Dependency Language Models for Transition-based Dependency Parsing

Juntao Yu (juntao.cn@gmail.com)
Bernd Bohnet
Overview

1. Semi-supervised techniques for parsing
2. Dependency language models
3. Experimental set-up
4. Empirical results
5. Conclusions
Outline of the problem

Expensive and time consuming

Labeled Data VS Unlabeled Data

Cheap and widely accessible
Semi-supervised approaches

• Boosting the training set
  • co-training (Sarkar, 2001; Steedman et al., 2003; Sagea and Tsujii, 2007)
  • self-training (McClosky et al., 2006a; Reichart and Rappoport, 2007; Petrov and McDonald, 2012)

• Using features learned from unlabeled data
  • word clustering (Koo et al., 2008; Cerisara, 2014)
  • word embeddings (Chen and Manning, 2014; Weiss et al., 2015; Dozat and Manning, 2017)
Dependency language models

- Annotated corpora
- Training
- Base model
- Labelling
- Auto-parsed data
- Dependency language models
- Annotated corpora
- Re-training
- Enhanced model
Dependency language models

- Annotated corpora
  - Training
- Base model
- Labelling
- Auto-parsed data
- Dependency language models
- Annotated corpora
  - Re-training
- Enhanced model
Dependency language models

1. Annotated corpora
   - Training
   - Labelling
   - Auto-parsed data

2. Base model
   - Dependency language models

3. Annotated corpora
   - Re-training
   - Enhanced model
Dependency language models

\[ P_u(x_{ch}|H) = \frac{\text{count}(x_{ch}, H)}{\sum x'_c \text{count}(x'_c, H)} \]

PH 10%
PM 10%-30%
PL 30%-100%
PO Unknown
Dependency language models

Annotated corpora → Training → Base model

Auto-parsed data → Labelling → Dependency language models

Dependency language models → Re-training → Annotated corpora

Annotated corpora → Enhanced model
Dependency language models

• Differences from Chen et al. (2012)

Graph-based VS Transition-based
(92.10 UAS) (93.38 UAS)
Dependency language models

- Differences from Chen et al. (2012)

Unlabeled parsing VS Labeled parsing
Dependency language models

• Differences from Chen et al. (2012)

Pipeline approach for tagging and parsing

VS

Joint tagging and parsing
Dependency language models

- Differences from Chen et al. (2012)

Single DLM

VS

Single and multiple DLMs
Dependency language models

• Differences from Chen et al. (2012)

  Single parsed text

  VS

  Single and double parsed text
Dependency language models

• Differences from Chen et al. (2012)

\[
\begin{align*}
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_0\_pos > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_0\_word > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_1\_pos > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_1\_word > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_0\_pos, s_1\_pos > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_0\_word, s_1\_word >
\end{align*}
\]
Experimental set-up:

• Mate transition-based parser (Bohnet et al., 2013)

• **DataSet (PTB and CTB5):**
  - English: Stanford v3.3.0
  - Chinese: Zhang and Nivre (2011)

• **Unlabeled set:**
  - English: Chelba et al. (2013) 30M sentences (800M words)
  - Chinese: Gigaword v5.0 20M sentences (450M words)

<table>
<thead>
<tr>
<th></th>
<th>train</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTB</td>
<td>2-21</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>CTB5</td>
<td>001-815,</td>
<td>886-931,</td>
<td>816-885,</td>
</tr>
<tr>
<td></td>
<td>1001-1136</td>
<td>1148-1151</td>
<td>1137-1147</td>
</tr>
</tbody>
</table>
Evaluating on dev set:

<table>
<thead>
<tr>
<th>$m$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>1-2</th>
<th>1-3</th>
<th>1-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>91.05</td>
<td>91.43</td>
<td>91.14</td>
<td>91.22</td>
<td>91.27</td>
<td>91.26</td>
<td>N/A</td>
</tr>
<tr>
<td>Chinese</td>
<td>78.95</td>
<td>79.85</td>
<td>79.42</td>
<td>79.06</td>
<td>79.97</td>
<td>80.11</td>
<td>79.73</td>
</tr>
</tbody>
</table>

Table 3: Effects (LAS) of different number of DLMs for English and Chinese. $m = 0$ refers the baseline.
Evaluating on dev set:

7M English high quality sentences: 91.56

<table>
<thead>
<tr>
<th>Size</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>91.05</td>
<td>91.43</td>
<td>91.38</td>
<td>91.13</td>
<td>91.28</td>
</tr>
<tr>
<td>Chinese</td>
<td>78.95</td>
<td>80.11</td>
<td>80.15</td>
<td>79.72</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 4: Effects (LAS) of DLMs extracted from different size (in million sentences) of corpus. Size = 0 refers the baseline.
Evaluating on test set (English):

<table>
<thead>
<tr>
<th>System</th>
<th>Beam</th>
<th>POS</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang and Nivre (2011)</td>
<td>32</td>
<td>97.44</td>
<td>90.95</td>
<td>93.00</td>
</tr>
<tr>
<td>Bohnet and Kuhn (2012)</td>
<td>80</td>
<td>97.44</td>
<td>91.19</td>
<td>93.27</td>
</tr>
<tr>
<td>Martins et al. (2013)</td>
<td>N/A</td>
<td>97.44</td>
<td>90.55</td>
<td>92.89</td>
</tr>
<tr>
<td>Zhang and McDonald (2014)</td>
<td>N/A</td>
<td>97.44</td>
<td>91.02</td>
<td>93.22</td>
</tr>
<tr>
<td>Chen and Manning (2014)†</td>
<td>1</td>
<td>N/A</td>
<td>89.60</td>
<td>91.80</td>
</tr>
<tr>
<td>Dyer et al. (2015)†</td>
<td>1</td>
<td>97.30</td>
<td>90.90</td>
<td>93.10</td>
</tr>
<tr>
<td>Weiss et al. (2015)†</td>
<td>8</td>
<td>97.44</td>
<td>92.05</td>
<td>93.99</td>
</tr>
<tr>
<td>Andor et al. (2016)†</td>
<td>32</td>
<td>97.44</td>
<td>92.79</td>
<td>94.61</td>
</tr>
<tr>
<td>Dozat and Manning (2017)†</td>
<td>N/A</td>
<td>N/A</td>
<td>94.08</td>
<td>95.74</td>
</tr>
<tr>
<td>Liu and Zhang (2017)†</td>
<td>N/A</td>
<td>N/A</td>
<td><strong>95.20</strong></td>
<td><strong>96.20</strong></td>
</tr>
<tr>
<td>Our Baseline</td>
<td>40</td>
<td>97.36</td>
<td>90.95</td>
<td>93.08</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>97.34</td>
<td>91.05</td>
<td>93.28</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>97.34</td>
<td>91.05</td>
<td>93.29</td>
</tr>
<tr>
<td>Our DLM</td>
<td>40</td>
<td>97.38</td>
<td>91.41</td>
<td>93.59</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>97.39</td>
<td>91.47</td>
<td>93.65</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>97.42</td>
<td>91.56</td>
<td>93.74</td>
</tr>
</tbody>
</table>

Table 5: Comparing with top performing parsers on English.
Evaluating on test set (Chinese):

<table>
<thead>
<tr>
<th>System</th>
<th>Beam</th>
<th>POS</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hatori et al. (2011)</td>
<td>64</td>
<td>93.94</td>
<td>N/A</td>
<td>81.33</td>
</tr>
<tr>
<td>Li et al. (2012)</td>
<td>N/A</td>
<td>94.60</td>
<td>79.01</td>
<td>81.67</td>
</tr>
<tr>
<td>Chen et al. (2013)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>83.08</td>
</tr>
<tr>
<td>Chen et al. (2015)</td>
<td>N/A</td>
<td>93.61</td>
<td>N/A</td>
<td>82.94</td>
</tr>
<tr>
<td>Our Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>93.99</td>
<td>78.49</td>
<td>81.52</td>
</tr>
<tr>
<td>80</td>
<td></td>
<td>94.02</td>
<td>78.48</td>
<td>81.58</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>93.98</td>
<td>78.96</td>
<td>82.11</td>
</tr>
<tr>
<td>Our DLM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>94.27</td>
<td>79.42</td>
<td>82.51</td>
</tr>
<tr>
<td>80</td>
<td></td>
<td>94.39</td>
<td>79.79</td>
<td>82.79</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>94.40</td>
<td><strong>80.21</strong></td>
<td><strong>83.28</strong></td>
</tr>
</tbody>
</table>

Table 6: Comparing with top performing parsers on Chinese.
Conclusions:

• Dependency language models significantly improved both English and Chinese accuracies (both LAS and UAS) for a strong transition-based parser.
  • 0.5% improvements on English.
  • 1% improvements on Chinese and achieved the state-of-the-art results.

• 0.4% improvements on Chinese part-of-speech tagging accuracy and constantly better performance on English part-of-speech tagging.
Questions?