Gaurav Nanda	Homework #3
Natural Language Processing	March 29, 2014

1 Implementation

1.1 Domain Adaptation Parser

This parser provides basic functionality to train with the seed data, annotate a pool of self-training data and combine the seed and self-trained data to create a new model. Further, performance of the new model is measured using the test dataset.

Seed data, self-training data and test dataset are passed as arguments to the parser. If there is no self-training data, pass "NOADAPTION" as an argument. Another optional argument has been provided to enable the multi-threading. Parsing of the self-training data is done in parallel using the number of threads argument.

Memory Tree Bank is used to combine self-training and seed data into the new model. Also, I am using *parseMultiple* function in *LexicalizedParser* class, which internally call *parse* method to convert input sentence to a Tree. However, there was an exceptional case that I handled for the sentences which are not parsed properly. If there is an exception while parsing, Parser add "X" as its root. Therefore, before adding self-trained trees back to the new model, we filter out all trees with "X" at its root.

1.2 Pre-Processors

There are two preprocessors implmented to produce various input files required for the Domain Adaption Parser.

First, Sentence Extractor, takes input location, output location and maximum number of sentences to be generated. Therefore, for WSJ we provide 35000 as max number and it generates various splits in range of 1 and 35,000 separted by 1000. Therefore, we go over input file only once.

Second, Split Brown, reads in each genre and takes 90-10 splits for each of them. Further, all 90s are clubbed together to form training file and rest are put in test file.

2 Experiments

Severals experiments were run in order to see the performance of domain adaptation or transfer learning. Here is the discussion of the experiments we ran:

2.1 Varying size of WSJ seed

Here we performed three different tests while varying the nunber of sentences in WSJ seed. Tests include unsupervised domain adaptation by normal training on WSJ, self-training on Brown, and

then testing on Brown, normal training on WSJ and testing on Brown and normal training and testing on WSJ. Training dataset is 90% of brown corpus and rest of 10% is test corpus.

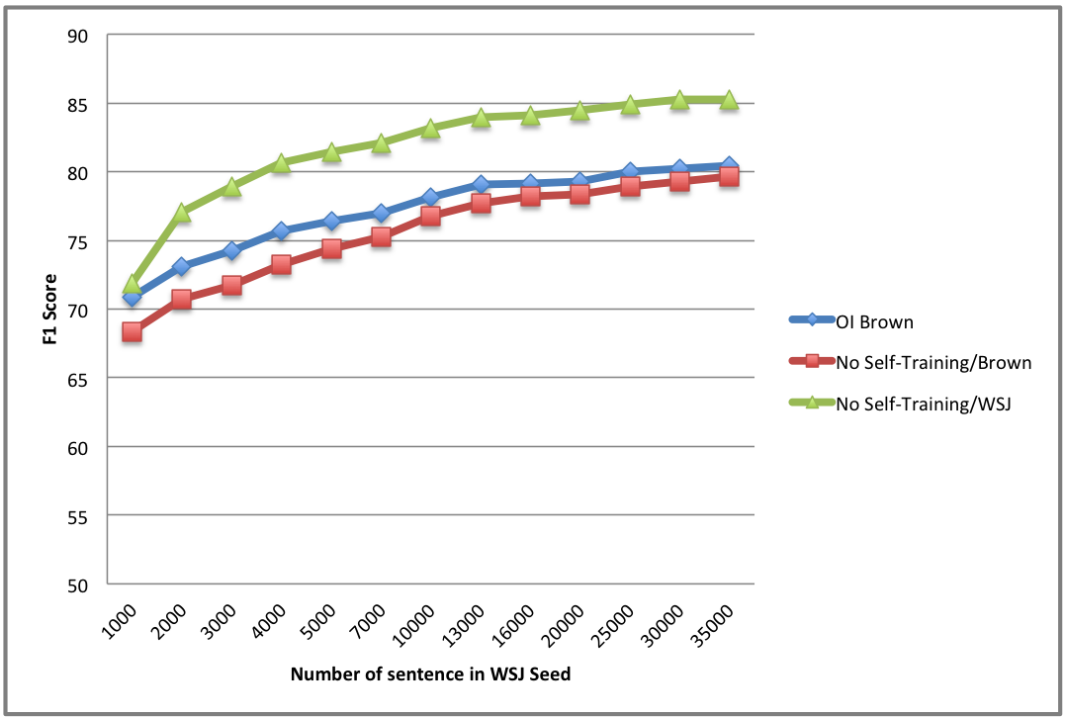

Figure 1: Learning curve of F1 score as a function of number of sentences in the seed set. (Blue:- Unsupervised domain adaptation by normal training on WSJ, self-training on Brown, and then testing on Brown). (Red:- Normal training on WSJ and testing on Brown). (Green:- Normal training and testing on WSJ).

In-domain testing Vs. out-of-domain testing

Without any domain adaptation, on average, there is 7.57% drop in F1 score from in-domain testing to out-of-domain testing. However, for 35,000 sentence we still see roughly 80 F1 score in out-of-domain testing. This reflects that domain of Brown and WSJ corpus, are distant, but not distant enough.

Impact of unsupervised domain adaptation on out-of-domain testing

On average there is an increase of 2.16% in F1 score, when we do unsupervised domain adaptation. However, rate of percentage increase drops, as we keep on increasing the seed-data. For instance, for 1000 seed sentences, percentage increase is 3.67% and for 35000 percentage increase is only 0.96%. This signifies the impact of domain adaptation is lesser when we have large amount of seed data.

Impact of seed size.

As we keep on increasing the seed size, F1 score keeps on increasing. However, as expected, percentage increase keeps on going lower with larger seed size. For example, for domain adaptation OI case, percentage increase from 1000 to 2000 sentence is 3.10%. However, percentage increase

from 30,000 to 35,000 is only 0.27%.

Here is the table having all the values for this experiment:

Seed Size	OI Brown	No Self-Training/Brown	No Self-Training/WSJ	% Drop I-O testing	% Adaptation Increase
1000	70.87	68.36	71.84	4.844097996	3.671737858
2000	73.07	70.69	77.03	8.230559522	3.366812845
3000	74.25	71.74	78.95	9.132362255	3.49874547
4000	75.68	73.25	80.66	9.186709645	3.317406143
5000	76.4	74.39	81.44	8.656679764	2.701976072
7000	76.96	75.23	82.12	8.39016074	2.299614515
10000	78.11	76.75	83.17	7.719129494	1.771986971
13000	79.06	77.71	83.96	7.444020962	1.737228156
16000	79.17	78.22	84.14	7.03589256	1.21452314
20000	79.25	78.35	84.48	7.256155303	1.148691768
25000	80.01	78.94	84.89	7.009070562	1.355459843
30000	80.22	79.31	85.26	6.97865353	1.147396293
35000	80.44	79.67	85.27	6.567374223	0.966486758
Average				**7.573143581**	**2.169081987**

Figure 2: Table showing F1 scores as a function of number of sentences in the seed set.

OI Brown: F1 score, where, normal training on WSJ, self-training on Brown, and then testing on Brown.

No Self-Training/Brown: F1 score, with normal training on WSJ and testing on Brown.

No Self-Training/WSJ: F1 score, where, normal training and testing on WSJ.

% Drop I-O testing: Percentage drop in F1 score from in-domain to out-domain testing.

% Adaption Increase: Percentage increase in F1 score with unsupervised domain adaptation on baseline out-of-domain testing.

2.2 Varying size of self-supervised Brown training sets.

Here, we fixed the WSJ seed size to 10,000 sentences and varied the size of self-supervised training data. Increasing size of self-supervised training data hardly increases F1 score. For 1000 sentences, F1 score is 76.91 and for 21,000 sentences, F1 score is 78.01. There is very little increase with increase in count of training data size.

These numbers come as no surprise and are totally expected. If we observe the previous experiment, F1 score with 10000 WSJ seed sentences, for baseline case is 76.75 and score with self-learning over entire Brown training set is 78.11. Therefore, in this experiment F1 score had to move between these two numbers only and that is exactly what is happening.

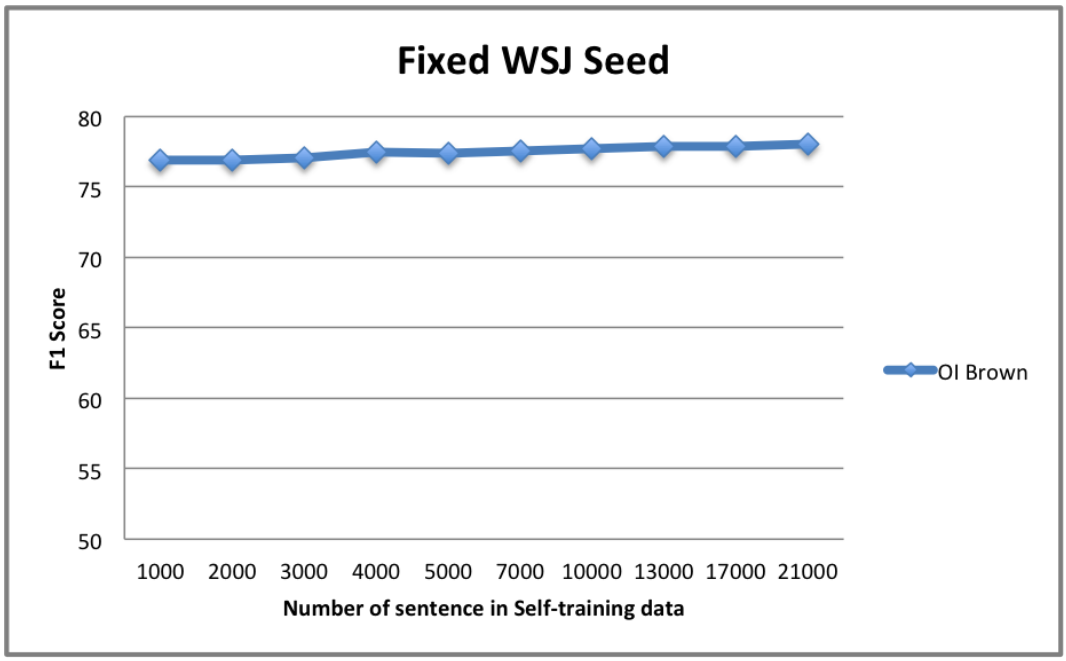

Figure 3: Table showing F1 scores as a function of number of sentences in the seed set.

2.3 Varying size of Brown seed.

Here we perform the same three tests, as we had performed while varying the WSJ seed size. However, we use previous 90% Brown data as the seed set, WSJ sections 02-22 as the self-training data and WSJ section 23 as the test set.

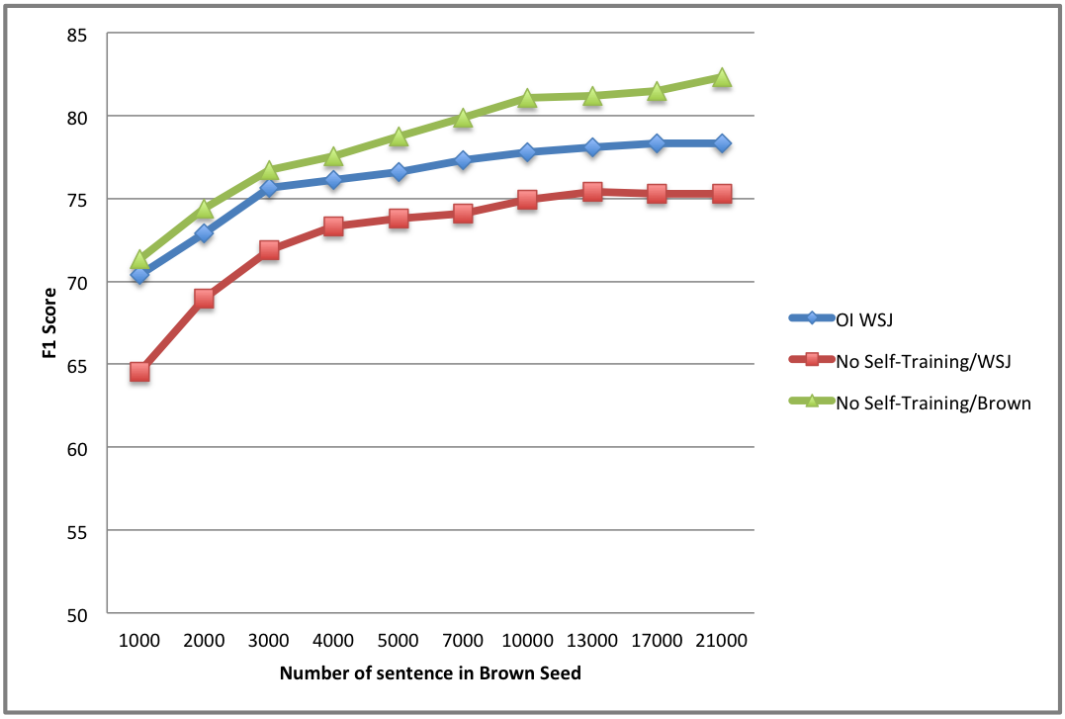

Figure 4: Learning curve of F1 score as a function of number of sentences in the seed set. (Blue:- Unsupervised domain adaptation by normal training on Brown, self-training on WSJ, and then testing on WSJ). (Red:- Normal training on Brown and testing on WSJ). (Green:- Normal training and testing on Brown).

Impact of inverting the "source" and "target"

Most of the observations are similar to our previous test for WSJ. However, if we observe the graph carefully, gap between blue and red lines has increased. This tells us that percentage increase of F1 values with self-trained data is higher in comparison to previous WSJ test.

This means that, domain adaptation is more helpful for WSJ data set in comparison to Brown data set. I surmise this happens because WSJ is not a diverse dataset and has data only pertaining to the financial world. Therefore, as the term "domain adaptation" suggests, when addition of self-trained WSJ data, F1 score on test data increases. Moreover, this increase is higher in comparison to self-adaptation on Brown data as that is a diverse data set.

In-domain testing Vs. out-of-domain testing

Without any domain adaptation, on average, there is 7.3% drop in F1 score from in-domain testing to out-domain testing. This number is roughly similar to number before inversion.

Impact of unsupervised domain adaptation on out-of-domain testing

On average there is an increase of 4.73% when we do unsupervised domain adaptation. However, percentage increase decreases, as we keep on increasing the seed-data. For instance, for 1000 seed sentences, percentage increase is 9.08% and for 21000 percentage increase is 3.98%. This signifies the impact of domain adaptation is less when we have large amount of seed data.

These numbers as we discussed earlier highlight the benefit of doing domain adaptation for WSJ in comparison to a diverse dataset like Brown.

Impact of seed size.

As we keep on increasing the seed size, F1 score keeps on increasing. However, as expected, rate of percentage increase keeps on going lower with larger seed size. For example, for domain adaptation OI case, percentage increase from 1000 to 2000 sentence is 3.56%. However, percentage increase from 17,000 to 21,000 is almost 0.

Here is the table having all the values for this experiment:

Seed Size	OI WSJ	No Self-Training/WSJ	No Self-Training/Brown	% Drop I-O testing	% Adaptation Increase
1000	70.38	64.52	71.33	9.547175102	9.082455053
2000	72.89	68.98	74.38	7.260016133	5.668309655
3000	75.63	71.86	76.69	6.298083192	5.246312274
4000	76.12	73.32	77.55	5.454545455	3.818876159
5000	76.59	73.77	78.76	6.335703403	3.822692151
7000	77.29	74.1	79.87	7.224239389	4.304993252
10000	77.77	74.93	81.05	7.55089451	3.790204191
13000	78.08	75.38	81.21	7.178918852	3.58185195
17000	78.32	75.26	81.5	7.656441718	4.065904863
21000	78.3	75.3	82.32	8.527696793	3.984063745
Average				**7.303371455**	**4.736566329**

Figure 5: Table showing F1 scores as a function of number of sentences in the seed set.

OI WSJ: F1 score, where, normal training on Brown, self-training on WSJ, and then testing on WSJ.

No Self-Training/WSJ: F1 score, with normal training on Brown and testing on WSJ.

No Self-Training/Brown: F1 score, where, normal training and testing on Brown.

% Drop I-O testing: Percentage drop in F1 score from in-domain to out-domain testing.

% Adaption Increase: Percentage increase in F1 score with unsupervised domain adaptation on baseline out-of-domain testing.

2.4 Varying size of self-supervised WSJ training sets.

Here, we fixed the Brown seed size to 10,000 sentences and varied the size of self-supervised WSJ training data. Increasing size of self-supervised training data hardly increases F1 score. For 1000 sentences, F1 score is 75.82 and for 35,000 sentences, F1 score is 77.86. There is very little increase with increase in count of training data size.

As we discussed earlier, these numbers also range between Baseline and self-training over entire training set.

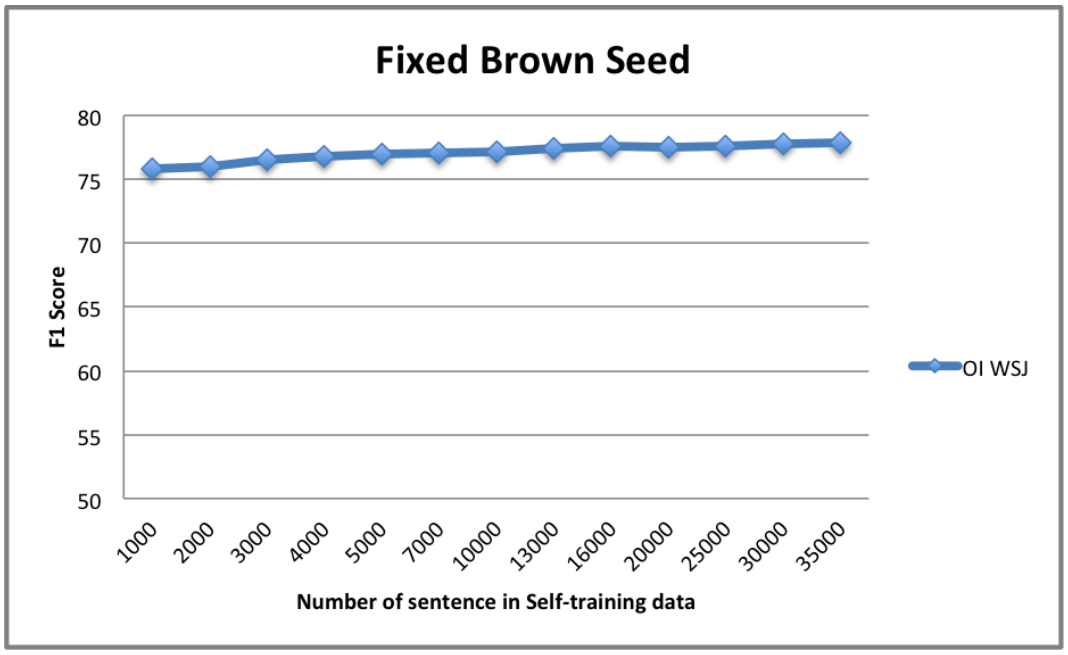

Figure 6: Table showing F1 scores as a function of number of sentences in the seed set.

3 Comparison with Reichart and Rappoport paper for OI setting

Our results corroborate with the results of paper. As discussed in the paper, there is consistent increase in the F1 score with self-training data for the OI setting. As we also observed, percentage increase in F1 number is higher for smaller seed data. This implies that, if we have less manually annotated data, we can self-train using un-annotated data from test domain, which would help to increase the F1 score.

Also, as the paper suggested, F1 score for self-training data would always be better than the baseline case and upper bounded by In-domain testing numbers. This is what we also observed in Figure 1 and Figure 4.