Data-driven Multilingual Coreference Resolution using Resolver Stacking

Anders Björkelund and Richárd Farkas
Institute for Natural Language Processing, University of Stuttgart
Available at http://www.ims.uni-stuttgart.de/~anders/

Approach

• Mention detection
• Non-referential classifier
• Coreference classifier
• Heavy feature engineering
• Disallowing transitive nesting
• Cluster mention decoder
• Resolver stacking

Mention Detection

Arabic: NP + PRP + PRP$
Chinese: NP + PN + NR
English: NP + PRP + PRP$ + NEs - NonRef

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th># Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>75.41</td>
<td>64.92</td>
<td>68.78</td>
<td>34.67</td>
</tr>
<tr>
<td>Chinese</td>
<td>76.53</td>
<td>63.67</td>
<td>69.01</td>
<td>34.67</td>
</tr>
<tr>
<td>English</td>
<td>75.33</td>
<td>63.67</td>
<td>68.78</td>
<td>34.67</td>
</tr>
</tbody>
</table>

Table 1: Performance of the non-referential classifier used for English. Precision, recall, and F-measure are given for each language. The left side uses a probability threshold of 0.5, and the right one a threshold of 0.95. The last column denotes the number of occurrences of the corresponding token. All numbers are computed on the development set.

Decoders and Stacking

• BestFirst (BF),
• Pronouns Closest First (PCF),
• Cluster mention decoder (AMP): score($m_i, m_j$) = $\prod_{m_k \in C} P(\text{coref}(m_i, m_j) | m_k)$

• Stacking: AMP + (BF/PCF)

Transitive Nesting

(1) ... she seemed to have such a good relationship with [[her], mother], like [[her], mother], treated her like a human being ...

(2) [[Taiwan]], ‘s.

Modified decoder to disallow transitive nesting, e.g. Skip linking $(a,d)$, if $(c,d)$ was negative

Additional Experiments

• Training on train+dev only minor improvement (Chinese, English)
• Training on gold syntax and testing on predicted is harmful (Arabic, Chinese, English)
• When testing on gold syntax, the models trained on predicted syntax are much better (Chinese, English)
• Gold boundaries are worse than predicted boundaries, even with gold syntax in test data (English)
• Ask for handout with detailed tables!