Incorporating Lexical Knowledge into Biomedical NE Recognition

Kyung-Mi Park Seon-Ho Kim Ki-Joong Lee Do-Gil Lee Hae-Chang Rim

Department of Computer Science & Engineering, Korea University

5-ka, Anam-dong, SEOUL, 136-701, KOREA

{kmpark, shkim, kjlee, dglee, rim}Onlp.korea.ac.kr

Abstract

In this paper, we propose a two-phase biomedical named entity(NE) recognition method based on SVMs. We first recognize biomedical terms, and then assign appropriate semantic classes to the recognized terms. In the two-phase NE recognition method, the performance of term recognition is critical to the overall performance of the system because term recognition errors can be propagated to the semantic classification phase. In this study, we try to improve the performance of term recognition by using lexical knowledge. We utilize salient NPs and collocations as lexical knowledge extracted from raw corpus. In addition, we use morphological knowledge extracted from training data as features for learning SVM classifiers. Experimental results show that our system obtains an F-measure of 62.97% on the test data, and that the performance can be improved up to 2.82%by using lexical knowledge.

1 Introduction

Biomedical NE recognition can be divided into two phases: the term recognition phase which finds the boundaries of terms for a given sentence, and the semantic classification phase which determines the semantic classes of the terms. With two-phase NE recognition method, we can alleviate the unbalanced class distribution problem by decreasing the number of negative examples, which is much larger than positive examples. Also, we can utilize different feature sets appropriate for each phase. However, since the output of recognition phase is used as the input of classification phase, the coverage of recognition phase needs to be high. Furthermore, since recognition errors can be propagated to the next phase, it is important to improve the performance of term recognition.

In this paper, we propose a two-phase biomedical NE recognition method based on SVMs (K. J. Lee, 2003). In order to improve the accuracy of term recognition, we try to incorporate lexical knowledge into pre-processing and post-processing steps of the term recognition phase. In these steps, we use lexical knowledge extracted from raw corpus. Also, we use morphological knowledge extracted from training data as features for the SVM learning.

The remaining part of this paper is organized as follows: In section 2, we explain the proposed method of two-phase biomedical NE recognition. In section 3, we describe an automatic extraction method of lexical and morphological knowledge. Finally, we show some experimental results in section 4, and conclude the paper in section 5.

2 Biomedical NE Recognition

As shown in Figure 1, the proposed NE recognition method consists of three-steps term recognition phase and the classification phase. In the pre-processing step, input sentences are POStagged and NP-chunked, and then noun phrases that include biomedically salient words such as biochemical, pathways, activation, human, cells, and R24 are identified. In the boundary identification step, one of B, I, O class is assigned to words in the identified noun phrases. In the post-processing step, collocation relations between words inside the boundary and words outside the boundary are identified. Finally, we classify the term identified by B and I as one of protein, DNA, RNA, cell_line, or cell_type in the semantic classification phase.

2.1 Pre-Processing

To recognize biomedical terms, we do not need to examine all words in a sentence. We filter out words which have a strong possibility of becoming a non-term. In this study, we find the terms only in noun phrases which include biomedically salient words. When some words are occurred more frequently in the training data than in out-of-domain corpus, we regard the words as



Figure 1: An example of biomedical NE recognition

salient words of the domain. Therefore, salient words can be extracted through corpus comparison. Most terms of the biomedical domain such as NF-kappa scarcely occur in out-of-domain text. Therefore, salient words can provide use-ful information for spotting the probable noun phrases which include terms. The italicized words in Figure 1 correspond to salient words.

In order to find noun phrases including salient words, we first identify noun phrases by using POS tagger (T. Brants, 2000) and NP chunker (L. Ramshaw, 1995), and then perform NP spotting by using salient words. As a result, about 95% of terms can be seen in the spotted NPs.

2.2 Boundary Identification

This step identifies boundaries of biomedical terms in noun phases that include salient words. In order to train a SVM for the identification task, we use the following features.

- word: We use 5 words: the target word, left two words, and right two words.
- **prefix/suffix**: Morphological patterns reflect orthographical characteristics of words as prefix or suffix. In case that a target word infrequently occurs in the training data, a morphological pattern can become an important feature as a normalized form of the word. Also, when a target word is only constituted by lower letters, a morphological pattern such as *-ase* can become a useful information to indicate whether the word is a term or not. Morphological patterns are sorted in the descending order of the amount of relative entropy of each

pattern. When a word is given, we find the most highly ranked morphological pattern that matches one of the word's substrings.

- word construction form: This feature indicates whether a target word contains capital letters, lower letters, numbers, or symbols.
- gene sequence: This binary feature indicates whether a target word represents a gene sequence or not.
- **Greek**: If a target word includes a Greek letter such as *alpha*, *beta*, or *kappa*, we use the Greek letter as a feature.
- dot, hyphen: This feature represents whether a target word has dot or hyphen.
- first word, capital start: This binary feature indicates whether the target word is the first word or not, and whether the target word starts with the capital letter.

Using SVM classifiers, we assign one of the following classes to each word: *B*-*TERM* class representing the beginning of a biomedical term, *I*-*TERM* class representing a part of a biomedical term, or *O* class indicating that the word does not belong to the biomedical term.

2.3 Post-Processing

In order to reduce term recognition errors, we extract collocations from Medline database, and extend a term boundary when the word inside the boundary and the word outside the boundary have a collocation relation. In Figure 1, a bigram human T is identified as a collocation, and human is included in the term. As

shown in Figure 1, since words such as *human*, *transcription*, and *binding* are frequently used as both a term and a non-term, it is difficult to determine whether the words become a term or not. Therefore, if the words occur outside the boundary, it can be useful to identify whether the words have a collocation relation with words inside the boundary.

2.4 Semantic Classification

In this phase, a proper semantic class is assigned to the recognized term by voting the SVM classifiers. We use the following features.

- word: Only 4 words on the rightside of the term are used as features.
- word variation: We also use 4 word variations like word features. To make a word variation, we alter capital letters to lower letters and substitute # for numbers. Also, to segment words into two parts, we regard symbols as space. Among the parts of the target words, we select the longest part as a word variation.

3 Automated Extraction of Lexical and Morphological Knowledge

3.1 Salient Word Extraction

In order to extract salient words from GENIA corpus, we compute each word's probabilities both in out-of-domain corpus and in GENIA corpus respectively. We use Penn II raw corpus as out-of-domain corpus. From the estimated probabilities, we compute the relative frequency ratio of a word w by equation (1). We regard the word as a salient word when its relative frequency ratio value is larger than a threshold. Table 1 shows the examples of salient words.

$$RFR(w) = \frac{P_{GENIA}(w)}{P_{Penn}(w)} \tag{1}$$

3.2 Morphological Pattern Extraction

In order to extract internal morphological patterns of words, we extract all substrings (*prefix* or suffix candidate) when the lengths are larger than two. We compute the relative entropy of each substring in order to discriminate informative prefixes and suffixes. The relative entropy of a substring str can be computed by equation (2) from training data. If words including the substring are more frequently used in terms than in non-terms, the substring can be

word	RFR	word	RFR
kappa	2935.54	synthesis	1034.11
receptors	2122.42	gamma	963.22
inhibited	1747.14	activate	959.05
hiv	1242.60	binding	958.01
expression	1206.58	$\operatorname{mobility}$	875.65
stimulation	1161.28	ra	854.80
receptor	1148.36	extracts	813.11
tcr	1138.35	induced	803.72
promoter	1127.51	alpha	782.25
replication	1104.99	stat	733.88

Table 1: Examples of salient words

prefix	RE	suffix	RE
rii	1.2308	gamma	1.0634
calcineurin	1.2291	alpha	1.0627
il-1	1.2291	-40	1.0594
c-re	1.2288	luciferase	1.0594
c-rel	1.2288	octamer	1.0560
hmg	1.2249	$\operatorname{te-like}$	1.0526
monoblast	1.2249	lipoproteins	1.0526
chemokine	1.2229	-encoded	1.0526
tat	1.2203	rins	1.0526
5-1	1.2187	kappa	1.0503

Table 2: Examples of informative prefixes and suffixes

an informative pattern. The table 2 shows the examples of informative suffixes and prefixes.

$$RE(str) = -P(term|str)log\frac{P(term|str)}{P(term)} -P(\neg term|str)log\frac{P(\neg term|str)}{P(\neg term)}$$
(2)

3.3 Collocation Extraction

To extract reliable collocations, we use a large size of Medline database. We use three key words such as human, blood cell, and transcription factor in order to retrieve similar abstracts analogous to the training data. The total number of extracted abstracts is 14,754. From the retrieved abstracts and the training data, we extract bigram collocations by applying the chisquare formula as equation (3). In the equation, α denotes $C(w_1, w_2)$, β denotes $C(\neg w_1, w_2)$, γ denotes the co-occurrence frequency of two words and N denotes the number of the total bigrams. The table 3 shows the examples of collocations.

$$\chi^2 = rac{N(lpha\delta-eta\gamma)^2}{(lpha+eta)(lpha+\gamma)(eta+\delta)(\gamma+\delta)} \quad (3)$$

collocation	chi-square
nylon wool	101040
academic press	101040
nitric oxide	101040
herpes simplex	88409
high-cholesterol diet	84199
heme oxygenase-1	80830
heat shock	76542
flow cytometry	74093
electrophoretic mobility	72561
navo3 instillation	70164

Table 3: Examples of collocations

4 Experiments

The term recognition phase generally finds the boundary of all terms in a given sentence. However, in a given training data, biomedical NEs which belong to the only 5 specific classes are tagged. We thought that it is proper to identify all terms and to assign appropriate semantic classes to the identified terms. In this study, we tagged all biomedical NEs in the training data by using GENIA 3.02p version. Thus, in the term recognition phase, we find all the biomedical terms represented in sentences. In the classification phase, we extract only the biomedical NEs which belong to 5 specific classes among the identified terms. The total number of biomedical NEs that belong to other classes is 25.703.

For experiments, we utilized the SVM light package. In both phases, we used a polynomial kernel (*degree* 2) with the one-vs-rest classification method.

Table 4 shows the overall performance on the test data. Our system obtains an F-measure of 62.97%. Table 5 shows the performance of each phase on the test data. Our system obtains an F-measure of 51.57% in the term recognition phase. Due to *other* class, our system obtains a recall of 76.66% and a precision of 38.85% in the recognition phase. For the classification task, our system obtains an accuracy of 85.11%.

Table 6 shows the effectiveness of the lexical knowledge. The baseline performance represents the result of the NE recognition without using any additional knowledge. We obtain the performance improvement of 2.82% by using additional knowledge.

5 Conclusion

In this study, we divide the task of biomedical NE recognition into two phases. To improve the

	R	Р	$F_{\beta=1}$
protein	69.71	59.37	64.12
DNA	60.80	51.57	55.80
RNA	66.10	49.68	56.73
cell_line	49.00	52.02	50.46
$cell_type$	65.85	69.58	67.67
overall	66.52	59.77	62.97
correct LEFT	71.60	64.34	67.77
correct RIGHT	77.42	69.56	73.28

Table 4: Experimental results on the test data: R(Recall), P(Precision)

	R	Р	$F_{\beta=1}$	Α
TR	76.66	38.85	51.57	-
\mathbf{SC}	-	-	-	85.11
overall	66.52	59.77	62.97	-

Table 5: Performance of term recognition(TR) and semantic classification(SC): A(Accuracy)

	R	Р	$F_{\beta=1}$
Baseline	63.85	56.85	60.15
Baseline+SW	64.12	56.90	60.29
Baseline+MP	65.57	59.13	62.19
Baseline+Co	64.06	57.73	60.73
Baseline+ALL	66.52	59.77	62.97

Table 6: Efficiency of lexical and morphological knowledge: SW(salient words), MP(morphological patterns), Co(collocations)

performance of term recognition, we propose a method of using additional knowledge. In the pre-processing, we perform NP spotting by using salient words extracted through corpus comparison. We also use morphological patterns as features for learning SVMs. In the postprocessing, we perform term extension using collocations. Experimental results show that introduction of additional knowledge increases the overall performance.

References

- K. J. Lee, Y. S. Hwang and H. C. Rim. 2003. *Two-Phase Biomedical NE Recognition based* on SVMs. In Proc. of ACL 2003 Workshop on Natural Language Processing in Biomedicine.
- T. Brants 2000. *TnT A Statistical Part-of-Speech Tagger*. Proceedings of the Sixth Applied Natural Language Processing.
- L. Ramshaw and M. Marcus 1995. Text Chunk using Transformation-Based Learning. Proceedings of the Third ACL Workshop on Very Large Corpora.