Reranking and Self-Training for Parser Adaptation

David McClosky, Eugene Charniak, and Mark Johnson

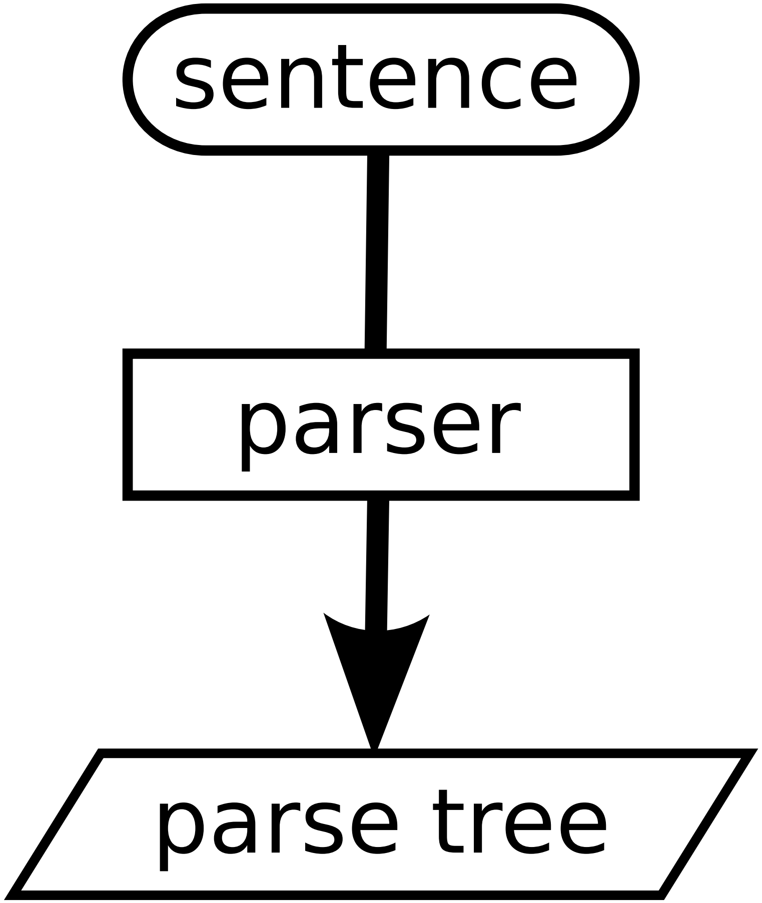
{dmcc|ec|mj}@cs.brown.edu

Brown Laboratory for Linguistic Information Processing (BLLIP)

Overview

- Introduction and Previous Work
- Parser portability
- Parser adaptation
- Reranker portability
- Analysis
- Future Work and Conclusions

Parsing



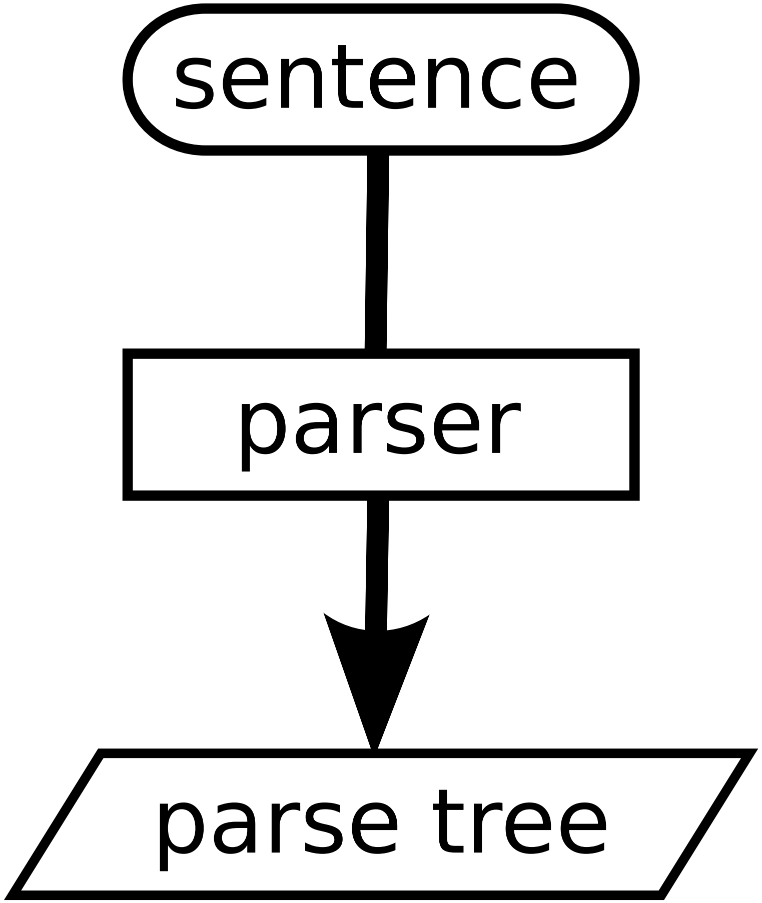
Parameters

Parser as in [Charniak and Johnson ACL 2005]

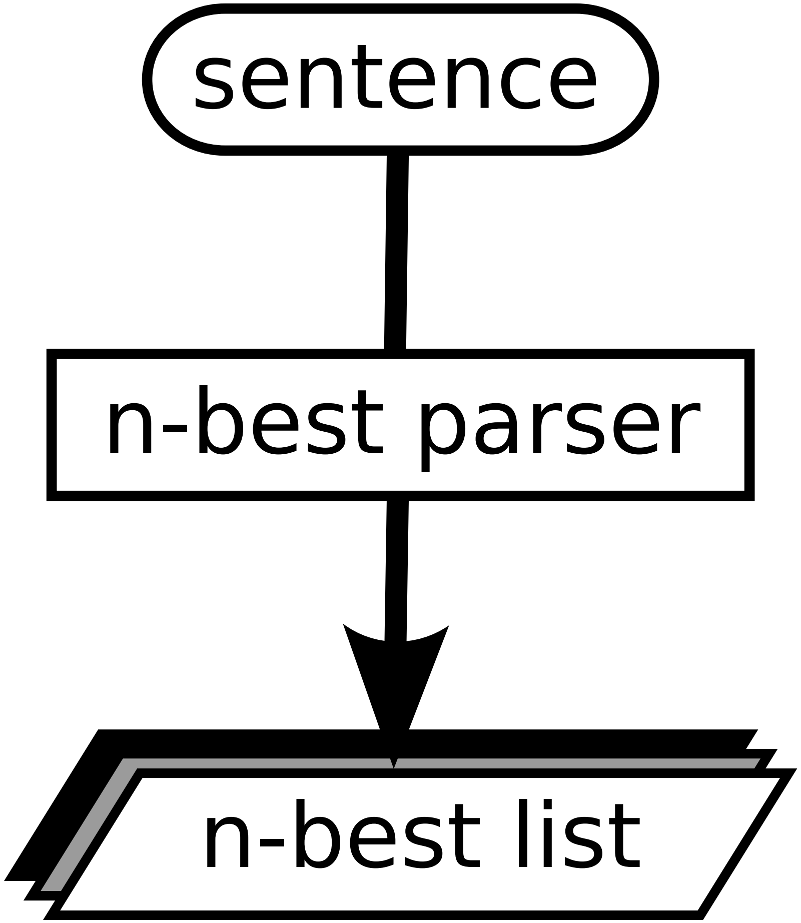
Corpus	# words	# sentences	Parameters
WSJ	950,028	39,832	$\sim$ 2,200,000
BROWN	373,152	19,740	$\sim$ 1,300,000

- Number of parameters is a function of training data.

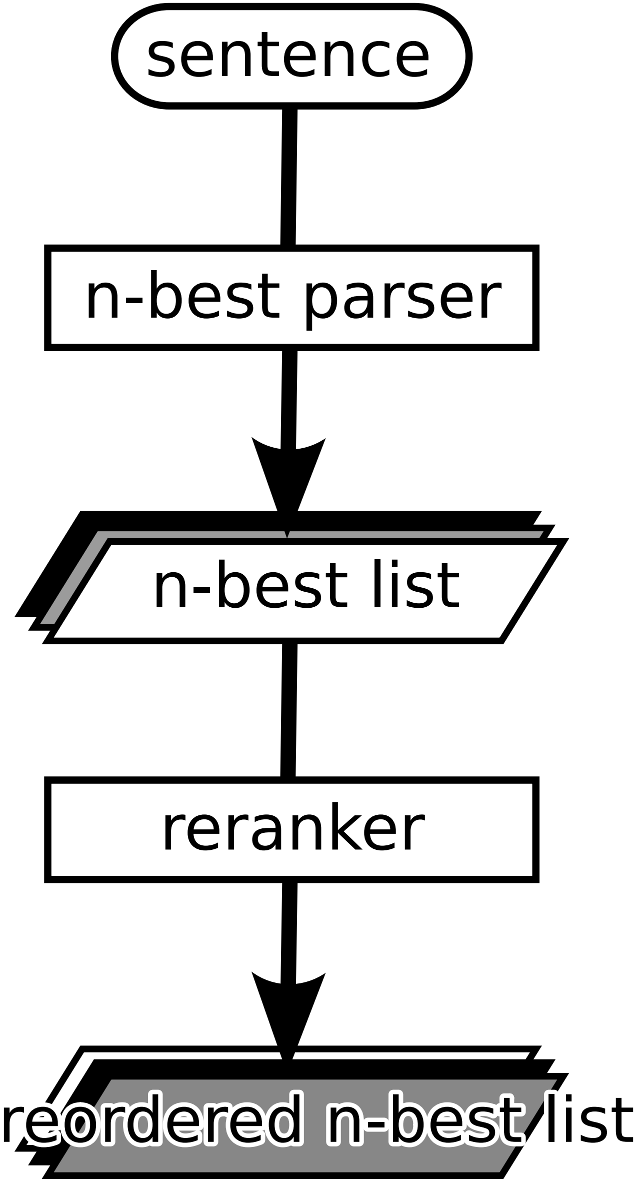
Parsing



n-best Parsing



Reranking Parsers



More Parameters

Reranking parser as in [Charniak and Johnson 2005]

- 14 feature schemas
- Extract features according to schemas then estimate feature weights

Corpus	Parser parameters	Reranker features
WSJ	$\sim$ 2,200,000	$\sim$ 1,300,000
BROWN	$\sim$ 1,300,000	$\sim$ 700,000

- Again, number of parameters is a function of training data.

Corpora and Domains

- WSJ: labeled news text, about 40,000 parses
- NANC: unlabeled news text, about 24 million sentences
- BROWN: labeled text from various domains, about 24,000 parses total

Corpora and Domains

- WSJ: labeled news text, about 40,000 parses
- NANC: unlabeled news text, about 24 million sentences
- BROWN: labeled text from various domains, about 24,000 parses total
  - Divisions as in [Bacchiani *et al.* 2006] (based on [Gildea 2001])
  - 19,740 train, 2,078 tune, 2,425 test
  - Treebanked sections are predominantly fiction
  - Each division of the corpus consists of sentences from all available genres

Self-Training

[McClosky, Charniak, and Johnson NAACL 2006]

- 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 parsed NANC data with WSJ training data

- Train a new model from the combined data

  train reranking parser on WSJ+NANC data

Overtrained?

Question: How does setting so many parameters from Wall Street Journal data affect parsing performance on the Brown corpus?

Previous Work

Training	Testing	f-measure	
		Gildea	Bacchiani
WSJ	WSJ	86.4	87.0
WSJ	BROWN	80.6	81.1
BROWN	BROWN	84.0	84.7
WSJ+BROWN	BROWN	84.3	85.6

[Gildea 2001], [Bacchiani *et al.* 2006]

Previous Work

Training	Testing	f-measure	
		Gildea	Bacchiani
WSJ	WSJ	86.4	87.0
WSJ	BROWN	80.6	81.1
BROWN	BROWN	84.0	84.7
WSJ+BROWN	BROWN	84.3	85.6

[Gildea 2001], [Bacchiani *et al.* 2006]

Previous Work

Training	Testing	f-measure	
		Gildea	Bacchiani
WSJ	WSJ	86.4	87.0
WSJ	BROWN	80.6	81.1
BROWN	BROWN	84.0	84.7
WSJ+BROWN	BROWN	84.3	85.6

[Gildea 2001], [Bacchiani *et al.* 2006]

Previous Work

Training	Testing	f-measure	
		Gildea	Bacchiani
WSJ	WSJ	86.4	87.0
WSJ	BROWN	80.6	81.1
BROWN	BROWN	84.0	84.7
WSJ+BROWN	BROWN	84.3	85.6

[Gildea 2001], [Bacchiani *et al.* 2006]

Previous Work

Training	Testing	f-measure	
		Gildea	Bacchiani
WSJ	WSJ	86.4	87.0
WSJ	BROWN	80.6	81.1
BROWN	BROWN	84.0	84.7
WSJ+BROWN	BROWN	84.3	85.6

[Gildea 2001], [Bacchiani *et al.* 2006]

Summary of findings

- The self-trained WSJ+NANC model does not appear to be overtrained.
- Both self-training and reranking techniques are fairly portable across domains.
- WSJ data with these techniques gives performance almost as good as actual BROWN corpus (does not work as well with more distant domains)

Overview

- Introduction and Previous Work
- Parser portability
- Parser adaptation
- Reranker portability
- Analysis
- Future Work and Conclusions

Parser Portability

Task: Use existing data/models from source domain to parse target domain.

Train: WSJ

Test: BROWN

Variables: Parser vs. reranker parser

Effect of self-training on NANC

Parser Portability

Train	Test	Parser	Reranking Parser
WSJ	WSJ	89.7	91.0
WSJ	BROWN	83.9	85.8

f-score on WSJ section 23 and BROWN development section

Parser Portability

Parsing model	Parser	Reranking Parser
WSJ baseline	83.9	85.8
WSJ+50k NANC	84.8	86.6
WSJ+250k NANC	85.7	87.2
WSJ+500k NANC	86.0	87.3
WSJ+1,000k NANC	86.2	87.3
WSJ+1,500k NANC	86.2	87.6
WSJ+2,500k NANC	86.4	87.7

f-score on BROWN development section

Parser Portability

Parsing model	Parser	Reranking Parser
WSJ baseline	83.9	85.8
WSJ+50k NANC	84.8	86.6
WSJ+250k NANC	85.7	87.2
WSJ+500k NANC	86.0	87.3
WSJ+1,000k NANC	86.2	87.3
WSJ+1,500k NANC	86.2	87.6
WSJ+2,500k NANC	86.4	87.7
BROWN baseline	86.4	87.7

f-score on BROWN development section

Parser Adaptation

Task: Use existing data/models from source domain with some target domain material to parse target domain.

Train: WSJ and/or BROWN

Test: BROWN

Variables: Number of self-trained sentences added

Amount of BROWN training data

Labeled In-domain Data

Parser model	Parser	Reranker
WSJ alone	83.9	85.8
BROWN alone	86.3	87.4
WSJ+BROWN	86.5	88.1

f-score on BROWN development section

Adding Self-Trained Data

Parser model	Parser	Reranker
WSJ alone	83.9	85.8
WSJ+2,500k NANC	86.4	87.7
BROWN alone	86.3	87.4
BROWN+250k NANC	86.8	88.1
WSJ+BROWN	86.5	88.1
WSJ+BROWN+250k NANC	86.8	88.1

f-score on BROWN development section

Reranker Portability

		Reranker	
Parser model	Parser alone	WSJ	BROWN
WSJ	82.9	85.2	85.2
WSJ+NANC	87.1	87.8	87.9
BROWN	86.7	88.2	88.4

f-scores on BROWN test section

Reranker Portability

		Reranker	
Parser model	Parser alone	WSJ	BROWN
WSJ	82.9	85.2	85.2
WSJ+NANC	87.1	87.8	87.9
BROWN	86.7	88.2	88.4

f-scores on BROWN test section

Reranker Portability

		Reranker	
Parser model	Parser alone	WSJ	BROWN
WSJ	82.9	85.2	85.2
WSJ+NANC	87.1	87.8	87.9
BROWN	86.7	88.2	88.4

f-scores on BROWN test section

Analysis Overview

- Oracle scores
- Parser agreement
- Per-category f-scores
- Factor analysis

Oracle Scores

Model	1-best	10-best	25-best	50-best
WSJ	82.6	88.9	90.7	91.9
WSJ+NANC	86.4	92.1	93.5	94.3
BROWN	86.3	92.0	93.3	94.2

f-score on BROWN development section

Oracle Scores

Model	1-best	10-best	25-best	50-best
WSJ	82.6	88.9	90.7	91.9
WSJ+NANC	86.4	92.1	93.5	94.3
BROWN	86.3	92.0	93.3	94.2

f-score on BROWN development section

Parser Agreement

Bracketing agreement f-score	88.03%
Complete match	44.92%
Average crossing brackets	0.94
POS Tagging agreement	94.85%

Agreement of parses from WSJ+NANC reranking parser with parses from BROWN reranking parser

Per-Category f-scores

Description	Size	BROWN	WSJ+NANC	Δ
Popular Lore	271	87.3	89.6	2.28
Letters	281	87.6	87.1	-0.45
General fiction	333	87.2	85.9	-1.29
Mystery	318	88.7	88.3	-0.45
Science fiction	76	87.7	88.8	1.17
Adventure	378	89.7	89.0	-0.64
Romance	338	88.0	86.6	-1.40
Humor	83	84.6	87.0	2.45

f-scores on BROWN development section

Factor Analysis

- Generalized linear model with binomial link with the predicted variable as

$$\text{BROWN } f\text{-score} > \text{WSJ+NANC } f\text{-score}$$

- Explanatory variables:
  - sentence length
  - number of prepositions
  - number of conjunctions
  - BROWN subcorpus ID

Factor Analysis

- Generalized linear model with binomial link with the predicted variable as

$$\text{BROWN } f\text{-score} > \text{WSJ+NANC } f\text{-score}$$

- Explanatory variables:
  - sentence length
  - number of prepositions ⋆
  - number of conjunctions
  - BROWN subcorpus ID ⋆

Per-Category f-scores

Description	Size	BROWN	WSJ+NANC	Δ
Popular Lore	271	87.3	89.6	2.28
Letters ★	281	87.6	87.1	-0.45
General fiction ★	333	87.2	85.9	-1.29
Mystery ★	318	88.7	88.3	-0.45
Science fiction	76	87.7	88.8	1.17
Adventure ★	378	89.7	89.0	-0.64
Romance ★	338	88.0	86.6	-1.40
Humor	83	84.6	87.0	2.45

f-scores on BROWN development section

Future Work

- Self-bridging corpora for harder domains
  - To parse BioMedical, self-train on biology text books
- Deeper comparison of BROWN and WSJ rerankers
- Parallel experiments for Switchboard and BioMedical domains
- Further analysis

Conclusions

- The self-trained WSJ+NANC model does not appear to be overtrained.
- Both self-training and reranking techniques are fairly portable across domains.
- WSJ data with these techniques gives performance almost as good as actual BROWN corpus (does not work as well with more distant domains)

Acknowledgements

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

We would like to thank the BLLIP team for their comments.

Questions?