A Neural Parsing Pipeline for Icelandic Using the Berkeley Neural Parser

Þórunn Arnardóttir
University of Iceland
Reykjavík, Iceland
tha86@hi.is

Anton Karl Ingason
University of Iceland
Reykjavík, Iceland
antoni@hi.is

Abstract

We present a machine parsing pipeline for Icelandic which uses the Berkeley Neural Parser and includes every step necessary for parsing plain Icelandic text, delivering text annotated according to IcePaHC. The parser is fast and reports an 84.74 F1 score. We describe the training and evaluation of the new parsing model and the structure of the parsing pipeline. All scripts necessary for parsing plain text using the new parsing pipeline are provided in open access via the CLARIN repository and GitHub.

1 Introduction

A parsed corpus has many applications but the process of making such a corpus manually can be long and time-consuming, making automatic parsers an appealing option. A corpus made with automatic parsing can never be as accurate as a manually corrected corpus but can produce a much larger corpus in a fraction of the time. The neural parsing pipeline introduced in this paper is practical within both linguistics and language technology. The pipeline can be used for parsing large amounts of texts and the resulting corpus can be used for research in Icelandic syntax, analyzing syntactic movements or changes in syntax over time. Within language technology, a parsing pipeline can be used for parsing sentences to decipher their meanings, creating a parsed corpus for training various software, and more.

The Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al., 2011; Rögnvaldsson et al., 2012) is a manually corrected Icelandic treebank which consists of one million words. By training the Berkeley Neural Parser using texts from IcePaHC, an Icelandic parsing pipeline, IceNeuralParsingPipeline (Arnardóttir and Ingason, 2020a), is created which takes in plain text and delivers the text parsed according to the IcePaHC annotation scheme. The parsing pipeline builds upon a previous pipeline released through CLARIN, IceParsingPipeline (Jökulsdóttir et al., 2019), which includes the Berkeley Parser, an older version of the Berkeley Neural Parser which is not based on a neural network. The new parsing model is a faster and more accurate parser, with an 84.74 F1 score compared to the 75.27 F1 score of the older parsing model. Crucially, the parsing pipeline is open-source, licensed under the MIT license and available at the Icelandic CLARIN repository[1] and GitHub.[2]

The structure of the paper is as follows. Section 2 focuses on previous work on Icelandic parsers and describes IcePaHC while Section 3 describes the Berkeley Neural Parser. In section 4, we describe the training process and the resulting model's accuracy is reported in section 5. Section 6 describes the updated parsing pipeline and finally, we conclude in section 7.

2 Related work and the Icelandic Parsed Historical Corpus

Language technology resources for Icelandic have grown in number over the last decade and as more resources become available, the making of further resources becomes more feasible. Two Icelandic non-data-driven parsers exist, IceParser (Loftsson and Rögnvaldsson, 2007), which was developed before

[1] http://hdl.handle.net/20.500.12537/17

[2] https://github.com/antonkarl/iceParsingPipeline

IcePaHC was created and is a shallow parser based on finite-state transducers, and Greynir's parser, which uses hand-written context-free grammar (CFG; Þorsteinsson et al., 2019). After IcePaHC became available, the first full phrase structure parser was developed, IceParsald, which utilizes the Berkeley Parser (Ingason et al., 2014) and is trained on IcePaHC. This parser was then further developed for use in the parsing pipeline discussed above. The Berkeley Parser preceding the Berkeley Neural Parser uses a probabilistic context-free grammar (PCFG; Petrov et al., 2006), which has been replaced by the state-of-the-art neural networks. Using neural networks instead of PCFGs or CFGs has resulted in more accurate parsers (Alberti et al., 2015; Chen and Manning, 2014), calling for an updated parsing pipeline using the Berkeley Neural Parser.

The data used for training the Berkeley Neural Parser is the Icelandic Parsed Historical Corpus (IcePaHC). It is a one-million-word, diachronic treebank which includes texts from the 12th to 21st centuries, distributed evenly (Rögnvaldsson et al., 2012). The texts included in IcePaHC have been manually corrected according to the Penn Parsed Corpora of Historical English (PPCHE) annotation scheme which uses labeled bracketing in the same way as the Penn Treebank. The part-of-speech tagset used is based on the one particular to the PPCHE, with minor changes to adapt it to Icelandic grammar. The annotation scheme employed in IcePaHC splits sentences into matrix clauses and marks their phrases. These phrases and their tokens are marked according to the tagset, which consists of 15 tags and their postdashial features.[3]

3 The Berkeley Neural Parser

The Berkeley Neural Parser is a constituency parser, relying on a repeated neural attention mechanism (Kitaev and Klein, 2018). The neural attention mechanism clarifies how information is transferred between different locations in a sentence. Different locations in a sentence can serve each other, both based on their positions and their contents, and these two types of attention are utilized in the parser with good results. Training the parser is possible when a gold corpus, such as IcePaHC, is available.

15 trained parser models for 11 different languages are distributed by the authors of the Berkeley Neural Parser, none of them being Icelandic. When training a model, several parameters can be used and tweaked. Three different word representation models are available: BERT, ELMo and fastText, BERT outperforming them both when used with the parser (Kitaev et al., 2019). BERT creates deep, contextualized word representations, delivering vectors for each word in the training set (Devlin et al., 2019). A few different BERT models are available, for example a multilingual model trained on 104 languages including Icelandic, but no models exclusively trained on Icelandic text have been published as of yet.

4 Training the model

IcePaHC consists of one million words in 73,012 matrix clauses. 80% of these clauses, 58,308 clauses, are used for training the parsing model, 10% for the development set and 10% for the test set, 7,302 clauses each. Since IcePaHC consists of data from different centuries (dated 1150–2008), an even distribution in the different sets is guaranteed by dividing every tenth part of the corpus between the training, development and test set.

Since IcePaHC tags are manually corrected, they can be used in training the model so that the parser predicts Part-of-Speech tags along with phrase structure. The cased multilingual BERT model is also utilized, having produced a more accurate parser for nine different languages (Kitaev et al., 2019). The neural network used for training consists of four layers, its learning rate is 0.00005, its batch size is 32 and the batch size during evaluation is 16. All computations were performed on resources provided by the Icelandic High Performance Computing Centre at the University of Iceland.

5 Evaluation

The parsing model is evaluated using EVALB, included in the Berkeley Neural Parser software. The 7,302 sentences of the test set are parsed using the parsing model and its output compared to the gold test set, delivering precision, recall, F-measure, complete match and tagging accuracy.

[3] A description of the tagset can be found at `https://linguist.is/icelandic_treebank/Tagset`

When trained using the multilingual BERT model, the parser achieves an 84.74 F1 score on the IcePaHC test set with 94.05% tagging accuracy. Its recall is 84.43%, its precision 85.07% and complete match is 46.61%. Some experiments were done with training a model without using word embeddings and using a different combination of training, development and test set. The highest accuracy reached without using word embeddings was an 82.18 F1 score. Changing the combination of data in the training, development and test set was also unfavorable. Using the oldest 80% of the data in the training set, the youngest 10% of the data in the test set and the 10% there in between in the development set, the model's accuracy dropped to a 77.01 F1 score. Reversing the data so that the youngest 80% of the data is in the training set, the oldest 10% in the test set and the 10% of data there in between in the development set proved to be more beneficial, reaching an 82.57 F1 score.

The parsing model is not only accurate, but also fast, as it is able to parse 228 sentences per second when run on NVIDIA Tesla V100 GPU on a Linux operating system. The model parses over five times the amount of sentences in the same amount of time when compared to the Berkeley Parser. When both parsers are run on two Intel Xeon E5-2680v3 CPUs on a Linux operating system, the Berkeley Neural Parser parses 4.8 sentences per second while the Berkeley Parser parses 0.85 sentences.

6 The Icelandic Neural Parsing Pipeline (IceNeuralParsingPipeline)

The parsing pipeline described in this article is based on iceParsingPipeline (Jökulsdóttir et al., 2019), with three changes. Instead of using Detector Morse for punctuation splitting, a tokenizer for Icelandic text, Greynir's Tokenizer (`https://github.com/mideind/Tokenizer`) is used for shallow tokenization. The software tokenizes the text, returning each sentence with its tokens separated. Using the tokenizer replaces two steps in the older pipeline. As mentioned, the Berkeley Neural Parser model is then used instead of the Berkeley Parser model in the parsing step.

The pipeline is structured as follows. First, the plain input text is tokenized and divided into sentences using Greynir's Tokenizer. Each sentence is then split up into matrix clauses using a matrix clause splitter, i.e. a sentence is split if a coordinating conjunction occurs. This step is illustrated in Figure 1, wherein a sentence has been split into two matrix clauses, shown in separate lines. The sentence translates to: "They believe in heaven and eternal life, and I'd preferably like to take up their religion", the coordinating conjunction *og* "and" marking the beginning of the second matrix clause. The splitting of sentences into matrix clauses is done because sentences in IcePaHC are divided into matrix clauses and each sentence parsed separately. The text is then parsed using the trained model. After having been parsed, the text is postprocessed in two steps. The first one consists of restoring dashes and removing extra labels and brackets created by the parser and in the second one, the text's format is changed. Before this step, each matrix sentence is displayed in a single line but to make the sentences more legible, they are formatted to conform to the IcePaHC format, wherein subphrases of a sentence are shown in separate lines.

Þeir trúa og himnaríki og eylífulífi,
og þeirra trúarbrögð vildi eg helst taka

Figure 1: A sentence split into two matrix clauses.

Because of the parsing model's speed, the pipeline can be used for parsing large corpora. Two treebanks have been created by using the pipeline, a historical one containing 2,7 million words, NeuralMIcePaHC (Arnardóttir and Ingason, 2020b), and a 500-million-word one containing mostly contemporary data, IceConTree (Arnardóttir et al., 2020).

7 Conclusion

In this paper, we have described an open-source parsing pipeline for Icelandic which includes a fast parser, the Berkeley Neural Parser. The parsing model is trained using the Icelandic Parsed Historical Corpus along with a BERT model and reports an 84.74 F1 score. Parsed text can be used in both language technology and research in syntax, in making a treebank or determining the phrase structure of a single sentence. In creating a parsing pipeline which can accept plain text and deliver its parsed counterpart, the parsing process is made accessible for those not specialized in computer science or linguistics and the parsing speed opens up the possibility of parsing large texts.

Acknowledgements

We would like to thank the anonymous reviewers for helpful comments as well as the High Performance Computing Centre at the University of Iceland for providing the resources to run our experiments.

References

C. Alberti, D. Weiss, G. Coppola, and S. Petrov. 2015. Improved transition-based parsing and tagging with neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1354–1359, Lisbon, Portugal, September. Association for Computational Linguistics.

Þ. Arnardóttir and A. K. Ingason. 2020a. IceNeuralParsingPipeline. CLARIN-IS, Stofnun Árna Magnússonar.

Þ. Arnardóttir and A. K. Ingason. 2020b. NeuralMIcePaHC. CLARIN-IS, Stofnun Árna Magnússonar.

Þ. Arnardóttir, A. K. Ingason, S. Steingrímsson, S. Helgadóttir, E. Rögnvaldsson, S. Barkarson, and J. Guðnason. 2020. The Icelandic contemporary treebank (IceConTree). CLARIN-IS, Stofnun Árna Magnússonar.

D. Chen and C. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October. Association for Computational Linguistics.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

A. K. Ingason, H. Loftsson, E. Rögnvaldsson, E. F. Sigurðsson, and J. Wallenberg. 2014. Rapid deployment of phrase structure parsing for related languages: A case study of insular scandinavian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

T. F. Jökulsdóttir, A. K. Ingason, and E. F. Sigurðsson. 2019. A parsing pipeline for Icelandic based on the IcePaHC corpus. In K. Simov and M. Eskevich., editors, *Proceedings of CLARIN Annual Conference 2019*, Leipzig, Germany.

N. Kitaev and D. Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July. Association for Computational Linguistics.

N. Kitaev, S. Cao, and D. Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, July. Association for Computational Linguistics.

H. Loftsson and E. Rögnvaldsson. 2007. IceParser: An incremental finite-state parser for Icelandic. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 128–135, Tartu, Estonia, May. University of Tartu, Estonia.

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

S. Petrov, L. Barrett, R. Thibaux, and D. 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, Sydney, Australia, July. Association for Computational Linguistics.

E. Rögnvaldsson, A. K. Ingason, E. F. Sigurðsson, and J. Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984, Istanbul, Turkey, May.

J. Wallenberg, A. K. Ingason, E. F. Sigurðsson, and E. Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. http://www.linguist.is/icelandic_treebankhttp://www.linguist.is/icelandic_treebank.