1. Implementation

The way we implement semi-supervised self-training is as follows: we train the parser on a given seed set and the trained parser parses the self-training set. The whole automatically annotated self-training set is then combined with the seed set to retrain the parser. The retrained parser is evaluated on the given test set.

2. Experiment

The experimented datasets are *wsj* and *brown*, both of which are sourced from Linguistic Data Consortium (LDC)'s Penn Treebank collection. We use section 02-22 of *wsj* dataset as the train seed set and its section 23 as the test set. For the *brown* dataset, we use first 90% sentences of each genre as the train seed set and the remained part as the test set.

Experiment 1 Train Seed Dataset: wsj. Unlabelled Datasets: brown.

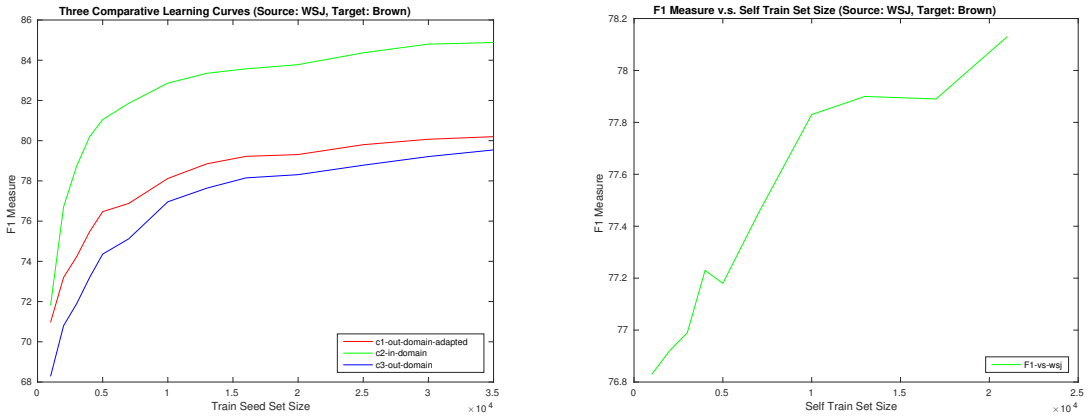

Figure 1: Left: Three Comparative Learning Curves. Right: Impact of Self-Train Set Size

Experiment 2 Train Seed Dataset: brown. Unlabelled Datasets: wsj.

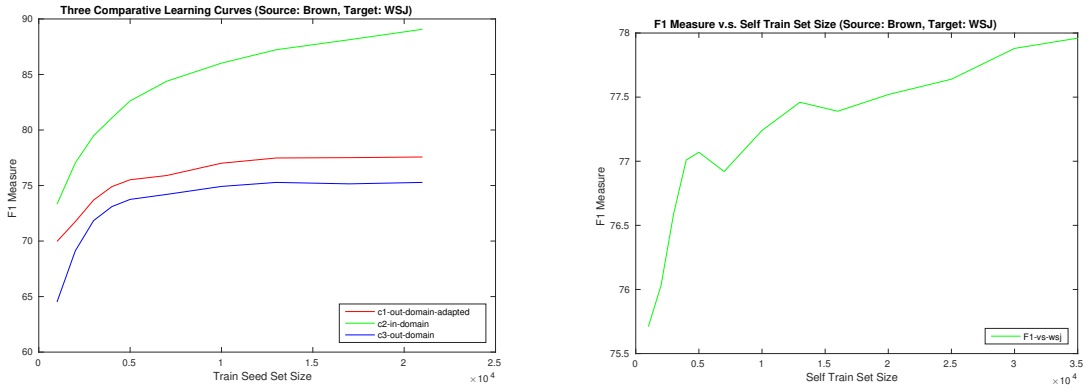

Figure 2: Left: Three Comparative Learning Curves. Right: Impact of Self-Train Set Size

3. Discussion

(a) How much does performance drop from in-domain testing to out-of-domain testing?

From the left side of Fig. 1 and Fig. 2, it can be observed that a large drop of F1 measure exists from in-domain testing to out-of-domain testing. If we use *wsj* as the source and *brown* as the target, the drop of F1 measure is around **7**. If we switch the source and the target (as shown in Fig. 2), then the drop would be around **15**.

(b) How does unsupervised domain adaptation impact performance on out-of-domain testing?

The unsupervised domain adaptation generally improves the out-of-domain testing. In Experiment 1, the out-of-domain testing with adapted training improves F1 measure by approximately **2**. And in Experiment 2, the out-of-domain testing with adapted training improves F1 measure by approximately **3**. This can be explained by the fact that self-training mechanism does help to transfer the knowledge of source domain to the target domain, and then improve the parsing performance of the target domain.

(c) How does increasing the size of seed and self-supervised training sets effect the relative performance?

Generally speaking, increasing the size of train seed set and self-train set does help improve the relative parsing performance. The behind reason is obvious (1) more train set instances better help the parser understands the source domain knowledge, which would in some sense be useful in parsing the target domain sentences, (2) more self-train set better help the parser to transfer the knowledge of the source domain to the target domain, and thus improve the overall performance of parsing in the target domain.

(d) How does inverting the "source" and "target" impact your results and why?

The F1 measure of out-of-domain testing in Experiment 1, is higher than that in Experiment 2, comparing two blue lines in the Fig. 1 and Fig. 2. That means the choice of source domain does impact the resulting parsing performance for the unsupervised domain adaptation task. The behind reason can be as follows: mastering the linguistic model of *wsj* could help to master the linguistic model of *brown*, much more than the reverse case.

(e) How do your results compare to the results described in the Reichart and Rappoport paper for the OI setting?

Overally, the experimental results I derived are consistent with those in the Reichart and Rappoport paper. The only difference that I got smaller numeric values for some testings. This might be because their preprocessing implementation includes some more suitable features.