Recognizing Nested Named Entities in GENIA corpus

Baohua Gu
School of Computing Science
Simon Fraser University, Burnaby, BC, Canada
bgu@cs.sfu.ca

Abstract

Nested Named Entities (nested NEs), one containing another, are commonly seen in biomedical text, e.g., accounting for 16.7% of all named entities in GENIA corpus. While many works have been done in recognizing non-nested NEs, nested NEs have been largely neglected. In this work, we treat the task as a binary classification problem and solve it using Support Vector Machines. For each token in nested NEs, we use two schemes to set its class label: labeling as the outmost entity or the inner entity. Our preliminary results show that while the *outmost labeling* tends to work better in recognizing the outmost entities, the *inner labeling* recognizes the inner NEs better. This result should be useful for recognition of nested NEs.

1 Introduction

Named Entity Recognition (NER) is a key task in biomedical text mining, as biomedical named entities usually represent biomedical concepts of research interest (e.g., protein/gene/virus, etc).

Nested NEs (also called embedded NEs, or cascade NEs) exhibit an interesting phenomenon in biomedical literature. For example, “human immuneodeficiency virus type 2 enhancer” is a DNA domain, while “human immunodeficiency virus type 2” represents a virus. For simplicity, we call the former the *outmost* entity (if it is not inside another entity), while the later the *inner* entity (it may have another one inside).

Nested NEs account for 16.7% of all entities in GENIA corpus (Kim, 2003). Moreover, they often represent important relations between entities (Nedadic, 2004), as in the above example. However, there are few results on recognizing them. Many studies only consider the outmost entities, as in BioNLP/NLPBA 2004 Shared Task (Kim, 2004).

In this work, we use a machine learning method to recognize nested NEs in GENIA corpus. We view the task as a classification problem for each token in a given sentence, and train a SVM model. We note that nested NEs make it hard to be considered as a multi-class problem, because a token in nested entities has more than one class label. We therefore treat it as a binary-class problem, using one-vs-rest scheme.

1.1 Related Work

Overall, our work is an application of machine learning methods to biomedical NER. While most of earlier approaches rely on handcrafted rules or dictionaries, many recent works adopt machine learning approaches, e.g, SVM (Lee, 2003), HMM (Zhou, 2004), Maximum Entropy (Lin, 2004) and CRF (Settles,2004), especially with the availability of annotated corpora such as GENIA, achieving state-of-the-art performance. We know only one work (Zhou,2004) that deals with nested NEs to improve the overall NER performance. However, their approach is basically rule-based and they did not report how well the nested NEs are recognized.

2 Methodology

We use SVM-light (http://svmlight.joachims.org/) to train a binary classifier on the GENIA corpus.

2.1 Data Set

The GENIA corpus (version 3.02) contains 97876 named entities (35947 distinct) of 36 types, and 490941 tokens (19883 distinct). There are 16672

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 112–113,
New York City, June 2006. ©2006 Association for Computational Linguistics

nested entities, containing others or nested in others (the maximum embedded levels is four). Among all the outmost entities, 2342 are protein and 1849 are DNA, while there are 9298 proteins and 1452 DNAs embedded in other entities.

2.2 Features and Class Label

For each token, we generate four types of features, reflecting its characteristics on orthography, part-of-speech, morphology, and special nouns. We also use a window of (-2, +2) as its context.

For each token, we use two schemes to set the class label: *outmost labeling* and *inner labeling*. In the outmost labeling, a token is labeled +1 if the *outmost* entity containing it is the target entity, while in the inner labeling, a token is labeled +1 if *any* entity containing it is the target entity. Otherwise, the token is labeled -1.

3 Experiment And Discussion

We report our preliminary experimental results on recognizing *protein* and *DNA* nested entities. For each target entity type (e.g., protein) and each labeling scheme, we obtain a data set containing 490941 instances. We run 5-fold cross-validation, and measure performance (P/R/F) of exact match, left/right boundary match w.r.t. outmost and inner entities respectively. The results are shown in Table 1 and Table 2.

		Outmost labeling (P/R/F)	**Inner labeling (P/R/F)**
Outmost Entities Recognized	Exact	0.772 /0.014 /0.028	0.705 /0.017 /0.033
	Left	0.363 /0.373 /0.368	0.173 /0.484 /0.254
	Right	0.677 /0.199 /0.308	0.674 /0.208 /0.318
	Overall	**0.60/0.20/0.23**	**0.52/0.24/0.20**
Inner Entities Recognized	Exact	0.692 /0.229 /0.344	0.789 /0.679 /0.730
	Left	0.682 /0.289 /0.406	0.732 /0.702 /0.717
	Right	0.671 /0.255 /0.370	0.769 /0.719 /0.743
	Overall	**0.68/0.26/0.37**	**0.76/0.70/0.73**

Table 1 Performance of nested protein entities

From the tables, we can see that while the outmost labeling works (slightly) better for the outmost entities, the inner labeling works better for the inner entities. This result seems reasonable in that each labeling scheme tends to introduces more entities of its type in the training set.

It is interesting to see that the inner labeling works much better in identifying inner proteins than in inner DNAs. The reason could be due to the fact that there are about three times more inner proteins than the outmost ones, while the numbers of inner DNAs and outmost DNAs are roughly the same (see Section 2.1).

Another observation is that the inner labeling gains significantly (over the outmost labeling) in the inner entities, comparing to its loss in the outmost entities. We are not sure whether this is the general trend for other types of entities, and if so, what causes it. We will address this issue in our following work.

		Outmost labeling (P/R/F)	**Inner labeling (P/R/F)**
Outmost Entities Recognized	Exact	0.853 /0.005 /0.009	0.853 /0.005 /0.009
	Left	0.682 /0.542 /0.604	0.543 /0.555 /0.549
	Right	0.324 /0.070 /0.114	0.321 /0.070 /0.115
	Overall	**0.62/0.21/0.24**	**0.57/0.21/0.22**
Inner Entities Recognized	Exact	0.269 /0.333 /0.298	0.386 /0.618 /0.475
	Left	0.272 /0.405 /0.325	0.336 /0.618 /0.435
	Right	0.237 /0.376 /0.290	0.350 /0.694 /0.465
	Overall	**0.26/0.37/0.30**	**0.36/0.64/0.46**

Table 2 Performance of nested DNA entities

We hope these results can help in recognizing nested NEs, and also attract more attention to the nested NE problem. We are going to further our study by looking into more related issues.

References

J. Kim, et al. 2003. GENIA corpus – a semantically annotated corpus for bio-textmining. Bioinformatics, Vol 19.

J. Kim, et al. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA. Proceedings of JNLPBA.

K. Lee, et al. 2003. Two-Phase Biomedical NE Recognition based on SVMS. Proceedings of ACL Workshop on NLP in Biomedical.

Y. Lin, et al. 2004. A Maximum Entropy Approach to Biomedical Named Entity Recognition. Proceedings of KDD Workshop on Data Mining and Bioinformatics.

G. Nenadic, et al. 2004. Mining Biomedical Abstracts: What's in a Term? Proceedings of IJCNLP 2004.

B. Settles. 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. Proceedings of Joint Workshop on NLPBA.

G. Zhou, et al. 2004. Recognizing Names in Biomedical Texts: a Machine Learning Approach. *Bioinformatics*, Vol. 20, no. 7.