Journal of AI and Data Mining
Vol 4, No 2, 2016, 203-212

10.5829/idosi.JAIDM.2016.04.02.08

Extraction of Drug-Drug Interaction from Literature through Detecting Linguistic-based Negation and Clause Dependency

B. Bokharaeian* and A. Diaz

Facultad de Informática, Universidad Complutense de Madrid, Calle del Prof. José G! Santesmases, Madrid, Spain.

Received 19 January 2016; Accepted 04 May 2016
**Corresponding author: behrou.bo@ucm.es (B. Bokharaeian).*

Abstract

Extracting biomedical relations such as drug-drug interaction (DDI) from text is an important task in biomedical natural language processing. Due to the large number of complex sentences in biomedical literature, researchers have employed some sentence simplification techniques to improve the performance of the relation extraction methods. However, no significant improvement has been reported in literature, since the task is difficult. This paper aims to explore clause dependency related features alongside to linguistic-based negation scope and cues to overcome complexity of the sentences. The results show through employing the proposed features combined with a bag of words kernel, the performance of the used kernel methods improves. Moreover, experiments show that the enhanced local context kernel outperforms other methods. The proposed method can be used as an alternative approach for sentence simplification techniques in biomedical area which is an error-prone task.

Keywords: *Drug-Drug Interaction, Relation Extraction, Negation Detection, Clause Dependency.*

1. Introduction

Although being relatively new, biomedical relation extraction from text is a fast-growing topic in Natural Language Processing (**NLP**) research field. Considering the ever increasing number of biomedical researches with the huge number of unstructured biomedical text resources being involved, it seems to be highly demanding to extract biomedical relations out of scientific texts and reports. Biomedical Natural Language processing or briefly BioNLP refers to the text mining applied to literature of the biomedical and molecular biology domain.

With **Drug-Drug interaction** being a serious event in medicine, automatic extraction of these interactions from text is an important task to be carried out in BioNLP. A drug-drug interaction (DDI) usually occurs when the activity level of one drug is changed by another drug. According to the reports by U.S. Food and Drug Administration (FDA) and other acknowledged studies, annual life-threatening DDI's occurring in the United States exceed two million cases [1]. With the purpose of recording DDIs, many academic researchers and pharmaceutical companies have tried to develop either relational or structural databases such as [2,3]. However, most valuable researches and information are still found only within unstructured text documents such as scientific publications and technical reports.

Moreover, since biomedical relations, such as *protein-protein* and *drug-drug interactions,* significantly contribute to identification of biological and medical processes, biomedical relation extraction is believed to be a very important research topic within the field. Many of the existing works on biomedical relation extraction task in the literature (including the DDI detection) are approached via supervised binary relation extraction [4]. As such, other types of algorithms including complex relation extraction algorithms and semi-supervised ones are expected to be incorporated into this kind of the relation extraction task [5].

Role identification of clauses incorporated in complex sentences in the course of DDI detection is another linguistic- driven task which is carried out in this research. According to linguistics, an

independent clause, or main clause, is the one that can be seen as a complete sentence, by itself, expressing a complete thought. Moreover, a **dependent clause** refers to a group of words, including a subject and a verb, which does not express a complete thought and cannot stand alone. It usually extends the meaning of the main clause [6]. Consequently, a **complex sentence** consists of one independent clause along with one or more dependent clauses. Moreover, the term clause connects or refers to a word used to join or to connect clauses to compose complex sentences. **Coordinators**, **conjunctive adverbs**, and **subordinators** are three types of connectors. As an example, in the following sentence:

- *Although* (specific drug or food interactions with mifepristone has not been studied), (it is possible) *that* < ketoconazole, itraconazole, erythromycin, and grapefruit juice may inhibit its metabolism.>

In this sample, “Although” and “that” are two subordinator connectors separating three different clauses identified in “(” and “<”. The main clause has been enclosed with “<”. Two other clauses are dependent clauses which complement the main clause.

One of the few researches on relation extraction task with clauses which had been taken into account, was the one carried out by [7]. They tried to select the best clauses to develop a sentence simplification algorithm and reported some improvements regarding different types of rules they used for the sake of simplification and clause selection procedures.
This research is an attempt to extend identified text or subtree features in three kernel-based methods, namely **global context kernel** (GCK) method, **local context kernel** (LCK) [8] method, and **subtree** [9] kernel. The extension has been carried out in two steps. First, several clause connectors are detected whether components of the kernel methods - token or subtree – exist in a dependent or independent clause and second, type of the clause was identified.

On the other hand, detecting negative assertions is essential in most BioMedical text mining tasks, where in general, the aim is to derive factual knowledge from textual data. According to linguistics [10], **negation** is a morphosyntactic operation in which a lexical item denies or inverts the meaning of another lexical item or construction. Likewise, a **negator** is a lexical item that expresses negation. Negation is commonly used in clinical and biomedical text documents and is an important origin of low precision in automated information retrieval systems [11]. Exploring efficiency of linguistic-based negation related features is another purpose of this research.
In the next section, some of the previous related works and resources will be reviewed.

2. Related works

The first Drug-Drug interaction corpus initially developed by [12] had a pile of 579 xml files describing DDIs randomly collected from the *DrugBank* database [13]. In 2011, the first DDI Extraction competition was held with the aim of encouraging researchers to explore new methods for extracting drug-drug interactions. The best results obtained in the course of detecting and classifying DDIs were a F-measure of 65.74%, a precision of 65.04% and a recall of 71.92% [14]. As a part of SemEval-2013 (International Workshop on Semantic Evaluation), the second competition was held in 2013. A novel corpus was developed which included not only the one used in 2011 [12] but some *Medline* abstracts. The participant teams developed solutions on the basis of either supervised or sentence-level relation extraction methods; the best F-measure achieved was 75%[15]. It is worth mentioning that the authors have participated in this challenge and the details of the developed system can be found at [16].
Additionally, several machine learning approaches have already been developed to extract relations from text including Sequence kernels [8], Tree kernels (parse tree based) [9], and Graph kernels (graph parsing-based) [17]. Two more recent approaches undertaken by [18] and [14] ranked first and second in DrugDDI challenge 2013, respectively. Chowdhury and his colleagues [18] have used linear combination of a feature based kernel, a Shallow Linguistic (SL) kernel and Path-enclosed Tree (PET) kernel to proposed a hybrid kernel. Defining a multiplicative constant they went for assigning a higher (or lower) weight to the information obtained by tree structures. In another work, Thomas and his colleagues [14] proposed a two-step approach starting with extraction of candidates using ensembles comprised of up to five different classifiers and then relabeling to one of the four categories. Moreover, other types of machine learning approaches such as maximal frequent sequence have been employed effectively in DDI extraction [19]. A survey about different machine learning tools in DDI related tasks can be found here [20].
Additionally, considering negation when addressing relation extraction task, Faisal

Chowdhury and colleagues [21] developed a list of such features as the nearest verb to the candidate entities in the parse tree and few negation cues, by which the SVM classifier was fed. Although, some improvement was observed but there was nothing providing how much the negation identification's performance has improved.

It is worth mentioning that two negation detection methods have been developed and employed for annotating the used corpora: a **linguistic-based** approach and an **event-oriented** approach. Two of the known negation annotated corpora are the linguistically-focused, scope-based **BioScope** and the event-oriented **Genia**. In BioScope, scopes aim to recognize the position of the key negated event in the sentence. Furthermore, all arguments of these key events are also under the scope [22]. In the Genia event, biological concepts (relations and events) have been annotated for negation, but no linguistic cues have been annotated for them [23]. In fact, the main objective of the BioScope corpus is to investigate this language phenomenon in a general, task-independent and linguistically-oriented manner.

As another subtask utilized to perform relation extraction task, sentence and clause simplification goes for modification, enhancement, classification or otherwise processing an existing piece text in such a way that the prose's grammar and structure is greatly simplified with the original meaning and information remained unchanged [24]. Moreover, being a text simplifying system, ISIMP [25] attempts to improve text mining tools including relation extraction tasks. Another research [26] performed on the same line, went for some simplification techniques to simplify complex sentences by splitting the clauses. They split the clauses before implementing some simplification rules to generate new simple sentences.

Throughout the rest of this paper, first the proposed method and its components are explained. It explains the process of employing the extracted features in combination with other kernel methods. Then, in the fourth section, the results are exhibited, and in the last section, the results are discussed and concluded, and some suggestions are given for future works.

3. Method

In this section, we begin our discussion by extending the DrugDDI corpus through negation scopes and cues [27]. Then feature extraction, general framework and other components of the implemented system are explained (Figure 2).

In order to use the negation effects in the course of relation extraction task, an extension of the two mentioned DrugDDI corpora, especially the Drug Bank section in the 2013 version, annotated with negation scope and negation cue was prepared (Figure 1).

All the sentences of the DrugDDI 2011, which consists of 5806 different sentence and 579 files, were used and automatically annotated and then due to possible mistakes that may have happened in the automatic process, a manual checking was carried out. Obviously, because every combination of drug names can be a DDI candidate in the corpus, each sentence may explain more than a DDI candidate. Therefore, as can be understood from table 1, DrugDDI and the produced NegDDI-DrugBank corpus have 31,270 DDI candidates. It is worth mentioning that, in this paper, "["is used to indicate the start point of the negation scope, and "]"to indicate the ending point; also "{"and "}"are used for identifying negation cue. One example is illustrated in the sentence below:

- [Concomitant use of bromocriptine mesylate with other ergot alkaloids is {not} recommended].

In passive sentences with the following structure "It (this or that) + finite format of to be + not + past participate"; scope opens at the beginning of the sentence.

The NegDDI-DrugBank corpus was prepared by adding new XML "negationtag" at the end of each sentence XML tag within which "negation cues" and "negation scopes" are used. The extended NegDDI-DrugBank corpus is available for public use[1].

3.1 Feature extraction

Researchers have proposed complicated kernel based methods to use different shallow or deep features to capture different types of complex sentences. However, most of the previous literatures [28] suggest that, rather than simple sentences of single clause, more errors are to be produced by complex compound sentences which are, by the way, very common in the biomedical literature.

Describing global context kernel, local context kernel, and subtree kernel, respectively, the three tables represent complex and compound sentences which are very commonly used in biomedical literature to produce higher error rates than those of simple sentences with just one clause. As it can

[1]http://nil.fdi.ucm.es/sites/default/files//NegDDI_DrugBank.zip

be seen in the tables where error analysis results are reported, these mistakes are more frequently undertaken when approaching via solely shallow linguistic processes [8]. Such approaches include tokenization, sentence splitting, Part-of-Speech (PoS) tagging and lemmatization.

```
<sentence id="DDI-DrugBank.d297.s4" text="Concurrent therapy with ORENCIA and TNF antagonists is not recommended.">
  <entity charOffset="24-30" id="DDI-DrugBank.d297.s4.e0" text="ORENCIA" type="brand"/>
  <entity charOffset="36-50" id="DDI-DrugBank.d297.s4.e1" text="TNF antagonists" type="group"/>
  <pair ddi="true" e1="DDI-DrugBank.d297.s4.e0" e2="DDI-DrugBank.d297.s4.e1" id="DDI-DrugBank.d297.s4.p0" type="advise"/>
  <negationtags><xcope> Concurrent therapy with ORENCIA and TNF antagonists is <cue>not</cue>
    recommended</xcope>.</negationtags>
</sentence>
```

Figure 1. The extended unified XML format of a sentence with negation cue in NegDDI-DrugBank corpus.

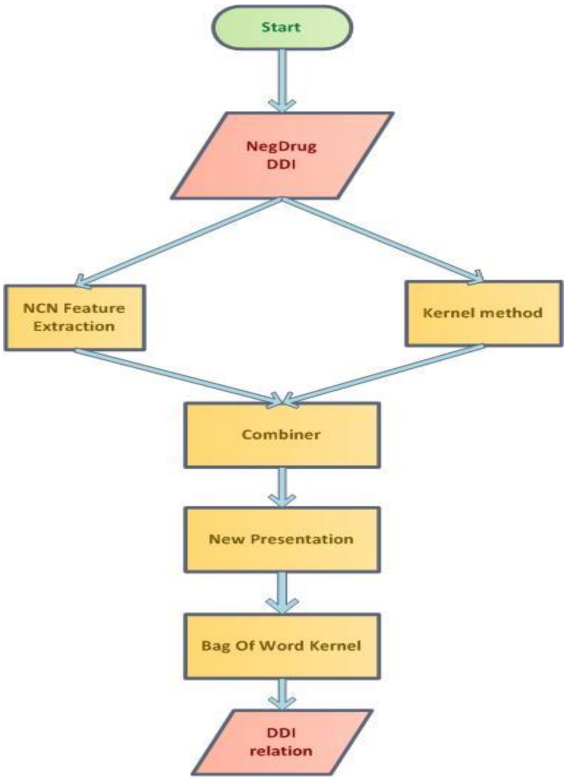

Figure 2. The Basic components of the implemented and proposed method.

In the rest of this section, the implemented algorithms for extracting clause related features along with the negation scope and cues are introduced.

3.1.1 Clause related features

As mentioned earlier, the position of independent or the dependent clauses, and also the type of dependent clause (e.g. Adverbial, adjective or noun) are known to be among major factors contributing into relation extraction process. This makes it of critical importance to distinguish independent clauses from the dependent ones.

As presented in table 1, more than 27% of the DDI candidates in the testing part of NegDDI-DrugBank corpus and 19% of those in the training part contain, at least, one dependent clause. Table I show that the frequency of subordinating clauses in sentences with negation cues is higher than that in other types of sentences. Therefore, due to the large number of sentences with more than one clause, the complex structure of these sentences, and their higher associated rate of error together with the important role they play when using negation concept, it will be very important to take clause dependency features into account.

Different types of dependent clauses can alter the sentence's overall meaning in distinct ways. For instance, a **concessive clause** is the one beginning with "although" or "even though", expressing an idea opposite to the main part of the sentence; as an example see the following sentence:

- *Although* there may be decreased *zalcitabine* activity because of lessened active metabolite formation, <the clinical relevance of these in vitro results is not known>.

Here the main clause (enclosed with "<") has opposite meaning to that of the dependent clause which indicates some changes in the *zalcitabine* activity. Another type of clause frequently seen within the NegDDI-DrugBank corpus is the **adverbial** clause. One of the connectors used within these adverbial clauses is while. See the following sentence as an example:

- The amount of metformin absorbed (**while** taking Acarbose) was bioequivalent to the amount absorbed (**when** taking placebo), (as indicated by the plasma AUC values.)

Table 1. Statistics of the ddi candidates with more than one clause in negddi-drugbank.

Category	Number of candidates with more than one clause	Total candidates	*Rate*
Test part	1401	5265	27%
Train part	5015	26005	19%
Have negation in test part	396	1409	28%
without negation in test part	1005	3898	26%

Analyzing different types of dependent clauses, two feature categories were extracted. The first category encompassed 28 Boolean features corresponding to 28 clause connectors. Consequently 28 Boolean features were extracted corresponding to the 28 clause connectors. A selected list of the most used connecters and their frequencies in the corpus can be found in table 2.

The other feature category was based on substructures (tokens or subtrees) used in the applied method; it identified whether substructure was inside the main clause or not. For instance, similar to features used in the Global context kernel, three new text features were extracted with "IDC" prefix for tokens inside the independent clause and "DC" for those inside the dependent clause. Similarly, in order to improve the subtree kernel, other new subtrees were defined corresponding to the usual subtrees. In short, when it is inside the dependent clause, this subtree comes with DC prefix added before its root name, while the IDC prefix was used for subtrees inside the independent clause.

Table 2. Statistics of the most frequent clause connectors in negddi-drugbank corpus.

Clause connector	Frequency.
Although	651
While	3358
When	511
Anywhere	29
Until	186
Till	710
Because	58
Even though	625
Since	1307
But	123
Unless	347
Total	7905

Constituent parse trees have been analyzed to detect whether a substructure is inside the main clause or within the dependent one. The proposed algorithm gives a "DC" prefix to the substructure provided that shortest path between the token, or the subtrees root, and the main root contains a subordinate clause node (SBAR) (Fig. 3); otherwise, it gives the substructure an "IDC" prefix. For example, for the sentence with its constituent parse tree shown in Fig. 3, "DC-clinical" is a new token made by the program due to existence of a "SBAR" along the path which connects the token to the main root.

Such a new token which can be placed on the left, right or between the two drug names within the original sentence beside other newly produced tokens, is the result of the proposed improved version of Global and Local context kernel. Via such an approach, one may create three new corresponding text features.

The improved version of the subtree kernel produced a new subtree with "IDC-IN" as its root based on the subtree on the upper left of figure 3 containing the leaf "although" and the root IN.

3.1.2. Negation related features

In addition to previous features, we conducted some experiments on negation related features. Regarding the position of drug names (inside or outside the negation scope), there are 6 different possibilities to be used as 6 features:

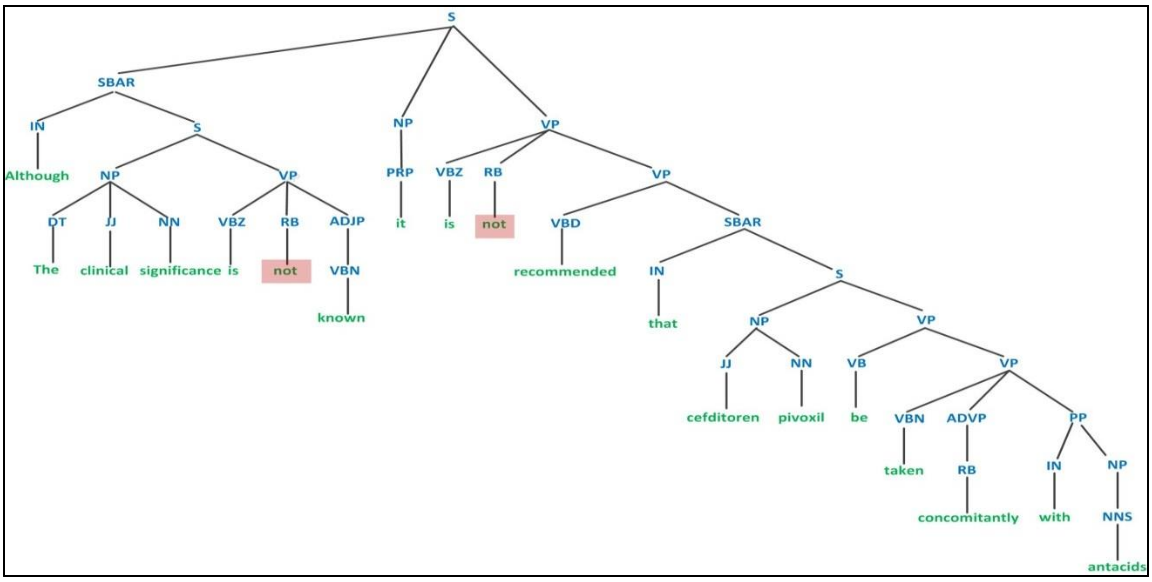

Figure 3. A sample of constituent parse tree of a complex sentence with two dependent clauses and two negation cues.

- BothinsideNegSc: A Boolean feature set to "true" when both drugs are inside the negation scope with other situations being false.
- BothLeftSNegSc: A Boolean feature set to "true" when both drugs are on the left side of the negation scope with other situations being false.
- BothRightNegSc: A Boolean feature set to "true" when both drugs are on the right side of the negation scope, while other situations are false.
- OneLeftOneInsideNegSc: A Boolean feature set to "true" when one drug is on the left side of the negation scope and the other is inside, with other situations being false.
- OneRightOneInsideNegSc: A Boolean feature set to "true" when one drug is on the right side of the negation scope and the other is inside, with other situations being false.
- OneLeftOneRightSc: A Boolean feature set to "true" when one drug is on the right side of the negation scope and the other one on the left, with other situations being false.

The features were used alongside with clause related features to enhance the performance of the used kernel methods. The next section will describe the proposed method based on the extracted features.

3.2 DDI extraction using bag of words kernel

The newly created presentation obtained from mentioned features was classified using a **bag-of-words** based kernel method which tries to find a polynomial combination of the features commonly known as the kernel function (Figure 2) support vector machine with *SMO* [**29**] implemented was used which outperformed other implementations of SVM, according to the performed experiments in the study, , e.g. *libSVM*, in terms of convergence rate and quality of results. *Weka* API was used as the implementation platform. Executed without a stemming step, the term minimum frequency of bag-of-words based kernel method was set to one.
For all the mentioned methods, every feature was considered blind, so as to replace all drug names within the generated features with two general terms, i.e. "DrugName" was used for the two drugs with their interaction being investigated while "OtherDrugNames" was the term used for the other drugs. In order to be aligned with pharmaceutical texts, the tokenization process was carried out using Stanford *BioNLPTokenizer* [30]; however, Stanford parser was used to parse constituents. Moreover, as the method via which the winning team of DDI extraction challenges 2011 was approached, *TreeTagger* was employed for the sake of **Lematization** and **Pos tagging**. Additionally, some guidelines which has been suggested in [31] has been employed for improving the performance.

4. Results

In this section, comparative results of the augmented methods are presented in terms of F-measure along with those of the original methods. In this section, the results of two different types of validation experiments are presented. Firstly, similar to SemEval DDI challenges, the training set of the Drug Bank corpus was used to train the system while the test set was utilized for testing the system. Secondly, the NegDDI-DrugBank corpus 2013 was 10-fold cross validated with the results being displayed. Subsequently, a statistical sign test is presented to show the significance of the improvements caused by the proposed method compared to other three methods used. The section is closed by presenting the error analyses performed on the results of the system.

Table 3 demonstrates associated results with the improved Global context kernel method (with **CLA** postfix for clause related and **NEG** for negation related features) undertaken along with the original ones. The first row in the table displays the results for those sentences with negation cue but no clause connectors, in the testing dataset. The second row shows the results for candidates with negation cue and clause connectors. The third and the fourth rows show the results for sentences without negation cue, but with and without clause connectors, respectively. The last row contains the results for entire testing dataset.

10-fold cross validation of NegDDI-DrugBank 2013 (both testing and training sets) led to the results reported in table 4. In addition to previous features, the negation related features are denoted by NEG postfix. As the table shows, F-measure was increased in the course of 10-fold cross validation experiments with the best F-measure for the proposed methods being the one obtained by the proposed local context kernel-(CLA) method (81.8%).

According to the reported results in tables 3, 4 and 5, the proposed method is proved to successfully improve the F-measure across all tested categories. In global context kernel method, sentences with no negation cues and with clause connectors exhibited the best improvements with an average increase of +4% in the value of F-measure. Similarly, in local context kernel, the best improvement was 3.9%.

On the other hand, similar experiments to those conducted by global context kernel were carried out by the modified local context kernel with the best performance exhibited by the dataset containing the sentences without negation cues and clause connectors, just like what was observed for the global context kernel (Table 4).

However, the system succeeded to realize satisfyingly enough improvement in the performance of the other three datasets in terms of detecting DDIs, satisfactorily.

Also, similar to global context kernel, by considering tokens in the original LCK, some duplicated features were generated by negation scope and cue, and clause dependency features; such feature generation is associated with system performance degradation.

In addition to the two mentioned sequence kernel methods, several experiments were carried out with subtree kernel which can be seen in table 5. With an average increase of +9.8% in the value of F-measure, sentences without negation cues but with clause connectors in the subtree kernel demonstrated the best improvements (Table 5). And an average of 3% was the best improvement which was obtained through employing both categories of features.

Last but not the least, the results of experiments undertaken on the improved subtree kernel with clause related features (Table 5) indicated the best performance and the highest improvement rate (2.1%) was obtained for clause related features which are the ones associated with the dataset containing sentences with clause connectors but no negation cues. Similarly, the highest improvement rate for both categories of features (9.8%) is the one associated with this dataset of sentences.

The results showed that, compared to other methods, the subtree kernel was relatively more effective in detecting DDIs within complex sentences; this seems to be because, rather than the other two used sequence kernels, it generally extracts a larger deal of structural and componential information from the sentences.

However, no significant improvement was observed for those sentences with negation cues, scopes and connectors. As explained before, this is possibly because of very good performance of the original subtree kernel when applied on this type of sentences, so that invented features are not so effective.

It is also worth mentioning that simple **cause-effect** and **time** connectors are considered to be the most frequent clause connectors. As shown in table 2, as the first and second most frequently used clause connectors in the NegDDI-DrugBank corpus, “when” and “but” are an adverbial clause connector and a coordinating conjunction, respectively.

Additionally, a set of additional experiments grounded on basic simplification methods were conducted in order to reduce complexity of the sentences. However, no better result was obtained through the additional experiments. For instance, substituting the dependent clause with an independent clause feature caused no improvements in the system performance.

With an F-measure of (comparable with 65.7% corresponding to the first system in DDI extraction challenge 2011 implemented by “Humboldt University of Berlin”), improved Local context kernel method (LCK-CLA) which produced the best obtained results for the testing set.

Table 3. Obtained results for global context kernel method with combination of negation and clause related feature sets in terms of f-measure.

Test Category	M (%)	M+ CLA (%)	Dif. (%)	M+ NEG+ CLA	Dif. (%)
With negation No connector	56.6	59.6	+3.0	57.8	+1.2
With negation With connector	51.7	53.9	+2.2	52.3	+0.6
No negation With connector	62.3	66.3	+4.0	63.7	+1.5
No negation No connector	64.7	65.8	+1.1	64.8	+0.1
Total	61.7	63.9	+2.2	62.4	+0.7

Table 4. Obtained results for subtree kernel method with combination of negation and clause related used feature sets in terms of f-measure.

Test Category	M (%)	M+ CLA (%)	Dif. (%)	M+ NEG+ CLA (%)	Dif. (%)
With negation No connector	60.9	60.9	+0.9	59.9	-1.0
With negation With connector	63.2	63.6	+0.4	63.2	0
No negation With connector	58.6	60.7	+2.1	68.4	+9.8
No negation No connector	36.3	37.3	+1	38.6	+2.3
Total	47.1	48.1	+1	50.1	+3.0

Table 5. Obtained results for local context kernel method with combination of negation and clause related used feature sets in terms of f-measure.

Test Category	M (%)	M+ CLA (%)	Dif. (%)	M+ NEG+ CLA (%)	Dif. (%)
With negation No connector	62.6	63.8	+1.2	61.5	-1.3
With negation With connector	58.0	61.9	+3.9	50.9	-7.2
No negation With connector	64.8	65.9	+1.1	65.7	+1.0
No negation No connector	63.9	64.9	+1	64.2	+0.2
Total	63.4	64.7	+1.3	63.0	-0.4

Table 6. Obtained results for 10-fold cross validation for three original kernel methods with combination of negation and clause related feature sets in terms of f-measure.

Method (M) name	M (%)	M+ NEG	Dif. (%)	M+ CLA	Dif. (%)	M+ NEG+ CLA	Dif. (%)
Global context Kernel	77.4	77.3	-0.1	78.9	+1.5	77.9	+0.2
Local context kernel	80.7	81.1	+0.4	81.8	+1.1	81.7	+1
Subtree kernel	71.9	72.1	+0.2	73.8	+1.9	74.9	+3

Error analysis

Some error identification analyses are presented in this section. The identified sources of error can be categorized as follows:

- Although most of the clause connectors can simply be identified by a superficial analysis (as in the experiments), there are, challenging clause connectors with possible alternative speech parts within a sentence. Such sorts of connectors are most problematic as there may be different speech parts (such as demonstrative pronouns) within the sentence. Thus, for the sake of simplicity, it was not used as a clause connector feature. Other similar clause connectors were either considered or ignored by whether they took common speech roles in scientific medical articles. For instance, “when” was considered as a connector only being a common speech role within the mentioned articles; it was, however, ignored as an information question word. In both cases, more precise in-depth procedures and experiments are required to achieve correct detections.

- As previously-mentioned, overlapping of the proposed features with those used in the original relation extraction method was another source of error. For instance, the studied clause connectors were associated with equivalent tokens within both original sequence kernels. Such situations downgrade the final performance of the system. Several experiments carried out on this problem revealed improvements once a manual **feature selection** method was undertaken. Therefore, undertaking an effective automatic feature selection method for each of the proposed methods can improve the system.

- All extraction systems (including the proposed system herein) suffer from **parentheses** as another source of inaccuracy. For instance, some parentheses contain a clause or explanation in which a drug name is used, see the sentence below:

– Although specific drug or food interactions with mifepristone have not been studied, on the basis of this drugs metabolism by CYP 3A4, it is possible that *ketoconazole*, itraconazole, erythromycin, and grapefruit juice may inhibit its metabolism (increasing serum levels of *mifepristone*).

Here, *ketoconazole* and *mifepristone* are two drug names interacting in the corpus. However, parentheses may prevent their interaction from being detected by the system. A simplification algorithm could be implemented to get rid of the parentheses issue.

5. Discussion and conclusion

Being an important task in the course of Biomedical Natural Language Processing, supervised biomedical relation extraction tries to extract relations between biomedical entities. Among other critical biomedical relations, this paper investigated Drug- Drug interactions. Many methods have been developed to extract DDI relations. However, substantial studies on the effects of clause dependency on the relation extraction task are yet to be reported, so as to propose adequate methods in this regard. Besides, sentences with negation cue(s) have more clause connectors compared to those with no negation cue; therefore it is very important to take clause connectors and dependent clauses in to consideration when trying to resolve a negation action.

In addition, the results confirmed that it is of great importance to take clause connectors and different types of clauses into account when performing relation extraction task.

This research undertook some experiments to use a few basic simplification methods (such as taking the main clause as a separate feature) to overcome issues associated with complex sentences; however, no significant improvement was achieved. It is believed that a combination of a simplification technique with a pronoun resolution method specifically-prepared-for-drug can improve the performance. To be used either in terms of a pre-processing step or along with other methods, such an algorithm may give better results. Moreover, the proposed method can be employed as an alternative approach to the sentence simplification error-prone task in the biomedical area.

Although the current results are promising, one of the challenging discussions is whether all kernel methods benefit from such features. As results of the subtree kernel for sentences with negation cues and clause connectors demonstrated, the authors believe that more advanced kernels deriving more informative features from different presentations of the sentence, may fail to benefit from the proposed features.

References

[1] Lazarou, J., Pomeranz, B. H. & Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. JAMA, vol. 279, no. 15, pp. 1200-1205.

[2] Gaulton, A. et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research, vol. 40, no. D1, pp. D1100--D1107.

[3] Ylva, B. et al. (2009). SFINX—a drug-drug interaction database designed for clinical decision support systems. European journal of clinical pharmacology, vol. 65, no. 6, pp. 627-633.

[4] Kadir, R. A. & Bokharaeian, B. (2013). Overview of biomedical relations extraction using hybrid rule-based approaches. Journal of Industrial and Intelligent Information Vol, vol. 1, no. 3.

[5] McDonald, R. et al. (2005). Simple Algorithms for Complex Relation Extraction with Applications to Biomedical IE. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, 2005, pp. 491-498.

[6] Rowan, M. & Harris, K. (1989). Explaining grammatical concepts. Journal of Basic Writing, pp. 21-41.

[7] Miwa, M., Saetre, R., Miyao, Y. & Tsujii, J. (2010). Entity-focused sentence simplification for relation extraction. In Proceedings of the 23rd international conference on computational linguistics, 2010, pp. 788-796.

[8] Giuliano, C., Lavellim ,A. & Romano, L. (2006). Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL '06), Trento, Italy, 2006, pp. 401-408.

[9] Vishwanathan, S. V. N. & Smola, A. (2003). Fast Kernels for String and Tree Matching. In Advances In Neural Information Processing Systems 15, 2003, pp. 569-576.

[10] Eugene, E. L., Anderson, S., Jr. Dwight D., & Douglas Wingate, J. (2004). Glossary of linguistic terms. Camp Wisdom Road Dallas: SIL International , 2004.

[11] Chapman, W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2002). Evaluation of Negation Phrases in Narrative Clinical Reports, 2002.

[12] Segura-Bedmar, I., Mart, P. (2011). The 1st DDIExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts. CEUR-WS, vol. 761, pp. 1-9.

[13] Wishart, D. S. et al. (2007). DrugBank: a knowledgebase for drugs, drug actions and drug targets," Nucleic acids research, 2007.

[14] Thomas, P., Neves, M. , Solt, I. , Tikk, D. & Leser, U. (2011). Relation Extraction for Drug-Drug Interactions using Ensemble Learning. In Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011), 2011, pp. 11-18.

[15] Segura-Bedmar, I. , Martinez, P. & Herrero-Zazo, M. (2013). SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, USA, 2013, pp. 341-350.

[16] Bokharaeian, B. & Diaz, A. (2013). NIL UCM: Extracting Drug-Drug interactions from text through combination of sequence and tree kernels. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Atlanta, Georgia, USA, 2013, pp. 644.

[17] Airola A. et al. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC bioinformatics, vol. 9 Suppl 11, 2008.

[18] Chowdhury, Md., Faisal, M. & Lavelli, A. (2013). Exploiting the Scope of Negations and Heterogeneous Features for Relation Extraction: A Case Study for Drug-Drug Interaction Extraction. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, June 2013, pp. 765-771.

[19] Blasco, G., Santiago S. M Mola-Velasco, Danger, R. & Rosso, P. (2011). Automatic drug-drug interaction detection: A machine learning approach with maximal frequent sequence extraction. Interactions, vol. 2397, no. 755, p. 3152.

[20] Danger, R., Segura-Bedmar, I., Martinez, P., & Rosso, P. (2010). A comparison of machine learning techniques for detection of drug target articles. Journal of biomedical informatics, vol. 43, no. 6, pp. 902-913.

[21] Chowdhury, F., Lavelli, A. & Kessler, F. (2013). Exploiting the Scope of Negations and Heterogeneous Features for Relation Extraction: A Case Study for Drug-Drug Interaction Extraction. In HLT-NAACL13, 2013, pp. 765-771.

[22] Szarvas, G., Vincze, V., Farkas, R. & Csirik, J. (2008). The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, 2008, pp. 38-45.

[23] Vincze, V., Szarvas, G., Mora, G., Ohta, T., & Farkas, R. (2011). Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. Journal of Biomedical Semantics, vol. 2, no. Suppl 5, p. S8, 2011.

[24] Siddharthan, A. (2014). A survey of research on text simplification. The International Journal of Applied Linguistics, pp. 259-98.

[25] Peng, Y., Tudor, C. O., Torii, M., Wu, C. H. & Vijay-Shanker, K. (2012). iSimp: A sentence simplification system for biomedicail text. In Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on, Oct 2012, pp. 1-6.

[26] Segura-Bedmar, I., Martinez, P. & Pablo-Sanchez, C. (2011). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. BMC Bioinformatics, vol. 12, no. Suppl 2, p. S1.

[27] Bokharaeian, B., Diaz, A., Neves, M., & Francisco,V. (2014). Exploring Negation Annotations in the DrugDDI Corpus. In Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, 2014, pp. 84-91.

[28] Segura-Bedmar, I., Martinez, P. & de Pablo-S, C. (2011). Using a Shallow Linguistic Kernel for Drug-Drug Interaction Extraction. Journal of Biomedical Informatics, vol. In Press, Corrected Proof, 2011.

[29] Joachims, T. (1999). Making large scale SVM learning practical. Universitat Dortmund, Tech. rep. 1999.

[30] McClosky, D., Surdeanu, M., & Manning, C. (2011). Event Extraction As Dependency Parsing for BioNLP 2011. In Proceedings of the BioNLP Shared Task 2011 Workshop, Stroudsburg, PA, USA, 2011, pp. 41-45.

[31] Pakzad, A. & Minaei Bidgoli, B. (2016). An improved joint model: POS tagging and dependency parsing. Journal of Artificial Intelligence & Data Mining, vol. 4, no. 1, pp. 1-8.

نشریه هوش مصنوعی و داده کاوی

استخراج تداخل دارویی از متن های پزشکی بوسیله تشخیص منفی سازی مبتنی بر زبان و وابستگی کلازی

بهروز بخاراییان[*] و آلبرتو دیاز

دانشکده مهندسی کامپیوتر، دانشگاه ملی مادرید ، مادرید ، اسپانیا .

ارسال ۲۰۱۶/۰۱/۱۹؛ پذیرش ۲۰۱۶/۰۵/۰۴

چکیده:

استخراج روابط پزشکی مانند تداخل داروها با یکدیگر (DDI) از متن یک زمینه تحقیق مهم در پردازش زبان‌های طبیعی زیست پزشکی است. با توجه به تعداد زیادی از جملات پیچیده در ادبیات زیست پزشکی ، محققان به کار بر روی برخی از جمله تکنیک‌های ساده‌سازی جملات برای بهبود کارآیی روش‌های استخراج رابطه مشغول بوده‌اند. از آنجا که این تسک، عمل دشواری است، بهبود قابل توجهی در ادبیات گزارش نشده است. هدف این مقاله به کشف نقش و توسعه‌مندی جهت بکارگیری وابستگی کلازی و منفی‌سازی مبتنی بر زبان در الگوریتم‌های استخراج رابطه می‌باشد.نتایج به دست آمده نشان می‌دهد که از طریق به کارگیری ویژگی‌های پیشنهادی عملکرد متدهای مبتنی بر کرنل با بکارگیری یک متد Bag-Of-Words، بهبود حاصل می‌شود. روش پیشنهادی می‌تواند به عنوان روش‌های جایگزین برای روش ساده‌سازی جملات در متن‌های پزشکی که یک کار مستعد خطا است استفاده شود.

کلمات کلیدی: تداخل دارو-دارو، استخراج رابطه، تشخیص منفی‌سازی، وابستگی کلازی.