Self-Training for Biomedical Parsing

David McClosky and Eugene Charniak
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University
Providence, RI 02912
{dmcc|ec}@cs.brown.edu

Abstract

Parser self-training is the technique of taking an existing parser, parsing extra data and then creating a second parser by treating the extra data as further training data. Here we apply this technique to parser adaptation. In particular, we self-train the standard Charniak/Johnson Penn-Treebank parser using unlabeled biomedical abstracts. This achieves an f-score of 84.3% on a standard test set of biomedical abstracts from the Genia corpus. This is a 20% error reduction over the best previous result on biomedical data (80.2% on the same test set).

1 Introduction

Parser self-training is the technique of taking an existing parser, parsing extra data and then creating a second parser by treating the extra data as further training data. While for many years it was thought not to help state-of-the art parsers, more recent work has shown otherwise. In this paper we apply this technique to parser adaptation. In particular we self-train the standard Charniak/Johnson Penn-Treebank (C/J) parser using unannotated biomedical data. As is well known, biomedical data is hard on parsers because it is so far from more "standard" English. To our knowledge this is the first application of self-training where the gap between the training and self-training data is so large.

In section two, we look at previous work. In particular we note that there is, in fact, very little data on self-training when the corpora for self-training is so different from the original labeled data. Section three describes our main experiment on standard test data (Clegg and Shepherd, 2005). Section four looks at some preliminary results we obtained on development data that show in slightly more detail how self-training improved the parser. We conclude in section five.

2 Previous Work

While self-training has worked in several domains, the early results on self-training for parsing were negative (Steedman et al., 2003; Charniak, 1997). However more recent results have shown that it can indeed improve parser performance (Bacchiani et al., 2006; McClosky et al., 2006a; McClosky et al., 2006b).

One possible use for this technique is for parser adaptation — initially training the parser on one type of data for which hand-labeled trees are available (e.g., Wall Street Journal (M. Marcus et al., 1993)) and then self-training on a second type of data in order to adapt the parser to the second domain. Interestingly, there is little to no data showing that this actually works. Two previous papers would seem to address this issue: the work by Bacchiani et al. (2006) and McClosky et al. (2006b). However, in both cases the evidence is equivocal.

Bacchiani and Roark train the Roark parser (Roark, 2001) on trees from the Brown treebank and then self-train and test on data from Wall Street Journal. While they show some improvement (from 75.7% to 80.5% f-score) there are several aspects of this work which leave its re-

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 101–104,
Columbus, Ohio, USA, June 2008.

sults less than convincing as to the utility of self-training for adaptation. The first is the parsing results are quite poor by modern standards.[1] Steedman et al. (2003) generally found that self-training does not work, but found that it does help if the baseline results were sufficiently bad.

Secondly, the difference between the Brown corpus treebank and the Wall Street Journal corpus is not that great. One way to see this is to look at out-of-vocabulary statistics. The Brown corpus has an out-of-vocabulary rate of approximately 6% when given WSJ training as the lexicon. In contrast, the out-of-vocabulary rate of biomedical abstracts given the same lexicon is significantly higher at about 25% (Lease and Charniak, 2005). Thus the bridge the self-trained parser is asked to build is quite short.

This second point is emphasized by the second paper on self-training for adaptation (McClosky et al., 2006b). This paper is based on the C/J parser and thus its results are much more in line with modern expectations. In particular, it was able to achieve an f-score of 87% on Brown treebank test data when trained and self-trained on WSJ-like data. Note this last point. It was not the case that it used the self-training to bridge the corpora difference. It self-trained on NANC, *not* Brown. NANC is a news corpus, quite like WSJ data. Thus the point of that paper was that self-training a WSJ parser on similar data makes the parser more flexible, not better adapted to the target domain in particular. It said nothing about the task we address here. Thus our claim is that previous results are quite ambiguous on the issue of bridging corpora for parser adaptation.

Turning briefly to previous results on Medline data, the best comparative study of parsers is that of Clegg and Shepherd (2005), which evaluates several statistical parsers. Their best result was an f-score of 80.2%. This was on the Lease/Charniak (L/C) parser (Lease and Charniak, 2005).[2] A close second (1% behind) was the parser of Bikel (2004). The other parsers were not close. However, several very good current parsers were not available when this paper was written (e.g., the Berkeley Parser (Petrov et al., 2006)). However, since the newer parsers do not perform quite as well as the C/J parser on WSJ data, it is probably the case that they would not significantly alter the landscape.

3 Central Experimental Result

We used as the base parser the standardly available C/J parser. We then self-trained the parser on approximately 270,000 sentences — a random selection of abstracts from Medline.[3] Medline is a large database of abstracts and citations from a wide variety of biomedical literature. As we note in the next section, the number 270,000 was selected by observing performance on a development set.

We weighted the original WSJ hand annotated sentences equally with self-trained Medline data. So, for example, McClosky et al. (2006a) found that the data from the hand-annotated WSJ data should be considered at least five times more important than NANC data on an event by event level. We did no tuning to find out if there is some better weighting for our domain than one-to-one.

The resulting parser was tested on a test corpus of hand-parsed sentences from the Genia Treebank (Tateisi et al., 2005). These are exactly the same sentences as used in the comparisons of the last section. Genia is a corpus of abstracts from the Medline database selected from a search with the keywords Human, Blood Cells, and Transcription Factors. Thus the Genia treebank data are all from a small domain within Biology. As already noted, the Medline abstracts used for self-training were chosen randomly and thus span a large number of biomedical sub-domains.

The results, the central results of this paper, are shown in Figure 1. Clegg and Shepherd (2005) do not provide separate precision and recall numbers. However we can see that the

[1] This is not a criticism of the work. The results are completely in line with what one would expect given the base parser and the relatively small size of the Brown treebank.

[2] This is the standard Charniak parser (without reranker) modified to use an in-domain tagger.

[3] `http://www.ncbi.nlm.nih.gov/PubMed/`

System	Precision	Recall	f-score
L/C	—	—	80.2%
Self-trained	86.3%	82.4%	84.3%

Figure 1: Comparison of the Medline self-trained parser against the previous best

Medline self-trained parser achieves an f-score of 84.3%, which is an absolute reduction in error of 4.1%. This corresponds to an error rate reduction of 20% over the L/C baseline.

4 Discussion

Prior to the above experiment on the test data, we did several preliminary experiments on development data from the Genia Treebank. These results are summarized in Figure 2. Here we show the f-score for four versions of the parser as a function of number of self-training sentences. The dashed line on the bottom is the raw C/J parser with no self-training. At 80.4, it is clearly the worst of the lot. On the other hand, it is already better than the 80.2% best previous result for biomedical data. This is solely due to the introduction of the 50-best reranker which distinguishes the C/J parser from the preceding Charniak parser.

The almost flat line above it is the C/J parser with NANC self-training data. As mentioned previously, NANC is a news corpus, quite like the original WSJ data. At 81.4% it gives us a one percent improvement over the original WSJ parser.

The topmost line, is the C/J parser trained on Medline data. As can be seen, even just a thousand lines of Medline is already enough to drive our results to a new level and it continues to improve until about 150,000 sentences at which point performance is nearly flat. However, as 270,000 sentences is fractionally better than 150,000 sentences that is the number of self-training sentences we used for our results on the test set.

Lastly, the middle jagged line is for an interesting idea that failed to work. We mention it in the hope that others might be able to succeed where we have failed.

We reasoned that textbooks would be a particularly good bridging corpus. After all, they are written to introduce someone ignorant of a field to the ideas and terminology within it. Thus one might expect that the English of a Biology textbook would be intermediate between the more typical English of a news article and the specialized English native to the domain.

To test this we created a corpus of seven texts ("BioBooks") on various areas of biology that were available on the web. We observe in Figure 2 that for all quantities of self-training data one does better with Medline than BioBooks. For example, at 37,000 sentences the BioBook corpus is only able to achieve and an f-measure of 82.8% while the Medline corpus is at 83.4%. Furthermore, BioBooks levels off in performance while Medline has significant improvement left in it. Thus, while the hypothesis seems reasonable, we were unable to make it work.

5 Conclusion

We self-trained the standard C/J parser on 270,000 sentences of Medline abstracts. By doing so we achieved a 20% error reduction over the best previous result for biomedical parsing. In terms of the gap between the supervised data and the self-trained data, this is the largest that has been attempted.

Furthermore, the resulting parser is of interest in its own right, being as it is the most accurate biomedical parser yet developed. This parser is available on the web.[4]

Finally, there is no reason to believe that 84.3% is an upper bound on what can be achieved with current techniques. Lease and Charniak (2005) achieve their results using small amounts of hand-annotated biomedical part-of-speech-tagged data and also explore other possible sources or information. It is reasonable to assume that its use would result in further improvement.

Acknowledgments

This work was supported by DARPA GALE contract HR0011-06-2-0001. We would like to thank the BLLIP team for their comments.

[4] `http://bllip.cs.brown.edu/biomedical/`

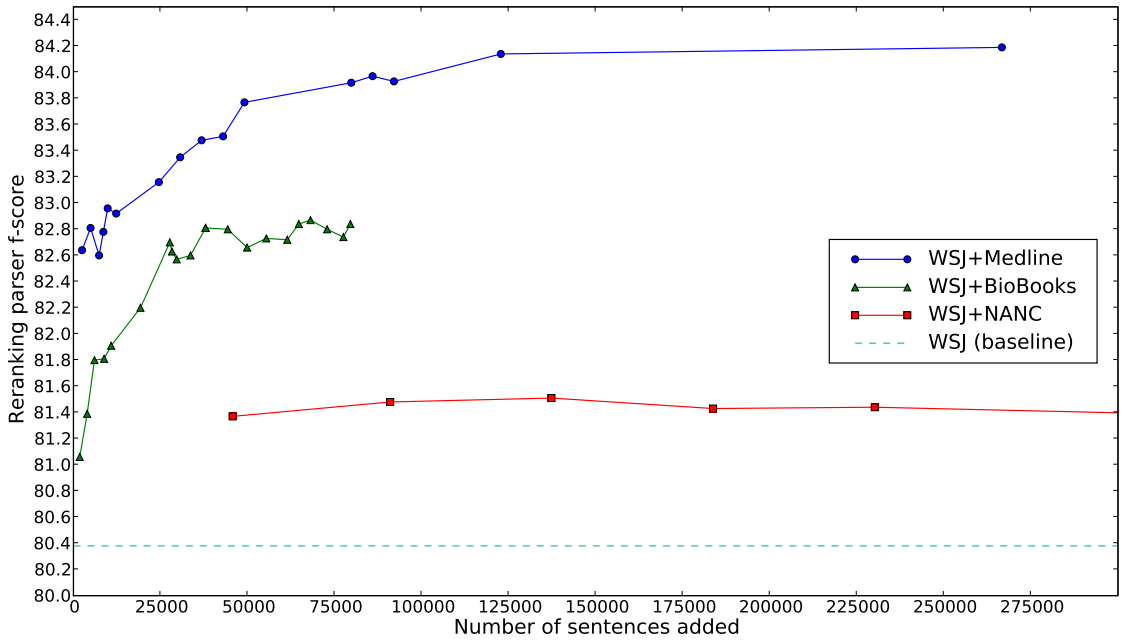

Figure 2: Labeled Precision-Recall results on development data for four versions of the parser as a function of number of self-training sentences

References

Michiel Bacchiani, Michael Riley, Brian Roark, and Richard Sproat. 2006. MAP adaptation of stochastic grammars. *Computer Speech and Language*, 20(1):41–68.

Daniel M. Bikel. 2004. Intricacies of collins parsing model. *Computational Linguistics*, 30(4).

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proc. AAAI*, pages 598–603.

Andrew B. Clegg and Adrian Shepherd. 2005. Evaluating and integrating treebank parsers on a biomedical corpus. In *Proceedings of the ACL Workshop on Software*.

Matthew Lease and Eugene Charniak. 2005. Parsing biomedical literature. In *Second International Joint Conference on Natural Language Processing (IJCNLP'05)*.

M. Marcus et al. 1993. Building a large annotated corpus of English: The Penn Treebank. *Comp. Linguistics*, 19(2):313–330.

David McClosky, Eugene Charniak, and Mark Johnson. 2006a. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159.

David McClosky, Eugene Charniak, and Mark Johnson. 2006b. Reranking and self-training for parser adaptation. In *Proceedings of COLING-ACL 2006*, pages 337–344, Sydney, Australia, July. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of COLING-ACL 2006*, pages 433–440, Sydney, Australia, July. Association for Computational Linguistics.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Bootstrapping statistical parsers from small datasets. In *Proc. of European ACL (EACL)*, pages 331–338.

Y. Tateisi, A. Yakushiji, T. Ohta, and J. Tsujii. 2005. Syntax Annotation for the GENIA corpus. *Proc. IJCNLP 2005, Companion volume*, pages 222–227.