

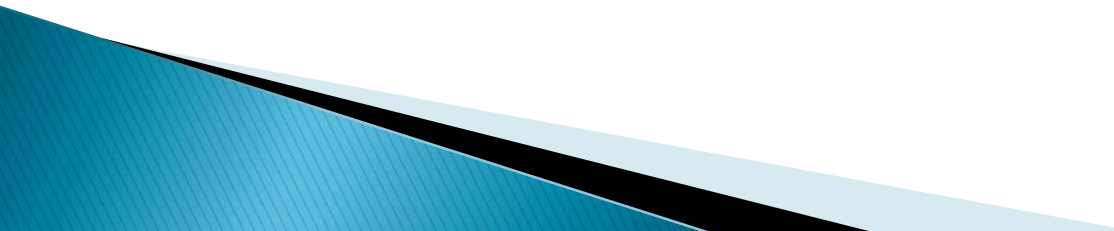
Tagging Icelandic text: A linguistic rule-based approach

Sebastian Hohmann



Introduction

- ▶ 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

Introduction

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

Words/Tagger	fnTBL	MXPOST	TnT
Unknown	54.03%	62.50%	71.60%
Known	91.36%	91.04%	91.74%
All	88.80%	89.08%	90.36%

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

Data-driven tagging methods

Words/Tagger	fnTBL	MXPOST	TnT
Unknown	54.03%	62.50%	71.60%
Known	91.36%	91.04%	91.74%
All	88.80%	89.08%	90.36%

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

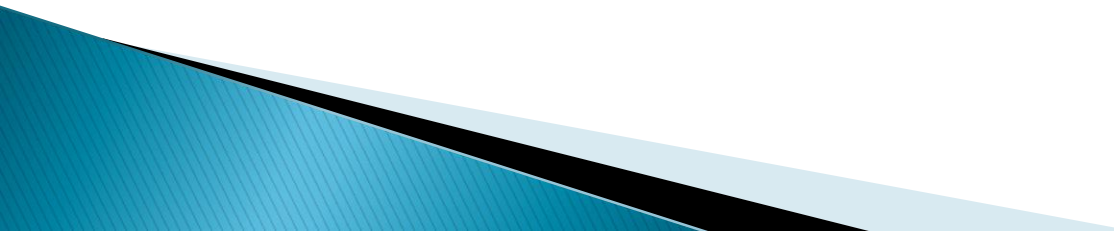
Data-driven tagging methods

Words/Tagger	fnTBL	MXPOST	TnT
Unknown	54.03%	62.50%	71.60%
Known	91.36%	91.04%	91.74%
All	88.80%	89.08%	90.36%

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)

Words/Tagger	MXPOST	fnTBL	TnT	IceTagger
Unknown	62.29%	55.51%	71.68%	75.09%
Known	91.00%	91.82%	91.82%	92.74%
All words	89.03%	89.33%	90.44%	91.54%

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

- ▶ Combination of taggers increases the accuracy

Words/Tagger	fnTBL*	TnT*	IceTagger	Simple voting
Unknown	66.30%	72.75%	75.09%	76.57%
Known	91.90%	92.53%	92.74%	94.15%
All words	90.15%	91.18%	91.54%	92.95%

Table 11. Tagging accuracy using features of IceMorphy.

Thanks for your attention!

A decorative graphic at the bottom of the slide consisting of a dark blue wavy shape on the left, a black horizontal bar in the middle, and a light blue wavy shape on the right.