Discriminants-based Treebanking of the WSJ sections of the Penn Treebank

Valia Kordoni, Yi Zhang, Iliana Simova

Conclusions and Outlook

We are nearing the end: Done annotating sections 00-20

Treebanking efficiency can be improved significantly by discriminant ranking without hurting annotation quality

Discriminant ranking models for training new annotators

Before ...

and after Discriminant Ranking

Goal

To (re-)treebank all sentences in the WSJ sections of the PTB with the latest version of ERG. It allows for a closer comparison with other parsing studies, and contributes independent deep syntactic and semantic annotations to (probably) the most annotated set of linguistically non-trivial real texts. There are initiatives to do the treebanking in multiple frameworks, also with the same texts translated into different languages.

Setup

- Started in October 2008
- Multiple annotators (2 on average)
- Intensive and long initial training period (~ half a year)
- Annotation speed peaked at around 35 sentences per hour

Discriminants-Based Treebanking

- Discriminant-based treebanking is efficient
- Observation shows more annotation time is spent on searching for the easily decidable discriminants
- On average ~100 derivation-tree based syntactic discriminants per 500-tree forest
- Only 6% are manually disambiguated

Evaluation

- Multiple timed annotation sessions with multiple annotators
- Annotating PARC 700
- Using different discriminant ranking models
- Measure both annotation speed and inter-annator agreement

Features

- discriminant type
- edge position
- edge category
- level of discrimination
- branch splitting

Evaluation

- Multiple timed annotation sessions with multiple annotators
- Annotating PARC 700
- Using different discriminant ranking models
- Measure both annotation speed and inter-annator agreement

Discriminants Discriminants Discriminants Discriminants

Based Treebanking of the Based Treebanking of the Based Treebanking of the Based Treebanking of the Penn Treebank sections of the Penn Treebank sections of the Penn Treebank sections of the Penn Treebank

Valia Kordoni, Yi Zhang, Iliana Simova

Results

<table>
<thead>
<tr>
<th>Ranking Model</th>
<th>Speed (s/h)</th>
<th>Speed-up (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>61.9</td>
<td>-</td>
</tr>
<tr>
<td>Own-model</td>
<td>96.1</td>
<td>55%</td>
</tr>
<tr>
<td>Peer-model</td>
<td>94.6</td>
<td>53%</td>
</tr>
<tr>
<td>Joint-model</td>
<td>95.0</td>
<td>53%</td>
</tr>
</tbody>
</table>

ITA

<table>
<thead>
<tr>
<th>Ranking Model</th>
<th>Cohen’s KAPPA (κ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5404</td>
</tr>
<tr>
<td>Own-model</td>
<td>0.5798</td>
</tr>
<tr>
<td>Peer-model</td>
<td>0.5567</td>
</tr>
<tr>
<td>Joint-model</td>
<td>0.5536</td>
</tr>
</tbody>
</table>

Conclusions and Outlook

- We are nearing the end: Done annotating sections 00-20
- Treebanking efficiency can be improved significantly by discriminant ranking without hurting annotation quality
- Discriminant ranking models for training new annotators