Tagging Icelandic text

An experiment with integrations and combinations of taggers

By Hrafn Loftsson

Published online: 13 December 2006
Abstract

- Use integrations and combinations of taggers to improve the tagging accuracy of Icelandic text.
- Combining five different taggers.
- Adding two linguistically motivated rules to the taggers.
Icelandic tagset

- Icelandic is a morphologically complex language.

- The Icelandic tagset consists of about 660 tags.
  - English
    - 45 tags in Penn TreeBank Corpus.
    - 87 tags in Brown Corpus.
  - 139 tags in Swedish.
Tags

- **Hann fpken (He)**
  - f-pronoun
  - p-personal
  - k-masculine
  - e-singular
  - n-nominative

- **borðaði sfg3ep (ate)**
  - s-verb
  - f-indicative
  - g-active
  - 3rd person
  - e-singular
  - þ-past
Taggers

- Data-Driven Taggers (DDT)
  - fnTBL (TBL)
    - Is transformation-based error-driven learning.
  - MXPOST (MXP)
    - Based on maximum entropy approach.
  - MBT
    - Is memory-based learning tagger.

- Hidden Markov Model (HMM)
  - Trigrams ‘n Tags (TnT)
Linguistic rule-based taggers

- IceTagger (Ice)
  - Ice uses hand-written local linguistic elimination rules, along with a list of idioms.
  - Ice also uses an integrated morphological analyser, IceMorphy, to obtain the possible tags for unknown words.

- Tri
  - Tri is a re-implementation of the TnT tagger.
  - Tri uses the same list of idioms as Ice.
## Results per tagger

### Table 1: The average tagging accuracy of Icelandic text using various taggers

<table>
<thead>
<tr>
<th>Words</th>
<th>Base(^a)</th>
<th>MXP</th>
<th>MBT</th>
<th>TBL</th>
<th>TnT</th>
<th>Tri</th>
<th>Ice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>4.39%</td>
<td>62.29%</td>
<td>59.40%</td>
<td>55.51%</td>
<td>71.68%</td>
<td>71.04%</td>
<td>75.09%</td>
</tr>
<tr>
<td>Known</td>
<td>81.84%</td>
<td>91.00%</td>
<td>91.47%</td>
<td>91.82%</td>
<td>91.82%</td>
<td>91.87%</td>
<td>92.74%</td>
</tr>
<tr>
<td>All</td>
<td>76.27%</td>
<td>89.03%</td>
<td>89.28%</td>
<td>89.33%</td>
<td>90.44%</td>
<td>90.46%</td>
<td>91.54%</td>
</tr>
<tr>
<td>(\Delta^b_{Err})</td>
<td>53.77%</td>
<td>54.83%</td>
<td>55.04%</td>
<td>59.71%</td>
<td>59.80%</td>
<td>64.35%</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) A base tagger which assigns each known word its most frequent tag, and the most frequent noun tag/proper noun tag to lower case/upper case unknown words

\(^b\) Error reduction with regard to the errors made by the base tagger for all words
Integration of taggers

- **TBL***
  - Is an improved version of TBL that lets IceMorphy provide the initial tag for each unknown word.

- **TnT***
  - Is also improved with integration of IceMorphy.
  - Uses IceMorphy to generate a filled lexicon.
  - Smoothing.
Integration of taggers

- Tri with IceMorphy (Tri*)
  - IceMorphy is called from within the Tri tagger to obtain possible tags for unknown words.
  - Also benefits from the lexicon filling mechanism as TnT*.

- Tri with Ice (Ice*)
  - If more than one tag is available. Tri lets Ice select the tag, instead of selecting the most frequent tag for the word.
Results with integration

<table>
<thead>
<tr>
<th>Words</th>
<th>TBL*</th>
<th>TnT*</th>
<th>Tri*</th>
<th>Ice*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown words</td>
<td>66.30%</td>
<td>72.80%</td>
<td>74.46%</td>
<td>75.33%</td>
</tr>
<tr>
<td>Known words</td>
<td>91.90%</td>
<td>92.54%</td>
<td>92.58%</td>
<td>93.00%</td>
</tr>
<tr>
<td>All words</td>
<td>90.15%</td>
<td>91.18%</td>
<td>91.34%</td>
<td>91.80%</td>
</tr>
<tr>
<td>$\Delta_{Err}^a$</td>
<td>7.69%</td>
<td>7.74%</td>
<td>9.13%</td>
<td>3.07%</td>
</tr>
</tbody>
</table>

$^a$ Error reduction with regard to the errors made by the unchanged version of the corresponding tagger for all words.
Combination of taggers

- It has been shown in previous papers that combining taggers will often result in higher tagging accuracy.

- Different taggers tend to produce different (complementary) errors.

- A number of different combination methods exits.
  - Weighted voting.
  - Stacking.
  - Simple voting.
Linguistic motivated rules (LMR)

- Two kinds of LMR where introduced in this paper.
- Both are based on specific strengths of Ice.
- Only used if all taggers dont agree.
## First rule of LMR

### DDT
- **Ég fp1en** (I)
  - f-pronoun
  - p-personal
  - 1st person
  - e-singular
  - n-nominative
- **borðaði sfg3ep** (ate)
  - s-verb
  - f-indicative
  - g-active
  - 3rd person
  - e-singular
  - þ-past

### LMR
- **Ég fp1en** (I)
  - f-pronoun
  - p-personal
  - 1st person
  - e-singular
  - n-nominative
- **borðaði sfg1ep** (ate)
  - s-verb
  - f-indicative
  - g-active
  - 1st person
  - e-singular
  - þ-past
A feature agreement constraint is used.

If all the tags, provided by the individual taggers for the current word, are nominal tags and the current tag provided by Ice agrees in gender, number and case with the preceding or following nominal tag.

Then it chooses the Ice’s tag.
Results with integration and combination

Table 3  Average tagging accuracy using combination of taggers

<table>
<thead>
<tr>
<th>#</th>
<th>Combination (simple voting\textsuperscript{a})</th>
<th>Rule</th>
<th>Accuracy of words</th>
<th>(\Delta_{E1})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Unkn.</td>
<td>Known</td>
</tr>
<tr>
<td>1.</td>
<td>MXP+TBL+TnT</td>
<td>None</td>
<td>71.80%</td>
<td>92.99%</td>
</tr>
<tr>
<td>2.</td>
<td>TBL+TnT+Ice</td>
<td>None</td>
<td>76.76%</td>
<td>93.77%</td>
</tr>
<tr>
<td>3.</td>
<td>MXP+MBT+TBL+TnT+Ice</td>
<td>None</td>
<td>76.74%</td>
<td>93.97%</td>
</tr>
<tr>
<td>4.</td>
<td>TBL*+TnT*+Ice</td>
<td>None</td>
<td>76.55%</td>
<td>94.13%</td>
</tr>
<tr>
<td>5.</td>
<td>MXP+MBT+TBL*+TnT*+Ice</td>
<td>None</td>
<td>78.70%</td>
<td>94.36%</td>
</tr>
<tr>
<td>6.</td>
<td>MXP+MBT+TBL*+TnT*+Ice*</td>
<td>None</td>
<td>78.65%</td>
<td>94.41%</td>
</tr>
<tr>
<td>7.</td>
<td>MXP+MBT+TBL*+TnT*+Ice*</td>
<td>1</td>
<td>78.66%</td>
<td>94.50%</td>
</tr>
<tr>
<td>8.</td>
<td>MXP+MBT+TBL*+TnT*+Ice*</td>
<td>1 &amp; 2</td>
<td>78.68%</td>
<td>94.56%</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Majority voting, in which ties are resolved by selecting the tag of the most accurate tagger in the tie

\textsuperscript{b} Error reduction with regard to the best single tagger in the combination
Conclusion

- Previous work got an average accuracy of 92.94%.

- In this paper the average accuracy was 93.48%.
  - 0.54% increase in accuracy.

- With integration and combinations of taggers, accuracy was improved greatly.
The End