Developing a Robust Part-of-Speech Tagger for Biomedical Text

Yoshimasa Tsuruoka[1,2], Yuka Tateishi[1,2], Jin-Dong Kim[1,2], Tomoko Ohta[1,2], John McNaught[3,5], Sophia Ananiadou[4,5], and Jun'ichi Tsujii[2,3]

[1] CREST, JST (Japan Science and Technology Agency),
Honcho 4-1-8, Kawaguchi-shi, Saitama 332-0012, Japan
{tsuruoka, yucca, jdkim, okap}@is.s.u-tokyo.ac.jp
[2] University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
tsujii@is.s.u-tokyo.ac.jp
[3] School of Informatics, University of Manchester,
P.O.Box 88, Sackville St, Manchester M60 1QD, UK
[4] School of Computing, Science and Engineering, Salford University,
Salford, Greater Manchester M5 4WT, UK
S.Ananiadou@salford.ac.uk
[5] The National Centre for Text Mining,
P.O.Box 88, Sackville St, Manchester M60 1QD, UK

Abstract. This paper presents a part-of-speech tagger which is specifically tuned for biomedical text. We have built the tagger with maximum entropy modeling and a state-of-the-art tagging algorithm. The tagger was trained on a corpus containing newspaper articles and biomedical documents so that it would work well on various types of biomedical text. Experimental results on the Wall Street Journal corpus, the GENIA corpus, and the PennBioIE corpus revealed that adding training data from a different domain does not hurt the performance of a tagger, and our tagger exhibits very good precision (97% to 98%) on all these corpora. We also evaluated the robustness of the tagger using recent MEDLINE articles.

1 Introduction

Since a huge amount of biomedical knowledge is described in the literature, automatic information extraction from biomedical documents is increasingly important for many researchers in this domain.

For extracting information from text, many natural language processing (NLP) techniques can be employed. For example, a simple approach to extracting information about protein-protein interactions would involve scanning the text for particular verbs and neighboring noun phrases by applying some linguistic patterns on words and their part-of-speech (POS) tags. A more sophisticated way would be to use parsers to deeply analyze the syntactic and semantic relations among the entities in the sentences.

P. Bozanis and E.N. Houstis (Eds.): PCI 2005, LNCS 3746, pp. 382–392, 2005.

In order to carry out noise-free information extraction, the very basic step in natural language processing of POS tagging must be performed with high precision. The precision of POS tagging not only directly affects the performance of pattern-based approaches but also influences the accuracy of parsing which in general uses the POS tags on the words as part of the input [1,2].

For documents like newspaper articles, there are a number of publicly available NLP tools including POS taggers, chunkers (shallow parsers), and syntactic parsers. However, the problem for researchers working on biomedical information extraction is that such tools do not necessarily work well on biomedical documents because the characteristics of biomedical text are considerably different from those of newspaper articles, which are often used as the training data for NLP tools [3,4] . Table 1 lists some examples of tagging errors made by the TnT tagger [5], a popular HMM-based POS tagger, which is trained on the Wall Street Journal corpus, when it is applied to biomedical text.

Recently, two large biomedical corpora that are annotated with POS tags have become publicly available: the GENIA corpus [6] and the PennBioIE corpus [3]. In building these corpora, the developers used a POS tagger to reduce manual annotation effort and reported that they could achieve better performance than with a standard tagger by using an already annotated portion of their corpus for training the tagger. Their observation clearly suggests that we might be able to build a good POS tagger for biomedical documents if we use their corpora as the training data.

However, since each corpus consists of text extracted from a particular domain (e.g. transcription factors for the GENIA corpus) and does not cover the entire characteristics of biomedical text, there are still remaining issues to be addressed: (1) Which corpus should we use for training? (2) Should we use a single corpus or combine two corpora? (3) Does the combination of corpora from different domains have a bad effect on trained tagger performance? if so, how much?

The purpose of this paper is to clarify these issues and develop a reliable POS tagger that can be used as a fundamental tool for biomedical text mining. In this paper we evaluate the performance of a part-of-speech tagger by using different combinations of corpora as the training data, and show how the domain of the training corpus affects the tagging performance. We also investigate the robustness of the trained taggers using recent MEDLINE articles.

2 POS Tagging Algorithm

As our POS tagging algorithm, we adopt a method based on a Cyclic Dependency Network proposed by Toutanova et al. [8], which is currently one of the best algorithms for English POS tagging. Unlike the popular Maximum Entropy Markov Modeling (MEMM) approach, this method can incorporate features about the tags on both sides of the classification target. Toutanova et al. achieved an accuracy of 97.24% on sections 22-24 in the Wall Street Journal corpus, using sections 0-18 for training. On the same sets for training and testing, Gimenez