Self-Training for Biomedical Parsing

David McClosky and Eugene Charniak

Brown Laboratory for Linguistic Information Processing (BLLIP)

{dmcc|ec}@cs.brown.edu

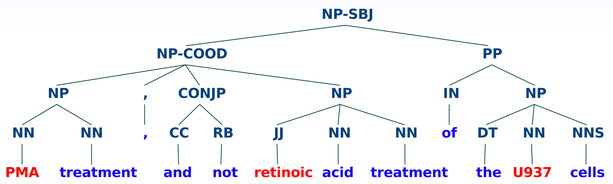
Biomedical Parsing

Goal

Parse biomedical articles for better document retrieval and automatic analysis.

Challenges

- Many unknown words (~25%)
- Little annotated data (~10k trees in GENIA [5])



Gold tree fragment from GENIA.
Words not seen in WSJ are **marked**.

Previous Work

Parser adaptation

- Train on source domain, self-train on target domain [1]

Self-training for portability

- Train and self-train on source domain [4]

Biomedical specific

- Biomedical tagger and other in-domain info [3]
- Parser comparison on GENIA corpus [2]

Approach

Self-training for adaptation

As in [1], train on source domain (WSJ) and self-train on unlabeled text. Experiment with unlabeled text in several domains:

NANC: North American News Text Corpus, 42 million sentences

BioBooks: 80k sentences from high school and college biology textbooks ("bridging corpus")

Medline: 270k sentences from abstracts of biomedical journal articles (same domain as GENIA test)

```
base = train(labeled)
autolabeled = label(base, unlabeled)
combined = labeled + autolabeled
selftrained = train(combined)
labeledtest = label(selftrained, test)
evaluate(test, labeledtest)
```

Pseudocode for self-training

Experiments

Charniak and Johnson reranking parser (2005) (WSJ reranker model)

Training data

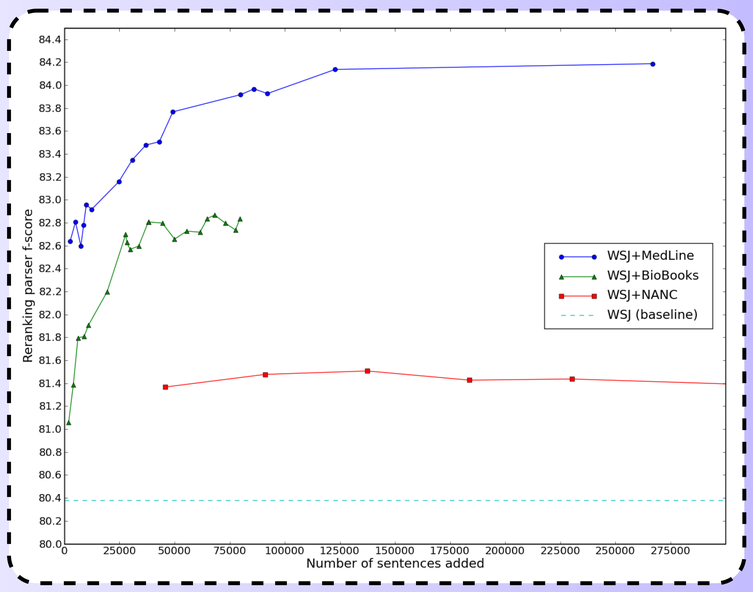
WSJ 2-21 and autolabeled data

Parameter tuning

WSJ 24

Evaluation data

GENIA development



Comparison

System	f-score
Charniak (2000)	78.0%
Collins (1999)/Bikel (2002)	79.4%
Lease and Charniak (2005)	80.2%
Charniak and Johnson (2005)	80.5%
Self-trained with Medline*	84.3%

f-scores on GENIA test set

* Charniak and Johnson reranking parser self-trained with 270k Medline sentences. Previous best on this test set was Lease and Charniak (2005) in [2].

Summary

- 20% error reduction over previous best without any in-domain labeled data.
- With labeled data, performance should be even higher.

Selected References

[1] M. Bacchiani, M. Riley, and B. Roark. *MAP adaptation of stochastic grammars.* CSL 2006

[2] A. Clegg and A. Shepherd. *Evaluating and integrating treebank parsers on a biomedical corpus*. ACL workshop on software, 2005.

[3] M. Lease and E. Charniak. *Parsing biomedical literature.* IJCNLP 2005.

[4] D. McClosky, E. Charniak, and M. Johnson. *Reranking and self-training for parser adaptation.* COLING-ACL 2006.

[5] Y. Tatesi, A. Yakushiji, T. Ohta, and J. Tsujii. *Syntax annotation for the GENIA corpus.* IJCNLP 2005.