Statistical Parsing with Domain Adaptation

Tyler McDonnell

I. Introduction

Though a wide body of literature has shown the effectiveness of supervised statistical parsers, learned statistical parsers can often be very specific to the genre or nuances of their training corpus. This is not ideal: labeled data is expensive to produce, so ideally we would like to be able to learn from the limited amounts available and generalize the learned knowledge across a wide body of domains. Here, I explore a previously proposed semi-supervised approach for domain adaptation, in which we attempt to transfer the expertise of a learned parser from one genre to another.

II. Methodology

For this experiment, I use a domain adaptation approach based on self-training, a form of semi-supervised learning in which we first train a statistical model using labeled data in the source domain, use this model to produce labeled output for an unlabeled set of data in the target domain (the self-training data), and then train a final statistical model using both the labeled source data and labeled self-training data. This approach is essentially single-iteration hard EM and was first used effectively for the problem of transfer learning in statistical parsing by Reichart and Rappoport [3].

Note that this methodology is not specific to any one problem space in natural language processing, and may be more broadly applicable to many problem spaces in machine learning. Here, I explore its utility specifically for the problem of statistical parsing, the process of predicting the the grammatical structure of sentences.

III. Experimental Setup

I developed my parser on top of the Stanford Statistical Parser [1], a highly optimized parser which combines an unlexicalized PCFG parser, which makes use of "linguistically motivated state splits" to refine notions of independence in the traditional PCFG model, with a lexicalized dependency parser [2]. In particular, for each of the experiments that follow, I trained a PCFG-based *LexicalizedParser* on a seed corpus; used the resultant parser to label a self-training corpus; and trained a final domain-adapted *LexicalizedParser* using the combined seed and self-training corpuses. I then evaluated the final model on the specified test set.

For experimentation, I used the classic Brown and WSJ datasets and investigated domain adaptation from WSJ to Brown and Brown to WSJ. For the former, I used WSJ Sections 2-22 as seed data; 90% of eight *genres* from the Brown corpus as the self-training data; and the remaining 10% of each genre as the out-of-domain test set. For the latter, I used the inverse: 90% of each of the genres of the Brown corpus was used as seed data; WSJ Sections 2-22 were used as self-training data; and WSJ Section 23 was used as test data.

IV. Results

Figure 1 and Figure 2 show the results of my experiments for each of the domain adaptations. For each figure, the left graph shows the F1 score for my parser for various seed sizes, while the right shows the F1 score for various sizes of self-training with a static seed set of 10,000 sentences from the source domain. The left graph also includes baseline in-domain and out-of-domain numbers for different training set sizes for comparison. For example, in Figure 1, the In-Domain curve represents the learning curve of the standard Stanford *LexicalizedParser* for different training sizes from WSJ Sections 2-22, evaluated on WSJ Section 23. Similarly, the Out-of-Domain curve for this figure represents the same trained parser evaluated on the Brown test corpus.

The remainder of this paper discusses these results and compares them to the original paper in which Reichart and Rappoport's applied this self-training method [3].

In-Domain vs. Out-of-Domain

F1 score drops remarkably from in-domain to out-of-domain testing. On average, the difference in F1 between in-domain and out-of-domain testing is about 8 points, but it can reach as high as 15 points. This should not be surprising, since a wide body of empirical evidence has shown that the performance of learned parsers can be very specific to the *genre*, or nuances, of the the corpus they were trained on. Indeed, this reality is the primary motivation behind domain adaptation!

Domain Adaptation Performance

Unsupervised domain adaptation improved out-of-domain performance for both experiments: WSJ $\rightarrow$ Brown and Brown $\rightarrow$ WSJ. This is both encouraging and expected, since domain adaptation methodologies were developed for this purpose. Minimally, we also expect gains since we are essentially doing semi-supervised EM. Even if the self-training set fails to capture any of the genre-specific aspects of the target domain, we would still expect some improvement from the increases in training data.

In general though, the self-trained out-of-domain performance was closer to the out-of-domain performance than it was to the in-domain performance, which suggests room for improvement. Unsupervised domain adaptation offered average gains of about 2 points in F1 score for WSJ $\rightarrow$ Brown and 4 points in the inverse.I now look at a few variables more carefully with regard to unsupervised domain adaptation performance.

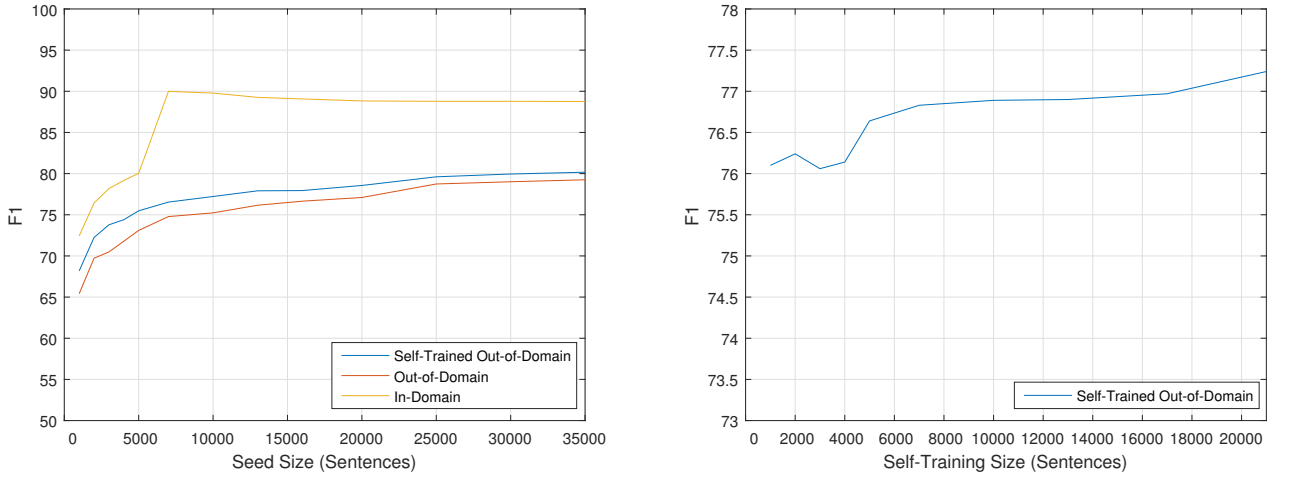

Fig. 1: WSJ → Brown Domain Adaptation

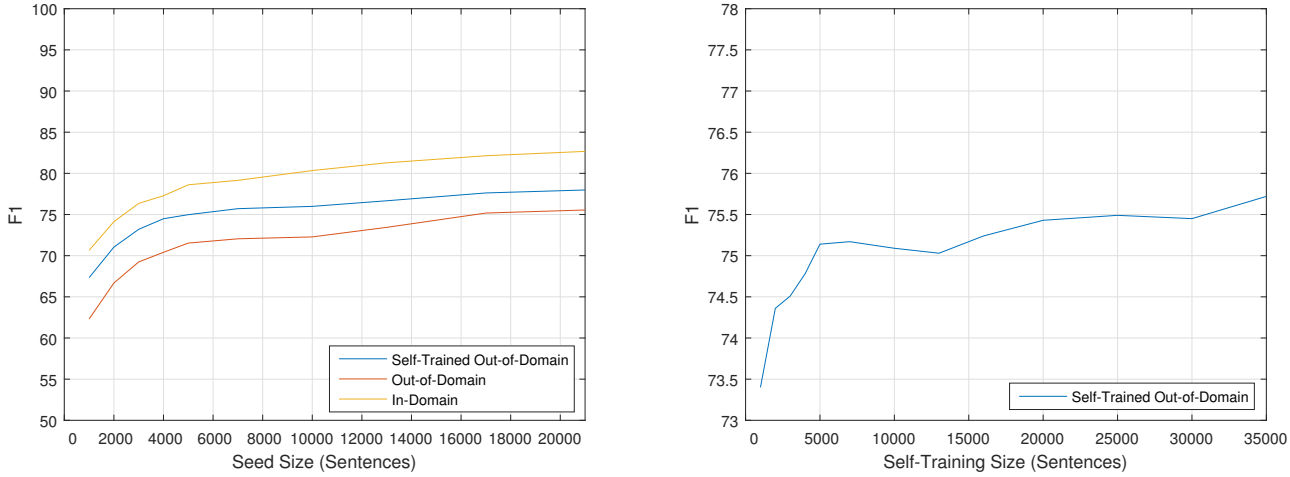

Fig. 2: Brown → WSJ Domain Adaptation

Effect of Seed Size

Domain adaptation yields higher performance benefits over out-of-domain testing at lower source seed sizes than higher seed sizes. This could either be interpreted as not very surprising or very telling. On the one hand, this should be expected, since we generally have a lot of statistical knowledge to gain by vastly increasing the amount of labeled data, even if the labels are imperfect, at small seed sizes. Again, this insight is related to the motivation of EM in the general semi-supervised case: we have some labeled data, but perhaps not enough to capture the characteristics of the domain, so we can use our labeled data to make a larger body of imperfectly labeled data. At the same time, this might suggest that most of the benefit of unsupervised domain adaptation is actually a result of the increased size of training data exclusively, rather than training data that captures the specific characteristics of the genre of the target domain.

Effect of Self-Training Size

The right graphs of Figure 1 and Figure 2 show that increasing the size of the self-training data generally increases the out-of-domain performance, but the gains are modest. For a static seed size, varying the size of the self-training data between one-tenth and two or more times the size of the seed data yielded F1 performance gains of only 0 to 2.5 points. Additionally, though the general trend is that a larger self-training size tends to yield slightly higher F1 scores, there were several data points where the performance dropped slightly.

In some sense, these results might further validate the hypothesis from the previous section: gains from unsupervised domain adaptation are primarily a result of the increased size of training data, rather than the advertised ability to capture the genre of the target domain. For this comparison, I used a static seed size of 10,000 sentences, which is rather large compared to the seed sizes used in the by Reichart and Rappoport [3]. Additionally, I previously noted that unsupervised domain adaptation yields higher performance gains over simple out-

of-domain testing at lower seed sizes rather than higher seed sizes. Indeed, a few tests boundary tests I carried out with self-training sizes of 1,000 and 20,000 on the WSJ $\rightarrow$ Brown domain showed that the gains from self-training are greater with smaller seed sets.

Effect of Source and Target Domain

In-Domain, Out-of-Domain, and Unsupervised Domain Adaptation performance were unilaterally higher when WSJ was used as the source corpus and Brown as the target. Perhaps the reason for this is obvious in the case of In-Domain training: the Brown Corpus includes different genres and thus presumably represents a much wider body of grammatical structures and styles than the WSJ corpus. Thus, it is more difficult for a model to parse across this varied corpus.

At the same time, this logic might suggest Brown would thus make a stronger source for Out-of-Domain testing, since it would capture grammatical structures and nuances of various genres. However, this does not appear to be the case for these experiments. One hypothesis for this result is that the Brown model may be actually bloated by the nuances of different genres, while the WSJ model is highly tuned to a large body of more or less uniformly structured grammar, allowing it to isolate and capture important features that happen to be generalize-able across many instances of many genres. Described another way, this might be seen a problem of overfitting. The parser must only learn one structure over all of the training examples in the case of the WSJ corpus, while it must learn 8 (crudely defined w.r.t. the number of genres in the experimental data used) over the training examples with the Brown corpus. One related point of interest that might support this view is the point between 5,000 and 7,000 training examples when the in-domain F1 score for the WSJ corpus tends rapidly toward 90 and declines afterward. This suggests a clear critical mass in the training data needed to capture the essential grammatical structure of the WSJ corpus, but also illustrates how further training data can introduce noise and bloating to the model.

Despite the fact that all models performed better in the case of WSJ $\rightarrow$ Brown, the average performance *gains* in F1 score offered by unsupervised domain adaptation were actually higher in the case of Brown $\rightarrow$ WSJ (4 points vs. 2 points). This makes sense: since training on a large body of text that conforms to a single genre, as in the case of our experiments with WSJ as seed, appears to learn a parser that is better able to isolate and discriminate based upon some subset of important features that are generalize-able, self-training on a corpus that conforms to one genre appears to have some of the same benefits. Alternatively, these increased gains when self-training on WSJ could be entirely due to the fact that the self-training data is able to more effectively capture the nuances of the single-genre WSJ domain than the multiple-genre Brown domain. An interesting experiment that might shed more light on these two possibilities might compare performance of domain adaptation using target, multi-genre self-training data with out-of-domain, single-genre self-training data. This is left for future work.

Comparison to Reichart and Rappoport

These experiments reinforce one of the claims of Reichart and Rappoport: the domain adapted parser strictly outperforms the baseline on out-of-domain test data. Unfortunately, the performance was still significantly lower than performance using in-domain training data, and the improvements offered by varying in-domain self-training data and holding the seed set constant were very small. This suggests that the gains are primarily due to introduction of *more* training data, rather than the introduction of *in-domain* training data that effectively captures the specific genre of the target domain.

Unfortunately, Reichart and Rappoport did not perform any experiments that varied the size of the self-training data, so I can not compare the latter of these results; however, it is worth noting that these results are not necessarily in disagreement with their claims. In fact, Reichart and Rappoport explicitly note prior art which has shown that self-training is usually not considered a valuable technique for improving the performance of statistical parsers with large seed sets. Instead, they show that in the specific case of small seed sets, domain adaptation can offer greater gains. This narrative is reinforce by my results, which show greater gains at small seed sizes.

Another difference is the use of the Stanford Statistical Parser, a much more sophisticated parser than the Collins' Parser, for my study. The linguistic heuristics integrated into the Stanford PCFG Parser, which negate some of the independence assumptions of PCFGs [2], may ultimately narrow the gap between in-domain and out-of-domain self-training sets.

My results may also suggest that training (self-training or seed) data from a single-genre source, such as WSJ, might actually be more valuable for modern statistical parsers than multi-genre data. Unfortunately, Richart and Rappoport only claim that their results were *similar* in the inverted case, where WSJ was used as the self-training data, but did not provide concrete results that might shed further light. Unfortunately, the results in this paper are not sufficient to draw conclusions.

V. Conclusion

Though advancements in statistical parsers have been one of the great achievements NLP, statistical parsers can often be very specific to the genre of the training corpus. In this study, I investigate a semi-supervised approach for domain adaptation of statistical parsers previously shown to be viable for small seed sets by Reichart and Rappoport [3]. I found that even for modest seed set sizes, the gains offered by domain adaptation are very small. For very small seed set sizes, the gains are slightly higher, though it appears that this might be a result of simply having more training data, rather than capturing domain-specific knowledge with the self-training data. Nevertheless, domain adaptation offers potential benefits when large amounts of annotated data are difficult to acquire.

References

[1] The Stanford Parser: A Statistical Parser http://nlp.stanford.edu/software/lex-parser.shtml
[2] D. Klein, C. D. Manning. Accurate Unlexicalized Parsing.
[3] R. Reichart, A. Rappoport. Self-Training for enhancement and domain adaptation of statistical parsers trained on small datasets. *ACL.* Vol. 7, 2007.