An Online Linguistic Analyser for Scottish Gaelic

Loïc Boizou [1], William Lamb [2]

[1]Centre of Computational Linguistics, Vytautas Magnus University, Kaunas, Lithuania

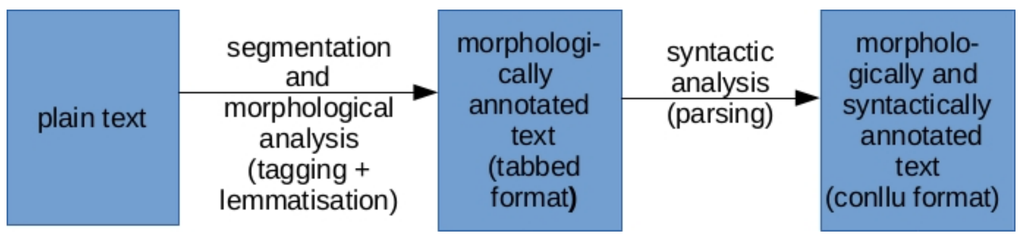
[2]University of Edinburgh, Scotland, United Kingdom

Scottish Gaelic: Current Situation

- Endangered Celtic language spoken by about 1% of the Scottish population.
- Current revitalisation efforts by different actors in various fields.
- Under-resourced language, but several key resources are available:
  - oral archives (www.tobarandualchais.co.uk/en/),
  - online dictionaries (www.faclair.com),
  - various corpora: DASG (dasg.ac.uk), ARCOSG (www.github.com/Gaelic-Algorithmic-Research-Group/ARCOSG), the UD Gaelic Treebank,
  - Google Translate.

Design Decisions for the Gaelic Linguistic Analyser (GLA)

- Building on free tools.
- Using ready-made resources to the largest possible extent to avoid work reduplication:
  - the morphologically annotated ARCOSG (www.github.com/Gaelic-Algorithmic-Research-Group/ARCOSG),
  - the UD Gaelic Treebank [1],
  - a lexicon provided by Michael Bauer and Will Robertson (www.faclair.com).
- Making the GLA freely available online.

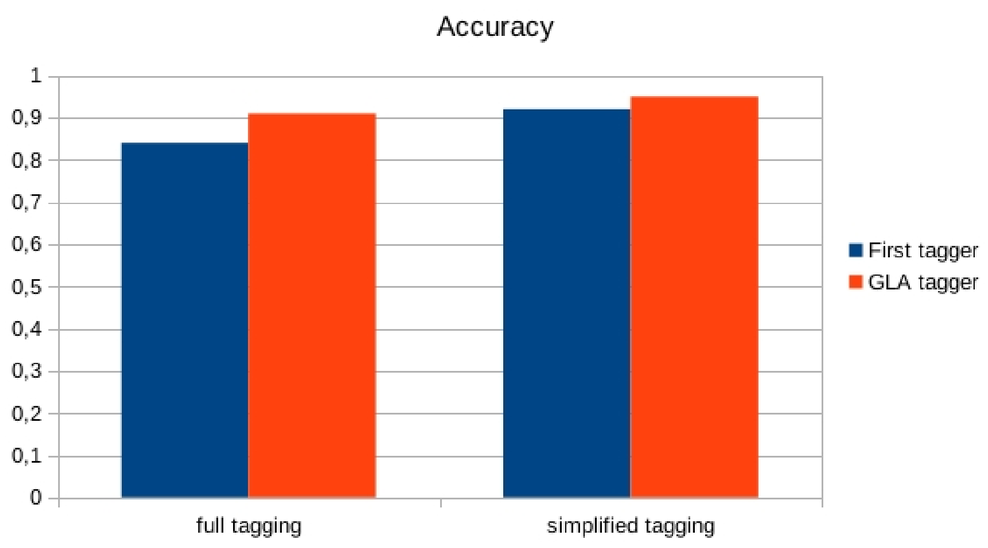


The Tagger

- The first Scottish Gaelic tagger developed by Lamb and Danso[1], [2] around 2013-2014. The tagger was trained on ARCOSG.
- The new tagger also uses the full ARCOSG as well as a few recent extra files as a training material (105,456 tokens in total)
- The model was trained on 96.6% of the corpus: one sentence in 20 was randomly picked for evaluation to ensure that all of the genres present in ARCOSG appear in the evaluation set.
- The tagger was developed in Python3 using the CRF method of the ML scikit-learn library (scikit-learn.org).
- Like the first tagger, the GLA tagger is provided with two options: with the full tagging and with a simplified tagging (with less categories).

Selected CRF features for each token:

1. original and lowercase word-forms,
2. prefix and suffix up to three letters,
3. information about symbols used in the token (e.g. capitals, numbers, hyphens, non-Gaelic letters),
4. position in the sentence (initial, final, intermediate),
5. the two previous and following tokens in the sentence.



The Lemmatiser

- The first tagger had no lemmatiser included, therefore it was important to add a lemmatiser to the GLA.
- Two testing lemmatisers were developed: 1) rule-based, 2) lexicon-based. Both of them used the results of the tagger to lead the lemmatisation process.
- The lexicon-based lemmatiser was selected as the GLA lemmatiser, since it avoids generating non-existing lemmas.
- The initial lexical list provided by Michael Bauer and Will Robertson amounts to 177,000 word-forms associated with their lemmas and parts of speech. For the lemmatisation process, it was restructured as a dictionary (or 'letter') tree simulated through a Python dictionary.
- There is no golden standard for lemmatisation (ARCOSG is not lemmatised yet), that is why the proper evaluation is not yet provided.

The Parser

- The GLA parser is a simple combination of the UD Scottish Gaelic Treebank[1] made by Colin Bachelor and the UDPipe Python Library[4].
- It operates on the morphologically analysed data (provided by the tagger and the lemmatiser) converted in the conllu format.
- The selected model was trained with the link2 algorithm of UDPipe, which obtained the best results (UAS: 97.11%, LAS: 96.40% on the training data, as evaluated by UDPipe).
- Given the scarcity of the data, the whole available treebank was used as a training material. As a consequence, a proper evaluation of the parser is still needed.

The Web Portal

- The GLA is the first component of the Scottish Gaelic Toolkit (SGT), which is accessible at the following adress: https://klc.vdu.lt/sgtoolkit/.
- The website is fully bilingual in Gaelic and English.
- The website is based on a Python server solution that relies on Flask (https://flask.palletsprojects.com) and Gunicorn (https://gunicorn.org/)
- It provides access to the GLA through a text area window, where Gaelic sentences can be written or pasted, or through a web service with a POST request.

References

[1] Batchelor C. Universal dependencies for Scottish Gaelic: syntax. In: Lynn T, Prys D, Batchelor C, Tyers F, editors. Proceedings of the Celtic Language Technology Workshop; 2019 August 19, Dublin, Ireland. European Association for Machine Translation; c2019. p. 7-9.

[2] Lamb W, Danso S. Developing an Automatic Part-of-Speech Tagger for Scottish Gaelic. In: Judge J, Teresa Lynn T, Ward M, Ó Raghallaigh B, editors. Proceedings of the First Celtic Language Technology Workshop; 2014 August 23, Dublin, Ireland. Association for Computational Linguistics and Dublin City University; c2014. p. 1-6.

[3] Lamb W, Danso S, Lawson A. Evaluating a Gaelic Part-of-Speech Tagger and Reference Corpus [Internet]. 2016. Available from: https://www.academia.edu/26589071/Evaluating_a_Gaelic_Part-of-Speech_Tagger_and_Reference_Corpus.

[4] Straka M, Straková J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Hajič J, Zeman D, editors. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies; 2017 August, Vancouver, Canada. Association for Computational Linguistics; c2017. p 88-12.

lboizou@gmail.com w.lamb@ed.ac.uk