Self-Training without Reranking for Parser Domain Adaptation and Its Impact on Semantic Role Labeling

Kenji Sagae
Institute for Creative Technologies
University of Southern California
Marina del Rey, CA 90292
sagae@ict.usc.edu

Abstract

We compare self-training with and without reranking for parser domain adaptation, and examine the impact of syntactic parser adaptation on a semantic role labeling system. Although self-training without reranking has been found not to improve in-domain accuracy for parsers trained on the WSJ Penn Treebank, we show that it is surprisingly effective for parser domain adaptation. We also show that simple self-training of a syntactic parser improves out-of-domain accuracy of a semantic role labeler.

1 Introduction

Improvements in data-driven parsing approaches, coupled with the development of treebanks that serve as training data, have resulted in accurate parsers for several languages. However, portability across domains remains a challenge: parsers trained using a treebank for a specific domain generally perform comparatively poorly in other domains. In English, the most widely used training set for parsers comes from the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993), and constituent parsers trained on this set are now capable of labeled bracketing precision and recall of over 90% (Charniak and Johnson, 2005; Huang, 2008) on WSJ testing sentences. When applied without adaptation to the Brown portion of the Penn Treebank, however, an absolute drop of over 5% in precision and recall is typically observed (McClosky et al., 2006b). In pipelined NLP applications that include a parser, this drop often results in severely degraded results downstream.

We present experiments with a simple self-training approach to semi-supervised parser domain adaptation that produce results that contradict the commonly held assumption that improved parser accuracy cannot be obtained by self-training a generative parser without reranking (Charniak, 1997; Steedman et al., 2003; McClosky et al., 2006b, 2008).[1] We compare this simple self-training approach to the self-training with reranking approach proposed by McClosky et al. (2006b), and show that although McClosky et al.'s approach produces better labeled bracketing precision and recall on out-of-domain sentences, higher F-score on syntactic parses may not lead to an overall improvement in results obtained in NLP applications that include parsing, contrary to our expectations. This is evidenced by results obtained when different adaptation approaches are applied to a parser that serves as a component in a semantic role labeling (SRL) system. This is, to our knowledge, the first attempt to quantify the benefits of semi-supervised parser domain adaptation in semantic role labeling, a task in which parsing accuracy is crucial.

2 Semi-supervised parser domain adaptation with self-training

Because treebanks are expensive to create, while plain text in most domains is easily obtainable, semi-supervised approaches to parser domain adaptation are a particularly attractive solution to the domain portability problem. This usually involves a manually annotated training set (a

[1] Reichart and Rappoport (2007) show that self-training without reranking is effective when the manually annotated training set is small. We show that this is true even for a large training set (the standard WSJ Penn Treebank training set, with over 40k sentences).

Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010, pages 37–44,
Uppsala, Sweden, 15 July 2010.

treebank), and a larger set of unlabeled data (plain text).

Bacchiani and Roark (2003) obtained positive results in unsupervised domain adaptation of language models by using a speech recognition system with an out-of-domain language model to produce an automatically annotated training corpus that is used to adapt the language model using a maximum *a posteriori* (MAP) adaptation strategy. In subsequent work (Roark and Bacchiani, 2003), this MAP adaptation approach was applied to PCFG adaptation, where an out-of-domain parser was used to annotate an in-domain corpus automatically with multiple candidate trees per sentence. A substantial improvement was achieved in out-of-domain parsing, although the obtained accuracy level was still far below that obtained with domain-specific training data.

More recent work in unsupervised domain adaptation for state-of-the-art parsers has achieved accuracy levels on out-of-domain text that is comparable to that achieved with domain-specific training data (McClosky et al., 2006b). This is done in a self-training setting, where a parser trained on a treebank (in a seed domain) is used to parse a large amount of unlabeled data in the target domain (assigning only one parse per sentence). The automatically parsed corpus is then used as additional training data for the parser. Although initial attempts to improve in-domain parsing accuracy with self-training were unsuccessful (Charniak, 1997; Steedman et al., 2003), recent work has shown that self-training can work in specific conditions (McClosky et al., 2006b), and in particular it can be used to improve parsing accuracy on out-of-domain text (Reichart and Rappoport, 2007).

2.1 Self-training with reraking

McClosky et al. (2006b) presented the most successful semi-supervised approach to date for adaptation of a WSJ-trained parser to Brown data containing several genres of text (such as religion, mystery, romance, adventure, etc.), obtaining a substantial accuracy improvement using only unlabeled data. Their approach involves the use of a first-stage n-best parser and a reranker, which together produce parses for the unlabeled dataset. The automatically parsed in-domain corpus is then used as additional training material. In light of previous failed attempts to improve generative parsers through self-training (Charniak, 1997; Steedman et al., 2003), McClosky et al. (2006a) argue that the use of a reranker is an important factor in the success of their approach. That work used text from the LA Times (taken from the North American News Corpus, or NANC), which is presumably more similar to the parser's training material than to text in the Brown corpus, and resulted not only in an improvement of parser accuracy on out-of-domain text (from the Brown corpus), but also in an improvement in accuracy on in-domain text (the standard WSJ test set of the Penn Treebank).

It can be argued that the McClosky et al. approach is not a pure instance of self-training, since two parsing models are used: the first-stage generative model, and a discriminative model for reranking. The generative parser is improved based on the output of the discriminative model, but McClosky et al. found that the discriminative model does not improve when retrained with its own output.

2.2 Self-training without reraking

Although there have been instances of self-training (or similar) approaches that produced improved parser accuracy without reranking, the success of these efforts are often attributed to other specific factors.

Reichart and Rappoport (2007) obtained positive results in in-domain and out-of-domain scenarios with self-training without reranking, but under the constant condition that only a relatively small set of manually labeled data is used as the seed training set. Sagae and Tsujii (2007) improved the out-of-domain accuracy of a dependency parser trained on the entire WSJ training set (40k sentences) by using unlabeled data in the same domain as the out-of-domain test data (biomedical text). However, they used agreement between different parsers to estimate the quality of automatically generated training instances and selected only sentences with high estimated accuracy. Although the parser improves when trained with its own output, the training instances are selected through the use of a separate dependency parsing model.

2.3 Simple self-training without reranking for domain adaptation

It is now commonly assumed that the simplest form of self-training, where a single parsing model is retrained with its own output (a single parse tree per sentence, without reranking or other means of training instance selection or estimation of parse quality), does not improve the model's accuracy.[2] This assumption, however, is largely based on previous attempts to improve *in-domain* accuracy through self-training (Steedman et al., 2003; Charniak, 1997; McClosky et al., 2006a, 2008). We will refer to this type of self-training as *simple self-training*, to avoid confusion with other self-training settings, such as McClosky et al.'s, where a reranker is involved.

We propose a simple self-training framework for domain adaptation, as follows:

1. A generative parser is trained using a treebank in a specific source domain.

2. The parser is used to generate parse trees from text in a target domain, different from the source domain.

3. The parser is retrained using the original treebank, augmented with the parse trees generated in step 2.

There are intuitive reasons that may lead one to assume that simple self-training *should not* work. One is that no additional information is provided to the model. In self-training with reranking, the generative model can be enriched with information produced by the discriminative model. When two parsers are used for training instance selection, one parser informs the other. In simple self-training, however, there is no additional source of syntactic knowledge with which the self-trained model would be enriched.

Another possible reason is that the output of the self-trained parser should be expected to include the same errors found in the automatically generated training material. If the initial parser has poor accuracy on the target domain, the training data it generates will be of poor quality, resulting in no improvement in the resulting trained model. The self-trained model may simply learn to make the same mistakes as the original model.

Conversely, there are also intuitive reasons for why it *might* work. A possible source of poor performance in new domains is that the model lacks coverage. Specific lexical items and syntactic structures in a new domain appear in a variety of contexts, accompanied by different words and structures. The parser trained on the source domain may analyze some of these new items and structures correctly, and it may also make mistakes. As long as errors in the automatically generated training material are not all systematic, the benefits of adding target-domain information could outweigh the addition of noise in the model.

Naturally, it may be that these conditions hold for some pairs of source and target domains but not others. In the next section, we present experiments that investigate whether simple self-training is effective for one particular set of training (WSJ) and testing (Brown) corpora, which are widely used in parsing research for English.

3 Domain adaptations experiments

In our experiments we use primarily the Charniak (2000) parser. In a few specific experiments we also use the Charniak and Johnson (2005) reranker; such cases are noted explicitly and are not central to the paper, serving mostly for comparisons. We follow the three steps described in section 2.3. The manually labeled training corpus is the standard WSJ training sections of the Penn Treebank (sections 02 to 21). Sections 22 and 23 are used as in-domain development and testing sets, respectively. The out-of-domain material is taken from the Brown portion of the Penn Treebank. We use the same Brown test set as McClosky et al. (2006b), every tenth sentence in the corpus. Another tenth of the corpus is used as a development set, and the rest of the Brown corpus is not used. The out-of-domain text then contains not one but several genres of text. The larger set of unlabeled data is composed of approximately 5.3 million words (320k sentences) of 20th century novels available from Project Gutenberg[3], which do not match exactly the target domain, but is closer to it in general than to the source domain (WSJ).

3.1 Simple self-training results

The precision, recall and F-score of labeled brackets of the initial parser, trained only on the WSJ Penn Treebank, are shown in the first row of results in Table 1 for the WSJ (in-domain) test set and the Brown (out-of-domain) test set. These figures serve as our baseline. The second row of results in Table 1 shows the results obtained with a model produced using simple self-training. The baseline model is used to parse the entire unlabeled dataset (320k sentences), and

[2] Except for in cases where the initial model is trained using a very small treebank.

[3] http://www.gutenberg.org

	WSJ			Brown		
	Precision	Recall	F-score	Precision	Recall	F-score
Baseline	89.49	88.78	89.13	83.93	83.19	83.56
Self-trained	88.26	87.86	88.06	**85.78**	**85.05**	**85.42**
MCJ			91.0			87.1

Table 1. Labeled constituent precision, recall and F-score for the WSJ and Brown test sets, obtained with the baseline model (trained only on the WSJ training set) and with the self-trained model. Results on Brown show an absolute improvement of almost 2%, while results on WSJ show a drop of about 1%. The last row shows the results obtained by McClosky et al. (2006a, 2006b) using self-training with reranking (denoted as MCJ), for comparison purposes.

the resulting parse trees are added to the WSJ training set to produce the self-trained model.

A substantial improvement is observed for the target test set (Brown), close to an absolute improvement of 2% in precision, recall and F-score. Table 1 also shows that parser accuracy fell by 1% on WSJ. Although we do not see this as a problem, since the our goal is to produce an improved model for parsing Brown, it is interesting that, unlike in the work of McClosky et al. (2006a, 2006b) where self-training includes reranking, simple self-training is effective specifically for domain adaptation, but not for improving the accuracy of the parser on in-domain data. At least in this case, simple self-training does not result in an absolutely improved parsing model (as appears to be the case with McClosky et al.'s self-training), although it *does* result in an improved model for the target data.

Finally, the last row in Table 1 shows the results on WSJ and Brown obtained by McClosky et al. (2006a, 2006b) using self-training with reranking. As they have shown, the discriminative reranker can be used to provide further improvements, as discussed in the next subsection.

Unlike McClosky et al. (2006a), we did not give different weights to the original and automatically generated training instances. In our experiments with the Brown development data, varying the weight of the gold-standard WSJ training data from 1 to 7, we observed only small differences in F-score (Table 2). The highest F-score, obtained when the WSJ training corpus is given a relative weight of 3, was only 0.07 higher than the F-score obtained when the WSJ training corpus is given a relative weight of 1.

WSJ relative weight	Brown dev F-score
1	84.51
2	84.52
3	84.58
4	84.53
5	84.51
6	84.55
7	84.57
Baseline (WSJ only)	82.91

Table 2: Brown development set F-scores obtained with self-trained models with different relative weights given to the gold-standard WSJ training data. The last row shows the F-score for the original model (without adaptation).

Table 3 shows results on the Brown development set when different amounts of unlabeled data are used to create the self-trained model. Although F-score generally increases with more unlabeled data, the effect is not monotonic. McClosky et al. observed a similar effect in their self-training experiments, and hypothesized that this may be due to differences between portions of the unlabeled data and the target corpus, and to varying parsing difficulty in portions of the unlabeled data, which results in varying quality of the parse trees produced automatically for training. A large improvement in F-score over the baseline is observed when adding only 30k sentences. Additional improvement is observed when additional sentences are added, but these are small in comparison. One interesting note is

Sentences added	Brown dev. F-score
0 (baseline)	82.91
10k	83.76
20k	84.02
30k	84.29
50k	84.26
100k	84.19
150k	84.38
200k	84.51
250k	84.42
300k	84.51

Table 3: Brown development set F-scores obtained with self-trained models created with different amounts of unlabeled data.

that, although self-training produced improved bracketing precision and recall, part-of-speech tagging accuracy of Brown remained largely unchanged from the baseline, in the range of 94.42% to 94.50% accuracy. It is possible that separate adaption for part-of-speech tagging may improve parsing F-score further.

The results in this section show that simple self-training is effective in adapting WSJ-trained parser to Brown, but more experiments are needed to determine if the same effects observed in our simple self-training experiments would also be observed with other pairs of seed training data and target datasets, and what characteristics of the datasets may affect domain adaptation.

3.2 Self-training with reranking results

To provide a more informative comparison between the results obtained with simple self-training and other work, we also performed McClosky et al.'s self-training with reranking using our unlabeled dataset. In this experiment, intended to provide a better understanding of the role of the unlabeled data (20th century novels vs. LA Times articles), we parse the unlabeled dataset with the Charniak (2000) parser and the Charniak and Johnson (2005) discriminative reranker to produce additional training material for the generative parser. The resulting generative parser produces slightly improved F-scores compared to the simple self-training setting (88.78% on WSJ and 86.01 on Brown), although a slight drop in WSJ F-score is still observed, indicating that the use of news text is likely an important factor in McClosky et al.'s superior F-score figures.

All of these models can be used to produce n-best parses with the Charniak parser, and these can be reranked with the Charniak and Johnson reranker, whether or not the self-training procedure that created the generative model involved reranking. McClosky et al. found that although their self-training procedure involves reranking, the gains in accuracy are orthogonal to those provided by a final reranking step, applied to the output of the self-trained model. As in their case, applying the WSJ-trained reranker to our self-trained model improves its accuracy. In the case of our simple self-trained model, the improvement is of about 1.7%, which means that if a reranker is used at run-time (but not during self-training), F-score goes up to 87.12%. Interestingly, applying a final pass of reranking to the model obtained with self-training with reranking brings F-score up only by less than 1.2%, to 87.17%. So at least in our case, improvements provided by the use the reranker appear not to be completely orthogonal.

4 Semantic Role Labeling with syntactic parser adaptation

To investigate the impact of parser domain adaptation through self-training on applications that depend on parser output, we use an existing semantic role labeling (SRL) system, the Illinois Semantic Role Labeler[4], replacing the provided parsing component with our (WSJ) baseline and (adapted) self-trained parsers.

We tested the SRL system using the datasets of the CoNLL 2005 shared task (Carreras and Màrquez, 2005). The system is trained on the WSJ domain using PropBank (Palmer et al. 2005), and the shared task includes WSJ and Brown evaluation sets. Using the baseline WSJ syntactic parser, the SRL system has an F-score of 77.49 on WSJ, which is a competitive result for systems using a single syntactic analysis per sentence. The highest scoring system (also a UIUC system) in the shared task has 79.44 F-score, and used multiple parse trees, which has been shown to improve results (Punyakanok et al., 2005). On the Brown evaluation, F-score is 64.75, a steep drop from the performance of the system on WSJ, which reflects that not just the syntactic parser, but also other system components, were trained with WSJ material. The

[4] http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=SRL

	Precision	Recall	F-score
Baseline (WSJ parser)	66.57	63.02	64.75
Simple self-trained parser (this paper)	**71.66**	**66.10**	**68.77**
MCJ self-trained parser	69.18	65.37	67.22
MCJ self-train and rerank	68.62	65.78	67.17

Table 4. Semantic role labeling results using the Illinois Semantic Role Labeler (trained on WSJ material from PropBank) using four different parsing models: (1) a model trained on WSJ, (2) a model built from the WSJ training data and 320k sentences from novels as unlabeled data, using the simple self-training procedure described in sections 2.3 and 3.1, (3) the McClosky et al. (2006a) self-trained model, and (4) the McClosky et al. self-trained model, reranked with the Charniak and Johnson (2005) reranker.

highest scoring system on the Brown evaluation in the CoNLL 2005 shared task had 67.75 F-score.

Table 4 shows the results on the Brown evaluation set using the baseline WSJ SRL system and the results obtained under three self-training parser domain adaptation schemes: simple self-training using novels as unlabeled data (section 3.1), the self-trained model of McClosky et al.[5], and the reranked results of the McClosky et al. self-trained model (which has F-score comparable to that of a parser trained on the Brown corpus).

As expected, the contributions of the three adapted parsing models allowed the system to produce overall SRL results that are better than those produced with the baseline setting. Surprisingly, however, the use of the model created using simple self-training and sentences from novels (sections 2.3 and 3.1) resulted in better SRL results than the use of McClosky et al.'s reranking-based self-trained model (whether its results go through one additional step of reranking or not), which produces substantially higher syntactic parsing F-score. Our self-trained parsing model results in an absolute increase of 4% in SRL F-score, outscoring all participants in the shared task (of course, systems in the shared task did not use adapted parsing models or external resources, such as unlabeled data). The improvement in the precision of the SRL system using simple self-training is particularly large. Improvements in the precision of the core arguments Arg0, Arg1, Arg2 contributed heavily to the improvement of overall scores.

We note that other parts of the SRL system remained constant, and the difference in the results shown in Table 4 come solely from the use of different (adapted) parsers.

5 Conclusion

We explored the use of simple self-training, where no reranking or confidence measurements are used, for parser domain adaptation. We found that self-training can in fact improve the accuracy of a parser in a different domain from the domain of its training data (even when the training data is the entire standard WSJ training material from the Penn Treebank), and that this improvement can be carried on to modules that may use the output of the parser. We demonstrated that a semantic role labeling system trained with WSJ training data can improve substantially (4%) on Brown just by having its parser be adapted using unlabeled data.

Although the fact that self-training produces improved parsing results without reranking does not necessarily conflict with previous work, it does contradict the widely held assumption that this type of self-training does not improve parser accuracy. One way to reconcile expectations based on previous attempts to improve parsing accuracy with self-training (Charniak, 1997;

[5] http://www.cs.brown.edu/~dmcc/selftraining.html

Steedman et al., 2003) and the results observed in our experiments is that we focus specifically on domain adaptation. In fact, the in-domain accuracy of our adapted model is slightly inferior to that of the baseline, more in line with previous findings.

This work represents only one additional step towards understanding of how and when self-training works for parsing and for domain adaptation. Additional analysis and experiments are needed to understand under what conditions and in what domains simple self-training can be effective.

One question that seems particularly interesting is why the models adapted using self-training with reranking and news text, which produce substantially higher parsing F-scores, did not outperform our model built with simple self-training in contribution to the SRL system. Although we do not have an answer to this question, two factors that may play a role are the domain of the training data and the use of the reranker, which may provide improvements in parse quality that are of a different kind of those most needed by the SRL system. This points to another interesting direction, where adapted parsers can be combined. Having different ways to perform semi-supervised parser adaptation may result in the creation of adapted models with improved accuracy on a target domain but different characteristics. The output of these parsers could then be combined in a voting scheme (Henderson and Brill, 1999) for additional improvements on the target domain.

Acknowledgments

We thank Andrew S. Gordon and the anonymous reviewers for valuable comments and suggestions. The work described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Michiel Bacchiani and Brian Roark. 2003. Unsupervised language model adaptation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the CoNLL 2005 shared task.*

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

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL).* Pages 132-139. Seattle, WA.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 2005 Meeting of the Association for Computational Linguistics (ACL)*, pages 173–180.

John C. Henderson, Eric Brill. 1999. Exploiting Diversity in Natural Language Processing: Combining Parsers. In *Proceedings of the Fourth Conference on Empirical Methods in Natural Language Processing (EMNLP-99),* pp. 187–194. College Park, Maryland.

Liang Huang (2008). Forest Reranking: Discriminative Parsing with Non-Local Features. In *Proceedings of the 2008 Meeting of the Association for Computational Linguistics (ACL)*. Columbus, OH.

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

David McClosky, Eugene Charniak and Mark Johnson. 2006a. Effective self-training for parsing. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL).* New York, NY.

David McClosky, Eugene Charniak and Mark Johnson. 2006b. Reranking and self-training for parser adaptation. In *Proceedings of the 21st international Conference on Computational Linguistics and the 44th Annual Meeting of the Association For Computational Linguistics (ACL).* Sydney, Australia.

David McClosky, Eugene Charniak and Mark Johnson. 2008. When is self-training effective for parsing? In *Proceedings of the 22nd international Conference on Computational Linguistics (COLING) - Volume 1.* Manchester, United Kingdom.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Vasin Punyakanok, Peter Koomen, Dan Roth and Wen-tau Yih. 2005. Generalized Inference with Multiple Semantic Role Labeling Systems. In *Proceedings of the CoNLL 2005 shared task.*

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL).* Pages 616-623. Prague, Czech Republic.

Brian Roark and Michiel Bacchiani. 2003. Supervised and unsupervised PCFG adaptation to novel domains. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – Volume 1 (NAACL-HLT).*

Kenji Sagae and Jun'ichi Tsujii. 2007. Multilingual dependency parsing and domain adaptation with data-driven LR models and parser ensembles. In *Proceedings of the CoNLL 2007 shared task*. Prague, Czech Republic.

Mark Steedman, Rebecca Hwa, Stephen Clark, Miles Osborne, Anoop Sarkar, Julia Hockenmaier, Paul Ruhlen, Steven Baker and Jeremiah Crim. 2003. Bootstrapping statistical parsers from small datasets. In *Proceedings of Tenth Conference of the European Chapter of the Association for Computational Linguistics (EACL) – Volume 1.* Budapest, Hungary.