# Effective Self-Training for Parsing

David McClosky

dmcc@cs.brown.edu

Brown Laboratory for Linguistic Information Processing (BLLIP)

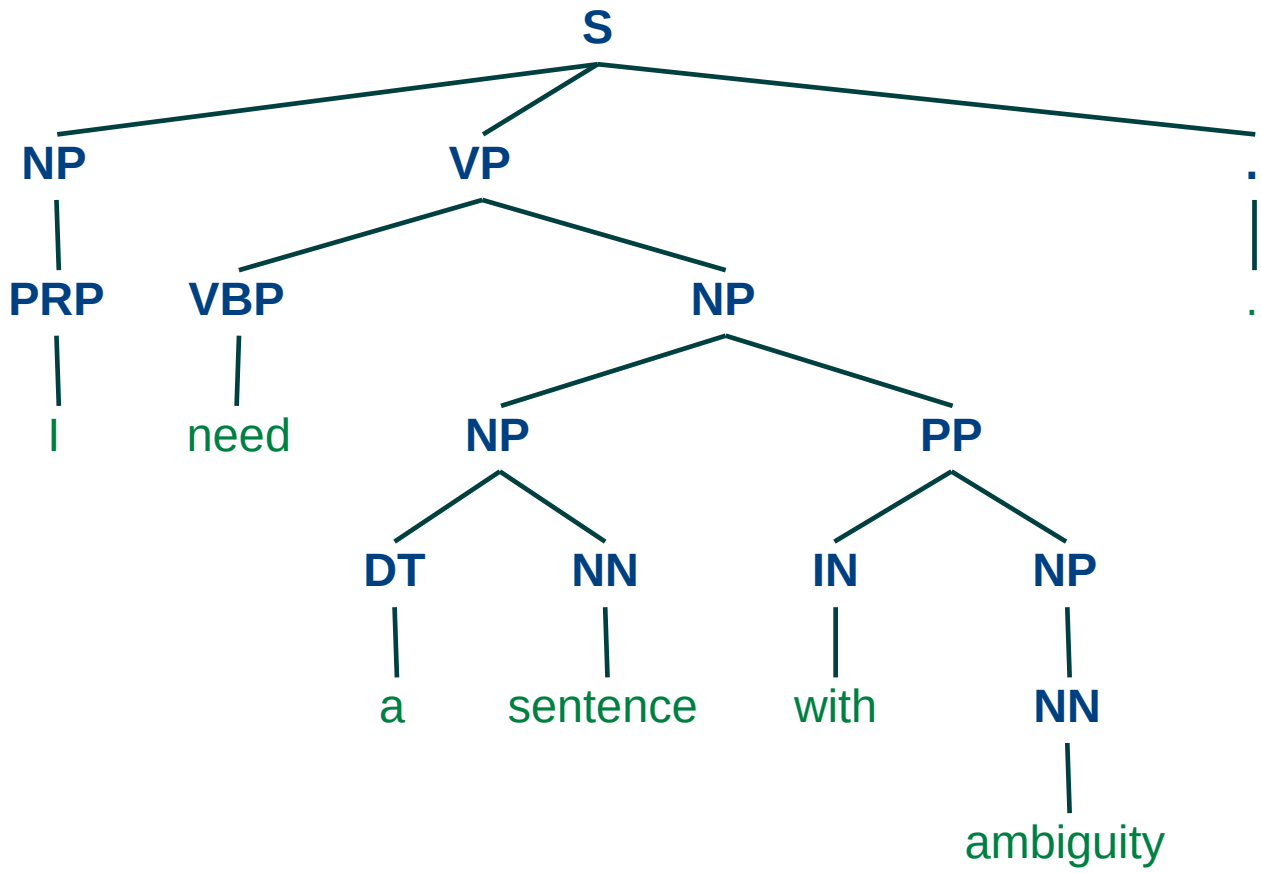
Joint work with Eugene Charniak and Mark Johnson

# Parsing

# Parsing

“I need a sentence with ambiguity.”

# Parsing




“I need a sentence with ambiguity.”

# Parsing

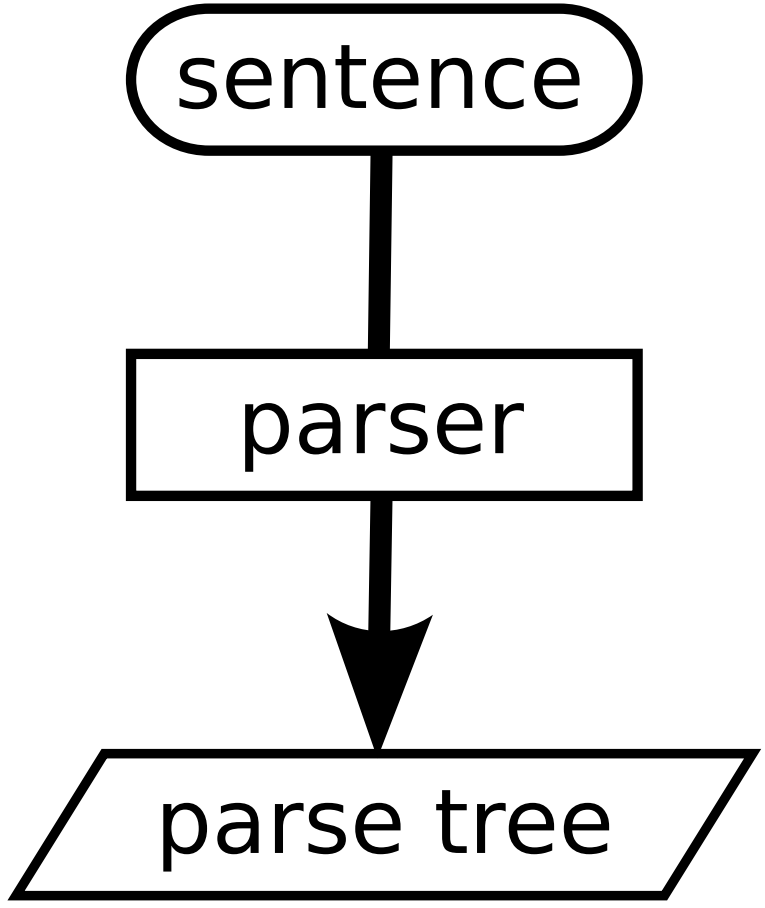
$s$ is a sentence

$\pi$ is a parse tree

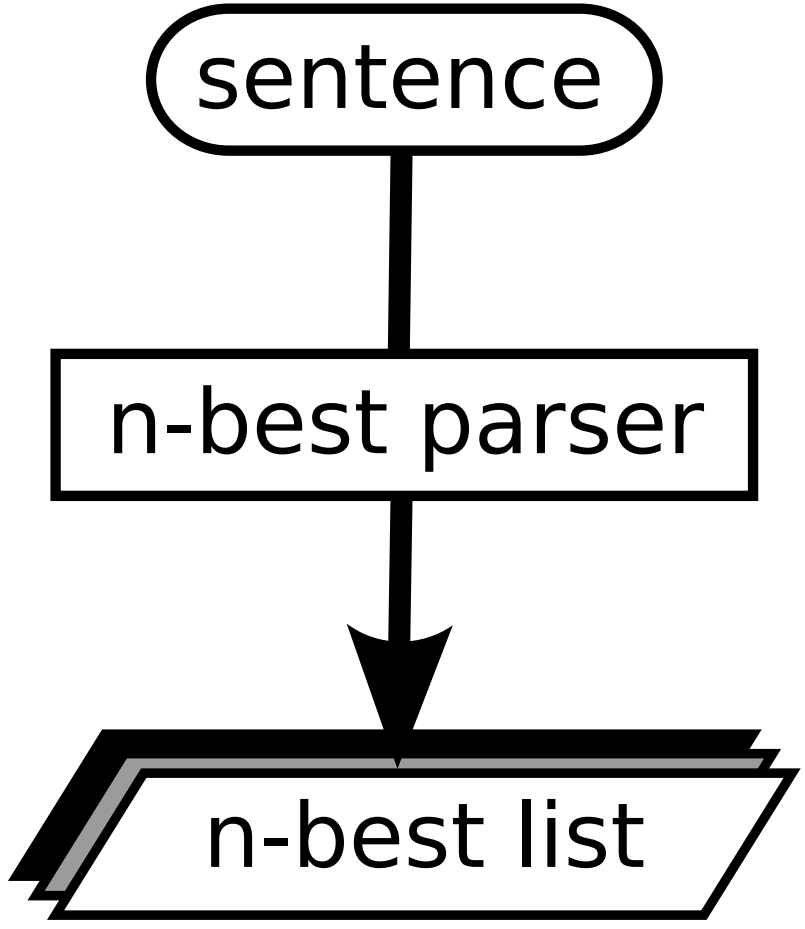
$$parse(s) = \arg\max_{\pi} p(\pi \mid s)$$

such that $yield(\pi) = s$

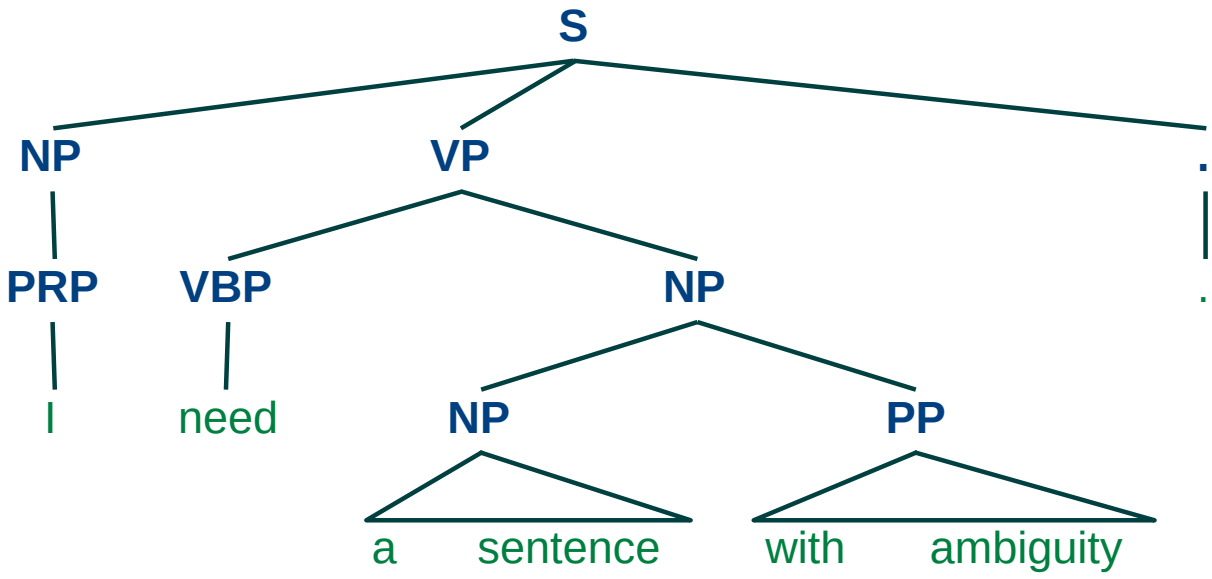
# Flow Chart



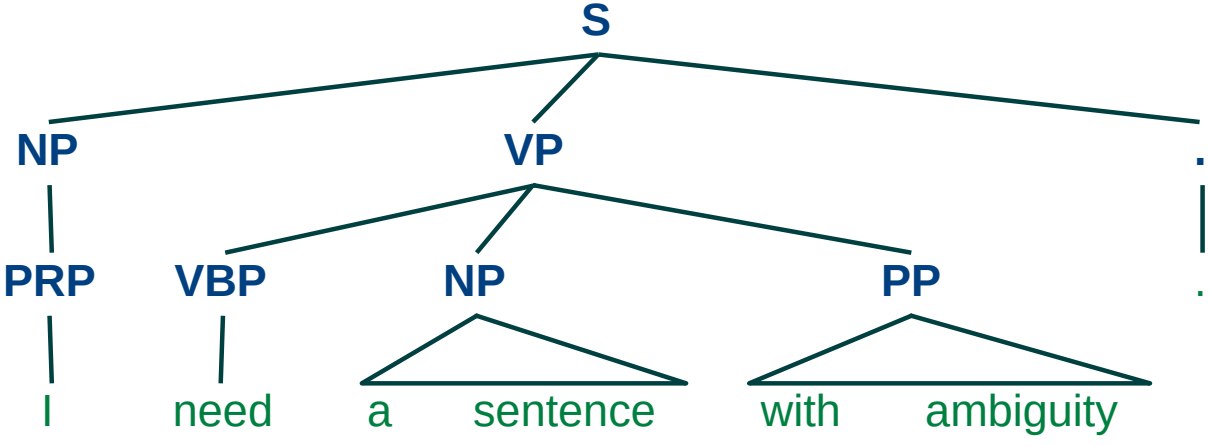
# Flow Chart



# $n$-best parsing




$p(\pi_1) = 7.25 \times 10^{-20}$

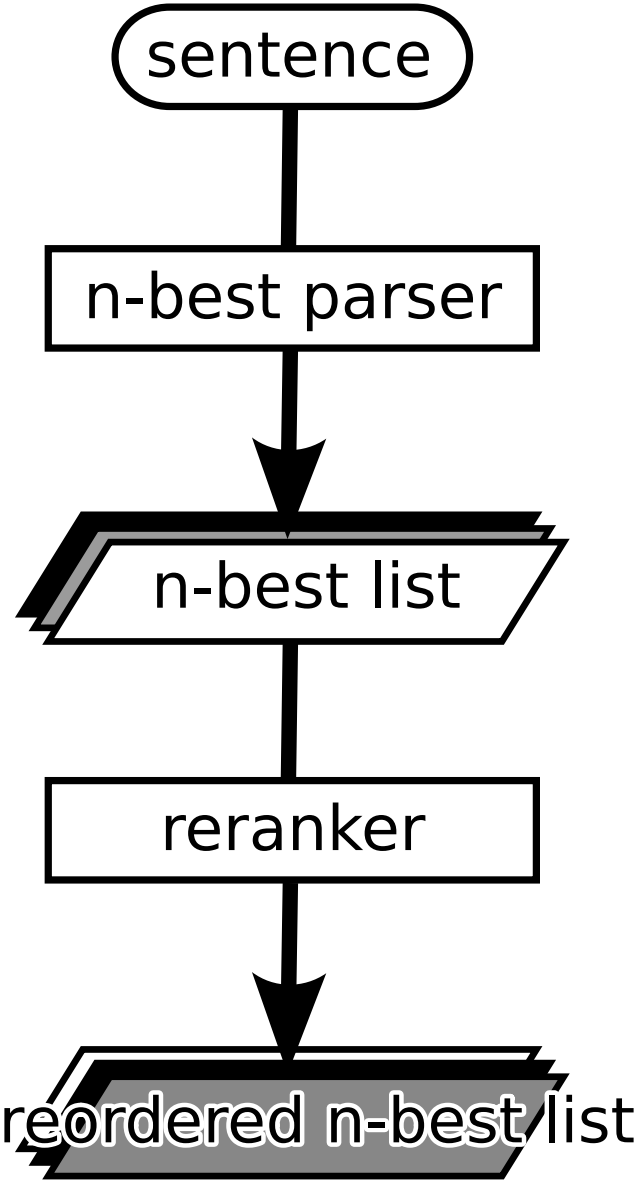



$p(\pi_2) = 7.05 \times 10^{-21}$

# Reranking Parsers

- Best parses are not always first, but the correct parse is often in the top 50
- Rerankers rescore parses from the $n$-best parser using more complex (not necessarily context-free) features
- Oracle rerankers on the Charniak parser's 50-best list can achieve over 95% $f$-score

# Flow Chart


sentence
n-best parser
n-best list
reranker
reordered n-best list

# Our reranking parser

- Parser and reranker as described in Charniak and Johnson (ACL 2005) with new features
- Lexicalized context-free generative parser, maximum entropy discriminative reranker
- New reranking features improve reranking parser's performance by 0.3% on section 23 over ACL 2005

# Unlabelled data

**Question:** Can we improve the reranking parser with cheap unlabeled data?

# Unlabelled data

**Question:** Can we improve the reranking parser with cheap unlabeled data?

- Self-training
- Co-training
- Clustering $n$-grams, use clusters as general class of $n$-grams
- Improve vocabulary, $n$-gram language model
- etc.

# Self-training

- Train model from labeled data

  train reranking parser on WSJ

- Use model to annotate unlabeled data

  use model to parse NANC

- Combine annotated data with labeled training data

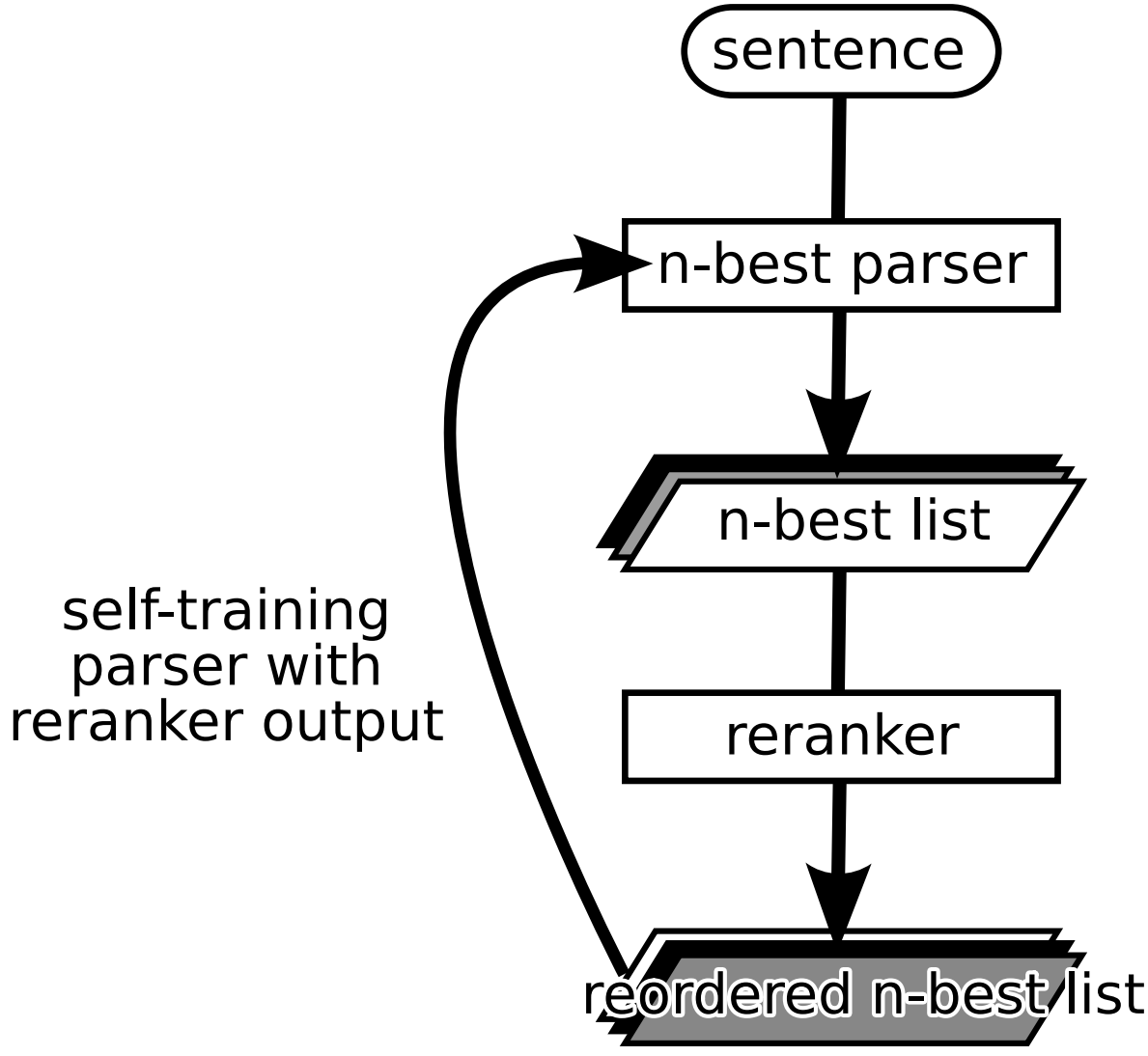
  merge WSJ training data with parsed NANC data

- Train a new model from the combined data

  train reranking parser on WSJ+NANC data

- Optional: repeat with new model on more unlabeled data

# Flow Chart


sentence
n-best parser
n-best list
reranker
reordered n-best list
self-training
parser with
reranker output

# Previous work

- Parsing: Charniak (1997), confirmed by Steedman et al. (2003)
  - insignificant improvement
- Part of speech tagging: Clark et al. (2003)
  - minor improvement/damage depending on amount of training data
- Parser adaptation: Bacchiani et al. (2006)
  - helps when parsing WSJ when training on Brown corpus and self-training on news data

# Experiments (overview)

- How should we annotate data? (parser or reranking parser)
- How much unlabelled data should we label?
- How should we combine annotated unlabeled data with true data?

# Annotating unlabeled data

| | Annotator | |
|---|---|---|
| Sentences added | Parser | Reranking parser |
| 0 (baseline) | 90.3 | |
| 50k | 90.1 | 90.7 |
| 500k | 90.0 | 90.9 |
| 1,000k | 90.0 | 90.8 |
| 1,500k | 90.0 | 90.8 |
| 2,000k | | 91.0 |

Parser (not reranking parser) $f$-scores on all sentences in section 22

# Annotating unlabeled data

| | WSJ Section | | |
|---|---|---|---|
| Sentences added | 1 | 22 | 24 |
| 0 (baseline) | 91.8 | 92.1 | 90.5 |
| 50k | 91.8 | 92.4 | 90.8 |
| 500k | 92.0 | 92.4 | 90.9 |
| 1,000k | 92.1 | 92.2 | 91.3 |
| 2,000k | 92.2 | 92.0 | 91.3 |

Reranking parser $f$-scores for all sentences

# Weighting WSJ data

- Wall Street Journal data is more reliable than the self-trained data
- Multiply each event in Wall Street Journal data by a constant to give it a higher relative weight

$$events = c \times events_{wsj} + events_{nanc}$$

- Increasing WSJ weight tends to improve $f$-scores.
- Based on development data, our best model is WSJ$\times$5+1,750k sentences from NANC

# Evaluation on test section

| Model | $f_{parser}$ | $f_{reranker}$ |
|---|---|---|
| Charniak and Johnson (2005) | – | 91.0 |
| Current baseline | 89.7 | 91.3 |
| Self-trained | 91.0 | 92.1 |

$f$-scores from all sentences in WSJ section 23

# The Story So Far...

- Retraining parser on its own output doesn't help
- Retraining parser on the reranker's output helps
- Retraining reranker on the reranker's output doesn't help

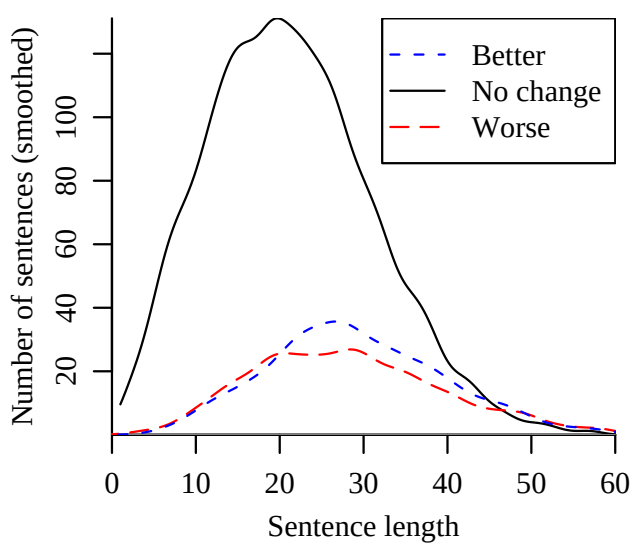
# Analysis: Global changes

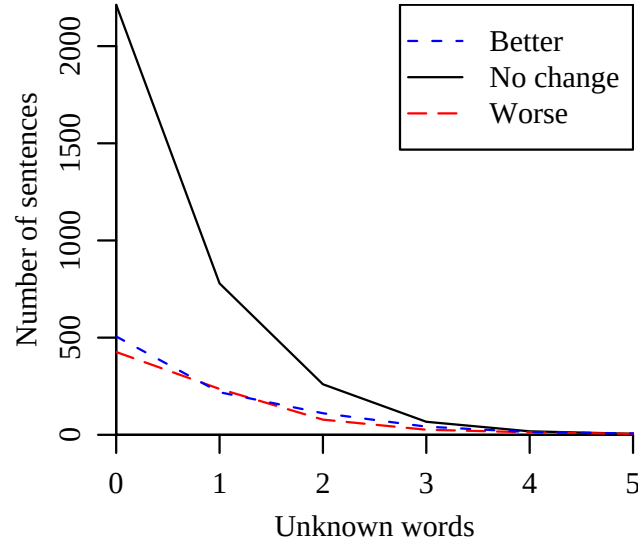
- Oracle $f$-scores increase, self-trained parser has greater potential

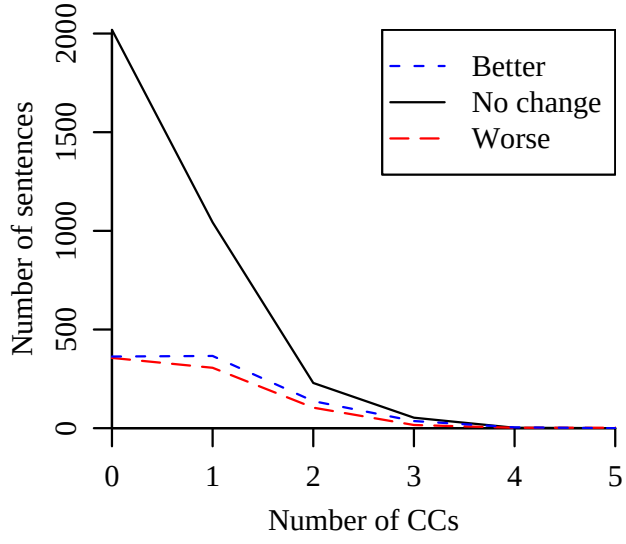
| Model | 1-best | 10-best | 50-best |
|---|---|---|---|
| Baseline | 89.0 | 94.0 | 95.9 |
| WSJ$\times$1 + 250k | 89.8 | 94.6 | 96.2 |
| WSJ$\times$5 + 1,750k | 90.4 | 94.8 | 96.4 |

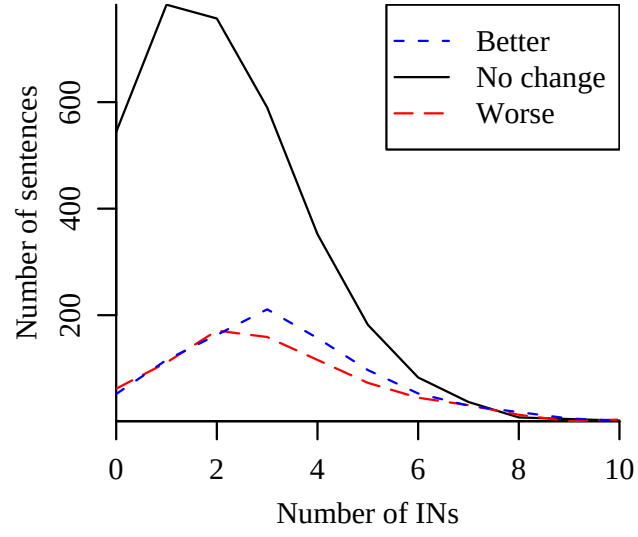
- Average of $\log_2 \frac{\text{Pr(1-best)}}{\text{Pr(50th-best)}}$ increases from 12.0 (baseline parser) to 14.1 (self-trained parser)

# Sentence-level Analysis

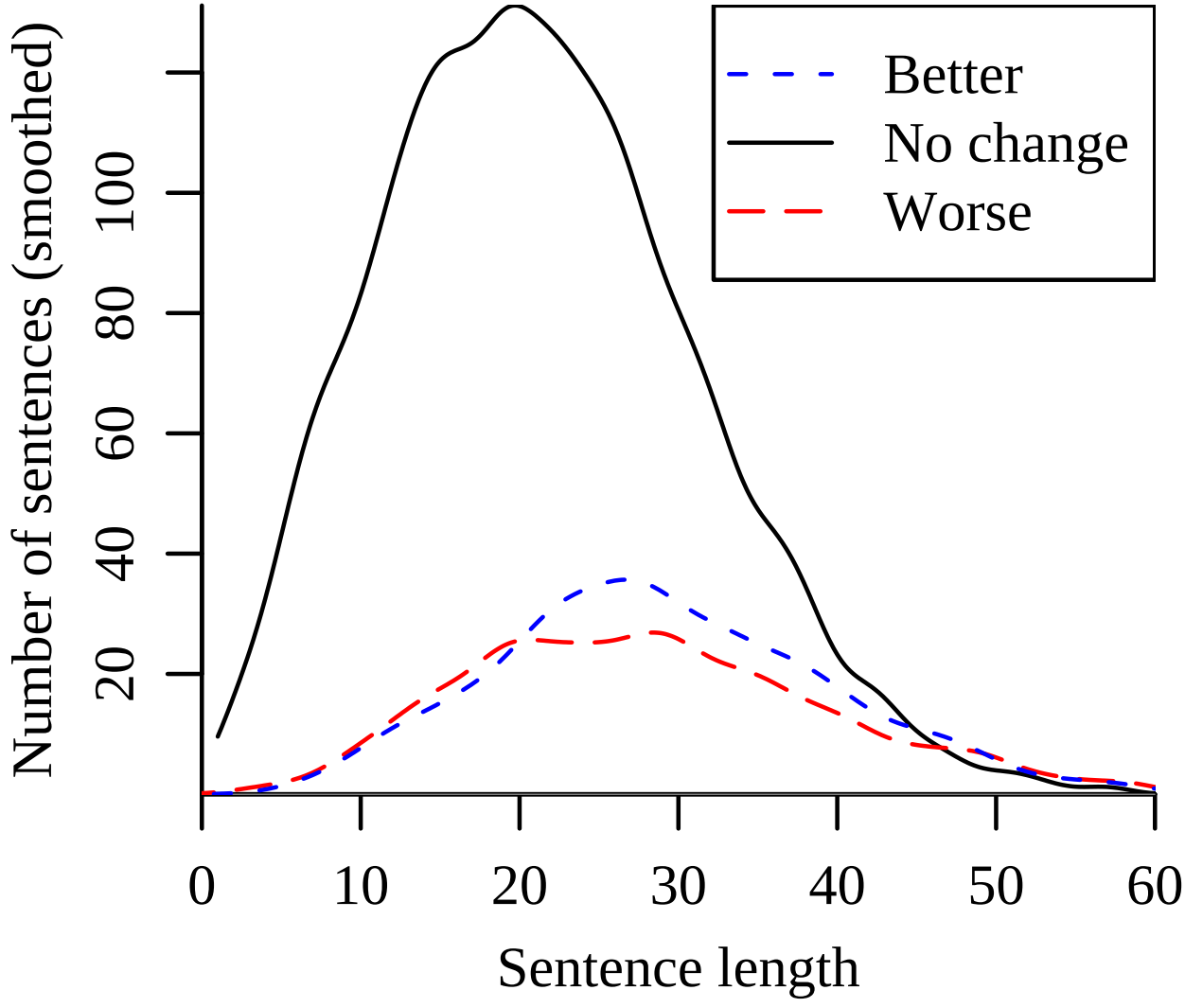


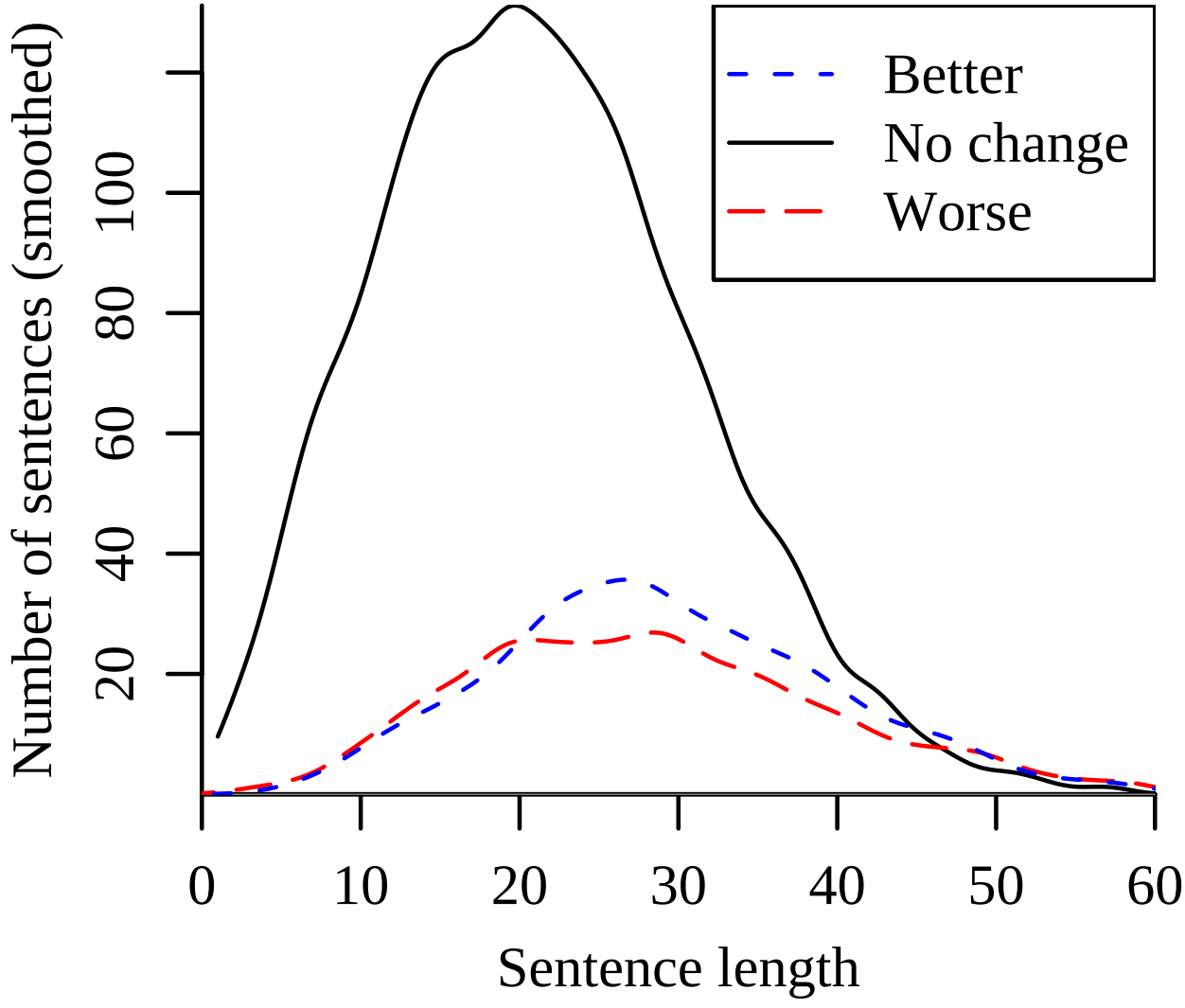



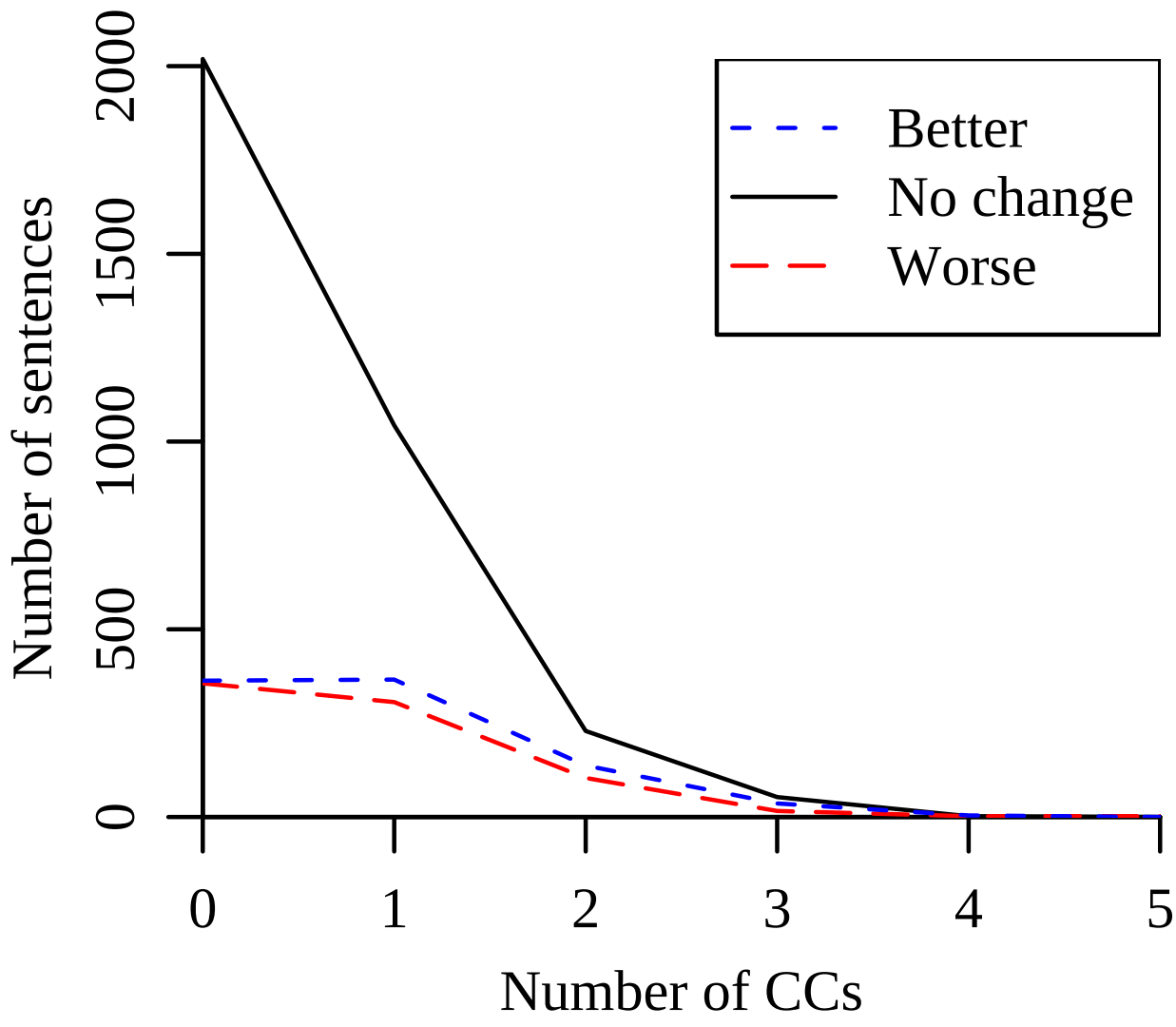
# Effect of Sentence Length



# The Goldilocks Effect™



# . . . and . . .


Number of sentences
0
500
1000
1500
2000
0
1
2
3
4
5
Number of CCs
Better
No change
Worse

# Ongoing work

- Parser adaptation (McClosky, Charniak, and Johnson ACL 2006)
- Sentence selection
- Clustering local trees
- Other ways of combining data

# Conclusions

- Self-training can improve on state-of-the-art parsing for Wall Street Journal
- Reranking parsers can self-train their first stage parser
- More analysis is needed to understand why reranking is necessary

Self-trained reranking parser available from:

`ftp://ftp.cs.brown.edu/pub/nlparser`

# Acknowledgements

This work was supported by NSF grants LIS9720368, and IIS0095940, and DARPA GALE contract HR0011-06-2-0001.

Thanks to Michael Collins, Brian Roark, James Henderson, Miles Osborne, and the BLLIP team for their comments.

## Questions?