A Supertag-Context Model for Weakly-Supervised CCG Parser Learning

Dan Garrette
Chris Dyer
Jason Baldridge
Noah A. Smith

U. Washington
CMU
UT-Austin
CMU
Contributions

1. A **new generative model** for learning CCG parsers from *weak supervision*

2. A way to select Bayesian *priors* that capture properties of CCG

3. A Bayesian *inference procedure* to learn the parameters of our model
Type-Level Supervision

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

the  lazy  dogs  wander
np/n  n/n  n  n/n
np  np  np/n  np/n
(np)/np  s
...
Type-Level Supervision

the
np/n
lazy
n/n
np
dogs
n
np
(s\np)/np
wander
n
n/n
np/n
s\np
...
PCFG: Local Decisions
PCFG: Local Decisions
PCFG: Local Decisions

A

B

C
PCFG: Local Decisions
PCFG: Local Decisions
PCFG: Local Decisions

\[ P(\frac{B}{D E} | \frac{B}{\_}) \quad P(\frac{C}{F G} | \frac{C}{\_}) \]
PCFG: Local Decisions

\[ P(\overline{BDE} \mid B) P(\overline{CFG} \mid C) \]
A New Generative Model

\[ P(\overline{D} \overline{E} | \overline{B} ) \]
A New Generative Model

\[
P(\frac{B}{D, E} | B) \times P_R(B \rightarrow F | B)
\]
A New Generative Model

\[\text{<S> \times P(D|B \wedge E) \times P_R(B \rightarrow F|B) \times P_L(S \leftarrow B|B)}\]
A New Generative Model

(This makes inference tricky… we’ll come back to that)
Why CCG?

• The grammar formalism *itself* can be used to guide learning
  
  • Given any two categories, we always know whether they are combinable.

• We can extract *a priori* context preferences, before we even look at the data
  
  • Adjacent categories *tend* to be combinable.
Why CCG?

universal, intrinsic grammar properties

all relationships must be learned
CCG Parsing

s
  /\  
 np /n n 
  /\  
 np/n the
  
  

n/n lazy

n

n

n

s\np


sleeps
the lazy dog sleeps
Supertag Context

the lazy dog sleeps
Supertag Context

np/n
the
n/n
dog
n
s\np/s
sleeps
Supertag Context

- np/n: the
- n: lazy dog
- s\np: sleeps
Supertag Context

The lazy dog sleeps
Constituent Context

• Klein & Manning showed the value of modeling context with the Constituent Context Model (CCM)
Constituent Context

DT ←―・― ( JJ NN ) ―→ VBZ

[Klein & Manning 2002]
Constituent Context

“substitutability”

lazy dog

[Klein & Manning 2002]
Constituent Context

“substitutability”

DT ( NN ) VBZ

dog
Constituent Context

“substitutability”

big lazy dog

[Klein & Manning 2002]
Constituent Context

“substitutability”

DT ←...... ( ~Noun ) ......→ VBZ

[Klein & Manning 2002]
Constituent Context

“substitutability”

[Klein & Manning 2002]
Supertag Context

the lazy dog sleeps
Supertag Context

- We know the constituent label
- We know if it’s a fitting context, even before looking at the data
This Paper

1. A **new generative model** for learning CCG parsers from *weak supervision*

2. A way to select Bayesian *priors* that capture properties of CCG

3. A Bayesian *inference procedure* to learn the parameters of our model
Supertag-Context Parsing

Standard PCFG

\[ P(A_{\text{root}}) \]

\[ P(A \rightarrow A_{\text{left}} A_{\text{right}} \text{ OR } w_i) \]
Supertag-Context Parsing

With Context

\[ P(A_{\text{root}}) \]
\[ P(A \rightarrow A_{\text{left}} A_{\text{right}} \text{ OR } w_i) \]
\[ P(A \rightarrow t_{\text{left}}) \]
\[ P(A \rightarrow t_{\text{right}}) \]
Prior on Categories

[Garrette, Dyer, Baldridge, and Smith, 2015]
Supertag-Context Prior

\[ P_{L\text{-prior}}(t_{\text{left}} | A) \propto \begin{cases} 10^5 & \text{if } t_{\text{left}} \text{ can combine with } A \\ 1 & \text{otherwise} \end{cases} \]
Supertag-Context Prior

\[ P_{R-prior}(t_{\text{right}} \mid A) \propto \begin{cases} 10^5 & \text{if } A \text{ can combine with } t_{\text{right}} \\ 1 & \text{otherwise} \end{cases} \]
This Paper

1. A **new generative model** for learning CCG parsers from *weak supervision*

2. A way to select Bayesian *priors* that capture properties of CCG

3. A Bayesian *inference procedure* to learn the parameters of our model
Type-Level Supervision

?  

the  lazy  dogs  wander
np/n  n/n  n  np

np  np  (s\np)/np
Type-Supervised Learning

unlabeled corpus

tag dictionary

universal properties of the CCG formalism
Posterior Inference

• A Bayesian inference procedure will make use of our linguistically-informed priors

• But we can’t do sampling like a PCFG
  • Can’t compute the inside chart, even with dynamic programming.
Sampling via Metropolis-Hastings

Idea:

• Sample tree from an efficient \textit{proposal} distribution
  • (PCFG parameters) (Johnson et al. 2007)

• Accept according to the \textit{full} distribution
  • (Context parameters)
Posterior Inference

Priors (prefer connections)

Model

the
np/n

lazy
n/n
np

dogs
n
np
(s\np)/np

wander
Posterior Inference

Priors (prefer connections)

Model

the
np/n

lazy
n/n
np

dogs
n
np
(np)/np

wander
Posterior Inference

Priors (prefer connections)

Model
Posterior Inference

Priors (prefer connections)

Model

Inside

the
np/n

lazy
n/n
np

dogs
n
np

(s\np)/np

wander
n
n/n

np/n

s\np

...
Posterior Inference

Priors (prefer connections)

Model

Sample

the
np/n

lazy
n/n

dogs
n

wander
n

…
Metropolis-Hastings

Priors (prefer connections)

Model
Metropolis-Hastings

Priors (prefer connections)

Model

Existing Tree

New Tree
Metropolis-Hastings

Priors (prefer connections)

Model

Existing Tree

New Tree
Metropolis-Hastings

Priors
(prefer connections)

Model

Existing Tree

New Tree
Metropolis-Hastings

Priors
(prefer connections)

Model
Posterior Inference

Priors (prefer connections)

Model
Metropolis-Hastings

- Sample tree based only on the pcfg parameters
- Accept based only on the context
- New worse than old $\Rightarrow$ less likely to accept
Experimental Results
Experimental Question

• When supervision is incomplete, does modeling context, and biasing toward combining contexts, help learn better parsing models?
English Results

<table>
<thead>
<tr>
<th>Size of Corpus</th>
<th>Parsing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>250k</td>
<td>65</td>
</tr>
<tr>
<td>200k</td>
<td>61</td>
</tr>
<tr>
<td>150k</td>
<td>64</td>
</tr>
<tr>
<td>100k</td>
<td>64</td>
</tr>
<tr>
<td>50k</td>
<td>63</td>
</tr>
<tr>
<td>25k</td>
<td>60</td>
</tr>
</tbody>
</table>

- **no context**
- **+context combinability**
Experimental Results

<table>
<thead>
<tr>
<th>Language</th>
<th>Parsing Accuracy (no context)</th>
<th>Parsing Accuracy (+context combinability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>55</td>
<td>58</td>
</tr>
<tr>
<td>Italian</td>
<td>52</td>
<td>54</td>
</tr>
<tr>
<td>Chinese</td>
<td>29</td>
<td>34</td>
</tr>
</tbody>
</table>

25k token TD corpus
Conclusion

Under weak supervision, we can use universal grammatical knowledge about context to find trees with a better global structure.
Deficiency

• Generative story has a “throw away” step if the context-generated nonterminals don’t match the tree.

• We sample only over the space of valid trees (condition on well-formed structures).

• This is a benefit of the Bayesian formulation.

• See Smith 2011.
Metropolis-Hastings

current tree

new tree

\[ P_{\text{context}}(y) = \frac{P_{\text{full}}(y)}{P_{\text{pcfg}}(y)} \]

\[ P_{\text{context}}(y') = \frac{P_{\text{full}}(y')}{P_{\text{pcfg}}(y')} \]

\[ z \sim \text{uniform}(0,1) \]

accept if \[ z < \frac{P_{\text{full}}(y') / P_{\text{pcfg}}(y')}{P_{\text{full}}(y) / P_{\text{pcfg}}(y)} = \frac{P_{\text{context}}(y')}{P_{\text{context}}(y)} \]