Adapting WSJ-Trained Parsers to the British National Corpus Using In-Domain Self-Training

Jennifer Foster, Joachim Wagner, Djamé Seddah and Josef van Genabith
National Centre for Language Technology
School of Computing, Dublin City University, Dublin 9, Ireland
{jfoster, jwagner, josef}@computing.dcu.ie, dseddah@paris4.sorbonne.fr*

Abstract

We introduce a set of 1,000 gold standard parse trees for the British National Corpus (BNC) and perform a series of self-training experiments with Charniak and Johnson's reranking parser and BNC sentences. We show that retraining this parser with a combination of one million BNC parse trees (produced by the same parser) and the original WSJ training data yields improvements of 0.4% on WSJ Section 23 and 1.7% on the new BNC gold standard set.

1 Introduction

Given the success of statistical parsing models on the Wall Street Journal (WSJ) section of the Penn Treebank (PTB) (Charniak, 2000; Collins, 2003, for example), there has been a change in focus in recent years towards the problem of replicating this success on genres other than American financial news stories. The main challenge in solving the parser adaptation problem are the resources required to construct reliable annotated training examples.

A breakthrough has come in the form of research by McClosky et al. (2006a; 2006b) who show that self-training can be used to improve parser performance when combined with a two-stage reranking parser model (Charniak and Johnson, 2005). Self-training is the process of training a parser on its own output, and earlier self-training experiments using generative statistical parsers did not yield encouraging results (Steedman et al., 2003). McClosky et al. (2006a; 2006b) proceed as follows: sentences from the *LA Times* newspaper are parsed by a first-stage generative statistical parser trained on some seed training data (WSJ Sections 2-21) and the *n*-best parse trees produced by this parser are reranked by a discriminative reranker. The highest ranked parse trees are added to the training set of the parser and the parser is retrained. This self-training method gives improved performance, not only on Section 23 of the WSJ (an absolute f-score improvement of 0.8%), but also on test sentences from the Brown corpus (Francis and Kučera, 1979) (an absolute f-score improvement of 2.6%).

In the experiments of McClosky et al. (2006a; 2006b), the parse trees used for self-training come from the same domain (American newspaper text) as the parser's original seed training material. Bacchiani et al. (2006) find that self-training is effective when the parse trees used for self-training (WSJ parse trees) come from a different domain to the seed training data and from the same domain as the test data (WSJ sentences). They report a performance boost of 4.2% on WSJ Section 23 for a generative statistical parser trained on Brown seed data when it is self-trained using 200,000 WSJ parse trees. However, McCloskey et al. (2006b) report a drop in performance for their reranking parser when the experiment is repeated in the opposite direction, i.e. with Brown data for self-training and testing, and WSJ data for seed training. In contrast, we report successful in-domain[1] self-training experiments with the BNC data as self-training and test material, *and* with the WSJ-trained reranking parser used by McCloskey et al. (2006a; 2006b).

We parse the BNC (Burnard, 2000) in its entirety

*Now affiliated to Lalic, Université Paris 4 La Sorbonne.

[1] We refer to data as being *in-domain* if it comes from the same domain as the test data and *out-of-domain* if it does not.

Proceedings of the 10th Conference on Parsing Technologies, pages 33–35,
Prague, Czech Republic, June 2007.

using the reranking parser of Charniak and Johnson (2005). 1,000 BNC sentences are manually annotated for constituent structure, resulting in the first gold standard set for this corpus. The gold standard set is split into a development set of 500 parse trees and a test set of 500 parse trees and used in a series of self-training experiments: Charniak and Johnson's parser is retrained on combinations of WSJ treebank data and its own parses of BNC sentences. These combinations are tested on the BNC development set and Section 00 of the WSJ. An optimal combination is chosen which achieves a Parseval labelled bracketing f-score of 91.7% on Section 23 and 85.6% on the BNC gold standard test set. For Section 23 this is an absolute improvement of 0.4% on the baseline results of this parser, and for the BNC data this is a statistically significant improvement of 1.7%.

2 The BNC Data

The BNC is a 100-million-word balanced part-of-speech-tagged corpus of written and transcribed spoken English. Written text comprises 90% of the BNC: 75% non-fictional and 25% fictional. To facilitate parsing with a WSJ-trained parser, some reversible transformations were applied to the BNC data, e.g. British English spellings were converted to American English and neutral quotes disambiguated. The reranking parser of Charniak and Johnson (2005) was used to parse the BNC. 99.8% of the 6 million BNC sentences obtained a parse, with an average parsing speed of 1.4s per sentence.

A gold standard set of 1,000 BNC sentences was constructed by one annotator by correcting the output of the first stage of Charniak and Johnson's reranking parser. The sentences included in the gold standard were chosen at random from the BNC, subject to the condition that they contain a verb which does not occur in the training sections of the WSJ section of the PTB (Marcus et al., 1993). A decision was made to select sentences for the gold standard set which differ from the sentences in the WSJ training sections, and one way of finding different sentences is to focus on verbs which are not attested in the WSJ Sections 2-21. It is expected that these gold standard parse trees can be used as training data although they are used only as test and development data in this work. Because they contain verbs which do not occur in the parser's training set, they are likely to represent a hard test for WSJ-trained parsers. The PTB bracketing guidelines (Bies et al., 1995) and the PTB itself were used as references by the BNC annotator. Functional tags and traces were not annotated. The annotator noticed that the PTB parse trees sometimes violate the PTB bracketing guidelines, and in these cases, the annotator chose the analysis set out in the guidelines. It took approximately 60 hours to build the gold standard set.

3 Self-Training Experiments

Charniak and Johnson's reranking parser (June 2006 version) is evaluated against the BNC gold standard development set. Labelled precision (LP), recall (LR) and f-score measures[2] for this parser are shown in the first row of Table 1. The f-score of 83.7% is lower than the f-score of 85.2% reported by McClosky et al. (2006b) for the same parser on Brown corpus data. This difference is reasonable since there is greater domain variation between the WSJ and the BNC than between the WSJ and the Brown corpus, and all BNC gold standard sentences contain verbs not attested in WSJ Sections 2-21.

We retrain the first-stage generative statistical parser of Charniak and Johnson using combinations of BNC trees (parsed using the reranking parser) and WSJ treebank trees. We test the combinations on the BNC gold standard development set and on WSJ Section 00. Table 1 shows that parser accuracy increases with the size of the in-domain self-training material.[3] The figures confirm the claim of McClosky et al. (2006a) that *self-training* with a reranking parsing model is effective for improving parser accuracy in general, and the claim of Gildea (2001) that *training* on in-domain data is effective for parser adaption. They confirm that *self-training* on *in-domain* data is effective for parser adaptation. The WSJ Section 00 results suggest that, in order to maintain performance on the seed training domain, it is necessary to combine BNC parse trees

[2] All scores are for the second stage of the parsing process, i.e. the evaluation takes place after the reranking. All evaluation is carried out using the Parseval labelled bracketing metrics, with `evalb` and parameter file `new.prm`.

[3] The notation *bnc500K+5wsj* refers to a set of 500,000 parser output parse trees of sentences taken randomly from the BNC concatenated with five copies of WSJ Sections 2-21.

	BNC Development			WSJ Section 00		
Self-Training	LP	LR	LF	LP	LR	LF
-	83.6	83.7	83.7	91.6	90.5	91.0
bnc50k	83.7	83.7	83.7	90.0	88.0	89.0
bnc50k+1wsj	84.4	84.4	84.4	91.6	90.3	91.0
bnc250k	84.7	84.5	84.6	91.1	89.3	90.2
bnc250k+5wsj	85.0	84.9	85.0	91.8	90.5	91.2
bnc500k+5wsj	85.2	85.1	85.2	91.9	90.4	91.2
bnc500k+10wsj	85.1	85.1	85.1	91.9	90.6	91.2
bnc1000k+5wsj	86.5	86.2	86.3	91.7	90.3	91.0
bnc1000k+10wsj	86.1	85.9	**86.0**	92.0	90.5	**91.3**
bnc1000k+40wsj	85.5	85.5	85.5	91.9	90.6	91.3
	BNC Test			**WSJ Section 23**		
-	84.0	83.7	83.9	91.8	90.9	91.3
bnc1000k+10wsj	85.7	85.4	**85.6**	92.3	91.1	**91.7**

Table 1: In-domain Self-Training Results

with the original seed training material during the self-training phase.

Of the self-training combinations with above-baseline improvements for both development sets, the combination of 1,000K BNC parse trees and Section 2-21 of the WSJ (multiplied by ten) yields the highest improvement for the BNC data, and we present final results with this combination for the BNC gold standard test set and WSJ Section 23. There is an absolute improvement on the original reranking parser of 1.7% on the BNC gold standard test set and 0.4% on WSJ Section 23. The improvement on BNC data is statistically significant for both precision and recall ($p < 0.0002$, $p < 0.0002$). The improvement on WSJ Section 23 is statistically significant for precision only ($p < 0.003$).

4 Conclusion and Future Work

We have introduced a set of 1,000 gold standard parse trees for the BNC. We have performed self-training experiments with Charniak and Johnson's reranking parser and sentences from the BNC. We have shown that retraining this parser with a combination of one million BNC parse trees (produced by the same parser) and the original WSJ training data yields improvements of 0.4% on WSJ Section 23 and 1.7% on the BNC gold standard sentences. These results indicate that self-training on in-domain data can be used for parser adaptation.

Our BNC gold standard set consists of sentences containing verbs which are not in the WSJ training sections. We suspect that this makes the gold standard set a hard test for WSJ-trained parsers, and our results are likely to represent a lower bound for WSJ-trained parsers on BNC data. When used as training data, we predict that the novel verbs in the BNC gold standard set add to the variety of training material, and will further help parser adaptation from the WSJ domain – a matter for further research.

Acknowledgments We thank the IRCSET Embark Initiative (basic research grant SC/02/298 and postdoctoral fellowship P/04/232), Science Foundation Ireland (Principal Investigator grant 04/IN.3/I527) and the Irish Centre for High End Computing for supporting this research.

References

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

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for treebank II style, Penn Treebank project. Technical Report MS-CIS-95-06, University of Pennsylvania.

Lou Burnard. 2000. User reference guide for the British National Corpus. Technical report, Oxford University.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best-parsing and maxent discriminative reranking. In *Proceedings of ACL-05*, pages 173–180, Barcelona.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL-00*, pages 132–139, Seattle.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):499–637.

W. Nelson Francis and Henry Kučera. 1979. Brown Corpus Manual. Technical report, Brown University, Providence.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of EMNLP-01*, pages 167–202, Barcelona.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

David McClosky, Eugene Charniak, and Mark Johnson. 2006a. Effective self-training for parsing. In *Proceedings of HLT-NAACL-06*, pages 152–159, New York.

David McClosky, Eugene Charniak, and Mark Johnson. 2006b. Reranking and self-training for parser adaptation. In *Proceedings of COLING-ACL-06*, pages 337–344, Sydney.

Mark Steedman, Miles Osbourne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Boot-strapping statistical parsers from small datasets. In *Proceedings of EACL-03*, Budapest.