

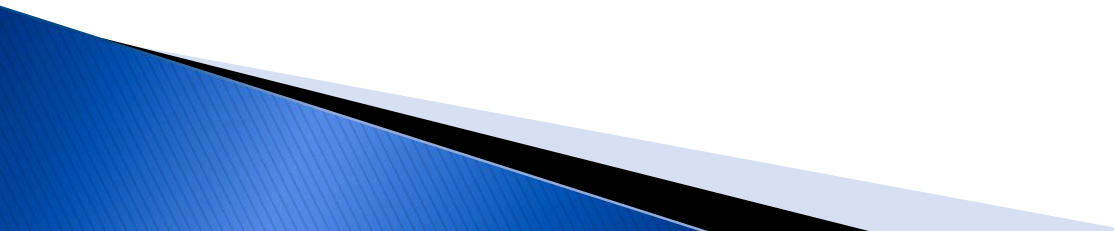
Tagging Icelandic text

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 consist of about 660 tags.
 - English
 - 45 tags in Penn TreeBank Corpus.
 - 87 tags in Brown Corpus.
 - 139 tags in Swedish.
 - 1000-2000 tags in Czech.

Tags

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

- ▶ borðaði **sfg3eþ** (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

Words	Base ^a	MXP	MBT	TBL	TnT	Tri	Ice
Unknown	4.39%	62.29%	59.40%	55.51%	71.68%	71.04%	75.09%
Known	81.84%	91.00%	91.47%	91.82%	91.82%	91.87%	92.74%
All	76.27%	89.03%	89.28%	89.33%	90.44%	90.46%	91.54%
Δ_{Err}^b		53.77%	54.83%	55.04%	59.71%	59.80%	64.35%

^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 2 Average tagging accuracy using integration of taggers

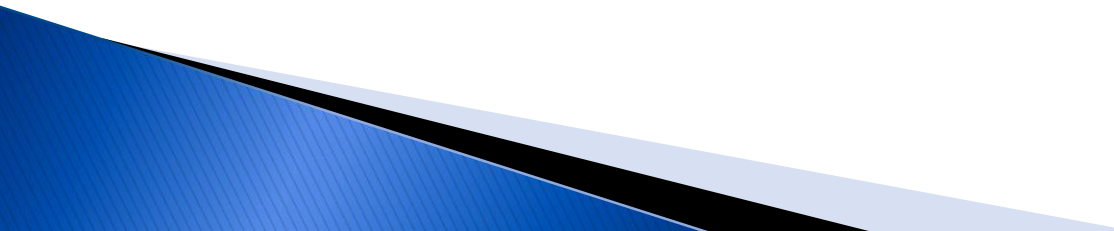
Words	TBL*	TnT*	Tri*	Ice*
Unknown words	66.30%	72.80%	74.46%	75.33%
Known words	91.90%	92.54%	92.58%	93.00%
All words	90.15%	91.18%	91.34%	91.80%
Δ_{Err}^a	7.69%	7.74%	9.13%	3.07%

^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 exists.
 - Weighted voting.
 - Stacking.
 - Simple voting.

Linguistic motivated rules (LMR)

- ▶ Two kinds of LMR were introduced in this paper.
 - ▶ Both are based on specific strengths of Ice.
 - ▶ Only used if all taggers don't agree.
- 

First rule of LMR

DDT

- ▶ **Ég fp₁en** (I)
 - ▶ f-pronoun
 - ▶ p-personal
 - ▶ 1st person
 - ▶ e-singular
 - ▶ n-nominative

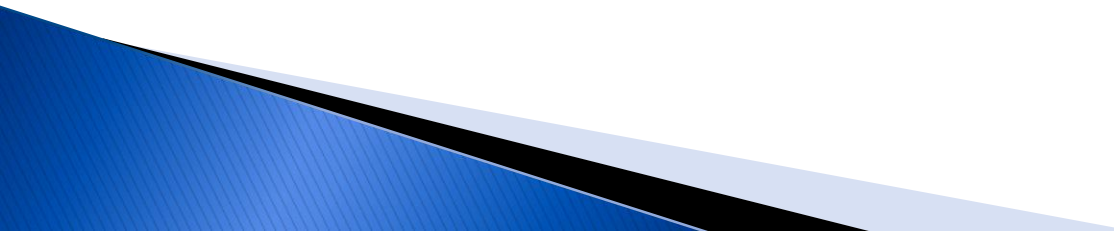
- ▶ **borðaði sfg₃eþ** (ate)
 - s-verb
 - f-indicative
 - g-active
 - 3rd person
 - e-singular
 - þ-past

LMR

- ▶ **Ég fp₁en** (I)
 - ▶ f-pronoun
 - ▶ p-personal
 - ▶ 1st person
 - ▶ e-singular
 - ▶ n-nominative

- ▶ **borðaði sfg₁eþ** (ate)
 - s-verb
 - f-indicative
 - g-active
 - 1st person
 - e-singular
 - þ-past

Second rule of LMR

- ▶ 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

#	Combination (simple voting ^a)	Rule	Accuracy of words			Δ_{E1}^b
			Unkn.	Known	All	
1.	MXP+TBL+TnT	None	71.80%	92.99%	91.54%	12.2%
2.	TBL+TnT+Ice	None	76.76%	93.77%	92.61%	12.7%
3.	MXP+MBT+TBL+TnT+Ice	None	76.74%	93.97%	92.80%	14.9%
4.	TBL*+TnT*+Ice	None	76.55%	94.13%	92.94%	16.6%
5.	MXP+MBT+TBL*+TnT*+Ice	None	78.70%	94.36%	93.29%	20.7%
6.	MXP+MBT+TBL*+TnT*+Ice*	None	78.65%	94.41%	93.34%	18.8%
7.	MXP+MBT+TBL*+TnT*+Ice*	1	78.66%	94.50%	93.43%	19.9%
8.	MXP+MBT+TBL*+TnT*+Ice*	1 & 2	78.68%	94.56%	93.48%	20.5%

^a Majority voting, in which ties are resolved by selecting the tag of the most accurate tagger in the tie

^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

