Adaptation of POS Tagging for Multiple BioMedical Domains

John E. Miller[1] **Manabu Torii**[2] **K. Vijay-Shanker**[1]

[1]Computer & Information Sciences
University of Delaware
Newark, DE 19716
{jmiller,vijay}@cis.udel.edu

[2]Biostatistics, Bioinformatics and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
mt352@georgetown.edu

1 Introduction

Part of Speech (POS) tagging is often a prerequisite for tasks such as partial parsing and information extraction. However, when a POS tagger is simply ported to another domain the tagger's accuracy drops. This problem can be addressed through hand annotation of a corpus in the new domain and supervised training of a new tagger. In our methodology, we use existing raw text and a generic POS annotated corpus to develop taggers for new domains without hand annotation or supervised training. We focus in particular on out-of-vocabulary words since they reduce accuracy (Lease and Charniak. 2005; Smith et al. 2005).

There is substantial information in the derivational suffixes and few inflectional suffixes of English. We look at individual words and their suffixes along with the morphologically related words to build a domain specific lexicon containing POS tags and probabilities for each word.

2 Adaptation Methodology

Our methodology is described in detail in Miller et al (2007) and summarized here: 1) Process generic POS annotated text to obtain state and lexical POS tag probabilities. 2) Obtain a frequency table of words from a large corpus of raw sub-domain text. 3) Construct a partial sub-domain lexicon matching relative frequencies of morphologically related words with words from the generic annotated text averaging POS probabilities of the k nearest neighbors. 4) Combine common generic words and orthographic word categories with the partial lexicon making the sub-domain lexicon. 5) Train a first order Hidden Markov Model (HMM) by Expectation Maximization (EM). 6) Apply the Viterbi algorithm with the HMM to tag sub-domain text.

3 Adaptation to Multiple Domains

Molecular Biology Domain: We used the Wall Street Journal corpus (WSJ) (Marcus et al, 1993) as our generic POS annotated corpus. For our raw un-annotated text we used 133,666 abstracts from the MEDLINE distribution covering molecular biology and biomedicine sub-domains. We split the GENIA database (Tateisi et al, 2003) into training and test portions and ignored the POS tags for training. We ran a 5-fold cross validation study and obtained an average accuracy of 95.77%.

Medical Domain: Again we used the WSJ as our generic POS annotated corpus. For our raw un-annotated text we used 164,670 abstracts from the MEDLINE distribution with selection based on 83 journals from the medical domain. For our HMM EM training we selected 1966 abstracts (same journals). For evaluation purposes, we selected 1932 POS annotated sentences from the MedPost (Smith et al, 2004) distribution (same journals). The MedPost tag set coding was converted to the Penn Treebank tag set using the utilities provided with the MedPost tagger distribution. We obtained an accuracy of 93.17% on the single medical test corpus, a substantial drop from the 95.77% average accuracy obtained in the GENIA corpus.

4 Coding Differences

We looked at high frequency tagging errors in the medical test set and found that many errors resulted directly from the differences in the coding styles between GENIA and MedPost. Our model reflects the coding style of the WSJ, used for our generic POS annotated text. GENIA largely followed the WSJ coding conventions. Annotation in the 1932 sentences taken from MedPost had some systematic differences in coding style from this.

BioNLP 2007: Biological, translational, and clinical language processing, pages 179–180,
Prague, June 2007.

Identified Differences: Lexical differences: 1) Words such as 'more' and 'less' are JJR or RBR in WSJ/GENIA but JJ or RB in MedPost. 2) Tokens such as %, =, /, <, > are typically NN or JJ in WSJ/GENIA but SYM in MedPost. 3)'be' is VB in WSJ/GENIA but VB or VBP in MedPost. 4) Some orthographic categories are JJ in WSJ/GENIA but NN in MedPost. Transition discrepancies: 1) Verbs are tagged VB following a TO or MD in WSJ/GENIA but only following a TO in MedPost. 2) MedPost prefers NN and NN-NN sequences.

***Ad Hoc* Adjustments**: We constructed a new lexicon accounting for some of the lexical differences and attained an accuracy of 94.15% versus the previous 93.17%. Next we biased a few initial state transition probabilities, changing P(VB|MD) from very high to a very low and increasing P(NN|NN), and attained an accuracy of 94.63%.

As the coding differences had nothing to do with suffixes and suffix distributions, the central part of our methodology, we tried some *ad hoc* fixes to determine what our performance might have been. We suffered at least a 1.46% drop in accuracy due to differences in coding, not language use.

5 Evaluation

The table shows the accuracy of our tagger and a few well-known taggers in our target biomedical sub-domains.

Molecular Biology	%Accuracy
- Our tagger (5-fold)	95.8%
- MedPost	94.1%
- Penn BioIE[1]	95.1%
- GENIA supervised	98.3%
Medical Domain	
- Our tagger	93.17%
- Our tagger (+ lex bias)	94.15%
- Our tagger (+ lex & trans bias)	94.63%
- MedPost supervised[2]	96.9%

The MedPost and Penn BioIE taggers used annotated text and supervised training in other biomedical domains, but they were not trained specifically for the GENIA Molecular Biology subdomain. Our tagger seems competitive with these taggers. We cannot claim superior accuracy as these taggers may suffer the same coding bias effects we have noted. The superior performance of the GENIA tagger (Tsuruoka et al. 2005) in the Molecular Biology/GENIA domain and the MedPost tagger (Smith et al. 2004) in its biomedical domain owes to their use of supervised training on an annotated training set with evaluation on a test set from the same domain. The approximate 1.5% bias effect due to coding differences is attributable to organizational differences in POS.

6 Conclusions

To cope with domain specific vocabulary and uses of vocabulary, we exploited the suffix information of words and related words to build domain specific lexicons. We trained our HMM using EM and un-annotated text from the specialized domains. We assessed accuracy versus annotated test sets in the specialized domains, noting discrepancies in our results across specialized domains, and concluding that our methodology performs competitively versus well-known taggers that used annotated text and supervised training in other biomedical domains.

References

M. Lease and E. Charniak. 2005. Parsing Biomedical Literature. IJCNLP-05: 58-69.

M. Marcus, B. Santorini, M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. Comp. Ling., 19:313-330.

J.E. Miller, M. Torii, K. Vijay-Shanker. 2007. Building Domain-Specific Taggers Without Annotated (Domain) Data. EMNLP-07.

L. Smith, T. Rindflesch, W.J. Wilbur. 2004. MedPost: a part-of-speech tagger for bioMedical text. Bioinformatics 20 (14):2320-2321.

L. Smith, T. Rindflesch, W.J. Wilbur. 2005. The importance of the lexicon in tagging biomedical text. Natural Language Engineering 12(2) 1-17.

Y. Tateisi, T. Ohta, J. Dong Kim, H. Hong, S. Jian, J. Tsujii. 2003. The GENIA corpus: Medline abstracts annotated with linguistic information. Third meeting of SIG on Text Mining, ISMB.

Y. Tsuruoka, Y. Tateishi, J.D. Kim, T. Ohta, J. McNaught, S. Ananiadou, J. Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text, Advances in Informatics, LNCS 3746: 382-392.

[1] PennBioIE. 2005. Mining The Bibliome Project. http://bioie.ldc.upenn.edu/.

[2] Based on Medpost test set of 1000 sentences, not on our test set of 1932 sentences.