Weakly-Supervised Bayesian Learning of a CCG Supertagger

Dan Garrette, Chris Dyer, Jason Baldridge, Noah A. Smith
Type-Level Supervision
Type-Level Supervision

- Unannotated text
- Incomplete tag dictionary: word $\mapsto$ \{tags\}
Type-Level Supervision

Used for POS tagging for 20+ years

[Kupiec, 1992]
[Merialdo, 1994]
Type-Level Supervision

Good POS tagger performance even with low supervision

[Das & Petrov 2011]
[Garrette & Baldridge 2013]
[Garrette et al. 2013]
Combinatory Categorial Grammar (CCG)
CCG

Every word token is associated with a category

Categories combine to categories of constituents

[Steedman, 2000]
[Steedman and Baldridge, 2011]
CCG

s

np dogs

s\np sleep
POS vs. Supertags

the dog sleeps

np/n n s\np

the dog sleeps
Supertagging

Type-supervised learning for supertagging is much more difficult than for POS

Penn Treebank POS
48 tags

CCGBank Supertags
1,239 tags
CCG

The grammar formalism *itself* can be used to guide learning
CCG Supertagging
CCG Supertagging

- Sequence tagging problem, like POS-tagging
- Building block for grammatical parsing
Supertagging

“almost parsing”
Why Supertagging?

the lazy dog sleeps
Why Supertagging?
CCG Supertagging

the lazy dog sleeps
CCG Supertagging

the /np\n
lazy /n\n
dog /n\n
sleeps /s\np
sleeps

the lazy dog

CCG Supertagging

np/n n/n n

s

np

sleeps
CCG Supertagging

np/n the lazy dog n
Principle #1

Prefer Connections
Supertags vs. POS

universal, intrinsic grammar properties

all relationships must be learned
Principle #2

Prefer Simplicity
Prefer Simplicity

\[
\text{buy} := (s_b \backslash np)/np \quad \text{appears 342 times in CCGbank}
\]

\[
\text{e.g. “Opponents don’t buy such arguments.”}
\]

\[
\text{buy} := (((s_b \backslash np)/pp)/pp)/np \quad \text{appears once}
\]

“Tele-Communications agreed to buy half of Showtime Networks from Viacom for $225 million.”
Weighted Tag Grammar

\[ a \rightarrow \{s, np, n, \ldots\} \quad p_{\text{atom}}(a) \times p_{\text{term}} \]

\[ A \rightarrow B / B \quad \overline{p_{\text{term}}} \times p_{\text{fwd}} \times p_{\text{mod}} \]

\[ A \rightarrow B \backslash B \quad \overline{p_{\text{term}}} \times p_{\text{fwd}} \times p_{\text{mod}} \]

\[ A \rightarrow B \backslash C \quad \overline{p_{\text{term}}} \times p_{\text{fwd}} \times p_{\text{mod}} \]
CCG Supertagging

np

(np/(np/n))/n  n

n/n

the  lazy  dog
HMM Transition Prior

\[ P(t \rightarrow u) = \lambda \cdot P(u) + (1-\lambda) \cdot P(t \rightarrow u) \]

simple is good connecting is good
Type-Supervised Learning

unlabeled corpus

universal properties of the CCG formalism

tag dictionary

same as POS tagging
Training
Posterior Inference

Forward-Filter Backward-Sample (FFBS)

[Carter and Kohn, 1996]
Posterior Inference

Unlabeled Data
_________________
_________________
_________________
_________________

Tag Dictionary
___  :  __, __, __
___  :  __, __, __
___  :  __, __, __
___  :  __, __, __

the  lazy  dogs  wander
np/n  n/n  np  (s\np)/np

n
n/n
np/n
s\np
...

Unlabeled Data
_________________
_________________
_________________
<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>lazy</th>
<th>dogs</th>
<th>wander</th>
</tr>
</thead>
<tbody>
<tr>
<td>np/n</td>
<td>np/n</td>
<td>np</td>
<td>(s\np)/np</td>
<td>np/n</td>
</tr>
</tbody>
</table>

**Posterior Inference**

Diagram:

- Priors
- HMM

- the (np/n)
- lazy (n/n)
- dogs (np)
- wander (n)

...
Posterior Inference

Priors

HMM

the lazy dogs wander

\textit{np/n} \textit{n/n} \textit{np} \textit{(s\,np)/np} \textit{n} \textit{np/n} \textit{s\,np} …
Posterior Inference

Priors

HMM

the lazy dogs wander

np/n n/n np np np np
(s\np)/np

...
Posterior Inference

Priors

HMM

the lazy dogs wander

np/n n/n np (s\np)/np n/n s\np
Posterior Inference

Priors -> HMM

the np/n lazy np dogs np (s
p)/np wander np/n s
p

...
Experiments
Baldridge 2008

Use universal properties of CCG to initialize EM

- Simpler definition of category complexity
- No corpus-specific information
English Supertagging

<table>
<thead>
<tr>
<th>Tag Dictionary Pruning Cutoff</th>
<th>Baldridge '08</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>78</td>
<td>80</td>
</tr>
<tr>
<td>0.01</td>
<td>67</td>
<td>80</td>
</tr>
<tr>
<td>0.001</td>
<td>55</td>
<td>73</td>
</tr>
<tr>
<td>None</td>
<td>41</td>
<td>51</td>
</tr>
</tbody>
</table>
Chinese Supertagging

<table>
<thead>
<tr>
<th>Tag Dictionary Pruning Cutoff</th>
<th>Baldridge '08</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>66</td>
<td>69</td>
</tr>
<tr>
<td>0.01</td>
<td>43</td>
<td>62</td>
</tr>
<tr>
<td>0.001</td>
<td>33</td>
<td>56</td>
</tr>
<tr>
<td>none</td>
<td>28</td>
<td>49</td>
</tr>
</tbody>
</table>

Tagging Accuracy

- Baldridge '08
- Ours
Italian Supertagging

<table>
<thead>
<tr>
<th>Tag Dictionary Pruning Cutoff</th>
<th>Baldridge '08</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>54</td>
<td>53</td>
</tr>
<tr>
<td>0.01</td>
<td>36</td>
<td>47</td>
</tr>
<tr>
<td>0.001</td>
<td>33</td>
<td>45</td>
</tr>
<tr>
<td>None</td>
<td>32</td>
<td>46</td>
</tr>
</tbody>
</table>

Tagging Accuracy
Code Available

GitHub repository linked from my website
Conclusion

Combining annotation exploitation with universal grammatical knowledge yields good models from weak supervision