A Parsing Pipeline for Icelandic Based on the IcePaHC Corpus

Tinna Frímann Jökulsdóttir,[1] Anton Karl Ingason[1] and Einar Freyr Sigurðsson[2]
[1]University of Iceland and [2]the Árni Magnússon Institute for Icelandic Studies

UNIVERSITY OF ICELAND

THE ÁRNI MAGNÚSSON INSTITUTE for Icelandic Studies

1. Introduction

- Icelandic is a less-resourced language in the context of the CLARIN goals of fostering language resources and technology infrastructure.
  - Thus it is crucial to create further Icelandic resources that facilitate the development and use of Icelandic Language Technology.
  - Both for research as well as practical applications.
- We describe a novel parsing pipeline for Icelandic that makes crucial use of the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al., 2011; Rögnvaldsson et al., 2012).
  - A corpus containing one million running words of manually corrected phrase structure annotation.
- A parsed corpus can be used to train an automatic parser in order to get access to fast machine annotation of the same type.
- But even though a parsed corpus is freely available, a researcher in the humanities or social sciences may not have the required knowledge to train it or set up a parsing pipeline.
- This project aims to ameliorate this situation by developing a pipeline that
  1. takes Icelandic in plain text format,
  2. converts it to an appropriate input format for the Berkeley parser,
  3. parses the data using a pre-trained model that we provide,
  4. cleans up the output.
- Thus the pipeline makes it straightforward for any future projects to take advantage of machine-annotation according to the IcePaHC annotation scheme.

4. Training the Berkeley Parser

- We use the Berkeley Parser (Petrov et al., 2006) for Part-of-Speech tagging as well as for parsing phrase structure.
- Why?
  - Its split-merge algorithm is known to yield accurate results.
  - It is relatively simple to train on data that are already in a labeled bracketing file format (like IcePaHC).
  - There exists a version of it that runs fast on massively multi-core GPU cards (Canny et al., 2013; Hall et al., 2014).
- The training data: The full IcePaHC corpus.
  - The training data are in expected format; labeled bracketing.
  - So in theory, the training process should be straightforward.
  - In practice, the process is a bit more complicated - few minor adjustments to the file format needed to be done.
  - With this pre-configured parsing pipeline the user can avoid this extra technical work that requires some specific knowledge.

2. The IcePaHC corpus

- IcePaHC is a dual purpose project.
  - A language technology tool and a syntactic research tool.
- The annotation scheme is based on the Penn Parsed Corpora of Historical English (Kroch and Taylor, 2000; Kroch et al., 2004).
- For most purposes, the English annotation guidelines are applied without modification to Icelandic.
  - In fact, the same search query can often be used for studying the same phenomenon in both languages.
- Some minor Icelandic-specific adjustments to the annotation scheme:
  - Somewhat larger tagset and lemmatization.
  - A way to annotate non-nominative subjects.
- A concrete example from the corpus:

(1) Þar kom að þeim Danaher.
there came to them Danish army.
'There, the Danish army came toward them.'

The annotated version:

```
(2) ( (IP-MAT (ADVP-LOC (ADV Þar-þar))
          (VBDI kom-koma)
          (PP (P að-að))
              (NP (PRO-D þeim-það)))
          (NP-SBJ (NPR-N Danaher-danaher)))
                    (ID 1275.MORKIN.NAR-HIS,.522) )
```

- Even though this project focuses on phrase structure annotation it does not eliminate Universal Dependencies.
  - Another ongoing project aims to implement a conversion from IcePaHC annotation to Universal Dependencies.

5. Post-processing and cleanup

- The pipeline includes scripts that makes some minor adjustments to the output of the Berkeley parser.
  - Does not change the information that is included in the output from the parser.
  - Only makes its format more similar to what scholars who study historical syntax are used to.
- The format of the files that are used for the raw data of the Penn Parsed Corpora of Historical English has become well known by researchers in the field.
  - This format is both machine-readable and conveniently human-readable.

3. Matrix clause boundary detection

- A crucial step in parsing involves boundary detection.
  - Separating segments that have a privileged status in the sense that they correspond to one tree in the annotation scheme.
  - The IcePaHC annotation: In most cases the matrix clause.
- Two subtasks:
  1. Punctuation-based sentence boundary detection.
  2. Conjunction-based matrix clause boundary detection.
- For both steps:
  - A feature extractor for potential boundaries was configured.
  - IcePaHC was used to train an implementation of the Averaged Perceptron classifier (Freund and Schapire, 1999) by Kyle Gorman (2019) to detect actual boundaries.
- For punctuations: The default configuration for English.
- For conjunctions: Determine if they are an actual matrix clause boundary or if it is a different use of the relevant conjunction.
  - Two words preceding the conjunction and two words following considered: To check for potential morphosyntactic indicators.
  - The indicators in question: comma, finite verb, non-finite verb, a word in the nominative case, a word that has any non-nominative (oblique) case value.

(3) John and Joshua walked to the store and bought some groceries.

6. Next steps and future directions

- The pipeline is already available online: https://github.com/antonkarl/iceParsingPipeline
- While it is very useful, much remains to be done.
  - The configuration that we use for individual steps will be improved in future work to yield even better results.
    - E.g. training the matrix clause splitter and the parser.
  - Proper evaluation will also be an essential ingredient of any iterative improvements.
  - Evaluation of different parser configurations.
- We have also started work on a Machine-parsed IcePaHC (MIcePaHC)
  - A corpus that does not have the manual corrections of IcePaHC but can grow much faster in size.
  - We aim at releasing the first version in the near future.
- At some point, we would also like to embed the functionality of the pipeline into our treebank software, treebankstudio.org.