



Context Sensitive Spelling Correction and Rich Morphology



Anton Karl Ingason

Skúli Bernhard Jóhannsson

Eiríkur Rögnvaldsson

University of Iceland

{antoni,skulib,eirikur}@hi.is

Hrafn Loftsson

Reykjavík University

hrafn@ru.is

Sigrún Helgadóttir

Árni Magnússon Institute for Icelandic Studies

sigruhel@hi.is



The Problem



- Traditional spelling correction
 - Looks for word forms which are not valid
- This is often insufficient
 - *himinn* ‘sky’ (nom.) vs. *himin* (acc.)
 - *farinn* ‘gone’ (masc.) vs. *farin* (fem.)
- Context sensitive spelling correction
 - Looks for word forms which are valid in isolation but not in a given context



Goals of the project



- A case study of Icelandic, a language with rich and often ambiguous morphology
- Ability to correct some types of errors with accuracy useful for word processing
- Provide an integration with a real world word processing system (results should not only be a table which says: 87.2% success)
- Make it easy for others to extend functionality



Methods

- This is a work in progress but we have tried two different approaches
 - A machine learning approach where classifiers are trained on corpora and used to disambiguate confusion sets (*pair of shoes vs. pear of shoes*)
 - Experiments with adding features and normalization to counter data sparseness
 - A rule-based approach with hand written rules that cover some common mistakes



Confusion Sets

- Confusion sets are an important part of how problem is formulated (*pair/pear*)
- They map the problem to a general classification problem
 - Choose between alternatives and assume one of them is correct
- Selected carefully to keep the user happy
 - Since accuracy is limited we need to choose confusion sets where spelling mistakes are probable



The Icelandic Data



- The standard tagset for Icelandic
 - About 700 POS-tags
 - Combinations of case, gender, number, definiteness, tense, mood, ...
- The information is encoded in the morphology using suffixes
 - Various kinds of ambiguities
 - Different contrasts/ambiguities depending on inflection class



Machine Learning Approach



- The context of a confusion word is tagged using IceTagger (Loftsson 2008) and lemmatized using Lemmald (Ingason et al. 2008)
- Features extracted from this information for every confusion word in the text
- Features used to train general purpose classifiers
 - Naïve Bayes, Winnow



Three types of features



- Our current approach is to combine three different types of features (work in progress)
 - Context Words: Word forms occurring at a distance ≤ 5 from the confusion word
 - Context Lemmas: Lemmas (base forms of words) occurring at a distance of ≤ 5 from the confusion word
 - Collocations with words and tags combined (all such possible tri-grams including the confusion word)



An example sentence

(1) *Listamaður frá Reykjavík hefur ákveðið að*
Artist from Reykjavík has decided to
sýna verk sín á listahátíð.
show work his at art festival

‘An artist from Reykjavík has decided to show his work at an art festival.’

- The confusion word here is *sýna* 'show' (vs. *sína* 'his')
- A typical Icelandic spelling mistake if *sína* is used
- No phonetic distinction between 'í' and 'ý' in Modern Icelandic





Collocation Extractor



- All possible tri-grams including confusion word, mixing word forms and tags

(2) *ákveðið að _ verk sín*
ssg cn _ nhfo fehfo
decided to _ work his

(3) *ákveðið að _ ; ssg að _ ; ákveðið cn _ ; ssg cn _ ;
cn _ verk ; að _ nhfo ; _ verk sín ; _ nhfo sín ;
_ verk fehfo ; _ nhfo fehfo ; að _ verk ; cn _ nhfo*



Confusion set	F_{Total}	F_1	F_2
sína ‘his’, sýna ‘show’	951	521	430
list ‘art’, lyst ‘appetite’	177	150	27
kvatt ‘said bye’, hvatt ‘encouraged’	170	100	70
mig ‘I-acc’, mér ‘I-dat’	895	558	337
vil ‘want-1.p.’, vill ‘want-3.p.’	803	480	322
fínn ‘fine-masc’, fín ‘fine-fem’	203	110	93
leiti ‘search,hill’, leyti ‘respect’	606	439	167
himinn ‘sky-nom’, himin ‘sky-acc’	192	101	91
deyi ‘die’, degi ‘day’	462	420	42
líkur ‘similar’, lýkur ‘finishes’	807	414	393
honum ‘he-dat’, hann ‘he-nom’	2829	2068	761

Table 1: Frequencies of confusion words in training corpus: F_{Total} =Total frequency of members of confusion set, F_1 =Frequency of more common member, F_2 =Frequency of less common member



Confusion set	F_T	F_S	F_1	F_2
sína, sýna	871	419	229	223
list, lyst	176	88	86	2
kvatt, hvatt	168	113	33	22
mig, mér	821	547	217	57
vil, vill	720	349	252	119
fínn, fín	169	116	25	28
leiti, leyti	567	319	58	190
himinn, himin	188	138	20	30
deyi, degi	447	101	331	15
líkur, lýkur	801	315	292	194
honum, hann	2674	1518	944	212

Table 2: Number of features extracted for each confusion set: F_T =Total number of features, F_S =Shared features (which belong to both members of the set), F_1 =Features which belong to the former member exclusively, F_2 =Features which belong to the latter member exclusively

Confusion set	Baseline	NaiveBayes	Winnow
sína, sýna	55.0%	96.0%	92.6%
list, lyst	85.0%	87.6%	71.8%
kvatt, hvatt	58.0%	77.6%	64.1%
mig, mér	62.0%	81.2%	77.8%
vil, vill	60.0%	95.3%	94.9%
fínn, fín	54.0%	80.8%	72.9%
leiti, leyti	72.0%	84.5%	83.0%
himinn, himin	53.0%	83.3%	73.4%
deyi, degi	91.0%	93.5%	92.2%
líkur, lýkur	51.0%	92.2%	87.0%
honum, hann	73.0%	87.5%	80.2%
Average	64.9%	87.2%	80.9%

Table 3: Evaluation of the performance of two classification algorithms from the Weka algorithm collection when given the task of disambiguating the members of each confusion set.



Rule-Based Approach



- When one of the two members of a confusion set occurs in a regular and well defined context a simple rule (or set of rules) can give high accuracy
- Some very common mistakes in Icelandic (as well as in other languages) fall into this category
- The two approaches can complement each other when trying to achieve practical results
- An open standardized framework is essential for rule based correction to allow non-programmers to contribute to the development

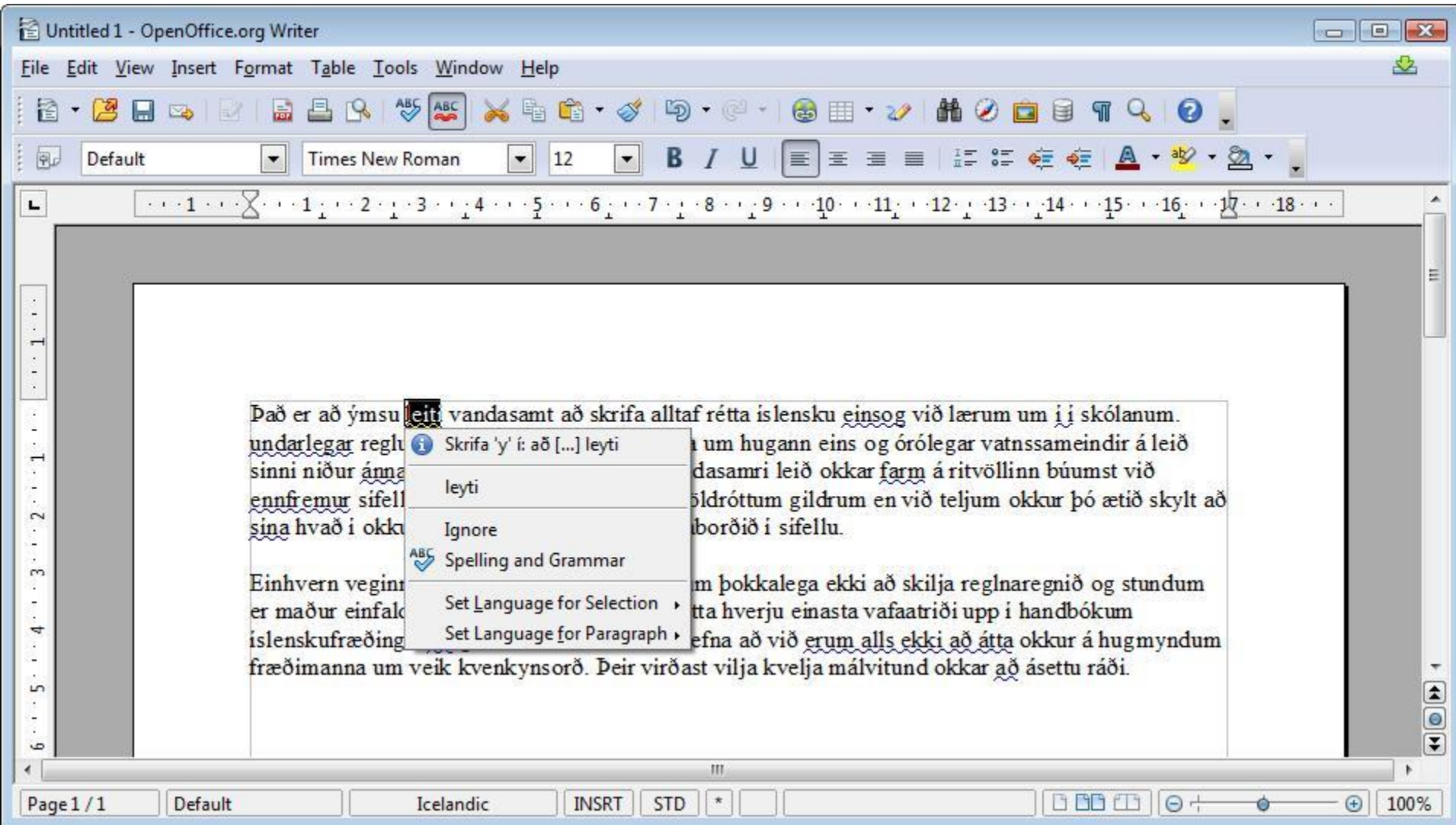


Real World Integration



- Work has started on integrating our solution with LanguageTool (LT), an open source proofreading API and an extension for OpenOffice.org (Naber 2003)
- LT Java rules allow linking with machine learning methods
- LT XML rules allow us to manually write rules for common mistakes
- Initial support for Icelandic context sensitive spelling correction in regular word processing is ready
 - Thanks to LT team for assistance
- Since this is an open framework other projects can contribute rules (Java or XML) to extend functionality

Screenshot from OO





References

Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In *GoTAL '08: Proceedings of the 6th international conference on Advances in Natural Language Processing*, pages 205-216, Berlin, Heidelberg. Springer-Verlag.

Hrafn Loftsson. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1):47-72.

Daniel Naber. 2003. A Rule-Based Style and Grammar Checker. Diploma thesis, University of Bielefeld.