Toward cross-lingual application of language-specific PoS tagging schemes

Hinrik Hafsteinsson and Anton Karl Ingason
University of Iceland
Reykjavík, Iceland
{hinhaf,antoni}@hi.is

Abstract

We describe the process of conversion between the PoS tagging schemes of two languages, the Icelandic MIM-GOLD tagging scheme and the Faroese Sosialurin tagging scheme. These tagging schemes are functionally similar but use separate ways to encode fine-grained morphological information on tokenised text. As Faroese and Icelandic are lexically and grammatically similar, having a systematic method to convert between these two tagging schemes would be beneficial in the field of language technology, specifically in research on transfer learning between the two languages. As a product of our work, we present a provisional version of Icelandic corpora, prepared in the Faroese PoS tagging scheme, ready for use in cross-lingual NLP applications.

1 Introduction

Part of Speech (PoS) tagging is the process of labelling words and symbols of running text based on their lexical category and morphological features. Text corpora that have been PoS-tagged in this way serve as a valuable tool in various fields of linguistic research and language technology. The specifics and format of the PoS tags used, the tagging scheme, varies greatly between languages and applications. In the current project, we focus on two languages with significant linguistic similarities, Icelandic and Faroese, and PoS tagging schemes for the two which overlap significantly in function; the MIM-GOLD tagging scheme (Barkarson et al., 2020) and the Sosialurin tagging scheme (Hansen et al., 2004), respectively.

Icelandic and Faroese are distinct yet relatively similar languages, with their similarities especially apparent in morphology and syntax. While Icelandic has seen significant gains in the field of language technology (LT) over the past few decades (Nikulásdóttir et al., 2017), the same is not true for Faroese. Due the similarities between the two, there is a real possibility that employing transfer learning, using Icelandic data in tandem with Faroese, to create effective LT tools and digital language resources for Faroese.

With the end goal of cross lingual transfer learning in mind, we focus on the task of PoS tagging. Our goal is to produce an effective way to map between the tagging schemes used for the two languages. This requires some revisions to one of the tagging schemes and assurance that a one-to-one mapping between tagsets is possible.

The paper is structured as follows. Section 2 discusses the possibilities of cross-lingual transfer learning between Faroese and Icelandic. Section 3 describes the Icelandic MIM-GOLD tagging scheme and Section 4 the Faroese Sosialurin tagging scheme. Section 5 discusses the current differences between the two tagging schemes and Section 6 details the procedure of converting between the two tagsets, while Section 7 discusses possible alternatives such a conversion. Section 8 concludes.

2 Faroese, Icelandic and transfer learning

The fundamental reason that makes Icelandic NLP implementations applicable for Faroese are the grammatical similarities between the two languages. These similarities are especially apparent in morphology, as both languages retain grammatical categories not apparent in other similar languages, e.g., four grammatical cases for nominals and an extensive conjugation system for verbs, to name a few. Furthermore, the similarities also extend to the syntax of the languages and orthographies, although with various systematic differences in both. With this in mind it can be supposed that NLP solutions that perform well for Icelandic may also perform well for Faroese, especially data-driven applications.

Some data already exists on the efficacy of cross-lingual transfer learning between Icelandic and Faroese. The FarParsald project (Ingason et al., 2014) focused on using a syntactically annotated corpus of Faroese, the Faroese Parsed Historical Corpus (FarPaHC; Sigurðsson et al. 2012), to train a syntactical parser, FarParsald, based on the data-driven Berkeley parser (Petrov et al., 2006; Petrov and Klein, 2007). The relatively small FarPaHC corpus, containing about 40,000 tokens, was supplemented with excerpts from its Icelandic counterpart, the one million word Icelandic Parsed Historical Corpus (IcePaHC; Rögnvaldsson et al. 2012). Using this approach, the overall parsing accuracy of FarParsald was raised from 75.44% to 78.06%, when 20% of the IcePaHC corpus, about 200,000 tokens, was added to the Faroese training data. In effect, a training set made of mostly Icelandic data returned better results than the Faroese-only data.

A similar approach may be taken in PoS tagging Faroese. ABLTagger (Steingrímsson et al., 2019), a recent Bi-LSTM driven PoS Tagger has shown impressive results in data-driven tagging of Icelandic. This implementation might well serve as a platform for further transfer learning between the two languages.

3 The MIM-GOLD tagging scheme

The Icelandic tagging scheme we use in our project is the MIM-GOLD tagging scheme, used in its eponymous corpus (Barkarson et al., 2020), a one million word, hand corrected corpus which serves as a gold standard for PoS tagging Icelandic. This tagging scheme is a modified version of the one used in the Icelandic Frequency Dictionary (IFD) corpus (Pind et al., 1991), with various revisions made to the tagset to improve and streamline machine tagging of texts.

In this tagging scheme, each token receives one PoS tag, constisting of a tag string. Each tag string consists of a series of characters, each having a particular morphosyntactic function, e.g., case, number, tense and grammatical gender. This is illustrated in Table 1, where the sentence in (1) is shown when tagged using the MIM-GOLD tagging scheme.

(1) Ég stökk á eftir strætó og veifaði.
I jumped on after bus and waved
'I jumped after the bus and waved.'

Token	PoS tag	Explanation

Ég	fp1en	**f**: pronoun; **p**: personal; **1**: 1st person;
		e: singular; **n**: nominative;

stökk	sfg1eþ	**s**: verb; **f**: indicative; **g**: active; **1**: 1st person; **e**: singular; **þ**: past tense
á	aa	**a**: adverb; **a**: doesn't govern case;
eftir	af	**a**: adverb; **þ**: governs case;

strætó	nkeþ	**n**: noun; **k**: masculine; **e**: singular;
		þ: dative;

og	c	**c**: conjunction;
veifaði	sfg1eþ	**s**: verb; **f**: indicative; **g**: active; **1**: 1st person; **e**: singular; **þ**: past tense
.	pl	**p**-punctuation, **l**-end of sentence

Table 1: A sentence tagged with the MIM-GOLD tagging scheme, with explanations.

4 The Sosialurin tagging scheme

The Faroese PoS tagging scheme we focus on is the one used in the Sosialurin corpus, devised by Hansen et al. (2004) as part of a larger project to create a PoS-tagged corpus for the language and train automatic PoS tagging software. This scheme is, to a large extent, based on the tagging scheme used in the IFD corpus for Icelandic (Pind et al., 1991). This was possible because of the many similarities between Icelandic and Faroese in morphology and grammar in general.

As in its Icelandic counterpart, the Faroese tagging scheme assigns each token a tag string, which contains a series of letters, each signifying relevant morphosyntactic information. The languages are not identical, however, and this is reflected in the Faroese tagging scheme. Furthermore, in a handful of grammatical categories, the Sosialurin tagging scheme encodes fewer details than the Icelandic one. In short, it is not as fine grained. Finally, the tagging schemes use different symbols in the tag strings themselves, rendering the tagging schemes superficially different. An example of the Sosialurin tagging scheme in practice is shown in Table 2, where the tokens of the sentence in (2) are shown with respective PoS tags.

(2) Hann er grivin undir Homrum.
he is buried under Hamrar
'He is buried at Hamrar.'

As discussed in Section 3 a number of revisions have been made to the IFD tagging scheme,

Token	Tag	Explanation
Hann	PPMSN	**P**-pronoun, **P**-personal **M**-masculine, **S**-singular, **N**-nominative
er	VNPS3	**V**-verb, **N**-indicative, **P**-present, **S**-singular, **3**-third person
grivin	VAMSN	**V**-verb, **A**-past participle, **M**-masculine, **S**-singular, **N**-nominative,
undir	ED	**E**-preposition, **D**-governs dative
Homrum	SMSDL	**S**-noun, **M**-masculine, **S**-singular, **D**-dative, **L**-location

Table 2: Example of the Sosialurin tagset, with explanations

Pronouns:	Added subcategories to tagstring
Adverbs:	Interjections and prepositions tagged as adverbs
Numerals:	New and reorganised subcategories
Abbreviations:	Subcategories for different types of abbreviations
Verbs:	Dedicated tag for supine removed
Nouns:	Place names and names of persons merged
Other:	New dedicated classes for punctuation and e-mail/web addresses

Table 3: Revisions applied to the Sosialurin tagging scheme based on the Icelandic *MIM-GOLD*.

mostly to improve tagging efficiency, culminating in the current MIM-GOLD tagging scheme for Icelandic. The same cannot be said about the Sosialurin tagging scheme, as no substantial revisions have been made to it since its inception. As such, we suggest a set of revisions to the Sosialurin tagging scheme, largely in step with the revisions made for the MIM-GOLD tagging scheme. These revisions are listed in Table 3.

The revisions applied to the Sosialurin tagging scheme include reworked numeral and punctuation tag strings, simplified case governance tagging for adverbs and the removal of a dedicated tag for past participles. Furthermore, various new tag strings were introduced, based on features from the original IFD tagging scheme which were omitted from the original Faroese scheme, e.g., distinction between different categories of pronouns.

In addition to the MIM-GOLD based revisions, we suggest a possible language-specific revision to the Faroese taggings scheme. This entails the removal of distinction between person (1st, 2nd or 3rd) from verb tags in the original tagset. In Faroese, person is never morphologically distinct in verbal plural forms, and may thus be reduntant in the tagging scheme, in theory. Such a revision would improve the accuracy of machine-tagging, but downstream effects, e.g., on syntactic parsing, are not clear. As such, we leave it as an open suggestion and do not apply it in our project.

With all revisions applied, the total number of theoretical tags in the Sosialurin tagset is about 600. When applied to the original Sosialurin corpus, 379 of these tags appear in the corpus, while the original corpus contained 390 unique tags. This is to be expected, mostly due to the simplified punctuation tags in the revised tagging scheme.

The revisions applied to the Faroese tagging scheme have been shown to positively affect overall PoS tagging accuracy. When applied to the Sosialurin corpus and evaluated using ten-fold cross validation, a Faroese implementation of ABLTagger achieved an overall error reduction rate of 7.51% (Hafsteinsson and Ingason, 2021).

5 Remaining tagging scheme differences

With the revisions based on MIM-GOLD, described in Section 4, to the Faroese tagging scheme, the function of the two tagging schemes has become markedly more similar. The remaining aspect separating the two are language-specific features of the two schemes, specifically concerning verbal PoS tags and the interpretation of article tags.

Both Icelandic and Faroese make a morphological distinction between two voices for verbs, the active and middle voices. The MIM-GOLD tagging scheme for Icelandic treats the verbal voice as a defining characteristic of all verbs. In the tag string, this is shown with the letter *g* for the active voice, and *m* for the middle voice. However, in the Sosialurin tagging scheme for Faroese, the verbal voice is instead treated as a verbal *mood*. This causes a discrepancy between the two tagging schemes, as the hierarchy of the verbal tag string is fundamentally different. A verb in Icelandic, tagged as being in the indicative mood, could either be in the active or middle voice. This is not possible in the Faroese tagging scheme, since the middle voice is considered a verbal mood; the hierarchical nature of the tag string doesn't allow two different mood labels.

The reason for this difference might be differences in the languages themselves. Although both Faroese and Icelandic exhibit what may be called a grammatical voice in verbs, the Faroese form is likely reduced compared to the Icelandic. In turn, the distinction between voice in Faroese verbs is not as fundamental as in Icelandic. With this in mind, the discrepancy as a whole may be tentatively circumvented in the tag conversion.

A more significant difference between the two tagging schemes concerns the article word class. The Icelandic tagging scheme tags uses a specific tag for definite articles, which reflects conventional analyses of Icelandic grammar, in which the free-standing definite article ‘hinn’ is classified as a distinct word class, with no indefinite article being used. This free-standing article is thought of as a literary device of irregular usage, with the more common suffixed definite article being in more general use. Conversely, Faroese uses both definite and indefinite free-standing articles; ‘tann’ and ‘hin’ as definite and ‘ein’ as indefinite, along with a suffixed definite article, like Icelandic (Þráinsson et al., 2004). Despite the apparent function of these words as articles within Faroese, these words are tagged as indicative pronouns in the Faroese tagging scheme, forgoing a distinct article tag altogether. Furthermore, this seems to be an inherent difference between the conventional analyses between the two languages, which discourages the approach of simply adding an article tag to the tagging scheme.

6 Conversion between tagsets

We suggest a partial solution to the effect of the inherent differences between the two tagging schemes, when converting between the two. Concerning the verbal tags, when converting from the Faroese tagging scheme to the Icelandic, all verb PoS tags *not* tagged as in the middle voice are mapped to equivalent verbal tags in the indicative mood, active voice. Faroese verbs tagged in the middle are, conversely, mapped to the indicative middle voice. The opposite is done when converting from Icelandic to the Faroese tagging scheme, with the information on mood being overwritten, in the case of verbs that are in the middle voice. This approach produces a one-to-one mapping between the two tagging schemes and mitigates the discrepancy between them. This is especially efficient when only converting from the Icelandic to Faroese, which suffices use in cross-lingual transfer learning, as described in Section 2.

Regarding the difference concerning the article class, further research is needed before an end result is settled on. The conversion between tagsets itself is not hampered by the absence of a distinct article tag in the Faroese tagging scheme, but it may have an effect when applying datasets with converted tagging schemes, e.g., in transfer learning. Future work will shed more light on this.

With this in mind, we have set up simple Python scripts which generate full tagsets for the tagging schemes and convert between the two. Furthermore, we have produced preliminary datasets for use in testing of cross-lingual transfer learning, based on the MIM-GOLD corpus for Icelandic, the tagset of which was used in the development of the conversion described above. The conversion scripts and training datasets are tentatively made available on GitHub[1] as products of this project.

7 Alternatives to conversion

Although we the main objective of the current project concerns the conversion between two tagging schemes, we are remain aware of the possibility of alternatives to this approach. One notable possibility would be to simply unify the two tagging schemes. With the modifications described in Section 4 applied to the Faroese tagging scheme, the two tagging schemes become near identical in function. If the end goal is to align one tagging scheme to the other, it begs the question whether a single tagging scheme would suit the needs of the two languages for use in NLP, e.g, by simply using the established Icelandic MIM-GOLD tagging scheme to describe both. The grammatical similarities between the two languages, discussed in Section 2 further supports this argument. However, as the remaining discrepancies between the tagging schemes suggests, this approach is at best inopportune. At the moment, the conventional analyses of the two languages differ in such a way that simply applying the Icelandic MIM-GOLD tagging scheme on Faroese text would be suboptimal. However, experimenting on this could be fruitful, and reconciling these differences in analysis at a future date may also be possible.

Circumventing the topic of the two tagging schemes discussed here, it should be noted that both Faroese and Icelandic have been described using the Universal Dependencies (UD) annotation scheme. Three UD corpora are available for Icelandic and two for Faroese, with considerable overlap in the production of the Faroese FarPaHC corpus and the Icelandic IcePaHC and Modern

[1] https://github.com/hinrikur/far-ice_corpora

corpora, each being converted to the UD format from existing datasets, as described by Arnardóttir et al. (2020). In this sense the two languages have already been described with a common annotation framework, although the UD annotating scheme is not strictly a dedicated PoS tagging scheme compared to the two tagging schemes used in our project.

8 Conclusion

We have described the process of conversion between the PoS tagging schemes of two grammatically similar languages, the Icelandic MIM-GOLD tagging scheme and the Faroese Sosialurin tagging scheme. Despite the two tagging schemes being functionally similar, they use separate ways to encode fine-grained morphological information on tokenised text. We described the differences between the two, along with revisions made to the Faroese tagging scheme, with the goal of streamlining automatic PoS tagging. We discussed grammatical differences between Faroese and Icelandic which result in minor discrepancies between the two tagging schemes and suggested a way to mitigate the effects of this when converting between the two. As a result, we produced a simple way to convert PoS tags between the languages. The results of our work have been made available for use, consisting of Python scripts for converting Icelandic and Faroese tagged corpora and preliminary converted training data, ready for application in cross-lingual NLP applications, with the end goal of it being of benefit in cross-lingual transfer learning.

References

Þórunn Arnardóttir, Hinrik Hafsteinsson, Einar Freyr Sigurðsson, Kristín Bjarnadóttir, Anton Karl Ingason, Hildur Jónsdóttir, and Steinþór Steingrímsson. 2020. https://www.aclweb.org/anthology/2020.udw-1.3 A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 16–25.

Starkaður Barkarson, Einar Freyr Sigurðsson, Eiríkur Rögnvaldsson, Hildur Hafsteinsdóttir, Hrafn Loftsson, Steinþór Steingrímsson, and Þórdís Dröfn Andrésdóttir. 2020. http://hdl.handle.net/20.500.12537/39 MIM-GOLD 20.05. CLARIN-IS, Stofnun Árna Magnússonar.

Hinrik Hafsteinsson and Anton Karl Ingason. 2021. Shared digital resource application within insular scandinavian. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 70–79.

Zakaris Svabo Hansen, Heini Justinussen, and Mortan Ólason. 2004. Marking av teldutøkum tekstsavni [Tagging of a digital text corpus].

Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Joel Wallenberg. 2014. Rapid Deployment of Phrase Structure Parsing for Related languages: A Case Study of Insular Scandinavian. In *Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 91–95.

Anna Björk Nikulásdóttir, Jón Guðnason, and Steinþór Steingrímsson. 2017. *Mál tækni fyrir íslensku 2018–2022: verkáætlun [Language Technology for Icelandic 2018-2022: Strategic Plan].*

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411.

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

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1977–1984.

Einar Freyr Sigurðsson, Anton Karl Ingason, Eiríkur Rögnvaldsson, and Joel C. Wallenberg. 2012. http://www.linguist.is/farpahc Faroese Parsed Historical Corpus (Far PaHC). Version 0.1.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM Tagger with a Morphological Lexicon and a Lexical Category Identification Step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168.

Höskuldur Þráinsson, Hjalmar P. Petersen, Jógvan í Lón Jacobsen, and Zakaris Svabo Hansen. 2004. *Faroese: An overview and reference grammar.* Føroya fróðskaparfelag, Torshavn.