in the project: CONACYT-SAGARPA 2003/245 Evaluation of the evapotranspiration of aquatic plants and its possible control in Lake Pátzcuaro through a chinampas module (2006), Associate Researcher in the project: Humedales de Pátzcuaro as a RAMSAR site. Secretariat of Urbanism and Environment of the Government of the State of Michoacán (2005).

Rubén Chávez Rivera

R. Chavez is Chemical Engineering (1989, Michoacana University) and PhD in Administration (2010). From 2016 he is a member Basic Academic Core of the Masters and Doctors degree in Science of the Desarrollo Regional of Instituto Investigaciones Economico y Empresariales; professor and researcher of Facultad de Quimico Farmacobiologia and Licenciatura de Biotecnologia, Michoacan University in Morelia (Mexico). He has 3 h-index (Web of Science), with 9 documents listed in Web of Science (1in Q2) and 36 citations (accessed on 18/11/2021). In the period 2016-2020, R. Chavez has contributed to 24 scientific publications (3 of them published in indexed journals in the Journal Citations Report), 26 national and international congress communications, 6 books; he has participated in 5 projects agro-industrial. He is an expert in economic analysis applied to regional agro-industrial development. In 2010 he defended the thesis “Strategic design for intellectual capital integration in organizations using diffuse logic”. One of his interests in research is the energy economy and local development. Another interesting research is the company's financial valuation and economic optimization with innovations in fuzzy logic.

COMPARATIVE STUDY OF JUSTIFICATION METHODS IN RECOMMENDER SYSTEMS: EXAMPLE OF INFORMATION ACCESS ASSISTANCE SERVICE (IAAS)

Kyelem Yacouba[1], Kabore Kiswendsida Kisito[1],
Ouedraogo Tounwendyam Frédéric[2] and Sèdes Florence[3]

[1]Department of Informatic, Université Joseph Ki-Zerbo,
Ouagadougou, Burkina Faso
[2]Department of Informatic, Université Norbert Zongo,
Koudougou, Burkina Faso
[3]IRIT, Toulouse, France

ABSTRACT

Justification of recommendations increases trust between users and the system but also generates more relevant recommendations than recommendation systems that do not incorporate it. That is why, we conducted a justification study of the recommendation for IAAS. Our comparative study shows that IAAS, which currently does not offer the opportunity to justify recommendations, needs to be improved. From the analysis of justification methods studied in this work, it appears that none of these methods can be used effectively in IAAS. That is why, we proposed a new IAAS architecture that deals separately with item classification and the extraction of the justification has added the item during recommendation generation. The item selection method remains unchanged as we plan to implement a new strategy to filter user's reviews should now be extended to four elements: the documentary unit, the group of users, the justification and the weight. Opinion A=(UD,G,J,a). Where UD represents the documentary unit, G the user group, J is the justification and a is the weight of the recommendation.

KEYWORDS

IAAS, Justification in Recommender Systems, users reviews, weight of reviews.

1. INTRODUCTION

In order to facilitate the access to the information contained in information systems, recommendation systems have been developed; these systems use the actions of users realized on the system to filter information. These recommendation systems have undergone a high evolution and have allowed the implementation of several approaches such as: the content-based filtering approach, the collaborative filtering approach, the hybrid approach, the demographic approach and the social approach etc [19]. All these approaches have been proposed in order to produce relevant recommendations to users. As for the collaborative filtering approach, the system uses the ratings of similar users to provide them recommendations [19]. With this approach, several algorithms allowing to provide more accurate recommendations to the user have been developed, for instance IAAS.

David C. Wyld et al. (Eds): NATL, NETWORKS, Fuzzy, CSEA - 2021
pp. 173-181, 2021. CS & IT - CSCP 2021 DOI:10.5121/csit.2021.112013

IAAS, Information Access Assistance Service is a collaborative filtering recommendation system. This system aims to be applied in several domains such as videos, audios, images and documents. It uses the notion of voting as a technique for evaluating items. This vote is carried out by the user after having taken note of the document. This user estimates that the document is important for one or more other users, and it materializes it through an opinion [2,3,4][20].

All recommendation approaches have produced algorithms to provide more accurate recommendations to the user. However, the accuracy of the recommendation and its acceptance improves when the user is able to understand the limitations and benefits of the recommendation. Otherwise, the user must receive the recommendation with the reasoning behind it [5],[18]. Thanks to these observations and to the evolution of recommendation systems which is to improve the interface through the justification of recommendations, the notion of justification has been introduced in recommendation systems. There are several types of justifications: keyword justification, influence based justification, content based justification, users reviews justification and comparative justification etc [1],[5],[7]. The justified recommendation gives credibility to the recommendation system.

As for IAAS, which does not take into account the notion of justification of recommended items must be thought of in order to improve the relevance of the recommendation. So, we will see through this study how to justify the recommendations with the IAAS algorithm. To achieve this, we will review the literature on the justification methods used in the recommendation systems in order to compare these methods. Then we position ourselves in relation to IAAS. To do this, we will first present our literature review on IAAS and the justification methods. Then, we will draw up the comparative table and we will finish by summarizing.

2. RELATED WORK

2.1. Information Access Assistance Service (IAAS)

In the IAAS recommendation system, the users appreciate the different documentary units during their consultation and this appreciation is carried out by giving a grade to the documentary unit. Thus a user gives his opinion (A) which is a grade 'a' ranging from 1 to 10, on a documentary unit (UD) and a given group (G). Hence for [2], [3], the opinion is defined by the following triplet A = (UD, G, a).

The system collects all these relevance notices and proceeds implicitly to the calculation and ranking of the relevant items for the user. As each documentary unit can receive several relevance notices, the notion of recommendation weight Pk(UDiGj) has been proposed by [2],[3],[4]. The calculation allows to give a weight to each item to be recommended. As a group of users can receive the recommendation of the same document through several users, the total weight is calculated from the following formula [2],[3],[4]:

$$Pk(UD_iG_j) = sum\ (UD_i, G_j, a) \quad (1)$$

In the case of a document that has no relevance notice by a user its recommendation weight pk is zero [2],[3],[4].

$$Pk(UD_iG_j) = 0 \quad (2)$$

In IAAS, a user's connection to the system is analyzed as a request to transmit recommendations. The recommendation transmitted to a user contains all the documents recommended to his group.

For each user group, after calculating the recommendation weight for each item, IAAS orders the list of documents based on the relevance value. The relevance value is expressed by the following formula [2]:

Relevance $P_{i,j}=ln(1+ Pk(UD_iG_j))$ *(3)*

The documents are ordered according to the decreasing values of $P_{i,j}$ and then reordered several times according to the users' profile to be personalized to each user of the group.

The notion of profile is very fundamental in IAAS. Indeed, in order to personalize the recommendations to the users, [2],[3],[4] have implemented the user profile and the document profile. These profiles are schemas that can be consulted in the works of [2], [3], [20].

2.2. Justification approaches

We present in this part of our work, the summary of the work already done on the approaches of justification in the systems of recommendations.

2.2.1. Approach Feature-Weighted Nearest Biclusters (FWNB)

The FWNB approach is built around four elements that are user group creation, keyword weighting, neighborhood formation, and recommendation and justification generation [5].

- *Creation of user groups*: it is based on the formation of user and item biclusters. The formation of these biclusters is done thanks to the similarity between users and items they have already evaluated. This bicluster formation is done automatically using the xMotif algorithm [6].

- *Weight of keywords*: The objective of [5] in constructing keyword weights is to find the distinct keywords that best describe the users' preferences. For this purpose, [5] used the similarity matrix between keywords and users. The weight of keyword f for a user u is calculated as follows:

$$W(u,f) = \mathrm{FF(u,f)} * \mathrm{IUF(f)} \quad (4)$$

$$IUF(f) = \log\frac{|\mathrm{U}|}{\mathrm{UF(f)}} \quad (5)$$

  |U| : total number of users and UF(f) : number of users in which the keyword f appears at least once.

  FF(u,f)=P(u,f) is the correlation between the user and the keyword f.

  Using the keyword user correlation matrix R_B (u,f), they generate the keyword weight matrix W_B from the formula W(u,f) [5].

- *Neighborhood formation*: This is the identification phase of the items and keywords to be recommended. All the items contained in the biclusters are candidates for recommendation as well as the keywords. Thus we determine the item and the justification for each user through the calculation of similarity between the user and his bicluster:
  $sim(u,b) = (1 - a) \cdot sim_I(u,b)+ a\ sim_F(u,b)$ *(6)*

simI(u,b): similarity between the user and his item; a∈[0, 1]; simF (u,b): similarity between the user and keyword

- *Recommendation and justification generation*: the generation of the item to be recommended as well as the corresponding keyword is done by simultaneous identification of the items in the neighborhood of the bicluster of a user u who :
  - ✓ are all preferred by other users according to the scores of the R_B matrix
  - ✓ contain the significant keywords of the matrix W_B.

  The generated recommendation is the form, 'item X is recommended to you because it contains the keyword f that you have already evaluated in item Y'.

2.2.2. Approach of Cataldo et al

[7] was interested in building an effective justification designed on the basis of the distinctive and relevant terms for the item starting from the users' reviews. For him, the effective justification must include the relevant and distinctive terms of the items that are discussed in the reviews. The approach of [7] is structured as follows: terms extraction, terms ranking, sentences filtering and the text summarization.

- *Feature extraction*: The first step is to identify the features that deserve to be included in the final justification. Thus the strategy of [7] takes as input a set of reviews R = {r1, r2 . . . rn} and produces a set of 4 -tuples ((ri, aij , rel(aij , ri), sent(aij , ri)). To extract the terms of the critics, [7] used the Kullback-Leibler divergence [8], which is a non-symmetric measure of the difference between two distributions to construct an algorithm.

- *Ranking the extracted terms*: [7] proceeded to calculate the score of the extracted terms by the following formula:

  $$score(aj) = \frac{\sum_{i=1}^{N} n_{aj,ri} * \mathrm{rel}(a_j, r_i) * sent(a_j, r_i)}{\mathrm{N}} \qquad (7)$$

  At the end, the terms are ranked in descending order and the K-first ones are labeled as main terms.

- *Sentence filtering*: After the identification of the terms, we proceed to a sentence filtering with the objective of filtering out the sentences that are considered not necessary in the final justification. To do this, we divide the criticisms ri ∈ R into sentences si1 . . . . sim. Then we check if the sentences respect in particular the criteria of content of the extracted terms. A top k of sentences are selected.

- *Text summarization*: the summary highlights the main contents of the item's reviews and maximizes both the coverage and the diversity of the justification while avoiding redundancy. The approach in [7] combines centroid-based text summarization [9], which has the advantage of being unsupervised, with a pre-trained neural language model, such as Word2Vec [10].

2.2.3. Approach of Jianmo et al

The generation of justifications for recommendations in [11] is done using a pipeline to identify candidate terms for justification and to form the users and item profiles from a corpus of reviews. The candidate terms are the reviews that the user had previously written. [11] constructs a dataset containing the custom's reviews for each user. The construction of the pipeline starts with the annotation of the reviews, then the classification of the annotated terms using the calculation of the distance between the selected terms and ends with the extraction of the justifications and the construction of the user and item profiles. Thus, for each user u we build the pipeline using the reference justifications D = (fd1; ...; dlr) consisting only of the justifications that the user had written.

Then we have the user profile composed of A = (a1; ... ; aK), we carry the most relevant ones and in the same way we build the items profile. For a user u and an item i as well as their reference justifications Du and Di, and the user profile Au and Ai of the item; we predict the justifications Ju;i = (w1;w2;...;wt) that explain why the item i is important for the user u. [11] identifies the terms or phrases using [13] and linguistic analysis. After identification, [11] uses the BERT [14] method for automatic classification of justifications and uses Fine-grained Aspect Extraction [15] for the extraction and the profile construction. The approach of [11] uses two models in its approach which are: Reference-based Seq2Seq Model and Aspect Conditional Masked Language Model. The experimentation has shown that the former produces high quality justifications and the latter produces diverse justifications.

2.2.4. Approach of Arpit et al

The explanation-based recommendation is a new approach that unifies recommendation and explanation. The recommendation is modeled as a path finding problem in the item-item similarity graph [12].

Once a chain has been constructed for each candidate item, the top-n chains are iteratively selected based on their total coverage of the candidate item terms and their dissimilarities with other top-n chains. The approach of [12] is built on the generation of explanation chains and the evaluation of this chain.

Generation of explanation chains*:*

$$rwd(j,i,C) = \frac{|(f_j \setminus \mathrm{covered}(i,C)) \cap f_i|}{|f_i|} + \frac{|(f_j \setminus \mathrm{covered}(i,C)) \cap f_i|}{|f_j|} \quad (8)$$

Evaluation of explanation chains:

$$score(\{C,i\}, C^*) = \frac{\sum_{j \in C} rwd(j,i,C)}{|C|+1} + \frac{|C \setminus \cup_{j' \in C*} J'|}{|C|+1} \quad (9)$$

Then comes the selection of items to be presented to the user. The technique of [12] does not compute separately the selection of items and justifications.

2.2.5. Approach of Or Biran et al

[16] proposes an automatic prediction method using machine learning to produce simple, short, quality natural language justifications; through the use of application domain critics [8]. This approach has a message prediction structure and architecture.

[16] uses the Semantic Template Typed (STT) message structure which is a small semantic network of typed entity slots and relationships for prediction. A set of STTs have been created for each justification domain and use specific STTs to a domain, as well as template sets, from text corpora for words extraction.

The template construction architecture of [16] consists of term selection and characterization, computation of certain quantities if any, and justification planning.

3. Architecture of Recommendation Justification in IAAS

The figure 1 below expresses the idea of how the recommendation justification in IAAS that we want to implement works. The diagram shows three main entities which are the users, the workstation and the Information Access Assistance Service.

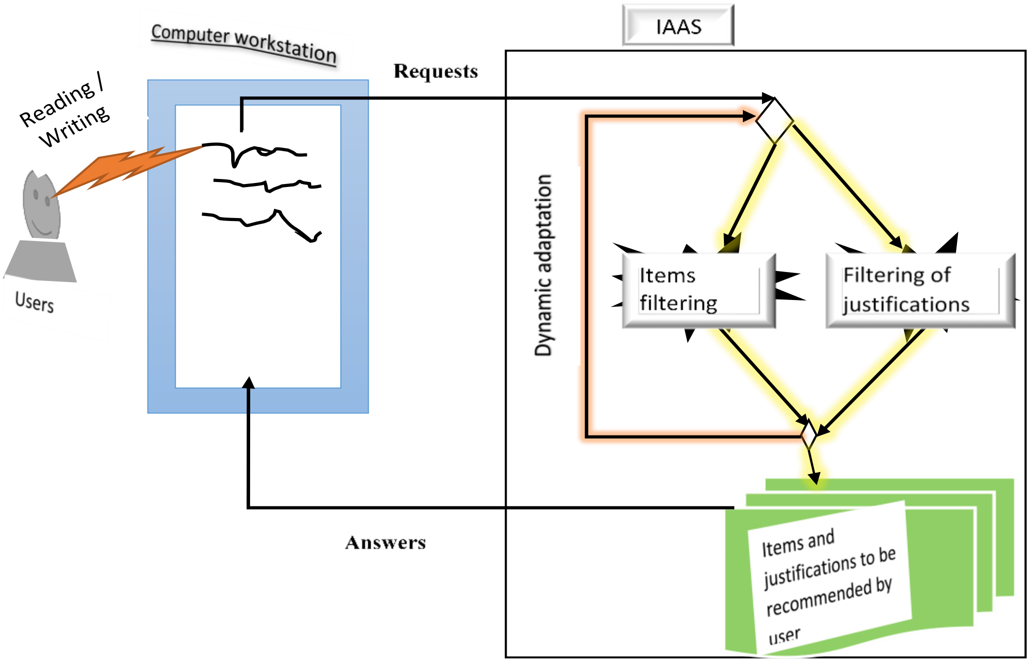


Figure. 1. Description of the recommendation justification in IAAS

Users give their opinions on the documents and at the same time note the justification of its opini on which is sent to IAAS for processing. In IAAS, we have separated the processing of the filtering of the documentary units and the justifications because the justifications are only texts that we want to process whereas the mechanism of filtering of the documents already developed in [2] is a manipulation of the numbers. After the separate processing of document and credential selections we collect them and then customize them to each user as [2] provides. Once a user logs in, the system returns his recommendations and it is still possible for him to make a new recommendation. In case a user recommends the same document to the same user group, the system uses the ranking already obtained to make a new readapatation of recommendation. This means that the recommendation is dynamic.

4. Synthesis of Justification Approaches in Recommender Systems

The table below summarizes the work already done in the area of justifications in recommender systems. The automatic summarization method column was realized after consulting [17].

4.1. Summary table

Table 1. Comparison of Some Justification Approaches in Recommender Systems

Justifying	**Methods**	**Justification filtering mechanism**	**Automatic text summarization methods**	**Justification style**
User-evaluated data [5]	Creation of groups Weight of terms Formation of neighbors [5]	Features weight Features frequency [5]	Digital approach: Learning-based methods	Keywords and influence [5]
Users reviews [7]	Words extraction Words classification Sentence filtering Text summarization [7]	Kl-divergence Term score Term ranking Sentences extraction [7]	Digital approach: Methods based on statistical calculations	Summary of reviews [7]
Key words written on the items [11]	Extration of explanation chains Evaluation of explanation chains [11]	Construction of item chains [11]	--------	Keywords [11]
User-evaluated data [12]	Pipeline Dataset [12]	Annotation, classification and extraction of terms [12]	Digital approach: Learning-based methods	Summary of reviews [12]
Corpus of texts [16]	Selection and characterization of terms Planning of the justification [16]	STT : it is used to predict the message [16]	Symbolic approach: Learning-based methods	Content and influence [16]

4.2. Positioning for justification in IAAS

Our analysis is conducted based on the IAAS literature review, the definition of recommendation justification in IAAS and the summary table of justification approaches. As in IAAS users give their opinions on the documents so they must provide their reviews at the same time. Also the fundamental concept in IAAS is that the user gives a weight to each item that is used to manage and filter documents. So it is better that we use these same weights to add to the user reviews that will be used as justification filtering strategy. As filtering mechanisms for existing evidence from the table, only [5] uses the weight of reviews and of the frequencies to filter the justification but does not use user reviews as justification. On the other hand, in the case of the approaches studied [16], it applies only in expert system, whereas IAAS is not an expert system, so this strategy does not interest us in this work. [11] cannot be used because the keywords written on the items are used to filter and order the list of recommendations to justify. The approaches of [5] and [12] use the evidence already assessed by users to automatically generate new justified recommendations using the numerical approach and learning based methods as a text summary tool.

The approach of [7] seems to have a very similarity because for [7], the reviews are entered by the users and the system collects all the reviews and then proceeds to process it separately and personalizes the recommendations. Only that in [7] the reviews are not accompanied by weight. We also have the summary method which is focused on statistical calculations which will be used in the case of IAAS since our reviews will carry weights. The test domain of [7] is different from that of IAAS because [7] is used in the domain of cinema while IAAS is used on the documents, videos, audios and images.

Therefore, of all the approaches that we studied, there is no case for giving weight to reviews. Based on this comparison, we are going to set up a new approach of recommendation justification which will take into account the opinions of users. Instead of a notice being a triplet as proposed by [2], it must be a quadruplet to take the justifications written by the users. Thus for the justification in IAAS a notice is now a quadruplet noted A=(UD,G,J,a) where J represents the justification. This message is sent to IAAS for processing as shown in figure 1.

This study does not question what is done on IAAS but aims to improve it by adding the justification. We will work on keeping the item selection technique and similarly we will develop a module for the processing of justifications.

5. CONCLUSIONS

In this paper, a literature review has been conducted on the different approaches to justification of recommendations and the Information Access Assistance Service. Then, a comparative study of these approaches allowed us to position ourselves on the justification approach in IAAS. Our contribution lies in the fact that at the end of our study we propose that the notices in IAAS should be a quadruplet instead of a triplet [2]. It also appears from our study that no approach can be used properly with IAAS, hence our immediate perspective to propose a specific justification approach to IAAS and then implement this approach.

REFERENCES

[1] Cataldo M., Alain D.S., Christoph T., Amon R., and Giovanni S. 2021. Exploring the Effects of Natural Language Justifications in Food Recommender Systems. In Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP ’21), June 21–25, 2021, Utrecht, Netherlands. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3450613.3456827

[2] Kabore, K. , Peninou, A., Sié, O. , Sèdes, F. Implementing The Information Access Assistant Service (IAAS) For An Evaluation. Int. J. Internet Technology and Secured Transactions , Vol. 6, No. 1, 2015 (2015)

[3] Kabore, K. , Sié, O. , Sèdes, F. Information Access Assistant Service (IAAS). In The 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013), IEEE UK/RI Computer Chapter, London, UK, December 9-12, (2013).

[4] Kiswendsida Kisito Kaboré : Système d’aide pour l’accès non supervisé aux unités documentaire. Thèse de doctorat du l’Université de Ouaga 1 Pr Joseph KI-ZERBO, Janvier 2018.

[5] Panagiotis S., Alexandros N., and Yannis M. Providing Justifications in Recommender Systems. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 38, NO. 6, NOVEMBER 2008 pp. 1262—1272 https://ieeexplore.ieee.org/abstract/document/4648950

[6] T. Murali and S. Kasif, "Extracting conserved gene expression motifs from gene expression data," in *Proc. Pacific Symp. Biocomputing Conf.*, 2003, vol. 8, pp. 77–88.

[7] Cataldo M., Gaetano R., Marco de G., Pasquale L., Giovanni S. Natural Language Justifications for Recommender Systems Exploiting Text Summarization and Sentiment Analysis . 2019 pp.63-73. http://ceur-ws.org/Vol-2495/paper8.pdf

[8] Or B. and Courtenay C. 2017. Explanation and Justification in Machine Learning: A Survey. In IJCAI-17 Workshop on Explainable AI (XAI), VOL. 8, NO.1. pp 1- 13

[9] Radev, D.R., Jing, H., Sty, M., Tam, D.: Centroid-based summarization of multiple documents. Inf. Process. Manage. 40(6), 919–938 (2004)

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. pp. 3111–3119 (2013)

[11] Jianmo N., Jiacheng L., and Julian M.. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In 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. 188–197.

[12] Arpit R. and Derek B. 2017. Explanation Chains: Recommendation by Explanation. RecSys '17 Poster Proceedings, Como, Italy, August 27–31, 2017, 2 pages.

[13] Mann, W.C. and Sandra A. T. (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization." Text 8 (3): 243-281.

[14] Jacob D., Ming-Wei C., Kenton L., and Kristina T. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019. pp. 2324—2335.

[15] Yongfeng Z., Guokun L., Min Z., Yi Z., Yiqun L., and Shaoping M. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval , pp. 83-92. ACM

[16] Or B., Kathleen M. Generating Justifications of Machine Learning Predictions. 1st International Workshop on Data-to-text 2015 http://www.cs.columbia.edu/~orb/papers/d2t_2015.pdf

[17] Mohamed Hédi Maâloul. Approche hybride pour le résumé automatique de textes. Application à la langue arabe. PhD thesis, pp 17- 43, 18 décembre 2012. https://tel.archives-ouvertes.fr/tel-00756111/file/These.pdf

[18] Mustafa B. and Raymond J. M.. 2005. Explaining recommendations: Satisfaction vs. promotion. In Beyond Personalization Workshop, IUI, Vol. 5. 153. pp 1,pp 7.

[19] Roza Lémdani. Système Hybride d'Adaptation dans les Systèmes de Recommandation. Thèse de Doctorat de l'Université Paris-Saclay préparée à CentraleSupelec. pp 23- 33. 11 juillet 2016. https://www.theses.fr/2016SACLC050.pdf

[20] Kyelem Y., Kabore K.K., Bassole D. (2022) Hybrid Approach to Cross-Platform Mobile Interface Development for IAAS. In: Shakya S., Bestak R., Palanisamy R., Kamel K.A. (eds) Mobile Computing and Sustainable Informatics. Lecture Notes on Data Engineering and Communications Technologies, vol 68. Springer, Singapore. https://doi.org/10.1007/978-981-16-1866-6_16

AUTHOR INDEX