Context-Sensitive Spelling Correction and Rich Morphology

Anton K. Ingason Skúli B. Jóhannsson
Eiríkur Rögnvaldsson
University of Iceland
Reykjavík, Iceland
{antoni,skulib,eirikur}@hi.is

Hrafn Loftsson
Reykjavik University
Reykjavík, Iceland
hrafn@ru.is

Sigrún Helgadóttir
Árni Magnússon Institute
for Icelandic Studies
Reykjavík, Iceland
sigruhel@hi.is

Abstract

Context-sensitive spelling correction is the task of correcting spelling errors which result in valid words. We present work in progress where we adapt established methods from English to a morphologically rich language and conclude that the rich morphology negatively affects performance. However, our system is still good enough to be useful in regular word processing.

1 Introduction

Context-sensitive spelling correction is the task of correcting spelling errors which result in valid words. For example, in the sentence *I want a peace of cake*, *peace* is a valid word in isolation but an error in this context (should be *piece*). Most spelling correction systems check one word at a time and do not correct such errors. Context-sensitive errors account for 25% to 50% of observed errors (in English data) (Kukich, 1992) and thus it is important to address this problem. A variety of methods have given good results for English but little attention has been paid to how well such methods perform on languages with very rich morphology. No earlier attempts at this task exist for the language of our study, Icelandic.

In this paper, we aim to shed light on the issue of morphological richness and ambiguity in context-sensitive spelling correction by presenting a system (a work in progress) for Icelandic, whose morphology is both rich and highly ambiguous. We adapt methods used in previous work on English to be used on Icelandic and evaluate the performance of the system.

The paper is organized as follows. In Sect. 2 we review some of the previous work carried out on the subject. Sect. 3 describes our method and its evaluation which estimates the accuracy of the system to be 87.2%. We conclude in Sect. 4.

2 Context-Sensitive Spelling Correction

In the literature, the problem of context-sensitive spelling correction is commonly formulated as a disambiguation task (Roth, 1998) where the ambiguity among words is modeled by confusion sets. A confusion set $C = \{W_1, ..., W_n\}$ means that each word W_i in the set is ambiguous with each other word in the set. Thus, if $C = \{piece, peace\}$ and either *piece* or *peace* is encountered in a text, the task is to decide which one was intended.

Such errors can be categorized in various ways (Kukich, 1992). One distinction is whether the contrast between the contexts of the members of the confusion set is semantic, grammatical or both. (1) Semantic contrast (piece/peace): Different words with different meaning but they belong to the same distributional class (in this case they are both nouns) and behave identically with respect to the syntactic context. (2) Grammatical contrast (he/him): Different forms of one word which behave differently with respect to the syntactic context. (3) Semantic and grammatical contrast (cite/site): Different words with different meanings and different syntactic properties (verb vs. noun). We evaluate the performance of a data-driven approach against all the above types of errors.

2.1 Related Work

Our focus is on data-driven systems which are able to handle disambiguation between members of confusion sets without relying entirely on syntactic structure. We can divide these into two categories based on whether they can be trained using only a corpus or whether they require external semantic databases like WordNet (Fellbaum, 1998).

Solutions which do not rely on external semantic databases extract semantic and grammatical features from the contexts of members of confusion sets using corpora and take advantage

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 231–234

of general purpose classifiers. Successful methods used for this purpose include Bayesian classifiers (Golding, 1995) and Winnow-based classifiers (Golding and Roth, 1999).

Solutions which do rely on external semantic databases take advantage of the fact that semantic relations within the lexicon of a given language provide useful evidence for semantic disambiguation. If we have a confusion set $C = \{a, b\}$ and many words in the context are semantically related to a but few are semantically related to b, then this is evidence that the intended form is a. A few such methods are compared in Budanitsky and Hirst (2006). It is a feasible option to take advantage of knowledge-rich resources in context-sensitive spelling correction but the problem is the cost of developing such resources. For example, there exists no WordNet-like resource for Icelandic suitable for this purpose.

2.2 Morphological Richness

We take Icelandic as our test-case for a morphologically rich language. Icelandic has rich inflection and morphosyntactic categories are encoded using affixes which are often quite ambiguous. This is reflected in the tagset normally used when PoS-tagging the language which has about 700 different PoS-tags (originally developed by Pind et al. (1991)). This leads to data sparseness when collecting evidence from a corpus. The sparseness of the features used for disambiguation can furthermore make it more difficult to effectively prune the number of features without losing important evidence, thus making scalability a more serious problem. To counter the data sparseness problem it is possible to normalize the data in the corpus. For normalization we used the Lemmald lemmatizer (Ingason et al., 2008). Lemmald is a data-driven system but it still employs some linguistic knowledge about Icelandic grammar. Note that the problems caused by the morphological richness also apply to the lemmatization itself. The lemmatizer achieves an accuracy of 98.54%.

The rich morphology and the corresponding large tagset also affect the accuracy of the PoS-tagging. Various taggers and combinations of taggers have been tried out for Icelandic PoS-tagging (Helgadóttir, 2005; Loftsson, 2006; Loftsson, 2008; Dredze and Wallenberg, 2008). The highest reported tagging accuracy for a data-driven solution is 92.06% (Dredze and Wallenberg, 2008) but the rule-based IceTagger achieves 91.54% accuracy (Loftsson, 2008) and runs considerably faster (2.700 tokens/sec vs. 179 tokens/sec). We used IceTagger in our experiments. Note that, while performance is low compared to the over 97% reported performance for state-of-the-art English taggers (Shen et al., 2007; Giménez and Màrquez, 2004), the tagging of only the word class actually has a similar accuracy as the state-of-the-art taggers for English – most mistakes are made in tagging other features such as case.

3 Machine Learning Approach

3.1 Feature Extraction

The choice of features is the result of experiments with different combinations of features intended to bring out important evidence of context. More work is needed to evaluate the actual contribution of specific types of features. We extracted three types of features from the corpus in our experiments. (1) Context Words: Word forms occuring at a distance of ≤ 5 from the confusion word; (2) Context Lemmas: Lemmas (base forms of words) occuring at a distance of ≤ 5 from the confusion word; (3) Collocations with words and tags combined (all such possible tri-grams including the confusion word). The Context Word and Context Lemma feature extractors simply collect all words and base forms of words which occur within a window of 5 tokens on either side of the confusion word. The Collocation feature extraction combines word forms and PoS-tags and every such possible tri-gram is a potential feature.

(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.'

Sentence (1) contains the word *sýna* 'show' which is sometimes confused with *sína* 'his/her/its' because in Modern Icelandic there is no phonetic difference between 'í' and 'ý' (and spelling mistakes are common in many words in which those letters occur). The window we use for our Collocation feature extractor is shown in (2) where the second line displays the PoS-tags for the corresponding tokens. The confusion word is represented as '_'.

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

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.

We then generate all possible tri-gram combinations of word forms and tags from the window. Those are shown in (3).

(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

For evaluation purposes we extracted features for 11 confusion sets from a selected part of the SÁ corpus[1] according to the methods described above. Table 1 shows the 11 confusion sets and their frequency in the corpus.

To reduce the number of features we remove all features which occur less than 4 times in the training data. Table 2 shows the number of features extracted for each confusion set, first the total number of features, then the features that occur in the context of both members of the confusion set, then the features which belong only to the former member and finally the features which belong only to the latter member.

As Table 2 shows, the amount of evidence varies quite a lot between confusion sets. For some of the confusion sets there are many features which belong exclusively to one of the two members but for others, such as *lyst* 'appetite', there is a serious data sparseness problem.

3.2 Evaluation

The features extracted according to the description in the previous section were fed to data-driven classification algorithms implemented in the Weka algorithm collection (Witten and Frank, 2005). We tried two methods that have performed well for English: Naive Bayes and Winnow. We also compared the result with a baseline classifier which always chooses the more common member of the confusion set. All tests were performed using a 10-fold cross validation on all sentences which contained the confusion set in question. The results of these tests are displayed in Table 3.

The results show lower accuracy than what has been reported for English (Golding, 1995; Golding and Roth, 1999) but the performance is still close to 90% which is probably enough for a real world application to be useful. It is unexpected that the Naive Bayes method outperforms Winnow which has been more successful for English (cf. the references above) and we do not have an ex-

[1] Textasafn Orðabókar Háskólans. [SÁ Corpus.] www.lexis.hi.is/corpus/leit.pl

planation for this.

It is not unexpected that the results are worse for a morphologically rich language like Icelandic than for morphologically simple English. Data sparseness and errors in PoS-tagging and normalization are the most likely reasons for this. Even if we include all the features described in the previous section, context-sensitive spell checking for Icelandic lags behind comparable systems for English.

4 Conclusion and Future Work

As expected, morphological complexity negatively affects performance. However, our system is still a viable option for everyday word processing. We have begun integrating our system into the LanguageTool platform (Naber, 2003) which provides easy integration into OpenOffice.org. Our system must still be viewed as work in progress and some issues require further study. We hope to gain a better understanding of why a Winnow-based classification method does not perform well for Icelandic. We also hope to construct semantic resources for Icelandic to complement the method we use for semantic disambiguation.

Acknowledgments

The work in this paper was supported by the Icelandic Research Fund, grant 070620022.

References

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32:13–47.

Mark Dredze and Joel Wallenberg. 2008. Further Results and Analysis of Icelandic Part of Speech Tagging. Technical report, University of Pennsylvania, Department of Computer and Information Science.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the* 4^{th} *International Conference on Language Resources and Evaluation*, pages 43–46.

Andrew R. Golding and Dan Roth. 1999. A Winnow-Based Approach to Context-Sensitive Spelling Correction. In J. Mooney and Claire Cardie, editors, *Machine Learning*, pages 107–130.

Andrew R. Golding. 1995. A Bayesian hybrid method for context-sensitive spelling correction. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 39–53.

Sigrún Helgadóttir. 2005. Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In H. Holmboe, editor, *Nordisk Sprogteknologi 2004*. Museum Tusculanums Forlag, Copenhagen.

Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and 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.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439.

Hrafn Loftsson. 2006. Tagging Icelandic text: An experiment with integrations and combinations of taggers. *Language Resources and Evaluation*, 40(2):175–181.

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.

Jörgen Pind, Friðrik Magnússon, and Stefán Briem. 1991. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavik, Iceland.

Dan Roth. 1998. Learning to Resolve Natural Language Ambiguities: A Unified Approach. In *AAAI '98/IAAI '98: Proceedings of the* 15^{th} *national/*10^{th} *conference on Artificial Intelligence/Innovative applications of Artificial Intelligence*, pages 806–813, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Libin Shen, Giorgio Satta, and Aravind Joshi. 2007. Guided Learning for Bidirectional Sequence Classification. In *Proceedings of the* 45^{th} *Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.