LETTER

Semantic Classification of Bio-Entities Incorporating Predicate-Argument Features

Kyung-Mi PARK[†a)] ***and*** **Hae-Chang RIM**[†b)]**,** ***Members***

SUMMARY In this paper, we propose new external context features for the semantic classification of bio-entities. In the previous approaches, the words located on the left or the right context of bio-entities are frequently used as the external context features. However, in our prior experiments, the external contexts in a flat representation did not improve the performance. In this study, we incorporate predicate-argument features into training the ME-based classifier. Through parsing and argument identification, we recognize biomedical verbs that have argument relations with the constituents including a bio-entity, and then use the predicate-argument structures as the external context features. The extraction of predicate-argument features can be done by performing two identification tasks: the biomedically salient word identification which determines whether a word is a biomedically salient word or not, and the target verb identification which identifies biomedical verbs that have argument relations with the constituents including a bio-entity. Experiments show that the performance of semantic classification in the bio domain can be improved by utilizing such predicate-argument features.

key words: *semantic classification, predicate-argument feature, biomedical verb, maximum entropy model*

1. Introduction

The clues for bio-entity class tagging are words themselves constituting a bio-entity, but about 21% of the words in the training data had characteristics of belonging to more than two classes. Therefore, for alleviating the ambiguous classification problem, it is necessary to utilize the words located on outside context of bio-entities. There were several related works of using external context in order to improve the performance of semantic classification system. [1] suggested a feature set with the inside and the outside contextual information of the bio-entity. In case of outside context, it chose the surrounding 2 words in a flat representation supporting the meaning of the given entity. [2] also defined the feature set as being the last five words in each bio-entity, together with the left and right two words around the bio-entity. However, according to our experimental results, the use of outside context was not effective on the improvement of performance.

In this study, we propose predicate-argument features extracted through parsing and argument identification. We assumed that verbs in a sentence are very helpful for classifying a bio-entity, since each semantic class which includes bio-entities is related to different biomedical actions, such as activation, inhibition, transcription, and dissociation, and such actions are mainly described by the verbs in the sentence. However, since a bio-entity is not related to all verbs in the sentence, it is necessary to identify whether the bio-entity is related to each verb in the sentence or not. Through parsing and argument identification, we try to identify *biomedical verbs** that have argument relations with the constituents including an entity. If a bio-entity occurs inside the boundaries of a parse constituent and the constituent has an argument relation with a biomedical verb, we regard that the bio-entity is related to the biomedical verb. In order to automatically recognize the *target verbs*** that are useful clues for estimating relatedness between biomedical verbs and bio-entities, we perform salient word identification and target verb identification. For identifying the biomedically salient words, we perform corpus comparison between a bio domain-specific corpus and a general corpus. For identifying the target verbs, we implement a predicate-argument recognizer based on the ME model by using PropBank corpus. Section 3 presents a detailed description of the identification tasks.

2. ME-Based Semantic Classification

In this section, we present a semantic classification method based on an ME model. We first explain the ME model, and then describe the proposed features for training the ME model for the classification task.

2.1 Maximum Entropy Model

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

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

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 [3]. In this study, the probability $P(o|h)$ is calculated by the

Manuscript received August 1, 2007.
Manuscript revised December 4, 2007.

[†]The corresponding authors are with the Department of Computer Science & Engineering, Korea University, Korea.
a) E-mail: kmpark@nlp.korea.ac.kr
b) E-mail: rim@nlp.korea.ac.kr

DOI: 10.1093/ietisy/e91–d.4.1211

*If some words are biomedically salient and POS type of the words given by the Charniak's parser is *verb*, we regard the words as *biomedical verbs*.

**We call biomedical verbs that have argument relations with the constituents including a bio-entity *target verbs*.

weighted sums of active features (i.e. $f_i(h, o)$=1). For instance, a feature for our task can be represented by the following feature function:

$$f_i(h, o) = \begin{cases} 1 & \text{if } w\_r_1 = gene, o = DNA \\ 0 & \text{otherwise} \end{cases}$$

It means that the semantic class of a bio-entity is likely to be DNA, when the first word from the rightmost of the entity is *gene*. The ME classifier for our task classifies the bio-entity into one of the following classes: *protein*, *DNA*, *RNA*, *cell_line*, *cell_type*. The set of feature functions will be described in the next section.

2.2 Features

In order to train the ME model, we use the following features. We first use the standard features introduced by previous works. Especially, we use word variations and morphological patterns for alleviating the data sparseness problem of the words. The features represent the most informative part of a word obtained by the orthographic characteristics of the word. We also use new features extracted from predicate-argument structures. The features represent proper external contexts for alleviating the ambiguous problem in which some words are used in more than two classes.

2.2.1 Standard Features

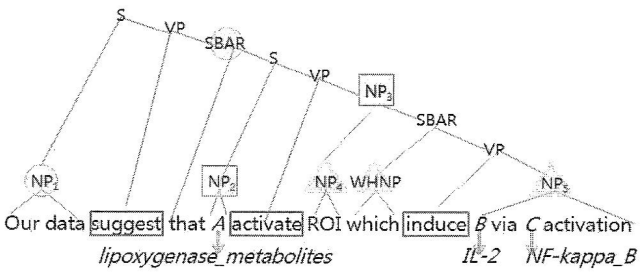
Word: The words themselves constituting the bio-entity play a significant role in classifying the bio-entity into a proper semantic class. Specially, the functional words play a key role in classification [1]. In general, functional words are often located on the rightmost of a bio-entity. Thus, we thought that the clues for class tagging are the words on the rightmost rather than the words on the leftmost of the bio-entity, and then used 1 word from the leftmost and 3 words from the rightmost of the bio-entity as features for learning the ME-based classifier. We proved that such intuition is correct in our experimentation as shown in Table 4.

Word Variation: To make a word variation, we alter all capital letters to lower letters and substitute # for numbers. Also, for segmenting a word into parts, we regard a symbol like hyphen as blank. Among the parts of the word, we select the longest part as a word variation which may contain more useful information. For example, *IL-2* changes to *il # #* using the above-mentioned rules, and *il* is extracted as a word variation.

Morphological Pattern: Morphological patterns reflect orthographical characteristics of words. When a target word infrequently occurs in the data, morphological patterns can alleviate the data sparseness problem of the word. In order to extract internal morphological patterns of words, we select every possible prefix or suffix of words when they consist of more than three characters, and then we compute the relative entropy of each substring in order to discriminate informative prefixes and suffixes [4], [5]. Table 1 shows the examples of patterns of words constituting bio-entities.

Table 1 Examples of morphological patterns.

word	prefix	suffix
IL-2	il-	il-2
gene	gen	ene
5-lipoxygenase	5-l	5-lipoxygenase
CD28	cd28	cd28
surface	surfac	rface
receptor	receptor	eceptor



Bio-entity	Features
lipoxygenase metabolites	tv=activate, tv_lw=activate_metabolites, tv=suggest, tv_lw=suggest_metabolites
IL-2	tv=induce, tv_lw=induce_IL-2, tv=activate, tv_lw=activate_IL-2, tv=suggest, tv_lw=suggest_IL-2
NF-kappa_B	tv=induce, tv_lw=induce_B, tv=activate, tv_lw=activate_B, tv=suggest, tv_lw=suggest_B

Fig. 1 An example of predicate-argument structures.

2.2.2 New Features

Predicate-Argument Structure: By using the Charniak's parser [6] and our system that performs argument identification, we obtain the syntactic structure as shown in Fig. 1. In Fig. 1, *suggest*, *activate*, and *induce* indicated by rectangle marks represent the biomedical verbs. We perform the syntactic analysis using the Charniak's parser by regarding words constituting a bio-entity as one word †. The positions *A*, *B*, and *C* indicating one word in Fig. 1 correspond to the bio-entities *lipoxygenase_metabolites*, *IL-2*, and *NF-kappa_B* respectively. In Fig. 1, the arguments of *suggest* are NP_1 and *S BAR* indicated by round marks, the arguments of *activate* are NP_2 and NP_3 indicated by square marks, and the arguments of *induce* are NP_4, *WHNP*, and NP_5 indicated by triangle marks. We use target verbs and the target verbs conjoined with the last word of the bio-entity as features. For example, since parse constituents NP_2 and *S BAR* include the given bio-entity *lipoxygenase_metabolites*, and the constituents are related to *activate* and *suggest* respectively, we extract the features for the bio-entity as shown in

†We assume that the biomedical named entity (NE) recognition task is divided into two phases and the boundaries of bio-entities are previously given. When we perform the syntactic analysis using the Charniak's parser, we want to assign a parse constituent to a bio-entity's parent node. Therefore, we regard words constituting the bio-entity as one word by using the boundaries of the bio-entity in the training data.

Table 2 Examples of biomedically salient words.

word	#WSJ	#GENIA	RFR
suggest	61	445	22.27
activate	1	201	613.71
induce	13	257	60.36
show	312	388	3.80
have	3,990	966	0.74

the lower part of Fig. 1.

3. Extraction of Predicate-Argument Features

In this section, we describe an automatic extraction method of predicate-argument features for semantic classification. First, we identify biomedically salient words occurring more frequently in the training corpus than in a general corpus. Next, if POS tag of the biomedically salient words is *verb*, and the words have argument relations with the constituents including the bio-entity, we identify the words as target verbs.

3.1 Biomedically Salient Word Identification

The words constituting bio-entities occur more frequently in a bio domain-specific corpus than in a general corpus. We regard the words as salient words. In order to recognize these salient words, we compute each word's probabilities both in the training corpus and in a general corpus respectively. From the estimated probabilities, we compute the relative frequency ratio (RFR) of a word w by Eq. (2).

$$RFR(w) = \frac{P_{GENIA}(w)}{P_{WSJ}(w)} \quad (2)$$

We regard the word as a biomedically salient word when its RFR is more than the predetermined threshold value [†]. The words that do not occur in the WSJ corpus are also regarded as salient words. Table 2 shows the RFR of the verbs in the example sentence of Fig. 1.

3.2 Target Verb Identification

In order to identify target verbs, we use the Charniak's parser and our system that performs argument identification. Our system utilized the output of full parser of Charniak. We implemented a predicate-argument recognizer based on an ME model [7]. Our system identified parse constituents in the sentence that represent valid semantic arguments of an identified biomedical verb. We assumed that most semantic arguments occur inside the specific distance on the parse tree from the parse constituent's parent node to the predicate's parent node. In order to reduce the number of candidate arguments, we incorporated tree distance restriction into pre-processing of argument identification task. In order to train the ME model for the identification task, we consider the following features.

- **path**: this is the syntactic path through the parse tree from the parse constituent to the predicate.
- **sub_cat**: this is the phrase structure rule expanding the predicate's parent node in the tree.
- **pred_POS**: this is POS of the predicate.
- **head_phr**: this is the syntactic category of the parse constituent's parent node.

The ME classifier for the task classifies each parse constituent into one of the following classes: ARG or NON-ARG. To test our system, we have experimented with CoNLL-2005 datasets which originated from the PropBank-1.0. By following the standard partition used in parsing, we used sections 02-21 for training and Sect. 23 for test. Experimental results show that our system obtains an F-score of 81.44% on the test data.

Table 3 Number of words belonging to more than two classes.

#class	#word
1	8,362(78.55%)
2	1,643(15.43%)
3	428(4.02%)
4	150(1.41%)
5	63(0.59%)

4. Experiments

To evaluate the proposed method, we have experimented with JNLPBA-2004 datasets [††] [8]. The training data came from the GENIA version 3.02 corpus, and the test data was a newly annotated collection of Medline abstracts from the GENIA project . Because the given training data considers only a small set of semantic classes of the GENIA corpus, biomedical entities not corresponding to 5 specific classes are ignored in training the ME-based classifier for semantic classification. To produce a classifier, we utilized the Zhang le's MaxEnt toolkit , and the L-BFGS parameter estimation algorithm with Gaussian Prior smoothing [9].

4.1 Experimental Results

As shown in Table 3, about 21% of the words in the training data have characteristics of belonging to more than two classes. For example, in order to describe bio-entity names, the number of words used in all classes is 63. Table 4 shows the performance according to the context variance used as word features on the semantic classification task. For example, *(L1, R3)* represents the performance of using 1 word from the leftmost and 3 words from the rightmost of the bio-entity, and *(R4)* represents the performance of using only 4 words from the rightmost of the bio-entity. *(L1, R3)* is the most effective on the performance. Also, experimental results show that the performance of using only words from

[†]We extract the words which occur more than 10 times in the training corpus than in the WSJ corpus.

[††]http://research.nii.ac.jp/ collier/workshops/JNLPBA04st.htm

Table 4 Performance according to the context variance.

#word	#Left,#Right	Accuracy
3	(L3)	88.06
	(R3)	91.62
	(L2, R1)	91.21
	(L1, R2)	91.02
4	(L4)	89.56
	(R4)	91.63
	(L3, R1)	91.28
	(L1, R3)	**91.84**
	(L2, R2)	91.01
5	(L5)	90.33
	(R5)	91.47
	(L4, R1)	91.49
	(L1, R4)	91.50
	(L3, R2)	91.28
	(L2, R3)	91.42

Table 5 Effects of using word variation and m-pattern.

method	Accuracy
word	91.84
word + word variation	91.54
word + m-pattern	**91.94**
word + word variation + m-pattern	91.61

Table 6 Classification accuracy when the words located on the external contexts of bio-entities are used.

method	Accuracy
word + m-pattern	91.94
word + m-pattern with adjacent one word	91.47
word + m-pattern with adjacent two words	90.90

Table 7 Effects of using predicate-argument features.

method	Accuracy
word + m-pattern	91.94
word + m-pattern + predicate-argument	**92.33**

the rightmost is better than the performance of using only words from the leftmost in our experimentation. Table 5 shows the effects of using word variation and morphological pattern (m-pattern) for alleviating the data sparseness problem of the word. We take the system with using only word features as a baseline. The word variation and morphological patterns are additionally used on the baseline system respectively. In our experiments, usage of word variation deteriorates the performance. However, we obtain the performance improvement by using morphological patterns. Table 6 shows the classification accuracy when the words located on the left and the right contexts of bio-entities are additionally used. According to the experimental results, the use of the external contexts in a flat representation are not effective on the improvement of performance. Table 7 shows the effects of using predicate-argument features. We obtain the performance improvement of about 0.42% by using predicate-argument features.

5. Conclusion

We have presented a semantic classification method based on an ME model. For alleviating the data sparseness problem of the word, we have added morphological patterns extracted from the training data to a feature set. For alleviating the ambiguous problem in which some words are used to more than two biomedical semantic classes, we have proposed predicate-argument features. Experimental results show that our system achieves an accuracy of 92.33% and that the introduction of predicate-argument features improves the performance of our system as compared with the baseline performance. The addition of morphological patterns is also effective on the improvement of performance. As a future work, we will apply predicate-argument features extracted through parsing and argument identification to biomedical relation extraction task.

References

[1] K.J. Lee, Y.S. Hwang, S.H. Kim, and H.C. Rim, "Biomedical named entity recognition using two-phase model based on SVMs," Journal of Biomedical Informatics, vol.37, no.6, pp.436–447, 2004.
[2] D. Yua and P. Srinivasan, "An applied machine learning technique for bio-entity classification," Information and Health at Iowa, 2004.
[3] A. Berger, S. Pietra, and V. Pietra, "A maximum entropy approach to natural language processing," Computational Linguistics, vol.22, no.1, pp.39–71, 1996.
[4] K.M. Park, S.H. Kim, and H.C. Rim, "ME-based biomedical NE recognition using lexical knowledge," ACM Trans. Asian Language Information Processing, vol.5, no.1, pp.4–21, 2006.
[5] J.D. Kim and J. Tsujii, "Word folding: Taking the snapshot of words instead of the whole," Proc. IJCNLP, 2004.
[6] E. Charniak, "A maximum-entropy-inspired parser," Proc. NAACL, 2000.
[7] K.M. Park and H.C. Rim, "Maximum entropy based semantic role labeling," Proc. CoNLL, pp.209–212, 2005.
[8] J.D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," Proc. JNLPBA, 2004.
[9] S. Chen and R. Rosenfeld, "A Gaussian prior for smoothing maximum entropy models," Technical Report CMUCS-99-108, Carnegie Mellon University, 1999.