CS388: Assignment-3

Shobhit Chaurasia
Department of Computer Science
UTEID: sc52987

1. INTRODUCTION

In this assignment, the goal is to incorporate domain adaptation in statistical parsing. A statistical parser trained on a one corpus might learn grammar rules that are specific to the underlying grammatical structure or genre of that corpus. Hence, it makes for an interesting experiment to evaluate the performance of a statistical parser on a different corpus.

One can expect that the parser trained on a "source" domain might not perform as well on a "target" domain especially if the two domains are significantly different in terms of genre or style of writing. However, instead of training a parser from scratch for the target domain, we can adapt a parser trained on the source domain to the target domain through *domain adaptation*. In domain adaptation, usually, a small set of labeled examples from the target domain is available to adapt the trained parser to the target domain. In this assignment, we will use a special case of this form of domain adaptation called unsupervised domain adaptation, which represents the worst case scenario - no labeled data from the target domain is made available to the parser.

2. UNSUPERVISED DOMAIN ADAPTATION

In unsupervised domain adaptation, we do not have access to labeled data from the target domain. A parser trained on a source domain is used to produce labels (in this case, parse trees) for sentences from the target domain. This "self-labeled" data is then used to re-train the parser along with the gold-labeled data from the source domain. One can repeat the above step multiple times, leading to an approach similar to semi-supervised (hard) EM. In this assignment, though, we re-train the parser using the self-training data only once, and then evaluate its performance on the target domain.

3. RESOURCES USED

Unlexicalized PCFG [2] from The Stanford Parser package [1] was used to perform the experiments. The two datasets used were the PennTreebank WSJ and Brown datasets.

4. EXPERIMENTS

The experiments involved using a part of WSJ (source domain) as seed set to train the parser, followed by using the trained parser to generate self-training data from unlabeled Brown corpus (target domain), followed by re-training the parser with the original seed set and the self-training data pooled together, and evaluating the parser on the Brown corpus. The experiments also involved switching the roles of the two corpus, and varying the size of the seed set and the self-training set. For comparisons, baseline experiments where the parser is trained and tested on the source domain without self-training, and control experiments where the parser is trained on source domain and tested on target domain without self-training were also performed.

5. EXPERIMENT SET-1

The first set of experiments included using sections 02-22 of WSJ as seed set, followed by using 90% of Brown corpus as self-training set, and performing evaluation on the remaining 10% of Brown corpus. The size of the seed set was varied from $1,000$ to $35,000$ to analyze the effect of increasing seed size on the performance of the parser in the target domain. F1 scores are plotted against the seed set size in Fig.1.

As evident from the red curve, the F1 score increases with increase in the size of seed set. This is expected, since a larger seed set trains a better parser initially, which enables better parses (i.e. fewer incorrect parses) to be generated for the self-training set, thereby increasing the quality of pseudo-supervision for the final training.

A Control experiment (green curve in Fig. 1) was also performed in which the parser was trained using WSJ as the seed set, and evaluated on the Brown corpus (the same 10% as before) without performing self-training. The curve shows a similar increasing pattern as the learning curve for the previous setting because a larger seed set trains a better parser. However, the performance without self-training is always worse than the performance with self-training (by $1 - 4.7$ points), thereby indicating that domain adaptation even in a completely unsupervised manner is useful. This should not be surprising, because the parser trained using the seed set does indeed generate a "decent" self-training set which has at least a small fraction of correctly parsed sentences. The pooling of the original seed set and this "decent" self-training induces some amount of target domain knowledge into the training set, thereby training a better parser overall (for the target domain). Further, the difference in the F1 score between the parser with and without self-training gradually decreases (from ≈ 4.7 points to ≈ 1 point) as the size of seed set is increased. This is because once the seed size is sufficiently large, the *difference* in the quality of pseudo-supervision due to addition of more seed examples fades out since this addition, while being useful, does not alter the parser's learned hypothesis significantly.

Further, an in-domain (Baseline) experiment (blue curve in Fig. 1) was performed to evaluate the performance of the

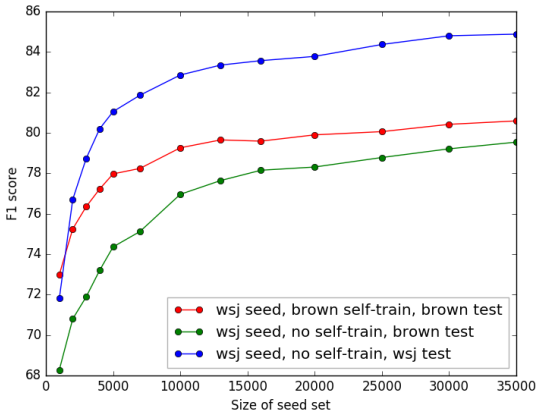

Figure 1: F1 score as a function of the size of seed set. The green curve corresponds to the Control setting. The blue curve corresponds to the Baseline setting.

Figure 2: F1 score as a function of size of self-training set, with $10,000$ seed examples.

parser which was trained on WSJ, on WSJ itself, without self-training. Once again, the performance increases with increase in the size of the seed set for the reasons mentioned above. Moreover, as expected, the in-domain F1 score is significantly better than the out-of-domain F1 score (by $4-7$ points), because the parser is primarily trained on WSJ corpus, and has better learned the underlying structure of WSJ as compared to that of a new domain. The self-training in the first case is merely pseudo-supervision, and not enough to adapt the parser as well to Brown as it is to WSJ on which it was originally trained (with real-supervision).

Something that's worth noting is that the rate of increase in F1 score in all the three settings decreases with increase in the size of seed set. This is because initially, when the size of the seed set is fairly small (e.g., less than $10,000$), the addition of more seed examples helps train a better parser. This is because a small seed set does not provide enough data to the parser to learn different nuances of the corpus, and addition of new seed examples provides significant amount of valuable supervision . However, when the parser is already trained on a sufficiently large seed set (e.g., greater than $20,000$), it has probably already learned quite a lot about the corpus (provided the seed set was reflective of the universe), and additional seed examples, while being useful, do not provide significantly newer information to the parser.

6. EXPERIMENT-2

This is similar to the first experiment in that sections 02-22 of WSJ were used as seed set to train a parser, followed by using 90% of Brown corpus for self-training, and evaluation was done on the remaining 10% of Brown corpus. However, instead of varying the size of seed set, the aim of this experiment was to analyze the effect of the size of self-training set on the parser's performance. The size of seed set was fixed at 10,000 parse trees from WSJ, while the size of self-training set was varied from 1000 to $21,000$. The F1 scores are plotted against the self-training set size in Fig.2.

There is a general increasing trend in the F1 score as the size of self-training set is increased which is because a larger self-training set provides the parser with more supervision to learn from. However, one must not forget that this supervision could be highly noisy, which is potentially why there is a meager drop in F1 score (79.13 to 79.08) on increasing the size of self-training set from $13,000$ to $17,000$, and a less than usual increase at 4000.

It should be noted that the magnitude of increase in F1 score as the size of self-training set is increased from 1000 to $21,000$ (keeping seed set fixed) is only ≈ 1.5 points (see Fig. 2) as compared to 7 points increase when the size of seed set is increased from 1000 to $20,000$ (keeping self-training set fixed; see Fig. 1). Although this comparison is not totally fair, it reflects how valuable seed-supervision (in the form of gold-labeled examples) is over the noisy pseudo-supervision (in the form of self-trained examples). Nonetheless, gold-labeled data is hard to collect, while self-training is free of cost, and the fact that it can boost the performance even by a small amount is amazing.

7. EXPERIMENT SET-3

The intent behind these set of experiments is, in essence, same as that behind the Experiment set-1. However, the source and target domains were flipped. This set of experiments included using 90% of the Brown corpus as the seed set, followed by using sections $02-22$ of WSJ as self-training set, and evaluating on section 23 of WSJ corpus. The size of the seed set was varied from $1,000$ to $21,000$ to analyze the effect of increasing seed size on the performance of the parser in target domain. The F1 scores are plotted against the seed set size in Fig.3.

The general trends are similar to those in experiment set-1, and hence, the corresponding observations will only be briefly stated here without full explanations to avoid repetition. An important observation is that the F1 scores in all the three settings in this set of experiments (with Brown as source domain and WSJ as target domain) are in general lower than their counter-parts in experiment set-1 (with WSJ as source domain, and Brown as target domain). This is because the Brown corpus has a wider spectrum of genres than WSJ, and hence is harder to learn.

As evident from the red curve, the F1 score increases with increase in the size of seed set. This is expected, since a larger seed set trains a better parser.

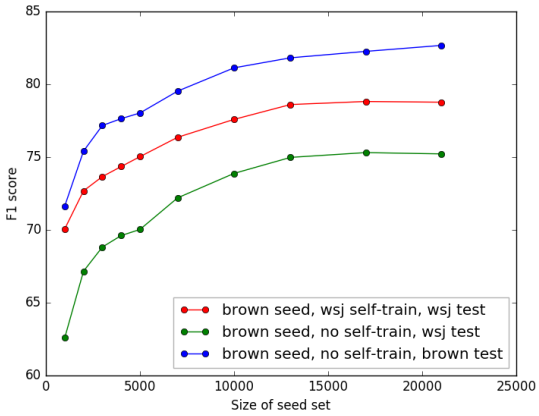

Figure 3: F1 score as a function of size of seed set. The green curve corresponds to the Control setting. The blue curve corresponds to the in-domain setting.

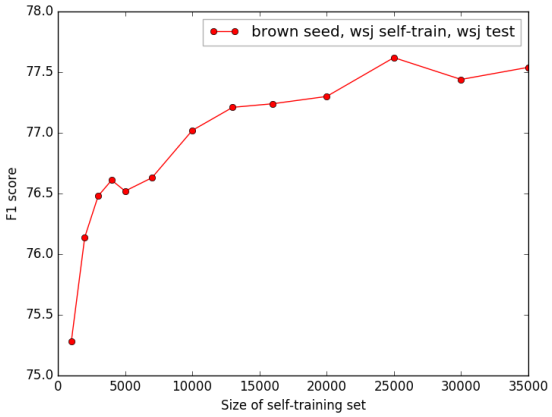

Figure 4: F1 score as a function of size of self-training set, with $10,000$ seed examples.

A Control experiment (green curve in Fig. 3) was also performed in which the parser was trained using Brown as the seed set, and evaluated on the WSJ corpus (section 23) without performing self-training. The curve shows a similar increasing pattern as the learning curve for the previous setting because a larger seed set trains a better parser. However, the performance without self-training is always worse than the performance with self-training (by $3.5-7.5$ points), thereby indicating that domain adaptation even in a completely unsupervised manner is useful. The difference is more pronounced here than the difference when WSJ was the source domain, indicating that self-training is more useful when Brown is the source and WSJ is the target domain, as compared to the setting with WSJ as source and Brown as target domain because of WSJ's "specificity" and Brown's generic nature.

The reason why unsupervised domain-adaptation works better for going from Brown to WSJ than the other way round is two-fold:

1. Brown is a more generic corpus with multiple genres, while WSJ is a highly specific one - it contains only news-wire articles. Domain adaptation from a generic corpus to a highly specific corpus is easier than the other way round.

2. The self-training data generated out of the Brown corpus by a parser trained on WSJ is expected to be of poorer quality than the one generated out of WSJ by a parser trained on Brown. This is because the parser trained on WSJ will have poorer generalization (in terms of performance on other genres) than the one trained on Brown.

Further, an in-domain experiment (blue curve in Fig. 3) was performed to evaluate the performance of the parser which was trained on Brown, on Brown itself, without self-training. Once again, the performance increases with increase in the size of the seed set. Here also, the in-domain F1 score is significantly better than the out-of-domain F1 score (by $8-10$ points)

8. EXPERIMENT-4

This is, in essence, similar to experiment-2, with source and target domains flipped - 90% of the Brown corpus was used as the seed set, followed by sections $02-22$ of WSJ used as self-training set, and evaluation done on section 23 of WSJ corpus. The aim of this experiment was to analyze the effect of the size of self-training set on the parser's performance. The size of seed set was fixed at 10,000 parse trees from Brown, while the size of self-training set was varied from 1000 to $35,000$. The F1 scores are plotted against the self-training set size in Fig.4.

Once again, the general observations are similar to that in experiment-2, and hence, will only be briefly stated here. As pointed out in the previous section, F1 scores in this setting (with Brown as seed and WSJ as self-training/testing) are in general lower than their counter-parts in experiment-2 (with WSJ as seed, and Brown as self-training/testing).

There is a general increasing trend in the F1 score as the size of self-training set is increased which is because a larger self-training set provides the parser with more supervision to learn from. However, one must not forget that this supervision could be highly noisy, which is potentially why there are meager drops in F1 scores in a couple of places on increasing the size of self-training set.

9. COMPARISON WITH PAPER

The observations drawn from the above experiments are in line with the results obtained in [3]. This is despite the fact that the parser used in [3] is the Collins parser, which is different from the one used in these experiments. This further affirms the conclusion that self-training is indeed useful for domain adaptation, and that this is not necessarily an artifact of the strength or weakness of a specific parser.

10. CONCLUSION

The results strongly suggest that self-training is a powerful method for domain adaptation. Its impact is especially profound when the seed training set is small. Even in settings with larger seed sets, an increase of even a couple of points of F1 score for free that it provides (without the need for additional labeled data from target domain) is impres-

sive. We also learned that not only is the size of seed set important, but also that the size of the self-training set can make a huge difference.

11. REFERENCES

[1] The stanford parser: a statistical parser. http://nlp.stanford.edu/software/lex-parser.shtml.

[2] Klein, D., and Manning, C. D. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (2003), Association for Computational Linguistics, pp. 423–430.

[3] Reichart, R., and Rappoport, A. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *ACL* (2007), vol. 7, pp. 616–623.