A Token Centric Part-of-Speech Tagger for Biomedical Text

Neil Barrett⋆ and Jens Weber-Jahnke

Department of Computer Science, University of Victoria, Victoria, Canada
{nbarrett,jens}@uvic.ca
www.simbioses.ca

Abstract. A difficulty with part-of-speech (POS) tagging of biomedical text is accessing and annotating appropriate training corpora. The latter may result in POS taggers trained on corpora that differ from the tagger’s target biomedical text. In such cases where training and target corpora differ tagging accuracy decreases. We present a POS tagger that is more accurate than two frequently used biomedical POS taggers (Brill and TnT) when trained on a non-biomedical corpus and evaluated on the MedPost corpus (our tagger: 81.0%, Brill: 77.5%, TnT: 78.2%). Our tagger is also significantly faster than the next best tagger (TnT). It estimates a tag’s likelihood for a token by combining prior probabilities (using existing methods) and token probabilities calculated in part using a Naive Bayes classifier. Our results suggest that future work should reexamine POS tagging methods for biomedical text. This differs from the work to date that has focused on retraining existing POS taggers.

Keywords: part-of-speech, POS, tag, TnT, Brill, Naive Bayes, MedPost, accuracy, biomedical.

1 Introduction

Biomedical texts include information related to human health. Several examples of biomedical texts are research papers, medical reference books and clinical documents. Automated processing of biomedical texts can support effective health tools (e.g. [22]). These tools may enhance research processes (e.g. information search) or directly impact quality of care.

Natural language processing (NLP) is computer processing of human language [11] and may be applied to biomedical texts. As an example, Zingmond and Lenert [22] apply NLP to chest X-ray reports to identify new and expanding neoplasms (abnormal tissue growth) for the purpose of monitoring patient follow-ups. Friedman, Knirsch, Shagina and Hripcsak [8] apply NLP to discharge summaries to determine the severity of a patient’s community acquired pneumonia.

⋆ This research was funded by the Natural Sciences and Engineering Research Council of Canada.
Part-of-speech (POS) tagging assigns tags to tokens, such as assigning the tag *noun* to the token *paper*. POS tags and tagging are components of NLP. POS tags supply information about words and surrounding words. For example, POS tags indicate which type of word, such as a noun or a verb, will occur in the vicinity of a tagged word \[11\]. POS tags affect word pronunciation in text to speech systems \[11\]. They also improve information retrieval from textual documents, such as the retrieval of names, times, dates and other named entities \[11\]. POS tags and tagging are an important component of NLP, including NLP of biomedical texts \[5,20\].

POS taggers are typically trained on linguistically annotated corpora created by human experts. A difficulty with POS tagging of biomedical text is accessing and annotating appropriate training corpora. The latter may result in POS taggers trained on corpora that differ from the tagger’s target biomedical text.

In general (see Section 2 for details of previous research), training POS taggers on non-biomedical corpora and applying these trained taggers to biomedical corpora results in an approximate 10% decrease in tagging accuracy. Tagging accuracy is the tagger’s ability to assign the correct POS tag to a token. Even in cases where a tagger is trained on one biomedical corpus and applied to a different biomedical corpus a decrease in tagging accuracy can occur \[5\]. There may be situations where linguistically annotating a corpus for training purposes is infeasible (e.g. financial cost). In such situations it would be beneficial to develop POS tagging methods that perform well across differing corpora.

In this paper, we present a POS tagger that is more accurate than two frequently used biomedical POS taggers when trained on a non-biomedical corpus and evaluated on the MedPost corpus \[19\]. Our algorithm is also significantly faster than the best tagger of the two frequently used biomedical taggers. Given our work, we suggest how to reduce the difference in accuracy observed when training and application corpora differ.

This paper is organized as follows. Section 2 discusses related work. Our tagger is presented in Section 3. It is evaluated using the method presented in Section 4 with results appearing in Section 5. We discuss these results and their implications to POS tagging of biomedical text in Section 6. Section 7 concludes this paper.

## 2 Background

Biomedical POS taggers are constructed by retraining existing taggers on biomedical corpora or by supplementing an existing tagger with a biomedical lexicon. In most cases, transformation based or Hidden Markov Model (HMM) taggers are used. These taggers are briefly explained. For more detail see Jurafsky and Martin \[11\].

Transformation based taggers, such as the Brill tagger \[3\], work by successively refining assigned tags. Each refinement applies a more specific rule with the goal of assigning a more appropriate tag. Transformation rules may be specified, or learned from templates and training corpora. An example template is: change