Parse Selection for Self-Training

*CLMS thesis under construction*

Jim White
March 14th 2013
Outline

• Introduction
• What is Self-Training?
• Proposed Questions for my Thesis
• Parse Quality Estimation for Statistical Parsing
  o Token-based Methods (sentence length, corpus n-grams)
  o Tree & Word Entropy
  o Ensemble Agreement (SEPA)
  o Methods for Dependencies (PUPA & ULISSE)
• Proposed Experiments
  o WSJ + NANC -> Brown, Brown + NANC -> WSJ, WSJ/Brown + ? -> GENIA
  o Iterated Self-Training?
• The X Factor
  o Exploit Inter-sentential Context?
  o Learn something from domain variation?
  o Data Adaptation?
• What is the Story with Reranking?
Introduction

• Statistical Constituency Parsing

• Domain Sensitivity
  - Gildea 2001, Corpus Variation and Parser Performance
  - The Brown Corpus has various genres of literature
  - Brown Test split is 9th of every 10 sentences, Dev is every 10th
  - Parsers have improved but degradation remains
    ▪ Train on WSJ: Test WSJ at 92%, Brown at 85.8%
    ▪ WSJ seems to be a good base corpus though
    ▪ Train on Brown, Test on Brown gets 87.4%
What Is Self-Training?

• Data Plentiful, Human Labels Scarce

• Eat Your Own Dog Food:
  o Train ParserA on Labeled Corpus
  o Parse Unlabeled Corpus with ParserA
  o Train ParserB on a Combination of Labeled Corpus and Output of ParserA
  o Some Weighting (and Waiting) May Be Involved…

• Early Efforts for Parsing Failed
  o Charniak 1997, Steedman 2003
Self-Training *Can* Work for Parsing

- McClosky, Charniak, and Johnson 2006, *Effective Self-Training for Parsing*
  - Train on WSJ (labeled) + 1,750k of NANC (unlabeled)
  - Test on WSJ improves from 91.0 to 92.1
  - Reduction in error of 8%
  - Used Charniak & Johnson 2005 reranking parser
Eeking Out an Edge

The effect of sentence length. Some sentences are “just right”. The Goldilocks Effect.
Works for Domain Adaptation Too

- McClosky 2010 PhD diss., *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*
  - Train on WSJ + X, Test on GENIA
  - WSJ-only is 77.9% w/o reranker and 80.5% with
  - Manual adaptation using three methods utilizing additional lexical resources w/o reranker gets 80.2%
    - Lease & Charniak 2005, *Parsing Biomedical Literature*
  - WSJ + 266k MEDLINE gets 84.3%
  - WSJ + 37k BIOBOOKS gets 82.8% (37k MEDLINE is 83.4%)
    - Shouldn’t BIOBOOKS work better since it seems like it is closer to WSJ than MEDLINE?
But Confusion Abounds...

- Reichart and Rappoport 2007, *Self-Training for Enhancement and Domain Adaptation of Statistical Parsers Trained on Small Datasets*
  - Unknown word rate *is* predictive of effectiveness when using a large amount of data

- McClosky et al 2008, *When is Self-Training Effective for Parsing?*
  - Unknown word rate *is not* predictive of effectiveness when using a large amount of data
  - But unknown bigrams and unknown biheads are
Too Much of a Good Thing?

- Sometimes using more self-training data hurts performance

<table>
<thead>
<tr>
<th>Parser model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>f-score</strong></td>
<td>Parser</td>
<td>Reranking parser</td>
</tr>
<tr>
<td>WSJ alone</td>
<td>83.9</td>
<td>85.8</td>
</tr>
<tr>
<td>WSJ + 2,500,000 NANC</td>
<td>86.4</td>
<td>87.7</td>
</tr>
<tr>
<td>BROWN alone</td>
<td>86.3</td>
<td>87.4</td>
</tr>
<tr>
<td>BROWN + 50,000 NANC</td>
<td>86.8</td>
<td>88.0</td>
</tr>
<tr>
<td>BROWN + 250,000 NANC</td>
<td>86.8</td>
<td>88.1</td>
</tr>
<tr>
<td>BROWN + 500,000 NANC</td>
<td>86.7</td>
<td>87.8</td>
</tr>
<tr>
<td>BROWN + 1,000,000 NANC</td>
<td>86.6</td>
<td>87.8</td>
</tr>
<tr>
<td>WSJ + BROWN</td>
<td>86.5</td>
<td>88.1</td>
</tr>
<tr>
<td>WSJ + BROWN + 50,000 NANC</td>
<td>86.8</td>
<td>88.1</td>
</tr>
<tr>
<td>WSJ + BROWN + 250,000 NANC</td>
<td>86.8</td>
<td>88.1</td>
</tr>
<tr>
<td>WSJ + BROWN + 500,000 NANC</td>
<td>86.6</td>
<td>87.7</td>
</tr>
<tr>
<td>WSJ + BROWN + 1,000,000 NANC</td>
<td>86.6</td>
<td>87.6</td>
</tr>
</tbody>
</table>

Table 3.9: f-scores from various combinations of WSJ, NANC, and BROWN corpora on BROWN development. The reranking parser used the WSJ-trained reranker model. The BROWN parsing model is naturally better than the WSJ model for this task, but combining the two training corpora results in a better model (as in Gildea (2001)). Adding small amounts of NANC further improves the results.
Parser Self-Training w/o Reranking


![Graphs showing performance improvement with self-training](image)

**Figure 2:** The performance of the PCFG-LA parser and Charniak’s parser evaluated on the test set, trained with different amounts of labeled training data, with and without self-training (ST).

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Unlabeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>39.8k</td>
<td>1.7k</td>
<td>2.4k</td>
<td>210k</td>
</tr>
<tr>
<td></td>
<td>(950.0k)</td>
<td>(40.1k)</td>
<td>(56.7k)</td>
<td>(5,082.1k)</td>
</tr>
<tr>
<td>Chinese</td>
<td>24.4k</td>
<td>1.9k</td>
<td>2.0k</td>
<td>210k</td>
</tr>
<tr>
<td></td>
<td>(678.8k)</td>
<td>(51.2k)</td>
<td>(52.9k)</td>
<td>(6,254.9k)</td>
</tr>
</tbody>
</table>

**Table 1:** The number of sentences (and words in parentheses) in our experiments.
Proposed Questions for my Thesis

• Can Self-Training Performance be Improved by Parse Selection?
  o Most work reported so far just uses all of the self-parsed data
  o Performance should improve if we only use the good stuff

• Can Iterated Self-Training Work Effectively?
  o Most reports so far show bad results for repeating the self-training steps
  o If the model moves only a little at a time then perhaps it could eventually go further

• Can Parser Self-Training Performance be Predicted?
  o Self-training tuning usually depends on labeled data
  o If we could use the current model to predict performance on unlabeled data effectively then all three of these things could work together
Parse Quality Assessment

• **Token-based Methods**
  - sentence length - a good baseline
  - corpus n-grams

• **Tree & Word Entropy**
  - Hwa 2000, *Sample Selection for Statistical Grammar Induction*

• **Ensemble Agreement (SEPA)**

• **Methods for Dependencies (PUPA & ULISSE)**
  - Reichart & Rappoport 2009, *Automatic Selection of High Quality Parses Created By a Fully Unsupervised Parser*
  - Dell’Orletta et al. 2011, *ULISSE: an Unsupervised Algorithm for Detecting Reliable Dependency Parses*
Active Learning for Human Annotation

Rank sentences in a new domain to minimize the cost of manual annotation for a given level of performance

- How to rank them?
  - Highest uncertainty
  - Lowest confidence
  - Lowest similarity
  - Best parse tree entropy
  - Should locality matter?

- Shorter, shallower sentences are easier for humans to annotate reliably
Oil the Squeaky Wheel(s) First

Lots of work done on this around a dozen years ago


Rank by highest parse tree entropy

- The sentence whose best parse tree has the highest entropy given the current model is most informative
- If annotating in batches (as humans are wont to do) then avoid selecting sentences that are very close to each other in each round
Parser Uncertainty & Tree Entropy

- Rebecca Hwa 2000, 2004
- Parser uncertainty is parse tree entropy divided by the log of the number of parse trees
- Parse tree entropy can be efficiently computed using dynamic programming from the parser’s model

\[
func(w, G) = \frac{TE(w, G)}{\lg(\|V\|)}
\]

\[
TE(w, G) = H(V) = -\sum_{v \in V} p(v) \lg(p(v))
\]
Word Entropy

- Tang, Luo, & Roukos 2002, Active Learning for Statistical Natural Language Parsing
- If you don’t have all the parse trees and their likelihood, compute using the ones you have (as in n-best parsing)

\[
H_S = \sum_{i=1}^{K} -p_i \log p_i
\]

\[
p_i = \frac{q_i}{\sum_{j=1}^{K} q_j}
\]

\[
H_w = \frac{H_S}{L_S}
\]
PLTIG parser: (a) A comparison of the evaluation functions’ learning curves. (b) A comparison of the evaluation functions for a test performance score of 80%. Rebecca Hwa, 2004, Sample Selection for Statistical Parsing
Not So Much for Self-Training Though

F-score on all of Brown vs % of self-parsed (WSJ) data using several metrics
Shorter Sentences are Easier

F-score on Brown sentences with < 40 words vs % of self-parsed (WSJ) data using several metrics
Accounting for Sentence Length

All: Log scale: -850

Medium: > 15 & <= 40 words Log scale: -600

Short: <= 15 words Scale: -250

Long: > 40 words, Log scale: -1000
The More The Merrier!

- Train N parsers on size S subsets of the training data
- For each sentence, parse it with each of the N parsers, choose one of the parses and compute an agreement score F for it
- Sort the chosen parses by F

\[
MF(s) = \frac{1}{N - 1} \sum_{i \in [1...N], i \neq l} f_{score}(m_i, m_l)
\]
Now We’re Getting Somewhere

• SEPA Mean F-score correlates much better with gold F-score than parser uncertainty

• Correlation Coefficient of 0.66

\[ \text{Correl}(X,Y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \]
Hmm…

• First Look at using SEPA for Self-Training Parse Selection

• EVALB on Brown Test only

F-score vs SEPA S %
Self-Training Experiment Design

- Same setup as McClosky et al 2006
  - Source Corpus: PTB WSJ, Target Corpus: Brown
  - Parser: BLLIP reranking parser (Charniak and Johnson 2005)
That Something Extra

• Exploit Inter-sentential Context?
  o Current methods treat sentences in isolation
  o Can we use contextual information such as semantic cohesion a la WSD by Yarowsky 1995?

• Learn Something from Domain Variation?
  o Lower SEPA grades for training data indicate variation from the corpus as a generalizable source
  o Can we use that information - maybe as a feature in learning local context

• Data Adaptation?
  o Modify the training data to generalize better w/o having to modify the parser models
  o Kundu & Roth 2011, *Adapting Text instead of the Model: An Open Domain Approach*
What is the Story with Reranking?

Kenji Sagae, *Self-Training Without Reranking for Parser Domain Adaptation and Its Impact on Semantic Role Labeling*

- What happens when the output from a semi-supervised parser is used in an application?
- Evaluate on a task extrinsic to syntactic parsing
- CoNLL 2005 Shared Task: Semantic Role Labeling (SRL)
“... attempt to quantify the benefits of semi-supervised parser domain adaptation in semantic role labeling, a task in which parsing accuracy is crucial.”

- Reranking has been shown to be effective for parser DA when used with self-training
  McClosky et al 2007, *Effective Self-Training for Parsing*

- Self-training w/o reranking has also been shown to be effective
  Reichart & Rappoport 2007, *Self-Training for Enhancement and Domain Adaptation of Statistical Parsers Trained on Small Datasets*
SRL DA Experiments

Train on WSJ, Test on Brown
  - Same setup as McClosky et al 2006 ("MCJ")
    Charniak parser with and w/o reranker
      - Tested and decided using different weights for different sources in self-training wasn’t worth the effort – “simple self-training”

Baseline: Top performing ST system trained on WSJ only.
  - UIUC SRL 79.44 F-score on WSJ and 64.75 on Brown.
The highest scoring system on the Brown evaluation in the CoNLL 2005 shared task had an F-score of 67.75.

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., 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.

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; Precision 66.57, Recall 63.02, F-score 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)
What’s the Verdict on Reranking?

What exactly got tested?

- “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.”
- UIUC SRL 79.44 F-score on WSJ and 64.75 on Brown.
- “... 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.”

What happens if the other SRL components are included in self-training?
Thank You!

http://depts.washington.edu/newscomm/photos/the-spring-cherry-blossoms-in-the-quad/