Correcting a PoS-tagged corpus using three complementary methods

Hrafn Loftsson
School of Computer Science
Reykjavik University
Reykjavik, Iceland
hrafn@ru.is

Abstract

The quality of the part-of-speech (PoS) annotation in a corpus is crucial for the development of PoS taggers. In this paper, we experiment with three complementary methods for automatically detecting errors in the PoS annotation for the Icelandic Frequency Dictionary corpus. The first two methods are language independent and we argue that the third method can be adapted to other morphologically complex languages. Once possible errors have been detected, we examine each error candidate and hand-correct the corresponding PoS tag if necessary. Overall, based on the three methods, we hand-correct the PoS tagging of 1,334 tokens (0.23% of the tokens) in the corpus. Furthermore, we re-evaluate existing state-of-the-art PoS taggers on Icelandic text using the corrected corpus.

1 Introduction

Part-of-speech (PoS) tagged corpora are valuable resources for developing PoS taggers, i.e. programs which automatically tag each word in running text with morphosyntactic information. Corpora in various languages, such as the English *Penn Treebank* corpus (Marcus et al., 1993), the Swedish *Stockholm-Umeå* corpus (Ejerhed et al., 1992), and the *Icelandic Frequency Dictionary* (IFD) corpus (Pind et al., 1991), have been used to train (in the case of data-driven methods) and develop (in the case of linguistic rule-based methods) different taggers, and to evaluate their accuracy, e.g. (van Halteren et al., 2001; Megyesi, 2001; Loftsson, 2006). Consequently, the quality of the PoS annotation in a corpus (the gold standard annotation) is crucial.

Many corpora are annotated semi-automatically. First, a PoS tagger is run on the corpus text, and, then, the text is hand-corrected by humans. Despite human post-editing, (large) tagged corpora are almost certain to contain errors, because humans make mistakes. Thus, it is important to apply known methods and/or develop new methods for automatically detecting tagging errors in corpora. Once an error has been detected it can be corrected by humans or an automatic method.

In this paper, we experiment with three different methods of PoS error detection using the IFD corpus. First, we use the *variation n-gram* method proposed by Dickinson and Meurers (2003). Secondly, we run five different taggers on the corpus and examine those cases where all the taggers agree on a tag, but, at the same time, disagree with the gold standard annotation. Lastly, we use *IceParser* (Loftsson and Rögnvaldsson, 2007) to generate shallow parses of sentences in the corpus and then develop various patterns, based on feature agreement, for finding candidates for annotation errors.

Once error candidates have been detected by each method, we examine the candidates manually and correct the errors. Overall, based on these methods, we hand-correct the PoS tagging of 1,334 tokens or 0.23% of the tokens in the IFD corpus. We are not aware of previous corpus error detection/correction work applying the last two methods above. Note that the first two methods are completely language-independent, and the third method can be tailored to the language at hand, assuming the existence of a shallow parser.

Our results show that the three methods are complementary. A large ratio of the tokens that get hand-corrected based on each method is uniquely corrected by that method[1].

[1]To be precise, when we say that an error is corrected by a method, we mean that the method detected the error candidate which was then found to be a true error by the separate error correction phase.

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 523–531,
Athens, Greece, 30 March – 3 April 2009.

After hand-correcting the corpus, we retrain and re-evaluate two of the best three performing taggers on Icelandic text, which results in up to 0.18% higher accuracy than reported previously.

The remainder of this paper is organised as follows. In Section 2 we describe related work, with regard to error detection and PoS tagging of Icelandic text. Our three methods of error detection are described in Section 3 and results are provided in Section 4. We re-evaluate taggers in Section 5 and we conclude with a summary in Section 6.

2 Related work

2.1 Error detection

The field of automatic error detection/correction in corpora has gained increased interest during the last few years. Most work in this field has focused on finding elements in corpora that violate consistency, i.e. finding inconsistent tagging of a word across comparable occurrences.

The variation n-gram algorithm is of this nature. This method finds identical strings (n-grams of words) in a corpus that are annotated differently. The difference in PoS tags between the strings is called a *variation* and the word(s) exhibiting the variation is called a *variation nucleus* (Dickinson and Meurers, 2003). A particular variation is thus a possible candidate for an error. The variation might be due to an error in the annotation or it might exhibit different (correct) tagging because of different contexts. Intuitively, the more similar the context of a variation, the more likely it is for the variation to be an error.

When Dickinson and Meurers applied their variation n-gram algorithm to the Wall Street Journal (WSJ) corpus of about 1.3 million words, it produced variations up to length $n = 224$. Note that a variation n-gram of length n contains two variation n-grams of length $n - 1$, obtained by removing either the first or the last word. Moreover, each variation n-gram contains at least two different annotations of the same string. Therefore, it is not straightforward to compute the precision (the ratio of correctly detected errors to all error candidates) of this method. However, by ignoring variation n-grams of length ≤ 5, Dickinson and Meurers found that 2436 of the 2495 distinct variation nuclei (each nucleus is only counted for the longest n-gram it appears in) were true errors, i.e. 97.6%. This resulted in 4417 tag corrections, i.e. about 0.34% of the tokens in the whole corpus were found to be incorrectly tagged[2].

Intuitively, the variation n-gram method is most suitable for corpora containing specific genres, e.g. business news like the WSJ, or very large balanced corpora, because in both types of corpora one can expect the length of the variations to be quite large. Furthermore, this method may not be suitable for corpora tagged with a large fine-grained tagset, because in such cases a large ratio of the variation n-grams may actually reflect true ambiguity rather than inconsistent tagging.

Another example of a method, based on finding inconsistent tagging of a word across comparable occurrences, is the one by Nakagawa and Matsumoto (2002). They use support vector machines (SVMs) to find elements in a corpus that violate consistency. The SVMs assign a weight to each training example in a corpus – a large weight is assigned to examples that are hard for the SVMs to classify. The hard examples are thus candidates for errors in the corpus. The result was a remarkable 99.5% precision when examples from the WSJ corpus were extracted with a large weight greater than or equal to a threshold value. However, the disadvantage with this approach is that a model of SVMs needs to be trained for each PoS tag, which makes it unfeasible for large tagsets.

A set of invalid n-grams can be used to search for annotation errors. The algorithm proposed by Květoň and Oliva (2002) starts from a known set of invalid bigrams, *[first,second]*, and incrementally constructs a set of *allowed inner tags* appearing between the tags *first* and *second*. This set is then used to generate the complement, *impossible inner tags* (the set of all tags excluding the set *allowed inner tags*). Now, any n-gram consisting of the tag *first*, followed by any number of tags from the set *impossible inner tags*, finally followed by the tag *second*, is a candidate for an annotation error in a corpus. When this method was applied on the NEGRA corpus (containing 350,000 tokens) it resulted in the hand-correction of 2,661 tokens or 0.8% of the corpus. The main problem with this approach is that is presupposes a set of invalid bigrams (e.g. constructed by a linguist). For a large tagset, for example the Icelandic one (see Section 2.2), constructing this set is a very hard task. Moreover, this method fails to detect annotation errors where a particular n-gram tag sequence

[2] In a more recent work, Dickinson (2008) has developed a method for increasing the recall (the ratio of correctly detected errors to all errors in the corpus).

is valid but erroneous in the given context.

PoS taggers have also been used to point to possible errors in corpora. If the output of a tagger does not agree with the gold standard then either the tagger is incorrect or the gold standard is incorrectly annotated. A human can then look at the disagreements and correct the gold standard where necessary. van Halteren (2000) trained a tagger on the written texts of the British National Corpus sampler CD (about 1 million words). In a random sample of 660 disagreements, the tagger was correct and the gold standard incorrect in 84 cases, i.e. the precision of this error detection method was 12.7%. A natural extension of this method is to use more than one tagger to point to disagreements.

2.2 PoS tagging Icelandic

The IFD corpus is a balanced corpus, consisting of 590,297 tokens. 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 corpus, is large (700 possible tags) 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 "*hestarnir*" ('(the) horses'). The corresponding tag is "*nkfng*", denoting noun (*n*), masculine (*k*), plural (*f*), nominative (*n*), and suffixed definite article (*g*).

The large tagset mirrors the morphological complexity of the Icelandic language. This, in turn, is the main reason for a relatively low tagging accuracy obtained by PoS taggers on Icelandic text, so far. The state-of-the art tagging accuracy, measured against the IFD corpus, is 92.06%, obtained by applying a bidirectional PoS tagging method (Dredze and Wallenberg, 2008). We have developed a linguistic rule-based tagger, *IceTagger*, achieving about 91.6% tagging accuracy (Loftsson, 2008). Evaluation has shown that the well known statistical tagger, *TnT* (Brants, 2000), obtains about 90.4% accuracy (Helgadóttir, 2005; Loftsson, 2008). Finally, an accuracy of about 93.5% has been achieved by using a tagger combination method using five taggers (Loftsson, 2006).

3 Three methods for error detection

In this section, we describe the three methods we used to detect (and correct) annotation errors in the IFD corpus. Each method returns a set of error candidates, which we then manually inspect and correct the corresponding tag if necessary.

3.1 Variation n-grams

We used the Decca software (`http://decca.osu.edu/`) to find the variation n-grams in the corpus. The length of the longest variation n-gram was short, i.e. it consisted of only 20 words. The longest variation that contained a true tagging error was 15 words long. As an example of a tagging error found by this method, consider the two occurrences of the 4-gram variation "henni datt í hug" (meaning 'she got an idea'):

1) henni/fpveþ datt/sfg3eþ í/a**þ** hug/nke**þ**
2) henni/fpveþ datt/sfg3eþ í/a**o** hug/nke**o**

In the first occurrence, the substring "í hug" (the variation nucleus) is incorrectly tagged as a preposition governing the dative case ("aþ"), and a noun in masculine, singular, dative ("nkeþ"). In the latter occurrence, the same substring is correctly tagged as a preposition governing the accusative case ("ao"), and a noun in masculine, singular, accusative ("nkeo"). In both cases, note the agreement in case between the preposition and the noun.

As discussed earlier, the longer variation n-grams are more likely to contain true errors than the shorter ones. Therefore, we manually inspected all the variations of length ≥ 5 produced by this method (752 in total), but only "browsed through" the variations of length 4 (like the one above; 2070 variations) and of length 3 (7563 variations).

3.2 Using five taggers

Instead of using a single tagger to tag the text in the IFD corpus, and compare the output of the taggers to the gold standard (as described in Section 2.1), we decided to use five taggers. It is well known that a combined tagger usually obtains higher accuracy than individual taggers in the combination pool. For example, by using simple voting (in which each tagger "votes" for a tag

and the tag with the highest number of votes is selected by the combined tagger), the tagging accuracy can increase significantly (van Halteren et al., 2001; Loftsson, 2006). Moreover, if all the taggers in the pool agree on a vote, one would expect the tagging accuracy for the respective words to be high. Indeed, we have previously shown that when five taggers all agree on a tag in the IFD corpus, the corresponding accuracy is 98.9% (Loftsson, 2007b). For the remaining 1.1% tokens, one would expect that the five taggers are actually correct in some of the cases, but the gold standard incorrectly annotated. In general, both the precision and the recall should be higher when relying on five agreeing taggers as compared to using only a single tagger.

Thus, we used the five taggers, MBL (Daelemans et al., 1996), MXPOST (Ratnaparkhi, 1996), fnTBL (Ngai and Florian, 2001), TnT, and IceTagger[3], in the same manner as described in (Loftsson, 2006), but with the following minor changes. We extended the dictionaries of the TnT tagger and IceTagger by using data from a full-form morphological database of inflections (Bjarnadóttir, 2005). The accuracy of the two taggers increases substantially (because the ratio of unknown words drops dramatically) and, in turn, the corresponding accuracy when all the taggers agree increases from 98.9% to 99.1%. Therefore, we only needed to inspect about 0.9% of the tokens in the corpus.

The following example from the IFD corpus shows a disagreement found between the five taggers and the gold standard: "fjölskylda spákonunnar í gamla húsinu" ('family (the) fortune-teller's in (the) old house').

3) fjölskylda/nven spákonunnar/nveeg í/a**o** gamla/lhe**þ**vf húsinu/nhe**þ**g

In this case, the disagreement lies in the tagging of the preposition "í". All the five taggers suggest the correct tag "aþ" for the preposition (because case agreement is needed between the preposition and the following adjective/noun).

3.3 Shallow parsing

In a morphologically complex language like Icelandic, feature agreement, for example inside noun phrases or between a preposition and a noun phrase, plays an important role. Therefore, of the total number of possible errors existing in an Icelandic corpus, feature agreement errors are likely to be prevalent. A constituent parser is of great help in finding such error candidates, because it annotates phrases which are needed by the error detection mechanism. We used IceParser, a shallow parser for parsing Icelandic text, for this purpose.

The input to IceParser is PoS tagged text, using the IFD tagset. It produces annotation of both constituent structure and syntactic functions. To illustrate, consider the output of IceParser when parsing the input from 3) above:

4) {*SUBJ [NP fjölskylda nven NP] {*QUAL [NP spákonunnar nveeg NP] *QUAL} *SUBJ} [PP í ao [NP [AP gamla lheþvf AP] húsinu nheþg NP] PP]

The constituent labels seen here are: *PP*=a preposition phrase, *AP*=an adjective phrase, and *NP*=a noun phrase. The syntactic functions are **SUBJ*=a subject, and **QUAL*=a genitive qualifier.

This (not so shallow) output makes it relatively easy to find error candidates. Recall from example 3) that the accusative preposition tag "ao", associated with the word "í", is incorrect (the correct tag is the dative "aþ"). Since a preposition governs the case of the following noun phrase, the case of the adjective "gamla" and the noun "húsinu" should match the case of the preposition. Finding such error candidates is thus just a matter of writing regular expression patterns, one for each type of error.

Furthermore, IceParser makes it even simpler to write such patterns than it might seem when examining the output in 4). IceParser is designed as a sequence of finite-state transducers. The output of one transducer is used as the input to the next transducer in the sequence. One of these transducers marks the case of noun phrases, and another one the case of adjective phrases. This is carried out to simplify the annotation of syntactic functions in the transducers that follow, but is removed from the final output (Loftsson and Rögnvaldsson, 2007). Let us illustrate again:

5) {*SUBJ [NP**n** fjölskylda nven NP] {*QUAL [NP**g** spákonunnar nveeg NP] *QUAL} *SUBJ}

[3]The first four taggers are data-driven, but IceTagger is a linguistic rule-based tagger.

[PP í ao [NP**d** [AP**d** gamla lheþvf AP] húsinu nheþg NP] PP]

In 5), an intermediate output is shown from one of the transducers of IceParser, for the sentence from 4). Note that letters have been appended to some of the phrase labels. This letter denotes the case of the corresponding phrase, e.g. “n”=nominative, “a”=accusative, “d”=dative, and “g”=genitive.

The case letter attached to the phrase labels can thus be used when searching for specific types of errors. Consider, for example, the pattern *PrepAccError* (slightly simplified) which is used for detecting the error shown in 5) (some details are left out)[4]:

```
PrepTagAcc = ao{WhiteSpace}+
PrepAcc = {Word}{PrepTagAcc}

PrepAccError =
  "[PP"{PrepAcc}("[NP"[nde]~"NP]")
```

This pattern searches for a string starting with “[PP” followed by a preposition governing the accusative case ({PrepAcc}), followed by a substring starting with a noun phrase “[NP”, marked as either nominative, dative or genitive case (“[nde]”), and ending with “NP]”.

We have designed three kinds of patterns, one for PP errors as shown above, one for disagreement errors inside NPs, and one for specific VP (verb phrase) errors.

The NP patterns are more complicated than the PP patterns, and due to lack of space we are not able to describe them here in detail. Briefly, we extract noun phrases and use string processing to compare the gender, number and case features in nouns to, for example, the previous adjective or pronoun. If a disagreement is found, we print out the corresponding noun phrase. To illustrate, consider the sentence “í þessum landshluta voru fjölmörg einkasjúkrahús” (’in this part-of-the-country were numerous private-hospitals’), annotated by IceParser in the following way:

6) [PP í aþ [NP þessum fa**kfþ** landshluta n**keþ** NP] PP] [VPb voru sfg3fþ VPb] {*SUBJ< [NP [AP fjölmörg lhfnsf AP] einkasjúkrahús nhfn NP] *SUBJ<}

In this example, there is a disagreement error in number between the demonstrative pronoun “þessum” and the following noun “landshluta”. The second “f“ letter in the tag “fakfþ” for “þessum” denotes plural and the letter “e” in the tag “nkeþ” for “landshluta” denotes singular.

Our VP patterns mainly search for disagreements (in person and number) between a subject and the following verb[5]. Consider, for example, the sentence “ég les meira um vísindin” (’I read more about (the) science’), annotated by IceParser in the following manner:

7) {*SUBJ> [NP ég fp**1e**n NP] *SUBJ>} [VP les sfg**3e**n VP] {*OBJ< [AP meira lheovm AP] *OBJ<} [PP um ao [NP vísindin nhfog NP] PP]

The subject “ég” is here correctly tagged as personal pronoun, first person, (“fp1en”), but the verb “les” is incorrectly tagged as third person (“sfg3en”).

By applying these pattern searches to the output of IceParser for the whole IFD corpus, we needed to examine 1,489 error candidates, or 0.25% of the corpus. Since shallow parsers have been developed for various languages, this error detection method may be tailored to other morphologically complex languages.

Notice that the above search patterns could potentially be used in a grammar checking component for Icelandic text. In that case, input text would be PoS tagged with any available tagger, shallow parsed with IceParser, and then the above patterns used to find these specific types of feature agreement error candidates.

4 Results

Table 1 shows the results of applying the three error detection methods on the IFD corpus. The column “Error candidates” shows the number of PoS tagging error candidates detected by each method. The column “Errors corrected” shows the number of tokens actually corrected, i.e. how many of the error candidates were true errors. The column “Precision” shows the ratio of correctly detected errors to all error candidates. The column “Ratio of corpus” shows the ratio of tokens corrected to all tokens in the IFD corpus. The column

[4]For writing regular expression patterns, we used the lexical analyser generator tool JFlex, `http://jflex.de/`.

[5]Additionally, one VP pattern searches for a substring containing the infinitive marker (the word “að” (’to’)), immediately followed by a verb which is not tagged as an infinitive verb.

Method	Sub-type	Error candidates	Errors corrected	Precision (%)	Ratio of corpus (%)	Uniqueness rate (%)	Feature agreement (%)
variation n-gram			254		0.04	65.0	4.7
5 taggers		5317	883	16.6	0.15	78.0	24.8
shallow parsing	All	1489	448	30.1	0.08	60.0	80.2
	PP	511	226	44.2	0.04	51.3	70.4
	NP	740	160	21.6	0.03	70.0	95.0
	VP	238	62	26.1	0.01	61.3	77.1
Total distinct errors			1334		0.23		

Table 1: Results for the three error detection methods

"Uniqueness rate" shows how large a ratio of the errors corrected by a method were not found by any other method. Finally, the column "Feature agreement" shows the ratio of errors that were feature agreement errors.

As discussed in Section 2.1, it is not straightforward to compute the precision of the variation n-gram method, and we did not attempt to do so. However, we can, using our experience from examining the variations, claim that the precision is substantially lower than the 96.7% precision obtained by Dickinson and Meurers (2003). We had, indeed, expected low precision when using the variation n-gram on the IFD corpus, because this corpus and the underlying tagset is not as suitable for the method as the WSJ corpus (again, see the discussion in Section 2.1). Note that as a result of applying the variation n-gram method, only 0.04% of the tokens in the IFD corpus were found to be incorrectly tagged. This ratio is 8.5 times lower than the ratio obtained by Dickinson and Meurers when applying the same method on the WSJ corpus. On the other hand, the variation n-gram method nicely complements the other methods, because 65.0% of the 254 hand-corrected errors were uniquely corrected on the basis of this method.

Table 1 shows that most errors were detected by applying the "5 taggers" method – 0.15% of the tokens in the corpus were found to be incorrectly annotated on the basis of this method. The precision of the method is 16.6%. Recall that by using a single tagger for error detection, van Halteren (2000) obtained a precision of 12.7%. One might have expected more difference in precision by using five taggers vs. a single tagger, but note that the languages used in the two experiments, as well as the tagsets, are totally different. Therefore, the comparison in precision may not be viable. Moreover, it has been shown that tagging Icelandic text, using the IFD tagset, is a hard task (see Section 2.2). Hence, even though five agreeing taggers disagree with the gold standard, in a large majority of the disagreements (83.4% in our case) the taggers are indeed wrong.

Consider, for example, the simple sentence "þá getur það enginn" ('then can it nobody', meaning 'then nobody can do-it'), which exemplifies the free word order in Icelandic. Here the subject is "enginn" and the object is "það". Therefore, the correct tagging (which is the one in the corpus) is "þá/aa getur/sfg3en það/fpheo enginn/foken", in which "það" is tagged with the accusative case (the last letter in the tag "fpheo"). However, all the five taggers make the mistake of tagging "það" with the nominative case ("fphen"), i.e. assuming it is the subject of the sentence.

The uniqueness ratio for the 5-taggers method is high or 78.0%, i.e. a large number of the errors corrected based on this method were not found (corrected) by any of the other methods. However, bear in mind, that this method produces most error candidates.

The error detection method based on shallow parsing resulted in about twice as many errors corrected than by applying the variation n-gram method. Even though the precision of this method as a whole (the subtype marked "All" in Table 1) is considerably higher than when applying the 5-taggers methods (30.1% vs. 16.6%), we did expect higher precision. Most of the false positives (error candidates which turned out not to be errors) are due to incorrect phrase annotation in IceParser. A common incorrect phrase annotation is one which includes a genitive qualifier. To illustrate, consider the following sentence "sumir farþeganna voru á heimleið" ('some of-the-passengers were on-their-way home'), matched

by one of the NP error patterns:

8) {*QUAL [NP sumir fo**kfn** farþeganna n**kfe**g NP] *QUAL} [VPb voru sfg3fþ VPb] [PP á aþ [NP heimleið nveþ NP] PP]

Here "sumir farþeganna" is annotated as a single noun phrase, but should be annotated as two noun phrases "[NP sumir fokfn NP]" and "[NP farþeganna nkfeg NP]", where the second one is the genitive qualifier of the first one. If this was correctly annotated by IceParser, the NP error pattern would not detect any feature agreement error for this sentence, because no match is carried out across phrases.

The last column in Table 1 shows the ratio of feature agreement errors, which are errors resulting from mismatch in gender/person, number or case between two words (e.g., see examples 6) and 7) above). Examples of errors not resulting from feature agreement are: a tag denoting the incorrect word class, and a tag of a an object containing an incorrect case (verbs govern the case of their objects).

Recall from Section 3.3 that rules were written to search for feature agreement errors in the output of IceParser. Therefore, a high ratio of the total errors corrected by the shallow parsing method (80.2%) are indeed due to feature agreement mismatches. 95.0% and 70.4% of the NP errors and the PP errors are feature agreement errors, respectively. The reason for a lower ratio in the PP errors is the fact that in some cases the proposed preposition should actually have been tagged as an adverb (the proposed tag therefore denotes an incorrect word class). In the case of the 5-taggers method, 24.8% of the errors corrected are due to feature agreement errors but only 4.7% in the case of the variation n-gram method.

The large difference between the three methods with regard to the ratio of feature agreement errors, as well as the uniqueness ratio discussed above, supports our claim that the methods are indeed complementary, i.e. a large ratio of the tokens that get hand-corrected based on each method is uniquely corrected by that method.

Overall, we were able to correct 1,334 distinct errors, or 0.23% of the IFD corpus, by applying the three methods (see the last row of Table 1). Compared to related work, this ratio is, for example, lower than the one obtained by applying the variation n-gram method on the WSJ corpus (0.34%). The exact ratio is, however, not of prime importance because the methods have been applied to different languages, different corpora and different tagsets. Rather, our work shows that using a single method which has worked well for an English corpus (the variation n-gram method) does not work particularly well for an Icelandic corpus but adding two other complementary methods helps in finding errors missed by the first method.

5 Re-evaluation of taggers

Earlier work on evaluation of tagging accuracy for Icelandic text has used the original IFD corpus (without any error correction attempts). Since we were able to correct several errors in the corpus, we were confident that the tagging accuracy published hitherto had been underestimated.

To verify this, we used IceTagger and TnT, two of the three best performing taggers on Icelandic text. Additionally, we used a changed version of TnT, which utilises functionality from IceMorphy, the morphological analyser of IceTagger, and a changed version of IceTagger which uses a hidden Markov Model (HMM) to disambiguate words which can not be further disambiguated by applying rules (Loftsson, 2007b). In tables 2 and 3 below, *Ice* denotes IceTagger, *Ice** denotes IceTagger+HMM, and *TnT** denotes TnT+IceMorphy.

We ran 10-fold cross-validation, using the exact same data-splits as used in (Loftsson, 2006), both before error correction (i.e. on the original corpus) and after the error correction (i.e. on the corrected corpus). Note that in these two steps we did not retrain the TnT tagger, i.e. it still used the language model derived from the original uncorrected corpus.

Using the original corpus, the average tagging accuracy results (using the first nine splits), for unknown words, known words, and all words, are shown in Table 2[6]. The average unknown word ratio is 6.8%.

Then we repeated the evaluation, now using the corrected corpus. The results are shown in Table 3. By comparing the tagging accuracy for all words in tables 2 and 3, it can be seen that the accuracy had be underestimated by 0.13-0.18 percentage points. The taggers TnT* and Ice* benefit the most from the corpus error correction – their

[6]The accuracy figures shown in Table 2 are comparable to the results in (Loftsson, 2006).

Words	TnT	TnT*	Ice	Ice*
Unknown	71.82	72.98	75.30	75.63
Known	91.82	92.60	92.78	93.01
All	90.45	91.25	91.59	91.83

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

Words	TnT	TnT*	Ice	Ice*
Unknown	71.88	73.03	75.36	75.70
Known	91.96	92.75	92.95	93.20
All	90.58	91.43	91.76	92.01

Table 3: Average tagging accuracy (%) using the corrected IFD corpus

accuracy for all words increases by 0.18 percentage points. Recall that we hand-corrected 0.23% of the tokens in the corpus, and therefore TnT* and Ice* correctly annotate 78.3% $(0.18/0.23)$ of the corrected tokens.

Since the TnT tagger is a data-driven tagger, it is interesting to see whether the corrected corpus changes the language model (to the better) of the tagger. In other words, does retraining using the corrected corpus produce better results than using the language model generated from the original corpus? The answer is yes, as can be seen by comparing the accuracy figures for TnT and TnT* in tables 3 and 4. The tagging accuracy for all words increases by 0.10 and 0.07 percentage points for TnT and TnT*, respectively.

The re-evaluation of the above taggers, with or without retraining, clearly indicates that the quality of the PoS annotation in the IFD corpus has significant effect on the accuracy of the taggers.

6 Conclusion

The work described in this paper consisted of two stages. In the first stage, we used three error detection methods to hand-correct PoS errors in an Icelandic corpus. The first two methods are language independent, and we argued that the third method can be adapted to other morphologically complex languages.

As we expected, the application of the first method used, the variation n-gram method, did result in relatively few errors being detected and corrected (i.e. 254 errors). By adding two new methods, the first based on the agreement of five taggers, and the second based on shallow parsing, we were able to detect and correct 1,334 errors in total, or 0.23% of the tokens in the corpus. Our analysis shows that the three methods are complementary, i.e. a large ratio of the tokens that get hand-corrected based on each method is uniquely corrected by that method.

Words	TnT	TnT*
Unknown	71.97	73.10
Known	92.06	92.85
All	90.68	91.50

Table 4: Average tagging accuracy (%) of TnT after retraining using the corrected IFD corpus

An interesting side effect of the first stage is the fact that by inspecting the error candidates resulting from the shallow parsing method, we have noticed a number of systematic errors made by IceParser which should, in our opinion, be relatively easy to fix. Moreover, we noted that our regular expression search patterns, for finding feature agreement errors in the output of IceParser, could potentially be used in a grammar checking tool for Icelandic.

In the second stage, we re-evaluated and retrained two PoS taggers for Icelandic based on the corrected corpus. The results of the second stage clearly indicate that the quality of the PoS annotation in the IFD corpus has a significant effect on the accuracy of the taggers.

It is, of course, difficult to estimate the recall of our methods, i.e. how many of the true errors in the corpus we actually hand-corrected. In future work, one could try to increase the recall by a variant of the 5-taggers method. Instead of demanding that all five taggers agree on a tag before comparing the result to the gold standard, one could inspect those cases in which four out of the five taggers agree. The problem, however, with that approach is that the number of cases that need to be inspected grows substantially. By demanding that all the five taggers agree on the tag, we needed to inspect 5,317 error candidates. By relaxing the conditions to four votes out of five, we would need to inspect an additional 9,120 error candidates.

Acknowledgements

Thanks to the Árni Magnússon Institute for Icelandic Studies for providing access to the IFD corpus and the morphological database of inflections, and to all the developers of the software used in this research for sharing their work.

References

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

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.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: a Memory-Based Part of Speech Tagger-Generator. In *Proceedings of the* 4^{th} *Workshop on Very Large Corpora*, Copenhagen, Denmark.

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

Markus Dickinson. 2008. Representations for category disambiguation. In *Proceedings of the* 22^{nd} *International Conference on Computational Linguistics (COLING-08)*, Manchester, UK.

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.

Eva Ejerhed, Gunnel Källgren, Ola Wennstedt, and Magnus Åström. 1992. *The Linguistic Annotation System of the Stockholm-Umeå Project*. Department of General Linguistics, University of Umeå.

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.

Pavel Květoň and Karel Oliva. 2002. Achieving an Almost Correct PoS-Tagged Corpus. In P. Sojka, I. Kopeček, and K. Pala, editors, *Proceedings of the* 5^{th} *International Conference on TEXT, SPEECH and DIALOGUE*, Brno, Czech Republic.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the* 16^{th} *Nordic Conference of Computational Linguistics (NoDaLiDa 2007)*, Tartu, Estonia.

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

Hrafn Loftsson. 2007b. *Tagging and Parsing Icelandic Text*. Ph.D. thesis, University of Sheffield, Sheffield, UK.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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.

Tetsuji Nakagawa and Yuji Matsumoto. 2002. Detecting errors in corpora using support vector machines. In *Proceedings of the* 19^{th} *International Conference on Computational Linguistics*, Taipei, Taiwan.

Grace Ngai and Radu 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.

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

Adwait Ratnaparkhi. 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.

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

Hans van Halteren. 2000. The Detection of Inconsistency in Manually Tagged Text. In A. Abeillé, T. Brants, and H. Uszkoreit, editors, *Proceedings of the* 2^{nd} *Workshop on Linguistically Interpreted Corpora*, Luxembourg.