A Verb-based Algorithm for Multiple-Relation Extraction from Single Sentences

Qi Hao[1], **Jeroen Keppens**[1], **and Odinaldo Rodrigues**[1]
[1]Department of Informatics, King's College London, London, United Kingdom

Abstract— *With the growing number of unstructured articles written in natural-language, automated extraction of knowledge, such as associations between entities, is becoming essential for many applications. In this paper, we develop an automated verb-based algorithm for multiple-relation extraction from unstructured data obtained on-line. Named Entity Recognition (NER) techniques were applied to extract biomedical entities and relations were recognized by algorithms with Natural Language Processing (NLP) techniques. Evaluation based on F-measure with a random sample of sentences from biomedical literature results an average precision of 90% and recall of 82%. We also compared the performance of the proposed algorithm with a single-relation extraction algorithm, indicating improvements of this work. In conclusion, the preliminary study indicates that this method for multiple-relation extraction from unstructured literature is effective. With different training dataset, the algorithm can be applied to different domains. The automated method can be applied to detect and predict hidden relationships among varying areas.*

Keywords: Multiple-relation extraction, Natural Language Processing (NLP), Named Entity Recognition (NER), verb-based algorithm

1. Introduction

A substantial amount of valuable knowledge is recorded in the form of unstructured text data, such as news, emails, journal articles and conference papers. The biomedical literature is one body of knowledge of this form. Although text documents provide an effective way to disseminate knowledge within a relatively small community or narrow field of study, it becomes very hard or impossible for humans to fully comprehend all the knowledge comprised in this form within much larger communities or within collections of related disciplines and specialties. This paper contributes to ongoing efforts to develop mechanisms for automated knowledge extraction from text data. In this work, we propose a verb-based algorithm to extract multiple relationships between entities from unstructured articles written in natural language. The algorithm was evaluated by an experiment using biomedical literature to extract bio-entities, including substance, symptom, disease and body part, and relations between them. The performance of proposed algorithm was also compared against single-relation extraction algorithm.

The ultimate goal of relation extraction is to construct networks from text data that indicate various associations among different entities across different areas. For example, the sentence *"The quality of magnesium status directly influences the Biological Clock function (BC)"* contains a relation between *magnesium* and *Biological Clock function*. The first step is to enable computers to analyze words and sentences through Natural Language Processing (NLP) techniques. Several open libraries and toolkits were developed recently, such as Stanford's CoreNLP [1] and OpenNLP [2]. They provide a rich set of tools common NLP tasks such as tokenization, lemmatization, part-of-speech (POS) tagging, parsing, etc. In addition, two essential steps – entity recognition and relationship extraction – have recently seen tremendous progress. Existing Named Entity Recognition (NER) tools can recognize not only general terms such as proper nouns, but also more specific entities such as diseases and symptoms [3], [4], [5], [6]. As for relation extraction, five main methods are currently used: extraction based on *co-occurrence*, *link-based* extraction, extraction using *machine learning*, *rule-based* extraction and *verb-based* extraction [7], [8]. The first four of these methods can deal with simple relations between two entities connected by some target words with relatively low precision and recall. Verb-based methods on the other hand normally have higher precision and can be applied in a variety of domains. However, current verb-based approaches can only extract a single relation embedded in a sentence composed of a verb phrase sandwiched between two entities of interest. If the sentence contains multiple relations, then existing verb-based algorithms can only extract one of these relations. More details about these five methods are discussed in Section 5.

In this work, we propose a verb-based algorithm using existing NLP techniques and NER tools to extract relations from text data, including multiple relations embedded in the same sentence. Data was automatically downloaded and then processed using standard NLP techniques to extract the entities. Subsequently, instead of identifying target verb phrases by POS tagging and parsing alone, we extract verb relations using semantically similar verbs. Single-relation extraction algorithms work well when extracting simple co-occurrence relations such as *Entity-Verb-Entity*. However,

we enhanced this process so that it can deal with three common sentence structures which embed multiple relations within a single sentence. By analyzing the structure of the clauses, our algorithm is able to extract multiple verb-based relations connected by a relative pronoun such as *which* or *that*. Similarly, an analysis of the sentence level conjunctive structure, allows the algorithm to extract multiple verb-based relations connected by conjunctions such as *and* or *but*. Finally, by analyzing the phrase level conjunctive structure, the algorithm is also able to extract one-to-many or many-to-one relations. In our experiments, our algorithm achieved an average precision of 90% and a recall rate of 82%. It is worth mentioning that our algorithm is not restricted to a fixed domain or a particular set of verbs.

This rest of the paper is organized as follows. Section 2 presents a brief review of the relevant NLP techniques and NER tools used to pre-process the texts. Section 3 explains the proposed algorithms and experiments. Section 4 provides an evaluation of the proposed algorithms. In Section 5 we discuss some related works and this is followed by some conclusions and future works in section 6.

2. Background

In order to process and analyze texts written in natural language, NLP techniques should be introduced. NLP is a cross disciplinary field in artificial intelligence and computational linguistics, investigating ways to enable computers to interact with humans and understand human natural languages. Some standard techniques include word segmentation, POS tagging, word sense disambiguation, parsing, and NER. Word segmentation enables computers identifying and extracting valid words from a continuous stream. POS tagging helps computers to classify words into categories such as noun, verb, and adjectives. Parsing determines structure of the sentences based on POS tags. NER helps computers to recognize and classify named entities that are rigid designators [9] in texts into pre-defined categories such as proper names of persons, organizations, and certain biological species and substances.

Many open source toolkits and libraries for NLP techniques have been developed these days. OpenNLP is a machine learning based toolkit that has been widely used for standard NLP tasks [10], [11], [12]. OpenNLP provides a command line script and an API as well. It can also be used as a package in Java program or R program.

Although OpenNLP can perform simple NER tasks such as recognizing person and company names, locations or times, NER remains a crucial and complex task for biomedical domain due to the complexity of bio-entities and lack of human-annotated data. Many NER systems have been designed to recognize and extract biomedical substances and diseases by supervised learning techniques, such as LingPipe [3], MetaMap [5] and Abner [4]. LingPipe mainly focuses on gene entities recognition using the GENETAG corpus for training. MetaMap is a highly developed software to map biomedical text to the UMLS Metathesaurus or equivalently, to recognize and extract Metathesaurus concepts in biomedical publications. Abner has multiple models for recognizing protein, DNA, RNA, cell line and cell type with the NLPBA and BioCreative corpora. Based on those existing approaches, the proposed methodology is discussed in Section 3.

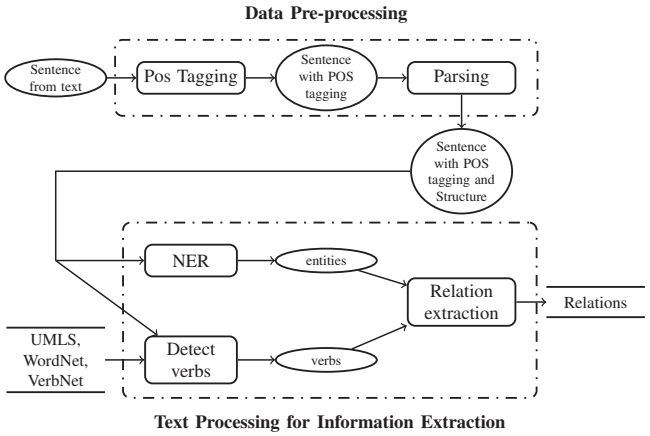

Fig. 1: Overview of a single iteration of the extraction process

3. Methodology

The main objective of this work is to extract relationships between entities from biomedical publications. The input of the system is a set of publication texts regarding some particular concepts. The texts are analyzed using standard NLP techniques and entities and verb relations are recognized and extracted by the proposed algorithms.

The system starts by identifying and extracting PubMed publication abstract records that contain the target biomedical terms such as "magnesium deficiency", "migraine attack". Those records are stored in text files with their corresponding PubMed IDs (PMID), titles and abstracts. Separated sentences from texts are the input of one iteration of the algorithm. An overview of a single iteration process is shown in Figure 1. The proposed algorithm is divided into two main tasks: Data pre-processing and text processing for information extraction. For data pre-processing, standard NLP techniques are applied in the selected sentences for POS tagging and parsing. As for text processing for information extraction, NER identifies relevant entities and a verb-detection algorithm identifies relevant verbs. Finally the system outputs the relations into a list of binary relation in the form of *Entity | Verb relationship | Entity*. In the following sections, we describe each of these steps in detail.

3.1 Data Collection

The biomedical literature data in PubMed consists of more than 26 million citations from MEDLINE, life science journals, and on-line books. The data pre-processing should be only focusing on papers that contain valued relationships between certain medical substances, diseases, or symptoms. Therefore, a automatic procedure was developed written in R with the help of an open source library called RISmed [13]. Both abstracts and full texts of biomedical articles are available on Pubmed. In this work, we only focused on abstracts because they conclude most of the important relations embedded in the whole texts. All the downloaded records with the PMIDs, titles and abstracts are in text formats so that are ready to be analyzed by NLP techniques.

3.2 Data Pre-processing

The original set of downloaded texts should be reduced based on the relevance for dimensional reduction. Therefore, some standard NLP techniques, such as sentence and word segmentation, POS tagging, parsing were introduced in this study. After preprocessing text data, biomedical entities of interest can be extracted using NER techniques and relations can be extracted by proposed algorithms in Section 3.3.

In this study, OpenNLP was used in an algorithm, written in R language, for POS tagging and parsing sentences into structured extended markup language (XML) format. The algorithm detects the sentence boundaries, determines the lemma of each word in the sentences and labels each word with grammatical roles such as noun, verb, and adjective. Finally, the algorithm determines the structures of the sentences based on POS tags and dependencies. The algorithm output consists of sentences with the corresponding POS tagged words in Penn-bank tree style, showing the syntactic relation to one another.

3.3 Text Processing for Information Extraction

After standard NLP tasks, NER was performed to identify the entities in the sentences. Three NER techniques, LingPipe, MetaMap and Abner were tried and their performances were evaluated. Since LingPipe training with GENETAG corpus can only recognize one bio-entity type, we only checked protein entity type in our experiments. Universal Medical Language System (UMLS) [14] was used for entity validation. Considering F-measure for evaluation, let True Positives represent an entity correctly identified, False Positives represent an incorrectly identified entity, False Negatives represent an entity not identified. By randomly selecting 100 sentences from 30 biomedical articles about "Magnesium deficiency" and "migraine attacks", experiments were repeated with the three NER techniques. The F-measures were as follows: LingPipe achieved 65.3%, MetaMap was 60.7% and Abner achieved 67.5%. Naturally, the higher the accuracy is, the more recognized entities are. Therefore, Abner was used for NER task in this work. Table 1 shows a sample of extracted entities and corresponding entity types from one text data with PMID 3006901.

Table 1: Example of extracted Entities

PMID	Candidate Entity	Entity Type
3006901	Fatty acid	Lipid
3006901	Vein graft	DNA
3006901	Cod liver	Body Part
3006901	Eicosapentaenoic acid	RNA
3006901	Cod-liver rich	cell Type

After entity extraction, verb relations were extracted from sentences. A verb-centric algorithm was developed with the help of a biomedical verb list. The UMLS semantic network [14] provides 54 bio-verbs or verb phrases which intend to cover the main relations that may exist between biomedical entities. In this work, we expanded the list of these 54 verbs by using WordNet [15] and VerbNet [16] so that the semantically similar verbs could be included.

The input of the algorithm is a Penn-bank style sentence with POS tags. The data processing procedure should only focus on sentences that contain valued relationships between extracted entities. Therefore, the algorithm starts by identifying relation bearing sentence based on two conditions: 1) there is more than one recognized bio-entity, 2) there is at least one verb that is semantically similar to one of the UMLS verb list.

If the current sentence satisfies these conditions, it is assumed to contain a relation of interest. The algorithm then determines the main verb in the sentence and extracts it. The basic relation, *Subject-Verb-Object*, called simple co-occurrence happens when a biomedical verb is detected between two entities. However, authors commonly use more complicated sentence structures. Three such structures were considered in this work:

(SS1) Clauses structures: *"... entity1 that/which ...entity2"* (using clauses to describe multiple verb-based relations in one sentence). For example, *"We propose that between attacks these metabolic shifts cause instability of neuronal function which enhances the susceptibility of brain to develop a migraine attack"*. There are two relation verbs, *cause* and *enhances*, which are connected by the relative pronoun word *which*.

(SS2) Sentence level conjunctive structures: *"...entity1 ...entity2 and/but ...entity3"* (using conjunctive structure to describe multiple verb-based relations in one sentence). For example, *"Female hormones lower magnesium but increase calcium levels which enhance migraine ubiquitousness."* There are two relation verbs, *lower* and *increase*, which are connected by the conjunctive word *but*.

(SS3) Phrase level conjunctive structures: *"...entity1 ...entity3, entity4, and entity5"* (using conjunctive structure to describe a single verb-based multi-to-multi relations). For example, *"Low magnesium intakes and blood levels have*

been associated with type 2 diabetes, metabolic syndrome, elevated C reactive protein, hypertension, atherosclerotic vascular disease, sudden cardiac death, osteoporosis, migraine headache, asthma, and colon cancer". There is a two-to-ten relation.

In order to overcome these cases, the input sentence should be divided into multiple semantic units if necessary to make sure each unit contains only one main verb. Since the relative pronoun structure and conjunctive structure have been recognized by the OpenNLP parser, the sentence can be partitioned based on the Penn-bank style tree. If the relative pronoun occurs just behind a noun or a noun phrase, the algorithm recognizes the clause sentence as a smaller unit and asserts the noun into the new unit. If the parent sub-tree of each conjunction corresponds to the entire sentence, the algorithm will recognize it as a sentence level word and break the sentence into smaller semantic units. The following sentences illustrate how the sentences structures above are dealt with in this work.

(SS1) Input sentence: *"We propose **that** between attacks these metabolic shifts cause instability of neuronal function **which** enhances the susceptibility of brain to develop a migraine attack."* The relative pronoun *"that"* appears after the verb *"propose"*, while *"which"* appears after the noun phrase *"instability of neuronal function"*. Therefore, the sentence is divided into two parts at the relative pronoun *"which"*, resulting in two smaller semantic units each containing an independent verb-based relation.

(SS2) Input sentence: *"Female hormones lower magnesium **but** increase calcium levels which enhance migraine ubiquitousness."* The conjunction *"but"* has the entire sentence recognized as its parent sub-tree. Therefore, the sentence is divided into two parts at the conjunction *but*, resulting in two smaller semantic units each containing an independent verb-based relation.

(SS3) Input sentence: *"Low magnesium intakes **and** blood levels have been associated with type 2 diabetes, metabolic syndrome, elevated C reactive protein, hypertension, atherosclerotic vascular disease, sudden cardiac death, osteoporosis, migraine headache, asthma, **and** colon cancer."* The two conjunctions *"and"* both return phrases as their parents. Therefore, the sentence will not be divided into smaller semantic units.

After partitioning the complex structure sentences into smaller semantic units, the main verb from each unit with a single verb-based relation is ready to be extracted. The algorithm extracts the word with verb tags if it is semantically similar to one of the UMLS verb list. This is described in Algorithm 1.

At this point, most of the main verbs in a sentence have been extracted. We are ready to construct the relations with bio-entities and bio-verbs. The algorithm scans the positions of each term in the semantic unit and recalls the locations of the main verb and bio-entities. Then, it extracts the bio-entities that are located before and after the main verb. Algorithm 2 describes the procedure for constructing the relation with bio-entities and bio-verbs.

```
Data: Penn-bank style semantic unit with POS tags
Result: Relation verbs
Split words;
if current sentence contains relations then
    while more words to process do
        read current word;
        if current word is the main verb then
            add the current word to relation-verbs;
        end
        go to the next word;
    end
else
    exit and go to next unit;
end
```

Algorithm 1: Algorithm for verb extraction when there is only a simple co-occurrence relation

```
Data: Penn-bank style semantic unit with POS tags
Result: Entities and relation verbs
Split words;
Scan each word and remember its position;
Construct relation from main verb;
while not the end of the semantic unit do
    if current word is a bio-entity then
        if it appears before the main verb then
            add the current word as subject of relation;
        else
            add the current word as object of relation;
        end
    end
    go to the next word;
end
```

Algorithm 2: Algorithm for constructing the relation with bio-entities and bio-verbs

For example, the sentence *"Female hormones lower magnesium but increase calcium levels which enhance migraine ubiquitousness",* from **(SS2)** will produce the three relationships shown in Table 2.

Table 2: Example of extracted relations from one sentence.

Subject	Verb	Object
Female hormones	lower	magnesium
Female hormones	increase	calcium levels
Calcium levels	enhance	migraine ubiquitousness

Note that the first and second relationship are connected by a conjunction and the third relationship is extracted from a clause unit.

For the sentence *"Low magnesium intakes and blood levels have been associated with type 2 diabetes, metabolic syndrome, elevated C reactive protein, hypertension, atherosclerotic vascular disease, sudden cardiac death, osteoporosis, migraine headache, asthma, and colon cancer"* from **(SS3)**, the algorithm first considers all the entities located before the main verb as one entity and then breaks them apart. The same happens to the entities located after the main verb. Therefore, the above sentence will produce 20 relationships. Table 3 shows four of the extracted relationships.

Table 3: Four example of extracted relations from one sentence.

Subject	Verb	Object
Low magnesium intakes	been associated with	type 2 diabetes
Low magnesium intakes	been associated with	metabolic syndrome
Blood levels	been associated with	type 2 diabetes
Blood levels	been associated with	metabolic syndrome

Finally, the semantic type of those entities should be obtained for a hierarchy relation network. UMLS contains a great number of bio-medical entities with the concept unique identifier (CUI) and their semantic types. Table 4 shows a sample of extracted relations from one text data with PMID 3006901.

Table 4: Example of extracted relations from one abstract.

Substance	Effect	Symptom	Disease	Body Part
Cod liver oil rich	Inhibit	Platelet aggregation	N/A	N/A
Cod liver rich	Found in	N/A	N/A	Cod liver
Eicosapentaenoic acid	Containing	N/A	N/A	Cod liver
Eicosapentaenoic acid	Inhibits	N/A	Intimal hyperplasia	N/A

4. Evaluation

In this section, we discuss the extraction effectiveness of the proposed algorithm in identifying relationships in biomedical texts, using benchmark datasets downloaded from the PubMed database. We downloaded three different topic datasets about "magnesium deficiency", "migraine attack" and "cancer", each of which consists of 100 article abstracts. A typical abstract contained roughly 8-10 sentences and the each dataset contained approximately 900 sentences. Samples of the text data downloaded are shown in Table 5.

Standard NLP techniques were used to obtain separated Penn-bank style sentence with POS tags from these three datasets. Subsequently, each sentence was fed to an Abner system to automatically recognize bio-entities and output the sentence with its entities labeled. The verb-based algorithm then took the labeled sentences with POS tags and the recognized bio-entities as input for verb extraction. Finally, after the bio-entities and bio-verbs were identified, the relations embedded in the sentences were extracted.

The performance of the algorithm was analyzed by randomly selecting 100 sentences from each dataset whose embedded relations were manually extracted. Then we ran the algorithm on those sentences and measured the precision, recall, and F-Score of the results. In what follows, by true positive we mean a correctly extracted relationship; by false positive we mean an incorrectly extracted relationship; and by false negative we mean a valid relationship that the algorithm failed to extract.

The precision, recall and F-score varied only slightly in our experiments. There was a higher incidence of false negatives resulting in relatively lower recall than precision rates. The proposed approach achieved an average precision of around 90% and recall of around 82%. Dataset 2 had slightly lower precision because the articles in it contained more complex sentences with multiple relationships. The algorithm ignored a few parts of some sentences whenever they could not be analyzed properly. These results are discussed in more detail in Section 4.

To demonstrate the advantage of the proposed multiple-relation extraction algorithm, we compared it with the single-relation extraction algorithm, which only extracts a single *Entity | Verb | Entity* relation from each sentence. We analyzed the effects of the three sentence structures **(SS1)**, **(SS2)**, and **(SS3)** on the single-relation extraction algorithm using the same collection of sentences. Without considering the three sentence structures **(SS1)**, **(SS2)**, and **(SS3)**, the single-relation extraction algorithm achieved a precision of around 73% and recall of around 65%. Our proposed algorithm improved the relation extraction performance significantly (precision of around 90% and recall of around 82%). The improvements highlight the importance of handling these issues. Table 6 summarizes the evaluation results of the single-relation extraction algorithm and the multiple-relation extraction algorithm running on the three datasets.

Discussion

During the evaluation experiment, we identified that the false positives and false negatives were caused by some similar issues. Most false positives occurred when the sentences appeared in the description of the work itself. For instance, the sentence *"This study aimed to assess whether magnesium deficiency can cause migraine headache"* describes the objective of the study rather than an actual relationship. However, the algorithm extracted the relationship *"magnesium deficiency | cause | migraine headache"*. In addition, the algorithm failed to correctly recognize relations with negative modifications, resulting in some false positives.

As for the false negatives, they were mainly caused by the nature of the verb-based algorithms. Pronouns are sometimes used to refer to entities appearing earlier in the text resulting

Table 5: Sample of the downloaded text data

PMID	Title	Abstract
25177816	Magnesium deficiency and dysregulation of vascular tone	Magnesium deficiency is associated with impaired vascular tone bidirectionally and can lead to both an abnormal increase and abnormal lowering of blood pressure...
25137281	There is chronic latent magnesium deficiency in apparently healthy university students	INTRODUCTION: Magnesium is an essential micronutrient for human body, and its deficiency has been associated with risk of non-communicable diseases...
3006901	Effects of cod-liver oil on intimal hyperplasia in vein grafts used for arterial bypass	Cod-liver oil rich in eicosapentaenoic acid, an unsaturated fatty acid, has been shown to inhibit platelet aggregation...

Table 6: Evaluation results of the single-relation extraction algorithm vs. the multiple-relation extraction algorithm

	Single-relation algorithm			Multiple-relation algorithm		
ID	Precision	Recall	F-Score	Precision	Recall	F-score
1	75.8%	66.9%	71.1%	90.3%	83.7%	86.9%
2	72.3%	65.8%	68.9%	88.3%	81.5%	84.8%
3	74.1%	65.5%	69.5%	91.4%	84.1%	87.6%

in the failure of the recognition of the entity. Some relations are described as noun phrases embedded in sentences instead of involving verbs. Three common scenarios are discussed as follows.

1) The use of pronouns such as *"it", "they"*. For example, in the sentence *"While the data extracted suggest that magnesium may be effective in treating all symptoms in patients experiencing migraine with aura across all migraine patients, its effectiveness seems to be limited to treating only photophobia and phonophobia"*, the pronoun "its" refers to the entity "magnesium" in the clause sentence. However, it is difficult for our algorithm to extract the relationship between "magnesium" and "photophobia" because the reference to magnesium is implicit in the term "its effectiveness".
2) Occurrence of prepositions such as *"by", "of"*. For example, the relation between "Mg and Mn" and "biosynthesis of terpenes and phenolics" was not extracted by the algorithm from the sentence *"These results suggest a profound effect of the combined supply of Mg and Mn on the biosynthesis of terpenes and phenolics"*.
3) Relations embedded in unconventional structures involving *"between ...and ...", "with ...and ..."* are difficult to extract. For example, the algorithm failed to recognize any verb indicating relationships in the sentence *"This experiment studied the positive influences between Vitamin D and the integrity of skeleton"*.

5. Related Work

Automatic relation extraction from unstructured texts has recently attracted considerable interest [7], [8], [17]. The main approaches used for this are described below.

Co-occurrence approaches provide the simplest way to detect relations if the two entities are frequently collocated with each other across a collection of texts or sentences. They result in high recalls but may have poor precisions. Now they are usually compared against other methods as a baseline method [18], [19]. *Link-based approaches* extend co-occurrence approaches if the two entities often co-occur with a common term across a collection of corpus. They usually improve the precision but the recall rate remains low [20]. Although in theory both approaches can be applied directly to raw texts, NLP techniques are employed in virtually all cases to pre-process the text.

Machine learning approaches label and segment sentences automatically by using Hidden Markov Model [21], Conditional Random Fields [22] and Naïve Bayes classifier [23]. However, they require manually annotated training data which can be expensive to obtain. In addition, they may result in a limited coverage in different domains.

Rule-based approaches use NLP techniques and templates generated manually by domain experts to identify semantic entities and extract associations connected by some specific verbs [24], [25]. Standard NLP techniques such as POS tagging parsing, and NER are used to generate the dependency trees and simple co-occur relation structures, such as *Entity-Verb-Entity*, *Entity binds Entity but not Entity*, are considered for relation extraction, resulting in a reasonable precision around 80% and recall around 85% [24], [25]. However, they are computationally costly if they are dealing with large size data [18]. In addition, most investigation of rule-based approaches has centered around specific types of relationships.

Verb-based approaches share some similarities with the rule-based approaches. They both highly rely on NLP techniques, while verb-based approaches cover a much wider range of complex relationship types [26]. However, existing

verb-based approaches only deal with single relations.

6. Conclusions and Future Work

In this work, we proposed an enhanced verb-based algorithm capable of extracting multiple relations embedded in a single sentence obtained from unstructured data (articles). Given a sentence written in natural language as input, the system processes it using standard NLP techniques and entities are extracted using a NER module. Unlike existing single-relation extraction algorithms, the proposed algorithm can handle complex sentences containing multiple relations such as clauses structure sentences **(SS1)** and conjunctive structure sentences **(SS2)** and **(SS3)**. After separating complex sentences into smaller semantic units, Algorithm 1 is applied to extract the main verbs and Algorithm 2 is applied to derive the relations from the extracted entities and verbs. Finally, the extracted relations are also classified into different semantic types such as substances, symptoms, diseases, body parts, etc. Our multiple relation extraction algorithm achieved an average precision of around 90% and a recall of around 82% when tested on three datasets obtained from the biomedical domain. For comparison, the single-relation extraction algorithm without our enhancements achieved precision of around 73% and recall of around 65% using the same datasets (see Table 6 for full results). Although the comparison was performed over a relatively small sample, it shows a significant improvement of the precision and recall rates of our algorithm over existing approaches. In future work, we intend to use publicly available datasets to evaluate the performance of our algorithm more directly against that of other relationship extraction approaches. It is worth emphasizing that with different NER training datasets and verb datasets, our algorithm can also extract relations in other domains such as news, etc.

We intend to improve the algorithm further so it can deal with the specific scenarios discussed in Section 4 and add polarity classification to it. Finally, the extracted relationships can be used for relation mapping and to discover novel hidden relations across research domains.

References

[1] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[2] J. Kottmann, B. Margulies, G. Ingersoll, I. Drost, J. Kosin, J. Baldridge, T. Goetz, T. Morton, W. Silva, A. Autayeu, *et al.*, "Apache opennlp," *Online (May 2011), www. opennlp. apache. org*, 2011.

[3] B. Carpenter, "Lingpipe for 99.99% recall of gene mentions," in *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, vol. 23, 2007, pp. 307–309.

[4] B. Settles, "Abner: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, 2005.

[5] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[6] L. V. Subramaniam, S. Mukherjea, P. Kankar, B. Srivastava, V. S. Batra, P. V. Kamesam, and R. Kothari, "Information extraction from biomedical literature: methodology, evaluation and an application," in *Proceedings of the twelfth international conference on Information and knowledge management*. ACM, 2003, pp. 410–417.

[7] A. Skusa, A. Rüegg, and J. Köhler, "Extraction of biological interaction networks from scientific literature," *Briefings in Bioinformatics*, vol. 6, no. 3, pp. 263–276, 2005.

[8] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "Frontiers of biomedical text mining: current progress," *Briefings in bioinformatics*, vol. 8, no. 5, pp. 358–375, 2007.

[9] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[10] J. Maršík and O. Bojar, "Trtok: a fast and trainable tokenizer for natural languages," *The Prague Bulletin of Mathematical Linguistics*, vol. 98, pp. 75–85, 2012.

[11] A. B. Abacha and P. Zweigenbaum, "Medical entity recognition: A comparison of semantic and statistical methods," in *Proceedings of BioNLP 2011 Workshop*. Association for Computational Linguistics, 2011, pp. 56–64.

[12] S. Tratz and A. Sanfilippo, "A high accuracy method for semi-supervised information extraction," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Association for Computational Linguistics, 2007, pp. 169–172.

[13] S. Kovalchik, "Download content from ncbi databases," 2014.

[14] B. L. Humphreys and D. Lindberg, "The umls project: making the conceptual connection between users and the information they need." *Bulletin of the Medical Library Association*, vol. 81, no. 2, p. 170, 1993.

[15] P. University, "Princeton university about wordnet," 2010. [Online]. Available: <http://wordnet.princeton.edu>

[16] K. K. Schuler, "Verbnet: A broad-coverage, comprehensive verb lexicon," 2005. [Online]. Available: http://verbs.colorado.edu/ mpalmer/projects/verbnet/

[17] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.

[18] J. R. Curran and M. Moens, "Scaling context space," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 231–238.

[19] R. Bunescu, R. Mooney, A. Ramani, and E. Marcotte, "Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from medline," in *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*. Association for Computational Linguistics, 2006, pp. 49–56.

[20] P. Srinivasan, "Text mining: generating hypotheses from medline," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 5, pp. 396–413, 2004.

[21] N. Collier, C. Nobata, and J.-i. Tsujii, "Extracting the names of genes and gene products with a hidden markov model," in *Proceedings of the 18th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2000, pp. 201–207.

[22] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H.-P. Kriegel, "Extraction of semantic biomedical relations from text using conditional random fields," *BMC bioinformatics*, vol. 9, no. 1, p. 207, 2008.

[23] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Computational linguistics*, vol. 28, no. 3, pp. 245–288, 2002.

[24] R. Feldman, Y. Regev, M. Finkelstein-Landau, E. Hurvitz, and B. Kogan, "Mining biomedical literature using information extraction," *Current Drug Discovery*, vol. 2, no. 10, pp. 19–23, 2002.

[25] J.-J. Kim, Z. Zhang, J. C. Park, and S.-K. Ng, "Biocontrasts: extracting and exploiting protein–protein contrastive relations from biomedical literature," *Bioinformatics*, vol. 22, no. 5, pp. 597–605, 2006.

[26] A. Sharma, R. Swaminathan, and H. Yang, "A verb-centric approach for relationship extraction in biomedical text," in *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*. IEEE, 2010, pp. 377–385.