POSBIOTM/W: A Development Workbench For Machine Learning Oriented Biomedical Text Mining System *

Kyungduk Kim, Yu Song, Gary Geunbae Lee
Department of Computer Science and Engineering
Pohang University of Science & Technology (POSTECH)
San 31, Hyoja-Dong, Pohang, 790-784, Republic of Korea
{getta, songyu, gblee}@postech.ac.kr

Abstract

The POSBIOTM/W[1] is a workbench for machine-learning oriented biomedical text mining system. The POSTBIOTM/W is intended to assist biologist in mining useful information efficiently from biomedical text resources. To do so, it provides a suit of tools for gathering, managing, analyzing and annotating texts. The workbench is implemented in Java, which means that it is platform-independent.

1 Introduction

Large amounts of biomedical literature exist and the volume continues to grow exponentially. Following the increase of literature, there is growing need for appropriate tools in support of collecting, managing, creating, annotating and exploiting rich biomedical text resources.

Especially, information on interactions among biological entities is very important for understanding the biological process in a living cell (Blascheke et. al., 1999). In our POSBIOTM/W workbench, we use a supervised machine learning method to generate rules automatically to extract biological events from free texts with minimum human effort. And we adopt the Conditional Random Fields (CRF) model (Lafferty et. al.,2001) for the biomedical named-entity recognition (NER) task. Finally, to reduce the labeling effort in a larger extent we incorporate an active learning idea into the workbench.

* The research was supported by Brain Neuro Informatics Research program by MOCIE.

[1] POSBIOTM/W stands for POSTECH Bio-Text Mining System Workbench

2 System Description

The POSBIOTM/W comprises a set of appropriate tools to provide users a convenient environment for gathering, managing and analyzing biomedical text and for named-entity annotation. The workbench consists of four components: Managing tool, NER tool, Event Extraction Tool and Annotation Tool. And we adopt an active learning idea into the workbench to improve the NER and the Event Extraction module's performance. The overall design is shown in Figure 1.

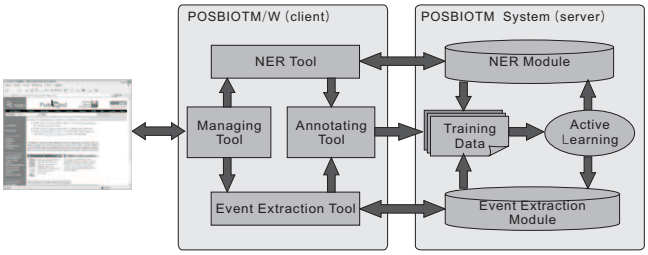

Figure 1: Overview of POSBIOTM/W

2.1 Managing tool

Main objective of the Managing tool is to help biologists search, collect and manage literatures relevant to their interest. Users can access to the PubMed database of bibliographic information using quick searching bar and incremental PubMed search engine.

Proceedings of HLT/EMNLP 2005 Demonstration Abstracts, pages 36–37,
Vancouver, October 2005.

2.2 NER tool

The NER tool is a client tool of POSBIOTM-NER module and able to automatically annotate biomedical-related texts. The NER tool provides access to three target-specific named entity models - GENIA-NER model, GENE-NER model and GPCR-NER model. Each of these model is trained based on GENIA-Corpus (Kim et. al., 2003), BioCreative data (Blaschke et. al., 2004) and POSBIOTM/NE corpus[2] respectively. In POSBIOTM-NER system, we adopt the Conditional Random Fields (CRF) model (Lafferty et. al., 2001) for the biomedical NER task.

2.3 Event Extraction tool

The Event Extraction tool extracts several biological events from texts using automatically generated rules. We use a supervised machine learning method to overcome a knowledge-engineering bottleneck by learning event extraction rules automatically. We modify the WHISK (Soderland, 1999) algorithm to provide a two-level rule learning method as a divide-and-conquer strategy. In two-level rule learning, the system learns event extraction rules which are inside of the noun chunk at first level, and then it learns the rules for whole sentence.

Since the system extracts biological events using automatically generated rules, we can not guarantee that every extracted event is always correct because many different rules can be applied to the same sentence. Therefore we try to verify the result with a Maximum Entropy (ME) classifier to remove incorrectly extracted events. For each extracted event, we verify each component of the event with the ME classifier model. If one component is contradicted to the class assigned by the classification model, we will remove the event. For detail event extraction process, please consult our previous paper (Kim et. al., 2004).

2.4 Annotation tool

Our workbench provides a Graphical User Interface based Annotation tool which enables the users to annotate and correct the result of the named-entity recognition and the event extraction. And users can upload the revised data to the POSBIOTM system, which would contribute to the incremental build-up of named-entity and relation annotation corpus.

2.5 Active learning

To minimize the human labeling effort, we employ the active learning method to select the most informative samples. We proposed a new active learning paradigm which considers not only the uncertainty of the classifier but also the diversity of the corpus, which will soon be published.

[2] POSBIOTM/NE corpus, our own corpus, is used to identify four target named entities: protein, gene, small molecule and cellular process.

References

Christian Blaschke, Andrade, M.A., Ouzouis, C., Valencia, A.. 1999. *Automatic extraction of biological information from scientific text : protein-protein interactions*. Intelligent Systems for Molecular Biology 60-67.

Christian Blaschke, L. Hirschman, and A. Yeh, editors. 2004. *Proceedings of the BioCreative Workshop, Granda, March.* `http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/handout/`

Eunju Kim, Yu Song, Gary Geunbae Lee, Byoung-Kee Yi. 2004. *Learning for interaction extraction and verification from biological full articles*. Proceedings of the ACM SIGIR 2004 workshop on search and discovery in bioinformatics, July 2004, Sheffield, UK

J.-D. Kim, T. Ohta, Y. Tateisi and J. Tsujii 2003. *GENIA corpus - a semantically annotated corpus for bio-textmining*. Bioinformatics, Vol 19 Suppl. 1 2003, pages i180-i182

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

Soderland S. 1999. *Learning information extraction rules for semi-structured and free text*. Machine Learning, volume 34, 233-272.