CLARIN Annual Conference Proceedings

2024

Edited by

Vincent Vandeghinste and Thalassia Kontino

15 – 17 October 2024
Barcelona, Spain

Please cite as:
CLARIN Annual Conference Proceedings, 2024. ISSN 2773-2177 (online).
Eds. Vincent Vandeghinste and Thalassia Kontino.
Barcelona, Spain, 2024.

Improving Phrase Structure Parsing for Icelandic

Ingunn Jóhanna Kristjánsdóttir
University of Iceland
ijk4@hi.is

Hafsteinn Einarsson
University of Iceland
hafsteinne@hi.is

Anton Karl Ingason
University of Iceland
antoni@hi.is

Abstract

We present an improved state-of-the-art in phrase structure parsing for Icelandic, building on the Icelandic Parsed Historical Corpus (IcePaHC), a CLARIN resource, as well as previous milestones presented at the CLARIN Annual Conference in the past. The present parsing system utilizes the Stanford Stanza system as well as IceBERT, a freely available Icelandic BERT Model. We describe previous work, the different configurations used for the present work, as well as the setup that yielded the best outcome, an F1 score of 90.38%.

1 Introduction

We present a new phrase structure parser for Icelandic, based on the Stanford Stanza system (Qi et al., 2020), which achieves a new state-of-the-art performance for the task in this language.

When considering resources such as syntactic parsers, Icelandic is a language with a relatively few speakers and it has lagged behind other bigger language communities in terms of developing crucial infrastructure. At the turn of the century, Icelandic Language Technology was virtually non-exisitent (Loftsson et al., 2009; Rögnvaldsson, 2010), at a time when English had resources like the Penn treebank (Marcus et al., 1993) and associated software innovations. However, around 2010, limited support for basic resources such as taggers (Loftsson, 2008) and lemmatizers (Ingason et al., 2008) had emerged for Icelandic and in 2011 a manually corrected constituency treebank, the Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2011), was released, leading the way for further development of parsing resources, and since then substantial progress has been made, thanks to a systematic Language Technology Programme (Nikulásdóttir et al., 2022), supported by Icelandic Government, an effort which has led to a large number of new CLARIN resources. While Icelandic still lags behind the major languages (Rehm & Way, 2023), this most recent period, along with international developments in new techniques, has involved fruitful development of new parsing tools.

In this paper, we present the Icelandic Stanza Phrase Structure Parser, an experiment where we train Stanza on the IcePaHC treebank and reach an F1 score of 90.38%, an improvement over earlier experiments. The paper is organized as follows. In Section 2, we review some background on Icelandic parsing resources and previous parsers for Icelandic. In Section 3, we describe our evaluation of the system and how it compares to other systems. In Section 4, we report on the findings. Finally, Section 5 concludes.

2 Background

2.1 Icelandic parsing resources

The Penn Treebank (Marcus et al., 1993) was the most significant resource for starting the research program that develops constituency parsers and it remains a key test case for new parsers to this date. Yet, some of the same experts who developed this treebank moved on to an annotation scheme that improves on some of design decisions in the Penn Historical Corpora (Kroch & Taylor, 2000; Kroch et al., 2004). This includes a relatively more flat structure that abstracts away from ambiguities such as

PP-attachment in many cases. This latter family of treebanks became the basis for the annotation scheme of the Icelandic Parsed Historical Corpus (IcePaHC), released in 2011 (Rögnvaldsson et al., 2011, 2012; Wallenberg et al., 2011). While we focus here on constituency treebanks, it is worth noting that datasets annotated with the Universal Dependencies annotation scheme became available later. One such treebank is described in Jónsdóttir and Ingason (2020), and another one, created by converting IcePaHC to the UD format, is described in Arnardóttir et al. (2020).

2.2 Previous parsers for Icelandic

While most parsers are data-driven and trained on corpora, the first solution for Icelandic was a rule-based shallow parser, IceParser (Loftsson & Rögnvaldsson, 2007). The first attempts at full phrase structure parsing involved a combination of Icelandic and Faroese data in Ingason et al. (2014), in which a section of data from IcePaHC was combined with the Faroese Parsed Historical Corpus (FarPaHC) (Sigurðsson et al., 2012). However, the first parsing experiment that used the full dataset of IcePaHC was described in Jökulsdóttir et al. (2019) using the Berkeley Parser as the central piece of a parsing pipeline. In the following year, an improved version of this system was created, based on the Berkeley Neural Parser (Arnardóttir & Ingason, 2020). This system reached an 84.74% F1 score, aided by a multilingual Bert model, with a recall of 84.43% and a precision of 85.07%. Without using the word embeddings, the system reached an F1 score of 82.18%.

In the realm of rule-based systems, Þorsteinsson et al. (2019) introduced a parser based on a wide-coverage context-free grammar for Icelandic. Furthermore, parsers that focus on dependency parsing have been released in recent years, building on the conversion of IcePaHC to UD. This includes the experiment in Arnardóttir et al. (2023) as well as the Stanford Stanza UD Parser (Qi et al., 2020).

As having trained word embeddings for a particular language is useful for improving performance, it has been a limiting factor in earlier parsing experiments that a contextual word embedding model for Icelandic was not available. Therefore, it is important for current work on Icelandic parsing that an Icelandic BERT model has now been made available, named IceBERT (Snæbjarnarson et al., 2022). We make use of IceBERT in our setup. Snæbjarnarson et al. (2022) also carried out a constituency parsing experiment on the GreynirCorpus test set and reached an F1 score of 90.02%. Their finding is similar to ours but not directly comparable because of the different nature of the GreynirCorpus test set.

3 Evaluation

Before training the Stanza parser on the Icelandic phrase structure data in IcePaHC, we had to make sure that the data was in an appropriate format. This meant that we had to clean up any empty nodes in the trees, as the parsing task is separate from recovering the identity of silent elements and traces of syntactic movement. We furthermore split the IcePaHC treebank into a training set, development set and test set. IcePaHC consists of one million words in 73,012 matrix clauses and 80% of these clauses are used for the training set, 10% for the development set and 10% for the test set. IcePaHC consists of data from different centuries (dated 1150–2008) and to guarantee an even distribution in the three sets, every tenth part of the corpus is divided between them. All computations were performed on resources provided by the Division of Information Technology of the University of Iceland through the Icelandic Research e-Infrastructure project, funded by the Icelandic Centre of Research.

4 Results

The results of the evaluation are shown in Table 1, contrasted with the earlier experiment of Arnardóttir and Ingason (2020). The table indicates, for each experimental setup, whether the IceBERT model was used (and whether it was the basic IceBERT model or the larger IceBERT Large), whether fine-tuning was applied, and whether the parser was run with custom settings vs. default settings. In the table, the baseline configuration involves no use of a BERT model, no fine-tuning and an in-order as opposed to a top-down transition scheme.

The best overall F1 score was 90.38% in the case when we used IceBERT Base, fine-tuning and a top-down transition scheme (as opposed to the default in-order transition scheme). The parser achieved

a slightly higher recall rate in the setup that was the same except that IceBERT large was used. This is a substantial improvement of the previous experiment in Arnardóttir and Ingason (2020). It is clear that the addition of an Icelandic BERT model makes a big difference for improving the results. The results mirror Snæbjarnarson et al. (2022) in that the best results are found with IceBERT base rather than IceBERT large, a somewhat surprising outcome. Snæbjarnarson et al. (2022) hypothesize that "this would change if the model is trained to convergence or better hyperparameter tuning".

Parser	IceBERT	Fine-tune	Top-down	F1 score	Precision	Recall
Berkeley Neural Parser (Arnardóttir & Ingason, 2020)				84.71%	85.07%	84.43%
Stanza						
Baseline	–	–	–	84.40%	84.92%	83.87%
IceBERT, Top-down	+Basic	–	+	88.66%	89.21%	88.11%
IceBERT, FT	+Basic	+	–	90.09%	90.32%	89.86%
IceBERT, FT, Top-down	+Basic	+	+	90.38%	90.54%	90.22%
IceBERT-Large, Top-down	+Large	–	+	88.80%	89.15%	88.46%
IceBERT-Large, FT	+Large	+	–	90.23%	90.40%	90.05%
IceBERT-Large, FT, Top-down	+Large	+	+	90.29%	90.39%	90.17%

Table 1: Results of the parsing experiment

5 Conclusion

In this paper we have described a new parsing setup for Icelandic that uses a CLARIN-available treebank for training and achieves better performance than the earlier system of Arnardóttir and Ingason (2020). The best F1 score in our experiments was 90.38% in the case when we used the IceBERT Base model, fine-tuning and a top-down transition scheme (as opposed to an in-order transition scheme). This is similar to the reported parsing accuracy in Snæbjarnarson et al. (2022), 90.02%, and slightly better, but the results are not directly comparable because different corpora were used for the experiments.

References

Arnardóttir, Þ., Hafsteinsson, H., Jasonarson, A., Ingason, A., & Steingrímsson, S. (2023, May). Evaluating a Universal Dependencies conversion pipeline for Icelandic. In T. Alumäe & M. Fishel (Eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 698–704). University of Tartu Library. https://aclanthology.org/2023.nodalida-1.69

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Bjarnadóttir, K., Ingason, A. K., Jónsdóttir, H., & Steingrímsson, S. (2020, December). A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In M.-C. de Marneffe, M. de Lhoneux, J. Nivre, & S. Schuster (Eds.), *Proceedings of the fourth workshop on universal dependencies (udw 2020)* (pp. 16–25). Association for Computational Linguistics. https://aclanthology.org/2020.udw-1.3

Arnardóttir, Þ., & Ingason, A. K. (2020). A neural parsing pipeline for icelandic using the berkeley neural parser. *Proceedings of CLARIN Annual Conference*, 48–51.

Ingason, A. K., Helgadóttir, S., Loftsson, H., & Rögnvaldsson, E. (2008). A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). *Advances in Natural Language Processing: 6th International Conference, GoTAL 2008 Gothenburg, Sweden, August 25-27, 2008 Proceedings*, 205–216.

Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., & Wallenberg, J. C. (2014, May). Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 91–95). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/855_Paper.pdf

Jökulsdóttir, T., Ingason, A., & Sigurðsson, E. (2019). A parsing pipeline for Icelandic based on the IcePaHC corpus. *Proceedings of CLARIN Annual Conference*, 138–141.

Jónsdóttir, H., & Ingason, A. K. (2020). Creating a parallel icelandic dependency treebank from raw text to universal dependencies. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2924–2931.

Kroch, A. S., Santorini, B., & Delfs, L. (2004). *Penn-Helsinki Parsed Corpus of Early Modern English. CD-ROM. First Edition.* [Size: 1.8 million words.].

Kroch, A. S., & Taylor, A. (2000). *Penn-Helsinki Parsed Corpus of Middle English. CD-ROM. Second Edition.* [Size: 1.3 million words.].

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

Loftsson, H., Bjarnadóttir, K., Helgadóttir, S., Whelpton, M., & Ingason, A. K. (2009). Icelandic language resources and technology: Status and prospects. *Nordic Perspectives on the CLARIN Infrastructure of Common Language Resources*, 27–33.

Loftsson, H., & Rögnvaldsson, E. (2007). Iceparser: An incremental finite-state parser for icelandic. *Proceedings of the 16th Nordic Conference of Computational Linguistics (NoDaLiDa 2007)*, 128–135.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, *19*(2), 313–330.

Nikulásdóttir, A. B., Arnardóttir, Þ., Barkarson, S., Guðnason, J., Gunnarsson, Þ. D., Ingason, A. K., Jónsson, H. P., Loftsson, H., Óladóttir, H., Rögnvaldsson, E., et al. (2022). Help yourself from the buffet: National language technology infrastructure initiative on clarin-is. *CLARIN Annual Conference*, 109–125.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Rehm, G., & Way, A. (Eds.). (2023). *European language equality - A strategic agenda for digital language equality*. Springer. https://doi.org/10.1007/978-3-031-28819-7

Rögnvaldsson, E. (2010). Icelandic language technology: An overview. *Language, Languages and New Technologies: ICT in the Service of Languages. Contributions to the Annual Conference*, 187–195.

Rögnvaldsson, E., Ingason, A. K., & Sigurðsson, E. F. (2011). Coping with variation in the icelandic parsed historical corpus (icepahc). *Language Variation Infrastructure*, *3*, 97–112.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., & Wallenberg, J. (2012). The icelandic parsed historical corpus (icepahc). *LREC*, 1977–1984.

Sigurðsson, E. F., Ingason, A. K., Rögnvaldsson, E., & Wallenberg, J. C. (2012). Faroese parsed historical corpus (FarPaHC) 0.1 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/92

Snæbjarnarson, V., Símonarson, H. B., Ragnarsson, P. O., Ingólfsdóttir, S. L., Jónsson, H., Thorsteinsson, V., & Einarsson, H. (2022, June). A warm start and a clean crawled corpus - a recipe for good language models. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 4356–4366). European Language Resources Association. https://aclanthology.org/2022.lrec-1.464

Wallenberg, J., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2011). *Icelandic Parsed Historical Corpus (IcePaHC)* [Version 0.9].

Þorsteinsson, V., Óladóttir, H., & Loftsson, H. (2019). A wide-coverage context-free grammar for icelandic and an accompanying parsing system. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1397–1404.