POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004

Yu Song, Eunju Kim, Gary Geunbae Lee, Byoung-kee Yi
Department of CSE,
Pohang University of Science and Technology (POSTECH)
Pohang, Korea 790-784
{songyu, hosuabi, gblee, bkyi} @postech.ac.kr

Abstract

Two classifiers -- Support Vector Machine (SVM) and Conditional Random Fields (CRFs) are applied here for the recognition of biomedical named entities. According to their different characteristics, the results of two classifiers are merged to achieve better performance. We propose an automatic corpus expansion method for SVM and CRF to overcome the shortage of the annotated training data. In addition, we incorporate a keyword-based post-processing step to deal with the remaining problems such as assigning an appropriate named entity tag to the word/phrase containing parentheses.

1 Introduction

Recently, with the rapid growth in the number of published papers in biomedical domain, many NLP (Natural Language Processing) researchers have been interested in a task of automatic extraction of facts from biomedical articles. The first and fundamental step is to extract the named entities. And recently several SVM-based named entity recognition models have been proposed. Lee et. al. ([Lee et. al., 2003]) proposed a two-phrase SVM recognition model. Yamamoto et. al. ([Yamamoto et. al., 2003]) proposed a SVM-based recognition method which uses various morphological information and input features such as base noun phrase information, stemmed forms of a word, etc. However, notable limitation of SVM is its low speed both for training and recognition.

On the other hand, conditional random fields (CRFs) ([Lafferty, 2001]) is a probabilistic framework for labelling and segmenting sequential data, which is much faster comparing with SVM. The conditional probability of the label sequence can depend on arbitrary, non-independent features of the observation sequence without forcing the model to account for the distribution of those dependencies. Named entity recognition problem can be taken as assigning the named entity class tag sequences to the input sentences. We adopt CRF to be the complementary scheme of SVM.

In natural language processing, supervised machine-learning based approach is a kind of standard and its efficiency is proven in various task fields. However, the most problematic point of supervised learning methods is that the size of training data is essential to achieve good performance, but building a training corpus by human labeling is time consuming, labor intensive, and expensive. To overcome this problem, various attempts have been proposed to acquire a training data set in an easy and fast way. Some approaches focus on minimally-supervised style learning and some approaches try to expand or acquire the training data automatically or semi-automatically. Using virtual examples, i.e., artificially created examples, is a type of method to expand the training data in an automatic way ([Niyogi et al, 1998] [Sasano, 2003] [Scholkopf et. al., 1996]. In this paper, we propose an automatic corpus expansion method both for SVM and CRF based biological named entity recognition using virtual example idea.

The remainder of this paper is organized as follows: Section 2 introduces named entity recognition (NER) part: two machine learning approaches with some justification, feature set used in NER and virtual examples generation. In section 3, we present some keyword-based post-processing methods. The experiment results and analysis will be presented in section 4. Finally, conclusion is provided in section 5.

2 Named Entity Recognition

The training corpus is provided in IOB notion. The IOB notation is used where named entities are not nested and therefore do not overlap. Words outside of named entities are tagged with “O”, while the first word in a named entity is tagged with B-[entity class], and further named entity words receive tag I-[entity class] for inside. We define the named entity recognition problem as a classification problem, assigning an appropriate classification tag for each token in the input sentences.

To simplify the classification problem, we assign each token only with I-[entity class]/O. Then we convert the tag of the initial token of a consecutive

sequence of predicted named entity tokens to B-[entity class].

2.1 SVM

Support Vector Machine (SVM) is a well-known machine learning technique showing a good performance in several classification problems. However, SVM has suffered from low speed and unbalanced distributed data.

Named entity token is a compound token that consists of the constituents of some other named entities, and all other un-related tokens are considered as outside tokens. Due to the characteristics of SVM, this unbalanced distribution of training data can cause a drop-off in classification coverage.

In order to resolve this low coverage and low speed problem together, we filter out possible outside tokens in the training data through two steps. First, we eliminate tokens that are not constituents of a base noun phrase, assuming that every named entity token should be inside of a base noun phrase boundary. Second, we exclude some tokens according to their part-of-speech tags. We build a stop-part-of-speech tag list by collecting tags which have a small chance of being a named entity token, such as predeterminer, determiner, etc.

2.2 CRF

Conditional random fields (CRFs) ([Wallach, 2004] is a probabilistic framework for labelling and segmenting a sequential data. Let $G(V,E)$ be a graph such that $Y=(Yv)v\in V$, and there is a node $v\in V$ corresponding to each of the random variable representing an element Yv of Y. Then (X,Y) is **a conditional random field**, and when conditioned on X, the random variables Yv obey the Markov property with respect to the graph: $p(Yv\,||\,X,Yw,w\neq v)=p(Yv\,|\,X,Yw,w\sim v)$, where $w\sim v$ means that w and v are neighbours in G.

Let X and Y be jointly distributed random variables respectively representing observation sequences and corresponding label sequences. A CRF is an undirected graphical model, globally conditioned on X (the observation sequence).

We try to use this CRF model to our NER as a complementary method for both speed and coverage. SVM predicts the named entities based on feature information of words collected in a predefined window size while CRF predicts them based on the information of the whole sentence. So, CRF can handle the named entities with outside tokens which SVM always tags as “O”.

2.3 Feature set

As an input to the classifier, we use a bit-vector representation, each dimension of which indicates whether the input matches with the corresponding feature.

The followings are the basic input features:

- Surface word - only in the case that the previous/current/next words are in the surface word dictionary.
- word feature - orthographical feature of the previous/current/next words.
- prefix/suffix - prefixes/suffixes which are contained in the current word among the entries in the prefix/suffix dictionary.
- part-of-speech tag - POS tag of the previous/current/next words.
- Base noun phrase tag - base noun tag of the previous/current/next words.
- previous named entity tag - named entity tag which is assigned for previous word. This feature is only for SVM.

The surface word dictionary is constructed from the words that occur more than one time in the training part of the corpus.

2.4 Automatic Corpus Expansion using Virtual Examples

To achieve good results in machine learning based classification, it is important to use training data which is sufficient not only in the quality but also in the quantity. But making the training data by hand requires considerable man-power and takes a long time. Expanding the training data using virtual examples is an attempt for corpus expansion in the biomedical domain.

We expand the training data by augmenting the set of virtual examples generated using some prior knowledge on the training data. We use the fact that the syntactic role of a named entity is a noun and the basic syntactic structure of a sentence is preserved if we replace a noun with another noun in the sentence. Based on this linguistic paradigmatic relation, we can generate a new sentence by replacing each named entity by another named entity which is in the named entity dictionary of the corresponding class. Then we augment the sentence into the original training data. If we apply this replacement processes n times for each sentence in the original corpus, then we can obtain a virtual corpus about n+1 times bigger than the original one. Since the virtual corpus strengthens the right information which may not be observed in the original corpus, it is helpful to extend the coverage of a recognition model and

also helpful to improve the recognition performance.

3 Keyword based post-processing

We notice that some words occur more frequently in the specific entity class. For example, the word "genes" appears in class DNA 590 times while in other entity class appears less than 10 times. The information provided by these key words not only impacts the named entity prediction part but also shows great power in post-processing part. Once keywords appear at specific position in a named entity, we can surely decide the entity class of this named entity.

3.1 Words containing parentheses or "and"

It is difficult but significant to decide whether parentheses or "and" are part of named entity or not. Parentheses occur in the named entity more than 700 times in the training data. Both SVM and CRF cannot work well while dealing with this problem.

Once a specific keyword appears at the right side of ")", we can tell that the parentheses belong to a named entity. The named entity tag information can also be determined by the keyword. For example, in Table 1, the left column is the result of the NER module. At post-processing stage, the word "genes" is detected on the right side of ")", then this pair of parentheses and keyword "genes" are included in the current named entity.

Before		After	
text	tag	text	tag
(	O	(	I-DNA
VH	I-DNA	VH	I-DNA
)	O	)	I-DNA
genes	O	genes	I-DNA

Table 1: An example for the usage of keywords.

A keyword list for parentheses is collected from the training corpus, including the named entity tag information. It not only solves the parentheses named entity tag problem but also helps to correct the wrong named entity assigned to the words between parentheses by the previous step. The word "and" can be treated similarly as the parenthesis case.

3.2 Correcting the wrong named entity tag

Some keywords occur in one specific type of named entities with high frequency. We employ the information provided by those keywords in correcting the wrongly assigned named entity tag.

First a list of high frequency keywords with class information is collected. Once a keyword is predicted as another type of named entity, all the words in the current named entity boundary will be corrected as the corresponding named entity type as the keyword. For example, the keywords "protein" and "proteins", in a very rare case, belong to other named entity class rather than the class "PROTEIN".

4 Experiment Result and analysis

4.1 Corpus

The shared task BioNLP/NLPBA 2004 provides 2000 MEDLINE abstracts from the GENIA ([Ohta et. al., 2002]) bio-named entity corpus version 3.02. There are total 5 entity classes: DNA, RNA, protein, cell_line and cell_type.

4.2 Experiment results and analysis

CLASS	Recall/Precision/F-score			
ALL	Full	R64.80	P67.82	F66.28
	Left	R69.99	P73.25	F71.58
	Right	R73.25	P76.67	F74.92
Protein	Full	R65.50	P73.04	F69.07
	Left	R71.26	P79.46	F75.13
	Right	R72.23	P80.54	F76.16
Cell_Line	Full	R53.77	P61.40	F57.33
	Left	R56.39	P64.40	F60.13
	Right	R63.57	P72.60	F67.79
DNA	Full	R58.60	P61.65	F60.08
	Left	R64.27	P67.61	F65.90
	Right	R66.79	P70.27	F68.48
RNA	Full	R65.49	P62.71	F64.07
	Left	R67.26	P64.41	F65.80
	Right	R75.22	P72.03	F73.59
Cell_Type	Full	R70.45	P59.45	F64.48
	Left	R74.46	P62.83	F68.15
	Right	R84.52	P71.32	F77.36

Table 2: Final result of POSBIOTM-NER (with no abstract boundary information).

Method	Full: Recall/Precision/F-score
SVM.base	R62.01 P65.80 F63.85
SVM+V	R63.91 P66.89 F65.37
CRF.base	R64.90 P61.33 F63.06
CRF+V	R65.78 P61.06 F63.34
Final	R64.80 P67.82 F66.28

Table 3: Step by step result

From table 3, we can see that after using virtual samples, both the precision and recall increased, especially for SVM. In CRF, even though the full f-score did not increase the full F-score much, but for RNA class, after using virtual samples, the f-score has increased 3%.

A CRF has different characteristics from SVM, and is good at handling different kinds of data. So, we simply merge the results of two machine learning approaches, by using the CRF results to extend the boundaries of named entities predicted by SVM. After merging the results of the baseline of SVM and CRF (without using virtual samples) the f-score reaches to 64.58, while the f-score of SVM alone is 63.85. The final score in Table 3 is the merged results with the virtual samples.

Although we have improved our system by using virtual samples, CRF and SVM as complementary means and post-processing, we still have some problems to solve, such as correct named entity boundary detection. It is more difficult to correctly predict the left boundary of named entities than the right boundary. From the analysis of the results, we usually predict "human" and "factor" as the beginning and end of a named entity, but, it is even difficult for human to decide correctly whether it is a part of a named entity or not.

5 Conclusion and Future Works

In this paper, we propose a general method for named entity recognition in the biomedical domain. Various morphological, part-of-speech and base noun phrase features are incorporated to recognise the named entities. We use two different kinds of machine learning techniques, SVM and CRF, with merged results. We also developed a virtual sample technique to overcome the training data shortage problem. Finally, we present a keyword-based heuristic post-processing which increases both precision and recall.

As shown in the experiment results, more correct detection of the named entity boundary is required, especially the detection of left boundary.

6 Acknowledgements

This research is supported by BIT Fusion project (by MOST).

References

J.Lafferty, A.McCallum, and F.Pereira. *Conditional random fields: probabilistic models for segmenting and labelling sequence data*. In International Conference on Machine Learning, 2001.

Ki-Joong Lee, Young-Sook Hwang, and Hae-Chang Rim. Two-phase biomedical NE recognition based on SVMs. *Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine,2003.*

P.Niyogi, F.Girosi, and T.Poggio. Incorporating prior information in machine learning by creating virtual examples. *Proceedings of IEEE volume 86, pages 2196-2207, 1998*

Manabu Sasano. Virtual examples for text classification with support vector machines. *Proceedings of 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003), 2003.*

Bernhard Scholkopf, Chris Burges, and Vladimir Vapnik. Incorporating invariances in support vector learning machines. *Artificial Neural Networks- ICANN96,1112:47-52,1996.*

Hanna M.Wallach. Conditional Random Fields: An Introduction. 2004

T.Ohta, Y. Tateisi, J.Kim, H. Mima and J.Tsujiii.2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. *In Proceedings of Human Language Technology Conference, 2002.*

Kaoru Yamamoto, Taku Kudo, Akihiko Konagaya, and Yuji Matusmoto. Protein name tagging for biomedical annotation in text. *Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine, 2003.*