BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 11 2005, pages 2794–2796
doi:10.1093/bioinformatics/bti414

Databases and ontologies

POSBIOTM–NER: a trainable biomedical named-entity recognition system

Yu Song*, Eunju Kim, Gary Geunbae Lee and Byoung-kee Yi
Department of CSE, POSTECH, Pohang, 790-784, Korea

Received on August 9, 2004; revised on March 4, 2005; accepted on March 28, 2005
Advance Access publication April 6, 2005

ABSTRACT

Summary: POSBIOTM–NER is a trainable biomedical named-entity recognition system. POSBIOTM–NER can be automatically trained and adapted to new datasets without performance degradation, using CRF (conditional random field) machine learning techniques and automatic linguistic feature analysis. Currently, we have trained our system on three different datasets. GENIA–NER was trained based on GENIA Corpus, GENE–NER based on BioCreative data and GPCR–NER based on our own POSBIOTM/NE corpus, respectively, which would be used in GPCR-related pathway extraction.

Availability: http://isoft.postech.ac.kr/Research/BioNER/POSBIOTM/NER/main.html

Contact: songyu@postech.ac.kr

1 INTRODUCTION

There exists a large amount of biomedical literature and the volume continues to grow exponentially. These publications embody a store of knowledge and information of interactions and relations among biological entities, which is very important for the understanding of biological processes. To perform interaction extraction from literature, the most elementary and core problem is the identification of material names concerned, which is known as named-entity recognition (NER) in natural language processing community.

Diversified named-entities are involved in different sorts of interactions in the specific domain. There is an increasing need for tools that can be adapted to multifarious biomedical domain to recognize the requisite entities.

Recent researchers mainly used machine learning models for the biomedical named-entity task. GENIA corpus (Kim *et al*., 2003) and BioCreative corpus (Blaschke *et al*., 2004, http://www.pdg.cnb.uam.es/BioLINK/workshop BioCreative 04/handout/) are freely available as training materials. The system performance of GNEIA corpus varies from 63.0 to 72.2, according to the report on Bio-Entity Recognition task at BioNLP/NLPBA 2004 (Kim *et al*., 2004). The performances of systems at BioCreative vary from 50.1 to 83.2. In the case of closed corpus, the best performance is 82.2 (Blaschke *et al*., 2004).

The named-entity recognition problem is regarded as a classification problem, marking up input sequences with named-entity category labels. The input sequences are represented as sets of features. The models based on hidden Markov model (HMM) and maximum entropy are generalizable and adaptable to new classes of words. HMM is a standard tool in the NER task. Collier *et al*. (2000) used an HMM to extract the names of genes and gene products. Zhou *et al*. (2004) proposed a rich set of features, integrated via HMM with back-off modeling. However, HMM suffers from the data sparseness problem and is not practical to represent multiple interacting features or long-range dependencies of the observations. Also it has a very strict independent assumption on the observation. An alternative is a conditional model, which allows arbitrary non-independent features on the observation sequence. Dingare *et al*. (2004) presents a maximum entropy Markov model (MEMM)-based system incorporating diverse set of features for identifying genes and proteins in biomedical abstracts for the BioCreative task1A. Still, MEMM is subject to the label bias[1] problem (Lafferty *et al*., 2001).

Our POSBIOTM–NER system is designed as a trainable Biomedical NER system. It can be adapted to new datasets without any further human effort. In our system, we adopt the Conditional Random Field (CRF) model (Lafferty *et al*., 2001) for the biomedical named-entity recognition task. The CRF model has all the advantages of conditional models and can handle the label bias problem. The CRF model is globally conditioned on the input sequences, resulting in a global optimal model. Various linguistic features are used to achieve better performance, such as part-of-speech[2] (POS) tag and base noun phrase[3] (NP) tag. Our system has been validated by three different datasets, including GENIA corpus and BioCreative corpus. For all the datasets, our system gives competitive performances. Our system also provides a client tool to access to our system and make use of modules. Please visit our system website for details.

2 SYSTEM DESCRIPTION

POSBIOTM–NER is a trainable biomedical named-entity recognition system. An overview of the system architecture is shown in Figure 1. The system is composed of named-entity (NE) detection Module and NER model training module. The training module can be adapted to a new training data and build a new NER model. The produced NER models are employed in the detection module to perform the named-entity recognition task. Currently, we provide three target-specific NER models: GENIA–NER model, GENE–NER model and GPCR–NER model. One easy and convenient way to make use of our system to perform the named-entity task is to utilize

*To whom correspondence should be addressed.

[1] Label bias: bias toward states with fewer outgoing transitions.

[2] Part-of-speech: the part of speech is the term used to describe how a particular word is used. For example, noun, verb, etc.

[3] NP: a noun phrase comprises a noun and any number of associated modifiers.

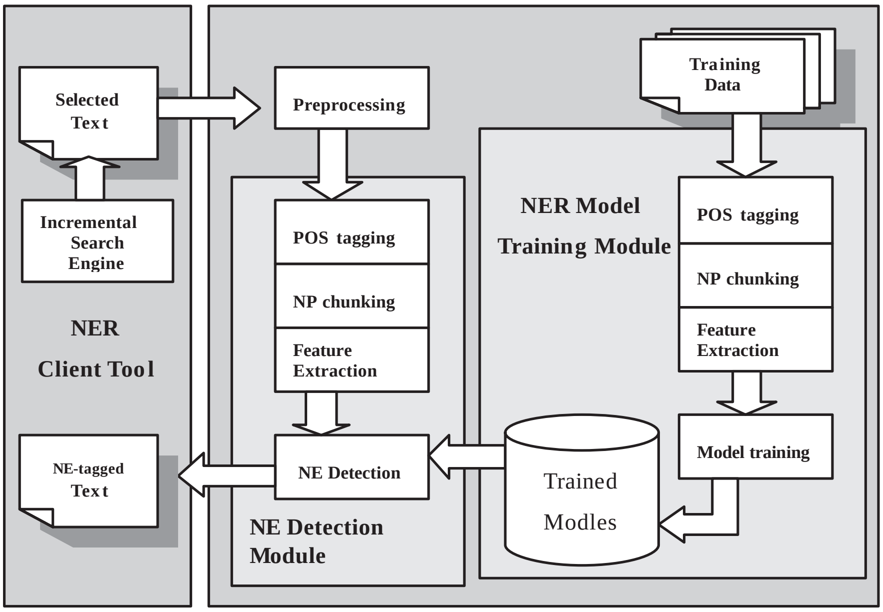

Fig. 1. An overview of the system architecture.

the client tool. Client tool is a part of our currently developing POSBIOTM workbench. The workbench provides a friendly graphical user interface and a set of tools which support collecting, managing, creating, annotating and exploiting biomedical text resources. The client tool and the whole binary code are freely available at http://isoft.postech.ac.kr/Research/BioNER/POSBIOTM/NER/main.html.

The client tool supports access to PubMed by using our incremental PubMed search engine. Users can maintain their own search keyword lists with specific search strategies, such as searching only in a predefined subset of Pubmed citations that meet the interest of users. The users can select one of the three NER models and transfer the chosen texts to the system, and then the detection module will perform the named-entity recognition task.

Before the selected texts are passed to the NE detection module, they are split into sentences and tokenized at the preprocessing step. Then each token in the sentences will be assigned with a POS tag and an NP tag. For each token, various kinds of features are extracted. These features are used to fully describe the local contexts of this token. The NER results are stored in the XML form. One example of the output is as follows:

```
The <protein>sphingosine-1-phosphate receptor </protein>
<protein>EDG-1 </protein> is essential for <protein>
platelet-derived growth factor </protein> -induced <cellular
process>cell motility </cellular process>.
```

Most of the current biomedical NER systems are designed to recognize only one given set of named entities, and moreover, handcrafted rules are used to enhance the performance. So, they can hardly be adapted to new datasets. Our system is a fully automatic trainable system. It can be adapted to different datasets without any modification. The only thing a user should provide is a training corpus. Feature sets, which are used to describe the distinctive characteristics of the tokens, as well as other useful information, will be obtained directly from the training corpus by using linguistic processing. The training corpus has to be provided in IOB notation. 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.

Our POSBIOTM–NER uses the following basic linguistic features:

- Surface word—word itself and word combination. Only in the case that the previous/current/next words are in the surface word dictionary.
- Word feature—orthographical features 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.
- POS tag—part-of-speech tag of the previous/current/next words.
- NP tag—base noun phrase tag of the previous/current/next words.

3 DIFFERENT NER MODELS

The GENIA–NER model uses GENIA corpus version 3.02 as a training corpus, which contains five target named-entity categories: protein, cell_line, cell_type, DNA and RNA. The GENIA corpus consists of 2000 MEDLINE abstracts, mainly concerning transcription factors in human blood cells and 404 abstracts are used as a test data.

The GENE–NER model uses BioCreative corpus. The aim of the GENE–NER model is the identification of terms in biomedical research articles, which are gene and/or protein names. The training corpus consists of 7.5 k sentences, selected from MEDLINE according to their likelihood of containing gene names. The test data consists of 2.5 k sentences.

GPCR–NER model aims at recognizing four target named-entity categories: protein, gene, small molecule and cellular process. The whole corpus consists of 50 full articles related to GPCR (G-protein coupled receptor) signal transduction pathway. From the corpus, 45 articles are randomly chosen as training data, and the remaining articles are used as test data. The outcome of this model is used directly in the GPCR-related pathway extraction.

4 PERFORMANCE

According to the different training data, we finally achieved a performance of 69.45 in F-measure for five categories of NER on the GENIA corpus (ver. 3.0), 79.82 in F-measure in gene recognition on the BioCreative data and 73.70 in F-measure for about four categories on our POSBIOTM/NE corpus, respectively. The feasibility and portability of our system were validated by its application to these three datasets. It takes ∼8–12 h to train an NER model on an Intel Pentium 4 GEON Linux server. To tag a 300-word abstract, it takes ∼0.3 s. The NER client tool is implemented in Java and should run on any operating system with Java runtime environment version 1.4 or higher. It has been tested under Windows and Linux.

ACKNOWLEDGEMENTS

This research was supported by BIT Fusion project (by MOST) and BK21 project (by Ministry of Education).

REFERENCES

Blaschke,C., Hirschman,L. and Yeh,A. (eds), (2004) *Proceedings of the BioCreative Workshop, Granada, March*.

Collier,N., Nobata,C. and Tsujii,J. (2000) Extracting the names of genes and gene products with a hidden Markov model. In: *Proceedings of International Conference on Computational Linguistics (COLING 2000)*, Saarbzucken, Germany, pp. 201–207.

Dingare,S., Finkel,J., Manning,C., Nissim,M. and Alex,B. (2004) Exploring the boundaries: gene and protein identification in biomedical text. In: *Proceedings of the BioCreative Workshop*, Granada, Spain.

Kim,J-D. *et al*. (2003) GENIA corpus—a semantically annotated corpus for biotextmining. *Bioinformatics*, **19** (Suppl. 1) i180–i182.

Kim,J-D., Ohta,T., Tsuruoka,Y., Tateisi Y. and Collier,N. (2004) Introduction to the bio-entity recognition task at JNLPBA. In: *Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Application (JNLPBA 2004)*, Geneva, Switzerland.

Lafferty,J., McCallum,A. and Pereira,F. (2001) Conditional random fields: probabilistic models for segmenting and labelling sequence data. In: *Proceedings of the 18th International Conference on Machine Learning (ICML, 2001)*, Williamstown, MA, Morgan Kauffmann Publishers Inc., pp. 282–289.

Zhou,G.D. *et al*. (2004) Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, **20**, 1178–1190.