Faster Parsing by Supertagger Adaptation

Jonathan K. Kummerfeld[a] Jessika Roesner[b] Tim Dawborn[a]
James Haggerty[a] James R. Curran[a*] Stephen Clark[c*]

University of Sydney[a] University of Texas at Austin[b]
University of Cambridge[c]

james@it.usyd.edu.au[a*] stephen.clark@cl.cam.ac.uk[c*]

ACL 2010

We Need Faster Parsers

Syntactic information is crucial for many tasks in NLP, such as QA and MT, but parsers are slow:

- State-of-the-art, typically < 1 sentence / sec
- Fastest state-of-the-art, < 50 sentences / sec

Far too slow to process the data available:

- $>$ 1,000,000,000,000 words of English, Web1T
- More coming

Combinatory Categorial Grammar

I ate pizza

Combinatory Categorial Grammar

$$
\dfrac{\dfrac{\text{I}}{NP} \quad \dfrac{\dfrac{\text{ate}}{(S\backslash NP)/NP} \quad \dfrac{\text{pizza}}{NP}}{S\backslash NP}>}{S}<
$$

Combinatory Categorial Grammar

$$\dfrac{\dfrac{\dfrac{\text{I}}{NP} \quad \dfrac{\dfrac{\text{ate}}{(S\backslash NP)/NP} \quad \dfrac{\text{pizza}}{NP}}{S\backslash NP}>}{S}<}{}$$

Combinatory Categorial Grammar

$$\begin{array}{ccc}
\text{I} & \text{ate} & \text{pizza} \\
\overline{NP} & \overline{(S\backslash NP)/NP} & \overline{NP} \\
 & \multicolumn{2}{c}{\underline{\hspace{8em}}{>}} \\
 & \multicolumn{2}{c}{S\backslash NP} \\
\multicolumn{3}{c}{\underline{\hspace{12em}}{<}} \\
\multicolumn{3}{c}{S}
\end{array}$$

Categories Encode Rich Lexical Information

I ate pizza with cutlery

I ate pizza with anchovies

Categories Encode Rich Lexical Information

$$
\begin{array}{ccccc}
\text{I} & \text{ate} & \text{pizza} & \text{with} & \text{cutlery}\\
\hline
NP & (S\backslash NP)/NP & NP & ((S\backslash NP)\backslash(S\backslash NP))/NP & NP\\
\cline{2-3}\cline{4-5}
 & \multicolumn{2}{c}{S\backslash NP \quad >} & \multicolumn{2}{c}{(S\backslash NP)\backslash(S\backslash NP) \quad >}\\
\cline{2-5}
 & \multicolumn{4}{c}{S\backslash NP \quad <}\\
\hline
\multicolumn{5}{c}{S \quad <}
\end{array}
$$

$$
\begin{array}{ccccc}
\text{I} & \text{ate} & \text{pizza} & \text{with} & \text{anchovies}\\
\hline
NP & (S\backslash NP)/NP & NP & (NP\backslash NP)/NP & NP\\
\cline{4-5}
 & & & \multicolumn{2}{c}{NP\backslash NP \quad >}\\
\cline{3-5}
 & & \multicolumn{3}{c}{NP \quad <}\\
\cline{2-5}
 & \multicolumn{4}{c}{S\backslash NP \quad >}\\
\hline
\multicolumn{5}{c}{S \quad <}
\end{array}
$$

Categories Encode Rich Lexical Information

$$
\begin{array}{ccccc}
\text{I} & \text{ate} & \text{pizza} & \text{with} & \text{cutlery} \\
\hline
NP & (S\backslash NP)/NP & NP & ((S\backslash NP)\backslash(S\backslash NP))/NP & NP \\
 & \multicolumn{2}{c}{\overline{\quad S\backslash NP \quad}^{>}} & \multicolumn{2}{c}{\overline{\quad (S\backslash NP)\backslash(S\backslash NP) \quad}^{>}} \\
 & \multicolumn{4}{c}{\overline{\quad S\backslash NP \quad}^{<}} \\
\multicolumn{5}{c}{\overline{\quad S \quad}^{<}}
\end{array}
$$

$$
\begin{array}{ccccc}
\text{I} & \text{ate} & \text{pizza} & \text{with} & \text{anchovies} \\
\hline
NP & (S\backslash NP)/NP & NP & (NP\backslash NP)/NP & NP \\
 & & & \multicolumn{2}{c}{\overline{\quad NP\backslash NP \quad}^{>}} \\
 & & \multicolumn{3}{c}{\overline{\quad NP \quad}^{<}} \\
 & \multicolumn{4}{c}{\overline{\quad S\backslash NP \quad}^{>}} \\
\multicolumn{5}{c}{\overline{\quad S \quad}^{<}}
\end{array}
$$

Taggers Constrain the Search Space

Divide parsing into two tasks, where for n words, each with k tags:

- Tagging – $O(nk^3)$
- Parsing – $O(n^5k^2)$

The tagger considers a set of 429 categories

Ideal World – Perfect Supertagging

Previously	,	watch	imports	were	denied
S/S	,	N/N	N	$(S[dcl]\backslash NP)/(S[pss]\backslash NP)$	$(S[pss]\backslash NP)/NP$

such	duty-free	treatment
NP/NP	N/N	N

Real World – Around 92% Accuracy

Previously	,	watch	imports	were	denied
S/S	,	**N**	N	$(S[dcl]\backslash NP)/(S[pss]\backslash NP)$	$(S[pss]\backslash NP)/NP$

such	duty-free	treatment
NP/NP	N/N	N

Real World – Multitagging Prevents Coverage Loss

Previously	,	watch	imports	were	denied
S/S	**,**	**N**	**N**	$(S[dcl]\backslash NP)/(S[pss]\backslash NP)$	$(S[pss]\backslash NP)/NP$
N				**$(S[dcl]\backslash NP)/NP$**	$S[pss]\backslash NP$
$S[adj]\backslash NP$				$(S[dcl]\backslash NP)/(S[adj]\backslash NP)$	$(S[pt]\backslash NP)/NP$
					$(S[dcl]\backslash NP)/NP$

such	duty-free	treatment
NP/NP	**N/N**	**N**
$((S\backslash NP)\backslash(S\backslash NP))/((S\backslash NP)\backslash(S\backslash NP))$		
$(N/N)/(N/N)$		
N/N		
$(NP/NP)/(NP/NP)$		

Adaptive Supertagging

Previously	,	watch	imports	were	denied
S/S	**,**	**N**	**N**	**(S[dcl]\NP)/NP**	$(S[pss]\backslash NP)/NP$
N				$(S[dcl]\backslash NP)/(S[adj]\backslash NP)$	$S[pss]\backslash NP$
					(S[dcl]\NP)/NP

such	duty-free	treatment
NP/NP	**N/N**	**N**
N/N		

Leave out Categories the Parser will not use

New Task:

- Target output – the categories the baseline would use
- To create target output to train on, run the parser
- 4 million sentences (limited by volume of WSJ in NANC)

Previous Work

Semi-supervised training has been used to improve parsing accuracy:

- Co-training, Sarkar (2001)
- Reranking, McClosky et al. (2006)
- Pipeline Iteration, Hollingshead and Roark (2007)

For efficiency improvement, van Noord (2009)

- Observe the parsing process for many sentences
- Only follow parsing steps observed for the training set

Baseline results for the C&C parser and supertagger

- Parse a large set of unannotated data
- Retrain the supertagger, using the parser annotated sentences
- Four discriminative training methods, GIS, BFGS, AP, MIRA

	F-score				Speed (sents / sec)			
NANC Data	0k	40k	400k	4m	0k	40k	400k	4m
Base	85.46				39.6			
GIS								
BFGS								
MIRA								

Speed increases of up to 85%

- Parse a large set of unannotated data
- Retrain the supertagger, using the parser annotated sentences
- Four discriminative training methods, GIS, BFGS, AP, MIRA

	F-score				Speed (sents / sec)			
NANC Data	0k	40k	400k	4m	0k	40k	400k	4m
Base	85.46				39.6			
GIS	85.44	85.46	85.58	85.62	37.4	44.1	51.3	54.1
BFGS	85.45	85.51	85.57	85.68	39.8	49.6	71.8	60.0
MIRA	85.44	85.40	85.38	85.42	34.1	44.8	60.2	73.3

Speed increases of up to 85%

- Parse a large set of unannotated data
- Retrain the supertagger, using the parser annotated sentences
- Four discriminative training methods, GIS, BFGS, AP, MIRA

	F-score				Speed (sents / sec)			
NANC Data	0k	40k	400k	4m	0k	40k	400k	4m
Base	85.46				39.6			
GIS	85.44	85.46	85.58	85.62	37.4	44.1	51.3	54.1
BFGS	85.45	85.51	85.57	85.68	39.8	49.6	71.8	60.0
MIRA	85.44	85.40	85.38	85.42	34.1	44.8	60.2	73.3

Speed increases of up to 85%

- Parse a large set of unannotated data
- Retrain the supertagger, using the parser annotated sentences
- Four discriminative training methods, GIS, BFGS, AP, MIRA

	F-score				Speed (sents / sec)			
NANC Data	0k	40k	400k	4m	0k	40k	400k	4m
Base	85.46				39.6			
GIS	85.44	85.46	85.58	85.62	37.4	44.1	51.3	54.1
BFGS	85.45	85.51	85.57	85.68	39.8	49.6	71.8	60.0
MIRA	85.44	85.40	85.38	85.42	34.1	44.8	60.2	73.3

Adjust Ambiguity to Trade Speed for Accuracy

- We are increasing speed by decreasing ambiguity
- Adjust system parameters to return ambiguity to baseline levels

	F-score				Speed (sents / sec)			
NANC Data	0k	40k	400k	4m	0k	40k	400k	4m
Baseline	85.46				39.6			
GIS	85.36	85.47	85.84	85.87	39.1	41.4	41.7	42.6
BFGS	85.45	85.55	85.64	85.98	39.5	43.7	43.9	42.7
Perceptron	85.28	85.39	85.64	-	45.9	48.0	45.2	-
MIRA	85.47	85.45	85.55	85.84	37.7	41.4	41.4	42.9

Adjust Ambiguity to Trade Speed for Accuracy

- We are increasing speed by decreasing ambiguity
- Adjust system parameters to return ambiguity to baseline levels

	F-score				Speed (sents / sec)			
NANC Data	0k	40k	400k	4m	0k	40k	400k	4m
Baseline	85.46				39.6			
GIS	85.36	85.47	85.84	85.87	39.1	41.4	41.7	42.6
BFGS	85.45	85.55	85.64	**85.98**	39.5	43.7	43.9	42.7
Perceptron	85.28	85.39	85.64	-	45.9	48.0	45.2	-
MIRA	85.47	85.45	85.55	85.84	37.7	41.4	41.4	42.9

Adjust Ambiguity to Trade Speed for Accuracy

- We are increasing speed by decreasing ambiguity
- Adjust system parameters to return ambiguity to baseline levels

	F-score				Speed (sents / sec)			
NANC Data	0k	40k	400k	4m	0k	40k	400k	4m
Baseline	85.46				39.6			
GIS	85.36	85.47	85.84	85.87	39.1	41.4	41.7	42.6
BFGS	85.45	85.55	85.64	**85.98**	39.5	43.7	43.9	**42.7**
Perceptron	85.28	85.39	85.64	-	45.9	48.0	45.2	-
MIRA	85.47	85.45	85.55	85.84	37.7	41.4	41.4	42.9

The Approach Works on Multiple Domains

Training Corpus	Speed (sents / sec) News	Wiki	Bio
Baseline	39.6	50.9	35.1
News	73.3	83.9	60.3
Wiki	62.4	73.9	58.7
Bio	66.2	90.4	59.3

The Approach Works on Multiple Domains

Training Corpus	Speed (sents / sec) News	Wiki	Bio
Baseline	39.6	50.9	35.1
News	73.3	83.9	60.3
Wiki	62.4	73.9	58.7
Bio	66.2	90.4	59.3

Adaptation is Domain Specific

Training Corpus	F-score News	Wiki	Bio
Baseline	85.46	80.8	75.0
News	85.84	80.1	75.2
Wiki	85.02	81.7	75.8
Bio	84.95	80.6	76.1

Adaptation is Domain Specific

Training Corpus	F-score News	Wiki	Bio
Baseline	85.46	80.8	75.0
News	85.84	80.1	75.2
Wiki	85.02	81.7	75.8
Bio	84.95	80.6	76.1

The Parsing Process

Previously , watch imports were denied

such duty-free treatment

Pass 1 – Minimum Ambiguity

Previously	,	watch	imports	were	denied
S/S	,	N	N	$(S[dcl]\backslash NP)/(S[pss]\backslash NP)$	$(S[pss]\backslash NP)/NP$
					$S[pss]\backslash NP$
					$(S[pt]\backslash NP)/NP$

such	duty-free	treatment
NP/NP	N/N	N

Pass 2 – More Ambiguity

Previously	,	watch	imports	were	denied
S/S	,	N	N	$(S[dcl]\backslash NP)/(S[pss]\backslash NP)$	$(S[pss]\backslash NP)/NP$
					$S[pss]\backslash NP$
					$(S[pt]\backslash NP)/NP$
					$(S[dcl]\backslash NP)/NP$

such	duty-free	treatment
NP/NP	N/N	N
$((S\backslash NP)\backslash(S\backslash NP))/((S\backslash NP)\backslash(S\backslash NP))$		
$(N/N)/(N/N)$		

Pass 3 – Further Ambiguity

Previously	,	watch	imports	were	denied
S/S	,	N	N	$(S[dcl]\backslash NP)/(S[pss]\backslash NP)$	$(S[pss]\backslash NP)/NP$
					$S[pss]\backslash NP$
					$(S[pt]\backslash NP)/NP$
					$(S[dcl]\backslash NP)/NP$

such	duty-free	treatment
NP/NP	N/N	N
$((S\backslash NP)\backslash(S\backslash NP))/((S\backslash NP)\backslash(S\backslash NP))$		
$(N/N)/(N/N)$		
N/N		

Pass 4 – Even More Ambiguity, Parsed at last!

Previously	,	watch	imports	were	denied
S/S	**,**	**N**	**N**	$(S[dcl]\backslash NP)/(S[pss]\backslash NP)$	$(S[pss]\backslash NP)/NP$
N				**(S[dcl]\NP)/NP**	$S[pss]\backslash NP$
$S[adj]\backslash NP$				$(S[dcl]\backslash NP)/(S[adj]\backslash NP)$	$(S[pt]\backslash NP)/NP$
					(S[dcl]\NP)/NP

such	duty-free	treatment
NP/NP	**N/N**	**N**
$((S\backslash NP)\backslash (S\backslash NP))/((S\backslash NP)\backslash (S\backslash NP))$		
$(N/N)/(N/N)$		
N/N		
$(NP/NP)/(NP/NP)$		

Parsing Sentences Earlier and/or With Lower Ambiguity

		Total Time Change (s)		
Pass	Ambiguity	Short	Medium	Long
Earlier	$<$	-1.1	-29	-26
	$=$	-0.095	-1.3	-0.44
	$>$	-0.40	-1.3	-0.31
Same	$<$	-2.8	-20	-30
	$=$	-0.28	0.30	0.44
	$>$	-0.037	0.34	0.099
Later	$<$	0.039	1.1	-2.5
	$=$	0.0019	0.0053	0.0
	$>$	-3.4e-5	0.033	0.16

Parsing Sentences Earlier and/or With Lower Ambiguity

		Total Time Change (s)		
Pass	Ambiguity	Short	Medium	Long
Earlier	<	-1.1	-29	-26
	=		-1.3	
	>		-1.3	
Same	<	-2.8	-20	-30
	=			
	>			
Later	<		1.1	-2.5
	=			
	>			

Improvement Relies on Parser Annotated Data

Annotation method	Cat. Acc.	F-score
Baseline	96.34	85.46
Parser	96.46	85.55
One-best super	95.94	85.24
Multi-tagger a	95.91	84.98
Multi-tagger b	96.00	84.99

Conclusion

Metric	Base	Adaptive	Ratio
Ambiguity	1.267	1.126	0.89
Newswire Accuracy			
Cat. Acc. (%)	96.34	95.18	n/a
F-score (%)	85.46	85.42	n/a
Speed			
WSJ (sents / sec)	39.6	73.3	1.85
Wikipedia (sents / sec)	50.9	83.9	1.65
Medline (sents / sec)	35.1	60.3	1.72

Conclusion

Adaptive training improves parsing speed while retaining accuracy

- Works across multiple domains
- No extra manually annotated data
- Enables accuracy gains while retaining high speed

Acknowledgements

- Australian Research Council, Discovery Grants DP0665973 and DP1097291
- Capital Markets Cooperative Research Centre
- Johns Hopkins University, CLSP Summer Workshop
- National Science Foundation, Grant Number IIS-0833652.

Acknowledgements

Model	Cat. Acc. (%)	F-score (%)	Speed (sents/sec)
Baseline	96.51	85.20	39.6
GIS, 4,000k NANC	96.83	**85.95**	42.6
BFGS, 4,000k NANC	96.91	85.90	42.7
MIRA, 4,000k NANC	96.84	85.79	42.9

Table: Evaluation of top models on Section 23 of CCGbank.

Acknowledgements

Corpus	Speed (sents / sec)		
Sent length	5-20	21-40	41-250
News	242	44.8	8.24
Wiki	224	42.0	6.10
Bio	268	41.5	6.48

Table: Cross-corpus speed for the baseline model on data sets balanced on sentence length.

Acknowledgements

Train Corpus	F-score
Rimell and Clark (2009)	81.5
Baseline	80.7
CCGbank + Genia	81.5
+ Newswire	81.9
+ Wikipedia	82.2
+ Biomedical	81.7
+ Bio with R&C models	82.3

Table: Performance comparison for models using extra gold standard biomedical data.

Acknowledgements