Improving the PoS tagging accuracy of Icelandic text

Hrafn Loftsson **Ida Kramarczyk**
School of Computer Science
Reykjavik University
Reykjavik, Iceland
{hrafn,ida07}@ru.is

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

Eiríkur Rögnvaldsson
Department of Icelandic
University of Iceland
Reykjavik, Iceland
eirikur@hi.is

Abstract

Previous work on part-of-speech (PoS) tagging Icelandic has shown that the morphological complexity of the language poses considerable difficulties for PoS taggers. In this paper, we increase the tagging accuracy of Icelandic text by using two methods. First, we present a new tagger, by integrating an HMM tagger into a linguistic rule-based tagger. Our tagger obtains state-of-the-art tagging accuracy of 92.31% using the standard test set derived from the IFD corpus, and 92.51% using a corrected version of the corpus. Second, we design an external tagset, by removing information from the internal tagset which reflects distinctions that are not morphologically based. Using the external tagset for evaluation, the tagging accuracy further increases to 93.63%.

1 Introduction

Icelandic is a morphologically complex language for which the task of part-of-speech (PoS) tagging has turned out to be difficult, both for data-driven and linguistic rule-based taggers (Helgadóttir, 2005; Loftsson, 2006; Loftsson, 2008; Dredze and Wallenberg, 2008). Before the work presented in this paper, the current state-of-the-art tagging accuracy was 92.06%, obtained using a bidirectional sequence classification method (Dredze and Wallenberg, 2008) and testing using the Icelandic Frequency Dictionary (IFD) corpus (Pind et al., 1991).

There are at least three reasons for this low accuracy – all of them are manifestations of the fact that the Icelandic language is morphologically complex. First, the large tagset used (about 700 tags) and the relatively small training corpus (about 590k tokens) causes data sparseness problems. Second, inherent long range tag dependencies in Icelandic text are difficult for many PoS tagging methods to resolve. Third, the tagset reflects distinctions which may be difficult to resolve at the level of PoS tagging, because some of them are not morphologically based.

The main material in this paper is threefold. First (in Section 2), we review previous tagging approaches for Icelandic and present a new tagger by integrating a Hidden Markov Model (HMM) tagger into a linguistic rule-based tagger in a novel way. Our tagger obtains an accuracy of 92.31%, which amounts to about a 3.2% error reduction rate compared to the previous best result. Furthermore, the accuracy increases to 92.51% when testing using a corrected version of the IFD corpus.

Second (in Section 3), we propose an external tagset (the tagset used for evaluation) by removing information from the internal tagset (the tagset used by a tagger) which reflects distinctions that are not morphologically based. These reductions should not affect the effectiveness of the tagset in practical applications. The tagging accuracy further increases to 93.63% using the external tagset.

Third (in Section 4), we discuss the results and provide directions for future work on tagging Icelandic.

2 Tagging Icelandic

In this section, we first describe the corpus used for training, developing and testing PoS taggers for Icelandic and the underlying tagset. Second, we review, in some detail, previous work on tagging Icelandic. Third, we describe our new tagging method, which results in a new state-of-the-art tagging accuracy. Finally, we evaluate our method using a corrected version of the original corpus.

2.1 The IFD corpus

All published tagging results hitherto for Icelandic have been based on the IFD corpus (Pind et al.,

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 103–110

1991). The IFD corpus is a balanced corpus, consisting of about 590k tokens. All 100 text fragments in the corpus were published for the first time in 1980–1989. The corpus comprises five categories of texts, i.e. Icelandic fiction, translated fiction, biographies and memoirs, non-fiction and books for children and youngsters. No two texts are attributed to the same person and all texts start and finish with a complete sentence. The corpus was semi-automatically tagged using a tagger based on linguistic rules and probabilities (Briem, 1989).

The main Icelandic tagset, constructed in the compilation of the IFD corpus, consists of about 700 possible tags, which is large compared to related languages. In this tagset, 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. To illustrate, consider the word "*strákarnir*" ('(the) boys'). The corresponding tag is "nkfng", denoting noun (*n*), masculine (*k*), plural (*f*), nominative (*n*), and suffixed definite article (*g*).

2.2 Previous tagging results

The first tagging results for Icelandic were based on an experiment using several data-driven taggers (Helgadóttir, 2005; Helgadóttir, 2007). The highest tagging accuracy, 90.4%, was obtained by the *TnT* tagger (Brants, 2000), a popular HMM tagger. By using a simplified version of the tagset the accuracy of TnT increased to 91.83%, and further to 98.14% when only considering the word class (the first letter of a tag). All results were obtained using 10-fold cross-validation and the corresponding data-splits now form the standard training (90%) and test corpora (10%) for evaluating taggers for Icelandic. The average unknown word ratio using this data-split is 6.8%.

Data sparseness, non-local tag dependencies and fine-grained distinctions in the tagset are mainly to blame for the relatively low tagging accuracy obtained by (at the time) state-of-the-art data-driven taggers. This motivated the development of a linguistic rule-based tagger for Icelandic (Loftsson, 2008). The tagger, *IceTagger*, is reductionistic in nature, i.e. it removes inappropriate tags from words in a given context. IceTagger first applies local rules (175 in total) for initial disambiguation and then uses a set of heuristics (global rules) for further disambiguation. The heuristics, for example, enforce feature agreement between subjects and verbs, between subjects and predicative complements, and between prepositions and the following nominals. If a word is still ambiguous after the application of the heuristics, the default heuristic is simply to choose the most frequent tag for the word.

An important part of IceTagger is the unknown word guesser, *IceMorphy* (Loftsson, 2008). It guesses the *tag profile* (the set of tags; sometimes called the *ambiguity class*) for unknown words by applying morphological analysis and ending analysis. In addition, IceMorphy can fill in the *tag profile gaps*[1] in the dictionary for words belonging to certain morphological classes.

For the sake of being easily able to compare the tagging accuracy between different methods, IceTagger and IceMorphy only use data resources based on the IFD corpus, i.e. data which is also available to data-driven taggers. The tagging accuracy of IceTagger is about 91.6%, a large improvement on the accuracy obtained by the TnT tagger. The tenth data file in the standard data-split was used for the development of IceTagger. Therefore, the average tagging accuracy is based on testing using the first nine test corpora.

Furthermore, by using the idea of a serial combination of a rule-based and a statistical tagger (Hajič et al., 2001), specifically making an HMM tagger, *TriTagger*, disambiguate words which IceTagger cannot fully disambiguate, the tagging accuracy increases to about 91.8% (Loftsson, 2006). In Table 1, we refer to this tagger as *Ice+HMM*[2].

Loftsson (2008) has also experimented with improving the tagging accuracy of the TnT tagger. The improvement consists of using IceMorphy to generate a "filled" dictionary, i.e. a dictionary for which tag profile gaps for certain words have been filled. Using such a dictionary significantly increases the tagging accuracy of TnT, from about 90.5% to about 91.3%. We refer to this tagger as the *TnT** tagger (see Table 1).

Before our current work, the state-of-the art

[1] A tag profile gap for a word occurs when a tag is missing from the tag profile. This occurs, for example, if not all possible tags for a given word are encountered during training.

[2] In (Loftsson, 2006), this tagger is called *Ice**.

Tagger	Unknown	Known	All
TnT	71.82	91.82	90.45
TnT*	72.98	92.60	91.25
IceTagger	75.30	92.78	91.59
Ice+HMM	75.63	93.01	91.83
BI+WC+CT	69.74	93.70	92.06
HMM+Ice	76.10	93.36	92.19
HMM+Ice+HMM	76.04	93.49	92.31

Table 1: Average tagging accuracy (%) using the original IFD corpus

tagging accuracy on Icelandic text[3] was obtained by Dredze and Wallenberg (2008) by applying a bidirectional sequence classification method (Shen et al., 2007). In this method, the classifier assigns the potential PoS tags (hypothesis) to a subsequence of words (called a span) based on features selected by the developer of the classifier. In each round, the highest scoring hypothesis is selected and the guessed tags are assigned to the span. Unassigned words are then reevaluated using the new information. Words either to the left or to the right of the previous assigned span can be chosen next – hence the name bidirectional classification.

Drezde and Wallenberg used the fact that data-driven methods are good at assigning correct word classes (the first letter of a tag in the IFD tagset) to words. Therefore, they divided the learning phase into separate learning problems. First, they constructed a word class (WC) tagger which classifies a word according to one of eleven word classes. Then the tagger only evaluates tags that are consistent with that class. This dramatically reduces the number of tags considered at each step during the bidirectional tagging algorithm. Secondly, noting that most tagging errors are due to errors in case, they constructed a case tagger (CT) that retags case on nouns, adjectives and pronouns, given the predicted tags from the WC tagger. Their combination of a bidirectional tagger, a WC tagger and a CT tagger (BI+WC+CT) resulted in an accuracy of 92.06% (see Table 1). The tenth data file was used for the development of the features used and the average accuracy is thus based on testing using the first nine test corpora.

[3]Note that in our review of previous tagging approaches we exclude results based on combination of taggers using voting schemes. For that part, the interested reader is referred to (Helgadóttir, 2005; Loftsson, 2006).

2.3 Our tagging method

The motivation behind our method is twofold. First, when only considering the word class we noted that the tagging accuracy of IceTagger (97.61%) is significantly lower than the corresponding tagging accuracy of an HMM tagger like TnT (98.14%). This may be due to the limited amount of local rules in IceTagger. Secondly, as discussed above, determining the word class first can simplify the remainder of the disambiguation task.

Thus, we borrow the word class tagger idea from Drezde and Wallenberg and apply it by developing a new tagger based on IceTagger and TriTagger. The main idea is to use TriTagger (the HMM tagger; see Section 2.2) for choosing the word class and then use IceTagger to perform tagging which is consistent with the chosen class, but based on the whole tag string. We are not aware of similar work, i.e. in which a data-driven tagger is integrated into a linguistic rule-based tagger in the form of a pre-processing step. More specifically, the following steps are carried out for each input sentence:

1. IceTagger starts by looking up the tag profile for known tokens in the dictionary and uses IceMorphy for filling in tag profile gaps and generating the tag profile for unknown tokens.

2. For each token and its tag profile, a copy is made. A version of TriTagger, trained on the complete tag strings, disambiguates the copied tokens by using the standard HMM method of finding the tag sequence that maximises the product of contextual probabilities and lexical probabilities (Brants, 2000). The result is one proposed tag for each token.

3. For each token, the proposed tag t from TriTagger is used to eliminate tags from the corresponding token in IceTagger that are not consistent with the word class of tag t.

4. Finally, the standard version of IceTagger is run using (possibly) a reduced tag profile for each token.

We refer to this new tagger as the HMM+Ice tagger. It is an integrated tagger and, consequently, runs like a single tagger. Note that our method

should be feasible for other morphologically complex languages for which an HMM tagger and a linguistic rule-based tagger already exist.

The tagging accuracy of HMM+Ice is 92.19% (see Table 1), which amounts to about a 7.1% and 1.6% error reduction rate compared to IceTagger and the BI+WC+CT tagger, respectively. As expected, the number of tags needed to be considered by IceTagger drops significantly when using TriTagger for initial disambiguation. The ambiguity rate (total number of tags divided by total number of tokens) for known ambiguous tokens in the standard version of IceTagger is 2.77. In the HMM+Ice tagger the corresponding number is 2.40, which amounts to a 13.4% drop in ambiguity rate.

Note that the HMM+Ice tagger applies the HMM before IceTagger runs, but, conversely, the Ice+HMM tagger (described in Section 2.2), applies the HMM after IceTagger. By combining these two methods, we obtain a more accurate tagger which runs in the following manner. It starts by following steps 1-3 described above. Then, in step 4, it runs the Ice+HMM tagger, instead of only running IceTagger. We refer to this method as the HMM+Ice+HMM tagger. The tagging accuracy of the HMM+Ice+HMM tagger is 92.31%, which amounts to about a 8.6% and 3.2% error reduction rate compared to IceTagger and the BI+WC+CT tagger, respectively. The difference between the HMM+Ice tagger and the HMM+Ice+HMM tagger is that the former chooses the most frequent tag for words which are still ambiguous after the application of IceTagger, whereas the latter applies the HMM model again to disambiguate those words.

Table 1 summarises the accuracy of all the PoS taggers discussed above (using the average from the first nine test corpora). The table shows that our HMM+Ice+HMM tagger outperforms the BI+WC+CT tagger because of higher accuracy for unknown words, but the accuracy obtained by the BI+WC+CT tagger for known words is superior by 0.21 percentage points. We hypothesised that this could partly be explained by the following. IceTagger uses a dictionary generated from a training corpus, consisting of each word encountered along with the tag profile for each word. Thus, the tag profile for a word w only contains tags that were found in a training corpus for w, in addition to missing tags generated by the tag profile gap filling mechanism of IceMorphy (discussed in Section 2.2). In contrast, a tagger based on the bidirectional classification method evaluates all possible tags in the tagset to select the top tag for a word. Consequently, during tagging it does not look up the tag profile in a dictionary for a given word. This means, for example, that the BI+WC+CT tagger is able to assign a noun tag to a word w even though w is never tagged as a noun in the training corpus.

To verify this hypothesis, we analysed the output generated by the BI+WC+CT tagger. For each test corpus, it assigns, on average, 559 tags that are not included in the corresponding dictionary (filled with tags from IceMorphy) derived during training. The average size of a test corpus is 59,081 tokens and therefore the "out-of-dictionary" tags are 1.02% of the total tag assignments. However, only 160 of the 559 tags are actually correct tag assignments. Nevertheless, 0.29% of the tagging accuracy for known words $(160/59,081)$ can be attributed to these 160 correct tags. This supports our hypothesis, because the tagging accuracy of the BI+WC+CT tagger for known words would be a little less than the corresponding accuracy of the HMM+Ice+HMM tagger if the former tagger could not use out-of-dictionary tag assignments.

It is important to note that tagging time is very important in practical applications. According to Dredze and Wallenberg (2008b), the WC tagger alone processes 179 tokens per second (processing time for the CT tagger is not given). In comparison, our HMM+Ice+HMM tagger processes about 2350 tokens per second[4] (running on a Dell Precision M4300 2 Duo CPU, 2.20 GHz).

2.4 Using the corrected corpus

Loftsson (2009) has produced a version of the IFD corpus in which a number of tagging errors (1,334 in total) have been corrected. His reevaluation of the taggers TnT, TnT*, IceTagger and Ice+HMM showed a significant improvement in tagging accuracy compared to using the original corpus. We repeat his tagging results in Table 2, along with the results for the BI+WC+CT tagger and our HMM+Ice and HMM+Ice+HMM taggers. For the taggers TnT, TnT*, Ice+HMM, HMM+Ice, and HMM+Ice+HMM the results are presented after

[4]The standard version of IceTagger (without HMM integration) processes more than 6600 tokens per second.

Tagger	Unknown	Known	All
TnT	71.97	92.06	90.68
TnT*	73.10	92.85	91.50
IceTagger	75.36	92.95	91.76
Ice+HMM	75.70	93.20	92.01
BI+WC+CT	69.80	93.85	92.21
HMM+Ice	76.17	93.59	92.40
HMM+Ice+HMM	76.13	93.70	92.51

Table 2: Average tagging accuracy (%) using the corrected IFD corpus

retraining on the corrected corpus. IceTagger does not need retraining because it does not derive a language model from a training corpus. Note that since we only had access to the output generated by the BI+WC+CT (not the tagger itself), we were not able to retrain that tagger. Thus, presumably, the accuracy of the BI+WC+CT in Table 2 is somewhat underestimated (and the same applies for the accuracy numbers which we present in Section 3).

Our HMM+Ice+HMM tagger achieves an accuracy of 92.51% for all words when testing using the corrected corpus. We suggest that researchers use the corrected version of the IFD corpus as a gold standard in future work[5].

3 Tagset Reduction

There are two main methods used when reducing tagsets in the context of PoS tagging – we refer to them as *tagset change* and *tagset mapping*. In the former method, the tagset is simplified and the training corpus updated to reflect the change in the tagset. Taggers are then retrained on the updated corpus and during testing the taggers thus produce tags according to the simplified tagset.

In the latter method, tagset mapping, the only change needed is in the testing (evaluation) part. When comparing a particular tag t_1 in the output of a tagger to a tag t_2 in the gold standard, the tags t_1 and t_2 are mapped to new simplified tags m_1 and m_2, respectively. Then, the tags m_1 and m_2 are compared instead of t_1 and t_2. When using the tagset mapping method, the tagset used by the tagger is called the *internal tagset* and the tagset used for evaluation called the *external tagset* (Brants, 1997). The motivation for using the tagset mapping method is that often the internal (larger) tagset encodes information that can help disambiguate words in context.

Char #	Category/ Feature	Symbol – signification
1	Word class	**n**–noun
2	Gender	**k**–masculine, **v**–feminine, **h**–neuter, **x**–unspecified
3	Number	**e**–singular, **f**–plural
4	Case	**n**–nominative, **o**–accusative, **þ**–dative, **e**–genitive
5	Article	**g**–with suffixed article
6	Proper noun	**m**–person, **ö**–place, **s**–other proper name

Table 3: The signification of the tags for nouns

The size of the current IFD tagset is a direct consequence of the morphological complexity of Icelandic and most of the distinctions that the tagset makes reflect morphosyntactic features which must be marked for the tagging to be useful. However, we believe that it is possible to make certain reductions which do not affect the effectiveness of the tagset in practical applications. In this section, we thus propose an external tagset, which can be used as an alternative to the original (internal) one used hitherto[6]. Our work is inspired by the tag simplification experiments by Helgadóttir (2005). We present four simplifications to the original tagset, implemented as tagset mappings[7], and evaluate taggers based on these different versions. In all cases, the tagging accuracy gained is presented relative to the accuracy obtained using the original tagset.

As discussed in Section 2.1, the current IFD tagset is large and makes fine-grained distinctions. Moreover, the tagset reflects distinctions which may be impossible (or at least very difficult) to resolve at the level of PoS tagging.

The most obvious example is the type of proper nouns, denoted by the sixth letter in the tags for nouns (see Table 3). This information is not of syntactic nature and to our knowledge this is not part of tagsets for other languages. Therefore, a separate natural language processing module, a

[5]The original IFD corpus and its corrected version is available for research purposes at The Árni Magnússon Institute for Icelandic Studies.

[6]We use linguistic knowledge when reducing the tagset. Another way, for example, would be to look at the precision and recall rates for each tag to motivate the tagset reduction.

[7]For the TnT tagger we indeed experimented with the tagset change method, but the tagging accuracy was either equivalent or substantially lower than using tagset mapping.

Tagger	Original tagset	Ignoring type of proper nouns	c=ct	Ignoring type of pronouns	Prep.= adverbs	All four mappings
TnT*	91.50	91.56	91.61	91.61	92.51	92.80
IceTagger	91.76	91.83	91.85	91.88	92.61	92.90
BI+WC+CT	92.21	92.27	92.27	92.31	92.89	93.12
HMM+Ice+HMM	92.51	92.57	92.62	92.62	93.35	93.63

Table 4: Average tagging accuracy (%) for all words using external tagsets

Named Entity Recogniser, is usually responsible for determining the type of proper nouns. Consequently, for the first simplification of the tagset, we remove the type information for proper nouns. During testing we thus perform a mapping which ignores the distinction made in the last letter of proper noun tags. This reduces possible proper noun tags from 144, in the internal tagset, to 48, in the external tagset. As can be seen by comparing columns 2 and 3 in Table 4, this increases the accuracy of the taggers by 0.06-0.07 percentage points.

In the IFD tagset, the tag "c" denotes a conjunction and "ct" a relativizer (a conjunction used to indicate a relative clause). The typical relativizer, "sem" ('that') can also be a comparative conjunction and it is often difficult, even for experienced linguists, to determine which function it has in a given sentence. Furthermore, this distinction must be based on syntactic and contextual information which is not available to a PoS tagger. The second simplification thus consists of mapping the "ct" tag to "c", i.e. removing the "ct" tag from the external tagset. This increases the tagging accuracy of the taggers by 0.06-0.11 percentage points (see column four of Table 4).

Tags starting with the letter "f" denote pronouns in the IFD tagset. The second letter, one of "[abeopst]" specifies type information, i.e. demonstrative, reflexive, possessive, indefinite, personal, interrogative or relative. In most cases, ignoring this type information does not lead to any loss of information, since most of the pronouns can only belong to one class anyway. In the few cases where a pronominal word form is ambiguous between pronoun classes, the distinction is either syntactically based or based on contextual information which is arguably beyond the realm of a PoS tagger. In the third simplification, we therefore perform a mapping which ignores the type of the pronoun. This reduces possible pronoun tags from 184, in the internal tagset, to 40, in the external tagset, and increases the tagging accuracy of the taggers by 0.10-0.12 percentage points (see column five of Table 4).

The three simplifications described above do not, however, help in reducing the most common tagging mistakes. Table 5 shows that out of the top six errors made by our HMM+Ice+HMM tagger, five are related to prepositions (tags "ao","aþ") and adverbs (tag "aa"), i.e. tagging words as prepositions governing the wrong case or tagging words as prepositions instead of adverbs, or vice versa. Notice that these tags are outsiders anyway, since they do not reflect any morphological distinctions in the words they are attached to, but only indicate the effect (case government) that these words have on their complements. However, the case is of course marked on the complement itself, so the case tag on the preposition/adverb is completely redundant but leads to a number of tagging errors. To illustrate, consider the phrase "í bæinn" ('to town') tagged as "ao nkeog". The second letter of the preposition tag "ao" denotes the case governed by the preposition and the fourth letter of the complement (noun) tag "nkeog" denotes the corresponding accusative case inflection. Only on the noun, therefore, does "o" signify morphologically marked grammatical information.

In the last simplification of the tagset, we therefore map the following seven tags "ao", "aþ", "ae", "aþm", "aþe", "aam", "aae" (preposition tags and adverbs in comparative and superlative form) to the adverb tag "aa", effectively disregarding the difference between prepositions and adverbs and reducing the external tagset by 7 tags. This increases the tagging accuracy by 0.68-1.01 percentage points (see column six of Table 4).

Finally, the last column of Table 4 shows the accuracy of the taggers when applying all the four tagset mappings at once. The overall tagging accuracy gain for the taggers is 0.91-1.30 percentage points when compared to using the original tagset. The size of the external tagset using all four map-

No.	Proposed tag > correct tag	Error rate	Cumulative rate
1.	aþ>ao	3.09%	3.09%
2.	aa>ao	1.69%	4.78%
3.	ao>aþ	1.68%	6.47%
4.	nveþ>nveo	1.66%	8.13%
5.	ao>aa	1.56%	9.70%
6.	aa>aþ	1.43%	11.13%
7.	nhen>nheo	1.00%	12.13%
8.	sfg3fn>sng	0.99%	13.12%
9.	nveo>nveþ	0.97%	14.09%
10.	nkeþ>nkeo	0.88%	14.97%

Table 5: The top ten most frequent errors made by the HMM+Ice+HMM tagger

pings is about 450 tags and our HMM+Ice+HMM tagger achieves an accuracy of 93.63% using this tagset.

4 Discussion and Future Work

Comparison in tagging accuracy between languages is difficult because of different levels of morphological complexity, different tagsets, different corpora, etc. However, for the sake of making one comparison to a related language, let us consider Swedish. An accuracy of about 95% was obtained for Swedish by a standard version of the TnT tagger, using a tagset consisting of 139 tags, a training corpus of 500k tokens, and an unknown word ratio of 8.1% (Megyesi, 2001). This can be compared to the 93.63% accuracy of our HMM+Ice+HMM tagger, obtained using a tagset of about 450 tags. According to this, there is still quite a large gap in tagging accuracy between the languages. Partly, it may be explained by the difference in tagset sizes, but, on the other hand, one would also expect that the tagging accuracy of Swedish could be increased by using a more sophisticated tagger than the standard version of TnT (e.g. a tagger similar to our HMM+Ice+HMM tagger). Due to the fact that Icelandic has considerably more complex inflectional morphology than Swedish, one may conclude that it will be difficult to achieve tagging accuracy numbers for Icelandic comparable to Swedish. Nevertheless, in order to further increase the tagging accuracy of Icelandic text, we foresee at least four possibilities.

First, one might try to minimise the ratio of unknown words. As mentioned in Section 2.2, the average unknown word ratio using the standard data-split is 6.8%. Since the tagging accuracy of all the taggers for unknown words is only about 70-76% (see Table 1), it is important to minimise this ratio (the experiment by Helgadóttir (2005) using "a backup lexicon" showed good results). One possibility is to use the comprehensive Morphological Database of Icelandic Inflections (MDII) (Bjarnadóttir, 2005) for this purpose. The MDII contains about 270,000 entries, over 5.8 million word forms. The database does not, however, contain any frequency information. The data from the MDII could be used to extend the dictionaries used by the taggers (for the HMM taggers a uniform distribution could be assumed in the tag profile for a word), which should result in a dramatic drop in the unknown word ratio and, presumably, an increased tagging accuracy for all words.

Second, one might consider implementing a tagger (and a parser) using the framework of Constraint Grammar (CG) (Karlsson et al., 1995), which has been applied to several languages. The main advantage of CG systems is high accuracy (Samuelsson and Voutilainen, 1997), but the main disadvantage is the labour-intensive development – for example, the Norwegian CG project took seven man labour years (Hagen et al., 2000). Regardless, we think that a CG system should be developed for Icelandic. Note that the existence of the MDII could reduce the development time, i.e. with regard to the morphological analyser which is a crucial part of a CG system.

Third, one could explore further combining data-driven and linguistic rule-based methods. For example, since the accuracy of the BI+WC+CT tagger for unknown words is the least of all the taggers (see Table 1), it can presumably be increased by integrating a morphological component like IceMorphy.

Finally, as pointed out by Dredze and Wallenberg (2008), a considerable proportion of the errors are mistakes in case assignments of verb subjects and objects (rows no. 4, 9, and 10 of Table 5 illustrate the latter). Finding ways to minimise these errors is therefore part of the challenge ahead.

5 Summary

In this paper, we first presented a new state-of-the-art tagger for Icelandic, HMM+Ice+HMM, by

integrating an HMM tagger into a linguistic rule-based tagger in a novel way. Our method should be feasible for other morphologically complex languages for which an HMM tagger and a linguistic rule-based tagger already exist. Evaluation shows that our HMM+Ice+HMM tagger obtains an accuracy of 92.31% using the standard test set derived from the IFD corpus. Furthermore, the accuracy increases to 92.51% using a corrected version of the corpus.

Second, we proposed an external tagset by removing information from the internal tagset which reflects distinctions that are not morphologically based. The accuracy of HMM+Ice+HMM increases to 93.63% using the external tagset.

Finally, we discussed the results and provided directions for future work.

Acknowledgments

The work in this paper was supported by the Icelandic Research Fund, grant 070025023. We thank Mark Drezde and Joel Wallenberg for providing access to the output of their tagger.

References

Kristín Bjarnadóttir. 2005. Modern Icelandic Inflections. In H. Holmboe, editor, *Nordisk Sprogteknologi 2005*, pages 49–50. Museum Tusculanums Forlag, Copenhagen.

Thorsten Brants. 1997. Internal and External Tagsets in Part-of-Speech Tagging. In *Proceedings of Eurospeech 97*, Rhodes, Greece.

Thorsten Brants. 2000. TnT: A statistical part-of-speech tagger. In *Proceedings of the* 6^{th} *Conference on Applied Natural Language Processing*, Seattle, WA, USA.

Stefán Briem. 1989. Automatisk morfologisk analyse af islandsk tekst. In *Papers from the Seventh Scandinavian Conference of Computational Linguistics*, Reykjavik, Iceland.

Mark Dredze and Joel Wallenberg. 2008. Icelandic Data Driven Part of Speech Tagging. In *Proceedings of the* 46^{th} *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, OH, USA.

Mark Dredze and Joel Wallenberg. 2008b. Further Results and Analysis of Icelandic Part of Speech Tagging. Technical report, Department of Computer and Information Science, University of Pennsylvania.

Kristin Hagen, Janne B. Johannessen, and Anders Nøklestad. 2000. A Constraint-Based Tagger for Norwegian. In C.-E. Lindberg and S.-N. Lund, editors, 17^{th} *Scandinavian Conference of Linguistics. Odense Working Papers in Language and Communication*, volume 19, pages 31–48. University of Southern Denmark, Odense.

Jan Hajič, Pavel Krbec, Karel Oliva, Pavel Květoň, and Vladimír 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.

Sigrún Helgadóttir. 2005. Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In H. Holmboe, editor, *Nordisk Sprogteknologi 2004*, pages 257–265. Museum Tusculanums Forlag, Copenhagen.

Sigrún Helgadóttir. 2007. Mörkun íslensks texta. [Tagging Icelandic text.]. *Orð og tunga*, 9:75–107.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

Hrafn Loftsson. 2006. Tagging Icelandic text: an experiment with integrations and combinations of taggers. *Language Resources and Evaluation*, 40(2):175–181.

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* 12^{th} *Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Greece.

Beáta Megyesi. 2001. Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, PA, USA.

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

Christer Samuelsson and Atro Voutilainen. 1997. Comparing a Linguistic and a Stochastic tagger. In *Proceedings of the* 8^{th} *Conference of the European Chapter of the ACL (EACL 1997)*, Madrid, Spain.

Libin Shen, Giorgio Satta, and Aravind K. Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proceedings of the* 45^{th} *Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206