University of Saarland

Seminar:Language Processing for Different Domains and Genres

Reranking and Self-Training for Parser Adaptation

Maria Sukhareva

Training Corpora

Statistical parsers are trained on treebanks.

The lack of corpora from different domains and genres lead to the parsers tuned to the particular corpora at expenses of other genres.

A statistical parser should aim at a broad coverage.

Parser adaptation

Parser adaptation is a process of leveraging existing labeled data from one domain and creating a parser capable of parsing a different domain.

Example: a parser trained on WSJ Penn Treebank, working for fiction texts.

The challenge for the statistical parser engineers:

Finding the resources required to construct reliable annotated training examples.

Therefore, there is need of a parser adaptation.

Available Corpora

WSJ (Wall Street Journal)

Cracking the Neanderthal Code

Our closest cousin's DNA may unravel secrets of evolution

MARKETPLACE | A9

A Mystery Multiplies Over a Math Prize

Why isn't a reclusive Russian pursuing a $1 million award?

SCIENCE JOURNAL | A9

The Peri
Of the 'F

How to keep t
on vacations

WEEKEND JO

THE WALL STREET JOURNAL.

© 2006 Dow Jones & Company. All Rights Reserved

WSJ.com   FRIDAY, JULY 21, 2006 - VOL. CCXLVIII NO. 17 - ★★ CONTINENTAL U.S. $1.00/HONOLULU $1.65

rce

; to Iran,
Builds
lentity

Leader Vows
ler Israeli Hit;
opaganda

'rofessional'

dropped 23 tons of
empt to kill him,
ssan Nasrallah ap-

Beirut, Lebanon,
Jerusalem

television yester-
tant pose. "Hezbol-
strike and retaken
Israelis, wagging

What's News–

Business and Finance

MICROSOFT'S NET FELL 24%, dragged down by legal expenses, as sales rose 16%. The software company also announced a stock-buyback program that includes a $20 billion tender offer. Google's earnings doubled and revenue surged 77% as the company controlled costs. SAP's profit climbed 43%.

■ **AMD said** net rose to $88.8 million on a 53% sales increase and predicted "seasonally strong" chip demand in the second half.

(Articles on Pages A3 and A11)

■ **Brocade's former CEO** and two other executives were charged with securities fraud in an alleged stock-options backdating scheme

World-Wide

■ **ISRAEL IS WARNING civilians in Southern Lebanon to leave the area.** In a possible prelude to a full-scale ground offensive, patrols were sent into the zone, but Hezbollah killed two soldiers and two Israeli helicopters collided, injuring four. Nasrallah appeared on TV to show he escaped Israeli bombs. The toll in Lebanon grew by at least 30, and food and fuel prices soared as supplies grew short. U.S. Marines landed in Beirut to assist evacuations for the first time since the 1984 barracks bombing. U.N. chief Annan called for an immediate cease-fire, further isolating the U.S., which wants to give Israel battlefield leeway. The crisis has been a political boon for Syria's Assad. (Column 1, Pages A3, A4)

Iran's Ahmadinejad wrote Germany's Merkel seeking advice on deal-

Burning Debate

As Emission Restrictions Loom, Texas Utility Bets Big on Coal

TXU to Build 11 New Plants
As Rivals Try Alternatives;
Global Warming Is Concern

Mr. Wilder Cites Demand

By Rebecca Smith

Top executives at many utility companies have reluctantly accepted that coal-fired power plants contribute to global warming, and they have begun planning for a more restrictive future.

Then there is C. John Wilder, chief executive of TXU Corp. The Dallas-based utility company is racing to build 11 big power plants in Texas that will burn pulverized coal. That process releases substantial amounts of carbon dioxide, the most worrisome of several heat-trapping

Power Source

Carbon-dioxide emissions from fossil-fuel combustion in the U.S. in 2004, by fuel type, in billions of tons:

*93.6% is from electricity generation

TXU Corp.'s carbon-dioxide emissions from coal are expected to rise:

	2005	2011
Emissions, millions of tons	54.9	133.5
Share of U.S. CO_2 emissions	2.7%	5.7%

Sources: Environmental Defense Fund; the company; Department of Energy

Brocad
2 Othe
In Opti

Authorities
As Backdat
Extend to O

Federal auth
criminal secu
against a forme
ecutive and tw
stock-options ba
ing they will t
widening scand
Prosecutors
the 43-year-old
Brocade Comm
of acting as a "c
dating option g
ing the options

By Charles For

Brown corpus

Many different genres of text.

(fiction and non-fiction)

North American News Corpus

British National Corpus

http://www.natcorp.ox.ac.uk/

Parser Improvement:Parse-Reranking

Paper 1: Reranking and Self-Training for Parser Adaptation

by David McClosky, Eugene Charniak, and Mark Johnson

Phase 1: “standard” generative parser generates n-best trees.

Phase 2: Discriminative Reranking. More detailed features are used to reorder the list of the best parsed trees.

Parser Improvement:Self-training

Unlabeled data are parsed.

Newly labeled data are treated as truth.

The data are added to the training corpus.

=>

is not normally effective:

The errors in the original model would be amplified in the new model.

Self-training Techniques for Parser Adaptation

McClosky et al. 2006:

Self-training requires labeled and unlabeled data.

The training data are of the same domain.

Labeled data: WSJ. Unlabeled data: NANC.

The experiment is performed on BROWN corpus.

Self-trained Parser:Briefly.

Unlabeled NANC sentences are parsed by reranking parser, producing 50 best sentences.

Labeled WSJ data, best parsed NANC data or reranked NANC data are mixed.

The parser is retrained. *(only the first stage is performed as the reranking didn't give significant results)*

Performance on Brown Corpus

NANC improves parsing performance from 83,9 % to 86,4% (n-best NANC parses)

As more NANC is added f-score approaches an asymptote.

NANC reduces data sparsity and fill in gaps in WSJ model.

The reranker adds 1-2 % to the f-score.

The results of the parser are similar to the results of the labeled train section of the BROWN corpus.

Incorporating In-Domain Data

Unlabeled In-Domain Data: WSJ trained reranking parser parsed BROWN data set, adding parsed BROWN sentences improved the performance for 2%.

Labeled In-Domain Data: Combination of the In-Domain Data with Out-of-Domain data:

BROWN model (as well as WSJ-combined) benefit only from a small amount of NANC sentences (250k)

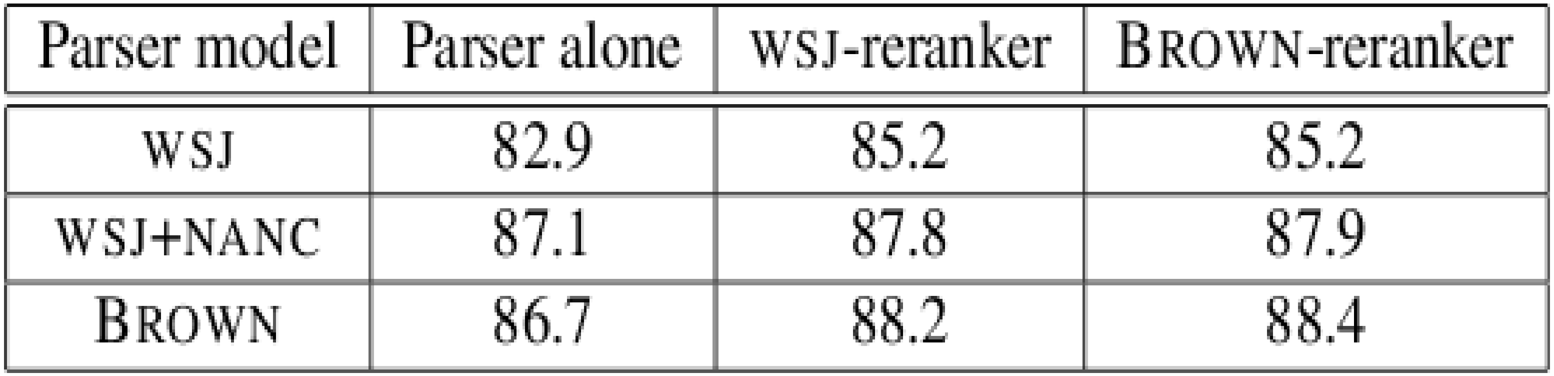
Tuning the parser back-off parameters on it.

Reranker Portability

The WSJ-trained reranker is portable to the BROWN fiction domain.

Applying the WSJ model to Switchboard corpus, showed the low performance of the parser but orthogonal benefit from self-training and reranking.

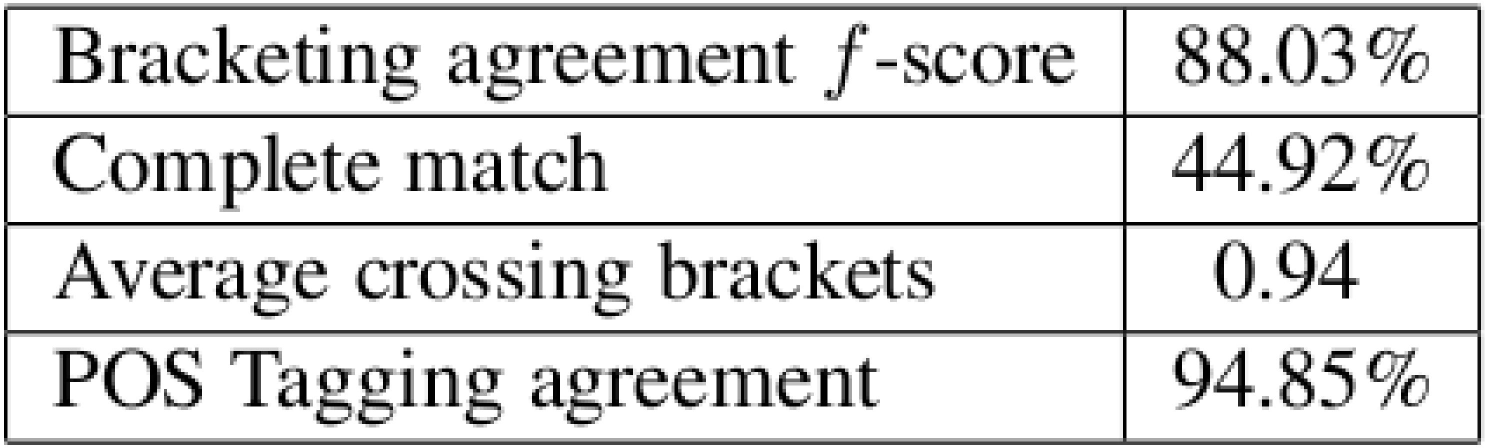
The BROWN reranker does not have a significant improvement over WSJ-reranker.



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

Parser Agreement

The output of the WSJ+NANC-trained and BROWN-trained reranking parser has a fairly high agreement.



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

Experiment with BNC

Adapting WSJ-Trained Parsers to the British National Corpus Using In-Domain Self-Training by Foster et al.

1000 BNC sentences are manually annotated.

(gold standard)

500 sentences are in the development set. 500 sentences are in the test set.

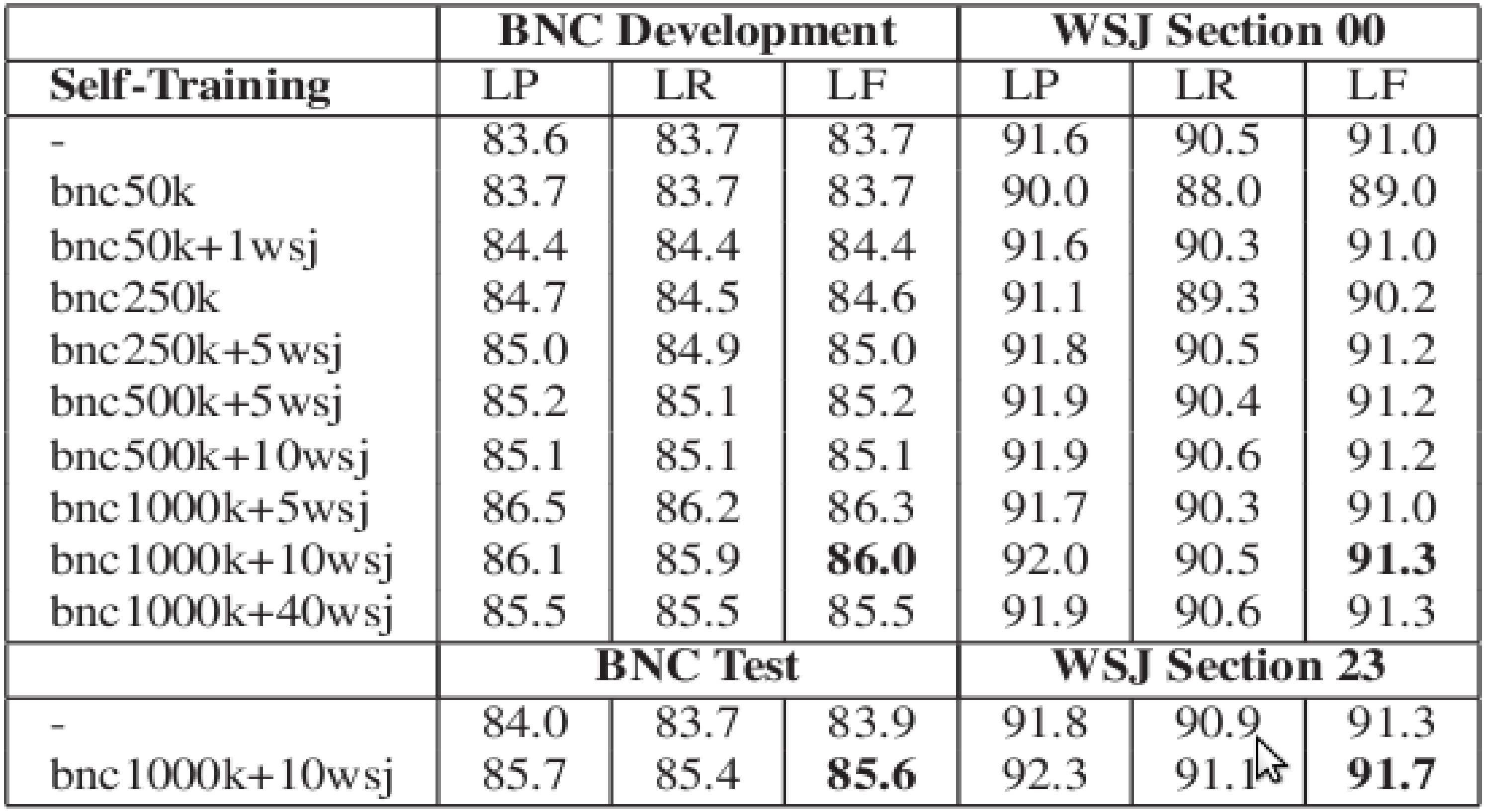
The parser is retrained on WSJ and its own parses of BNC sentences.

The combinations are tested on WSJ and on the development set of BNC.

The result: Parseval labeled bracketing f-score is 91.7 % on WSJ (S23) and 85,6% on BNC.

Self-Training Experiment with BNC

Retrain the first-stage of generative statistical parser of Charniak and Johnson using combinations of BNC.



	BNC Development			WSJ Section 00		
Self-Training	LP	LR	LF	LP	LR	LF
-	83.6	83.7	83.7	91.6	90.5	91.0
bnc50k	83.7	83.7	83.7	90.0	88.0	89.0
bnc50k+1wsj	84.4	84.4	84.4	91.6	90.3	91.0
bnc250k	84.7	84.5	84.6	91.1	89.3	90.2
bnc250k+5wsj	85.0	84.9	85.0	91.8	90.5	91.2
bnc500k+5wsj	85.2	85.1	85.2	91.9	90.4	91.2
bnc500k+10wsj	85.1	85.1	85.1	91.9	90.6	91.2
bnc1000k+5wsj	86.5	86.2	86.3	91.7	90.3	91.0
bnc1000k+10wsj	86.1	85.9	**86.0**	92.0	90.5	**91.3**
bnc1000k+40wsj	85.5	85.5	85.5	91.9	90.6	91.3
	BNC Test			**WSJ Section 23**		
-	84.0	83.7	83.9	91.8	90.9	91.3
bnc1000k+10wsj	85.7	85.4	**85.6**	92.3	91.1	**91.7**

Overview

	Seed Data	Training data	Test Data	Parser	Reranker
McClosky et al. (2006)	WSJ	----	BROWN	83,9%	85,8%
	WSJ+BROWN	----	BROWN	86,5%	88,1%
	WSJ+BROWN	NANC250k	BROWN	86,8%	88,1%
	BROWN	NANC250k	BROWN	86,8%	88,1%
Foster et al. (2007)	WSJ	BNC1000k	BNC		85,6%
	WSJ	BNC1000k	WSJ23		91,7%

Conclusion

Rerankers and self-trained combined models work well across domains.

The training of reranker on out-of-domain parses achieves approximately the same result as training on in-domain parses.

Corpora differences affect the result, as well as the domain proximity.

Self-training on in-domain data can be used for parser adaptation.