Analysing Inconsistencies and Errors in PoS Tagging in two Icelandic Gold Standards

Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
steinst@hi.is

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

Eiríkur Rögnvaldsson
University of Iceland
Reykjavík, Iceland
eirikur@hi.is

Abstract

This paper describes work in progress. We experiment with training a state-of-the-art tagger, *Stagger*, on a new gold standard, *MIM-GOLD*, for the PoS tagging of Icelandic. We compare the results to results obtained using a previous gold standard, *IFD*. Using *MIM-GOLD*, tagging accuracy is considerably lower, 92.76% compared to 93.67% accuracy for *IFD*. We analyze and classify the errors made by the tagger in order to explain this difference. We find that inconsistencies and incorrect tags in *MIM-GOLD* may account for this difference.

1 Introduction

For some years a new gold standard, *MIM-GOLD*, for training PoS taggers has been under development (Loftsson et al., 2010; Helgadóttir et al., 2012). This corpus contains approximately one million tokens of text from various sources from the period 2000–2009.

State-of-the-art PoS tagging accuracy for Icelandic, 92.82%, was achieved by Loftsson and Östling (2013), using the Averaged Perceptron Tagger *Stagger* without an external lexicon. All PoS taggers tested so far for Icelandic have been developed or trained and tested on the Icelandic Frequency Dictionary (*IFD*) (Pind et al., 1991).

In this paper we describe the training and testing of *Stagger* on *MIM-GOLD*. Results are compared to the results for training and tagging *IFD* reported by Loftsson and Östling (2013). Tagging errors made by *Stagger*, when tagging the two gold standards, are examined and classified to explain the difference in tagging accuracy.

In Section 2 we describe the two corpora used for training and tagging. In Section 3 we describe training and tagging of *MIM-GOLD* with *Stagger*. In Section 4 we discuss the results. In Section 5 we report on the analysis of errors made by *Stagger* when tagging the two corpora, and in Section 6 we conclude.

2 Resources

2.1 The IFD corpus

The *IFD* contains 100 fragments of text published for the first time in 1980–1989. Each fragment contains about 5,000 tokens. There are five categories of text in the corpus, approximately equally sized, four of which (80%) are literary texts from published books. The fifth category contains non-fictional texts from various sources. The tagset used in *IFD* is based on the traditional Icelandic analysis of word classes and grammatical categories, with some exceptions where that classification has been rationalized. The tagset contains about 700 possible tags of which 639 occur in *IFD*. The size of the tagset mirrors the morphological complexity of Icelandic. The corpus was tagged with a combination of automatic methods and manual checking. In the work on training and testing *Stagger* on *IFD* (Loftsson and Östling, 2012), the authors used a reduced tagset of 565 tags and a corrected version of *IFD* (Loftsson, 2009). 15.9% of the word forms in the *IFD* are ambiguous as to the tagset within the *IFD*. This figure is quite high, which illustrates the complex inflectional morphology of Icelandic. We will show in Section 5 that many of the errors made by the taggers when tagging Icelandic are due to this high ambiguity rate.

2.2 The MIM-GOLD corpus

The foundation for the building of *MIM-GOLD* is the Tagged Icelandic Corpus (*MIM*), which was released in the spring of 2013, both for search[1]

[1] http://mim.arnastofnun.is/

and download[2]. This corpus contains 25 million running words from various genres dating from the first decade of the 21*st* century (Helgadóttir et al., 2012). The compilation of *MIM-GOLD* has been described in two papers (Loftsson et al., 2010; Helgadóttir et al., 2014). *MIM-GOLD* contains about one million tokens which were sampled from 13 text types in *MIM*. The largest contributions are newspaper texts, text from published books and blog text. Other text classes include text from various websites, law text, text from school essays, text written-to-be-spoken, text from adjudications, text from radio news scripts and e-mails. *MIM-GOLD* is thus twice the size of *IFD* and the texts are more varied. About 80% of the texts in *IFD* are literary texts compared to less than 25% in *MIM-GOLD*.

The texts were tagged with the program *CorpusTagger*, which was developed for sentence segmentation, tokenization and tagging of *MIM-GOLD* (Loftsson et al., 2010). Five different individual taggers were used, after which *CombiTagger* (Henrich et al., 2009) was applied to select a single tag. All the taggers used were trained or developed using *IFD*. The *IFD* tagset was therefore used with some adjustments. Three different correction phases have been applied to the tagging of *MIM-GOLD*. In the first phase, systematic ways of error detection were applied in the form of noun phrase (NP), prepositional phrase (PP), and verb phrase (VP) error detection programs described by Loftsson (2009). In the second correction phase, all the tags in *MIM-GOLD* were checked manually. The third phase was carried out in a semi-automatic manner using *IceTagger* (Loftsson, 2008). The tags output by *IceTagger* were compared with the (presumed) correct tags in the corpus. If a difference was found, the line containing the discrepancy was marked as an error candidate and inspected manually. The total number of tags corrected in all three phases was just under 130.000.

Some adjustments were made to the tagset of *IFD* for *MIM-GOLD*, in line with the reduction of the *IFD*-tagset reported for the experiment with *Stagger*. Named entity classification for proper nouns was removed and all number constants were labelled with a single tag. Two other modifications were made. Tagging and tokenization of abbreviations was modified, and foreign names in *IFD* were tagged as proper nouns and provided with a gender in a similar fashion to Icelandic names. There is, however, considerable inconsistency in the tagging of foreign names in *MIM-GOLD*. During the second correction phase foreign names were tagged as proper nouns and marked for gender, if they were common and exhibited Icelandic inflection. When gender was difficult to decide they were tagged with unspecified gender. During the third correction phase this decision was modified such that foreign names were simply classified as foreign words. As a result foreign names in *MIM-GOLD* are classified in three different ways: As foreign words; as proper nouns with gender specified or as proper nouns with gender unspecified. As a part of further correcting the tagging of *MIM-GOLD* it is necessary to tackle this inconsistency. Since the texts in *IFD* date from the period 1980–1989 and are mainly literary texts no e-mail addresses or web addresses occur in the text. For *MIM-GOLD* a new tag was used for these entities.

2.3 The Database of Icelandic Inflection

In experiments with tagging Icelandic, extended lexicons have been derived from the Database of Icelandic Inflection (*BÍN*)[3] (Bjarnadóttir, 2012). This was done in the experiment with training and testing *Stagger* on *IFD* and is also used in the experiment reported here.

3 Experiment

The experiment with training and testing *Stagger* on *IFD* reported by Loftsson and Östling (2013) was repeated for *MIM-GOLD*. We evaluated the version of *Stagger* using linguistic features (LF) and the unknown word guesser *IceMorphy* (Loftsson, 2008) and added an extended lexicon based on *BÍN*. We did not add word embeddings (WE) as was done in the original experiment. Results are shown in Table 1. Average unknown word ratio for the *IFD* corpus when using *BÍN* is 0.97% and for the *MIM-GOLD* corpus 3.43%.[4]

4 Results

As shown in Table 1, overall accuracy for *IFD* is 93.67%. Comparable result for *MIM-GOLD* is

[2] `http://malfong.is/`

[3] *BÍN* contains about 270,000 paradigms with about 5.8 million inflectional forms. It is available at `http://bin.arnastofnun.is/`

[4] Loftsson and Östling used folds 1–9 of the *IFD* corpus for training and testing and fold 10 for development. In the present experiment we used all folds for training and testing.

Corpus	Unknown words	Known words	All words
IFD	58.31	94.02	93.67
MIM-GOLD	68.97	93.61	92.76

Table 1: Tagging accuracy when tagging *IFD* and *MIM-GOLD* using 10-fold cross-validation.

92.76%. Accuracy for known words is higher for *IFD* (94.02%) than for *MIM-GOLD* (93.61%), but accuracy for unknown words is higher for *MIM-GOLD* (68.97%) than for *IFD* (58.31%). The higher accuracy for known words in *IFD* can be explained by a greater number of tagging errors and more inconsistencies in *MIM-GOLD*. Higher accuracy for unknown words in *MIM-GOLD* are explained, at least in part, by a higher number of unknown tokens in *MIM-GOLD* that are relatively easy for the tagger to tag correctly, such as web addresses, e-mails and foreign words. In the next Section we will perform error analysis to try to explain this difference in accuracy.

5 Error analysis

Manning (2011) trained and tested the Stanford PoS tagger (Toutanova et al., 2003) on standard splits of the Wall Street Journal (WSJ) portion of the Penn Treebank for PoS tagging. In order to understand how tagging accuracy could be improved, Manning analyzed the errors made by the tagger, and suggested that the largest opportunity for further progress comes from improving the linguistic resources from which taggers are trained. To get a rough breakdown of how the linguistic resources can be improved, Manning did a small error analysis, taking a sample of 100 errors which the Stanford tagger made when tagging the WSJ. We did a similar analysis to try to explain the difference in tagging accuracy when *Stagger* was trained and tested on the two Gold standards, as described in Section 3. We took a random sample of 300 errors from each corpus. The errors were divided into six classes, as shown in Table 2.

Class of error	IFD (%)	MIM-GOLD (%)
Unknown	8.00	16.33
Improvable tagging	38.00	31.33
Insufficient context	36.00	29.67
Ambiguous tags	11.33	7.00
Inconsistency	1.00	4.67
Gold standard error	5.67	11.00
Total	**100.00**	**100.00**

Table 2: Percentage of different PoS tagging error types.

1. **Unknown words/word forms:** The word either did not appear in the training data, so the tagger had to rely on context features, or the word form did not appear with the tag it has in this context. The most common errors in this category are proper nouns tagged as common nouns. Other errors include adverbs tagged as nouns, and incorrect gender or case for words with case inflection.
2. **Improvable tagging:** This category has errors for which we could imagine a tagger finding the right tag, either by looking at the context of a few more words or looking at particular features of surrounding words. The most common errors in this category are incorrect case or gender tags for nouns and adjectives. Often there were wrong tags, even though the tagger tagged adjacent words correctly, and there should be agreement for case or gender with the word in question. In many of these errors the case is determined by a preceding verb.
   Example of failure in agreement in two adjacent words: "í guðrækilegum *[correct: singular; tagged as: plural]* umvöndunartóni *[correct: singular; tagged as: singular]*" (e. *in a tone of religious disapproval*).
3. **Insufficient contextual knowledge:** The determination of the correct tag requires broad contextual knowledge, such as (i) incorrect case with prepositions where semantics are required for the correct case; (ii) long distance assignment of case, gender or person; (iii) incorrect tagging of lower case word forms in multiword named entities.
   Example of long distance assignment of person: "Ég nefndi *[correct: 1st person; tagged as: 1st person]* síðast tvö af þessum orðum og boðaði *[correct: 1st person; tagged as: 3rd person]* ..." (e. *The last time, I talked about two of these words and announced ...*)
4. **Ambiguous tags:** Unclear or ambiguous tags, in the context, such as (i) verb tense, where a verb has the same form for past and present tense and it is unclear which is being

used; (ii) words that are commonly used in either of two genders, and (iii) examples where it is not clear whether plural or singular forms are being used, e.g. in headlines.
Example of the homonymic past and present form: "Meðan ég elti *[In corpus: past tense; Tagged as: present tense]* hann." (e. *While I chase/chased him.*)

5. **Gold standard inconsistency:** Due to discrepancy in annotation of particular word classes.
6. **Gold standard errors:** In *MIM-GOLD* there are two common error types responsible for the majority of errors in this category. 11 of the 33 errors were unanalyzed tags where a correct tag could easily be determined and is determined correctly by the tagger. 6 errors were due to split sentences, because of incorrectly determined sentence breaks. Other error types were found in both *MIM-GOLD* and *IFD*.

The above classification is subjective and other researchers might have classified a few of the errors differently, in particular when choosing between the categories *improvable tagging* and *insufficient contextual knowledge*.

For our purposes, the most important categories are the last two, where the gold standard is wrong or inconsistent.

	IFD (%)	**MIM-GOLD** (%)
Correct tag	93.67	92,76
Unknown	0.51	1.18
Improvable tagging	2.40	2.27
Insufficient context	2.28	2.14
Ambiguous tags	0.72	0.51
Inconsistency	0.06	0.34
Gold standard error	0.36	0.80
Total	**100.0**	**100.0**

Table 3: Tagging categorization in the corpora.

6 Discussion and further work

Generalizing from our sample and looking at the percentage of tags in the tagger output that falls into each category, we see a clear difference between the corpora (Table 3). We confirmed that there is statistically significant difference ($p<0.001$) in error types between the two corpora by performing a chi-square test. The proportion of words falling into the category *insufficient contextual knowledge* is roughly the same. The same applies to *improvable tagging*. Unknown words are more common in *MIM-GOLD*. This can be explained by the fact that the texts in this corpus come from more varied sources than the texts in *IFD*. *Ambiguous tags* are somewhat more common in *IFD*, this can possibly be explained by *IFD* containing mostly literary texts. The lower score for 10-fold validation is likely explained by the high rate of wrong tags and inconsistencies in *MIM-GOLD*, 1.14% of the total compared to 0.42% in *IFD*, a difference of 0.72% compared to 0.91% difference in tagging accuracy.

Results from the tagging experiment show lower tagging accuracy for *MIM-GOLD* than *IFD*. We have shown that this may, at least in part, be explained by a higher number of inconsistencies and incorrect tags in *MIM-GOLD* than *IFD*. To determine the most cost-efficient way of reducing these errors, a further error analysis should be carried out and decisions made, based on that data, as to where we should focus our efforts. When the tagging accuracy in *MIM-GOLD* has been improved, experiments will be made to merge the two corpora in training data-driven taggers.

Acknowledgments

The correction of MIM-GOLD was funded in part by the Institute of Linguistics at the University of Iceland and the Ministry of Education, Science and Culture. The authors would also like to thank Hrafn Loftsson for assistance in training Stagger.

References

Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. In *Proceedings of "Language Technology for Normalization of Less-Resourced Languages", workshop at the* 8^{th} *International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey.

Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir, and Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MIM). In *Proceedings of the workshop Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 – AfLaT, LREC 2012*, pages 67–72, Istanbul, Turkey.

Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2014. Correcting errors in a new gold standard for tagging icelandic text. In *Proceedings*

of the 9th *International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Verena Henrich, Timo Reuter, and Hrafn Loftsson. 2009. Combitagger: A system for developing combined taggers. In *Proceedings of the* 22nd *International FLAIRS Conference, Special Track: "Applied Natural Language Processing"*, Florida, USA.

Hrafn Loftsson and Robert Östling. 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), NEALT Proceedings Series 16*, Oslo, Norway.

Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the* 7th *International Conference on Language Resources and Evaluation, LREC 2010*, Valetta, Malta.

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

Hrafn Loftsson. 2009. Correcting a PoS-tagged corpus using three complementary methods. In *Proceedings of the* 12th *Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Greece.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. Springer.

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

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL, Edmonton, Canada.