Utilizing Various Natural Language Processing Techniques for Biomedical Interaction Extraction

Kyung-Mi Park*, Han-Cheol Cho* and Hae-Chang Rim*

Abstract—The vast number of biomedical literature is an important source of biomedical interaction information discovery. However, it is complicated to obtain interaction information from them because most of them are not easily readable by machine. In this paper, we present a method for extracting biomedical interaction information assuming that the biomedical Named Entities (NEs) are already identified. The proposed method labels all possible pairs of given biomedical NEs as INTERACTION or NO-INTERACTION by using a Maximum Entropy (ME) classifier. The features used for the classifier are obtained by applying various NLP techniques such as POS tagging, base phrase recognition, parsing and predicate-argument recognition. Especially, specific verb predicates (activate, inhibit, diminish and etc.) and their biomedical NE arguments are very useful features for identifying interactive NE pairs. Based on this, we devised a two-step method: 1) an interaction verb extraction step to find biomedically salient verbs, and 2) an argument relation identification step to generate partial predicate-argument structures between extracted interaction verbs and their NE arguments. In the experiments, we analyzed how much each applied NLP technique improves the performance. The proposed method can be completely improved by more than 2% compared to the baseline method. The use of external contextual features, which are obtained from outside of NEs, is crucial for the performance improvement. We also compare the performance of the proposed method against the co-occurrence-based and the rule-based methods. The result demonstrates that the proposed method considerably improves the performance.

Keywords—Biomedical Interaction Extraction, Natural Language Processing, Interaction Verb Extraction, Argument Relation Identification

1. INTRODUCTION

With the growing number of published biomedical research articles, medical researchers are faced with the difficulty of reading hundreds or thousands of research articles to find advances in the fields of their interest [1]. Although much work has been done to alleviate this problem by using search engines such as PubMed, the situation does not seem to be getting better due to the increasing number of newly published papers. Moreover, the redundant information problem in the biomedical domain is more severe than in the general domain and the traditional information retrieval techniques do not provide satisfactory solutions. To overcome these difficulties in the biomedical domain, the application of various natural language processing techniques are required.

Manuscript received November 25, 2010; accepted April 12, 2011.
Corresponding Author: Kyung-Mi Park
* Corresponding Authors: Dept. of Computer Science and Engineering, Korea University, Seoul, Korea ({kmpark; johanc;rim}@nlp.korea.ac.kr)

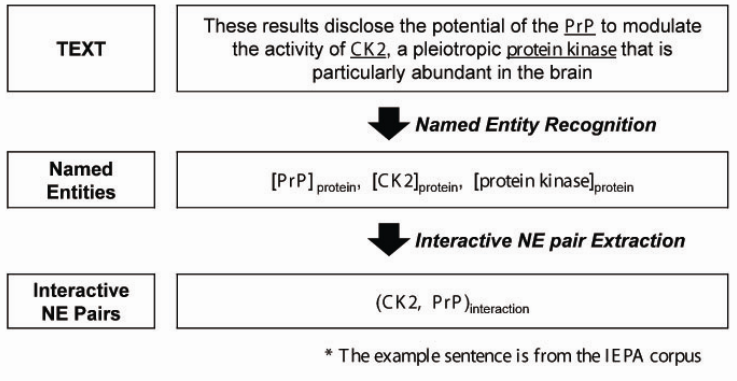

Fig. 1. The results of image after denoising with our method (δ=20)

The main tasks of biomedical information extraction can be summarized as: 1) the recognition of named-entities (NEs) and 2) the extraction of interactive NE pairs like Fig. 1. The NEs are the names of biomedically significant terms like *protein*, *gene*, *disease*, *treatment* and so on. The interactive NE pairs are the pairs of recognized NEs that have biomedical interaction between two NEs such as *activate*, *inhibit*, *diminish* and etc. In this paper, we propose a method for extracting the interactive NE pairs assuming that biomedical Named Entities (NEs) are already identified.

For the interaction information extraction, the proposed method firstly generates all possible pairs of identified NEs. Then it determines whether candidate NE pairs have interactions or not. The clues for identifying interaction between NEs usually exist in the context of each NE as shown in Table 1.

In the sentence (1), we can see the binary relation between two NEs, *insulin* and *phospholipase Cgamma 1*, located in the subject and object positions of the biomedical verb *stimulated*. It is relatively simple to recognize the relation between these NEs because *stimulated* is a definite clue word to identify the interaction. Sentence (2) is an example of the interaction that is represented by the noun, *prevention*. Nouns representing biomedical actions are also important clue words. The sentence (3) meaning *disease occurring after treatment* has an interaction between the NEs related to *side effects*. However, it is very difficult to identify the interaction between the given NE pair since explicit clue words do not exist for the *side effect* relationship. In this case, the word *following* can be an important clue word for the interaction extraction even though it is not a biomedical word and less decisive against *stimulated* or *prevention* in sentences (1) and (2). The sentence (4) meaning that *in order to reduce disease, treatment should be initiated* is an example that two NEs do not modify the same word. The first NE modifies *reduce* in the infinitive phrase and the second NE modifies *initiated* in the main clause. Therefore, it is

Table 1. Example sentences including an interaction NE pair

(1) Both EGF and *insulin*$_{\text{protein}}$ stimulated the accumulation of *phospholipase Cgamma 1*$_{\text{protein}}$ at the actin arc.
(2) *Statins*$_{\text{treatment}}$ for prevention of *stroke*$_{\text{disease}}$
(3) *Malignant mesodermal mixed tumor*$_{\text{disease}}$ of the uterus following *irradiation*$_{\text{treatment}}$
(4) To reduce the rate of *macrosomic infants in gestational diabetes*$_{\text{disease}}$, *glycemic control*$_{\text{treatment}}$ should be initiated before 34 gestational weeks.

necessary to discover that the two verbs *initiated* and *reduce* share a relationship in this sentence.

As we can see in the above example sentences, it is a complicated process to locate the right clue words in the context of NEs because they can appear in diverse natural language expressions. We employ various NLP techniques such as POS tagging, base phrase recognition, parsing, and predicate-argument recognition to solve this problem. These NLP techniques help to extract useful features from the biomedical texts to determine whether biomedical NE pairs have interactions or not. In the experiments, we show that these NLP techniques improve the performance of interaction extraction compared to the conventional approaches such as the co-occurrence-based and the rule-based approaches.

The remaining parts of this paper are organized as follows: Section 2 discusses previous approaches in biomedical interaction extraction. Section 3 explains the proposed method with its features extracted by using the NLP techniques mentioned above. Sub-section 3.3 especially elucidates how to produce partial predicate-argument structures with interaction verb extraction and argument relation identification methods. Finally, Section 4 provides experimental results of the proposed system and Section 5 concludes the paper.

2. RELATED WORK

There are several related works performing biomedical interaction extraction based on the co-occurrence-based approach, the rule-based approach, and the machine learning-based approach.

2.1 Co-occurrence-based Approach

(Craven, 1999) [2] presented a relation extraction method by training separate interaction extractors for each biomedical relation. This method first distinguishes the sentences including a specific interaction like *subcellular-location* relation from the other sentences and then extracts binary biomedical relations from these distinguished sentences. However, the target of the method is restricted to the sentences that have exactly two NEs of a specific binary relation. If a sentence has more than two NEs or more than one relation, this method cannot extract interaction pairs from the given sentence.

In (Ray and Craven, 2001) [3], they proposed an approach that incorporates the grammatical structure of sentences in the states of the Hidden Markov Models (HMMs). Compared with the previous work of Craven in [2], it also eliminates the sentences not including interactions. However, this method uses a phrase as a unit of state in a HMM model rather than a bag of words representation. Although the interaction extraction performance is improved against [2], it has two big shortcomings. First, it does not provide any means for finding the exact boundary of the NEs. The method only finds phrases that include NEs and considers extracted phrases as correct NEs if the NEs are included in the phrases. Therefore, the true performance is expected to be lower than the evaluated one. Second, we found that the interaction extraction method was task-dependent when it was applied to the Interaction Extraction Performance Assessment (IEPA) corpus during the experiment of comparison. The original interaction extraction method showed a lower performance than the basic extraction method, which only extracts interactions from the test corpus that existed in the training corpus.

Finally, the methods suggested by (Hahn et. al, 2002) [4] and (Srinivasan and Rindflesch, 2002) [5] assume that two NEs frequently co-occur in a biomedical text if one NE has a signifi-

cant influence over the other NE. The relations between NE pairs can be identified if the co-occurrence frequency of the NE pairs exceeds the predetermined threshold. Particularly, (Srinivasan and Rindflesch, 2002) uses the co-occurrence frequency of the MeSH[1] terms instead of the biomedical terms in order to alleviate the data sparseness problem.

The co-occurrence based approach shows good extraction performance for the known NE pairs that exist in the training corpus or in other external resources. However, it also has two weaknesses. First, the precision of extracted interaction NE pairs is low because this approach assumes that two NEs are related if they appear in the same sentence, even though they do not have a real relationship to one another. Secondly, the co-occurrence based methods do not seem to properly extract new interaction NE pairs because of the lack of co-occurrence information.

2.2 Rule-based Approach

The rule-based approach uses biomedical action words as key clues to find relations between NE pairs. For example, it extracts the *inhibit*-relation if two NEs, *I kappa B-alpha* and *transcription factor*, appear in the subject and the object positions of the biomedical action word *inhibit*.

In [6-8], the methods extract interactive NE pairs by recognizing noun phrases in the specific argument positions of predefined stem words like *bind* or *inhibit*. The interrelationships between two NEs are identified by using sequential automata for recognizing noun phrases, verb phrases, and their dependency. In the experiments, high precision can be achieved for the extracted NE pairs. However, the coverage of these methods is limited because the test data consists of the sentences involving specific stems such as *bind* or *inhibit*. The methods cannot recognize the NE pairs that are related to the binding or inhibitive action if these stems do not appear in the sentences.

The previous works in [9-10] extract NE pairs by applying patterns of the specific verbs that express biomedical actions. The patterns consist of lexical, syntactic, and semantic restrictions and are obtained manually and semi-automatically. Particularly, learning new patterns from the training corpus expands these patterns. They also adopted sentence normalization to reduce the problem of data sparseness. The semantic classes substitute for NEs, and POS tags or types of syntactic phrases replace words other than NEs. Although the pattern expansion step increases the coverage of the methods, the recall is still low due to the various sentence structures.

The rule-based approach usually shows high precision against the co-occurrence based and the machine learning based approaches. However, the recall of the rule-based approach is quite low because it is impossible to make patterns manually or automatically, in which can cover almost all kinds of diverse natural language expressions.

2.3 The Machine Learning-Based Approach

The machine learning-based approach obtains various lexical, syntactic, and semantic features from the training data and utilizes the features for training a classifier. A lot of machine learning techniques have been employed to make use of these features such as Naive Bayes, the decision tree, the neural network, the Hidden Markov Model (HMM), the Support Vector Machine (SVM) and the Maximum Entropy (ME) model. The trained classifier can determine whether

1 It is the ontology for biomedical domain, and is similar to the WordNet which is the ontology for general domain. It includes about nineteen thousands of the biomedical terms. The number of the top concepts in MeSH is 15 concepts

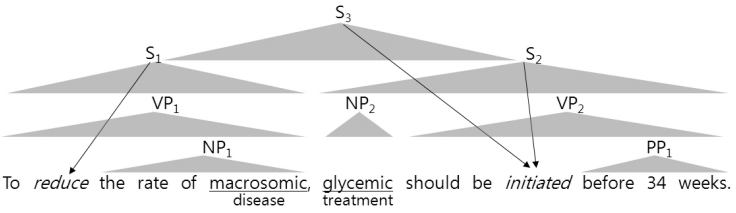

Fig. 2. An example of sentence structure

candidate NE pairs are interactive NE pairs or not.

In the previous work [11], Barbara Rosario and Marti A. Hearst proposed a neural network based method for distinguishing among seven relation types that can occur between NEs, and *treatment* and *disease*, in bioscience text. The features used for the classifier are as follows: words constituting NEs, POS-tags, syntactic categories, orthographical features, the MeSH ID of each word, and the sub-hierarchical information of MeSH. Although the method exploits rich features, it does not make use of the contexts outside NEs and mainly depends on MeSH. The experiment to assess the performance contribution of each feature shows that the MeSH features are the most important for performance improvement. However, manually built biomedical resources make it difficult to extract the latest interactive NE pairs.

The machine learning based methods utilizing contexts outside of NEs have also been suggested. In the previous work [12], their method uses the surrounding words of two NEs in a flat representation and a headword from a sub-tree including the NEs as features. These contextual features are very useful, especially for extracting the latest interactive NE pairs. The example sentence in Fig. 2 is a good example showing that contextual features help to determine the interactive relations between NEs. The example sentence has an interactive relation between two NEs *macrosomic* and *glycemic*. There are significant clue words such as *reduce* modified by *macrosomic* and *initiated* modified by *glycemic* outside the NEs. Particularly, *reduce* is the key contextual clue word since the *treatment-cure-disease* relation exists between two NEs.

However, the previous method fails to obtain useful contextual features for the existing interaction NE pair in this example. We are going to show how the previous method obtains contextual features and why it fails to work with the example sentence in Fig. 2. First, let us suppose that the features surrounding the NEs are defined as left adjacent three words and right adjacent three words. The features for the *glycemic* are *rate, of, macrosomic, should, be, initiated* and the features for the *macrosomic* are *the, rate, of, glycemic, should, be*. Although the features for the NE *glycemic* include the correct clue word *initiated*, the features for the NE *macrosomic*, do not have the correct clue word *reduce*. Secondly, the sub-tree including two NEs, *glycemic* and *macrosomic*, is S_3 containing the whole example sentence. The headword of S_3 is *initiated*, although the contextual feature is expected to be *reduce*. These problems happen because this method does not fully exploit possible contextual features.

3. NLP Features for the Machine Learning Based Biomedical Interaction Extraction

In this section, we propose a biomedical interaction extraction method based on the machine learning approach. This machine learning based approach has several merits that the co-

occurrence based and the rule-based approaches lack. First, it extracts the latest interaction NE pairs relatively well compared to the co-occurrence based approach since it uses contextual information rather than co-occurrence information between two NEs. Secondly, its coverage is broader than the rule-based approach because it can handle diverse natural language expressions by exploiting various linguistic features, rather than just applying highly precise but less broad rules. It also does not need human efforts to make a set of extraction rules like the rule-based approach.

The proposed method exploits various features obtained by applying NLP techniques such as part-of-speech (POS) tagging, base phrase recognition, parsing, and predicate-argument recognition. These rich features make the proposed method better for extracting interactive NE pairs precisely and broadly.

Section 3.1 briefly introduces the Maximum Entropy (ME) Model as a classification model. Section 3.2 explains the basic NLP features that have been employed and it has proven its usefulness in many applications. In Section 3.3, we describe the predicate-argument features in detail. Although the predicate-argument features are very useful and important high-level information, it is difficult to obtain. We elucidate a two-step method generating partial predicate-argument structures between biomedically salient verbs and their arguments.

3.1 Maximum Entropy Model

In the maximum entropy framework, the conditional probability of predicting an outcome o given a history h is defined as follows:

$$P(o \mid h) = \frac{1}{Z_\lambda(h)} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(h,o)\right) \tag{1}$$

where $f_i(h,o)$ is a binary-valued feature function, λ_i is the weighting parameter of $f_i(h,o)$, k is the number of features, and $Z_\lambda(h)$ is a normalization factor for $\Sigma_o P(o|h)=1$ [13]. In this study, the probability $P(o|h)$ is calculated by the weighted sums of active features (i.e. $f_i(h,o)=1$). For instance, a feature for our task can be represented by an indicating feature function, as follows:

$$f_i(h,o) = \begin{cases} 1 & if \quad Path = NP \uparrow S \downarrow VP \downarrow NP, \quad o = RELATION \\ 0 & otherwise \end{cases}$$

It means that it is likely to be the interaction between the NEs, when the syntactic path including the two NEs is $NP \uparrow S \downarrow VP \downarrow NP$. The maximum entropy classifier for the interaction extraction task classifies each NE pair into one of the following classes:

- *RELATION* class - a NE pair has a relation, or
- *NONE-RELATION* class – a NE pair does not have a relation.

3.2 Basic Features

In this section, we explain basic features obtained from words, base phrases, and parsed trees in detail. Table 2 also summarizes these features used in our study.

Table 2. Summary of Basic Features

WORD: - Two leftmost and rightmost words constituting each NE - A combination of first words and a combination of last words of two NEs - A maximum four words nearest to each NE between two NEs - The left two words of a left NE and the right two words of a right NE
POS: - POS of the above words
CHUNK: - The maximum of four headwords of base phrases appearing between the two NEs - The two headwords in the left side of the left NE and the right side of the right NE - The syntactic path from the left NE to the right NE - The number of VP, NP, and SBAR appearing between two NEs
PARSE TREE: - A syntactic path through the parse tree from the left NE to the right NE - A headword and its POS-tag of a sub-tree, which includes two NEs

WORD: Words constituting or neighboring NEs are informative features to determine a relation between NEs. We use all words satisfying the following condition as WORD features.

(1) Two leftmost words and two rightmost words of each NE
(2) A combination of first words and a combination of last words of two NEs of a candidate NE pair
(3) A maximum four words nearest to each NE between two NEs of a candidate NE pair
(4) The left two words of a left NE and the right two words of a right NE of a candidate NE pair

POS-tag: POS-tag features consist of POS-tags of WORD features' values. The POS-tag information alleviates the data sparseness problem resulting from WORD features. For the tagging, we re-implemented the TnT tagger [14] based on the trigram HMM model by using Penn II corpus and GENIA corpus as training corpora.

CHUNK: Base phrase recognition simplifies the syntactic structures and enables the selective use of the important words. Headwords neighboring NEs, the syntactic path between two NEs of a candidate NE pair, and the number of specific phrases between two NEs of a candidate NE pair are useful contextual information. We use the following information as CHUNK features:

(1) A maximum of four headwords of base phrases between two NEs of a candidate NE pair
(2) Two headwords in the left side of the left NE and two headwords in the right side of the right NE of a candidate NE pair
(3) The syntactic path between the two NEs of a candidate NE pair
(4) The number of specific base phrases such as VP, NP, and SBAR between the two NEs of a candidate NE pair

The range of value for the number of specific base phrases is from 0 to 2. Even if a specific base phrase appears more than twice, we use the value 2 in order to exclude unreliable values. We try to capture the characteristics of sentence structure that include two NEs. For example, if verb phrases do not appear between two NEs, it seems to represent that the two NEs are connected by a combination of noun phrases and prepositional phrases. If one VP phrase appears,

two NEs are likely to appear in the subject and the object positions of the corresponding verb. If more than one verb phrase exists, two NEs do not seem to have a relation in a given sentence structure. We use the YamCha chunker [15] with the CoNLL-2000 datasets as the training corpus. During the syntactic analysis, each NE identified during the identification process is regarded as one word to assign a phrase type to the NE.

PARSE TREE: By using Charniak's parser [16], we obtain the syntactic structure as shown in Fig. 2. Each NE is regarded as one word during the parsing. The following features are obtained from a parsed tree:

(1) The syntactic path of a parsed tree from a left NE to a right NE of a candidate NE pair

(2) A headword and its POS-tag of a sub-tree including two NEs of a candidate NE pair

Syntactic path information is a very informative feature. For example, we can easily assume that two NEs appear in the subject and the object positions of a verb if the syntactic path, $NP \uparrow S \downarrow VP \downarrow NP$, is given. In order to alleviate the data sparseness problem, we use a non-terminal symbol once when it repeats continuously. Therefore, the tree path in an example of Fig. 2 is $NP \uparrow S \downarrow VP \downarrow NP$.

3.3 Predicate-Argument Features

In this section, we describe the features obtained from the partial predicate-argument structures of a given biomedical sentence. The features are summarized in Table 3. In addition, the following subsections explain the interaction verb extraction step and the argument relation identification step. By using these two steps, the partial predicate-argument structures can be generated that indicate the dependencies between biomedically salient verbs and their arguments.

Table 3. Summary of Predicate-Argument Features

THE PREDICATE-ARGUMENT STRUCTURE:
- Verbs that have argument relations with constituents including a NE
- A conjoined feature of the biomedical verb and the last word of the NE

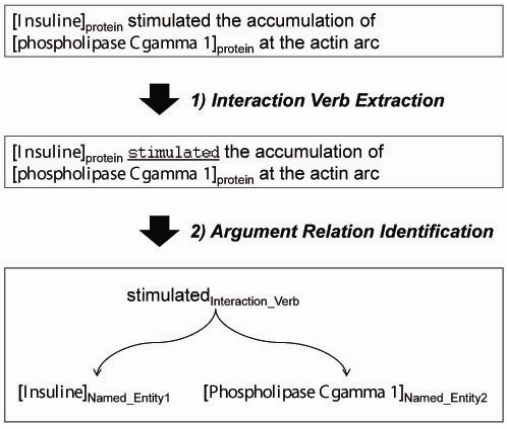

Fig. 3. Generating a Partial Predicate-Argument Structure

The predicate-argument structure provides very useful and important information to determine the existence of a relation between NEs. The predicate-argument features are as follows:

(1) Interaction verbs that have argument relations with constituents including a NE

(2) A conjoined feature of a biomedical verb and the last word of a NE

However, predicate-argument recognition is quite a challenging issue and the predicate-argument features used in the proposed method are a part of the entire predicate-argument structure. The example sentence in Fig. 3 shows how predicate-argument features can be obtained by using interaction verb extraction and argument relation identification steps. The details about these methods are described in the subsequent subsections.

3.3.1 Interaction Verb Extraction

The purpose of interaction verb extraction is to extract biomedically salient verbs that are likely to represent interactions between arguments. We consider the verbs appearing in a biomedical-domain corpus more frequently than in a general corpus as interaction verbs. The interaction verbs can provide key clues to find the proper NE pairs that include interactions. These interaction verbs also can be determined through corpus comparison.

In order to retrieve interaction verbs that appear more frequently in a biomedical-domain corpus rather than in a general corpus, we compute each verb's appearance probability both in a domain-specific corpus and in a general corpus respectively. We use the GENIA corpus as a biomedical-domain corpus. The number of words in the GENIA corpus is 441,886. We use the Wall Street Journal (WSJ) corpus as a general corpus. The number of words in the WSJ corpus is 1,349,202. From the estimated probabilities, we compute the relative frequency ratio of a verb *v* by the following equation:

$$RFR(v) = \frac{P_{GENIA}(v)}{P_{WSJ}(v)} \tag{2}$$

We regard a verb as an interaction verb if its relative frequency ratio is larger than the predetermined threshold value. We extract the verbs that appear more than 10 times in the GENIA corpus than in the WSJ corpus. The verbs that do not appear in the WSJ corpus are also regarded as interaction verbs.

3.3.2 Argument Relation Identification

In the argument relation identification step, retrieved interaction verbs are matched with their arguments. Since we do not need to know argument types (such as subject, object, etc.), argument relation identification only examines the existence of a relationship between an interaction verb and an argument candidate. Argument relation identification is implemented based on both partial parsing and full parsing.

Partial-Parsing Based Argument Relation Identification

For argument relation identification based on partial parsing, POS-tagging and chunking [17-18] are performed as a preprocessing step. Then, the dependencies between verb chunks and other chunks are analyzed and proper tags are assigned to dependent chunks. For the partial parsing based argument identifier, we employ SVM as a classification model. In addition, the

words that cause the data sparseness problem are replaced with the top concepts of WordNet or the POS-tags. For example, if an infrequent word has been assigned to two or more top concepts in WordNet, the first top concept is assigned. If it does not exist in WordNet, the POS-tag is used. Moreover, the proposed argument relation identifier only analyzes the existence of relations between verbs and noun phrases, and verbs and prepositional phrases including following noun phrases. Since a verb is not likely to be related to other chunks far from the verb, we limited the target chunks as follows: there may be a maximum of 1 intervening verb chunk between a target chunk to the left of the verb, and 0 between a target chunk to the right of the verb chunk. Detailed features are explained in the previous work [17].

We performed a small experiment to assess the effect of replacing an infrequent word with its WordNet top concept or POS-tag. The test data is in sections 10 to 19 of the WSJ corpus of the Penn Treebank II. These sections of the Penn Treebank II resulted in 239,472 vectors, and we conducted tenfold cross validation. In the experimental result, it was efficient to adopt the POS tags rather than the top concepts of WordNet in order to alleviate the data sparseness problem. Ultimately, the performance of argument relation extraction was a recall of 86.86%, a precision of 90.00% and an F1-score of 88.40%.

Full-Parsing Based Argument Relation Identification

Argument relation identification can be done with full parse information in a similar manner. We implemented the full parsing based argument relation identifier with the maximum entropy model. The features used for the identifier are as follows:

(1) The POS of the predicate
(2) The syntactic category of the parse constituent's parent node
(3) The syntactic path from the parse constituent to the predicate
(4) The phrase structure rule that expands the predicate's parent node in the tree

In this approach, we also applied *tree distance restriction*. This restriction reduces the search space and hardly sacrifices the performance of argument relation identification.

To test the implemented identifier, we performed a small experiment with CoNLL-2005 datasets (Wall Street sections 02-21 as training set, Charniak' trees.) By following the standard partition used in parsing, we used sections 02-21 for training and section 23 for test. Experimental results showed that our system obtained an F1-score of 81.44% on the test data [19].

4. EXPERIMENTS

To test the proposed method, we have experimented with Iowa State University's Interaction Extraction Performance Assessment (IEPA) corpus [20-21]. This corpus has 486 sentences taken from 303 abstracts. Each sentence contains at least one pair of biochemicals of interest. For building classifiers, we utilized Zhang le's MaxEnt toolkit, and the L-BFGS parameter estimation algorithm with Gaussian Prior smoothing [22].

4.1 Experimental Results

For analyzing the performance improvement of the proposed system, it is useful to estimate the relative contribution of the each feature. Table 4 shows the performance of various feature

Table 4. Performance of various feature combinations in the interaction extraction task

Method	Recall	Precision	F1-score
ALL	88.91	85.36	87.10
ALL-{POS}	88.96	85.47	87.18
ALL-{chunk1_between}	88.63	84.79	86.67
ALL-{chunk2_external}	88.74	85.01	86.83
ALL-{chunk3_path}	88.97	85.44	87.17
ALL-{chunk4_vp,np,sbar}	88.67	85.48	87.05
ALL-{parse1_path}	88.92	85.38	87.11
ALL-{parse2_head_POS}	88.87	85.34	87.07
ALL-{arg1_verb}	88.67	85.00	86.80
ALL-{arg2_verb_last}	88.63	84.93	86.74

combinations on the interaction extraction task. We can notice that the performance deteriorates by leaving out one feature at a time in this table. A brief explanation about the leaving out features in Table 4 is provided as follows:

POS: Features obtained from POS-tags of WORD features
chunk1_between: A maximum of four headwords of base chunks appearing between the two NEs
chunk2_external: Two headwords in the left side of the left NE and the right side of the right NE respectively
chunk3_path: Syntactic path from the left NE to the right NE
chunk4_vp,np,sbar: The number of VP, NP, and SBAR chunks appearing between two NEs
parse1_path: The syntactic path through the parse tree from the left NE to the right NE
parse2_head_POS: A headword and its POS-tag of a sub-tree which includes two NEs
arg1_verb: Verbs that have argument relations with two NEs
arg2_verb_last: Conjoined strings of biomedical verbs and the last words of their argument NEs

The removing *chunk1_between* feature has the most effect on the performance, while removing the *POS* feature has the least effect. According to the experimental results, we found that the left and right contexts of NEs are effective features to improve the interaction extraction performance.

Table 5 shows the effect of each NLP step on the interaction extraction performance. The *WORD* feature set explained in the Section 3 is used for the baseline system without utilizing any NLP techniques. The *POS*, *chunk*, *parse tree* and *predicate-argument structure* feature sets in Tables 2 and 3 also correspond to the feature sets explained in Section 3. In the experiment, the *POS* feature set deteriorates the interaction extraction performance. The *chunk* feature set is most effective and the *predicate-argument structure* feature set is also effective. The interaction extraction performance is improved by about 2% by using the *chunk* and *predicate-argument structure* feature sets.

Table 6 compares the performance of the interaction extraction with the predicate-argument feature structure set based on the two different parsing methods. The full-parsing based method utilizes the output of the full parser of Charniak. The partial-parsing based method is based on the parser, which is implemented by performing the pre-processing with POS tagging and chunking. The experimental results show that the full-parsing based method is better than the partial-parsing based method. However, the performance improvement of the full-parsing based method is not satisfactory in spite of the high computational cost.

Table 5. Performance of various feature combinations on the interaction extraction task

Method	Recall	Precision	F1-score
Baseline	87.24	83.17	85.16
Baseline + POS feature set	87.17	83.02	85.04
Baseline + chunk feature set	89.00	84.75	86.82
Baseline + chunk and parse tree feature sets	89.01	84.78	86.84
Baseline + chunk and predicate-argument structure feature sets	89.21	85.30	87.21

Table 6. Comparison of interaction extraction performance by using the predicate-argument feature set based on full-parsing and partial-parsing manners

Method	Recall	Precision	F1-score
Full-parsing based method	89.21	85.30	87.21
Partial-parsing based method	89.11	85.07	87.04

4.2 Results of Comparison

The performance of the proposed system is also compared to the co-occurrence based and the rule based systems as shown in Table 7. We re-implemented Ray and Craven's system [3] and Pustejovsky's system [7-8] for the comparison. A brief explanation of these re-implemented systems is provided as follows:

The co-occurrence based system: For the reimplementation of Ray and Craven's system [3], we used the TnT tagger [14] and the YamCha chunker [15]. The Sundance parser [23], which was originally used for their system, was not available for us at the time of the experiment. We also used a small fixed probability for each unknown word for the smoothing instead of using m-estimates, because the paper gave no information about a specific m value. Lastly, we tested a re-implemented method with a 10-fold cross validation. The test was performed twice with different interaction extraction methods. The performance for the new corpus, IEPA corpus, was quite low because the original interaction extraction method was task dependent; the average recall is 47.30%, precision is 39.75%, and the F1-score is 43.20%. The basic interaction extraction method, which only extracts interactions from the test corpus that existed in the training corpus, demonstrated the better performance; the average recall is 65.42%, precision is 51.90%, and the F1-score is 57.88%.

The rule-based system: For the experiment of Pustejovsky's system [7-8], the sentences that include the stems *activate*, *inhibit* and *bind* were extracted. The POS-tagging and chunking were performed on these sentences and base phrases were identified. To find the probable phrase pairs, which include an interrelationship between NEs, we applied phrase patterns extracted from the training data. For example, if the adjacent NP phrases of the biomedical stem *inhibit* includes NEs and the extraction pattern, *NP-VP-NP*, exists, it extracts the relation.

Table 7. Comparison of the proposed system with the co-occurrence-based and the rule-based systems

Method	Recall	Precision	F1-score
The co-occurrence-based system	65.42	51.90	57.88
The rule-based system	71.65	88.16	79.05
The proposed system	89.21	85.30	87.21

5. CONCLUSION

In this paper, we proposed an interaction extraction method that exploits rich features. These features can be obtained by utilizing various NLP techniques such as POS tagging, base phrase recognition, parsing, and predicate-argument recognition. Especially, we devised a two-step method to generate the partial predicate-argument structure. It finds interactive verb predicates first, and then locates their arguments. This method effectively reduces the computational cost of generating the entire predicate-argument structure of a given sentence and barely harms the interaction extraction performance.

Experimental results show that the proposed system obtains an 87.21% F1-score. Based on the results, we can conclude that various NLP techniques generating features that have different characteristics consistently improve the interaction extraction performance. Moreover, we precisely analyzed the contribution of each feature and NLP technique for the performance improvement as shown in Tables 4 and 5.

The results of comparison demonstrate that the performance of the proposed method is better than the performance of the re-implemented systems. We believe that the machine learning approach using rich features complements the weaknesses of the co-occurrence based and the rule based approaches.

REFERENCES

[1] A. Madkour, K. Darwish, H. Hassan, A. Hassan, and O. Emam, "BioNoculars: Extracting Protein-Protein Interactions from Biomedical Text", Association for Computational Linguistics, 2007.

[2] M. Craven, "Learning to extract relations from Medline", In Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction, 1999.

[3] S. Ray, and M. Craven, "Representing sentence structure in Hidden Markov Models for information extraction", In Proceedings of the International Joint Conference on Artificial Intelligence, 2001.

[4] U. Hahn, M. Romacker, and S. Schulz, "Creating Knowledge Repositories from Biomedical Reports : The MEDSYNDIKATE Text Mining System", In Proceedings of the Pacific Symposium on Biocomputing, 2002.

[5] P. Srinivasan, and T. Rindflesch, "Exploring text mining from Medline", In Proceedings of the American Medical Informatics Association Symposium, 2002.

[6] T. Rindflesch, L. Hunter, and A. Aronson, "Mining molecular binding terminology from biomedical text", In Proceedings of the American Medical Informatics Association Symposium, 1999.

[7] J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, and B. Cochran, "Robust relational parsing over biomedical literature: Extracting inhibit relations", In Proceedings of the Pacific Symposium on Biocomputing, 2002.

[8] J. Pustejovsky, J. Castano, R. Sauri, A. Rumshinsky, J. Zhang, and W. Luo, "Medstract: Creating large-scale information servers for biomedical libraries", In Proceedings of the ACL-02 the Workshop on Natural Language Processing in the Biomedical Domain, 2002.

[9] C. Friedman, P. Kra, H. Yu, and M. Krauthammer, A. Rzhetsky, "GENIES : A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles", Bioinformatics, 2001.

[10] R. Feldman, Y. Regev, M. Finkelstein-Landau, E. Hurvitz, and B. Kogan, "Mining biomedical literature using information extraction", Current Drug Discovery, 2002.

[11] B. Rosario, and M. Hearst, "Classifying semantic relations in bioscience texts", In Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2004.

[12] J. Xiao, J. Su, G. Zhou, and C. Tan, "Protein-Protein Interaction Extraction: A Supervised Learning Approach", In proceedings of the 1st International Symposium on Semantic Mining in Biomedicine,

2005.
[13] A. Berger, S. Pietra, and V. Pietra, "A maximum-entropy approach to natural language processing", Computational Linguistics, 1996.
[14] T. Brants, "TnT - A statistical Part-of-Speech Tagger", In Proceedings of the 6th Applied Natural Language Processing, 2000.
[15] T. Kudoh, and Y. Matsumoto, "Use of support vector learning for chunk identification", In Proceedings of the 3rd Conference on Natural Language Learning, 2000.
[16] E. Charniak, "A Maximum-Entropy-Inspired Parser", In Proceedings of the North American Chapter of the Association for Computational Linguistics, 2000.
[17] K.M. Park, Y.S. Hwang, and H.C. Rim, "Two-Phase Semantic Role Labeling based on Support Vector Machines", In Proceedings of the 7th Conference on Natural Language Learning, 2004.
[18] S. Buchholz, "Memory-Based Grammatical Relation Finding", PhD. Thesis, Tilburg University, 2002.
[19] K.M. Park, and H.C. Rim, "Maximum Entropy based Semantic Role Labeling", In Proceedings of the 8th Conference on Natural Language Learning, 2005.
[20] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele, "Mining Medline: abstracts, sentences, or phrases?", In Proceedings of the Pacific Symposium on Biocomputing, 2002.
[21] J. Ding, D. Berleant, J. Xu, and A. Fulmer, "Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser", In Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, 2003.
[22] S. Chen, and R. Rosenfeld, "A Gaussian prior for smoothing maximum entropy models", Technical Report CMUCS-99-108, Carnegie Mellon University, 1999.
[23] E. Riloff, "The Sundance sentence analyzer", http://www.cs.utah.edu/projects/nlp/, 1998.
[24] J.D. Kim, T. Ohta, and J. Tsujii, "Corpus annotation for mining biomedical events from literature", BMC Bioinformatics, 2008.

Kyung-Mi Park
She received her BS and MS degrees in Computer Science and Engineering from Yonsei University in 2000 and 2002, respectively. She received a Ph.D. degree in Computer Science and Engineering from Korea University in 2008. Her research interests include natural language processing, information extraction, and text mining.

Han-Cheol Cho
He received his BS and MS degrees in Computer Science and Engineering from Korea University in 2004 and 2006, respectively. His research interests are in the area of natural language processing, information extraction, and text mining. This includes topics such as word segmentation, machine learning, and knowledge representation.

Hae-Chang Rim
He received his Ph.D. degree in Computer Science from University of Texas in Austin, Texas in 1990. He has been a professor at Korea University since 1991. His research interests are in the area of natural language processing, Korean language engineering, and information retrieval. This includes topics such as information extraction, machine translation, and dialogue management.