Disambiguating Confusion Sets in a Language with Rich Morphology

Steinunn Rut Friðriksdóttir[a] and Anton Karl Ingason[b]
Faculty of Icelandic and Comparative Cultural Studies, University of Iceland, Sæmundargata 2, 102 Reykjavík, Iceland
{srf2, antoni}@hi.is

Keywords: Confusion Sets, Homophones, Context Dependency, Rich Morphology, Disambiguation, Icelandic.

Abstract: The processing of strings which are semantically distinct but can be easily confused with each other, often on account of being pronounced identically, is a prime example of context dependency in Natural Language Processing. This problem arises when a system needs to distinguish whether a *bank* is a 'river bank' or a 'financial institution' and it also challenges systems for context-sensitive spelling and grammar correction because pairs like *their/there* and *I/me* are one common source of issues that such systems must address. In practice, this type of context-dependency can be especially prominent in languages with rich morphology where large paradigms of inflected word forms lead to a proliferation of such confusion sets. In this paper, we present our novel confusion set corpus for Icelandic as well as our findings from an experiment that uses well-known classification algorithms to disambiguate confusion sets that appear in our corpus.

1 INTRODUCTION

Spelling mistakes in high resource languages such as English can be corrected by a wide variety of available spell checkers and proofreading software. Traditionally, this task involves looking up an individual word and making sure it exists in the vocabulary. If not, an error message is prompted. While this is very beneficial for correcting typographical errors, the problem remains that this does not detect mistakes that involve confusing valid words in a language. By taking context into consideration, the probability of a word given its context can be evaluated. Context sensitive spell checkers use confusion sets which specify a list of confusable words, e.g., *then/than*, each occurrence of which is represented as a vector of features obtained from the target word's surrounding context. A classifier is then trained on sentences containing the confusion set, generating both positive and negative examples of each context. Once trained, the classifier predicts the most likely candidate of the confusion set given an unseen sentence containing the target words.

In morphologically rich languages such as Icelandic, whose part of speech tags (comprising of both word classes and morphological information) are several hundred, the need to disambiguate confusable word pairs becomes particularly apparent. As there is often minimal orthographic difference between grammatical genders or cases for example, the possibility of confusion is high. The aim of this paper is to experiment with machine learning approaches to context sensitive spelling correction for the highly ambiguous morphology of the Icelandic language. The morphological richness of the language has also been noted in the literature in the context of other tasks in Natural Language Processing such as lemmatization (Ingason et al., 2008). It should be noted that these type of systems could also prove beneficial for grammar correction. We briefly discuss this in Sect. 3. In addition, while our research focuses solely on Icelandic, we hope that this approach could prove useful for other low resource languages.

The paper is organized as follows: The next section describes the task of context-sensitive spelling correction and the case of a morphologically rich language such as Icelandic. In Sect. 3, we present the Icelandic Confusion Set Corpus (ICoSC) and describe its contents. In Sect. 4, we present our experiment of disambiguating Icelandic by feeding the corpus to a handmade feature extractor to the machine learning algorithm. The results of the experiment are presented in Sect. 5. We conclude in Sect. 6.

2 BACKGROUND

For high resource languages such as English, there is a wide variety of spell checkers and proofreading

[a] https://orcid.org/0000-0002-3675-7975
[b] https://orcid.org/0000-0002-2069-5204

software available for commercial use. The idea behind the simplest ones is to look up an isolated word in a predefined dictionary, prompting an error message if no such word exists. The database can even be expanded by adding non-existent word to the personal dictionary of the user. The predominant type of spelling mistakes that go undetected in this type of software are therefore the kind that result in a real but unintended word, often distinguished only semantically from the intended word, such as when *then* is written in place of *than*.

2.1 Confusion Sets

Another approach is needed to tackle this type of mistakes. Rather than looking at the word in isolation, it is necessary to look at the context to determine which word is most likely to have been intended given the morphological and semantic aspects of the surrounding words (Golding and Roth, 1999). In morphologically rich languages such as Icelandic, whose combined word class and morphological tags are several hundred, the need to disambiguate confusable word pairs becomes particularly apparent. As there is often minimal difference in writing between grammatical genders or cases for example, the possibility of confusion is high, not least for dyslexic people or immigrants learning the language.

To solve this task, a confusion set is defined which specifies words that commonly get confused, e.g. *then, than* or *your, you're*. Each of these words is then represented as a feature vector derived from a small context window around the target word (Rozovskaya and Roth, 2010). In our case, the considered context is obtained from the two words that immediately precede the target word as well as the (single) word that immediately follows the target word. A binary classifier is trained on multiple sentence examples containing each word of the confusion set, and then made to predict the most likely candidate in the confusion set when faced with previously unseen sentence examples.

2.2 Related Work

The problem of correcting spelling errors resulting in valid words has been addressed for high resource languages such as English, which is morphologically rather simple. In recent years, NLP specialists have been working on solving this problem for low resource languages as well. In their 2011 paper, Petros et al. present an automatic spelling correction for Modern Greek homophones using several different algorithms such as Naive Bayes and Random Forest (Spiridonidou, 2014). In 2015, Rokaya combined the use of statistical methods and confusion sets for the purpose of disambiguating semantic errors in Arabic, (Rokaya, 2015) and in the same year, Samani M.H., Rahimi Z. and Rahimi S. address real-word spelling mistakes in Persian using n-gram based context retrieval for confusion sets (Samani et al., 2015). All these researches show promising results. In 2009, Ingason et al. conducted a small-scale experiment addressing semantic disambiguation for Icelandic, where features extracted from the context of confusion sets were fed to the Naïve Bayes and Winnow algorithms (Ingason et al., 2009). This experiment showed promising results and we hope to further expand this research in our experiment, using a much larger database than previously available.

2.3 Usefulness for Non-native Speakers and Dyslexic People

In her pilot study, conducted in 2017, Arnórsdóttir explored which mistakes non-native speakers are most likely to make when speaking Icelandic (Arnórsdóttir, 2017). The participants were either Francophones or native German speakers. According to her results, Francophone speakers struggle more with grammatical gender and case agreement than German speakers, which may indicate that language transfer is easier between Icelandic and other Germanic languages than between Icelandic and Roman languages. In any case, these types of mistakes, where grammatical genders or cases are confused, are more likely to be made by non-native speakers learning Icelandic as a second language. With the constantly growing number of immigrants in Iceland, a context-sensitive spell checker could prove very useful when encouraging L2-learners to communicate in Icelandic. This could also potentially benefit dyslexic people, who typically struggle with spelling (Morris et al., 2002), as inadvertently jumbling letters can result in unintended, valid words (e.g. confusing *dog* with *god* or *box* with *pox*).

3 CONFUSION SET CORPUS

The first part of our experiment was on collecting the necessary data, a task only made possible through the release of the *Icelandic Gigaword Corpus* (Steingrímsson et al., 2018), hereinafter referred to as IGC, which was compiled and tagged during the years 2015 to 2017 and consists of about 1300 million running words of text, tagged using *IceStagger* (Loftsson and Östling, 2013). The IGC is categorized into

six types of text, taken from various available media, the text collection of the Árni Magnússon Institute for Icelandic studies and official documents. In the current project, we cross-referenced the IGC with the Database of Icelandic Morphology (Bjarnadóttir et al., 2019). These texts have now become the foundation for the compilation of the Icelandic Confusion Set Corpus (ICoSC), which was constructed during the course of three months during the year 2019. The final result will be made available under a CC-BY licence for anyone wanting to run their own experiment or replicate ours.

The ICoSC consists of three categories of confusion sets, selected for their linguistic properties as homophones, separated orthographically by a single letter. The categories are:

- 197 pairs containing y/i *(leyti 'extent' / leiti 'search')*: In modern Icelandic, there is no phonetic distinction between these sounds (both of which are pronounced as [ɪ]) and thus their distinction is purely historical. The use of *y* refers to a vowel mutation from another, related word, some of which are derived from Danish. Confusing words that differ only by these letters is therefore very common when writing Icelandic.
- 150 pairs containing ý/í *(sýn 'vision' / sín 'theirs (possessive reflexive)')*: The same goes for these sounds, which are both pronounced as [i]. The original rounding of *y* and *ý* started merging with the unrounded counterparts of these sounds in the 14th century and the sounds in question have remained merged since the 17th century (Gunnlaugsson, 1994).
- 1203 pairs containing nn/n *(forvitinn 'curious (masc.)' / forvitin 'curious (fem.)'*: The alveolar nasal [n] is not elongated in pronunciation and therefore there is no real distinction between these sounds in pronunciation (although the preceding vowel to a double n is often elongated). The distinction between them is often grammatical and refers to whether the word has a feminine or masculine grammatical gender. However, the rules on when to write each vary and have many exceptions, many of which are taught as something to remember by heart. It is therefore common for both native and non-native speakers to make spelling and/or grammar mistakes in these type of words.
- 8 pairs commonly confused by Icelandic speakers: These confusion sets could prove useful in grammar correction as their difference is in their morphological information rather than their orthography. These include for example *mig/mér* (*me* (accusative) / *me* (dative)) which commonly get confused when followed by experiencer-subject verbs (Jónsson and Eythórsson, 2005; Ingason, 2010; Thráinsson, 2013; Nowenstein, 2017).

It is worth noting that although various spelling and grammar mistakes are well suited for a confusion set approach, some mistakes, for examples patterns that are very general and abstract require different methods. For example, use of the so-called New Passive in Icelandic (Ingason et al., 2013) is usually corrected to a traditional passive in proofreading but as this pattern applies to the passives of a wide range of verbs and arguments and the paraphrase involves changing both word forms and word order, other methods are better suited for this purpose.

Included in the ICoSC are spreadsheets containing all collected confusion sets of each category and their frequencies. The spreadsheets are organized so that for each set, the total frequency of each candidate is calculated along with the frequency of each possible PoS tag for that candidate. The seventh and eight column of the tables contain binary values referring to whether the confusion set is grammatically disjoint or grammatically identical. The final column shows the frequency of the less frequent candidate of the set which can be used to determine which sets are viable in an experiment. Also included are text files containing the list of words from each category (as well as three categories not used in this experiment due to data sparsity) and text files containing all sentence examples from the IGC including the words for each category. As the n/nn examples are by far the most frequent confusion sets, the corpus also includes a word list and sentence examples for the 55 most frequent sets. All files have UTF-8 encoding.

4 DISAMBIGUATION METHOD

In our experiment, we mainly focused on comparing three distinct categories of confusion sets.

- Grammatically disjoint word pairs *(they/them)*: The PoS tags for each word never overlap with the other. This is very common for Icelandic;
- Grammatically identical word pairs *(principle/principal)*: Both words within the pair belong to the same distributional class and differ only by semantics. Somewhat surprisingly, this turned out to be the smallest category in our research where only six word pairs had high enough frequency to be of value;

- Word pairs that fall under neither aforementioned category and thus the words within the pair can differ both in their semantic and syntactic properties, *(lose/loose)*.

The Icelandic language has a very rich morphology. This is reflected in the 565 tags used in the IGC, which contain information both on the word class and the morphological aspects of each word. Examples of this can be seen in Table 1. The release of the IGC is revolutionary to the development of NLP tools in Icelandic and has made it possible to conduct research on a much larger scale. Nonetheless, this great number of tags leads to data sparseness where some tags appear significantly less often than others. Careful grammatical feature selection is therefore very important and should be considered beforehand for each task at hand. As our results show, it is difficult to generalize feature selection for different types of confusion sets and accuracy could be significantly improved by adding more features.

Table 1: Examples of confusion sets.

	Word form	Possible tags
WF1	sýna 'show/vision'	6 (verb, noun)
WF2	sína 'his, hers, etc.'	3 (pron.)
WF1	einn 'one (masc.)'	7 (num., pron.)
WF2	ein 'one (fem.)'	14 (num., pron.)
WF1	breytt 'changed'	7 (verb, adj.)
WF2	breitt 'wide/cover'	4 (adj., verb)

In our experiment, we use the decision tree algorithm provided by Scikit learn (Pedregosa et al., 2011) to create a binary classifier that can determine which of the candidates from our two-word confusion sets is more likely to be the intended word. A key property of a decision tree is that it is very easily human-interpretable (Bishop, 2006), which in theory should prove useful for a morphologically complex language such as Icelandic as it should make it easier to keep the feature selection scalable (we will explore using different algorithms in future research). All tests were done using 10-fold cross validation on all the sentences in the data which contained the confusion set being observed. The splitting of the trees can be observed by using Graphviz' connection to Scikit learn, see Figure 1.

The feature selection for this experiment consists of only 12 binary features, handpicked by the authors, and the context words considered are the two words immediately preceding the target word and the (single) word immediately following the target word. The features are as follows (true/false): Left context word is nominal (words with grammatical case, such as nouns and pronouns); Right context word is nominal; Left context word is finite (a verb that inflects for person agreement); Right context word finite; Left context word is nominative; Right context word is nominative; Left context word is oblique (has some grammatical case other than nominative); Right context word is oblique, Left context word is a particle; Right context word is a particle; The context word two words to the left of the target word is feminine; The context word two words to the left of the target word is masculine. The importance of each feature for a confusion set can be examined using *feature importance* from Scikit learn, see Figure 2. These features were chosen due to their expected generalizability but could be significantly improved by looking at the grammatical properties of each confusion set category separately. Future research could also include the significance of context lemmas and n-grams including the target word, as explored by Ingason et al. (2009). Although not applied here, methods that employ semantic relatedness (Budanitsky and Hirst, 2006) of words in the context can also be invoked for this kind of a task.

5 EVALUATION

The decision tree algorithm was run on all viable confusion sets in the ICoSC. Due to overall data sparseness and uneven word count between categories, we only considered confusion sets where the less common candidate occurred at least 25 times in the data, except in the case of grammatically identical word pairs which included confusion sets where the less common word occurred at least 10 times. Due to the high number of nn/n-pairs, their limit was raised to at least 50 occurrences of the less frequent word. Other categories considered contained too little data to be of use. We evaluated the accuracy, precision, recall and f-score of the algorithm for each of our sets.

Table 2: Example sets evaluation.

Set	Accuracy	Precision	Recall	F-score
neytt/neitt 'consumed'/'anything'	0.99	0.99	0.99	0.99
ynni/inni 'work'/'inside'	0.99	0.99	0.99	0.99
einna/eina 'about'/'one'	0.98	0.98	0.99	0.99
munnur/munur 'mouth'/'difference'	0.98	0.99	0.99	0.99
mynni/minni 'mouth of a river'/'mine'	0.98	0.98	0.99	0.99
rýkur/ríkur 'steams'/'rich'	0.95	0.94	0.94	0.93
sýna/sína 'show'/'theirs'	0.92	0.94	0.94	0.94

Table 2 shows examples of high-scoring confusion

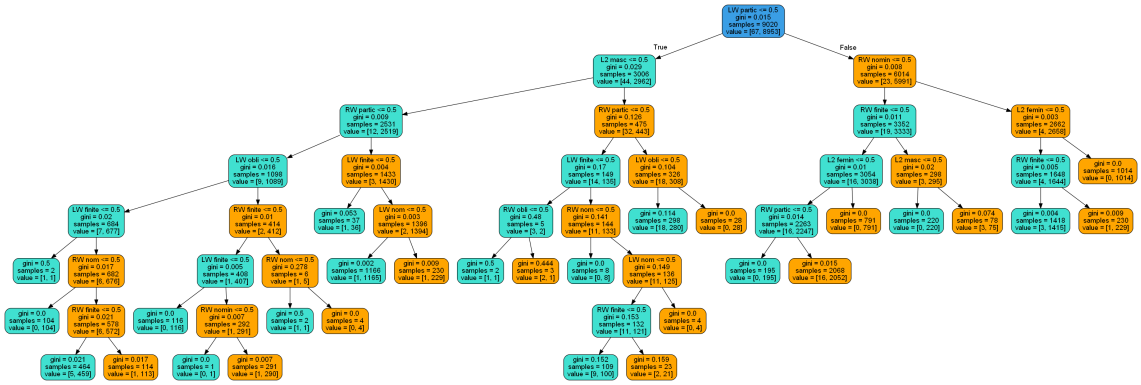

Figure 1: Decision tree for *neytt* 'consumed'/*neitt* 'anything'.

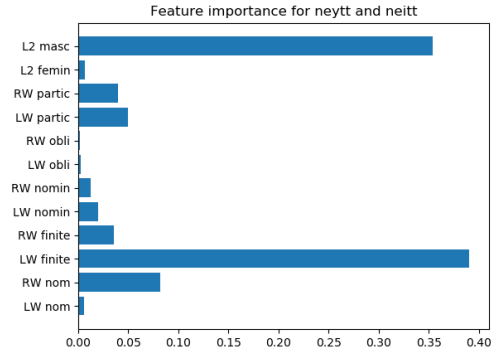

Figure 2: Feature importance for *neytt 'consumed' / neitt 'anything'*.

sets and indeed, 20 out of 91 pairs scored over 90% in all measures. Table 3 shows the average scores for each of the categories. The algorithm performs best on grammatically disjoint confusion sets, where there is no overlap between the candidates' PoS tags, which suggests that the contextual features of individual candidates is less likely to overlap and that results could be perfected by examining their linguistic properties. On the other hand, the poorest performance is on the grammatically identical sets, where both candidates have exactly the same PoS tags. This may indicate that more work is needed to distinguish between candidates separated only by semantics. The reader should keep in mind however that the number of sets in the grammatically identical category is much smaller than of the other two categories and may not be properly representative.

6 CONCLUSION

Throughout the years, the lack of data has been the biggest Achilles' heel for the development of Icelandic NLP tools. Fortunately, thanks to The Icelandic language technology programme 2018-2022 (Nikulásdóttir et al., 2017) and the release of the IGC, there are a number of reasons to be optimistic about the future. It's our hope that the ICoSC will aid in the creation of Icelandic language technology. The decision tree experiment should be considered as a work in progress and by no means as a finalized tool. Results could undoubtedly be improved by a more careful choice of linguistic features and by taking into consideration a wider context. However, it is clear from the sheer amount of confusable words within the data that a context sensitive spell checker could prove tremendously useful for Icelandic. With increased generalization comes increased usability and we hope that our research can be expanded to other morphologically rich, low resource languages. We aspire to better our results in future research.

Table 3: Average scores for categories.

Type	Accuracy	Precision	Recall	F-score
Disjoint	0.78	0.77	0.76	0.75
Identical	0.73	0.68	0.66	0.64
Overlap	0.79	0.75	0.68	0.68
y/i	0.86	0.76	0.74	0.73
ý/í	0.79	0.82	0.79	0.78
nn/n	0.75	0.74	0.73	0.70
Various	0.75	0.71	0.66	0.66

REFERENCES

Arnórsdóttir, A. L. (2017). *Je parle très bien l'islandais, surtout à l'écrit: recherche sur les transferts du français vers l'islandais chez les apprenants francophones*. Unpublished BA-thesis, University of Iceland.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. (2019). DIM: The Database of Icelandic Mor-

phology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, NODALIDA 2019, Turku, Finland.

Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Golding, A. R. and Roth, D. (1999). A winnow-based approach to context-sensitive spelling correction. *Machine learning*, 34(1-3):107–130.

Gunnlaugsson, G. M. (1994). *Um afkringingu á/y, ỳ, ey/í íslensku*. Málvísindastofnun Háskóla Íslands.

Ingason, A. K. (2010). Productivity of non-default case. *Working papers in Scandinavian syntax*, 85:65–117.

Ingason, A. K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. (2008). A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In *Proceedings of Sixth International Conference on Natural Language Processing*, GoTAL 2008, Gothenburg, Sweden.

Ingason, A. K., Jóhannsson, S. B., Rögnvaldsson, E., Loftsson, H., and Helgadóttir, S. (2009). Context-Sensitive Spelling Correction and Rich Morphology. In *Proceedings of the 17th Nordic Conference of Computational Linguistics*, NODALIDA 2009, Odense, Denmark.

Ingason, A. K., Legate, J. A., and Yang, C. (2013). The evolutionary trajectory of the Icelandic New Passive. *University of Pennsylvania Working Papers in Linguistics*, 19(2):11.

Jónsson, J. G. and Eythórsson, T. (2005). Variation in subject case marking in Insular Scandinavian. *Nordic Journal of Linguistics*, 28.2:223–245.

Loftsson, H. and Östling, R. (2013). Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 105–119, Oslo, Norway. Linköping University Electronic Press, Sweden.

Morris, B., Munoz, L., and Neering, P. (2002). Overcoming dyslexia. *Fortune-European edition-*, 145(10):46–51.

Nikulásdóttir, A. B., Guðnason, J., and Steingrímsson, S. (2017). *Language Technology for Icelandic. Project Plan*. Icelandic Ministry of Science, Culture and Education.

Nowenstein, I. (2017). Determining the nature of intra-speaker subject case variation. In Thráinsson, Höskuldur, C. H. H. P. P. and Hansen, Z. S., editors, *Syntactic Variation in Insular Scandinavian*, pages 91–112. John Benjamins.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rokaya, M. (2015). Arabic semantic spell checking based on power links. *International Information Institute (Tokyo). Information*, 18(11):4749–4770.

Rozovskaya, A. and Roth, D. (2010). Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 961–970, Cambridge, MA. Association for Computational Linguistics.

Samani, M. H., Rahimi, Z., and Rahimi, S. (2015). A content-based method for persian real-word spell checking. In *2015 7th Conference on Information and Knowledge Technology (IKT)*, pages 1–5.

Spiridonidou, A. (2014). Knowledge-poor context-sensitive spelling correction for modern greek.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.

Thráinsson, H. (2013). Ideal speakers and other speakers. the case of dative and other cases. In Fenández, B. and Etxepare, R., editors, *Variation in Datives – A Micro-Comparative Perspective*, pages 161–188. Oxford University Press.