An Empirical Comparison of Unsupervised Constituency Parsing Methods

Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, Kewei Tu

{lijun2, caoyf, caijiang, tukw}@shanghaitech.edu.cn
yongjiang.jy@alibaba-inc.com
Background

● **Goal**: To learn a constituency parser without parse tree annotations
Background

● **Goal**: To learn a constituency parser without parse tree annotations

● **Trends**: This task receives a lot of attention recently (2019)
  ○ increasing number of accepted papers: NACCL*2, ACL*5, EMNLP*3
  ○ with high quality: ICLR 2019 best paper (Shen et al., 2019)
Background

- **Goal**: To learn a constituency parser without parse tree annotations
- **Trends**: This task receives a lot of attention recently (2019)
  - increasing number of accepted papers: NACCL*2, ACL*5, EMNLP*3
  - with high quality: ICLR 2019 best paper (Shen et al., 2019)
- **Problems**: No unified experimental standard has been adopted
  - making the results across papers incomparable
Background

- **Goal**: To learn a constituency parser without parse tree annotations
- **Trends**: This task receives a lot of attention recently (2019)
  - increasing number of accepted papers: NACCL*2, ACL*5, EMNLP*3
  - with high quality: ICLR 2019 best paper (Shen et al., 2019)
- **Problems**: No unified experimental standard has been adopted
  - making the results across papers incomparable
- **Our contributions**:
  - Propose a standardized experimental setup
  - Conduct a systematic experiments on
    - **PRPN** (Shen et al., 2018)
    - **URNNG** (Kim et al., 2019b)
    - **DIORA** (Drozdov et al., 2019)
    - **CCM** (Klein and Manning, 2002)
    - **CCL** (Seginer, 2007)
Experimental setup

- Language
Experimental setup

- **Language**

Different languages have different syntactic properties

Japanese (mostly left branching)  English (mostly right branching)
Experimental setup

- **Language**: Use KTB and PTB for training and evaluation
- **Dataset pre-processing**

Data. For English, §2–21 of the Penn Treebank are used as training corpus for both, with §24 held out as validation, and §23 used for evaluation. Singleton words in the training corpus with unknown word classes using the the Berkeley parser’s mapping rules. Orthographic case distinctions are preserved, and numbers (besides singletons) are not tokenized.

We performed most experiments on the 7422 sentences in the Penn treebank Wall Street Journal section which contained no more than 10 words after the removal of punctuation and null elements (WSJ-10). Evaluation was done by measuring un-
Experimental setup

- **Language**: Use KTB and PTB for training and evaluation
- **Dataset pre-processing**: Train on length $\leq 10/40$; Split into train/dev/test
- **Punctuation post-processing**:

```
On Monday the stock market crashed
```

```
On Monday, the stock market crashed.
```
Experimental setup

- **Language**: Use KTB and PTB for training and evaluation
- **Dataset pre-processing**: Train on length <= 10/40; Split into train/dev/test
- **Punctuation post-processing**: Attach to root or least common ancestor
- **Evaluation**

```
       S
     /   \
    NP    VP
   /  \   /  \   \
  I   VPD NP   PP
   \    \   \  \
    shot DT NN IN PRP$ NN
     an elephant in my pajamas
```

\[(0, 0) (1, 1) (2, 2) \ldots (2, 3) (4, 6) (1, 6) (0, 6)\]
Experimental setup

- **Language**: Use KTB and PTB for training and evaluation
- **Dataset pre-processing**: Train on length \( \leq 10/40 \); Split into train/dev/test
- **Punctuation post-processing**: Attach to root or least common ancestor
- **Evaluation**
Experimental setup

- **Language**: Use KTB and PTB for training and evaluation
- **Dataset pre-processing**: Train on length \( \leq 10/40 \); Split into train/dev/test
- **Punctuation post-processing**: Attach to root or least common ancestor
- **Evaluation**: Report Micro/Macro/Evalb F1
- .....  
- More details can be found in our paper
### Experimental results (English)

<table>
<thead>
<tr>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
</tr>
<tr>
<td></td>
<td>± 3.65</td>
<td>± 3.45</td>
<td>± 3.45</td>
</tr>
<tr>
<td></td>
<td>± 3.65</td>
<td>± 3.45</td>
<td>± 3.45</td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
</tr>
<tr>
<td></td>
<td>± 1.33</td>
<td>± 1.83</td>
<td>± 1.83</td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
</tr>
<tr>
<td></td>
<td>± 3.65</td>
<td>± 3.45</td>
<td>± 3.45</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length ≤ 10.

| PRPN  | 18.08  | 21.73  | 22.85 | 41.99  | 45.50  | 45.36 | 33.25 | 33.92  | 36.85 |
|       | ± 3.66 | ± 3.69 | ± 3.45| ± 4.05 | ± 3.73 | ± 3.82| ± 3.20| ± 3.27 | ± 3.03|
| URNNG | 34.62  | 38.58  | 38.43 | 35.88  | 39.58  | 39.62 | 36.7  | 38.44  | 40.11 |
|       | ± 2.19 | ± 1.65 | ± 2.07| ± 0.00 | ± 0.00 | ± 0.00| ± 0.00| ± 3.84 | ± 4.03|
| DIORA | 20.44  | 23.72  | 25.08 | 46.27  | 47.81  | 49.39 | 41.48 | 41.56  | 44.63 |
|       | ± 1.53 | ± 1.66 | ± 1.44| ± 0.31 | ± 0.33 | ± 0.29| ± 0.43| ± 0.37 | ± 0.41|
| CCL   | 19.08  | 21.56  | 18.68 | 37.41  | 41.67  | 37.98 | n/a   | n/a    | n/a   |
| CCM   | 49.54  | 52.60  | 52.48 | 40.90  | 43.62  | 44.34 | 0.09  | 0.54   | 5.48  |
|       | ± 3.65 | ± 3.45 | ± 3.45| ± 4.05 | ± 3.73 | ± 3.82| ± 3.20| ± 3.27 | ± 3.03|
| LBranch | 6.00   | 8.98   | 11.49 | 6.00   | 8.98   | 11.49 | 4.88  | 6.36   | 10.01 |
|       | ± 1.53 | ± 1.66 | ± 1.44| ± 0.31 | ± 0.33 | ± 0.29| ± 0.43| ± 0.37 | ± 0.41|
| RBranch | 35.88  | 39.58  | 39.61 | 35.88  | 39.58  | 39.61 | 0.07  | 0.52   | 5.45  |
|       | ± 3.65 | ± 3.45 | ± 3.45| ± 4.05 | ± 3.73 | ± 3.82| ± 3.20| ± 3.27 | ± 3.03|
| UBound | 84.41  | 83.32  | 85.34 | 84.41  | 83.32  | 85.34 | 77.76 | 75.06  | 78.96 |

Evaluated on all test sentences.
## Experimental results

<table>
<thead>
<tr>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29 ± 4.49</td>
<td>37.29 ± 5.04</td>
<td>44.72 ± 3.59</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77 ± 1.11</td>
<td>53.67 ± 0.83</td>
<td>60.41 ± 0.89</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55 ± 2.50</td>
<td>37.90 ± 2.13</td>
<td>44.93 ± 2.00</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31 ± 3.55</td>
<td>36.61 ± 3.85</td>
<td>33.55 ± 3.67</td>
</tr>
<tr>
<td>CCM</td>
<td>62.97 ± 6.35</td>
<td>70.14 ± 5.03</td>
<td>50.29 ± 5.00</td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43 ± 54.20</td>
<td>60.79 ± 51.43</td>
<td>54.20 ± 60.79</td>
</tr>
<tr>
<td>UBound</td>
<td>83.20 ± 78.74</td>
<td>86.64 ± 83.20</td>
<td>86.64 ± 86.64</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length ≤ 10:

| PRPN   | 18.08 ± 3.66     | 21.73 ± 3.69     | 22.85 ± 3.45    | 41.99 ± 4.05    | 45.50 ± 3.73    | 45.36 ± 3.82    | 33.25 ± 4.21    | 33.92 ± 4.35    | 36.85 ± 4.43    |
| URNNG  | 34.62 ± 2.19     | 38.58 ± 1.65     | 38.43 ± 2.07    | 35.88 ± 0.00    | 39.58 ± 0.00    | 39.62 ± 0.00    | 36.7 ± 3.67     | 38.44 ± 3.84    | 40.11 ± 4.03    |
| DIORA  | 20.44 ± 1.53     | 23.72 ± 1.66     | 25.08 ± 1.44    | 46.27 ± 0.31    | 47.81 ± 0.33    | 49.39 ± 0.29    | 41.48 ± 4.69    | 41.56 ± 4.67    | 44.63 ± 4.98    |
| CCL    | 19.08 ± 21.56    | 18.68 ± 37.41    | 41.67 ± 37.98   | n/a (49.70)     | n/a (51.51)     | n/a (47.46)     |
| CCM    | 49.54 ± 52.60    | 52.48 ± 40.90    | 43.62 ± 44.34   | 0.09 ± 33.15    | 0.54 ± 36.88    | 5.48 ± 35.65    |
| LBranch| 6.00 ± 8.98      | 11.49 ± 6.00     | 8.98 ± 11.49    | 4.88 ± 5.55     | 6.36 ± 8.30     | 10.01 ± 11.07   |
| RBranch| 35.88 ± 39.58    | 39.61 ± 35.88    | 39.58 ± 39.61   | 0.07 ± 35.54    | 0.52 ± 38.98    | 5.45 ± 39.3     |
| UBound | 84.41 ± 83.32    | 85.34 ± 84.41    | 85.34 ± 85.34   | 77.76 ± 75.06   | 78.96           |
## Experimental results

<table>
<thead>
<tr>
<th>Metric</th>
<th>micro</th>
<th>macro</th>
<th>evalb</th>
<th>micro</th>
<th>macro</th>
<th>evalb</th>
<th>micro</th>
<th>macro</th>
<th>evalb</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
<td>56.98</td>
<td>58.79</td>
<td>65.23</td>
<td>38.07</td>
<td>52.17</td>
<td>51.56</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
<td>± 3.66</td>
<td>± 2.85</td>
<td>± 2.92</td>
<td>± 3.94</td>
<td>± 3.08</td>
<td>± 3.08</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
<td>51.43</td>
<td>54.20</td>
<td>60.94</td>
<td>47.95</td>
<td>49.07</td>
<td>41.65</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
<td>50.26</td>
<td>52.92</td>
<td>59.86</td>
<td>42.66</td>
<td>47.13</td>
<td>55.15</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
<td>± 0.72</td>
<td>± 0.68</td>
<td>± 0.58</td>
<td>± 0.98</td>
<td>± 1.92</td>
<td>± 0.77</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
<td>53.67</td>
<td>57.45</td>
<td>53.67</td>
<td>n/a</td>
<td>62.39</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
<td>50.29</td>
<td>53.73</td>
<td>60.03</td>
<td>1.04</td>
<td>54.30</td>
<td>22.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>1.03</td>
<td>56.80</td>
<td>22.63</td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>68.19</td>
<td>56.85</td>
<td>75.15</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length ≤ 10.

<table>
<thead>
<tr>
<th>Metric</th>
<th>micro</th>
<th>macro</th>
<th>evalb</th>
<th>micro</th>
<th>macro</th>
<th>evalb</th>
<th>micro</th>
<th>macro</th>
<th>evalb</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRPN</td>
<td>18.08</td>
<td>21.73</td>
<td>22.85</td>
<td>41.99</td>
<td>45.50</td>
<td>45.36</td>
<td>33.25</td>
<td>42.17</td>
<td>36.85</td>
</tr>
<tr>
<td></td>
<td>± 3.66</td>
<td>± 3.69</td>
<td>± 3.45</td>
<td>± 4.05</td>
<td>± 3.73</td>
<td>± 3.82</td>
<td>± 3.20</td>
<td>± 1.82</td>
<td>± 3.03</td>
</tr>
<tr>
<td>URNNG</td>
<td>34.62</td>
<td>38.58</td>
<td>38.43</td>
<td>35.88</td>
<td>39.58</td>
<td>39.62</td>
<td>36.7</td>
<td>36.72</td>
<td>38.44</td>
</tr>
<tr>
<td></td>
<td>± 2.19</td>
<td>± 1.65</td>
<td>± 2.07</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>20.44</td>
<td>23.72</td>
<td>25.08</td>
<td>46.27</td>
<td>47.81</td>
<td>49.39</td>
<td>41.48</td>
<td>46.94</td>
<td>44.63</td>
</tr>
<tr>
<td></td>
<td>± 1.53</td>
<td>± 1.66</td>
<td>± 1.44</td>
<td>± 0.31</td>
<td>± 0.33</td>
<td>± 0.29</td>
<td>± 0.43</td>
<td>± 1.59</td>
<td>± 0.41</td>
</tr>
<tr>
<td>CCL</td>
<td>19.08</td>
<td>21.56</td>
<td>18.68</td>
<td>37.41</td>
<td>41.67</td>
<td>37.98</td>
<td>n/a</td>
<td>49.70</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>49.54</td>
<td>52.60</td>
<td>52.48</td>
<td>40.90</td>
<td>43.62</td>
<td>44.34</td>
<td>0.09</td>
<td>33.15</td>
<td>5.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBranch</td>
<td>6.00</td>
<td>8.98</td>
<td>11.49</td>
<td>6.00</td>
<td>8.98</td>
<td>11.49</td>
<td>4.88</td>
<td>5.55</td>
<td>6.36</td>
</tr>
<tr>
<td>RBranch</td>
<td>35.88</td>
<td>39.58</td>
<td>39.61</td>
<td>35.88</td>
<td>39.58</td>
<td>39.61</td>
<td>0.07</td>
<td>35.54</td>
<td>5.45</td>
</tr>
<tr>
<td>UBound</td>
<td>84.41</td>
<td>83.32</td>
<td>85.34</td>
<td>84.41</td>
<td>83.32</td>
<td>85.34</td>
<td>77.76</td>
<td>75.06</td>
<td>78.96</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences.

Trained on PTB with length <= 40, with punctuation

Trained on PTB with length <= 40, without punctuation
## Experimental results

<table>
<thead>
<tr>
<th>Metric</th>
<th>PTB len10_nopunct</th>
<th>PTB len40_nopunct</th>
<th>PTB len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td><strong>62.97</strong></td>
<td><strong>63.35</strong></td>
<td><strong>70.14</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length ≤ 10.

<table>
<thead>
<tr>
<th>Metric</th>
<th>PTB len10_nopunct</th>
<th>PTB len40_nopunct</th>
<th>PTB len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>18.08</td>
<td>21.73</td>
<td>22.85</td>
</tr>
<tr>
<td></td>
<td>± 3.66</td>
<td>± 3.69</td>
<td>± 3.45</td>
</tr>
<tr>
<td>URNNG</td>
<td>34.62</td>
<td>38.58</td>
<td>38.43</td>
</tr>
<tr>
<td></td>
<td>± 2.19</td>
<td>± 1.65</td>
<td>± 2.07</td>
</tr>
<tr>
<td>DIORA</td>
<td>20.44</td>
<td>23.72</td>
<td>25.08</td>
</tr>
<tr>
<td></td>
<td>± 1.53</td>
<td>± 1.66</td>
<td>± 1.44</td>
</tr>
<tr>
<td>CCL</td>
<td>19.08</td>
<td>21.56</td>
<td>18.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td><strong>49.54</strong></td>
<td><strong>52.60</strong></td>
<td><strong>52.48</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBranch</td>
<td>6.00</td>
<td>8.98</td>
<td>11.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>35.88</td>
<td>39.58</td>
<td>39.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBound</td>
<td>84.41</td>
<td>83.32</td>
<td>85.34</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences.
Experimental results (English)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>± 4.49</td>
<td>37.29</td>
<td>± 5.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>± 1.11</td>
<td>53.67</td>
<td>± 0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>± 2.50</td>
<td>37.90</td>
<td>± 2.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
<td>53.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
<td>50.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>51.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>83.20</td>
</tr>
</tbody>
</table>

**Evaluated on all test sentences.**

<table>
<thead>
<tr>
<th>Metric</th>
<th>PRPN</th>
<th>URNNG</th>
<th>DIORA</th>
<th>CCL</th>
<th>CCM</th>
<th>LBranch</th>
<th>RBranch</th>
<th>UBound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18.08</td>
<td>34.62</td>
<td>20.44</td>
<td>19.08</td>
<td>49.54</td>
<td>6.00</td>
<td>35.88</td>
<td>84.41</td>
</tr>
<tr>
<td></td>
<td>± 3.66</td>
<td>± 2.19</td>
<td>± 1.53</td>
<td>18.68</td>
<td>52.60</td>
<td>11.49</td>
<td>39.58</td>
<td>85.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.99</td>
<td>40.90</td>
<td>8.98</td>
<td>35.88</td>
<td>84.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>± 4.05</td>
<td>± 3.82</td>
<td>± 3.66</td>
<td>± 0.31</td>
<td>± 0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.73</td>
<td>39.75</td>
<td>± 0.00</td>
<td>± 0.33</td>
<td>± 1.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>± 3.82</td>
<td>± 3.62</td>
<td>± 0.00</td>
<td>± 0.29</td>
<td>± 0.29</td>
</tr>
</tbody>
</table>
## Experimental Results (English)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
<th>10.00</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
<td>56.98</td>
<td>58.79</td>
<td>65.23</td>
<td>38.07</td>
<td>33.75</td>
<td>51.56</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
<td>56.98</td>
<td>58.79</td>
<td>65.23</td>
<td>38.07</td>
<td>33.75</td>
<td>51.56</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 5.39</td>
<td>± 3.66</td>
<td>± 2.85</td>
<td>± 2.92</td>
<td>± 3.94</td>
<td>± 3.33</td>
<td>± 3.08</td>
<td>± 3.94</td>
<td>± 5.04</td>
<td>± 5.39</td>
<td>± 3.66</td>
<td>± 2.85</td>
<td>± 2.92</td>
<td>± 3.94</td>
<td>± 3.33</td>
<td>± 3.08</td>
</tr>
<tr>
<td>URNN</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
<td>51.43</td>
<td>54.20</td>
<td>60.94</td>
<td>47.95</td>
<td>41.65</td>
<td>59.34</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
<td>51.43</td>
<td>54.20</td>
<td>60.94</td>
<td>47.95</td>
<td>41.65</td>
<td>59.34</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
</tr>
<tr>
<td>Diora</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
<td>50.26</td>
<td>52.92</td>
<td>59.86</td>
<td>42.66</td>
<td>37.77</td>
<td>55.15</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
<td>50.26</td>
<td>52.92</td>
<td>59.86</td>
<td>42.66</td>
<td>37.77</td>
<td>55.15</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
<td>± 0.72</td>
<td>± 0.68</td>
<td>± 0.58</td>
<td>± 0.98</td>
<td>± 0.84</td>
<td>± 0.77</td>
<td>± 0.98</td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
<td>± 0.72</td>
<td>± 0.68</td>
<td>± 0.58</td>
<td>± 0.98</td>
<td>± 0.84</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
<td>53.67</td>
<td>57.45</td>
<td>53.67</td>
<td>n/a</td>
<td>52.33</td>
<td>n/a</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
<td>53.67</td>
<td>57.45</td>
<td>53.67</td>
<td>n/a</td>
<td>52.33</td>
<td>n/a</td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
<td>50.29</td>
<td>53.73</td>
<td>60.03</td>
<td>1.04</td>
<td>4.30</td>
<td>22.70</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
<td>50.29</td>
<td>53.73</td>
<td>60.03</td>
<td>1.04</td>
<td>4.30</td>
<td>22.70</td>
</tr>
<tr>
<td>Rbranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>1.03</td>
<td>4.30</td>
<td>22.63</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>1.03</td>
<td>4.30</td>
<td>22.63</td>
</tr>
<tr>
<td>Ubound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>68.19</td>
<td>56.85</td>
<td>75.15</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>68.19</td>
<td>56.85</td>
<td>75.15</td>
</tr>
</tbody>
</table>

Scores with (without) punctuation post-processing.
### Experimental results (English)

<table>
<thead>
<tr>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th></th>
<th></th>
<th>ptb_len40_nopunct</th>
<th></th>
<th></th>
<th>ptb_len40_punct</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
<td>56.98</td>
<td>58.79</td>
<td>65.23</td>
<td>38.07 (52.17)</td>
<td>33.75 (46.1)</td>
<td>51.56 (60.59)</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
<td>± 3.66</td>
<td>± 2.85</td>
<td>± 2.92</td>
<td>± 3.94 (± 3.08)</td>
<td>± 3.33 (± 2.75)</td>
<td>± 3.08 (± 1.94)</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
<td>51.43</td>
<td>54.20</td>
<td>60.94</td>
<td>47.95 (49.07)</td>
<td>41.65 (44.61)</td>
<td>59.34 (59.78)</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00 (± 0.00)</td>
<td>± 0.00 (± 0.00)</td>
<td>± 0.00 (± 0.00)</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
<td>50.26</td>
<td>52.92</td>
<td>59.86</td>
<td>42.66 (47.13)</td>
<td>37.77 (41.37)</td>
<td>55.15 (57.87)</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
<td>± 0.72</td>
<td>± 0.68</td>
<td>± 0.58</td>
<td>± 0.98 (± 1.92)</td>
<td>± 0.84 (± 1.30)</td>
<td>± 0.77 (± 1.36)</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
<td>53.67</td>
<td>57.45</td>
<td>53.67</td>
<td>n/a (62.39)</td>
<td>n/a (52.33)</td>
<td>n/a (62.00)</td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
<td>50.29</td>
<td>53.73</td>
<td>60.03</td>
<td>1.04 (54.30)</td>
<td>4.30 (54.68)</td>
<td>22.70 (58.02)</td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>1.03 (56.80)</td>
<td>4.30 (56.19)</td>
<td>22.63 (67.74)</td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>68.19</td>
<td>56.85</td>
<td>75.15</td>
</tr>
</tbody>
</table>

Evaluating on test sentences with length $\leq 10$.

| PRPN  | 18.08 | 21.73 | 22.85 | 41.99 | 45.50 | 45.36 | 33.25 (42.17) | 33.92 (43.55) | 36.85 (44.43) |
|       | ± 3.66 | ± 3.69 | ± 3.45 | ± 4.05 | ± 3.73 | ± 3.82 | ± 3.20 (± 1.82) | ± 3.27 (± 1.95) | ± 3.03 (± 1.60) |
| URNNG | 34.62 | 38.58 | 38.43 | 35.88 | 39.58 | 39.62 | 36.7 (36.72) | 38.44 (38.84) | 40.11 (40.03) |
|       | ± 2.19 | ± 1.65 | ± 2.07 | ± 0.00 | ± 0.00 | ± 0.00 | ± 0.00 (± 0.00) | ± 0.00 (± 0.00) | ± 0.00 (± 0.00) |
| DIORA | 20.44 | 23.72 | 25.08 | 46.27 | 47.81 | 49.39 | 41.48 (46.94) | 41.56 (46.73) | 44.63 (49.38) |
|       | ± 1.53 | ± 1.66 | ± 1.44 | ± 0.31 | ± 0.33 | ± 0.29 | ± 0.43 (± 1.59) | ± 0.37 (± 1.50) | ± 0.41 (± 1.44) |
| CCL   | 19.08 | 21.56 | 18.68 | 37.41 | 41.67 | 37.98 | n/a (49.70) | n/a (51.51) | n/a (47.46) |
| CCM   | 49.54 | 52.60 | 52.48 | 40.90 | 43.62 | 44.34 | 0.09 (33.15) | 0.54 (36.88) | 5.48 (35.65) |
| LBranch | 6.00 | 8.98 | 11.49 | 6.00 | 8.98 | 11.49 | 4.88 (5.55) | 6.36 (8.30) | 10.01 (11.07) |
| RBranch | 35.88 | 39.58 | 39.61 | 35.88 | 39.58 | 39.61 | 0.07 (35.54) | 0.52 (38.98) | 5.45 (39.3) |
| UBound | 84.41 | 83.32 | 85.34 | 84.41 | 83.32 | 85.34 | 77.76 | 75.06 | 78.96 |

Evaluating on all test sentences.
Experimental results (English)

<table>
<thead>
<tr>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
</tr>
<tr>
<td>DIO RA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
</tr>
<tr>
<td>CCM</td>
<td><strong>62.97</strong></td>
<td><strong>63.35</strong></td>
<td><strong>70.14</strong></td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences:

| PRPN  | 18.08 | 21.73 | 22.85 | 41.99 | 45.50 | 45.36 | 33.25 | 42.17 | 33.92 | 43.55 | 36.85 | 44.43 |
|       | ± 3.66 | ± 3.69 | ± 3.45 | ± 4.05 | ± 3.73 | ± 3.82 | ± 3.20 | ± 1.82 | ± 3.27 | ± 1.95 | ± 3.03 | ± 1.60 |
| URNNG | 34.62 | 38.58 | 38.43 | 35.88 | 39.58 | 39.62 | 36.7 | 36.72 | 38.44 | 38.84 | 40.11 | 40.03 |
|       | ± 2.19 | ± 1.65 | ± 2.07 | ± 0.00 | ± 0.00 | ± 0.00 | ± 0.00 | ± 0.00 | ± 0.00 | ± 0.00 | ± 0.00 |
| DIO RA| 20.44 | 23.72 | 25.08 | **46.27** | **47.81** | **49.39** | **41.48** | **46.94** | **41.56** | **46.73** | **44.63** | **49.38** |
|       | ± 1.53 | ± 1.66 | ± 1.44 | ± 0.31 | ± 0.33 | ± 0.29 | ± 0.43 | ± 1.59 | ± 0.37 | ± 1.50 | ± 0.41 | ± 1.44 |
| CCL   | 19.08 | 21.56 | 18.68 | 37.41 | 41.67 | 37.98 | n/a | (49.70) | n/a | (51.51) | n/a | (47.46) |
| CCM   | **49.54** | **52.60** | **52.48** | 40.90 | 43.62 | 44.34 | 0.09 | 33.15 | 0.54 | 36.88 | 5.48 | 35.65 |
| LBranch | 6.00 | 8.98 | 11.49 | 6.00 | 8.98 | 11.49 | 4.88 | 5.55 | 6.36 | 8.30 | 10.01 | 11.07 |
| RBranch | 35.88 | 39.58 | 39.61 | 35.88 | 39.58 | 39.61 | 0.07 | 35.54 | 0.52 | 38.98 | 5.45 | 39.3 |
| UBound | 84.41 | 83.32 | 85.34 | 84.41 | 83.32 | 85.34 | 77.76 | 75.06 | 78.96 |
# Experimental results

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
<td>56.98</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
<td>± 3.66</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
<td>51.43</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
<td>50.26</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
<td>± 0.72</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
<td>53.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
<td>50.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>51.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>83.20</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length $\leq 10$.

<table>
<thead>
<tr>
<th>Metric</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>18.08</td>
<td>21.73</td>
<td>22.85</td>
</tr>
<tr>
<td></td>
<td>± 3.66</td>
<td>± 3.69</td>
<td>± 3.45</td>
</tr>
<tr>
<td>URNNG</td>
<td>34.62</td>
<td>38.58</td>
<td>38.43</td>
</tr>
<tr>
<td></td>
<td>± 2.19</td>
<td>± 1.65</td>
<td>± 2.07</td>
</tr>
<tr>
<td>DIORA</td>
<td>20.44</td>
<td>23.72</td>
<td>25.08</td>
</tr>
<tr>
<td></td>
<td>± 1.53</td>
<td>± 1.66</td>
<td>± 1.44</td>
</tr>
<tr>
<td>CCL</td>
<td>19.08</td>
<td>21.56</td>
<td>18.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>49.54</td>
<td>52.60</td>
<td>52.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBranch</td>
<td>6.00</td>
<td>8.98</td>
<td>11.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>35.88</td>
<td>39.58</td>
<td>39.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBound</td>
<td>84.41</td>
<td>83.32</td>
<td>85.34</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences.

Punctuation post-processing almost always improve performance.
### Experimental results (English)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
<td>56.98</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
<td>± 3.66</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>52.67</td>
<td>60.41</td>
<td>51.43</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
<td>50.26</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
<td>± 0.72</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
<td>53.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
<td>50.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>51.43</td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>83.20</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length $\leq 10$.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
</tr>
<tr>
<td>PRPN</td>
<td>18.08</td>
<td>21.73</td>
<td>22.85</td>
<td>41.99</td>
</tr>
<tr>
<td></td>
<td>± 3.66</td>
<td>± 3.69</td>
<td>± 3.45</td>
<td>± 4.05</td>
</tr>
<tr>
<td>URNNG</td>
<td>34.62</td>
<td>38.58</td>
<td>38.43</td>
<td>35.88</td>
</tr>
<tr>
<td></td>
<td>± 2.19</td>
<td>± 1.65</td>
<td>± 2.07</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>20.44</td>
<td>23.72</td>
<td>25.08</td>
<td>46.27</td>
</tr>
<tr>
<td></td>
<td>± 1.53</td>
<td>± 1.66</td>
<td>± 1.44</td>
<td>± 0.31</td>
</tr>
<tr>
<td>CCL</td>
<td>19.08</td>
<td>21.56</td>
<td>18.68</td>
<td>37.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>49.54</td>
<td>52.60</td>
<td>52.48</td>
<td>40.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBranch</td>
<td>6.00</td>
<td>8.98</td>
<td>11.49</td>
<td>6.00</td>
</tr>
<tr>
<td>RBranch</td>
<td>35.88</td>
<td>39.58</td>
<td>39.61</td>
<td>35.88</td>
</tr>
<tr>
<td>UBound</td>
<td>84.41</td>
<td>83.32</td>
<td>85.34</td>
<td>84.41</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences.

Different metrics produce very different scores.
### Experimental results (English)

<table>
<thead>
<tr>
<th></th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metric</strong></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
</tr>
</tbody>
</table>

**Evaluated on test sentences with length ≤ 10.**

<table>
<thead>
<tr>
<th></th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metric</strong></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>18.08</td>
<td>21.73</td>
<td>22.85</td>
</tr>
<tr>
<td></td>
<td>± 3.66</td>
<td>± 3.69</td>
<td>± 3.45</td>
</tr>
<tr>
<td>URNNG</td>
<td>34.62</td>
<td>38.58</td>
<td>38.43</td>
</tr>
<tr>
<td></td>
<td>± 2.19</td>
<td>± 1.65</td>
<td>± 2.07</td>
</tr>
<tr>
<td>DIORA</td>
<td>20.44</td>
<td>23.72</td>
<td>25.08</td>
</tr>
<tr>
<td></td>
<td>± 1.53</td>
<td>± 1.66</td>
<td>± 1.44</td>
</tr>
<tr>
<td>CCL</td>
<td>19.08</td>
<td>21.56</td>
<td>18.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>49.54</td>
<td>52.60</td>
<td>52.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBranch</td>
<td>6.00</td>
<td>8.98</td>
<td>11.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>35.88</td>
<td>39.58</td>
<td>39.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBound</td>
<td>84.41</td>
<td>83.32</td>
<td>85.34</td>
</tr>
</tbody>
</table>

**Evaluated on all test sentences.**

Neural methods tend to have high variance.
### Experimental results (English)

<table>
<thead>
<tr>
<th>Metric</th>
<th>prb_len10_nopunct</th>
<th>prb_len40_nopunct</th>
<th>prb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>Evaluated on test sentences with length $\leq 10$.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
</tr>
<tr>
<td>URNN</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences.

| PRPN   | 18.08  | 21.73  | 22.85 | 41.99  | 45.50  | 45.36 | 33.25  | 42.17  | 36.85 |
|        | ± 3.66 | ± 3.69 | ± 3.45 | ± 4.05 | ± 3.73 | ± 3.82 | ± 3.20 | ± 1.82  | ± 3.03 |
| URNN   | 34.62  | 38.58  | 38.43 | 35.88  | 39.58  | 39.62 | 36.7   | 36.72  | 40.11 |
|        | ± 2.19 | ± 1.65 | ± 2.07 | ± 0.00 | ± 0.00 | ± 0.00 | ± 0.00 | ± 0.00  | ± 0.00 |
| DIORA  | 20.44  | 23.72  | 25.08 | 46.27  | 47.81  | 49.39 | 41.48  | 46.94  | 44.63 |
|        | ± 1.53 | ± 1.66 | ± 1.44 | ± 0.31 | ± 0.33 | ± 0.29 | ± 0.43 | ± 1.59  | ± 0.41 |
| CCL    | 19.08  | 21.56  | 18.68 | 37.41  | 41.67  | 37.98 | n/a    | (49.70) | n/a    |
|        |        |        |       |        |        |       |        |        |       |
| CCM    | 49.54  | 52.60  | 52.48 | 40.90  | 43.62  | 44.34 | 0.09   | 33.15  | 5.48  |
|        |        |        |       |        |        |       |        |        |       |
| LBranch| 6.00   | 8.98   | 11.49 | 6.00   | 8.98   | 11.49 | 4.88   | 5.55   | 10.01 |
|        |        |        |       |        |        |       |        |        |       |
| RBranch| 35.88  | 39.58  | 39.61 | 35.88  | 39.58  | 39.61 | 0.07   | 35.54  | 5.45  |
|        |        |        |       |        |        |       |        |        |       |
| UBound | 84.41  | 83.32  | 85.34 | 84.41  | 83.32  | 85.34 | 77.76  | 75.06  | 78.96 |

URNNG seems to degenerate to right branching probably because it’s not fully tuned due to the limited computational budget.
## Experimental results (English)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
<td>56.98</td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
<td>± 3.66</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
<td>51.43</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
<td>50.26</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
<td>± 0.72</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>32.55</td>
<td>52.67</td>
</tr>
<tr>
<td>CCM</td>
<td>62.97</td>
<td>63.35</td>
<td>70.14</td>
<td>50.29</td>
</tr>
<tr>
<td></td>
<td>± 2.27</td>
<td>± 0.27</td>
<td>± 0.27</td>
<td>± 0.72</td>
</tr>
<tr>
<td>LBChance</td>
<td>12.22</td>
<td>22.20</td>
<td>30.27</td>
<td>13.32</td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>51.43</td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>83.20</td>
</tr>
</tbody>
</table>

Evaluating on all test sentences:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
</tr>
<tr>
<td>PRPN</td>
<td>18.08</td>
<td>21.73</td>
<td>22.85</td>
<td>41.99</td>
</tr>
<tr>
<td></td>
<td>± 3.66</td>
<td>± 3.69</td>
<td>± 3.45</td>
<td>± 4.05</td>
</tr>
<tr>
<td>URNNG</td>
<td>34.62</td>
<td>38.58</td>
<td>38.43</td>
<td>35.88</td>
</tr>
<tr>
<td></td>
<td>± 2.19</td>
<td>± 1.65</td>
<td>± 2.07</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>20.44</td>
<td>23.72</td>
<td>25.08</td>
<td>46.27</td>
</tr>
<tr>
<td></td>
<td>± 1.53</td>
<td>± 1.66</td>
<td>± 1.44</td>
<td>± 0.31</td>
</tr>
<tr>
<td>CCL</td>
<td>19.08</td>
<td>21.56</td>
<td>18.68</td>
<td>37.41</td>
</tr>
<tr>
<td>CCM</td>
<td>49.54</td>
<td>52.60</td>
<td>52.48</td>
<td>40.09</td>
</tr>
<tr>
<td>LBChance</td>
<td>6.00</td>
<td>8.98</td>
<td>11.49</td>
<td>6.00</td>
</tr>
<tr>
<td>RBranch</td>
<td>35.88</td>
<td>39.58</td>
<td>39.61</td>
<td>35.88</td>
</tr>
<tr>
<td>UBound</td>
<td>84.41</td>
<td>83.32</td>
<td>85.34</td>
<td>84.41</td>
</tr>
</tbody>
</table>

Statistical methods also perform well under several settings.
### Experimental results (English)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
</tr>
<tr>
<td>PRPN</td>
<td>31.29</td>
<td>37.29</td>
<td>44.72</td>
<td><strong>56.98</strong></td>
</tr>
<tr>
<td></td>
<td>± 4.49</td>
<td>± 5.04</td>
<td>± 3.59</td>
<td>± 3.66</td>
</tr>
<tr>
<td>URNNG</td>
<td>50.77</td>
<td>53.67</td>
<td>60.41</td>
<td>51.43</td>
</tr>
<tr>
<td></td>
<td>± 1.11</td>
<td>± 0.83</td>
<td>± 0.89</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>31.55</td>
<td>37.90</td>
<td>44.93</td>
<td>50.26</td>
</tr>
<tr>
<td></td>
<td>± 2.50</td>
<td>± 2.13</td>
<td>± 2.00</td>
<td>± 0.72</td>
</tr>
<tr>
<td>CCL</td>
<td>28.31</td>
<td>36.61</td>
<td>33.55</td>
<td>53.67</td>
</tr>
<tr>
<td><strong>CCM</strong></td>
<td><strong>62.97</strong></td>
<td><strong>63.35</strong></td>
<td><strong>70.14</strong></td>
<td>50.29</td>
</tr>
<tr>
<td>RBranch</td>
<td>51.43</td>
<td>54.20</td>
<td>60.79</td>
<td>51.43</td>
</tr>
<tr>
<td>UBound</td>
<td>83.20</td>
<td>78.74</td>
<td>86.64</td>
<td>83.20</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length ≤ 10.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>ptb_len10_nopunct</th>
<th>ptb_len40_nopunct</th>
<th>ptb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
</tr>
<tr>
<td>PRPN</td>
<td>18.08</td>
<td>21.73</td>
<td>22.85</td>
<td>41.99</td>
</tr>
<tr>
<td></td>
<td>± 3.66</td>
<td>± 3.69</td>
<td>± 3.45</td>
<td>± 4.05</td>
</tr>
<tr>
<td>URNNG</td>
<td>34.62</td>
<td>38.58</td>
<td>38.43</td>
<td>35.88</td>
</tr>
<tr>
<td></td>
<td>± 2.19</td>
<td>± 1.65</td>
<td>± 2.07</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>20.44</td>
<td>23.72</td>
<td>25.08</td>
<td><strong>46.27</strong></td>
</tr>
<tr>
<td></td>
<td>± 1.53</td>
<td>± 1.66</td>
<td>± 1.44</td>
<td>± 0.31</td>
</tr>
<tr>
<td>CCL</td>
<td>19.08</td>
<td>21.56</td>
<td>18.68</td>
<td>37.41</td>
</tr>
<tr>
<td><strong>CCM</strong></td>
<td><strong>49.54</strong></td>
<td><strong>52.60</strong></td>
<td><strong>52.48</strong></td>
<td>40.90</td>
</tr>
<tr>
<td>LBranch</td>
<td>6.00</td>
<td>8.98</td>
<td>11.49</td>
<td>6.00</td>
</tr>
<tr>
<td>RBranch</td>
<td>35.88</td>
<td>39.58</td>
<td>39.61</td>
<td>35.88</td>
</tr>
<tr>
<td>UBound</td>
<td>84.41</td>
<td>83.32</td>
<td>85.34</td>
<td>84.41</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences.

CCM even generalizes to longer sentences.
# Experimental results (Japanese)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Train</th>
<th>ktb_len10_nopunct</th>
<th>ktb_len40_nopunct</th>
<th>ktb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
<td>micro</td>
</tr>
<tr>
<td>PRPN</td>
<td>10.18</td>
<td>23.72</td>
<td>30.48</td>
<td>14.29</td>
</tr>
<tr>
<td></td>
<td>± 2.75</td>
<td>± 2.17</td>
<td>± 2.12</td>
<td>± 11.95</td>
</tr>
<tr>
<td>URNNG</td>
<td>1.37</td>
<td>16.60</td>
<td>23.43</td>
<td>1.93</td>
</tr>
<tr>
<td></td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>21.96</td>
<td>32.37</td>
<td>39.60</td>
<td>34.69</td>
</tr>
<tr>
<td></td>
<td>± 6.59</td>
<td>± 5.35</td>
<td>± 5.10</td>
<td>± 6.51</td>
</tr>
<tr>
<td>CCL</td>
<td>18.49</td>
<td>30.31</td>
<td>32.28</td>
<td>2.74</td>
</tr>
<tr>
<td></td>
<td>± 2.55</td>
<td>± 1.71</td>
<td>± 1.72</td>
<td>± 1.98</td>
</tr>
<tr>
<td>CCM</td>
<td>24.69</td>
<td>36.32</td>
<td>41.72</td>
<td>32.67</td>
</tr>
<tr>
<td></td>
<td>± 6.59</td>
<td>± 5.35</td>
<td>± 5.10</td>
<td>± 6.51</td>
</tr>
<tr>
<td>LBranch</td>
<td>23.86</td>
<td>34.69</td>
<td>41.07</td>
<td>23.86</td>
</tr>
<tr>
<td></td>
<td>± 1.71</td>
<td>± 1.72</td>
<td>± 1.72</td>
<td>± 1.98</td>
</tr>
<tr>
<td>RBranch</td>
<td>1.37</td>
<td>16.60</td>
<td>23.67</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
</tr>
<tr>
<td>UBound</td>
<td>57.68</td>
<td>60.82</td>
<td>67.25</td>
<td>57.68</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length ≤ 10.

| Metric | PRPN   | 8.01 ± 1.19 | 13.92 ± 1.28 | 15.61 ± 1.09 | 11.11 ± 8.06 | 17.25 ± 8.82 | 18.45 ± 7.39 | 5.83 ± 7.15 | 10.16 ± 12.17 | 13.1 ± 14.07 |
|        | ± 2.63 | ± 2.63 | ± 2.63 | ± 2.63 | ± 2.63 | ± 2.63 | ± 2.63 | ± 2.63 | ± 2.63 | ± 2.63 |
| URNNG  | 0.24 ± 0.00 | 6.44 ± 0.00 | 8.47 ± 0.00 | 0.68 ± 0.00 | 6.94 ± 0.00 | 8.87 ± 0.00 | 0.33 ± 0.26 | 5.08 ± 5.6 | 8.01 ± 7.95 | ± 6.00 ± 0.00 |
| CCL    | 12.62 | 19.43 | 18.03 | 1.20 | 7.69 | 12.60 | n/a (8.63) | n/a (14.18) | n/a (18.44) |
| CCM    | 12.21 | 21.70 | 19.46 | 20.21 | 28.60 | 26.80 | 1.33 (1.42) | 5.91 (6.78) | 8.94 (8.98) |
| RBranch| 0.22 | 6.43 | 8.46 | 0.22 | 6.43 | 8.46 | 0.20 (0.17) | 4.83 (5.45) | 7.89 (8.54) |
| UBound | 64.38 | 62.52 | 67.32 | 64.38 | 62.52 | 67.32 | 59.40 | 56.44 | 62.53 |
# Experimental results (Japanese)

<table>
<thead>
<tr>
<th>Metric</th>
<th>ktb_len10_nopunct</th>
<th>ktb_len40_nopunct</th>
<th>ktb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>10.18</td>
<td>23.72</td>
<td>30.48</td>
</tr>
<tr>
<td></td>
<td>± 2.75</td>
<td>± 2.17</td>
<td>± 2.12</td>
</tr>
<tr>
<td>URNNG</td>
<td>1.37</td>
<td>16.60</td>
<td>23.43</td>
</tr>
<tr>
<td></td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>21.96</td>
<td>32.37</td>
<td>39.60</td>
</tr>
<tr>
<td></td>
<td>± 6.59</td>
<td>± 5.35</td>
<td>± 5.10</td>
</tr>
<tr>
<td>CCL</td>
<td>18.49</td>
<td>30.31</td>
<td>32.28</td>
</tr>
<tr>
<td>CCM</td>
<td>24.69</td>
<td>36.32</td>
<td><strong>41.72</strong></td>
</tr>
<tr>
<td></td>
<td>± 7.15</td>
<td>± 4.45</td>
<td>± 7.82</td>
</tr>
<tr>
<td>LBranch</td>
<td>23.86</td>
<td>34.69</td>
<td>41.07</td>
</tr>
<tr>
<td></td>
<td>± 6.59</td>
<td>± 5.35</td>
<td>± 5.10</td>
</tr>
<tr>
<td>RBranch</td>
<td>1.37</td>
<td>16.60</td>
<td>23.67</td>
</tr>
<tr>
<td>UBound</td>
<td>57.68</td>
<td>60.82</td>
<td>67.25</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length ≤ 10.

<table>
<thead>
<tr>
<th>Metric</th>
<th>ktb_len10_nopunct</th>
<th>ktb_len40_nopunct</th>
<th>ktb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>8.01</td>
<td>13.92</td>
<td>15.61</td>
</tr>
<tr>
<td></td>
<td>± 1.19</td>
<td>± 1.28</td>
<td>± 1.09</td>
</tr>
<tr>
<td>URNNG</td>
<td>0.24</td>
<td>6.44</td>
<td>8.47</td>
</tr>
<tr>
<td></td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>14.95</td>
<td>21.90</td>
<td><strong>21.97</strong></td>
</tr>
<tr>
<td></td>
<td>± 3.22</td>
<td>± 4.19</td>
<td>± 2.95</td>
</tr>
<tr>
<td>CCL</td>
<td>12.62</td>
<td>19.43</td>
<td>18.03</td>
</tr>
<tr>
<td>CCM</td>
<td>12.21</td>
<td>21.70</td>
<td>19.46</td>
</tr>
<tr>
<td></td>
<td>± 6.59</td>
<td>± 4.45</td>
<td>± 7.82</td>
</tr>
<tr>
<td></td>
<td>± 6.59</td>
<td>± 4.45</td>
<td>± 7.82</td>
</tr>
<tr>
<td>RBranch</td>
<td>0.22</td>
<td>6.43</td>
<td>8.46</td>
</tr>
<tr>
<td>UBound</td>
<td>64.38</td>
<td>62.52</td>
<td>67.32</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences.

Left branching is strong on Japanese
### Experimental results (Japanese)

<table>
<thead>
<tr>
<th>Metric</th>
<th>ktb_len10_nopunct</th>
<th>ktb_len40_nopunct</th>
<th>ktb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>10.18</td>
<td>23.72</td>
<td>30.48</td>
</tr>
<tr>
<td></td>
<td>± 2.75</td>
<td>± 2.17</td>
<td>± 2.12</td>
</tr>
<tr>
<td>URNNG</td>
<td>1.37</td>
<td>16.60</td>
<td>23.43</td>
</tr>
<tr>
<td></td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
</tr>
<tr>
<td>DIORA</td>
<td>21.96</td>
<td>32.37</td>
<td>39.60</td>
</tr>
<tr>
<td></td>
<td>± 6.59</td>
<td>± 5.35</td>
<td>± 5.10</td>
</tr>
<tr>
<td>CCL</td>
<td>18.49</td>
<td>30.31</td>
<td>32.28</td>
</tr>
<tr>
<td>CCM</td>
<td>24.69</td>
<td>36.32</td>
<td>41.72</td>
</tr>
<tr>
<td></td>
<td>± 1.12</td>
<td>± 1.28</td>
<td>± 1.09</td>
</tr>
<tr>
<td>LBranch</td>
<td>23.86</td>
<td>34.69</td>
<td>41.07</td>
</tr>
<tr>
<td>RBranch</td>
<td>1.37</td>
<td>16.60</td>
<td>23.67</td>
</tr>
<tr>
<td>UBound</td>
<td>57.68</td>
<td>60.82</td>
<td>67.25</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences.

<table>
<thead>
<tr>
<th>Metric</th>
<th>PRPN</th>
<th>URNNG</th>
<th>DIORA</th>
<th>CCL</th>
<th>CCM</th>
<th>LBranch</th>
<th>RBranch</th>
<th>UBound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8.01</td>
<td>0.24</td>
<td>14.95</td>
<td>12.62</td>
<td>12.21</td>
<td>11.15</td>
<td>0.22</td>
<td>64.38</td>
</tr>
<tr>
<td></td>
<td>± 1.19</td>
<td>± 0.00</td>
<td>± 3.22</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 1.12</td>
<td>± 0.00</td>
<td>± 2.52</td>
</tr>
<tr>
<td></td>
<td>± 1.28</td>
<td>± 0.00</td>
<td>± 4.19</td>
<td>± 1.09</td>
<td>± 1.09</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.46</td>
</tr>
<tr>
<td></td>
<td>15.61</td>
<td>8.47</td>
<td>21.97</td>
<td>18.03</td>
<td>19.46</td>
<td>18.49</td>
<td>8.46</td>
<td>67.32</td>
</tr>
<tr>
<td></td>
<td>± 1.09</td>
<td>± 0.00</td>
<td>± 2.95</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 0.00</td>
<td>± 3.72</td>
</tr>
</tbody>
</table>

Performance of PRPN drops a lot, due to its right branching bias during inference.
## Experimental results (Japanese)

<table>
<thead>
<tr>
<th>Metric</th>
<th>ktb_len10_nopunct</th>
<th>ktb_len40_nopunct</th>
<th>ktb_len40_punct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>micro</td>
<td>macro</td>
<td>evalb</td>
</tr>
<tr>
<td>PRPN</td>
<td>10.18 (2.75)</td>
<td>23.72 (2.17)</td>
<td>30.48 (2.12)</td>
</tr>
<tr>
<td>URNNG</td>
<td>1.37 (0.00)</td>
<td>16.60 (0.00)</td>
<td>23.43 (0.00)</td>
</tr>
<tr>
<td><strong>DIORA</strong></td>
<td><strong>21.96 (6.59)</strong></td>
<td><strong>32.37 (5.35)</strong></td>
<td><strong>39.60 (5.10)</strong></td>
</tr>
<tr>
<td>CCL</td>
<td>18.49</td>
<td>30.31</td>
<td>32.28</td>
</tr>
<tr>
<td>CCM</td>
<td><strong>24.69</strong></td>
<td><strong>36.32</strong></td>
<td><strong>41.72</strong></td>
</tr>
<tr>
<td>LBranch</td>
<td>23.86</td>
<td>34.69</td>
<td>41.07</td>
</tr>
<tr>
<td>RBranch</td>
<td>1.37</td>
<td>16.60</td>
<td>23.67</td>
</tr>
<tr>
<td>UBound</td>
<td>57.68</td>
<td>60.82</td>
<td>67.25</td>
</tr>
</tbody>
</table>

Evaluated on test sentences with length ≤ 10.

<table>
<thead>
<tr>
<th></th>
<th>micro</th>
<th>macro</th>
<th>evalb</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRPN</td>
<td>8.01 (1.19)</td>
<td>13.92 (1.28)</td>
<td>15.61 (1.09)</td>
</tr>
<tr>
<td>URNNG</td>
<td>0.24 (0.00)</td>
<td>6.44 (0.00)</td>
<td>8.47 (0.00)</td>
</tr>
<tr>
<td><strong>DIORA</strong></td>
<td><strong>14.95 (3.22)</strong></td>
<td><strong>21.90 (4.19)</strong></td>
<td><strong>21.97 (2.95)</strong></td>
</tr>
<tr>
<td>CCL</td>
<td>12.62</td>
<td>19.43</td>
<td>18.03</td>
</tr>
<tr>
<td>CCM</td>
<td>12.21</td>
<td>21.70</td>
<td>19.46</td>
</tr>
<tr>
<td>LBranch</td>
<td>11.15</td>
<td>20.62</td>
<td>18.49</td>
</tr>
<tr>
<td>RBranch</td>
<td>0.22</td>
<td>6.43</td>
<td>8.46</td>
</tr>
<tr>
<td>UBound</td>
<td>64.38</td>
<td>62.52</td>
<td>67.32</td>
</tr>
</tbody>
</table>

Evaluated on all test sentences.

DIORA and CCM have the best performance.
Conclusion

- We propose a standardized experimental setup for unsupervised constituency parsing
- We empirically compare five methods and find that recent models do not show a clear advantage over decade-old models
Thank you!