Tagging Icelandic text: An experiment with integrations and combinations of taggers

Hrafn Loftsson
Department of Computer Science, 211 Regent Court, Portobello Street, S1 4DP, Sheffield, United Kingdom (h.loftsson@dcs.shef.ac.uk)

Abstract. We use integrations and combinations of taggers to improve the tagging accuracy of Icelandic text. The accuracy of the best performing integrated tagger, which consists of our linguistic rule-based tagger for initial disambiguation and a trigram tagger for full disambiguation, is 91.80%. Combining five different taggers, using simple voting, results in 93.34% accuracy. By adding two linguistically motivated rules to the combined tagger, we obtain an accuracy of 93.48%. This method reduces the error rate by 20.5%, with respect to the best performing tagger in the combination pool.

Keywords: combination of taggers, integration of taggers, linguistically motivated rules, simple voting, tagging accuracy

Abbreviations: DDT – Data-Driven Taggers; HMM – Hidden Markov Model; IFD – Icelandic Frequency Dictionary; LMR – Linguistically Motivated Rules

1. Introduction

Icelandic is a morphologically complex language, whose main part-of-speech tagset consists of about 660 tags.

We have previously developed a linguistic rule-based tagger, *IceTagger* (hereafter referred to as *Ice*), which achieves 91.54% average tagging accuracy. Moreover, we have used tagger integration (i.e. making one tagger use a feature or a functionality of another tagger), and a combination of three taggers, using simple voting, to achieve 92.94% accuracy (Loftsson, 2006a, 2006b).

In this paper, we present additional tagger integration methods and build a combined tagger using five taggers. Furthermore, we show how simple linguistically motivated rules (LMR) can improve the tagging accuracy.

Our best performing integrated tagger achieves 91.80% tagging accuracy. By combining five taggers, using simple voting, we obtain 93.34% accuracy. When adding two LMR to the combined tagger, the accuracy increases to 93.48% and reduces the error rate by 20.5%, with respect to the best performing tagger in the combination pool.

This paper is organised as follows. In Section 2, we briefly describe the Icelandic language, the tagset and the corpus used. The individual

taggers used in this research are described in Section 3. Section 4 is devoted to our integration methods and Section 5 describes the combination methods. Evaluation results are presented in Section 6, and we conclude, in section 7, with a summary and direction for future work.

2. The Icelandic language, the tagset and the corpus

The Icelandic language is one of the Nordic languages which comprise the North-Germanic branch of the Germanic language tree. The language is morphologically rich, mainly due to inflectional complexity.

Due to the morphological richness of the language, the main tagset, constructed in the compilation of the Icelandic Frequency Dictionary (*IFD*) corpus (Pind et al., 1991), is large (about 660 tags) compared to tagsets of related languages. Each character in a tag has a particular function. The first character denotes the word class. For each word class there is a predefined number of additional characters (at most six) which describe morphological features, like gender, number and case for nouns; degree and declension for adjectives; voice, mood and tense for verbs, etc. The reader is referred to (Loftsson, 2006a; Pind et al., 1991) for a more complete description of the tagset.

For the purpose of using ten-fold cross-validation, ten different disjoint pairs of files have been created using the *IFD* corpus. Each pair consists of a training set, containing about 90% of the tokens from the corpus, and a test set, containing about 10% of the tokens. The test corpora do not share any examples, whereas the training corpora overlap (Helgadóttir (2004) describes the corpus more thoroughly).

3. Individual taggers used

The data-driven taggers (DDT) used in this research are state-of-the art: *fnTBL* (hereafter referred to as *TBL*) (Ngai and Florian, 2001), based on transformation-based error-driven learning; *MXPOST* (hereafter referred to as *MXP*) (Ratnaparkhi, 1996), based on a maximum entropy approach; *MBT* (Daelemans et al., 1996), based on memory-based learning; and *TnT* (Brants, 2000), based on a Hidden Markov Model (HMM). Additionally, we used the taggers *Ice* and *Tri*, described briefly below.

Ice, a linguistic rule-based tagger, uses hand-written local linguistic elimination rules (the idea is borrowed from the well known *Constraint Grammar* framework (Karlsson et al., 1995)), along with a list of idioms (derived semi-automatically from the *IFD* corpus), for initial disambiguation. Thereafter, various heuristics (algorithmic procedures) are

used to force feature agreement between words, effectively eliminating more tags. At the end, for a word not fully disambiguated, the default rule is to select the word's most frequent tag. In addition to a lexicon derived from the *IFD* corpus, *Ice* uses a special lexicon which mainly includes tags for irregular verb forms. When testing *Ice* on the *IFD* corpus, this has the effect that the average unknown word ratio is slightly lower than the corresponding ratio when testing the DDT, i.e. 6.79% vs. 6.84%. *Ice* uses an integrated morphological analyser, *IceMorphy*, to obtain the possible tags for unknown words. *Ice* and *IceMorphy* are described in detail in (Loftsson, 2006a, 2006b).

Tri is our re-implementation of the *TnT* tagger. The difference between these two taggers is that *Tri* uses the same list of idioms as *Ice*, and the special lexicon described above, as a backup lexicon.

All taggers were trained and tested, with their default options on the *IFD* corpus using ten-fold cross-validation. The only exception is the *MBT* tagger, for which we conducted an experiment to select the optimal settings: features *-p ddwfaa* and *-P cndFasssss*, search algorithm *IB1-IG* (k=5) and the *modified value distance metric*; for details consult (Daelemans et al., 2003).

When implementing *Ice*, 10% of the *IFD* corpus, i.e. the tenth test corpus, was used to develop rules. Therefore, the accuracy figures presented for all taggers in table I (and henceforth) are average figures computed using only the first nine test corpora. All differences in tagging accuracy in table I (and subsequently in tables II and III) are significant at $\alpha < 0.05$, using McNemar's chi-squared test as described by Dietterich (1998).

4. Integration of taggers

We define tagger integration as enabling one tagger to use a feature or a functionality of another tagger. In this section, we describe four integration methods, all of which have resulted in an improved tagging accuracy of Icelandic text. The first two methods, which consist of integrating our morphological analyser with state-of-the-art DDT, are described in more detail in (Loftsson, 2006a). The latter two methods are new.

First, in order to improve the relatively poor tagging accuracy of *TBL* for unknown words (see table I), we made *IceMorphy* provide *TBL* with an initial tag (the most probable tag from the set of guessed tags) for each unknown word. This increased the overall accuracy of *TBL* from 89.33% to 90.15%.

Table I. The average tagging accuracy of Icelandic text using various taggers.

Words	Base[a]	MXP	MBT	TBL	TnT	Tri	Ice
Unknown	4.39%	62.29%	59.40%	55.51%	71.68%	71.04%	75.09%
Known	81.84%	91.00%	91.47%	91.82%	91.82%	91.87%	92.74%
All	76.27%	89.03%	89.28%	89.33%	90.44%	90.46%	**91.54%**
Δ_{Err}[b]		53.77%	54.83%	55.04%	59.71%	59.80%	64.35%

[a] A base tagger which assigns each known word its most frequent tag, and the most frequent noun tag/proper noun tag to lower case/upper case unknown words.
[b] Error reduction with regard to the errors made by the base tagger for all words.

Table II. Average tagging accuracy using integration of taggers.

Words	TBL*	TnT*	Tri*	Ice*
Unknown words	66.30%	72.80%	74.46%	75.33%
Known words	91.90%	92.54%	92.58%	93.00%
All words	90.15%	91.18%	91.34%	**91.80%**
Δ_{Err}[a]	7.69%	7.74%	9.13%	3.07%

[a] Error reduction with regard to the errors made by the unchanged version of the corresponding tagger for all words.

Second, we improved the accuracy of the *TnT* tagger in the following manner. *IceMorphy* is able to generate missing tags in a tag profile for a word belonging to a particular morphological class. We used this feature of *IceMorphy* to generate a "filled" lexicon, to be used by the *TnT* tagger. Each generated missing tag is marked with the frequency 1. This improved *TnT's* accuracy from 90.44% to 91.18%.

The third integration method is an integration of our *Tri* tagger with *IceMorphy*. In order to improve the accuracy of this tagger, we call *IceMorphy* from within the *Tri* tagger to obtain possible tags for unknown words. Moreover, we made the *Tri* tagger benefit from the lexicon filling mechanism described above. This version of the *Tri* tagger achieves an accuracy of 91.34%.

Lastly, we integrated our linguistic rule-based tagger with the *Tri* tagger. By making *Ice* call the *Tri* tagger for full disambiguation (instead of simply selecting the most frequent tag for a word not fully disambiguated) the overall tagging accuracy increases from 91.54% to

91.80%. A similar approach has, for example, been used for tagging text in the highly inflected Czech language (Hajiĉ et al., 2001).

Henceforth, we will refer to the *TBL+IceMorphy* tagger as *TBL**, the *TnT+IceMorphy* tagger as *TnT**, the *Tri+IceMorphy* tagger as *Tri** and the *Ice+Tri* tagger as *Ice**. Note that all our integrated systems run like a single tagger, i.e. the text to be tagged is processed and tagged only once. The change in accuracy between the unchanged versions of the taggers and the integrated taggers can be seen by comparing tables I and II.

5. Combination of taggers

It has been shown that combining taggers will often result in higher tagging accuracy than achieved by individual taggers (van Halteren et al., 2001; Sjöbergh, 2003). The reason is that different taggers tend to produce different (complementary) errors and the differences can be exploited to yield better results.

A number of different combination methods exists, e.g. simple voting, weighted voting and stacking (see (van Halteren et al., 2001) for a good overview), as well as combinations using LMR (Borin, 2000). In this experiment, we combine taggers using simple voting and LMR.

In simple voting, each tagger gets an equal vote when voting for a tag and the tag with the highest number of votes is selected. When combining taggers using LMR the relative strength of a given tagger, in a particular linguistic context, is utilised in the combination.

6. Evaluation

6.1. Simple voting

In the first tagger combination experiment for Icelandic, the *MXP*, *TBL* and *TnT* taggers were used in a simple voting scheme, obtaining an average accuracy of 91.54% (Helgadóttir, 2004) (see row 1 in table III). By using *Ice* instead of the relatively low accuracy tagger *MXP*, the accuracy increases substantially, to 92.61% (see row 2). By adding the two least accurate taggers, *MXP* and *MBT*, to the combination pool, the overall accuracy increases further to 92.80% (see row 3).

In (Loftsson, 2006a), we had improved the first simple voting result for Icelandic text by combining *TBL**, *TnT** and *Ice* – obtaining an accuracy of 92.94% (see row 4 in table III). Here, we improve this result by adding the taggers *MXP* and *MBT* to the combination pool, resulting

in an accuracy increase to 93.29% (see row 5). This time, the addition of the two taggers is about twice as effective than before, mainly because of higher accuracy for unknown words. The errors made by these two taggers for unknown words are probably, in many cases, complementary to the corresponding errors proposed by *TBL** (which receives "help" from *IceMorphy* for unknown words), but less complementary to *TBL*, which was used in the combination pool in row 3.

The benefit of using our integrated taggers is clear by comparing the accuracy of the combined taggers in rows 2 and 4, and in rows 3 and 5, in table III.

Table III. Average tagging accuracy using combination of taggers.

			Accuracy of words			
#	Combination (simple voting[a])	Rule	Unkn.	Known	All	Δ_{E1}[b]
1.	MXP+TBL+TnT	None	71.80%	92.99%	**91.54%**	12.2%
2.	TBL+TnT+Ice	None	76.76%	93.77%	**92.61%**	12.7%
3.	MXP+MBT+TBL+TnT+Ice	None	76.74%	93.97%	**92.80%**	14.9%
4.	TBL*+TnT*+Ice	None	76.55%	94.13%	**92.94%**	16.6%
5.	MXP+MBT+TBL*+TnT*+Ice	None	78.70%	94.36%	**93.29%**	20.7%
6.	MXP+MBT+TBL*+TnT*+Ice*	None	78.65%	94.41%	**93.34%**	18.8%
7.	MXP+MBT+TBL*+TnT*+Ice*	1	78.66%	94.50%	**93.43%**	19.9%
8.	MXP+MBT+TBL*+TnT*+Ice*	1 & 2	78.68%	94.56%	**93.48%**	20.5%

[a] Majority voting, in which ties are resolved by selecting the tag of the most accurate tagger in the tie.

[b] Error reduction with regard to the best single tagger in the combination.

Finally, we replaced the standard version of *Ice* with *Ice**, i.e. *Ice* with the *Tri* tagger for full disambiguation. This slightly improved the overall tagging accuracy (see row 6 in table III).

6.2. Linguistically motivated rules

We wrote two kinds of LMR, both of which are based on specific strengths of *Ice*, and which are only fired if not all taggers agree.

First, we have noticed that the DDT have difficulties providing the correct tag in a particular context, whereas *Ice* performs considerably better for the same context. This occurs, for example, where there are "long" dependencies between a subject and a verb and the verb has the same lexical form for 1^{st} and 3^{rd} person. A typical example is "*ég opnadi dyrnar, steig inn ...*" (I opened door, stepped inside ...). The correct tag for the verb "*steig*" includes a 1^{st} person feature, but all the DDT

propose a 3^{rd} person tag. The reason is that the 3^{rd} person tag is more frequent and the DDT have a limited context window size.

Another example of a long dependency is between a subject and a reflexive pronoun, e.g. "*... sagdi konan og færdi sig*" (...said woman and moved herself), in which the reflexive pronoun has the same lexical form in all genders.

In both these examples, *Ice* provides the correct tag, because of its built-in feature agreement functionality, but is outvoted by the DDT. We, thus, built a simple rule which always selects the 1^{st} person verb tags if they are suggested by *Ice*, and the tags suggested by *Ice* for the reflexive pronouns "*sig*", "*sér*" and "*sín*".

For the second rule, we used a feature agreement constraint: "If all the tags, provided by the individual taggers for the current word, are nominal tags and the current tag provided by *Ice* agrees in gender, number and case with the preceding (selected) nominal tag or the following (yet to be selected) nominal tag, then choose *Ice's* tag". Using this rule improves the tagging accuracy, because disambiguating using nominal feature agreement is one of the strengths of *Ice*.

Row 8 of table III shows that using simple voting along with the two LMR results in an overall tagging accuracy of 93.48%.

7. Conclusion

We have used integrations and combinations of taggers to improve the tagging accuracy of Icelandic text. Accuracy of the best performing integrated tagger, consisting of using *IceTagger*, for initial disambiguation, along with a HMM tagger, for full disambiguation, is 91.80%.

The best performing simple voting method, using five individual taggers, achieves 93.34% tagging accuracy. Furthermore, when adding two LMR to the combined tagger, the accuracy increases to 93.48%.

We envision several ways to improve the accuracy further. First, increasing the training corpus size for the DDT might be a feasible option, because our best combination method could be used for initial tagging, followed by manual corrections. Second, adding more taggers to the combination pool might improve the accuracy. Third, adding more linguistic knowledge to *Ice* is possible, especially with the purpose of fixing frequent errors. Fourth, reducing the tagset, without too much loss of information, is worthwhile. Lastly, we would like to experiment with using stacking methods, e.g. using a memory-based method which learns from the tagged output of the individual taggers.

Acknowledgements

Thanks to the Institute of Lexicography at the University of Iceland, for providing access to the *IFD* corpus, and Professor Y. Wilks for valuable comments and suggestions in the preparation of this paper.

References

Borin, L.: 2000, 'Something borrowed, something blue: Rule-based combination of POS taggers'. In: *Proceedings of the* 2^{nd} *International Conference on Language Resources and Evaluation*. Athens, Greece.

Brants, T.: 2000, 'TnT: A statistical part-of-speech tagger'. In: *Proceedings of the* 6^{th} *Conference on Applied natural language processing*. Seattle, WA, USA.

Daelemans, W., J. Zavrel, P. Berck, and S. Gillis: 1996, 'MBT: a Memory-Based Part of Speech Tagger-Generator'. In: *Proceedings of the* 4^{th} *Workshop on Very Large Corpora*. Copenhagen, Denmark.

Daelemans, W., J. Zavrel, and A. van den Bosch: 2003, 'MBT: Memory-Based Tagger'. Reference Guide: ILK Technical Report - ILK 03-13.

Dietterich, T. G.: 1998, 'Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms'. *Neural Computation* **10**(7), 1895–1924.

Hajič, J., P. Krbec, K. Oliva, P. Květoň, and V. Petkevič: 2001, 'Serial Combination of Rules and Statistics: A Case Study in Czech Tagging'. In: *Proceedings of the* 39^{th} *Association of Computational Linguistics Conference*. Toulouse, France.

Helgadóttir, S.: 2004, 'Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic'. In: H. Holmboe (ed.): *Nordisk Sprogteknologi 2004*. Museum Tusculanums Forlag.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila: 1995, *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Berlin, Germany: Mouton de Gruyter.

Loftsson, H.: 2006a, 'Tagging Icelandic text: A linguistic rule-based approach'. Technical Report CS-06-04, Department of Computer Science, University of Sheffield.

Loftsson, H.: 2006b, 'Tagging a Morphologically Complex Language Using Heuristics'. In: T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala (eds.): *Advances in Natural Language Processing,* 5^{th} *International Conference on NLP, FinTAL 2006, Proceedings*. Turku, Finland.

Ngai, G. and R. Florian: 2001, 'Transformation-Based Learning in the Fast Lane'. In: *Proceedings of the* 2^{nd} *Conference of the North American Chapter of the ACL*. Pittsburgh, PA, USA.

Pind, J., F. Magnússon, and S. Briem: 1991, *The Icelandic Frequency Dictionary*. Reykjavik, Iceland: The Institute of Lexicography at the University of Iceland.

Ratnaparkhi, A.: 1996, 'A Maximum Entropy Part-of-Speech Tagger'. In: *Proceedings of the Empirical Methods in Natural Language Processing Conference*. Philadelphia, PA, USA.

Sjöbergh, J.: 2003, 'Combining POS-taggers for improved accuracy on Swedish text'. In: *Proceedings of NoDaLiDa 2003*. Reykjavik, Iceland.

van Halteren, H., J. Zavrel, and W. Daelemans: 2001, 'Improving Accuracy in Wordclass Tagging through Combination of Machine Learning Systems'. *Computational Linguistics* **27**(2), 199–230.