

Exploring the human effect on the accuracy of PoS taggers

Martha Dís Brandt

Reykjavík University, Iceland

marthab08@ru.is

May 17, 2009

Abstract

This paper explores the impact of inconsistencies stemming from human mistakes on the accuracy of part-of-speech (PoS) taggers in Icelandic. It reveals previously unpublished information regarding work on the original corpus that is now used as a gold standard for Icelandic. With this new knowledge I look at recent efforts to correct said corpus, what that work has done for the tagging accuracy and contemplate possible ripple effects.

1 Introduction

My goal with this paper was to look at the evaluation of accuracy of PoS taggers using the Icelandic language, in particular with regards to the accuracy of the original tagging of the gold standard, i.e. the Icelandic Frequency Dictionary (IFD).

Therefore I have sought counsel from one of the authors of the IFD to determine the extent of the human factor in regards to its creation. My hope is that the hitherto unpublished information can shed further light on the research recently done by Loftsson (2009) on correcting a PoS tagged corpus.

In the second chapter of this paper I take a closer look at the creation of the Icelandic gold standard. In chapter three I consider the human factor, and chapter four looks at the work on error detection and correction of the IFD corpus, while I tie the previous chapters together in the summary in chapter five.

2 Creating a Gold Standard

The Icelandic Frequency Dictionary (IFD) is the gold standard for tagging Icelandic text. Work began on the IFD in 1989 and was completed in 1991, it is a balanced corpus with about 500,000 tokens. The texts selected for the corpus were equally distributed throughout five different genres, twenty from each, and were all roughly the same size, i.e. about 5,000 tokens per text (see Table 1). Furthermore, the authors and translators of any selected text could not be the author nor translator of a second text in the corpus.

Genre	No. of texts	Ca. no. of tokens
Fiction by Icelandic authors	20	100,000
Fiction translated into Icelandic	20	100,000
Biographies and memoirs	20	100,000
Educational texts (soft sciences)	10	50,000
Educational texts (hard sciences)	10	50,000
Icelandic books for children and adolescents	10	50,000
Translated books for children and adolescents	10	50,000

Table 1: Distribution of text selections for the IFD corpus

As has been established, the IFD is used as the gold standard to measure the accuracy of taggers when tagging Icelandic text. What hasn't been brought to light is how the IFD was affected by the human factor.

3 The Human Factor

From Pind et al. (1991), Briem (1990) and Briem (2009) I have determined the following regarding the creation of the IFD:

The texts:

- All of the texts were proofread and apparent misspellings and typing errors were corrected.
- Authors' spelling eccentricities were not corrected.¹

¹I must say that I had my doubts about this decision, for in my opinion these eccentricities are just another form of misspelling. However, I can accept that the meaning of these words are the same as those correctly spelled and I will assume that the authors were consistent in their eccentricities.

- Words containing 'z' were changed according to current Icelandic spelling and grammatical rules.²

The tagger:

- Stefán Briem developed a tagger to facilitate the job of tagging.
- Said tagger uses 75 rules, mostly morphological, and a word-collection.
- Each rule assigns points in order to select the correct tag for the token.
- The first version of the tagger provided ca. 70% correct tags.

The tagger used the results of a previous word frequency analysis of about 54,000 tokens (Magnússon, 1988) to process the first half of the corpus. This portion was then proofread and manually corrected either by Stefán Briem or Friðrik Magnússon, i.e. they divided the task and did not overlap each other's work while still collaborating closely (Briem, 2009).

At this point the corrected first half of the corpus was added to the tagger's word-collection which improved tagging by ca. 10% before completing the tagging of the second half of the corpus. The same method was used for manual corrections of this portion as was used on the first half of the corpus.

Stefán Briem explained to me that he and Friðrik Magnússon divided the task of proofreading and manually correcting the output of the tagger. He also said that since Friðrik was a highly educated Icelandic linguist while he himself was not, that the two collaborated closely on said task. Therefore, it is my opinion that the corpus was reviewed by only one human, half by an expert and half by a professional, as none of the text was revised by both of them. This means that there were no inconsistencies of tags for particular tokens, as each token will only have one final result regardless of where the tag came from.

Since there are no inconsistencies in the IFD then there cannot be instances of taggers agreeing with any of the human experts where the humans disagree. This also means that a tagger can easily be tailored to come up with the same result as the IFD, giving the tagger a higher accuracy rating than it should receive.

²The letter 'z' was removed from the Icelandic alphabet and where it was formerly used the letter 's' should now be in its place. There are a few exceptions to this rule, e.g. if 'z' was used in a name it may remain unchanged.

4 Detecting Errors in PoS-tagged corpora

It is common knowledge that humans make mistakes, this cannot be avoided. Even when humans are being excruciatingly meticulous, errors will occur. Therefore it is inevitable that when computerized results are compared to human results that the accuracy of such calculations is affected.

Loftsson (2009) used three different methods to detect errors in a PoS-tagged corpus, i.e. the Icelandic Frequency Dictionary. First a *variation n-gram* method proposed by (Dickinson and Meurers, 2003), then a combination of five different taggers in contrast with the IFD gold standard, and finally he used shallow parsing with *IceParser* (Loftsson and Rögnvaldsson, 2007) to identify error candidates.

752 error candidates (0.13% of the corpus) were manually inspected from the *variation n-gram* method and an additional 9,633 were browsed through. 0.9% of the corpus was inspected after the combination method (ca. 5,300 tokens). 1,489 error candidates (0.25% of the corpus) were inspected as a result of patterns derived from the shallow parsing.

There were 1,334 true and unique errors that were discovered with these 3 complementary methods, allowing 0.23% of the IFD corpus to be corrected, which provided grounds for establishing whether the human factor had indeed affected the calculations of the tagging accuracy.

In order to verify this, Loftsson (2009) compared the original accuracy calculations of *IceTagger* (Loftsson, 2008) and *TnT* (Brants, 2000), found in columns 1 & 2 on Figure 1, with the accuracy of the unmodified programs using the corrected corpus, found in columns 5 & 6 on the same Figure.

Then both taggers were augmented, a Hidden Markov Model (HMM) for *IceTagger* and *TnT* made use of *IceMorph* (which is the morphological analyser of *IceTagger*), for additional disambiguation functionality and run on both the original IFD corpus and the corrected one. The accuracy results can be found in columns 3 & 4 and 7 & 8, respectively, also on Figure 1.

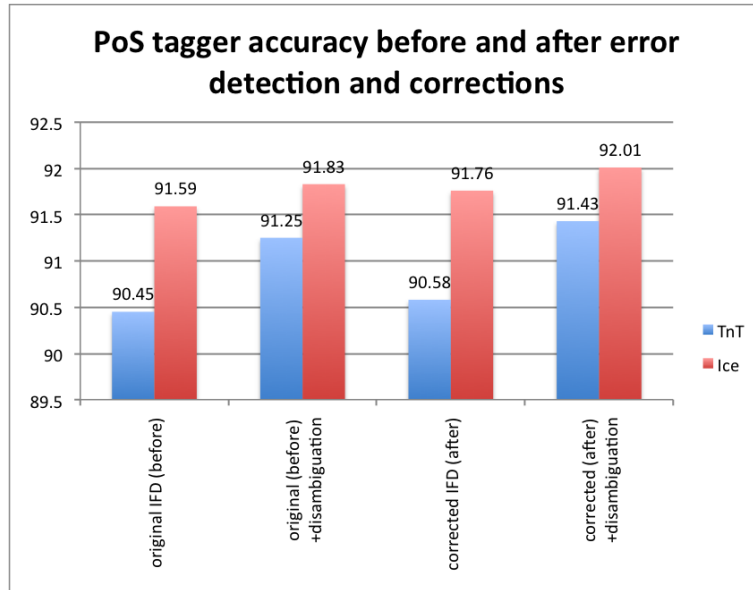


Figure 1: Exaggerated graph comparing accuracy of TnT and IceTagger for all words, with and without added disambiguation components. *Reproduced from (Loftsson, 2009).*

5 Summary

The subject of this paper was whether the human aspect of PoS tagging has had any effect on the calculated accuracy of PoS taggers for Icelandic. It is my opinion that the accuracy has been affected, and the study by Loftsson (2009) substantiates that view.

It is human to err, as the saying goes, and with that in mind, consider that computerized PoS taggers are being compared to the outcome of human PoS tagging. I believe that where only one human expert reviews the annotation of a corpus, it is likelier for errors to slip past unnoticed. Also, if sections of the corpus are reviewed by different humans then it is likelier that whatever errors do occur will vary depending on which person makes them.

I am sure that if the 5-tagger method implemented by Loftsson (2009) were to be used with 4 out of 5 taggers agreeing, instead of all 5, that more annotation errors could be corrected in the IFD corpus. Which in turn would ripple back to allow re-evaluation of the tagging accuracy for Icelandic again. However, I think it is highly unlikely that 100% correct tagging is achievable.

Acknowledgments

I would like to thank Joakim Nivre for awakening my curiosity on this aspect of Natural Language Processing, Hrafn Loftsson for his guidance, and Stefán Briem for providing invaluable information on the early stages of creating the Icelandic Frequency Dictionary.

References

- Brants, T. (2000). TnT: A statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, Seattle, Washington, USA.
- Briem, S. (1990). Automatísk morfologísk analyse af íslandskum tekst. In Pind, J. and Rögnvaldsson, E., editors, *Papers from the Seventh Scandinavian Conference of Computational Linguistics Reykjavik 1989*, pages 3–13, Reykjavik, Iceland. Institute of Lexicography, Institute of Linguistics.
- Briem, S. (2009). Personal email communication.
- Dickinson, M. and Meurers, W. D. (2003). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, Budapest, Hungary. Association for Computational Linguistics.
- 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 12th Conference of the European Chapter of the ACL*, pages 523–531, Athens, Greece. Association for Computational Linguistics.
- Loftsson, H. and Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NoDaLiDa)*, Tartu, Estonia.
- Magnússon, F. (1988). Hvað er títt? Tíðnikönnun Orðabókar Háskólans. *Orð og tunga*, 1:1–49.
- Pind, J., Magnússon, F., and Briem, S. (1991). *The Icelandic Frequency Dictionary*. The Institute of Lexicography, University of Iceland.