Faster Parsing and Supertagging Model Estimation

Jonathan K. Kummerfeld\textsuperscript{a} \hspace{1cm} James R. Curran\textsuperscript{a} \hspace{1cm} Jessika Roesner\textsuperscript{b}

School of Information Technologies\textsuperscript{a} \hspace{1cm} Department of Computer Science\textsuperscript{b}
University of Sydney \hspace{1cm} University of Texas at Austin
Australia \hspace{1cm} USA

\{jkum0593,james\}@it.usyd.edu.au \hspace{1cm} jessi@mail.utexas.edu

ALTW 2009
Motivation – Parsing

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

- State-of-the-art, usually < 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 online
- More coming
Tagging and Parsing

One claims he is pro–choice
Part of Speech Tagging

One claims he is pro-choice

NN VBZ PRP VBZ JJ
Combinatory Categorial Grammar (CCG) – Supertagging

One claims he is pro–choice

\[
\overline{N} \rightarrow (S\backslash NP)/S \\
\overline{NP} \rightarrow (S\backslash NP)/(S\backslash NP)
\]
Combinatory Categorial Grammar (CCG) – Parsing

One claims he is pro–choice

\[
\begin{array}{cccc}
N & (S\backslash NP)/S & \bar{NP} & (S\backslash NP)/(S\backslash NP) & S\backslash NP \\
\bar{NP} & S\backslash NP & S & S \\
S & S\backslash NP & S
\end{array}
\]
Supertagging Ambiguity

I ate pizza with cutlery

I ate pizza with anchovies
Supertagging Ambiguity

\[
\begin{align*}
I & \\
NP & \rightarrow (S \backslash NP)/NP \\
S \backslash NP & \rightarrow \bullet
\end{align*}
\]

\[
\begin{align*}
I & \\
NP & \rightarrow (S \backslash NP)/NP \\
NP & \rightarrow (S \backslash NP)/NP \\
S \backslash NP & \rightarrow \bullet
\end{align*}
\]

\[
\begin{align*}
I & \\
NP & \rightarrow (S \backslash NP)/NP \\
NP & \rightarrow (NP \backslash NP)/NP \\
NP & \rightarrow \bullet
\end{align*}
\]

\[
\begin{align*}
I & \\
NP & \rightarrow (S \backslash NP)/NP \\
NP & \rightarrow NP \backslash NP \\
NP & \rightarrow \bullet
\end{align*}
\]

\[
\begin{align*}
I & \\
NP & \rightarrow NP \backslash NP \\
NP & \rightarrow \bullet
\end{align*}
\]

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation
Supertagging Ambiguity

\[
\begin{align*}
\text{I ate pizza with cutlery} \\
NP \ (S\NP)/NP \ NP \ ((S\NP)\,(S\NP))/NP \ NP \\
S\NP \ (S\NP) \\
S
\end{align*}
\]
Motivation – Parsing

The key idea behind the speed of the fastest parsers today is to shift work from parsing to tagging:
For $n$ words, each with $k$ tags

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

- Provide fewer tags, but still include the tags the parser would have used anyway

Implementation

- Perceptron Algorithms
- Parallelisation

Results

- Modified rule usage
- Training data type and volume
- Algorithm comparison
- Feature extension
### Ideal World

One claims he is pro-choice.

| $N$ | $(S\backslash NP)/S$ | $NP$ | $(S\backslash NP)/(S\backslash NP)$ | $S\backslash NP$ | pro-choice |
Current World – Problem

One claims he is pro-choice

\[
\begin{array}{c}
N \\
(S\setminus NP)/NP \\
NP \\
(S\setminus NP)/(S\setminus NP) \\
S\setminus NP
\end{array}
\]
## Current World – Solution

<table>
<thead>
<tr>
<th>One</th>
<th>claims</th>
<th>he</th>
<th>is</th>
<th>pro – choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N/N$</td>
<td>$(S\backslash NP)/NP$</td>
<td>$NP$</td>
<td>$(S\backslash NP)/(S\backslash NP)$</td>
<td>$S\backslash NP$</td>
</tr>
<tr>
<td>$N$</td>
<td>$N$</td>
<td>$(S\backslash NP)/NP$</td>
<td>$(S\backslash NP)/(S\backslash NP)$</td>
<td>$(S\backslash NP)/(S\backslash NP)$</td>
</tr>
<tr>
<td>$(S/S)/(S/S)$</td>
<td>$(S\backslash NP)/NP$</td>
<td>$(S\backslash NP)/(S\backslash NP)$</td>
<td>$(S\backslash NP)/S$</td>
<td>$(S\backslash NP)/PP$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$N$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$(S\backslash NP)/NP$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$N/N$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$(S\backslash NP)/(S\backslash NP)$</td>
</tr>
</tbody>
</table>
Adaptive Supertagging

One claims he is pro–choice

\[
\begin{align*}
\text{One} & \quad \text{claims} & \quad \text{he} & \quad \text{is} & \quad \text{pro–choice} \\
N/N & \quad N & \quad NP & \quad (S\backslash NP)/(S\backslash NP) & \quad S\backslash NP \\
&S\backslash NP/\quad NP & \quad (S\backslash NP)/PP & \quad (S\backslash NP)/NP \\
&S\backslash NP/(S\backslash NP) & \quad (S\backslash NP)/NP & \quad N/N
\end{align*}
\]

How do we teach the supertagger to produce these tags?
Use the parser!
Outline

Core Idea

- Provide fewer tags, but still include the tags the parser would have used anyway

Implementation

- Perceptron Algorithms
- Parallelisation

Results

- Modified rule usage
- Training data type and volume
- Algorithm comparison
- Feature extension
## Implementation

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial System</th>
<th>Additions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Feature Extraction</td>
<td>3 Types</td>
<td>+9 Types</td>
</tr>
<tr>
<td></td>
<td>Single thread</td>
<td>Parallel</td>
</tr>
<tr>
<td>Parameter Estimation</td>
<td>BFGS, GIS</td>
<td>AP, MIRA</td>
</tr>
<tr>
<td></td>
<td>Single thread</td>
<td>Parallel</td>
</tr>
</tbody>
</table>

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation
Implementation – Extra Constraint

Added a constraint that only allows Backward Composition to occur if both children are type raised
## Implementation – AP and MIRA

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Training Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40k</td>
</tr>
<tr>
<td>GIS</td>
<td>7,200</td>
</tr>
<tr>
<td>BFGS</td>
<td>6,300</td>
</tr>
<tr>
<td>AP</td>
<td>76</td>
</tr>
<tr>
<td>MIRA</td>
<td>96</td>
</tr>
</tbody>
</table>

* Kummerfeld et al. Faster Parsing and Supertagging Model Estimation altw 2009
Implementation – Initial System

- Data
  - Extract features
  - Contexts
  - Estimate weights
  - Model

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation
Implementation – Parallelised

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation
Implementation – Parallelised Weight Estimation

Figure: Information flow for parallel model estimation
Core Idea

- Provide fewer tags, but still include the tags the parser would have used anyway

Implementation

- Perceptron Algorithms
- Parallelisation

Results

- Modified rule usage
- Training data type and volume
- Algorithm comparison
- Feature extension
## Extra Constraint on Rule Application

<table>
<thead>
<tr>
<th>Parser</th>
<th>F-score (%)</th>
<th>Speed (sent / sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&amp;C 1.02</td>
<td>83.22</td>
<td>31.7</td>
</tr>
<tr>
<td>Modified</td>
<td>83.41</td>
<td>47.8</td>
</tr>
</tbody>
</table>

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation

ALTW 2009
Plan

- Acquire a large set of unannotated data – Wikipedia
- Parse the corpus
- Retrain the supertagger, using the parsed sentences

Variations

- Amount of data
- Estimation algorithms
- Feature set
Training Data Type and Volume

Figure: Evaluation on the Wall Street Journal

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation

Baseline
Figure: Evaluation on the Wall Street Journal

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation
Training Data Type and Volume

Figure: Evaluation on the Wall Street Journal
Training Data Type and Volume

Figure: Evaluation on Wikipedia

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation
Algorithm Comparison

Figure: Evaluation on the Wall Street Journal

Kummerfeld et al.
Faster Parsing and Supertagging Model Estimation
Algorithm Comparison

Figure: Evaluation on Wikipedia

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation
Feature Extension

Figure: Evaluation on the Wall Street Journal

Kummerfeld et al. Faster Parsing and Supertagging Model Estimation
Feature Extension

Figure: Evaluation on Wikipedia
Future Work

- Other domains
- Expanded training sets
- Co-training
- Online learning
Conclusion

Improved training:
- Enabled access to more text
- Constructed an effective source of more text

Improved parsing speed:
- Added an extra constraint on rule usage
- Trained models that are adapted to the parser

Improved parsing accuracy:
- Constructed statistical models using more evidence
- Expanded the set of statistical features
# Conclusion

<table>
<thead>
<tr>
<th>Metric</th>
<th>Initial</th>
<th>Final</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>40k</td>
<td>80k</td>
<td>2</td>
</tr>
<tr>
<td>Time (secs)</td>
<td>6,300</td>
<td>160</td>
<td>1/40</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-score (%)</td>
<td>83.22</td>
<td>83.79</td>
<td>n/a</td>
</tr>
<tr>
<td><strong>Speed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WSJ (sents / sec)</td>
<td>31.7</td>
<td>62.8</td>
<td>2.0</td>
</tr>
<tr>
<td>Wikipedia (sents / sec)</td>
<td>30.8</td>
<td>69.7</td>
<td>2.3</td>
</tr>
</tbody>
</table>
Acknowledgements

- Johns Hopkins University, CLSP Summer Workshop
- Capital Markets Cooperative Research Centre Limited