Correcting Errors in a New Gold Standard for Tagging Icelandic Text

Sigrún Helgadóttir*, Hrafn Loftsson‡, Eiríkur Rögnvaldsson†
*The Árni Magnússon Institute for Icelandic Studies, ‡School of Computer Science, Reykjavík University,
†Department of Icelandic, University of Iceland
Reykjavík, Iceland
sigruhel@hi.is, hrafn@ru.is, eirikur@hi.is

Abstract

In this paper, we describe the correction of PoS tags in a new Icelandic corpus, *MIM-GOLD*, consisting of about 1 million tokens sampled from the Tagged Icelandic Corpus, *MÍM*, released in 2013. The goal is to use the corpus, among other things, as a new gold standard for training and testing PoS taggers. The construction of the corpus was first described in 2010 together with preliminary work on error detection and correction. In this paper, we describe further the correction of tags in the corpus. We describe manual correction and a method for semi-automatic error detection and correction. We show that, even after manual correction, the number of tagging errors in the corpus can be reduced significantly by applying our semi-automatic detection and correction method. After the semi-automatic error correction, preliminary evaluation of tagging accuracy shows very low error rates. We hope that the existence of the corpus will make it possible to improve PoS taggers for Icelandic text.

Keywords: Corpus creation, Gold standard, semi-automatic error correction

1. Introduction

Since the beginning of serious work on Icelandic language technology around 2000, a corpus built for the making of the Icelandic Frequency Dictionary (*IFD*) (Pind et al., 1991) has been used as a gold standard for tagging Icelandic text. The *IFD* corpus consists of text segments from 100 texts, published for the first time in 1980–1989. Each text segment contains about 5,000 tokens, and the corpus has a total of about 590k tokens. About 80% of the texts are literary texts.

The *IFD* tagset has about 700 tags, consisting of character strings where each character in a tag has a particular function. The first character denotes the word class and the remaining characters (up to 5) denote various morphological features, such as gender, number and case. The corpus was tagged with a program that used a combination of grammatical rules and frequency information. Subsequently, all tags were manually corrected. Currently, all taggers used for part-of-speech (PoS) tagging Icelandic have either been trained on the *IFD* corpus (data-driven taggers) or developed (rule-based taggers) by using the corpus.

There are, however, three main problems associated with using the *IFD* corpus as a gold standard for tagging Icelandic. First, the corpus is small in relation to the size of the tagset. This results in data sparseness problems when the corpus is used for training data-driven taggers. Second, most of the text excerpts in the corpus are literary texts, which are in many cases relatively easy to tag and do not contain many of the anomalities and irregularities that are abundant in less formal styles and often pose problems for taggers. Third, the text does not contain any material produced after 1989. The performance of data-driven taggers may, therefore, be too low when tagging text from different genres and when they encounter new words or recent linguistic phenomena, especially in recent informal texts.

For these reasons, it was decided that a new corpus that could serve as a gold standard for tagging Icelandic had to be built. The first stages of our work on this corpus, henceforth referred to as *MIM-GOLD*, are reported on in (Loftsson et al., 2010). In that paper, we described in detail the individual phases of the corpus construction, i.e. text selection and cleaning, sentence segmentation and tokenization, PoS tagging with a tagger combination method, error detection, and error correction. Furthermore, we discussed some problems that had emerged and highlighted which software tools we found to be useful. Our preliminary evaluation results showed that the error detection programs were effective and that our tagger combination method is crucial with regard to the amount of hand-correction that must be carried out in later stages of the work.

In this paper, we describe the methods that were used in further correcting the corpus, and our preliminary evaluations of the resulting tagging accuracy. Our results show that, even after the corpus has been corrected manually, tagging errors in the corpus can be reduced significantly by applying a semi-automatic error correction method.

2. Previous work

The foundation for the building of *MIM-GOLD* is the Tagged Icelandic Corpus (*MÍM*), which was released in the spring of 2013, both for search[1] and for download[2]. This corpus contains 25 million words from various genres dating from the first decade of the 21st century (Helgadóttir et al., 2012).

While the *MÍM* corpus was being compiled, one million tokens were sampled from the corpus texts to form a new gold standard, *MIM-GOLD* (Loftsson et al., 2010), thus almost doubling the training material compared to the *IFD* corpus. The texts were sampled from 13 text types, the largest contributions being newspaper text (34.3%), text from books (23.5%) and blog text (13.4%). Other smaller text classes are text from the University of Iceland Science

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

Web, text from various websites, law text, text from school essays, text written-to-be-spoken, text from adjudications, text from radio news scripts and text from web media and e-mails.

A special program, *CorpusTagger*,[3] was developed for sentence segmentation, tokenization and tagging of *MIM-GOLD*. The program uses *IceNLP* (Loftsson and Rögnvaldsson, 2007) for tokenization and sentence segmentation. The text was then tagged with five different taggers (see (Loftsson et al., 2010) for information about the individual taggers), after which *CombiTagger* (Henrich et al., 2009) was applied to select a single tag.

The quality of the PoS annotation in a corpus is crucial for the development of PoS taggers. Therefore, the field of automatic error detection/correction in corpora has gained increased interest during the last 10 years or so. Most work in this field has focused on finding elements in corpora that violate consistency, i.e. finding inconsistent tagging of a word across comparable occurrences (see, for example, (Dickinson and Meurers, 2003)).

In the original work on *MIM-GOLD*, systematic ways in the form of noun phrase (NP), prepositional phrase (PP) and verb phrase (VP) error detection programs, described by (Loftsson, 2009), were used to detect specific tagging errors. About 7,200 error candidates were detected and 82.2% of those were inspected in the work described in (Loftsson et al., 2010). To estimate the accuracy of the tagging of *MIM-GOLD*, about 1% (every 100th word) of the tags were inspected. A tag was considered correct if the whole tag string was correct. Tagging accuracy was estimated to be between 88.1% and 95.5%, depending on text type.

3. Correcting the corpus

According to the estimation of the tagging accuracy discussed in Section 2., it was imperative to reduce the tagging errors, in order for *MIM-GOLD* to be used as reliable training and testing material for PoS taggers. In this section, we describe the current work in further correcting tagging errors in the corpus.

In the first correction phase (Phase 1), during 2010–2011, a student was employed full-time during the summers and part-time during term time to manually check and correct the tags in *MIM-GOLD*. The text was arranged in the files such that one <token, tag> pair occupies a line. In order to check the tags, it is necessary to go through all the texts line by line. After Phase 1, the corpus was made available for download in 13 files as version 0.9 on the website `http://malfong.is/`, which was established in connection with the META-NORD project (Helgadóttir and Rögnvaldsson, 2013). Metadata about the corpus was entered into the META-SHARE node at Tilde,[4] also as a part of the META-NORD project.

The second phase (Phase 2) of correcting the corpus started at the end of 2012, and was carried out in the following semi-automatic manner. First, the corpus was automatically tagged with *IceTagger* (Loftsson, 2008). A script was written that compares the tags output by *IceTagger* with the (presumed) correct tags in the corpus. If a difference is found, the line containing the discrepancy is marked as an *error candidate*. A second student was employed during the summer of 2013 and part-time after that to inspect the error candidates. For each error candidate, the student was instructed to i) select the tag in the corpus; or ii) select the tag proposed by *IceTagger*; or iii) select a new correct tag when neither *IceTagger* nor the corpus contained the correct tag. At the time of writing about 80% of the texts have been checked and corrected. A second student was employed late in 2013 to help with checking the error candidates. It is anticipated that Phase 2 will be finished during the summer of 2014.

This method of using a single tagger to point to error candidates in a PoS-tagged corpus has, for example, been used by (van Halteren, 2000). Note that there may be cases where an error in the corpus coincides with an incorrect prediction of the tagger being used, i.e. when the human annotator and the tagger make the same mistake.

4. Results

In this section, we present evaluation results on the tagging accuracy in *MIM-GOLD*.

The results are shown in Table 1 which gives information about the 13 text types in the corpus. The number of tokens (second column) is not exactly the same as was reported in (Loftsson et al., 2010) since some adjustments were made to the tokenization of the text during the manual correction (Phase 1). The third column in the table, "Phase 0 accuracy", shows the estimated tagging accuracy after the correction method reported in (Loftsson et al., 2010). The fourth and the fifth columns show the size of the evaluation sample for each text type and the accuracy after Phase 1, respectively. The sixth column contains the number of error candidates found by *IceTagger*. The number and ratio of error candidates corrected (true positves) is shown in columns seven and eight. The ninth column contains the number of errors corrected, but not detected by *IceTagger*, and, finally, the last column reports the accuracy after the semi-automatic error detection and correction (Phase 2).

As can be seen from Table 1, the texts with the lowest (Phase 0) accuracy, 87.6%, were taken from Newspaper 2 (*Fréttablaðið*). The text with the highest accuracy, 95.5%, was taken from web media.

When evaluting the correctness of tags, a tag was considered correct if the whole tag string was correct as was reported in (Loftsson et al., 2010).

Let us first look at results of work performed on the Newspaper 2 text. During Phase 1, changes were made to 12,298 tags in the Newspaper 2 text, or 13.0% of the tags. To estimate the accuracy of the tagging after Phase 1, 476 tags were inspected, resulting in an estimation of 89.9% accuracy.

During Phase 2, 14,182 error candidates were detected by *IceTagger* in the Newspaper 2 text. During inspection, 5,318 of those were considered to be true errors (37.5%) and the corresponding tags in the corpus were corrected. The student changed a total of 6,343 tags during this phase, i.e. 1,025 changes were made to tags not marked as error

[3] CorpusTagger was also used for the development of the *MÍM* corpus.

[4] `http://metashare.tilde.lv/`

Text type	Tokens	Phase 0 accuracy	Evaluation sample	Phase 1 accuracy	Error candidates	Error cand. corr.	% corrected	Corr. not detected	Phase 2 accuracy
Newspaper 1[a]	248,879	92.3	1061	96.7	28,443	3,614	16.1	1,703	99.5
Books	237,065	95.1	2,510	97.7	20,677	3,453	16.7	655	99.7
Blogs	135,489	90.0	725	95.7	16,885	3,682	21,8	587	99.6
Newspaper 2[b]	94,487	87.6	476	89.9	14,182	5,318	37.5	1,025	99.8
www.visindavefur.is[c]	92,202	92.8	521	97.1	7,444	713	14.2	311	99.8
Websites[d]	65,177	94.0	164	98.2	3,826	634	16.6	164	100.0
Laws[e]	41,217	94.0			4,692				
School essays	34,357	94.2	361	95,0	3,709	753	20.3	271	99.5
Written-to-be-spoken	19,354	92.1	203	98.5	2,273	405	17.8	41	98.5
Adjudications	12,936	88.1	136	94.1	1,973	559	28.3	87	100.0
Radio news scripts[f]	11,194	92.3	119	97.5	1,161	164	14.1	31	100.0
Web media	8,524	95.5	90	96.7	1,046	167	16.0	26	100.0
E-mail	5,512	89.7	59	91.5	959	238	24.8	45	100.0
Total:	1,006,393	92.3	6,425	96.4	107,270	19,700	19.9	4,946	99.6

Table 1: Information about the various text types in the new gold standard

[a] The newspaper *Morgunblaðið*. About 80% of error candidates have been inspected.

[b] The newspaper *Fréttablaðið*.

[c] A website operated by the University of Iceland where the public can post questions on any subject. Error candidates for about half the text have been inspected.

[d] Manual correction and checking of error candidates for half the text has been accomplished.

[e] Error candidates not inspected.

[f] The Icelandic National Broadcasting Service.

candidates by *IceTagger*. The same evaluation sample (476 tags) estimates the tagging accuracy to be 99.8% after this correction phase.

Let us next look at results of work on the web media text. In the original work (Loftsson et al., 2010), accuracy for this text type was estimated to be 95.5%. During Phase 1, changes were made to 841 tags in the web media text, or 9.9% of the tags. After this phase, tagging accuracy was estimated to be 96.7% by inspecting 90 tags (about 1% sample). In Phase 2, *IceTagger* produced 1,046 error candidates. During inspection, 167 of those were considered to be true errors (16.0%) and the corresponding tags in the corpus were corrected. The student changed a total of 193 tags during this phase, i.e. 26 changes were made to tags not marked as error candidates by *IceTagger*. No errors were found in the evaluation sample after Phase 2.

As one would expect, considerably lower percentage of tags had to be corrected for the text showing higher tagging accuracy in the original work (Phase 0).

It might also be interesting to look at results for correcting tagging errors in book texts. In the original work (Loftsson et al., 2010), accuracy for this text type was estimated to be 95.1%. During Phase 1, 16,686 errors were corrected, or 7.0% of all running words. Accuracy after Phase 1 was estimated to be 97.7%, by inspecting a 1% sample (2,510 tags). During Phase 2, 20,677 error candidates were detected by *IceTagger*. Of those, 3,453 or 16.7% were considered to be true errors and were corrected. For 2,345 (68%) of the errors detected and corrected, the tag suggested by IceTagger was chosen (ii in Section 3.) and for the remaining errors (1,108) a new tag was selected (iii in Section 3.). The student changed a total of 4,108 tags, i.e. 655 tags that were not marked as error candidates by *IceTagger*.

Accuracy after Phase 2 was estimated to be 99.7% by inspecting 2,510 tags. Errors were only found in 8 tags and it is difficult to draw any conclusion about the type of errors made from such a small sample.

We will also mention results for the text type "adjudications". During Phase 1 changes were made to 1,636 tags in the adjudications text, or 12.6% of the tags. To estimate the accuracy of the tagging after Phase 1, about 1% (136) of the tags were inspected, resulting in an estimation of 94.1% accuracy.

During Phase 2, 1,973 error candidates were detected in the adjudications text. During inspection, 559 of those were considered to be true errors (28.3%) and the corresponding tags in the corpus were corrected. The student changed a total of 646 tags during this phase, i.e. 87 changes were made to tags not marked as error candidates by *IceTagger*. No errors were found in the evaluation sample after Phase 2.

As one would expect, text types with low original tagging accuracy were more difficult to check and hand-correct. The text from e-mails has original tagging accuracy of 89.7% and during Phase 1 15.2% of tags were corrected. After Phase 1 accuracy was estimated to be 91.5%, but was raised to 100% after Phase 2.

It should be pointed out that it is not always possible to decide with complete certainty which tag should be assigned to a certain token. As is well known, even trained linguists may disagree on the correct tagging of a small percentage of tokens. For instance, it is often difficult to decide whether a

certain token should be tagged as a past participle or as an adjective. The morphological form is often the same, and thus the decision will have to be made on syntactic grounds. In some cases, the syntactic environment is not decisive, as in the following example:

"... eru þessar tegundir sýnu verst á sig komnar." (are these species by far in the worst condition)

In such cases, the word in question (here *komnar*) could be tagged either as an adjective or as a past participle, and the person doing the tagging would have to depend entirely on his or her linguistic intuition to select either possibility. In the above example the PoS tagger tagged the word *komnar* as a past participle whereas both students correcting the tagging chose to tag it as an adjective.

In a few instances, the rules for deciding the correct tag were redefined between the two correction phases. This is for example true for foreign names which in some instances were originally tagged as proper names but in Phase 2 were tagged as foreign words. This may explain the poor tagging accuracy after Phase 1 for Newpaper 2 where one would expect a fair number of foreign names.

Another case where the classification was changed in Phase 2 is the tagging of prepositions/adverbs that take a clause as their complement. In Phase 1, these tokens were tagged as adverbs (*aa*) since the clause bears no overt case-marking. An example of this is the following:

"... að leitast við að bjarga andlitinu..." (try to save one's face)

In Phase 2, it was decided to tag these tokens in the same way as they would be tagged if their complement was a noun phrase instead of a clause, that is, as case-governing prepositions (*ao* or *aþ*). In the above example, *við* was tagged as *aa* during Phase 1 and as *ao* (governing accusative) in Phase 2.

One student took care of the manual correction phase (Phase 1). He estimates that on average he may have checked about 1,500 lines per hour but reached about 2,500 lines per hour at the most.

Two other students took care of the semi-automatic correction phase (Phase 2) which will be completed in the summer of 2014. One of the students did most of the work. She estimates that on the average she checked about 4,000–6,000 lines per hour but depending on the text the performance could vary between 2,000 and 9,000 lines per hour.

Both students working on Phase 2 checked 15,453 lines in a file containing part of the text from websites. In this file 1993 errors were detected by *IceTagger*. The students disagreed on 175 tags, or 8.8% of the error candidates.

5. Discussion and future work

The results presented in this paper show that, even after the *MIM-GOLD* corpus has been corrected manually (Phase 1), tagging errors in the corpus can be reduced significantly by applying a semi-automatic error correction method (Phase 2). In our semi-automatic correction method, a single tagger produces error candidates. The error candidates are then inspected manually and true errors corrected. After applying this correction method, our preliminary estimation of the tagging accuracy shows very low error rates.

Our plan is to finish the second phase of the error correction during the summer of 2014. The corpus will be made available for download as version 1.0 on the website `http://malfong.is/`. Furthermore, the corpus will be lemmatized and made available for search on the *MÍM* website (`http://mim.arnastofnun.is/`.

The corpus texts will be made available for training of data-driven taggers as ten pairs of training and test sets. More detailed analysis of errors in the corpus at different stages of processing will be performed at a later stage. The analysis may give some indication on how to improve automatic taggers for Icelandic text. A tagged corpus where tagging has been corrected, manually or with semi-automatic means, is also useful for teaching grammar, especially for students at secondary school level.

6. Acknowledgments

The original work on creating the *MIM-GOLD* corpus was partly supported by both the Icelandic Student Innovation Fund and the Icelandic Research Fund, grant 090662012. The work on correcting the tags has been supported by the META-NORD project (supported by the EU ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, grant agreement no 270899 (META-NORD)) and by the Icelandic Ministry of Education, Science and Culture as a part of the Icelandic Government's IT Policy Programme.

Kristján Friðbjörn Sigurðsson checked the tags in the whole corpus manually, Steinunn Valbjörnsdóttir took care of the bulk of the Phase 2 error checking and Brynhildur Stefánsdóttir did a part of the Phase 2 error checking.

7. References

Dickinson, M. and Meurers, W. D. (2003). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the* 11^{th} *Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary.

Helgadóttir, S. and Rögnvaldsson, E. (2013). Language Resources for Icelandic. In Smedt, K. D., Borin, L., Lindén, K., Maegaard, B., Rögnvaldsson, E., and Vider, K., editors, *Proceedings of the Workshop on Nordic Language Research Infrastructure at NODALIDA 2013*, pages 60–76. NEALT Proceedings Series 20. Linköping Electronic Conference Proceedings, Linköping, Sweden.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (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.

Henrich, V., Reuter, T., and Loftsson, H. (2009). CombiTagger: A System for Developing Combined Taggers. In *Proceedings of the* 22^{nd} *International FLAIRS Conference, Special Track: "Applied Natural Language Processing"*, Sanibel Island, Florida, USA.

Loftsson, H. and Rögnvaldsson, E. (2007). IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Proceedings of Interspeech 2007, Special Session: "Speech and language technology for less-resourced languages"*, Antwerp, Belgium.

Loftsson, H., Yngvason, J. H., Helgadóttir, S., and Rögnvaldsson, E. (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 7^{th} International Conference on Language Resources and Evaluation, LREC 2010*, Valetta, Malta.

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

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

Pind, J., Magnússon, F., and Briem, S. (1991). *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavik, Iceland.

van Halteren, H. (2000). The detection of inconsistency in manually tagged text. In A. Abeillé, T. B. and Uszkoreit, H., editors, *Proceedings of the 2nd Lingustically Interpreted Corpora*, Luxembourgh.