Multi-Model and Crosslingual Dependency Analysis

Johannes Heinecke, Munshi Asadullah
Orange Labs, Lannion, France

Architecture

BistParser modifications: One dependency tree per sentence

Training:
- hidden layer size 40, 50 or 100, depending on language
- other BistParser options used: --k 3 --lstmdims 125
  --lstmlayers 2 --bibi-lstm --usehead --userl
- word embeddings for all languages (except Gothic)
  - all words in lowercase (if applicable)
  - punctuation separated from words
  - word2vec standard options except -size {300,500} and -window 10

Surprise Languages

Two crosslingual approaches: training (1) on a mix of 23 languages and (2) on a typologically close language (hsb → cs, sme → fi, kmr → fa, bxr → hi), both without word embeddings: (2) gave much better results.

Example of modified CONLL (cols. 1, 2 and 4) used for training (i.e. cs, shown below left) and prediction (in this case hsb, below right):

<table>
<thead>
<tr>
<th>Training data (cs)</th>
<th>Test data (hsb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Manažeři</td>
<td>1 Njejsu</td>
</tr>
<tr>
<td>2 rozhodují</td>
<td>2 DET</td>
</tr>
<tr>
<td>3 ADV</td>
<td>3 archeologiské</td>
</tr>
<tr>
<td>4 ADP</td>
<td>4 doklady</td>
</tr>
<tr>
<td>5 mistě</td>
<td>5 ADP</td>
</tr>
<tr>
<td>6 PUNCT</td>
<td>6 PUNCT</td>
</tr>
</tbody>
</table>

Results

- 10th position with LAS 68.61% (improved to 69.75% after bug fixes)
- 9th position with Content Word LAS (CLAS) as evaluation metrics: 64.15%
- 8th position on surprise languages: 38.72% (7th position with CLAS: 34.28%)

Runtime (on Tira VM, Ubuntu Xenial): 3 hours (all treebanks), using <16GB

Contact: johannes.heinecke@orange.com