**Supervised Treebank Conversion: Data and Approaches (ACL-2018)**

Xinzhou Jiang†, Bo Zhang†, Zhenghua Li†, Min Zhang†, Sheng Li†, Luo Si‡,
† Soochow University, Suzhou, China; ‡ Alibaba Inc., Hangzhou, China
zhli13@suda.edu.cn; http://hlt.suda.edu.cn/~zhli

**Motivation**
- Many languages have multi-heterogeneous treebanks, with new ones being annotated all the time.
- To meet the needs of certain applications, or
- Motivated by different linguistic philosophies

**Questions**
- Can we accurately convert a guideline-A tree to a guideline-B tree? (the conversion task)
- Can we improve parsing accuracy by enlarging training data with converted data? (the parsing task)

**Previous works on multi-treebank exploitation**
- The indirect guide-feature method (stacked learning) (Li+, ACL-2012)
- The indirect multi-task learning (MTL) method (Guo+, COLING-2016)
- The direct treebank conversion method
- Unsupervised (Niu+, ACL-2009)
- Using pseudo bi-tree aligned data (Zhu+ ACL-2011; Li+ ACL-2013)

**Contributions of this work**
- First propose the task of supervised treebank conversion
- Manually annotate a bi-tree aligned dataset for training and evaluation
- Propose two simple yet effective approaches

**Annotation of bi-tree aligned data**
- Guideline (62 pages)
  http://hlt.suda.edu.cn/index.php/SUCDT
- Annotation platform
  http://101.132.166.249/anno-sys
- Annotators: 15 part-time students
- Cost (~10K sentences): ~500 person-hours
- Overall Accuracy: 78.6%
- Inter-annotator consistency
  - Dependency-wise: 78.5%
  - Sentence-wise: 43.7%

**Approaches: pattern embedding & treeLSTM**
- The key is how to make full use of the source-side tree as guidance.

**Experiments**

<table>
<thead>
<tr>
<th>Data</th>
<th>#Sent</th>
<th>#Tok (HIT)</th>
<th>#Tok (our)</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>7,768</td>
<td>119,707</td>
<td>36,348</td>
</tr>
<tr>
<td>dev</td>
<td>998</td>
<td>14,863</td>
<td>4,839</td>
</tr>
<tr>
<td>test</td>
<td>1,995</td>
<td>29,975</td>
<td>9,679</td>
</tr>
<tr>
<td>train-HIT</td>
<td>52,450</td>
<td>980,791</td>
<td>36,348</td>
</tr>
</tbody>
</table>

**Conversion results on the test data**

<table>
<thead>
<tr>
<th>Data</th>
<th>Training data</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-task</td>
<td>train &amp; train-HIT</td>
<td>79.29</td>
<td>74.51</td>
</tr>
<tr>
<td>Pattern</td>
<td>train</td>
<td>86.66</td>
<td>82.03</td>
</tr>
<tr>
<td>TreeLSTM</td>
<td>train</td>
<td>86.69</td>
<td>82.09</td>
</tr>
<tr>
<td>Combined</td>
<td>train</td>
<td>86.66</td>
<td>81.82</td>
</tr>
</tbody>
</table>

- Treebank conversion is better than MTL due to the explicit use of the source-side tree.
- The two approaches are similarly effective.

**Results on the fully annotated test data (372 sent)**

- The real parsing and conversion accuracies are very high (over 90%).