POSBIOTM-NER: A Machine Learning Approach for Bio-Named Entity Recognition

Yu Song, Eunji Yi, Eunju Kim, Gary Geunbae Lee,
Department of CSE, POSTECH, Pohang, Korea 790-784
Soo-Jun Park
Bioinformatics Research Team, Computer and Software Research Lab., ETRI, Taejon, Korea 305-350

Abstract

Two main difficulties in SVM (Support-Vector Machine) and other machine-learning based biological named entity recognition are the existence of many different spelling variants and a lack of annotated corpus for training. Attempts are made to resolve these two difficulties respectively, which turn out to be rewarding. We automatically expand the annotated corpus in a fast, efficient, and easy way to achieve better results. In addition, we propose the use of edit-distance as a significant contributing feature for SVM.

1 Introduction

Recently, with the rapid growth in the number of published papers in the biomedical domain, many NLP (Natural Language Processing) researchers have been interested in the task of automatic extraction of facts from biomedical articles. The first step is to extract named entity. There are several SVM-based named entity recognition models. Lee et al. [1] proposed a two-phrase recognition model. Yamamoto et al. [7] proposed a SVM-based recognition method which uses various morphological information and input features such as base noun phrase information, the stemmed form of word, etc.

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 [2] [3] [4]. In this paper, we propose an automatic corpus expansion method for SVM-based biological named entity recognition using virtual example idea.

One of the main difficulties in biological named entity recognition (NER) is that many variant forms exist for each named entity in biomedical articles. It is difficult to recognize them even if we meet a named entity already defined in the named entity dictionary. Edit distance, a useful metric to measure the similarity between two strings, has been used to help with such problems in fields such as computational biology, error correction, and pattern matching in large databases. A named entity recognition method using an edit-distance is suggested by Tsuruoka and Tsujii[5]. We suggest using the edit distance as a contributing feature for SVM in this paper.

2 Method

2.1 Overview

Figure 1 depicts an overview of our POSBIOTM-NER system based on the method to be described in this section. Preprocess part contains BIO (Begin, Inside, Outside) tagging, POS (part-of-speech) tagging and NP (Noun Phrase) chunking.

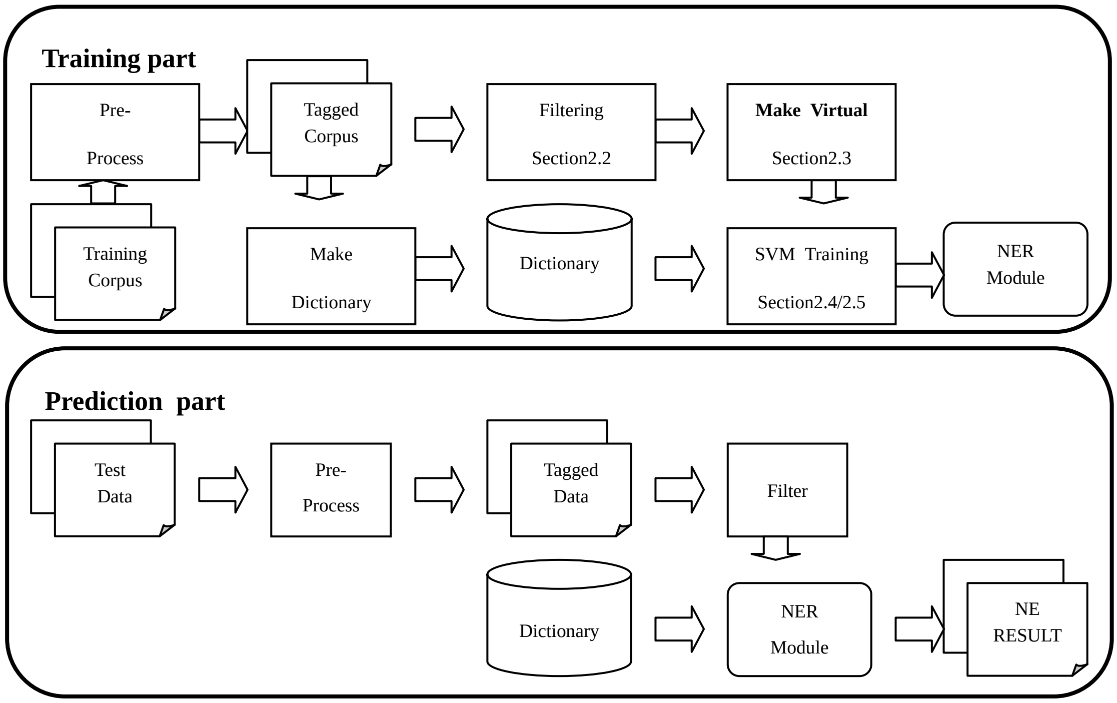

FIG-1 System Architecture

We make use of BIO representation, in which B-G, I-G, O stand for begin, inside and outside of a gene, respectively. We define the named entity boundary identification problem as the classification problem to assign a proper B-G/I-G/O tag to each token. Possible outside tokens are filtered out using POS tagging and NP chunking information. Dictionaries such as surface word dictionary are generated based on the Training Corpus. Original training corpus will be automatically expanded in order to get better results. **Edit-Distance** feature and other basic features are used for SVM training and prediction.

2.2 Filtering of Outside Tokens

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, unbalanced distribution of training data can cause a drop-off of classification coverage.

In order to resolve this problem, 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.

2.3Automatic 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 the quality but also 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 a new attempt of 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 paradigmatic relation, we can generate a new sentence by replacing each named entity in the given sentence by another named entity which is in the named entity dictionary of the corresponding class and then augment the sentence into the original training data. If we apply this replacement process 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 contains more context information which is not in the original corpus, it is helpful to extend the coverage of a recognition model and also helpful to improve the recognition performance.

2.4 Basic Features of SVM-based Recognition

As an input to the SVM, we use 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

The surface word dictionary is constructed from the words that occur more than one time in the training part of the corpus. The prefix/suffix dictionary is constructed by collecting overlapped character sequences longer than two characters that occur more than two times in the named entity token collection.

2.5 Use of Edit-Distance

Edit distance is a useful metric measure for the similarity of two strings [6]. Consider two strings X and Y over a finite alphabet, whose length is m and n respectively with m >= n. The edit distance between X and Y is defined as the weighted sum of all the sequences of edit operations (insertions, deletions, and substitutions of characters) that transform X into Y.

To incorporate the effect of the edit-distance with SVM-based recognition method, we adopt edit-distance features as additive input features. Each token has N edit-distance features where N is the number of named entity category. To calculate the value of each edit-distance features, we define candidate tokens CTi for each i = 1, …, N and initialize them as an empty string, and then do the following for each token in the input sentence.

- For each i=1,…,N, do the following:

1. (a) If the previous token was begin or inside of the named entity category NEi, then concatenate the current word to CTi.
   (b) Otherwise, copy the current word to CTi.
2. Calculate the minimum value among the edit-distances from CTi, to each entry of the named entity dictionary for category NEi, and store it as MEDi.
3. Calculate medScorei by MEDi/length(CTi).
4. Set the ith edit-distance feature value for the current token the same as medScorei

In the calculation of the edit-distance, we use a cost function suggested by Tsurouka and Tsujii [5]. The cost function includes some consideration for different lexical variations such as hyphen, lower-case and upper-case letter, which are all appropriate to biomedical named entity recognition.

3 BioCreative Results

There are total 196,620 tokens in the BioCreative training data. After filtering step, 98,686 tokens remain, but only 0.0523% of the actual named entity tokens are filtered out. Since many outside tokens are filtered out during the filtering step, a large portion of the training and recognition time is reduced.

One problem remains that some tokens which appear many times in bio-named entity are filtered out as they are not considered to be included in a noun phrase, such as "(" , "." and ",". After simple modification of the filtering step, this problem will be resolved.

We generated four virtual example sentences for each sentence in the original training corpus. To show the usefulness of virtual examples, we trained 3 different models. First model only bases on the training corpus. Second one bases on the training corpus augmented by the virtual corpus. Finally, third one bases on the training corpus augmented by the unique virtual corpus, in which the reduplicated virtual samples are removed. Table 1 shows the final results from these 3 models. The two models using the virtual samples outperform the original model in F-measure, but the precision decreases due to the virtual samples.

	Precision	Recall	Balanced-F-measure
Training Only	0.800	0.685	0.613
Training + Virtual	0.637	0.697	0.666
Training + Virtual.Unique	0.632	0.705	0.667

Table-1 Final Results

4 Conclusion

We propose a method for named entity recognition in the biomedical domain that adopts an edit-distance measure to resolve the spelling variant problem. Our model uses the edit-distance metric as additive input features of SVM, which is a well-known machine learning technique showing a good performance in several classification problems. Moreover, to expand the training corpus which is always scarce, in an automatic and effective way, we propose an expansion method using virtual examples. This is a rewarding attempt and helpful to improve the recognition performance, especially in recall.

The current form of POSBIOTM-NER is to recognize not only the gene name, but also other significant biological entities at the same time, like protein, small molecule and the cellular process. At present we are searching for other features for a future performance improvement. Also we are working on a compensation method which can minimize the precision drops in the corpus expansion method using virtual examples.

5 Acknowledgements

This research is supported by 21C Frontier project on Human-Robot Interface (by MOST).

6 References

[1] 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.*

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

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

[4] Bernhard Scholkopf, Chris Burges, and Vladimir Vapnik. Incorporating invariances in support vector learning machines. *Artifical Neural Networds- ICANN96,1112:47-52,1996.*

[5] Yoshimasa Tsuruoka and Jun'ichi Tsujii. Boosting precision and recall of dictionary-based protein name recognition. *Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine,2003.*

[6] Robert A. Wagner and Michael J.Fisher. The string-to-string correction problem. *Journal of the Association for Computing Machinery, 21(1), 1974.*

[7] 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.*