Summary of Discussion for

Comparison of character-level and part of speech features for name recognition in biomedical texts

by Nigel Collier and Koichi Takeuchi

Journal of Biomedical Informatics 37(2004) 423-435

INLS 279: Bioinformatics Research Review

Presented by Christopher Maier

2005-03-08

This paper explored the extent to which different word representations (orthography or POS) affected named-entity (NE) recognition in the biomedical text domain. Orthography, the way a word "looks" on a per-character level, is a popular representation choice in this NE domain because of the very distinctive orthographies of gene and protein names. Part-of-speech (POS) tags are also a leading choice. By training a POS tagger on an annotated domain corpus, the tagger can help demarcate term and phrase boundaries, and disambiguate word senses. The goal of this paper was to examine the extent to which each of these representations contributed to NE recognition, as well as to determine the effect of these representations used in combination.

The learning algorithm employed in this paper was the support vector machine (SVM). This was chosen due to SVM's ability to handle sparse datasets (text is a notorious example of such a dataset), as well as its ability to handle relatively small numbers of patterns.

We felt that one of the strong points of this article was its very thorough background section. It did a very good job of explaining the current state-of-the-art in the NE field. The rationale and usage of orthographic and POS representations was clearly laid out. The importance of the GENIA corpus in the biomedical NE field was explained, as well. The figure showing the differences between the two POS taggers used, as well the performance improvements that come with in-domain retraining was appreciated. Though retraining often resolved erroneous POS tagging, concern was expressed about both taggers' propensities to split components of a biological NE's name apart into separate tokens. For example, the protein name I?Ba was split into four tokens by both taggers. True, re-training led to all four tags receiving a 'noun' designation, but the inability to recognize such a name as a single token was worrisome. Perhaps such behavior was contributory to the ultimate finding that POS information actually *reduced* system performance when paired with orthographic information?

Table 4 was helpful for demonstrating the predictive power of several orthograhpic classes. The discussion on deterministic versus non-deterministic orthographic tagging was a bit vague, however. The consensus is that deterministic tagging results in one, somehow "best fitting", orthographic tag being applied to a token, whereas non-deterministic tagging seems to result in as many applicable tags being applied to a token as possible. More description concerning the process by which orthographic tags were determined by both these methods would have been appreciated, as it is not at all clear from the brief mention they received in the text.

The description of the 'base' experimental treatment was vague as well. We assumed that 'surface word' meant the particular token under examination. The meaning of 'lemma' in this context was less clear.

It was universally acknowledged that the result tables were poorly and confusingly designed. Far too much information was put into single tables, where we felt that several tables and possibly some line graphs would have been more clear.

A more explicit description of both the input to as well as the output from the SVM algorithm would have been helpful. It was unclear exactly what these would look like. The derivation of the final statistics (precision, recall, F-score) was thus hazy as well. Intuitively we understand that the orthographic and POS information is being combined somehow, but exactly how was not clearly explained.

Aside from the above mentioned issues, we felt that this was a good article. It explained the current understanding of the field, posed an important research question, and attempted an in-depth analysis of that question. The bibliography of the paper is a particularly rich one, offering several jumping-off points for further examination of the biological NE field.