Second-Order Neural Dependency Parsing with Message Passing and End-to-End Training

Xinyu Wang and Kewei Tu
ShanghaiTech University
Motivation and Contributions

• Higher-order approaches have achieved state-of-the-art performance

• Our work:
  • Apply second-order semantic parser (Wang et al., 2019) to syntactic dependency parsing.

• Our observation:
  • Higher-order decoding is effective even with contextual word embeddings.
  • Parsers without head-selection constraint can match the accuracy of parsers with the head-selection constraint and can even outperform the latter when using BERT embedding

Structure

Edge Prediction

Label Prediction

Embedding

BiLSTM

FNN

Biaffine or Trilinear Function

MFVI Recurrent Layers

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)

Copyright © 2023 by the authors.

This paper is distributed under the Creative Commons Attribution 4.0 License.

No commercial use is permitted.

This paper is a viable extension of previous works on...

(Edge prediction)

Q(t) [s(sib), s(gp)]

Q(T)

(w_i, o_i)

(h_{i, edge-h/d}, h_{i, label-h/d})

h_{i, (sib), (gp)}

(s(sib), s(gp))

(s(edge), s(label))

(w_j, o_j)

(h_{j, edge-h/d}, h_{j, label-h/d})

h_{j, (sib), (gp)}

Q(T)
Approach

Binary Classification (Single):

\[
M_{ij}^{(t-1)} = \sum_{k \neq i,j} Q_{ik}^{(t-1)}(1)s_{ij,ik}^{(sib)} + Q_{jk}^{(t-1)}(1)s_{ij,jk}^{(gp)}
+ Q_{ki}^{(t-1)}(1)s_{ki,ij}^{(gp)}
\]

\[
Q_{ij}^{(t)}(0) \propto 1
\]

\[
Q_{ij}^{(t)}(1) \propto \exp\{s_{ij}^{(edge)} + M_{ij}^{(t-1)}\}
\]
Conditional Random Field

Nodes: Edges between two words

3,4 = not Poles

<TOP> 0
They 1
were 2
not 3
Poles 4

Sibling factor

Grandparent factor

Unary factor
Approach

Binary Classification (Single):

\[
\mathcal{M}_{ij}^{(t-1)} = \sum_{k \neq i, j} Q_{ik}^{(t-1)} (1)s_{ij, ik}^{(sib)} + Q_{jk}^{(t-1)} (1)s_{ij, jk}^{(gp)}
+ Q_{ki}^{(t-1)} (1)s_{ki, ij}^{(gp)}
\]

\[
Q_{ij}^{(t)}(0) \propto 1
\]

\[
Q_{ij}^{(t)}(1) \propto \exp\{s_{ij}^{(edge)} + \mathcal{M}_{ij}^{(t-1)}\}
\]

Head-selection (Local):

\[
\mathcal{M}_{j}^{(t-1)} (i) = \sum_{k \neq i, j} Q_{k}^{(t-1)} (i)s_{ij, ik}^{(sib)}
+ Q_{k}^{(t-1)} (j)s_{ij, jk}^{(gp)} + Q_{i}^{(t-1)} (k)s_{ki, ij}^{(gp)}
\]

\[
Q_{j}^{(t)} (i) = \frac{\exp\{s_{ij}^{(edge)} + \mathcal{M}_{j}^{(t-1)}(i)\}}{\sum_{k=0}^{n} \exp\{s_{kj}^{(edge)} + \mathcal{M}_{j}^{(t-1)}(k)\}}
\]
## Results

<table>
<thead>
<tr>
<th></th>
<th>PTB</th>
<th></th>
<th>CTB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UAS</td>
<td>LAS</td>
<td>UAS</td>
</tr>
<tr>
<td>Dozat and Manning (2017)</td>
<td>95.74</td>
<td>94.08</td>
<td>89.30</td>
</tr>
<tr>
<td>Ma et al. (2018)♦</td>
<td>95.87</td>
<td>94.19</td>
<td>90.59</td>
</tr>
<tr>
<td>F&amp;G (2019)♦</td>
<td>96.04</td>
<td>94.43</td>
<td>-</td>
</tr>
<tr>
<td>GNN</td>
<td>95.87</td>
<td>94.15</td>
<td>90.78</td>
</tr>
<tr>
<td>Single1O</td>
<td>95.75</td>
<td>94.04</td>
<td>90.53</td>
</tr>
<tr>
<td>Local1O</td>
<td>95.83</td>
<td>94.23</td>
<td>90.59</td>
</tr>
<tr>
<td>Single2O</td>
<td>95.86</td>
<td>94.19</td>
<td>90.75</td>
</tr>
<tr>
<td>Local2O</td>
<td>95.98</td>
<td>94.34</td>
<td>90.81</td>
</tr>
<tr>
<td>Ji et al. (2019)†</td>
<td>95.97</td>
<td>94.31</td>
<td>-</td>
</tr>
<tr>
<td>Zhang et al. (2020)††</td>
<td>96.14</td>
<td>94.49</td>
<td>-</td>
</tr>
<tr>
<td>Local2O††</td>
<td>96.12</td>
<td>94.47</td>
<td>-</td>
</tr>
<tr>
<td>+BERT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhou and Zhao (2019)♦</td>
<td>97.20</td>
<td>95.72</td>
<td></td>
</tr>
<tr>
<td>Clark et al. (2018)♦</td>
<td>96.60</td>
<td>95.00</td>
<td>-</td>
</tr>
<tr>
<td>Single1O</td>
<td>96.82</td>
<td>95.20</td>
<td>92.73</td>
</tr>
<tr>
<td>Local1O</td>
<td>96.86</td>
<td>95.32</td>
<td>92.47</td>
</tr>
<tr>
<td>Single2O</td>
<td>96.86</td>
<td>95.31</td>
<td>92.78</td>
</tr>
<tr>
<td>Local2O</td>
<td>96.91</td>
<td>95.34</td>
<td>92.55</td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th></th>
<th>PTB</th>
<th>CTB</th>
<th>bg</th>
<th>ca</th>
<th>cs</th>
<th>de</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>it</th>
<th>nl</th>
<th>no</th>
<th>ro</th>
<th>ru</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNN</td>
<td>94.15</td>
<td>89.50</td>
<td>90.33</td>
<td>92.39</td>
<td>90.95</td>
<td>79.73</td>
<td>88.43</td>
<td>91.56</td>
<td>87.23</td>
<td>92.44</td>
<td>88.57</td>
<td>89.38</td>
<td>85.26</td>
<td>91.20</td>
<td>89.37</td>
</tr>
<tr>
<td>Single1O</td>
<td>94.04</td>
<td>89.28</td>
<td>90.05</td>
<td>92.72</td>
<td>92.07</td>
<td>81.73</td>
<td>89.55</td>
<td>92.10</td>
<td>88.27</td>
<td>92.64</td>
<td>89.57</td>
<td>91.81</td>
<td>85.39</td>
<td>92.60</td>
<td>90.13</td>
</tr>
<tr>
<td>Local1O</td>
<td>94.23</td>
<td>89.28</td>
<td>90.30</td>
<td>92.56</td>
<td>92.15</td>
<td>81.42</td>
<td>89.43</td>
<td>91.99</td>
<td>88.26</td>
<td>92.49</td>
<td>89.76</td>
<td>91.91</td>
<td>85.27</td>
<td>92.72</td>
<td>90.13</td>
</tr>
<tr>
<td>Single2O</td>
<td>94.19</td>
<td>89.55</td>
<td>90.24</td>
<td>92.82</td>
<td>92.13</td>
<td>81.99</td>
<td>89.64</td>
<td>92.17</td>
<td>88.69</td>
<td>92.83</td>
<td>89.97</td>
<td>91.90</td>
<td>85.53</td>
<td>92.58</td>
<td>90.30</td>
</tr>
<tr>
<td>Local2O</td>
<td>94.34</td>
<td>89.57</td>
<td>90.53</td>
<td>92.83</td>
<td>92.12</td>
<td>81.73</td>
<td>89.72</td>
<td>92.07</td>
<td>88.53</td>
<td>92.78</td>
<td>90.19</td>
<td>91.88</td>
<td>85.88</td>
<td>92.67</td>
<td>90.35</td>
</tr>
</tbody>
</table>

† means that the model is statistically significantly better than the Local1O model with a significance level of p<0.05
‡ represents winner of the significant test between the Single2O and Local2O models

- Our second-order approaches outperform GNN and the first-order approaches both with and without BERT embeddings
- Without BERT, Local approaches slightly outperforms Single approaches, although the difference between the two is quite small
- When BERT is used, Single approaches clearly outperforms Local approaches
- The relative strength of Local and Single approaches varies over treebanks, suggesting varying importance of the head-selection constraint
Speed Comparison

<table>
<thead>
<tr>
<th>System</th>
<th>Train</th>
<th>Test</th>
<th>Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNN</td>
<td>392</td>
<td>464</td>
<td>$O(n^2d)$</td>
</tr>
<tr>
<td>Zhang et al. (2020)</td>
<td>200</td>
<td>400</td>
<td>$O(n^3)$</td>
</tr>
<tr>
<td>Single1O</td>
<td>616</td>
<td>1123</td>
<td>$O(n^2)$</td>
</tr>
<tr>
<td>Local1O</td>
<td>625</td>
<td>1150</td>
<td>$O(n^2)$</td>
</tr>
<tr>
<td>Single2O</td>
<td>481</td>
<td>966</td>
<td>$O(n^3)$</td>
</tr>
<tr>
<td>Local2O</td>
<td>486</td>
<td>1006</td>
<td>$O(n^3)$</td>
</tr>
</tbody>
</table>
Conclusion

• Second-order graph-based dependency parsing based on message passing and end-to-end neural networks
• Design a new approach that incorporates the head-selection structured constraint
• Show the effectiveness of second-order parsers against first-order parsers even with contextual embeddings
• Competitive accuracy with recent SOTA second-order parsers and significantly faster speed
• The limited usefulness of the head-selection constraint