TWO LEVEL SELF-SUPERVISED RELATION EXTRACTION FROM MEDLINE USING UMLS

Huda Banuqitah, Fathy Eassa, Kamal Jambi and Maysoon Abulkhair

Faculty of Computing & Information Technology,
King Abdulaziz University, Jeddah- Saudi Arabia

ABSTRACT

The biomedical research literature is one among many other domains that hides a precious knowledge, and the biomedical community made an extensive use of this scientific literature to discover the facts of biomedical entities, such as disease, drugs,etc.MEDLINE is a huge database of biomedical research papers which remain a significantly underutilized source of biological information. Discovering the useful knowledge from such huge corpus leads to various problems related to the type of information such as the concepts related to the domain of texts and the semantic relationship associated with them. In this paper, we propose a Two-level model for Self-supervised relation extraction from MEDLINE using Unified Medical Language System (UMLS) Knowledge base. The model uses a Self-supervised Approach for Relation Extraction (RE) by constructing enhanced training examples using information from UMLS. The model shows a better result in comparison with current state of the art and naïve approaches.

KEYWORDS

Relation Extraction, Self-supervised, Machine Learning, Knowledge base.

1. INTRODUCTION

In the last two decades, usage of medical computing systems showed an explosive growth. The vast amount of Information they store potentially contains new knowledge that can provide decision support to enhance the quality of medical care. An enormous database and repository of biomedical literature is available for researcher community and may contain the required knowledge. MEDLINE is one example of the online bibliographic database from a biomedical domain that contains more than 22 million biomedicine journal articles [1]. Knowledge Discovery from Databases (KDD) from such a biomedical courps as MEDLINE is a complicated process, and it takes several processes [2]. The efficient exploitations of these resources require Information Extraction (IE) techniques that transform unstructured information into the structured form. An example of such techniques is Relation Extraction (RE) which is an automatic mining of relations between biomedical entities in text. The extraction of the relationship between biomedical entities is the process to determine the semantic link between those entities and characterizing the nature of this relationship[1]. Recently RE has found growing interest among IE community and many studies focused on it because it helps to find new relations and interactions between biomedical entities from raw text and minimize intervention of a human resource. RE includes multiple techniques such as rule–based approach, Natural Language Processing (NLP) and Machine Learning (ML) methods [3, 4]. There are three

DOI : 10.5121/ijdkp.2016.6302

main types of RE approaches which are: Unsupervised method which needs no labeling, Supervised that uses a corpus of labelled data and Self-supervised that uses a small set of labelled examples. The Unsupervised method extracts strings of words that exist between the entities in large amounts of text, and then clusters and simplifies these word strings to produce relation. Unsupervised methods can use enormous quantities of data and extract very large numbers of relationships, but the resulting relations may not be easy to map to relations needed for a particular knowledge base.

On the other hand, Supervised relation extraction method uses ML techniques to address this problem, which requires a sufficiently annotated training data that consist of positive and negative examples. Furthermore, constructing the annotated data set for training is expensive, required expert knowledge and consumed plenty of time. Where the Self-supervised approach overcomes this bottleneck by using a significant knowledge base which contains information about the target relation to automatically annotate a data set. The main assumptions are the sentences contain an entity pairs either representing or not representing a relation will also express the relationship as well. Furthermore, Self-supervised approaches combine the advantages of supervised approaches, by including noisy pattern features in a probabilistic classifier, and advantage of Unsupervised methods, by extracting large numbers of relations from large corpora. It is generally believed that Self-supervised techniques would benefit the relation extraction in a generic domain. However, these techniques are not fully explored in the biomedical domain, because of two reasons. The first reason is that the main source of knowledge of Self-supervised approaches is Freebase, which is for the general domain and lack of biomedical knowledge. The second reason is, the developed Self-supervised learning models assume that each entity instance is independent which is violated and not applicable in the biomedical domain[5].

Thus, we proposed a biomedical relational model of Two-level for automatic Self-supervised Relation Extraction from Biomedical domain using UMLS. As we mentioned previously, KDD is iterative and interactive multiphase processes that include different steps like data selection, data preparation and preprocessing, data transformation, Data Mining (DM) and evaluation process. For that, we developed our model to discover knowledge from the biomedical domain by integrated diverse techniques for data mining including Self-supervision, natural language processing, and machine learning to build a Relation Extraction system from MEDLINE that requires minimal supervision using Unified Medical Language system (UMLS[i]).

UMLS is a collection of files and software that include different biomedical knowledge base and vocabularies. Metathesaurus is a database in UMLS that contains millions of health and biomedical related concepts names and the relationship between them. All the concepts categorized by their semantic type[ii] and all names of the concept are unified by Concept Unique Identifier (CUI). MRREL[iii] is a subset of the Metathesaurus and contains different relationships between different biomedical concepts defined by a pair of CUIs.

The aim of this paper is to enhance the Self-supervised Relation Extraction in MEDLINE biomedical domain by using semantic types of entities from UMLS knowledge base to construct training examples. Our contribution to the overall solution, based on a mature architecture and with a proof-of-concept implementation by using a semantic type of concept to construct the examples of the relation of interest that as training examples of the Self-supervised approach that improves the result comparing with others in term of Precision, Recall and F-Score performance.

The rest of the paper organized as the follow. Section 2 presents the related work of the study, while section 3, describe the details of our knowledge discovery model. Section 4 shows the experiments procedure, with methods of training and test set construction. Section 5 shows the results and the discussion while the final section is the conclusion and future work.

2. RELATED WORK

Current biomedical research needs to exploit the enormous amount of information reported in the scientific literature using DM techniques. In particular, those techniques aimed at finding relationships between entities, which is the key for identification of actionable knowledge from these kinds of literature which are called Relation Extraction (RE). This section presents the different efforts that have been achieved in RE from the biomedical domain which used Self-supervised approach.

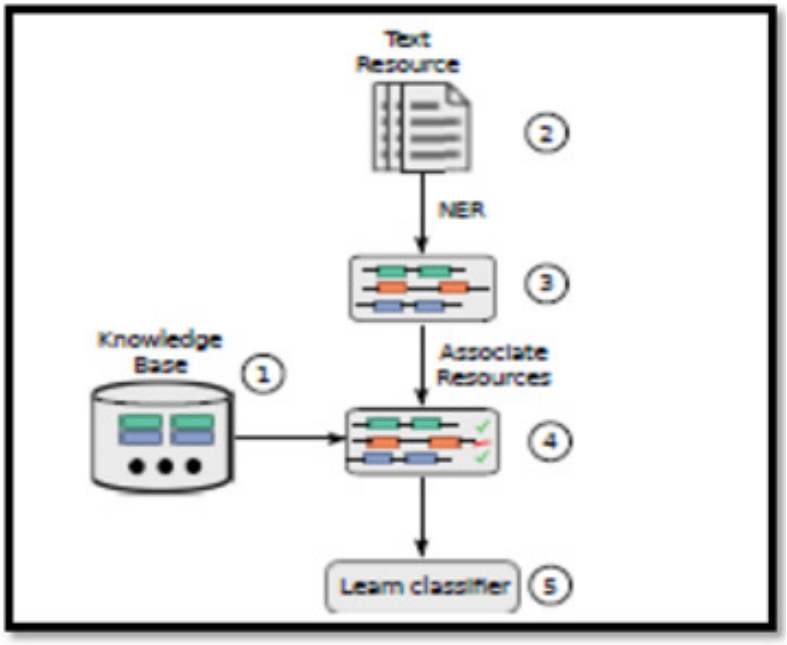

Figure 1: Self-Supervised workflow[6].

The author in [6] presented a general distant supervision approach for relationship extraction as shown in Figure 1. The details are summarized in the following steps:

1. Identify a knowledge base which includes pairs of entities about the relationship-type in question (*e.g.,* PPI-database).

2. Compile a large text (not annotated) resource relevant for the target domain (*e.g.,* MEDLINE abstracts).

3. Recognize and normalize related named entities (*e.g.,* protein names).

4. Associate entity-pairs from the knowledge base with previously identified instances in the text corpus.

5. Entity pairs contained in the knowledge base are labelled as positive examples. Negative examples are labelled by following the closed world assumption. The closed world assumption states that entity pairs lacking in the knowledge base do not feature the relationship type in question.

There are limited works which used Self-supervised approaches in biomedical domain, the authors in [7] proposed a Self-supervised relation extraction by using Yeast Protein Database (YPD) that contains subcellular location fields for many proteins, they collected a set of instances of subcellular locations of proteins from the Yeast Protein Database and then identified sentences from the associated PubMed abstracts in order to get an annotated corpora.

By using the coordination structure of an entity in the sentences, [5] developed a distant supervised model that combine the results from open information extraction techniques, to perform a task of relation extraction from biomedical literature. The model incorporates a grouping strategy to take into consideration the coordinating structure among entities co-occurred in one sentence. They apply the approach to extract gene expression relationship between genes and brain regions from literature. The Results showed that the methods can achieve better performance over baselines of Transductive Support Vector Machine and with non-grouping strategy.

In [8] the authors use Self-supervision learning to train a classifier for Protein-Protein Interactions (PPI). They use a Support Vector Machine (SVM) classification algorithm with a shallow linguistic kernel as a classifier. The knowledge about interacting proteins is taken from the IntAct database.

Using UMLS as Knowledge base, the authors in [9] proposed a Self-supervised relation extraction from the biomedical domain in MEDLINE abstracts using UMLS to annotate automatically the training data which is then used to train the classifier. To generate the training examples with positive and negative examples, all CUI pairs for the target relation are extracted from MRREL and considered as a set of positive instance pairs. Thus, the occurrence of a positive entity pair in a sentence will represent the relation of interest. Any CUI pairs which also occur in another MRREL relations are removed from the list of positive instance pairs. In contrast, negative examples will be generated based on the positive instance pair set; new CUI pair combinations will be generated by combining all CUIs from the first position with all CUIs from the second position. Only if a newly generated CUI pair is not in the positive list and not contained in another MRREL relation, then it will be used as negative instance pair. The model evaluated using two techniques Held-out and manual evaluation. On manual evaluation, the relation classifier was trained using (may_treat) relation that created using Self-supervised process and evaluated by using manually annotated corpus using test data set, and the result outperforms naïve approach with an F-Score of 0.571, 0.600 Precision and 0.545 Recall. The result indicated that UMLS is a useful resource for Self-supervised relation extraction. Also by using UMLS to train a distant supervised relational classifier, [10] presented the first results using UMLS knowledge base and the model evaluated using existing evaluation data sets since there were no resources directly annotated with UMLS relations available. Their results showed that using a distantly supervised classifier trained on MRREL relations similar to those found in the evaluation data set provides promising results.

The authors in [11] demonstrated the potential of distant learning in constructing a fully automated relation extraction process. They produced two distantly labelled corpora for protein–protein and drug to drug interaction extraction, with knowledge found in databases such as IntAct for genes and Drug Bank for drugs. They labelled approximately 50,000 MEDLINE abstracts using the shallow linguistic classifier trained on a distantly labelled corpus. In other words, the classifier trained on five manually annotated corpora and the same classifier trained on a distantly labelled corpus agree on 86.4 % of all 50,000 predictions.

There are some works done in Self-supervised approach outside the biomedical domain. Mintz and others in [12] use Freebase to provide distant supervision for relation extraction. They utilized a similar heuristic by matching Freebase tuples with unstructured sentences from Wikipedia articles in their experiments to create features for learning relation extractors. Matching Freebase with arbitrary sentences instead of matching Wikipedia infobox with corresponding Wikipedia articles will potentially increase the size of matched sentences at a cost of accuracy. They conclude that their results suggest that syntactic features are indeed useful in distantly supervised information extraction. Also, the authors [13] used Freebase knowledge base to annotate the New York Times corpus with the entity pairs. They focused on the three relations which are nationality, place of birth, and contains. To train the classifier, the authors introduced the usage of a multi-instance learning approach for this context. In contrast, the authors in [14] annotated the information in the articles of Wikipedia using the infoboxes of Wikipedia as a knowledge source.

3. MODEL DESCRIPTION

3.1. Model Architecture

The First level deals with data extraction, preparation and relation examples extraction by using UMLS knowledge base. The second level deals with feature extraction, constructing the training set and then train and evaluate the classifier using SVM classification to extract the relation of unlabelled data, SVM found to be the best and for multiple classification problems[15].

As shown in Figure 2 of our model architecture, the user enters the query using the system interface then the model coordinates execution of user query and displays the result. Then, the sentences that contain a pair of entities with a semantic type that matches the user query are retrieved. We use MySQL database in our model to store and query annotated sentences in addition to retrieve relation examples that match user query from the UMLS knowledge base.

The features extraction from the sentences is done by tokenizing, lemmatizing and parsing sentences; we built our model based on the CoreNLP library [16]. Finally, constructing the training examples that will be used by the classifier to train the data set using the Linear SVM classification algorithm and then extract relation for the unlabelled sentences.

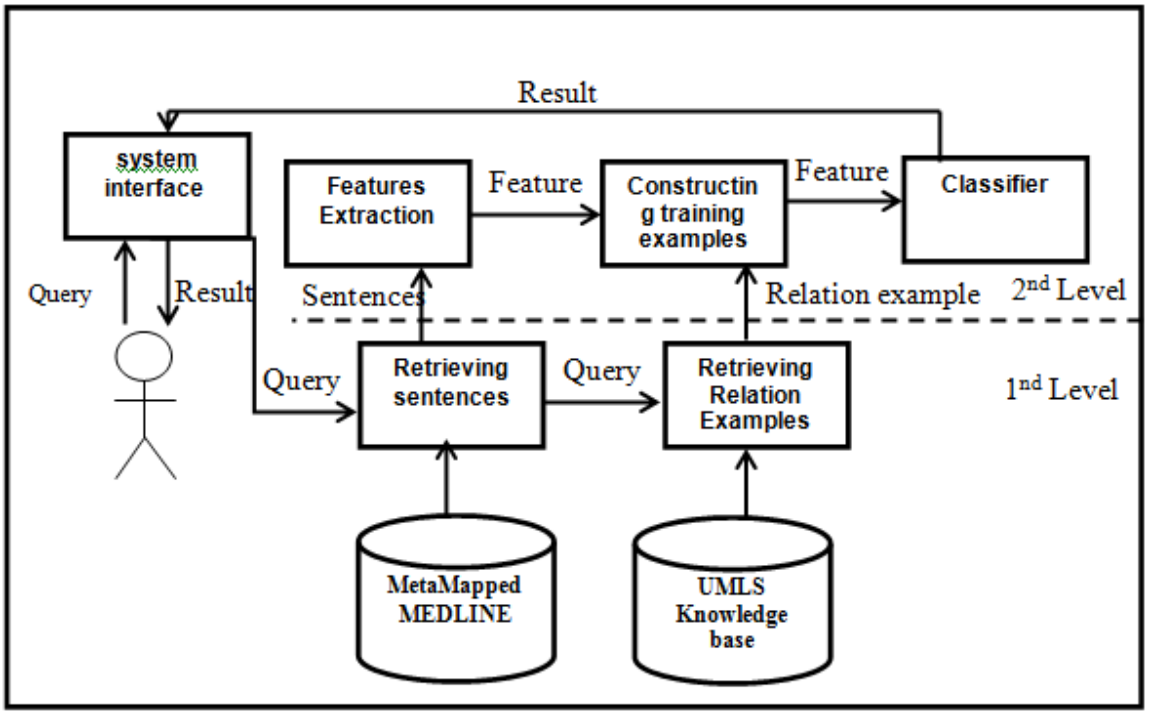

Figure 2: Model Architecture

3.2 Model Workflow

The whole process of our approach can be detailed as follows:

3.2.1 Pre-processing

We used MetaMapped MEDLINE corpus[iv] as initial data, which consists of articles and abstracts of millions of research papers and publication from medicine. The sentences of MEDLINE contain the information of interest are used to generated the training data set examples for the Self-supervised model. Therefore, it is important to identify related information. The relations in UMLS are identified by a pair of Concept Unique Identifier (CUI), so we need a mapping of UMLS concepts to the MEDLINE sentences. For that, we used a MetaMapped MEDLINE, which is annotated by MetaMap tool[v]. Each sentence in MEDLINE annotated with UMLS concepts, and the annotations are represented in MetaMap machine output format[vi].

The whole MetaMapped MEDLINE is about 165 GB in size and can be downloaded in two ways: as one file of the entire corpus or as 779 small files each about 250 MB in volume. In our experiment, we used a subset of those files.

The basic unit of MetaMap machine output is an utterance that represents annotation for a single sentence. Each utterance consists of phrases – subset of initial sentence. For each phrase, MetaMap provides several different mappings. Each mapping consists of several entity values that represent matched concept and score assigned by MetaMap tool. The value of the assigned score is reflected the mapping confidence in which the lower the score value means higher confidence in mapping.

In more details we populated the database with only mappings that match our test user query and have the best score. We use two semantic types which are "**bacs**" and "**dsyn**" pairs, where "bacs" is Biologically Active Substance and "dsyn" is refer to Disease or Syndrome. The database was populated with more than 503151 sentences from initial 30 GB of MetaMapped MEDLINE file.

3.2.2 Feature extraction

We adopted the same features implemented by [10, 12, 17] because they clearly represent the relation between the entities in the sentence also they help in determining the accurate class of the relation between disease and treatment. The adopted features are: The sequence of words between entities, Post Of Speech tag (POS) of words between entities and Words on the semantic path between entities. For constructing the above lexical and syntactic features, we annotated each sentence in the training set with part of speech tags and dependency tree using Stanford CoreNLP library. Consider the following example sentence: "Multiple doses of METHOTREXATE used in the treatment of ECTOPICUPREGNANCY", where METHOTREXATE is the first entity in the sentence And ECTOPICUPREGNANCY is the second entity in the sentence .

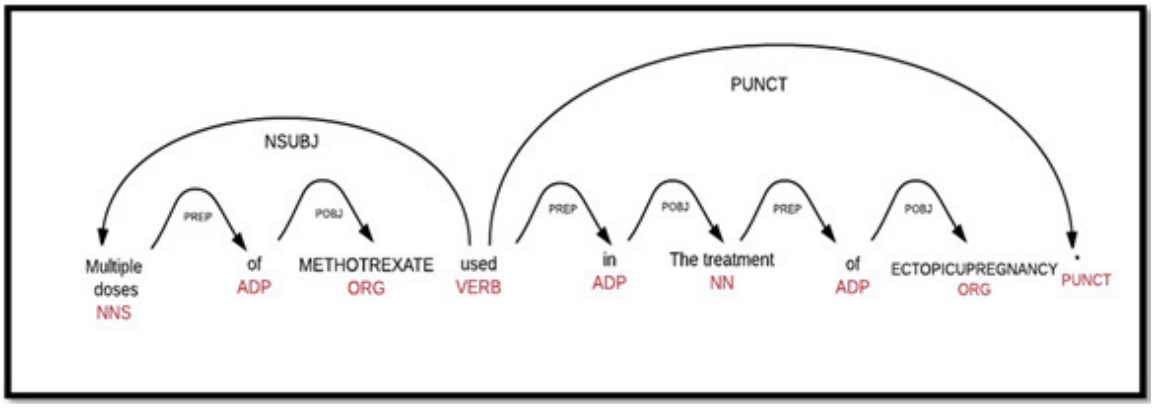

Figure 3: parsing tree of sentences notes that: NSUBJ is Nominal subject, PREP is Prepositional modifier, PUNCT is Punctuation, POBJ is Object of a preposition,NNS is Noun- plural, ADP is adpositions (prepositions and postpositions) and ORG – organization.

Figure 3 show the parsing tree of the sentences where:

1. Words between entities feature for example sentence will be: "used in the treatment of".

2. Words on the semantic path between entities. The semantic path is a path on the semantic graph (arrows on Figure 3). For example sentence, we start from first entity than go to by the arrow to word "of", than to "multiple doses", than to "used", than to "in", than to "the treatment", "of" and finally to second entity".

3.2.3 Data set annotations and relation extraction

For data set annotation with relations, we used MRREL relation subset of UMLS knowledge base. Each entry in MRREL contains a pair of entities and relation between them in form (CUI_1, CUI_2, relation name). The relation annotating algorithm works in two steps. First, it looks for all entries in MRREL that match the semantic types from user query. Second, it

matches those entries to sentences by CUI. If sentence contains the same pair of CUIs as MRREL entry it is annotated with corresponding relation. For relation extraction, we trained and evaluated our classifier using SVM classification algorithm on our data set to extract and predict the relations between disease and treatment for the rest of unlabelled sentences. The classification process is based on the extracted feature that we mentioned previously that constitute the features vectors used to train the classifier.

4. EXPERIMENT

4.1. Training set construction

The training set was constructed from sentences that matched MRREL relations with our own method as mentioned in section 3.2.3 and inspired by [9]. However, our model differs in two aspects: we used a semantic type of the entities to get all the relations between the biomedical entities in UMLS, and we used general relation examples that appear between our given semantic types to construct the negative examples. In contrast authors of [9] used only pairs that participate in ”may_treat” relation regardless of their semantic type.

To enhance the training set quality, we applied filtering by part of speech tag. MetaMap tool has a most common error that is annotating verbs or adjectives as if they were nouns as observed by manual check. Using CoreNLP library as in [16] we annotated each sentence in the training set with part of speech tags and threw away those sentences which concept was not marked as nouns.

For the training set labeling, all relations were divided into two groups: specific relations that labelled with “RO” in MRREL, where RO relation described as a relationship other than synonymous, narrower, or broader, and other than RO relation groups that represent more general relations. General relations were considered as negative examples for classification and labelled as "other." Sentences with multiple "RO" relations were not included in a training set because they could represent any of those relations but classifier needs the exact match with label and ground truth. We also discard non-frequent relations.

Another observation was that “RO may_treat” relation almost include“RO=may_prevent” relation and all most of the sentences labelled with “may_prevent” were also labelled with “may_treat”. Manual analysis showed that ground truth for such sentence could be either of both relations as shown in example 1 that the treatment “desferrioxamine” treats the” iron overload”, and they are indistinguishable by MRREL. We decided to unite such relations into one more general.

Example 1: [Intensified **desferrioxamine** (TREATMENT) treatment (by either subcutaneous or intravenous route) or use of other oral iron chelators, or both, remains the established treatment to reverse cardiac dysfunction due to **iron overload** (DISEASE)]

So from 503151, only 291575 sentences remain with a single pair of semantic type “dsyn-bacs” and 6699 were labelled with relation from MRREL. Then the left is 284876 unlabelled sentences. Final training set after removing the sentences with more than one relation consisted of 4171 (positive and negative) examples with specific relationship as follow:

- RO=null, 1493 examples
- RO=gene_product_malfunction_associated_with_disease, 1246 examples
- RO=related_to, 539 examples
- RO=may_treat, 311 examples
- Other, 582 examples.

Since our target examples of relation is "may_treat" we observed that "null" and "related_to" relations will not serve this relation between treatment and disease entities, if we consider example 2, we can observe that "METABOLIC SYNDROME" does not treat or prevent the disease "CHOLESTEROL", but they are related to each other in another way. For that, we exclude "null" and "related_to" examples from the training data set examples.

Example 2: [BACKGROUND: To establish the rate of agreement in predicting **METABOLIC SYNDROME** (TREATMENT) (ms) in different pediatric classifications using percentiles or fixed cut-offs, as well as exploring the influence of **CHOLESTEROL** (DISEASE)]

After excluding "related_to" and "null" relation from the positive training example, we got 2139 labelled examples.

4.2. Testing set construction

We used two test data sets for evaluation of our classifier. The first test set constructed by combining different relation mining data sets so that it could be similar to a training set. The second test set we used the same test set presented in [9] after their permission.

In the first test set, we employed three most specific and frequent relations: "may_treat", "gene_product_malfunction_associated_with_disease " and "other" to serve our training set that contains these relations. Further, we identify this data set as "Triple relation" test set (for simplicity). For this test set, 70 examples of "other" relation were labelled manually. 500 "may_treat" examples and 60 "other" examples were obtained from disease-treatment relations test set in[17]. 500 examples of "gene_product_malfunction_associated_with_disease" were randomly chosen among positive examples of gene-disease relation test set in [18].

The second test set from [9] contains 227 examples of "other" relations and 173 examples of "may_treat" relations. We will call this set "may_treat." test set. And science it is important to keep in the training set only those relations that presented in the test set, we exclude the relation "gene_malfunction_is_associated_with_disease" from the training set examples to evaluate using the test set "may_treat" from [9].

5. RESULT AND DISCUSSION

The Results of the model proposed by Roller and other in [9] are shown in Table1.

Table1. result of [9] on evaluation based on "may_treat" test set.

	Precision	Recall	F-Score
5,000 training instances	0.273	0.273	0.273
10,000 training instances	**0.600**	**0.545**	**0.571**
20,000 training instances	0.417	0.455	0.435

For evaluation we used the most common metrics that used in classifier evaluation in: Precision, Recall and F-Score which defined in equations (1), (2), and (3) respectively:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F - Score = 2 * \frac{precision*recall}{precision+recall} \quad (3)$$

Where: TP is the true positive results of classification, FP is the false positive results of classification and FN is the false negative.

By applying a different combination of features, that discussed in section 3.2.2. The model shows better results in comparison with [9] work, in term of Precision, Recall, and F-Score when using Linear SVM as the algorithm of classification and Words between entities as basic feature based on "Triple relation" test set as shown in Table 2. Furthermore, as shown in Table 3 and based on "may_treat" test set, the better result achieved in the term of Recall and F-Score when using Words between entities features with words on semantic path features using Linear SVM algorithm which outperform the best result in [9], this indicated the efficiency of the proposed approach in constructing training examples using semantic types of biomedical entities in UMLS.

Table 2. Model results of evaluation based on "Triple relation" test set.

Method	Accuracy	Average Precision	Average Recall	Average F-Score
Naive approach	0.38	0.37	3	0.31
Words between entities, using Linear SVM	**0.62**	**0.76**	**0.62**	**0.64**

Table 3. Model results of evaluation based on "may_treat" test set.

Method	Accuracy	Precision	Recall	F-Score
Naive approach	0.57	0.32	0.57	0.41
Words between entities + words on semantic path, using Linear SVM	0.60	0.54	**0.72**	**0.62**

Figure 4 shows the comparison between the result of the model based on "may_treat" test set by using entities semantic type, and the result in [9] which did not use semantic type in their approach. As we can see that the proposed model achieved best results among the other in term of Recall and F-Score.

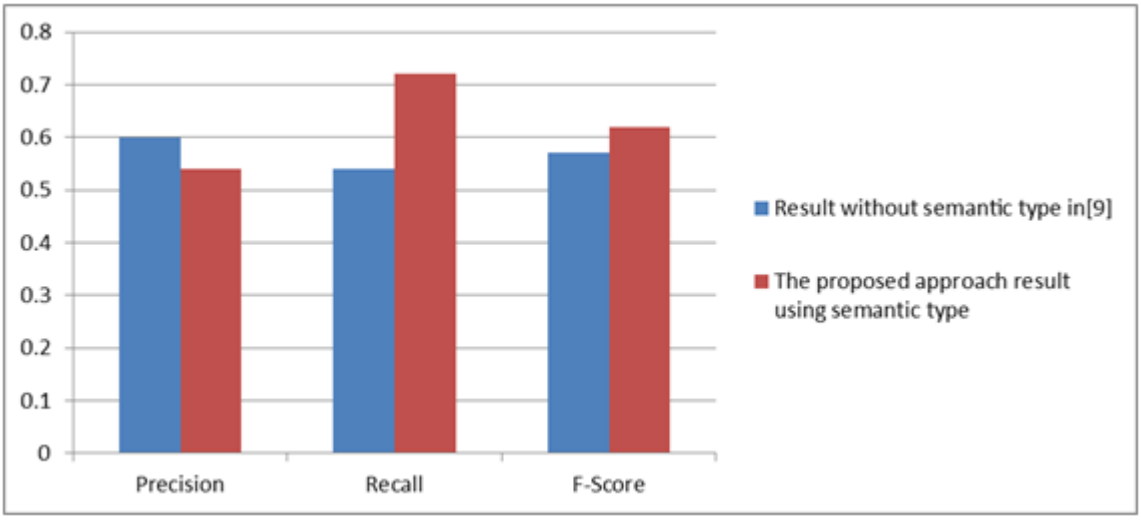

Figure 4: Comparison of Precision, Recall, and F-Score of proposed model and other.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose Two-Level Knowledge Discovery for Relation Extraction using UMLS knowledge base and demonstrated model performance on MEDLINE data. We used a Self-supervised approach for relation extraction by incorporate data mining and machine learning techniques. Additionally, we proposed our own approach to constructing the training set examples with positive and negative instances based on entities semantic type from MRREL section in UMLS. The approach achieved better result in term of Precisions, Recall and F-Score with 0.76, 0.62 and 0.64 respectively on "Triple relation" test set, and 0.72 of Recall and 0.62 F-Score on "may_treat" test set. The model also demonstrates an approach to minimize the cost of relation extraction by using a weekly labelled training example using UMLS. Our future plan is to deal with multiple data and knowledge sources by developing an algorithm for prioritizing relation examples from different corpus and knowledge base.

REFERENCES

[1] A. Bchir and W. B. A. Karaa, "Extraction of drug-disease relations from MEDLINE abstracts," in Computer and Information Technology (WCCIT), 2013 World Congress on, 2013, pp. 1-3.

[2] S. Benomrane, M. Ben Ayed, and A. M. Alimi, "An agent-based Knowledge Discovery from Databases applied in healthcare domain," in Advanced Logistics and Transport (ICALT), 2013 International Conference on, 2013, pp. 176-180.

[3] V. N. Romero, S. Kudama, and R. B. Llavori, "Towards the Discovery of Semantic Relations in Large Biomedical Annotated Corpora," in Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on, 2011, pp. 465-469.

[4] L. Yao, C. J. Sun, X. L. Wang, and X. Wang, "Relationship extraction from biomedical literature using Maximum Entropy based on rich features," in Machine Learning and Cybernetics (ICMLC), 2010 International Conference on, 2010, pp. 3358-3361.

[5] L. Mengwen, L. Yuan, A. Yuan, H. Xiaohua, A. Yagoda, and R. Misra, "Relation extraction from biomedical literature with minimal supervision and grouping strategy," in Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on, 2014, pp. 444-449.

[6] P. Thomas, "Robust relationship extraction in the biomedical domain," Mathematisch-Naturwissenschaftliche Fakultät, 2015.

[7] M. Craven and J. Kumlien, "Constructing Biological Knowledge Bases by Extracting Information from Text Sources," presented at the Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, 1999.

[8] P. Thomas, I. Solt, R. Klinger, and U. Leser, "Learning protein protein interaction extraction using distant supervision," Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing, pp. 34-41, 2011.

[9] R. Roller and M. Stevenson, "Self-supervised Relation Extraction Using UMLS," in Information Access Evaluation. Multilinguality, Multimodality, and Interaction. vol. 8685, E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, et al., Eds., ed: Springer International Publishing, 2014, pp. 116-127.

[10] R. Roller and M. Stevenson, "Applying UMLS for Distantly Supervised Relation Detection," in Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi), 2014, pp. 80-84.

[11] P. Thomas, T. Bobic, M. Hofmann-Apitius, U. Leser, and R. Klinger, "Weakly Labelled Corpora as Silver Standard for Drug-Drug and Protein-Protein Interaction," Third Workshop on Building and Evaluating Resources for Biomedical Text Mining Workshop Programme, p. 63, 2012.

[12] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labelled data," presented at the Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, Suntec, Singapore, 2009.

[13] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labelled text," presented at the Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III, Barcelona, Spain, 2010.

[14] R. Hoffmann, C. Zhang, and D. S. Weld, "Learning 5000 relational extractors," presented at the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 2010.

[15] O. Frunza, D. Inkpen, and T. Tran, "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts," IEEE Transactions on Knowledge and Data Engineering, vol. 23, pp. 801-814, 2011.

[16] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in ACL Demonstrations, 2014.

[17] B. Rosario and M. A. Hearst, "Classifying semantic relations in bioscience texts," presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, 2004.

[18] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research," BMC Bioinformatics, vol. 16, pp. 1-17, 2015.

AUTHORS

Huda Banuqitah is Teaching Assistant in Information Technology Department , Faculty of Computing and Information Technology, King Abdulaziz University, where she is Currently Master Student .

Fathy Essa :received the B.Sc degree in electronics and electrical communication engineering from Cairo University, Egypt in 1978, and the M. Sc. degree in computers and Systems engineering from Al Azhar University, cairo, Egypt in 1984, and Ph.D degree in computers and systems engineering from Al-Azhar University , Cairo, Egypt with joint supervision with University of Colorado, U.S.A, in 1989. He is a full professor with computer Science dept, Faculty of Computing and Information technology, King Abdullaziz University, Saudi Arabia. His research interests include agent based software engineering, cloud computing, software engineering, big data, distributed systems, exascale system testing

Kamal M Jambi received the B.Sc degree with honor Computer Science from University of Petroleum and Mineral, KSA in 1982, and the M. Sc. from Michigan State University, MI, USA in 1986, and Ph.D degree from Illinois Institute of Technology, IL, U.S.A, in 1991. He is a full professor with Computer Science dept, Faculty of Computing and Information technology, King Abdullaziz University, Saudi Arabia. His research interests include OCR, NLP, Image Processing, software engineering, big data, distributed systems

Maysoon Abulkhair is an Assistant Professor and the Supervisor of IT department at King Abdulaziz University, Jeddah, KSA. The major interested research field is HCI, associating it with different knowledge area such as artificial intelligent, machine learning, and data mining.

---

[i] Unified Medical Language System (UMLS)
[ii] Semantic Types of UMLS Concepts
[iii] MRREL table description

[iv] - http://ii.nlm.nih.gov/MMBaseline/
[v] https://metamap.nlm.nih.gov
[vi] https://metamap.nlm.nih.gov/Docs/2012_MMO.pdf