Straight to the Tree: Constituency Parsing with Neural Syntactic Distance

Yikang Shen*, Zhouhan Lin*, Athul Paul Jacob, Alessandro Sordoni, Aaron Courville, Yoshua Bengio

University of Montreal, Microsoft Research, University of Waterloo
Overview

- Motivation
- Syntactic Distance based Parsing Framework
- Model
- Experimental Results
Overview

- Motivation
- Syntactic Distance based Parsing Framework
- Model
- Experimental Results
ICLR 2018: Neural Language Modeling by Jointly Learning Syntax and Lexicon

Supervised Constituency Parsing with Syntactic Distance? [Shen et al. 2018]
Chart Neural Parsers

1. High computational cost:
   Complexity of CYK is $O(n^3)$.
2. Complicated loss function:

$\max_0 \max_T \left[ s(T) + \Delta(T, T^*) - s(T^*) \right]$

Transition based Neural Parsers

1. Greedy decoding:
   Incompleted tree (the shift and reduce steps may not match).
2. Exposure bias
   The model is never exposed to its own mistakes during training

[Stern et al., 2017; Cross and Huang, 2016]
Overview

- Motivation
- **Syntactic Distance based Parsing Framework**
- Model
- Experimental Results
Only the order of split (or combination) matters for reconstructing the tree.

Can we model the order directly?
Syntactic distance

**Definition 2.1.** Let $T$ be a parse tree that contains a set of leaves $(w_0, \ldots, w_n)$. The height of the lowest common ancestor for two leaves $(w_i, w_j)$ is noted as $\tilde{d}_j^i$. The syntactic distances of $T$ can be any vector of scalars $d = (d_1, \ldots, d_n)$ that satisfy:

$$\text{sign}(d_i - d_j) = \text{sign}(\tilde{d}_i^{i-1} - \tilde{d}_j^{j-1})$$

(1)

For each **split point**, their **syntactic distance** should share the same order as the height of **related node**
Convert to binary tree

[Stern et al., 2017]
Tree to Distance

The height for each non-terminal node is the maximum height of its children plus 1

Algorithm 1 Binary Parse Tree to Distance

(∪ represents the concatenation operator of lists)

1: function DISTANCE(node)
2: if node is leaf then
3: \[
4: \text{d} \leftarrow []
5: \text{c} \leftarrow []
6: \text{t} \leftarrow [\text{node.tag}]
7: \text{h} \leftarrow 0
8: \]
9: else
10: \[
11: \text{child}_l, \text{child}_r \leftarrow \text{children of node}
12: \text{d}_l, \text{c}_l, \text{t}_l, \text{h}_l \leftarrow \text{Distance(child}_l)
13: \text{d}_r, \text{c}_r, \text{t}_r, \text{h}_r \leftarrow \text{Distance(child}_r)
14: \text{h} \leftarrow \text{max}(\text{h}_l, \text{h}_r) + 1
15: \text{d} \leftarrow \text{d}_l \cup [\text{h}] \cup \text{d}_r
16: \text{c} \leftarrow \text{c}_l \cup [\text{node.label}] \cup \text{c}_r
17: \text{t} \leftarrow \text{t}_l \cup \text{t}_r
18: end if
19: return d, c, t, h
20: end function
Tree to Distance

S
  NP
    She
  VP
    S-VP
      O
      enjoys
    O
    playing
  NP
      tennis
S  VP  S-VP  O

NP  VP  S-VP  Ø

d0  d1  d2  d3  d4  d5
She  enjoys  playing  tennis  .  .
Distance to Tree

Split point for each bracket is the one with maximum distance.

**Algorithm 2** Distance to Binary Parse Tree

1: function TREE(d,c,t)
2:     if d = [] then
3:         node ← Leaf(t)
4:     else
5:         $i \leftarrow \text{arg max}_i(d)$
6:         child$_L \leftarrow \text{Tree}(d_{<i}, c_{<i}, t_{<i})$
7:         child$_R \leftarrow \text{Tree}(d_{>i}, c_{>i}, t_{\geq i})$
8:         node ← Node(child$_L$, child$_R$, c$_i$)
9:     end if
10: return node
11: end function
Distance to Tree

She enjoys playing tennis.
Overview

- Motivation
- Syntactic Distance based Parsing Framework
- Model
- Experimental Results
Framework for inferring the distances and labels

Labels for non-leaf nodes

Labels for leaf nodes

Distances

She

enjoys

playing

tennis

PRP

VBZ

VBG

NN

.
Inferring the distances

Distances

- d0
- d1
- d2
- d3
- d4
- d5
Inferring the distances
Pairwise learning-to-rank loss for distances

\[ L_{\text{dist}}^{\text{rank}} = \sum_{i, j > i} [1 - \text{sign}(d_i - d_j)(\hat{d}_i - \hat{d}_j)]^+ \]

\[
\text{sign}(x) = \begin{cases} 
1, & x > 0 \\
0, & x = 0 \\
-1, & x < 0 
\end{cases}
\]

a variant of hinge loss
Pairwise learning-to-rank loss for distances

$$L_{\text{dist}}^{\text{rank}} = \sum_{i,j>i} [1 - \text{sign}(d_i - d_j)(\hat{d}_i - \hat{d}_j)]^+$$

While $d_i > d_j$:

$$L$$

While $d_i < d_j$:

$$L$$
Framework for inferring the distances and labels

Labels for non-leaf nodes

Labels for leaf nodes

Distances
Framework for inferring the distances and labels

Labels for non-leaf nodes

Labels for leaf nodes
Inferring the Labels
Inferring the Labels
Inferring the Labels

S

Ø

VP

S-VP

<s> She enjoys playing tennis. </s>
Putting it together

\[ L = L_{\text{label}} + L_{\text{rank}}^{\text{dist}} \]
Putting it together
Overview

- Motivation
- Syntactic Distance based Parsing Framework
- Model
- Experimental Results
Experiments: Penn Treebank

<table>
<thead>
<tr>
<th>Model</th>
<th>LP</th>
<th>LR</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vinyals et al. (2015)</td>
<td>-</td>
<td>-</td>
<td>88.3</td>
</tr>
<tr>
<td>Zhu et al. (2013)</td>
<td>90.7</td>
<td>90.2</td>
<td>90.4</td>
</tr>
<tr>
<td>Dyer et al. (2016)</td>
<td>-</td>
<td>-</td>
<td>89.8</td>
</tr>
<tr>
<td>Watanabe and Sumita (2015)</td>
<td>-</td>
<td>-</td>
<td>90.7</td>
</tr>
<tr>
<td>Cross and Huang (2016)</td>
<td>92.1</td>
<td>90.5</td>
<td>91.3</td>
</tr>
<tr>
<td>Liu and Zhang (2017b)</td>
<td>92.1</td>
<td>91.3</td>
<td>91.7</td>
</tr>
<tr>
<td>Stern et al. (2017a)</td>
<td>93.2</td>
<td>90.3</td>
<td>91.8</td>
</tr>
<tr>
<td>Liu and Zhang (2017a)</td>
<td>-</td>
<td>-</td>
<td>91.8</td>
</tr>
<tr>
<td>Gaddy et al. (2018)</td>
<td>-</td>
<td>-</td>
<td>92.1</td>
</tr>
<tr>
<td>Stern et al. (2017b)</td>
<td>92.5</td>
<td>92.5</td>
<td>92.5</td>
</tr>
<tr>
<td><strong>Our Model</strong></td>
<td>92.0</td>
<td>91.7</td>
<td>91.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ensemble</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Shindo et al. (2012)</td>
<td>-</td>
<td>-</td>
<td>92.4</td>
</tr>
<tr>
<td>Vinyals et al. (2015)</td>
<td>-</td>
<td>-</td>
<td>90.5</td>
</tr>
<tr>
<td><strong>Semi-supervised</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhu et al. (2013)</td>
<td>91.5</td>
<td>91.1</td>
<td>91.3</td>
</tr>
<tr>
<td>Vinyals et al. (2015)</td>
<td>-</td>
<td>-</td>
<td>92.3</td>
</tr>
<tr>
<td><strong>Re-ranking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charniak and Johnson (2005)</td>
<td>91.8</td>
<td>91.2</td>
<td>91.5</td>
</tr>
<tr>
<td>Huang (2008)</td>
<td>91.2</td>
<td>92.2</td>
<td>91.7</td>
</tr>
<tr>
<td>Dyer et al. (2016)</td>
<td>-</td>
<td>-</td>
<td>93.3</td>
</tr>
</tbody>
</table>
# Experiments: Chinese Treebank

<table>
<thead>
<tr>
<th>Model</th>
<th>LP</th>
<th>LR</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charniak (2000)</td>
<td>82.1</td>
<td>79.6</td>
<td>80.8</td>
</tr>
<tr>
<td>Zhu et al. (2013)</td>
<td>84.3</td>
<td>82.1</td>
<td>83.2</td>
</tr>
<tr>
<td>Wang et al. (2015)</td>
<td>-</td>
<td>-</td>
<td>83.2</td>
</tr>
<tr>
<td>Watanabe and Sumita (2015)</td>
<td>-</td>
<td>-</td>
<td>84.3</td>
</tr>
<tr>
<td>Dyer et al. (2016)</td>
<td>-</td>
<td>-</td>
<td>84.6</td>
</tr>
<tr>
<td>Liu and Zhang (2017b)</td>
<td>85.9</td>
<td>85.2</td>
<td>85.5</td>
</tr>
<tr>
<td>Liu and Zhang (2017a)</td>
<td>-</td>
<td>-</td>
<td>86.1</td>
</tr>
<tr>
<td><strong>Our Model</strong></td>
<td>86.6</td>
<td>86.4</td>
<td>86.5</td>
</tr>
<tr>
<td><strong>Semi-supervised</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhu et al. (2013)</td>
<td>86.8</td>
<td>84.4</td>
<td>85.6</td>
</tr>
<tr>
<td>Wang and Xue (2014)</td>
<td>-</td>
<td>-</td>
<td>86.3</td>
</tr>
<tr>
<td>Wang et al. (2015)</td>
<td>-</td>
<td>-</td>
<td>86.6</td>
</tr>
<tr>
<td><strong>Re-ranking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charniak and Johnson (2005)</td>
<td>83.8</td>
<td>80.8</td>
<td>82.3</td>
</tr>
<tr>
<td>Dyer et al. (2016)</td>
<td>-</td>
<td>-</td>
<td>86.9</td>
</tr>
</tbody>
</table>
Experiments: Detailed statistics in PTB and CTB

<table>
<thead>
<tr>
<th>dev/test result</th>
<th>Prec.</th>
<th>Recall</th>
<th>F1</th>
<th>label accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>labeled</td>
<td>91.7/92.0</td>
<td>91.8/91.7</td>
<td>91.8/91.8</td>
<td>94.9/95.4%</td>
</tr>
<tr>
<td>unlabeled</td>
<td>93.0/93.2</td>
<td>93.0/92.8</td>
<td>93.0/93.0</td>
<td></td>
</tr>
<tr>
<td>CTB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>labeled</td>
<td>89.4/86.6</td>
<td>89.4/86.4</td>
<td>89.4/86.5</td>
<td>92.2/91.1%</td>
</tr>
<tr>
<td>unlabeled</td>
<td>91.1/88.9</td>
<td>91.1/88.6</td>
<td>91.1/88.8</td>
<td></td>
</tr>
</tbody>
</table>
Experiments: Ablation Test

<table>
<thead>
<tr>
<th>Model</th>
<th>LP</th>
<th>LR</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>92.0</td>
<td>91.7</td>
<td>91.8</td>
</tr>
<tr>
<td>w/o top LSTM</td>
<td>91.0</td>
<td>90.5</td>
<td>90.7</td>
</tr>
<tr>
<td>w. Char LSTM</td>
<td>92.1</td>
<td>91.7</td>
<td>91.9</td>
</tr>
<tr>
<td>w. embedding</td>
<td>91.9</td>
<td>91.6</td>
<td>91.7</td>
</tr>
<tr>
<td>w. MSE loss</td>
<td>90.3</td>
<td>90.0</td>
<td>90.1</td>
</tr>
</tbody>
</table>
Experiments: Parsing Speed

<table>
<thead>
<tr>
<th>Model</th>
<th># sents/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petrov and Klein (2007)</td>
<td>6.2</td>
</tr>
<tr>
<td>Zhu et al. (2013)</td>
<td>89.5</td>
</tr>
<tr>
<td>Liu and Zhang (2017b)</td>
<td>79.2</td>
</tr>
<tr>
<td>Stern et al. (2017a)</td>
<td>75.5</td>
</tr>
<tr>
<td>Our model</td>
<td>111.1</td>
</tr>
<tr>
<td>Our model w/o tree inference</td>
<td>351</td>
</tr>
</tbody>
</table>
Conclusions and Highlights

- **A novel constituency parsing scheme**: predicting tree structure from a set of real-valued scalars (syntactic distances).
- Completely **free from compounding errors**.
- **Strong performance** compare to previous models, and
- **Significantly more efficient** than previous models
- **Easy deployment**: The architecture of model is no more than a stack of standard recurrent and convolutional layers.
One more thing...

- The research in rank loss is well-studied in the topic of learning-to-rank, since 2005 (Burges et al. 2005).
- Models that are good at learning these syntactic distances are not widely known until the rediscovery of LSTM in 2013 (Graves 2013).
- Efficient regularization methods for LSTM didn’t become mature until 2017 (Merity 2017).
Thank you!

Yikang Shen, Zhouhan Lin
MILA, Université de Montréal
{yikang.shn, lin.zhouhan}@gmail.com

Questions?

Code: [QR Code]

Paper: [QR Code]