Biomedical Named Entity Recognition: A Poor Knowledge HMM-Based Approach

Natalia Ponomareva, Ferran Pla, Antonio Molina, and Paolo Rosso

Departamento de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia, Spain,
{nponomareva,fpla,amolina,prosso}@dsic.upv.es

Abstract. With a recent quick development of a molecular biology domain it becomes indispensable to promote different resources as databases and ontologies that represent the formal knowledge of the domain. As these resources have to be permanently updated, due to a constant appearance of new data, the Information Extraction (IE) methods become very useful. Named Entity Recognition (NER), that is considered to be the easiest task of IE, still remains very challenging in molecular biology domain because of the special phenomena of biomedical entities. In this paper we present our Hidden Markov Model (HMM)-based biomedical NER system that takes into account only parts-of-speech as an additional feature, which are used both to tackle the problem of non-uniform distribution among biomedical entity classes and to provide the system with an additional information about entity boundaries. Our system, in spite of its poor knowledge, has proved to obtain better results than some of the state-of-the-art systems that employ a greater number of features.

1 Introduction

Recently the molecular biology domain has been getting a massive growth due to many discoveries that have been made during the last years and due to a great interest to know more about the origin, structure and functions of living systems. It causes to appear every year a great deal of articles where scientific groups describe their experiments and report about their achievements.

Nowadays the largest biomedical database resource is MEDLINE that contains more than 14 millions of articles of the world's biomedical journal literature. To deal with such an enormous quantity of biomedical texts different biomedical resources as databases, ontologies, search engines adapted to this domain have been created.

In fact, NER is the first step to order and structure all the existing domain information. In molecular biology it is used to identify within the text which words or phrases refer to biomedical entities, and then to classify them into relevant biology concept classes.

Although NER in biomedical domain has received attention by many researchers, the task remains very challenging and the results achieved in this

Z. Kedad et al.(Eds.): NLDB 2007, LNCS 4592, pp. 382–387, 2007.

area are much poorer than in the newswire one. Its difficulty is caused principally by the complex structure of molecular names and the lack of naming convention [1].

In this paper, we present our HMM-based biomedical Named Entity (NE) recognizer which uses only POS tags as an additional feature. We will show that POS information is very useful in biomedical NER task and that only applying this rather poor knowledge we may archieve good results and, moreover, surpass the performance of the systems exploiting a large set of features.

The paper is organized as follows. Section 2 is dedicated to illustrate some important characteristics of the Genia corpus that have been used during the construction of our model and the experiments. In Section 3, our biomedical NE recognizer is described and its comparison with the best state-of-the-art systems is made. Finally, Section 4 draws our conclusions and discusses the future work.

2 The Genia Corpus

Any supervised machine-based model depends on a corpus that has been used to train it. At the moment the largest and, therefore, the most popular biomedical annotated corpus is Genia corpus v. 3.02 which contains 2,000 abstracts from the MEDLINE collection annotated with 36 biomedical entity classes. In our experiments, we have used its JNLPBA version [2].

The JNLPBA corpus is annotated with 5 classes of biomedical entities: protein, RNA, DNA, cell type and cell line. Biomedical entities are tagged using the IOB2 notation. In Table 1 a tag distribution within the training and test corpora is shown. It can be seen that the majority of words (about 80%) does not belong to any biomedical category. Furthermore, the biomedical entities themselves also have an irregular distribution: the most frequent class (protein) contains about 10% of words approximately, whereas the most rare one (RNA) - less than 0.5%. The tag irregularity may cause a confusion among different types of entities with a tendency for any word to be referred to the most numerous class.

Table 1. Entity tag distribution in the training and test corpora

Corpus	Protein, %	DNA, %	RNA, %	cell type, %	cell line, %	no-entity, %
Training	11.2	5.1	0.5	3.1	2.3	77.8
Test	9.7	2.8	0.3	4.9	1.5	80.8

3 Preliminary Results: The HMM Approach

The HMM approach has been proved to be successfully employed in many NLP tasks, such as speech recognition, machine translation, POS tagging, NER, etc. In this section, our HMM-based NER system will be introduced together with the description of its main characteristics. Then, we will present experimental results and their comparison with some other state-of-the-art NER systems based on the same approach.

3.1 HMM-Based Biomedical NE Recognizer Description

Let $\mathbf{w} = (w_1 w_2 ... w_n)$ be a sequence of observed words of length n. Let $\mathbf{t} = (t_1 t_2 ... t_n)$ be a sequence of biomedical entity tags assigned to words from the word sequence $\mathbf{w}$. We denote as $\mathbf{T}$ a collection of various sequences of biomedical entity tags of length n. The solution of the NE recognition task using a second order HMM approach can be presented as follows:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t} \in \mathbf{T}} [\log P(t_1) + \log P(t_2|t_1) + \sum_{i=3}^{n} \log P(t_i|t_{i-1}t_{i-2}) + \sum_{i=1}^{n} \log P(w_i|t_i)]$$

It is common to incorporate into NER systems different features which are considered to be useful for the correct classification. Our system exploits only POS feature supplied by the Genia Tagger[1]. It is significant that this tagger was trained on the Genia corpus in order to provide better results in the biomedical texts annotation. As it has been shown by [3], the use of the POS tagger adapted to the biomedical task may greatly improve the performance of the NER system than the use of the tagger trained on any general corpus as, for instance, Penn TreeBank.

In our system, the POS information serves both to provide an additional knowledge about entity boundaries and to diminish an entity class irregularity. As we have seen in Section 2, the majority of words in the corpus does not belong to any entity class. Such data irregularity can provoke errors, which are known as false negatives, and, therefore, may diminish the recall of the model. Besides, there also exists a non-uniform distribution among biomedical entity classes: e.g. class "protein" is more than 20 times larger than class "RNA" (see Table 1).

To solve the above problem we have decided to split the most numerous categories by means of POS tags of words. The idea of splitting or specializing tags was previously successfully applied to other NLP tasks, such as POS tagging, chunking or clause detection [4]. For the biomedical NER task, a similar idea was proposed by [5] who employed it for the SVM approach.

We have constructed three models using different sets of POS tags:

(1) only the non-entity class has been splitted;
(2) the non-entity class and two most numerous entity categories (protein and DNA) have been splitted;
(3) all the entity classes have been splitted.

It may be observed that each following model includes the set of entity tags of the previous one. Thus, the last model has the greatest number of states.

Besides, we have carried out various experimens with a different number of boundary tags, and we have concluded that only adding two tags (E - end of an entity and S - a single word entity) to a standard IOB2 set can notably improve the performance of the system.

Consequently, each entity tag of our models contains the following components:

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

(i) entity class (protein, DNA, RNA, etc.);
(ii) entity boundary (B - beginning of an entity, I - inside of an entity, E - end of an entity, S - a single word entity);
(iii) POS information.

The key point, that should be paid attention to, is the POS set used in the splitting procedure. Kazama et al. [5] applied all the POS tags of the Penn TreeBank tag set. We think that the whole set of POS is rather redundant and contributes neither to the system accuracy, no to its stability.

In order to split a non-entity class, the distribution of its in-class POS tags has been analyzed (Table 2). We have realized several experiments to choose the best set of POS tags. As a result, the POS with a relative frequency of more than 1% have been selected to participate in the entity tag balancing.

Table 2. POS distribution inside of a no-entity category

POS	NN	IN	DT	JJ	<.>	NNS	<,>	CC	VBN	RB	VBD	VBZ	TO	VBP	CD	)	(	VB
%	19.6	16.3	9.6	8.6	4.8	4.7	4.6	4.2	3.8	3.1	2.7	2.4	2.0	1.8	1.7	1.6	1.5	1.4

The classes of biomedical entities have been divided according to the POS distribution within the class "Protein". In order to participate in the splitting procedure, the most frequent POS tags have been chosen (Table 3). As it may be noticed from Table 3, some parts-of-speech can appear only in certain parts of a biomedical entity (e.g. coma, brackets or conjunction never stay at the beginning of an entity).

Table 3. List of POS tags participated in the biomedical entity category splitting

POS	POS position in the entity
NN, JJ, NNS	Everywhere
(, CC, <,>	Inside
CD,)	Inside or at the end

3.2 Experiments

The first experiments we have carried out were devoted to compare our three HMM-based models in order to analyze what entity class splitting provides the best performance. In Table 4, our baseline (i.e., the model without class balancing procedure) is compared with our three models. The results seem to be promissing taking into account the poor additional information we have employed. Although all our models have improved the baseline, there is a significant difference between the first model and the other two models, which have shown rather similar results.

Our system has been compared to those that are based on the same approach and used the same training and test corpora (Table 5). The system developed by Zhao et al. [6] deserves special attention because it exploits nothing else but

Table 4. Analysis of the influence of different sets of POS to the system performance

Model	Tags number	Recall, %	Precision, %	F-score
Baseline	21	63.7	60.2	61.9
Model (1)	40	68.4	61.4	64.7
Model (2)	95	69.1	62.5	65.6
Model (3)	135	69.4	62.4	65.7

a huge unlabel corpus extracted from the MEDLINE collection. The other system developed by Zhou et al. [3] achieved the best perfomance in the JNLPBA task. It exploites a large set of features and some deep knowledge resources and techniques, e.g. post-processing operations which serve to correct entity boundaries. Zhou et al. have analyzed a contribution of features, rules and external resources into the system performance and thanks to this information we can compare results of our best model with their system before applying the deep knowledge techniques.

Table 5. Comparison of biomedical NER systems based on the HMM approach

System	F-score
Zhou complete	72.6
Zhou w/o deep knowledge	64.1
Zhao	64.8
Our best model	65.7

Analyzing the results shown in Table 5, it can be appreciated that our system, which only uses in-domain POS information has obtained better results than the Zhou (w/o deep knowledge) and Zhao systems which employed many features or external resources.

We would like to remark the role of post-processing operations for the improvement of NER systems performance [7,3]. Actually, as it can be seen in Table 5, Zhou has increased the F-score on 8.5 after using deep knowledge techniques. Patrick et al. [7], who employed ME approach, also have obtained a great improvement of the F-score from 61.1 to 68.2 after applying a set of post-processing rules. All the above shows the importance of post-processing procedures for the biomedical NER task.

4 Conclusions and Future Work

In this paper, we have presented our biomedical NE recognizer. In order to tackle the problem of non-uniform distribution among biomedical entity classes, the possibility of splitting the most numerous categories by means of POS tags has been investigated. We have explored different sets of POS to realize a splitting procedure. As a result, a splitting of only the non-entity class has improved the performance of our system on about 3 points of F-score. The best result was

obtained by the model, when all the entity classes were splitted (about 4 points of improvement). Despite the poor knowledge which has been used, we were able to obtain better performance than some of the state-of-the-art systems that exploited much more additional information.

As future work we plan to develop a rule-based post-processing module for our NER system. Futhermore, we will investigate different sets of features in order to find the optimal one. In fact, as it was already shown by some researchers, a rich set of features does not always help to achieve good results and could even worsen the system performance [8,9].

Acknowledgements

This work has been partially supported by MCyT TIN2006-15265-C06-04 research project.

References

1. Zhang, J., Shen, D., Zhou, G., Jian, S., Tan, C.L.: Enhancing hmm-based biomedical named entity recognition by studying special phenomena. Journal of Biomedical Informatics 37(6) (2004)
2. Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y.: Introduction to the bio-entity recognition task at jnlpba. In: Proceedings of the Int. Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004), pp. 70–75 (2004)
3. Zhou, G., Su, J.: Exploring deep knowledge resources in biomedical name recognition. In: Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004), pp. 96–99 (2004)
4. Molina, A., Pla, F.: Shallow parsing using specialized hmms. JMLR Special Issue on Machine Learning approaches to Shallow Pasing (2002)
5. Kazama, J., Makino, T., Ohta, Y., Tsujii, J.: Tuning support vector machines for biomedical named entity recognition. In: Proceedings of the Workshop on NLP in the Biomedical Domain (at ACL 2002), pp. 1–8 (2002)
6. Zhao, S.: Name entity recognition in biomedical text using a hmm model. In: Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004) (2004)
7. Patrick, J., Wang, Y.: Biomedical named entity recognition system. In: Proceedings of the Tenth Australasian Document Computing Symposium (ADCS 2005) (2005)
8. Settles, B.: Biomedical named entity recognition using conditional random fields and novel feature sets. In: Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004), pp. 104–107 (2004)
9. Collier, N., Takeuchi, K.: Comparison of character-level and part of speech features for name recognition in bio-medical texts. Journal of Biomedical Informatics 37(6), 423–425 (2004)