

# Using Unlabeled MEDLINE Abstracts for Biological Named Entity Classification

Manabu Torii      K. Vijay-Shanker  
torii@cis.udel.edu      vijay@cis.udel.edu

Department of Computer and Information Sciences, University of Delaware, Newark,  
DE 19716, USA

**Keywords:** natural language processing, named entity classification, bootstrapping

## 1 Introduction

Named Entity Recognition is a crucial step for Information Extraction from biological texts. By using surface clues such as capitalization, numbers, and special symbols, existing tools extract names of protein and other biological entities well. However, names of different entities share surface characteristics, and it is difficult to classify detected names based only on that attribute. As pointed out by Narayanaswamy *et al.* [3], extraction of names other than just protein and their classification can be important in two ways. First, in order to extract rich information, we want to identify specific classes of named entities. Second, classification of detected names can help in improving precision of protein name recognition as it can be used to eliminate the others.

Like Narayanaswamy *et al.* [3], we consider both lexical and contextual features for classification. They use hand created rules to exploit contextual information. However, it is difficult to develop rules that cover all the cases by hand, and it would be appropriate to consider Machine Learning (ML) methods to automatically obtain rules. Due to the lack of annotated texts, we use an unsupervised learning method with bootstrapping technique. In particular, we take Yarowsky's approach [4] used in Word Sense Disambiguation. In this paper, we classify detected names as *protein*, *chemical*, or *source*. Liu *et al.* [2] and Hatzivassiloglou *et al.* [1] also applied ML approach for disambiguation of biomedical and biological terms. Their methods use external information source to annotate texts, and do not require human annotation either. However, as far as we are aware, we are the first to introduce an unsupervised method using bootstrapping for named entity recognition in the biological domain.

## 2 Method and Results

First, we download MEDLINE abstracts, and apply the extraction system of Narayanaswamy *et al.* [3] to detect names. This part of the system also classifies name instances that have surface clues. Typically, the system detects names in texts with high precision and recall, and about 40% of detected names are classified. Next, Decision List (DL) [4] based on rule templates for lexical and contextual information is learned over name instances already classified. The learned DL is subsequently applied to the remaining instances, and once classified, they become a part of the training data for the next round. We have focused on contextual information, and provided rule templates that consider neighboring words and phrases. Instead of using exact words, we apply stemming (Porter Stemmer [5]), and do not distinguish numbers and Greek symbols. Some of the rules automatically learned are listed in Table 1. For example, an instance preceded by “*express(ion) of*” is classified as a protein name, but an instance preceded by “*express(ed) in*” is classified as a source name.

We downloaded over 6000 MEDLINE abstracts, and annotated them using the bootstrapping method described above. DL was learned over the annotated abstracts, and then tested on 50 abstracts annotated by domain experts. We downloaded over 6000 MEDLINE abstracts, and annotated them

Table 1: Contextual rules learned by decision list algorithm.

Protein	Preceded by <i>express of</i> , Preceded by <i>induct of</i> , Preceded by <i>matur</i> , Followed by <i>activ</i>
Chemical	Preceded by <i>dose</i> , Preceded by <i>exposur to</i> , Followed by <i>dimer</i> , Followed by <i>induc</i>
Source	Preceded by <i>express in</i> , Preceded by <i>activ in</i> , Followed by <i>line</i> , Followed by <i>synthesi in</i>

Table 2: Evaluation of the unsupervised learning method.

	Extraction System			Recognition System			Our System			Recog. + Our Sys.		
	prec.	rec.	f-val.	prec.	rec.	f-val.	prec.	rec.	f-val.	prec.	rec.	f-val.
Protein	0.88	0.39	0.54	0.92	0.62	0.74	0.78	0.79	0.79	0.82	0.83	0.83
Chemical	0.83	0.45	0.58	0.86	0.62	0.72	0.78	0.62	0.69	0.80	0.71	0.75
Source	0.78	0.31	0.45	0.80	0.46	0.59	0.72	0.44	0.54	0.74	0.52	0.61

The system by Narayanaswamy *et al.* [3] consists of the extraction module and the classification module. In the table above, *Extraction System* consists only of the extraction module, which we use for generating a seed corpus. *Recognition System* is the full system with the both modules. *Recog. + Our Sys.* is the system that applies rules learned by *Our System* to the output of *Recognition System*.

using the bootstrapping described above. The resulted DL was evaluated on 50 Medline abstracts annotated by domain experts.

### 3 Discussion

As hoped for, our ML method learned good contextual rules, and achieved high recall. Especially, those rule templates exploiting prepositions were known to be effective. Even more encouraging was that combined with human developed system, the results appear to be very good (“Recog. + Our Sys.” in Table 2). However, precision of the system could be improved. Inherently, our system should more sensitive to extraction errors, since it assigns specific classes to mis-extracted instances if a contextual clue exists nearby. By counting off those error cases, precision of our system jumps to 0.92, 0.93, and 0.92 for protein, chemical, and source. For the remaining errors, our quick inspection suggests possibility of improvement by imposing more constraints in use of contextual rules. For example, a contextual clue distant from a name instance should be overridden by a lexical clue or any linguistic clue that is more nearby. We are currently looking into this matter, and also examining the interaction between lexical and contextual information.

### Acknowledgment

We thank M. Narayanaswamy and K.E. Ravikumar for providing their extraction system, and helping us with the hand annotation of the test corpus.

### References

- [1] Hatzivassiloglou, V., Duboue, P.A., and Rzhetsky, A., Disambiguating proteins, genes, and RNA in text: A machine learning approach, *Bioinformatics*, 17(1):97–106, 2001.
- [2] Liu, H., Lussier, Y., and Friedman, C., Disambiguating biomedical terms in biomedical narrative text: An unsupervised method, *J. Biomedical Informatics*, 34(4):249–261, 2001.
- [3] Narayanaswamy, M., Ravikumar, K.E., and Vijay-Shanker, K., A biological named entity recognizer, *Proc. Pacific Symp. Biocomputing*, 2003. (in press)
- [4] Yarowsky, D., Unsupervised word sense disambiguation rivaling supervised methods, *Proc. Assoc. Computational Linguistics*, 189–196, 1995.
- [5] <http://www.tartarus.org/~martin/PorterStemmer/>