Tagging Icelandic text: A linguistic rule–based approach

Sebastian Hohmann
Main functions of a tagger:
  ◦ Disambiguation
  ◦ Unknown Word guessing

Two main approaches for disambiguation:
  ◦ Data–Driven
  ◦ Linguistic rule–based
Introduction

- **Data-driven:**
  - Use of a pre-tagged training corpus
  - Language independent
  - Very popular in the last 10–15 years

- **Linguistic rule-based:**
  - Use hand-crafted rules
  - For one specific language
  - Complex and time-consuming
Is it useful to implement a Linguistic rule–based tagger for a morphologically complex language as Icelandic?
Introduction

- IceNLP:
  - OpenSource NLP Toolkit for analyzing and processing Icelandic text
  - Linguistic rule-based tagger: IceTagger
  - Unknown word guesser: IceMorphy
Data-driven tagging methods

- Data-driven:
  - Use of a pre-tagged training corpus
  - Language independent

- TnT tagger
  - Probabilistic trigram tagger

- MXPOST tagger
  - Maximum Entropy approach

- fnTBL tagger
  - Brill tagger (automatically created rules)
Data-driven tagging methods

Tagging Icelandic:
- University of Iceland – Institute of Lexicography
- Using of IFD corpus
  - 590,000 tokens
  - 639 different tags

<table>
<thead>
<tr>
<th>Words/Tagger</th>
<th>fnTBL</th>
<th>MXPOST</th>
<th>TnT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>54.03%</td>
<td>62.50%</td>
<td>71.60%</td>
</tr>
<tr>
<td>Known</td>
<td>91.36%</td>
<td>91.04%</td>
<td>91.74%</td>
</tr>
<tr>
<td>All</td>
<td>88.80%</td>
<td>89.08%</td>
<td>90.36%</td>
</tr>
</tbody>
</table>

Table 3. Average tagging accuracy in the Icelandic tagging experiment.
Data-driven tagging methods

<table>
<thead>
<tr>
<th>Words/Tagger</th>
<th>fnTBL</th>
<th>MXPOST</th>
<th>TnT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>54.03%</td>
<td>62.50%</td>
<td>71.60%</td>
</tr>
<tr>
<td>Known</td>
<td>91.36%</td>
<td>91.04%</td>
<td>91.74%</td>
</tr>
<tr>
<td>All</td>
<td>88.80%</td>
<td>89.08%</td>
<td>90.36%</td>
</tr>
</tbody>
</table>

*Table 3. Average tagging accuracy in the Icelandic tagging experiment.*

- Experiment for Swedish with smaller corpus: 93.55%
- Tags per token in Swedish: 2.05
- Tags per token in Icelandic: 2.74
Data-driven tagging methods

<table>
<thead>
<tr>
<th>Words/Tagger</th>
<th>fnTBL</th>
<th>MXPOST</th>
<th>TnT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>54.03%</td>
<td>62.50%</td>
<td>71.60%</td>
</tr>
<tr>
<td>Known</td>
<td>91.36%</td>
<td>91.04%</td>
<td>91.74%</td>
</tr>
<tr>
<td>All</td>
<td>88.80%</td>
<td>89.08%</td>
<td>90.36%</td>
</tr>
</tbody>
</table>

*Table 3. Average tagging accuracy in the Icelandic tagging experiment.*

- Ambiguity in IFD corpus: 59.7%
- Ambiguity in Brown corpus (English): 35%
Rule-based tagging methods

- Linguistic rule-based:
  - Use hand-crafted rules
  - For one specific language

- IceTagger (175 rules)
- Swedish CG project (2,100 rules)
- EngCG-2 (3,600 rules)
Rule-based tagging methods

- Evaluation for 9 test corporas (containing 90% of the IDF corpus)

<table>
<thead>
<tr>
<th>Words/Tagger</th>
<th>MXPOST</th>
<th>fnTBL</th>
<th>TnT</th>
<th>IceTagger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>62.29%</td>
<td>55.51%</td>
<td>71.68%</td>
<td>75.09%</td>
</tr>
<tr>
<td>Known</td>
<td>91.00%</td>
<td>91.82%</td>
<td>91.82%</td>
<td>92.74%</td>
</tr>
<tr>
<td>All words</td>
<td>89.03%</td>
<td>89.33%</td>
<td>90.44%</td>
<td>91.54%</td>
</tr>
</tbody>
</table>

*Table 6. Average tagging accuracy of IceTagger in comparison to the three data-driven taggers.*
Unknown word guesser

- IceMorphy:
  - Morphological analysis

- Can be called as a stand-alone module
Combination of taggers

- 'simple voting' combination scheme:
  - Each tagger votes for a specific tag
  - The tag with the highest number of votes 'wins'
Combination of taggers increases the accuracy

<table>
<thead>
<tr>
<th>Words/Tagger</th>
<th>fnTBL*</th>
<th>TnT*</th>
<th>IceTagger</th>
<th>Simple voting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>66.30%</td>
<td>72.75%</td>
<td>75.09%</td>
<td>76.57%</td>
</tr>
<tr>
<td>Known</td>
<td>91.90%</td>
<td>92.53%</td>
<td>92.74%</td>
<td>94.15%</td>
</tr>
<tr>
<td>All words</td>
<td>90.15%</td>
<td>91.18%</td>
<td>91.54%</td>
<td>92.95%</td>
</tr>
</tbody>
</table>

Table 11. Tagging accuracy using features of IceMorphy.
Thanks for your attention!