Subdomain adaptation of a POS tagger with a small corpus

Yuka Tateisi
Faculty of Informatics
Kogakuin University
Nishishinjuku 1-24-2
Shinjuku-ku, Tokyo, 163-8677, Japan

Yoshimasa Tsuruoka
School of Informatics
University of Manchester
Manchester M60 1QD, U.K.

Jun-ichi Tsujii
Dept. of Computer Science
University of Tokyo
Hongo 7-3-1, Bunkyo-ku,
Tokyo 113-0033, Japan
School of Informatics
University of Manchester
Manchester M60 1QD, U.K.

1 Introduction

For the domain of biomedical research abstracts, two large corpora, namely GENIA (Kim et al 2003) and Penn BioIE (Kulik et al 2004) are available. Both are basically in *human* domain and the performance of systems trained on these corpora when they are applied to abstracts dealing with other species is unknown. In machine-learning-based systems, re-training the model with addition of corpora in the target domain has achieved promising results (e.g. Tsuruoka et al 2005, Lease et al 2005). In this paper, we compare two methods for adaptation of POS taggers trained for GENIA and Penn BioIE corpora to *Drosophila melanogaster* (fruit fly) domain.

2 Method

Maximum Entropy Markov Models (MEMMs) (Ratnaparkhi 1996) and their extensions (Tutanova et al 2003, Tsuruoka et al 2005) have been successfully applied to English POS tagging. Here we use second-order standard MEMMs for learning POS. where the model parameters are determined with maximum entropy criterion in combination a regularization method called inequality constraints (Kazama and Tsujii 2003). This regularization method has one non-negative meta-parameter called width-factor that controls the "fitness" of the model parameters to the training data.

We used two methods of adapting a POS tagging model. One is to add the domain corpus to the training set. The other is to use the reference distribution modeling, in which the training is performed only on the domain corpus and the information about the original training set is incorporated in the form of the r*eference distribution* in the maximum entropy formulation (Johnson et al 2000, Hara et al 2005).

A set of 200 MEDLINE abstracts on *D. melanogaster*, was manually annotated with POS according to the scheme of the GENIA POS corpus (Tateisi et al 2004) by one annotator. The new corpus consists of 40,200 tokens in 1676 sentences. From this corpus which we call "Fly" hereafter, 1024 sentences are randomly taken and used for training. Half of the remaining is used for development and the rest is used for testing.

We measured the accuracy of the POS tagger trained in three settings:

Original: The tagger is trained with the union of Wall Street Journal (WSJ) section of Penn Treebank (Marcus et al 1993), GENIA, and Penn BioIE. In WSJ, Sections 0-18 for training, 19-21 for development, and 22-24 for test. In GENIA and Penn BioIE, 90% of the corpus is used for training and the rest is used for test.

Combined: The tagger is trained with the union of the Original set plus *N* sentences from Fly.

Refdist: Tagger is trained with *N* sentences from Fly, plus the Original set as reference.

In Combined and Refdist settings, *N* is set to 8, 16, 32, 64, 128, 256, 512, 1024 sentences to measure the learning curve.

3 Results

The accuracies of the tagger trained in the Original setting were 96.4% on Fly, 96.7% on WSJ,

This work is partially supported by SORST program, Japan Science and Technology Agency.

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 136–137,
New York City, June 2006.

98.1% on GENIA and 97.7% on Penn BioIE corpora respectively. In the Combined setting, the accuracies were 97.9% on Fly, 96.7% on WSJ, 98.1% on GENIA and 97.7% on Penn BioIE. With Refdist setting, the accuracy on the Fly corpus was raised but those for WSJ and Penn BioIE corpora dropped from Original. When the width factor *w* was 10, the accuracy was 98.1% on Fly, but 95.4% on WSJ, 98.3% on GENIA and 96.6% on Penn BioIE. When the tagger was trained only on WSJ the accuracies were 88.7% on Fly, 96.9% on WSJ, 85.0% on GENIA and 86.0% on Penn BioIE. When the tagger was trained only on Fly, the accuracy on Fly was even lower (93.1%). The learning curve indicated that the accuracies on the Fly corpus were still rising in both Combined and Refdist settings, but both accuracies are almost as high as those of the original tagger on the original corpora (WSJ, GENIA and Penn BioIE), so in practical sense, 1024 sentences is a reasonable size for the additional corpus. When the width factor was smaller (2.5 and 5) the accuracies on the Fly corpus were saturated with N=1024 with lower values (97.8% with *w*=2.5 and 98.0% with *w*=5).

The amount of resources required for the Combined and the Refdist settings were drastically different. In the Combined setting, the learning time was 30632 seconds and the required memory size was 6.4GB. On the other hand, learning in the Refdist setting took only 21 seconds and the required memory size was 157 MB.

The most frequent confusions involved the confusion between FW (foreign words) with another class. Further investigation revealed that most of the error involved Linnaean names of species. Linnaean names are tagged differently in the GENIA and Penn BioIE corpora. In the GENIA corpus, tokens that constitute a Linnaean name are tagged as FW (foreign word) but in the Penn BioIE corpus they are tagged as NNP (proper noun). This seems to be one of the causes of the drop of accuracy on the Penn BioIE corpus when more sentences from the Fly corpus, whose tagging scheme follows that of GENIA, are added for training.

4 Conclusions

We compared two methods of adapting a POS tagger trained on corpora in human domain to fly domain. Training in Refdist setting required much smaller resources to fit to the target domain, but the resulting tagger is less portable to other domains. On the other hand, training in Combined setting is slower and requires huge memory, but the resulting tagger is more robust, and fits reasonably to various domains.

References

Tadayoshi Hara, Yusuke Miyao and Jun'ichi Tsujii. 2005. Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In Proceedings of IJCNLP 2005, LNAI 3651, pp. 199-210.

Mark Johnson and Stefan Riezler. 2000. Exploiting auxiliary distributions in stochastic unification-based grammars. In Proceedings of 1st NAACL.

Jun'ichi Kazama and Jun'ichi Tsujii. 2003. Evaluation and extension of maximum entropy models with inequality constraints. In Proceedings of EMNLP 2003.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus – a semantically annotated corpus for bio-textmining. Bioinformatics, 19(Suppl. 1):i180–i182.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Pete White. 2004. Integrated annotation for biomedical information extraction. In Proceedings of BioLINK 2004, pp. 61–68.

Matthew Lease and Eugene Charniak. 2005. Parsing Biomedical Literature, In Proceedings of IJCNLP 2005, LNAI 3651, pp. 58-69.

Mitchell P. Marcus, Beatrice Sanorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, Vol.19, pp. 313-330.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. In Proceedings of EMNLP 1996.

Yuka Tateisi and Jun'ichi Tsujii. (2004). Part-of-Speech Annotation of Biology Research Abstracts. In the Proceedings of LREC2004, vol. IV, pp. 1267-1270.

Kristina Toutanova, Dan Klein, Christopher Manning and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 173-180.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In Proceedings of 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392.