Improving Neural Parsing by Disentangling Model Combination and Reranking Effects

Daniel Fried*, Mitchell Stern* and Dan Klein
UC Berkeley
Top-down generative models
Top-down generative models

The man had an idea.
The man had an idea.
Top-down generative models

(S (NP The man) had (NP an idea) .)
The man had an idea.
The man had an idea.
The man had an idea.

(S (NP The man )

VP

NP

The man

NP

had

an

idea

. 
Top-down generative models

The man had an idea.

(S (NP The man) (VP had an idea))
Top-down generative models

The man had an idea.

(S (NP The man) (VP had (NP an idea) ). )
The man had an idea.

(S (NP The man) (VP had (NP an idea) .))

$G_{LSTM}$ [Parsing as Language Modeling, Choe and Charniak, 2016]
Top-down generative models

(S (NP The man) (VP had (NP an idea) ) ).

\[ G_{\text{LSTM}} \] \textbf{[Parsing as Language Modeling, Choe and Charniak, 2016]}\n
\[ G_{\text{RNNG}} \] \textbf{[Recurrent Neural Network Grammars, Dyer et al. 2016]}
Generative models as rerankers
Generative models as rerankers

base parser  generative neural model

B  \rightarrow  G
Generative models as rerankers

**Base parser** \( B \) \rightarrow **Generative neural model** \( G \)

\[ y \sim p_B(y|x) \]
Generative models as rerankers

base parser  generative neural model

\[
y \sim p_B(y|x)
\]

\[
\text{argmax}_y \ p_G(x, y)
\]
Generative models as rerankers

base parser → generative neural model
Generative models as rerankers

base parser \rightarrow \text{generative neural model}

F1 on Penn Tree Bank
Generative models as rerankers

base parser \rightarrow \text{generative neural model} \rightarrow \text{Charniak parser}

Choe and Charniak 2016

89.7 \rightarrow 92.6

LSTM language model ($G_{\text{LSTM}}$)
Generative models as rerankers

- Base parser: Charniak parser
- Generative neural model: LSTM language model
  - Choe and Charniak 2016: 89.7, LSTM language model ($G_{LSTM}$)
  - Dyer et al. 2016: 91.7, RNNG-discriminative
  - RNNG-discriminative: 93.3, RNNG-generative ($G_{RNNG}$)

F1 on Penn Tree Bank
B: Necessary evil, or secret sauce?

base parser  generative neural model

B  →  G
B: Necessary evil, or secret sauce?

base parser \hspace{1cm} \text{generative neural model}

\[ \mathbf{B} \rightarrow \mathbf{G} \]

Should we try to do away with B?
B: Necessary evil, or secret sauce?

Should we try to do away with B?

No, better to combine B and G more explicitly.
**B: Necessary evil, or secret sauce?**

base parser \rightarrow \text{generative neural model}

\begin{center}
\begin{tikzpicture}
  \node [draw] (b) {B};
  \node [draw, right of=b, xshift=1cm] (g) {G};
  \draw[->] (b) -- (g);
\end{tikzpicture}
\end{center}

Should we try to do away with B?

No, better to combine B and G more explicitly

93.9 F1 on PTB; 94.7 semi-supervised
Using standard beam search for G

True Parse: (S (NP The man)

Beam
Using standard beam search for \( G \)

True Parse:

- (S)
- (NP)
- The
- man

Beam:

- (S)
Using standard beam search for G

True Parse

(S)  (NP)  The  man

Beam

(S)  (NP)  (VP)  (PP)
Using standard beam search for G

True Parse
(S) (NP) The man

Beam
(S) (NP) → (NP)
(VP) → (NP)
(NP) → (NP)
Using standard beam search for $G$

<table>
<thead>
<tr>
<th>True Parse</th>
<th>Beam</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S)</td>
<td>(S)</td>
</tr>
<tr>
<td>(NP)</td>
<td>(NP)</td>
</tr>
<tr>
<td>The</td>
<td>(VP)</td>
</tr>
<tr>
<td>man</td>
<td>(PP)</td>
</tr>
<tr>
<td></td>
<td>(NP)</td>
</tr>
<tr>
<td></td>
<td>(NP)</td>
</tr>
</tbody>
</table>
Using standard beam search for $G$

True Parse

Beam

(S) → (NP) → (NP) → (NP) → ... → (NP)

The man

The
Using standard beam search for $G$

Beam Size 100

$G_{RNNG}$  29.1 F1

$G_{LSTM}$  27.4 F1
Word generation is lexicalized:

(S (NP The man ) (VP had (NP an idea ) ) ) .
Word-synchronous beam search

\[ w_0 \]

\( (S) \)

[Roark 2001; Titov and Henderson 2010; Charniak 2010; Buys and Blunsom 2015 ]
Word-synchronous beam search

[S (NP (VP (PP (S (NP The) (NP The) The)) The))]

[Roark 2001; Titov and Henderson 2010; Charniak 2010; Buys and Blunsom 2015]
Word-synchronous beam search

\[ \text{(S)} \rightarrow \text{(NP)} \rightarrow \text{(VP)} \rightarrow \text{(PP)} \rightarrow \text{(NP)} \rightarrow \text{The} \rightarrow \text{(NP)} \rightarrow \text{man} \]

[Roark 2001; Titov and Henderson 2010; Charniak 2010; Buys and Blunsom 2015]
Word-synchronous beam search

![Graph showing F1 on PTB vs Beam Size for \(G_{\text{LSTM}}\) and \(G_{\text{RNN-G}}\).]
Word-synchronous beam search

![Graph showing F1 on PTB vs Beam Size for different models: LSTM and RNNG. The graph illustrates the performance improvement as the beam size increases.]
Finding model combination effects
Finding model combination effects

The man had an idea.
Add G’s search proposal to candidate list:

*The man had an idea.*
Finding model combination effects

Add G’s search proposal to candidate list:

\[ \text{The man had an idea.} \]
Finding model combination effects

Add G’s search proposal to candidate list:
Finding model combination effects

Add G’s search proposal to candidate list:

\[ G \cup B \rightarrow G \]
Finding model combination effects

F1 on PTB

93.5

B

G_{RNNG} \cup B

RNNG Generative Model

93.7

B

G_{LSTM} \cup B

LSTM Generative Model
Finding model combination effects

F1 on PTB
Reranking shows implicit model combination

B hides model errors in G
Can we do better by simply combining model scores?

\[ \log p_G(x, y) \]
Can we do better by simply combining model scores?

\[ \log p_G(x, y) \]
Can we do better by simply combining model scores?

\[ \lambda \log p_G(x, y) + (1 - \lambda) \log p_B(y|x) \]
Making model combination explicit

F1 on PTB

score with G

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNNG Generative Model ($G=\text{G}_{\text{RNNG}}$)</td>
<td>93.5</td>
</tr>
<tr>
<td>LSTM Generative Model ($G=\text{G}_{\text{LSTM}}$)</td>
<td>93.7</td>
</tr>
</tbody>
</table>

B

$\text{G}_{\text{RNNG}} \cup \text{B}$

$\text{G}_{\text{LSTM}} \cup \text{B}$
Making model combination explicit

F1 on PTB

- Score with G + B
- Score with G

RNNG Generative Model ($G = G_{RNNG}$)

- B: 93.9
- $G_{RNNG}$: 93.5
- $U$: 92.8

LSTM Generative Model ($G = G_{LSTM}$)

- B: 94.0
- $G_{LSTM}$: 93.7
- $U$: 93.5
Explicit score combination prevents errors

\[ \text{fast} \quad \text{best} \]
Comparison to past work

F1 on PTB
Comparison to past work

F1 on PTB

92.6
Choe & Charniak
2016
Comparison to past work

F1 on PTB

- Choe & Charniak, 2016: 92.6
- Dyer et al., 2016: 93.3
Comparison to past work

F1 on PTB

- Choe & Charniak 2016: 92.6
- Dyer et al. 2016: 93.3
- Kuncoro et al. 2017: 93.6
Comparison to past work

F1 on PTB

Choe & Charniak 2016

Dyer et al. 2016

Kuncoro et al. 2017

Ours

92.6

93.3

93.6

93.5

$G_{RNNG} \cup B \rightarrow G_{RNNG} + B$
Comparison to past work

F1 on PTB

<table>
<thead>
<tr>
<th>Model</th>
<th>2016 F1</th>
<th>2016 F1</th>
<th>2017 F1</th>
<th>Ours F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choe &amp; Charniak</td>
<td>92.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyer et al.</td>
<td>93.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kuncoro et al.</td>
<td>93.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td>93.9</td>
</tr>
</tbody>
</table>

| Ours                          |         |         |         | 93.9    |
Comparison to past work

F1 on PTB

- Choe & Charniak 2016: 92.6
- Dyer et al. 2016: 93.3
- Kuncoro et al. 2017: 93.6
- Ours: 93.9

Add silver data: 93.8
Add $G_{\text{LSTM}}$: 93.9
$G_{\text{RNNG}} \cup B \rightarrow G_{\text{RNNG} + B}$: 93.5
Comparison to past work

F1 on PTB

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choe &amp; Charniak 2016</td>
<td>93.8</td>
</tr>
<tr>
<td>Dyer et al. 2016</td>
<td>93.3</td>
</tr>
<tr>
<td>Kuncoro et al. 2017</td>
<td>93.6</td>
</tr>
<tr>
<td>Ours</td>
<td>93.9</td>
</tr>
</tbody>
</table>

*94.7*  
(add silver data)  
(add $G_{LSTM}$)  
$G_{RNNG} \cup B \rightarrow G_{RNNG} + B$
Conclusion

Search procedure for G
Conclusion

Search procedure for G

(more effective version forthcoming: Stern et al., EMNLP 2017)
Conclusion

Search procedure for $G$.

(more effective version forthcoming: Stern et al., EMNLP 2017)

Found model combination effects in $B \rightarrow G$. 
Conclusion

Search procedure for $\mathbb{G}$

(more effective version forthcoming: Stern et al., EMNLP 2017)

Found model combination effects in $\mathbb{B}$ → $\mathbb{G}$

Large improvements from simple, explicit score combination:

$\mathbb{B}$ → $\mathbb{G}$ + $\mathbb{B}$
Thanks!