Tagging the Past: Experiments using the Saga Corpus

Hrafn Loftsson

School of Computer Science, Reykjavik University, Iceland

hrafn@ru.is

ABSTRACT

There is an increasing interest in the NLP community in developing tools for annotating historical data, for example, to facilitate research in the field of corpus linguistics. In this work, we experiment with several PoS taggers using a sub-corpus of the Icelandic Saga Corpus. This is carried out in three main steps. First, we evaluate taggers, which were trained on Modern Icelandic, when tagging Old Icelandic. Second, we semi-automatically correct errors in the training corpus using a bootstrapping method. Finally, we evaluate the taggers on the corrected training corpus. The best performing single tagger is Stagger, a tagger based on the averaged perceptron algorithm, obtaining an accuracy of 91.76%. By combining the output of three taggers, using a simple voting scheme, the accuracy increases to 92.32%.

KEYWORDS: Historical Data, Icelandic Saga Corpus, Part-of-Speech Tagging.

1 Introduction

Most Natural Language Processing (NLP) tools, for various languages, have been developed for processing and analyzing modern texts, as opposed to historical (cultural heritage) texts. This is due to the abundance of modern texts in digital form, and, often, the lack of availability of historical texts. Another reason is that when the first NLP tools are developed for a given language, the emphasis is usually on producing tools for processing and analyzing the modern language.

More and more historical texts are now gradually becoming available in digital form. Consequently, there is an increasing interest in the NLP community in developing annotated historical resources, and tools for analyzing historical texts.

Examples of recent annotated resources are: Penn Parsed Corpora of Historical English (Kroch and Taylor, 2000), Icelandic Parsed Historical Corpus (Wallenberg et al., 2011), and Corpus of Early Modern German (Scheible et al., 2011a). These three example resources are all tagged with Part-of-Speech (PoS), while the first two are also syntactically annotated.

Examples of recent experiments with NLP tools for historical texts are: an identification of verb constructions in Swedish (Pettersson et al., 2012), a study of the performance of basic NLP tools for Italian (Pennacchiotti and Zanzotto, 2008), an adaptation of existing NLP tools for Spanish (Sánchez-Marco et al., 2011), and an evaluation of an "off-the-shelf" PoS tagger for German (Scheible et al., 2011b).

Recently, Rögnvaldsson and Helgadóttir (2011) developed the first tagger for Old Icelandic. They bootstrapped a Hidden Markov Model (HMM) tagger by creating a tagged sub-corpus (95,000 tokens) from the Saga Corpus (Old Icelandic Sagas).[1] Hereafter, we refer to the tagged sub-corpus as *Saga-Gold*.

The aim of our work is to complement the work of Rögnvaldsson and Helgadóttir (2011). The overall goal is similar, i.e. developing a high accuracy tagger for Old Icelandic texts. We carry this out in the following three main steps. First, we evaluate several PoS taggers, which were trained on Modern Icelandic, on Saga-Gold produced by Rögnvaldsson and Helgadóttir (2011). Second, we semi-automatically correct tagging errors in Saga-Gold, with a bootstrapping method using the same taggers.[2] Finally, we perform 10-fold cross-validation on the corrected corpus, again using the same taggers and a combination method. All the PoS taggers and corpora used in our work are freely available and open-source.

The best performing single tagger is Stagger (Östling, 2012), a tagger based on the averaged perceptron algorithm, obtaining an accuracy of 91.76%. By combining the output of three taggers using a simple voting scheme, the accuracy increases to 92.32%. We intend to use our combination method to annotate the whole of the Saga Corpus.

The problem of domain adaptation has received increasing attention in recent years. The problem arises in a variety of NLP applications where the distribution of the training data differs in some way from that of the test data. Our work, as well as, for example, (Rögnvaldsson and Helgadóttir, 2011; Sánchez-Marco et al., 2011), essentially deals with the issue of adapting a PoS tagging model based on a modern language to a different domain, an older language.

[1] Available for download at `http://www.malfong.is`

[2] Although Saga-Gold is a gold corpus, we found that it contained many errors that needed to be corrected. The corrected training corpus will be made available at `http://www.malfong.is`

Several other experiments with domain adaptation within the field of PoS tagging have been described in the literature, e.g. adapting a model based on finanical data to biomedical data (Blitzer et al., 2006) and to dialogues (Kübler and Baucom, 2011), respectively.

This paper is structured as follows. In Section 2, we describe the individual PoS taggers used in our experiments. We discuss previous work in tagging both Modern and Old Icelandic in Section 3. Our development and evaluation work is described in Section 4. Error analysis is performed in Section 5, and, finally, we draw conclusions and propose future work in Section 6.

2 PoS Taggers Used

We use four different PoS taggers for tagging Old Icelandic texts in Section 4: Stagger, TriTagger, IceTagger and HMM+Ice+HMM. These taggers are freely available, open-source, and, importantly, fast during training and testing.[3]

Stagger (Östling, 2012) is an implementation of the averaged perceptron algorithm by Collins (2002). Stagger uses feature-vector representations commonly used in maximum entropy taggers (Ratnaparkhi, 1996; Toutanova et al., 2003). The feature vectors represent "histories", the context in which a tagging decision is made. For every feature, the perceptron algorithm calculates integer weight coefficients, which are updated for every training sentence. After the final update, these coefficients are stored with the corresponding features. When tagging new texts, the perceptron algorithm sums up all the coefficients of the features in a given context and returns the highest scoring sequence of tags for an input sentence.

TriTagger is a HMM tagger, a re-implementation of the well-known TnT tagger (Brants, 2000). TriTagger uses a trigram model to find the sequence of tags for words in a sentence, which maximizes:

$$P(t_1)P(t_2|t_1)\prod_{i=3}^{n}P(t_i|t_{i-2},t_{i-1})\prod_{i=1}^{n}P(w_i|t_i) \quad (1)$$

In equation 1, w_i denotes word i in a sentence of length n ($1 \leq i \leq n$) and t_i denotes the tag for w_i. The probabilities are derived using maximum likelihood estimation based on the frequencies of tags found during training.

IceTagger (Loftsson, 2008) is a linguistic rule-based tagger. It is reductionistic in nature, i.e. it removes inappropriate tags from the set of possible tags for a specific word in a given context. IceTagger first applies local rules for initial disambiguation and then uses a set of heuristics for further disambiguation. 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 given word.

HMM+Ice+HMM (Loftsson et al., 2009) is a hybrid tagger, comprising both IceTagger and TriTagger. It works as follows. First, TriTagger (the HMM) performs initial disambiguation only with regard to the word class. Then, the rules of IceTagger are run. Finally, the HMM disambiguates words that IceTagger is not able to fully disambiguate.

In addition to these four taggers, we use CombiTagger[4] (Henrich et al., 2009), a system for developing combined taggers. Tagger combination methods are a means of correcting for the biases of individual taggers, and they are especially suitable when tagging a corpus, i.e.

[3] TriTagger, IceTagger and HMM+Ice+HMM are all part of the IceNLP toolkit, available for download at `http://icenlp.sourceforge.net`. Stagger is available for download at `http://www.ling.su.se/stagger`

[4] CombiTagger is open-source – available for download at `http://combitagger.sourceforge.net`

when effectiveness (accuracy) is more important than efficiency (running time). It has been shown that combining taggers will often result in a higher tagging accuracy than is achieved by individual taggers (Brill and Wu, 1998; van Halteren et al., 2001; Loftsson, 2006). The reason is that different taggers tend to produce different errors, and the differences can often be exploited to yield better results.

3 Previous Work on Tagging Icelandic

The Icelandic language is one of the Nordic languages which comprise the North-Germanic branch (Danish, Swedish, Norwegian, Icelandic, Faroese) of the Germanic language tree. Linguistically, Icelandic is most closely related to Faroese and the dialects of Western Norway.

The Icelandic language is morphologically rich, mainly due to inflectional complexity. From a syntactic point of view, Icelandic has a basic subject-verb-object (SVO) word order, but, in fact, the word order is relatively flexible, because morphological endings carry a substantial amount of syntactic information.

The main change in Modern Icelandic since Old Icelandic is in the phonological system.

> On the other hand, the inflectional system and the morphology has in all relevant respects remained unchanged from the earliest texts up to the present, although a number of nouns have shifted inflectional class, a few strong verbs have become weak, one inflectional class of nouns has been lost, and the dual in personal and possessive pronoun has disappeared. The syntax is also basically the same, although a number of changes have occurred. The changes mainly involve word order, especially within the verb phrase, and the development of new modal constructions. (Rögnvaldsson et al., 2012)

In this section, we describe previous work on tagging Iceland texts – first Modern Icelandic, and then Old Icelandic.

3.1 Tagging Modern Icelandic

A few years ago, no PoS tagger existed for tagging Modern Icelandic. Now, however, various PoS taggers have been developed, i.e. data-driven taggers (Helgadóttir, 2005; Dredze and Wallenberg, 2008; Loftsson et al., 2009) and a rule-based tagger (Loftsson, 2008). All these taggers have been trained and developed using the Icelandic Frequency Dictionary[5] (IFD) (Pind et al., 1991), a corpus of about 590,000 tokens of Modern Icelandic. The tagset used in the compilation of the IFD has become the standard tagset for tagging Icelandic. It contains about 700 possible tags, of which 639 appear in the IFD. Thus, the tagset mirrors the morphological complexity of the language.

The PoS tags in the IFD are character strings where each character 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 form “markari” ‘tagger’. The corresponding tag is *nken*, denoting noun (*n*), masculine (*k*), singular (*e*), and nominative (*n*) case.

[5]Available for download at `http://www.malfong.is`

Tagger	Unknown	Known	All
TriTagger	72.98	92.18	**90.86**
Stagger	63.77	93.02	**91.04**
IceTagger	77.02	93.07	**91.98**
HMM+Ice+HMM	77.47	93.84	**92.73**

Table 1: Average tagging accuracy (%) of four taggers when tagging the IFD corpus (Modern Icelandic) using 10-fold cross-validation. Average unknown word rate (UWR) in testing is 6.8%.

In recent work on tagging Icelandic, the tagset has been reduced, by removing named-entity classification for proper nouns and labeling all number constants with a single tag – resulting in 565 tags appearing in the changed version of the IFD (Loftsson et al., 2011). Nevertheless, the PoS (morphosyntactic) tagging of modern Icelandic texts is a challenging task. The reason is, for example, that the tagset is large in relation to the size of the available training corpus, and the tagset makes very fine distinctions.

Table 1 shows the accuracy of the four taggers, described in Section 2, when tagging the IFD corpus (565 tags) using 10-fold cross-validation. The accuracy figures for TriTagger, IceTagger, and HMM+Ice+HMM are copied from (Loftsson et al., 2011), whereas we trained and tested Stagger ourselves.[6] All the four taggers were run using default options. A PoS tag predicted by a tagger is correct only if it agrees in the whole tag string with the gold (correct) tag.

It is noteworthy that Stagger's accuracy for known words is significantly higher than the corresponding figure for the other purely data-driven tagger, i.e. TriTagger. This may be explained by the fact that the HMM model used by TriTagger only conditions on the (already assigned) tags to the left of the current word w when predicting the tag for w, whereas a model based on the averaged perceptron algorithm (Stagger) can, in addition to the left hand tag features, use word features to the right of w.

Note, however, that the accuracy of Stagger for unknown words is much lower than for the other taggers. TriTagger's handling of unknown words is based on an effective suffix analysis algorithm proposed by Brants (2000). IceTagger (and thus HMM+Ice+HMM) uses a morphologically-based guesser, IceMorphy (Loftsson, 2008), for providing the set of possible tags for an unknown word.[7]

3.2 Tagging Old Icelandic

The Saga Corpus contains old Icelandic narrative texts in modern Icelandic spelling, assumed to be written in the 13^{th} and 14^{th} centuries. It contains text from four different categories of stories: "Íslendingasögur" (Family Sagas), "Sturlunga" (Sturlunga Saga), "Heimskringla" (Sagas of the Kings of Norway), and "Landnámabók" (The Book of Settlement). In total, the Saga Corpus contains about 1,650,000 tokens.

Using the TnT tagger, Rögnvaldsson and Helgadóttir (2011) have semi-automatically annotated a third (about 95,000 tokens) of the 283,000 tokens from Sturlunga Saga. This annotated

[6] When training Stagger on the IFD, we used 12 iterations.

[7] The newest experiments using Stagger for tagging Modern Icelandic show that Stagger indeed obtains state-of-the-art tagging accuracy when enriched with language-dependent linguistic features and given access to IceMorphy (Loftsson and Östling, 2013).

sub-corpus is referred to as Saga-Gold and was used by Rögnvaldsson and Helgadóttir (2011) as a training corpus for developing a tagger for Old Icelandic. For testing, they used 1,000 tokens from each of the four different texts in the Saga Corpus, i.e. 4,000 tokens in total.

Three tagging experiments using TnT were performed by Rögnvaldsson and Helgadóttir (2011). First, training the tagger on the IFD, i.e. on modern texts only, resulted in an accuracy of 88.0%. Second, training on Saga-Gold, i.e. on old texts only, resulted in an accuracy of 91.7%. Finally, by training on the union of the IFD and Saga-Gold, the accuracy increased to 92.7%.

Note that the tagging accuracy increases substantially when training on Saga-Gold, a small training corpus, compared to when only training on the IFD, a corpus whose size is more than 6 times larger. There are two reasons for this, as explained by Rögnvaldsson and Helgadóttir (2011). First, many of the tagging errors made in the first experiment are due to constructions found only in Old Icelandic, and by training on Saga-Gold the tagger learns the correct tagging of these constructions. Second, the unknown word rate (UWR) was much lower in the second experiment (9,6%) than in the first experiment (14.6%), reflecting the fact that many words in Old Icelandic do not appear in Modern Icelandic.

The accuracy of 92.7%, obtained by training TnT on the union of the IFD and Saga-Gold, is high compared with the accuracy of 90.4% obtained by the same tagger when tested on Modern Icelandic only (Helgadóttir, 2005). Texts from the Saga Corpus are much less diversified and simpler than the texts in the IFD corpus, and therefore, in principle, one should be able to achieve higher accuracy on Old Icelandic texts compared to Modern Icelandic. It has to be noted, however, that the test data of only 4,000 tokens, used for evaluating TnT on Old Icelandic, may be too small for obtaining reliable tagging figures. In our experiments, we more than double the size of the test data (see Section 4.4).

4 Development and Evaluation

Our main goal was to the develop a high accuracy tagger for Old Icelandic. We carried this out in the following three main steps. First, we evaluated four PoS taggers trained on Modern Icelandic on Saga-Gold (see Section 4.2). Second, we semi-automatically corrected tagging errors in Saga-Gold, with a bootstrapping method using the same four taggers (see Section 4.3). Finally, we performed 10-fold cross-validation on the corrected corpus, again using the same taggers and a combination method (see Section 4.4).

Our work complements the work of Rögnvaldsson and Helgadóttir (2011), described in Section 3.2. Our work is different from (or extends) the previous work in that i) we correct errors in the Saga-Gold corpus; ii) we evaluate many taggers, as opposed to a single one; iii) we perform testing using cross-validation, as opposed to testing on a single (small) file; and iv) we present results of error analysis.

4.1 The training corpora

We used two training corpora: the IFD corpus, described in Section 3.1, and the Saga-Gold corpus, described in Section 3.2. For both corpora, we used a version in which the tagset has been reduced as explained in Section 3.1. The number of unique tags appearing in Saga-Gold is 459, whereas 565 unique tags appear in the IFD.

In the IFD corpus, the first letter of the first word in each sentence is a lower case letter, except for proper nouns. This is not the case in Saga-Gold. For the sake of consistency, we changed

the first letter of the first word in each sentence to an upper case letter in the IFD corpus. Henceforth, when referring to the IFD corpus, we mean this changed version.

4.2 Evaluation of taggers trained on Modern Icelandic

We started our evaluation work by testing PoS taggers, that had been trained or developed for tagging Modern Icelandic (see Section 3.1), on Old Icelandic (Saga-Gold). We trained TriTagger and Stagger on the IFD. IceTagger comes "off-the-shelf" with dictionaries derived from the IFD, and thus does not need training. The HMM+Ice+HMM tagger uses the trained model generated by TriTagger.

The results of the evaluation are shown in columns 2-4 in Table 2. The tagging accuracy is much lower than shown in Table 1. The ordering of the taggers, from lowest accuracy to highest, is also different. When tagging Modern Icelandic, TriTagger obtained the lowest accuracy of the four taggers. In contrast, when tagging the Saga-Gold corpus, it obtains the highest accuracy for all words. Nevertheless, its accuracy drops by 4.17 percentage points.

	Original Saga-Gold			Corrected Saga-Gold			
Tagger	Unknown	Known	All	Unknown	Known	All	Increase
IceTagger	63.99	85.88	**83.55**	65.47	87.03	**84.74**	1.19
Stagger	56.58	87.44	**84.15**	57.02	88.27	**84.94**	0.79
HMM+Ice+HMM	63.71	88.07	**85.48**	65.20	89.17	**86.62**	1.14
TriTagger	65.56	89.29	**86.69**	65.11	89.55	**86.87**	0.18

Table 2: Average tagging accuracy (%) of four taggers, trained on the IFD corpus (Modern Icelandic), when tagging the original Saga-Gold corpus (Old Icelandic) and the corrected version. Average UWR in testing is 10.7%.

IceTagger, which obtained the second highest accuracy for Modern Icelandic, performs badly when tagging Saga-Gold. Its accuracy drops by 8.43 percentage points. This was to be expected, because the hand-crafted rules of IceTagger have been developed to tag modern texts. This poor performance of a rule-based tagger, developed for contemporary texts, when tagging historical text is consistent with the results of (Pennacchiotti and Zanzotto, 2008) when tagging Italian historical texts.

The HMM+Ice+HMM tagger benefits from using the HMM generated by TriTagger and therefore performs much better than IceTagger alone. Stagger performs significantly worse than TriTagger, partly due to much lower accuracy for unknown words.

The results of this experiment show that using taggers trained on Modern Icelandic is hardly a viable option when tagging the whole of the Saga Corpus – the accuracy is simply not high enough. The drop in accuracy is in line with results from related work on tagging historical data with taggers trained on modern texts, e.g. (Rögnvaldsson and Helgadóttir, 2011; Pennacchiotti and Zanzotto, 2008; Scheible et al., 2011b).

On the other hand, these results are better (i.e. not as bad) than found by Scheible et al. (2011b) for German. When they used a tagger trained on Modern German to tag texts (58,000 tokens) with normalized modern spelling from the period 1650-1800, the accuracy dropped from about 97%, for the modern texts, to 79.7% for the older texts. This may partly be explained by the

fact that the German experiment used a variety of genres for testing, while we use texts from one genre, Sturlunga Saga. However, the main reason is probably the fact that "[...] Icelandic is often claimed to have undergone relatively small changes from the oldest written sources up to the present", and "[the changes] have not affected the inflectional system, which has not changed in any relevant respects" (Rögnvaldsson and Helgadóttir, 2011). Therefore, we do not witness such a dramatic drop in tagging accuracy as found in the German experiment.

4.3 Correcting tagging errors

When we looked at the errors made by the taggers when tagging the Saga-Gold corpus, we noticed that the gold tags in Saga-Gold were incorrect in many cases. In order to obtain more reliable evaluation results, we thus corrected some of the errors in Saga-Gold. Instead of inspecting each and every word-tag pair in the corpus (about 95,000 pairs), we only looked at those pairs for which a tagger predicts a different tag compared to the gold. We inspected these mismatches (error candidates) and manually corrected the true positives.

Correcting errors in corpora is a time consuming task, and therefore it is important to apply methods that can speed up the process. We carried out the error correction based on a general bootstrapping method, i.e. i) manually annotate/correct a small part of a corpus C; ii) train a tagger T using the annotated/corrected training corpus; iii) use the resulting tagging model to tag more (unannotated or uncorrected) data from C; iv) hand-correct the tagging of the new data and add it to the training set; iv) repeat steps ii)-iv), until all the data of C has been tagged by T and corrected (cf. (Zavrel and Daelemans, 2000; Forsbom, 2009)).

Step/Tagger	Trained on	Tokens tagged in Saga-Gold
1. HMM+Ice+HMM	IFD	1-30,000 = A
2. TriTagger	IFD ∪ A	30,000-50,000 = B
3. Stagger	IFD ∪ A ∪ B	50,000-70,000 = C
4. CombiTagger	IFD ∪ A ∪ B ∪ C	70,000-95,000 = D

Table 3: The error correction bootstrapping method.

We devised an error correction bootstrapping method using several taggers. We define an error candidate, produced by a tagger T, as a word-tag pair (w, t), such that T predicts the tag t, which is different from the corresponding gold tag in the corpus.

Table 3 shows which taggers were used during each phase of the error correction, which part of the corpus they were trained on, and the part they tagged.[8] In step 1, we used the HMM+Ice+HMM tagger, trained on the IFD corpus, to tag the first 30,000 tokens in Saga-Gold, and then we inspected/corrected the error candidates generated by the tagger for these tokens. In step 2, we used TriTagger, trained on the union of the IFD corpus and the first 30,000 corrected tokens from Saga-Gold, to tag tokens 30,000-50,000 in the corpus. Then we inspected/corrected the error candidates generated by the tagger for these 20,000 tokens. In step 3, we used Stagger, trained on the union of the IFD corpus and the first 50,000 corrected tokens in Saga-Gold (corrected in steps 1 and 2), to tag tokens 50,000-70,000 in the corpus. Then we inspected/corrected the error candidates generated by Stagger for these 20,000 tokens.

[8] The error correction phase took 40-50 hours.

In the last step, we trained TriTagger, Stagger and HMM+Ice+HMM on the union of the IFD corpus and the first 70,000 corrected tokens in Saga-Gold (corrected in steps 1-3). Using the resulting models, we tagged the last 25,000 tokens in Saga-Gold. Then, we applied CombiTagger on the output of the three taggers in a simple voting scheme. If at least two taggers out of three agree on a tag then the corresponding tag is selected. If all taggers disagree, then the tag of the best performing tagger is selected. Finally, we inspected/corrected the error candidates generated by the combined tagger, for the last 25,000 of the 95,000 tokens in Saga-Gold.

In total, we corrected the tags for 2,144 tokens in Saga-Gold, i.e. 2.3% of the total number of tokens. Note that we used different taggers at different stages to point to error candidates. If we had used a single tagger T, then the accuracy of T might have been overestimated when testing on the corrected corpus. The reason is that we only look at those instances where T predicts a tag which is different from the gold tag. This means that we miss those cases where T agrees with the gold tag, but the gold tag is indeed incorrect!

Columns 5-7 in Table 2 show the accuracy of the four taggers when tested against the corrected version of Saga-Gold. We can see that all the taggers obtain higher accuracy on all words when tagging the corrected corpus.

The four taggers benefit differently from the error correction. Considering all words, the tagging accuracy of IceTagger, HMM+Ice+HMM, Stagger and TriTagger increases by 1.19, 1.14, 0.79, and 0.18 percentage points, respectively. The reason why IceTagger and HMM+Ice+HMM benefit the most is probably that in many cases the rules of IceTagger had indeed predicted a correct tag, but the taggers were "penalized" because of incorrect annotation in Saga-Gold (i.e. before the correction was carried out).

4.4 Cross-validation using Saga-Gold

The previous section showed that using taggers trained on Modern Icelandic to tag Old Icelandic resulted in accuracies below 87%. In the next experiment, we trained and tested the taggers on the corrected Saga-Gold corpus only, using 10-fold cross-validation.[9] We split Saga-Gold into 10 folds, such that the 1^{st} sentence of Saga-Gold was put into the 1^{st} fold, the 2^{nd} sentence into the 2^{nd} fold, ..., the 11^{th} sentence into the 1^{st} fold, the 12^{th} sentence into the 2^{nd} fold, etc. The resulting test files have 9,520 tokens, on average. We only evaluated TriTagger and Stagger in this experiment, since the dictionaries of IceTagger (and thus also of HMM+Ice+HMM) are derived from the IFD corpus.

Tagger	Unknown	Known	All
TriTagger	61.07	90.96	**89.26**
Stagger	52.67	92.53	**90.29**

Table 4: Average tagging accuracy (%) of two taggers when tagging the corrected Saga-Gold using 10-fold cross-validation. Average UWR in testing is 5.7%.

In Table 4, we can see that TriTagger obtains 89.26% accuracy for all words, and that Stagger performs better, i.e. it obtains an accuracy of 90.29%. The accuracies for both taggers increases substantially compared to when trained on Modern Icelandic. Note also that the UWR in Table

[9] When training Stagger on Saga-Gold, we used 8 iterations.

4 is 5.7%, only about half of the UWR in Table 2. We had indeed expected that Stagger would out-perform TriTagger, since this is what we found when the taggers were trained and tested using modern texts only (see Table 1).

The accuracy of 90.29% obtained by Stagger is less than 1 percentage points lower than the accuracy of the same tagger when trained and tested on Modern Icelandic (see Table 1), despite a large difference in the size of the training material available in the two corpora, Saga-Gold and the IFD. Note, however, that Saga-Gold only contains one genre, whereas the IFD contains several genres. The tagging of the former corpus is thus, presumably, easier than the tagging of the latter, given the same amount of training data.

In order to reduce the UWR of 5.7%, and to increase the training material, we, next, added data from the IFD. We thus trained the three taggers, TriTagger, Stagger and HMM+Ice+HMM, on the union of Saga-Gold and the IFD, i.e. we added the whole of the IFD corpus to each fold from Saga-Gold. Furthermore, we used CombiTagger in a simple voting scheme on the output of the three taggers.

The results are shown in Table 5. The UWR drops down to 3.6% by adding data from the IFD. Stagger is the best performing single tagger, obtaining an accuracy of 91.76% for all words. By combining the output of three taggers, the accuracy increases to 92.32% for all words.

Tagger	Unknown	Known	All
TriTagger	71.50	90.96	**90.26**
HMM+Ice+HMM	70.91	91.29	**90.58**
Stagger	64.01	92.77	**91.76**
CombiTagger	72.38	93.09	**92.32**

Table 5: Average tagging accuracy (%) of three taggers, and a combination method, when tagging the corrected Saga-Gold using 10-fold cross-validation. The IFD corpus is added to each training fold. Average UWR in testing is 3.6%.

Even though the accuracy of 92.32% can likely be improved (see Section 5), we believe that it is high enough for applying the combination method for tagging the whole of the Saga Corpus. Recall from our discussion in Section 3.1, that the tagset is large and makes fine distinctions. However, this level of detail in the tags might not be necessary for all research in corpus linguistics.

In two additional experiments, we relaxed the condition that the whole tag string needs to be correct. First, we allowed the gold tag and the tag of the combined tagger to differ in only one of the morphological features, given that the word class was correct. This results in an accuracy of 96.63%. Second, when only demanding that the word class is correct (thus ignoring all morphological features), the accuracy increases to 97.55%.

5 Discussion and Error Analysis

Since research on tagging Old Icelandic is currently in its starting phase, the accuracy can most likely be improved in future work. For the reason mentioned at the end of Section 3.2, one should be able to achieve higher accuracy on Old Icelandic texts compared to Modern Icelandic. Furthermore, since a tagger for Old Icelandic does not (necessarily) need to be able to handle

modern texts, it should not need to be trained on large amount of the IFD corpus. In future work, we would thus like to experiment with using only a part of the IFD for training our tagger.

In the remainder of this section, we discuss some of the most frequent tagging errors. We performed error analysis on the output of Stagger, when trained on the union of Saga-Gold and the IFD. We combined the tagging errors from the test sets of all of the 10 folds.

We define an error type as a pair (x, y), where x is the predicted tag and y is the gold tag. Stagger makes 1876 different error types. 1121 of those, or 59.8%, appear only once. The 10 most frequent errors account for 18.5% of the total errors, as shown in Table 6.

Error type	Rate	Cumulative rate
(c,aa)	2.80	2.80
(aþ,ao)	2.70	5.50
(sfg3en,ct)	2.49	7.99
(ct,c)	1.81	9.80
(ao,aþ)	1.66	11.46
(c,ct)	1.49	12.95
(aa,aþ)	1.49	14.44
(sfg3en,c)	1.48	15.92
(nken-s,nkeo-s)	1.31	17.23
(aa,ao)	1.30	18.53

Table 6: The 10 most frequent error types and their rate of occurrence in % in the output of Stagger. An error type is a pair (x, y): x is the predicted tag and y is the gold tag. aa=adverb; ao/aþ=preposition governing accusative/dative case; c=conjunction; ct=relative particle; nken-s/nkeo-s=noun, masculine, singular, nominative/accusative, proper noun; sfg3en=verb, indicative mood, active voice, third person, singular, present tense.

The rate of the most frequent error type, (c,aa), is 2.80%. This error occurs when Stagger predicts a conjunction (*c*), while an adverb (*aa*) is correct. The word "og" is to blame for this error type. In Modern Icelandic it usually denotes the coordinating conjunction 'and', while in Old Icelandic it often denotes the adverb 'also'. Below, one such tagging error made by Stagger is shown, when tagging the sentence "Þar var og Eyvindur prestur Þórarinsson" 'There was also Eiríkur priest Þórarinsson':

Word	Stagger	Gold tag
Þar	aa	aa
var	sfg3eþ	sfg3eþ
og	c	aa
Eyvindur	nken-s	nken-s
prestur	nken	nken
Þórarinsson	nken-s	nken-s

The rate of the third most frequent error type, (sfg3en,ct), is 2.49%. The *sfg3en* tag denotes: verb (*s*), indicative mood (*f*), active voice (*g*), third person (*3*), singular (*e*), and present tense (*n*). The *ct* tag denotes a relative particle. This particular error type occurs with the word "er". In Modern Icelandic this word most often means 'is, am', while in Old Icelandic it is most often

used as a temporal conjunction ('when') (tag *c*) or a relative particle ('that, which, who') (tag *ct*) (Rögnvaldsson and Helgadóttir, 2011).

Below, one such tagging error made by Stagger is demonstrated, when tagging the sentence part "Maður hét Haukur er kallaður var Víga-Haukur" 'Man named Haukur who called was Víga-Haukur':

Word	Stagger	Gold tag
Maður	nken	nken
hét	sfg3eþ	sfg3eþ
Haukur	nken-s	nken-s
er	sfg3en	ct
kallaður	sþgken	sþgken
var	sfg3eþ	sfg3eþ
Víga-Haukur	nken-s	nken-s

When Stagger was trained on the IFD only and tagged Saga-Gold (see Section 4.2), the aggregate rate of the error types (sfg3en,ct) and (sfg3en,c) was 5.67%. In contrast, when Stagger is trained on the union of Saga-Gold and the IFD, the rate of the same two error types is 2.49% + 1.48% = 3.97% (see Table 6). Thus, by adding training data from Saga-Gold, the tagger learns to tag these constructions correctly in some instances, but still tags many of them incorrectly, due to the large amount of training data coming from the IFD. In contrast, when Stagger is trained on Saga-Gold only (no data from IFD), the rate of the same error types is only 0.2%.

Finally, in Table 6, the following four error types appear: *(aþ,ao)*, *(ao,aþ)*, *(aa,aþ)*, and *(aa,ao)*. The first two differentiate between a preposition governing the accusative (*ao*) or dative case (*aþ*). The latter two differentiate between an adverb (*aa*) and a prepositions (*ao*/*aþ*). The underlying constructions, from which these four error types originate, are difficult to annotate correctly, even for humans.

6 Conclusion

In this paper, we first evaluated taggers, which were trained on Modern Icelandic, when tagging Saga-Gold, a sub-corpus of the Icelandic Saga Corpus. Second, we described a bootstrapping method used to correct 2.3% of the tokens in Saga-Gold. Third, we performed experiments in using several taggers to tag the corrected corpus. Finally, we discussed the results of error analysis.

Stagger is the best performing single tagger, obtaining an accuracy of 91.76% when trained on the union of Saga-Gold and the IFD corpus. By combining the output of three different taggers using a simple voting scheme, the accuracy increases to 92.32%.

In future work, we intend to: i) experiment with increasing the accuracy of Stagger for unknown words; ii) find ways to tackle the most frequent error types; and iii) experiment with using only a part of the IFD corpus as training material.

We would also like to find and correct more errors in Saga-Gold, using our combined tagger to point to error candidates. We intend to use the combined tagger to tag the whole of the Saga Corpus and make the results public – to facilitate corpus linguistics research on Old Icelandic.

Acknowledgments

We thank Sigrún Helgadóttir and Eiríkur Rögnvaldsson for making the training corpus, Saga-Gold, available to us. We are also grateful to Robert Östling for his willingness in improving Stagger's functionality based on our suggestions.

References

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, Sydney, Australia.

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.

Brill, E. and Wu, J. (1998). Classifier Combination for Improved Lexical Disambiguation. In *Proceedings of the* 36^{th} *Annual Meeting of the Association for Computational Linguistics and* 17^{th} *International Conference on Computational Linguistics*, COLING-ACL, Montreal, Quebec, Canada.

Collins, M. (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the ACL-02 Conference on Empirical methods in Natural Language Processing*, Philadelphia, PA, USA.

Dredze, M. and Wallenberg, J. (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*, ACL-HLT, Columbus, OH, USA.

Forsbom, E. (2009). Extending the View: Explorations in Bootstrapping a Swedish PoS Tagger. In *Proceedings of the* 17^{th} *Nordic Conference of Computational Linguistics*, NoDaLiDa, Odense, Denmark.

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

Helgadóttir, S. (2005). Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In Holmboe, H., editor, *Nordisk Sprogteknologi 2004*. Museum Tusculanums Forlag, Copenhagen.

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.

Kroch, A. and Taylor, A. (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition.

Kübler, S. and Baucom, E. (2011). Fast Domain Adaptation for Part of Speech Tagging for Dialogues. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP, Hissar, Bulgaria.

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

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

Loftsson, H., Helgadóttir, S., and Rögnvaldsson, E. (2011). Using a morphological database to increase the accuracy in PoS tagging. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP, Hissar, Bulgaria.

Loftsson, H., Kramarczyk, I., Helgadóttir, S., and Rögnvaldsson, E. (2009). Improving the PoS tagging accuracy of Icelandic text. In *Proceedings of the* 17^{th} *Nordic Conference of Computational Linguistics*, NoDaLiDa, Odense, Denmark.

Loftsson, H. and Östling, R. (2013). Tagging a Morphologically Complex Language Using an Averaged Perceptron Tagger: The Case of Icelandic. In *Proceedings of the* 19^{th} *Nordic Conference of Computational Linguistics*, NoDaLiDa, Oslo, Norway.

Pennacchiotti, M. and Zanzotto, F. M. (2008). Natural Language Processing across time: an empirical investigation on Italian. In Nordström, B. and Ranta, A., editors, *Advances in Natural Language Processing,* 6^{th} *International Conference on NLP, GoTAL 2008, Proceedings*. Springer, Berlin.

Pettersson, E., Megyesi, B., and Nivre, J. (2012). Parsing the past – identification of verb constructions in historical text. In *EACL 2012 workshop on: Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Avignon, France.

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

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA.

Rögnvaldsson, E. and Helgadóttir, S. (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In Sporleder, C., van den Bosch, A., and Zervanou, K., editors, *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*. Springer, Berlin.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the* 8^{th} *International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey.

Sánchez-Marco, C., Boleda, G., and Padró, L. (2011). Extending the tool, or how to annotate historical language varieties. In *Proceedings of the* 5^{th} *ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR, USA.

Scheible, S., Whitt, R. J., Durrell, M., and Bennett, P. (2011a). A Gold Standard Corpus of Early Modern German. In *Proceedings of the ACL-HLT 2011 Linguistic Annotation Workshop (LAW V)*, Portland, OR, USA.

Scheible, S., Whitt, R. J., Durrell, M., and Bennett, P. (2011b). Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In *Proceedings of the* 5^{th} *ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR, USA.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (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.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., and Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9.

Zavrel, J. and Daelemans, W. (2000). Bootstrapping a Tagged Corpus through Combination of Existing Heterogeneous Taggers. In *Proceedings of the* 2^{nd} *International Conference on Language Resources and Evaluation*, LREC, Athens, Greece.

Östling, R. (2012). Stagger: A modern POS tagger for Swedish. In *Proceedings of the* 4^{th} *Swedish Language Technology Conference*, SLTC, Lund, Sweden.