Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200612

An Online Linguistic Analyser for Scottish Gaelic

Loïc BOIZOU [a] and William LAMB [b,1]

[a] *Centre of Computational Linguistics, Vytautas Magnus University, Lithuania*
[b] *Celtic and Scottish Studies, University of Edinburgh, Scotland, UK*

Abstract. This paper describes the Gaelic Linguistic Analyser, a new resource for the Scottish Gaelic language. The GLA includes a tagger, a lemmatiser and a parser, which were developed largely on the basis of existing resources. This tool is available online as the first component of the Scottish Gaelic Toolkit.

Keywords. Scottish Gaelic language, lemmatiser, tagger, parser, web portal, language technology

1. Introduction

This paper presents a new resource for the under-resourced Celtic language, Scottish Gaelic (henceforeth, Gaelic). Although only spoken by about 1 % of the Scottish population, there have been efforts to revitalise Gaelic recently, in domains such as culture, media, education and government. Hereditary Gaelic communities are endangered [1], but the language's profile has been boosted of late by its inclusion in a major television program (Outlander) and within certain language technology platforms (e.g. Duolingo and Google Translate). As native domains contract, it has become crucial for the language's long-term survival to develop data and cornerstone tools to support its use in technologically mediated ones.

In the field of Natural Language Processing, Gaelic is less advanced than two related languages, Irish [2,3] and Welsh [2]. However, some key resources are available, such as oral archives [3], online dictionaries [4], and corpora such as DASG [5], ARCOSG [6] and the UD Gaelic Treebank. The Gaelic Linguistic Analyser (GLA) – which combines a tagger, a lemmatiser, and a parser – strongly relies on these resources. Sharing existing resources widely and building new tools on top of them is of crucial importance for lesser-resourced languages, to minimise reduplication and fully exploit work already done.

[1] Corresponding Author: William Lamb; University of Edinburgh, 50 George Square, Scotland, UK, EH8 9LH; E-mail: w.lamb@ed.ac.uk.
[2] http://techiaith.cymru/cloud/api/parts-of-speech-tagger-api/?lang=en
[3] www.tobarandualchais.co.uk/en/
[4] www.faclair.com
[5] sg.ac.uk
[6] www.github.com/Gaelic-Algorithmic-Research-Group/ARCOSG

2. The Part-of-Speech Tagger

Lamb and Danso [4] developed the first tagger for Gaelic using a small part of ARCOSG, later extending it with the full corpus [5]. The present analyser uses the whole of ARCOSG for training and evaluation, in addition to further training files (105,456 tokens in total). The tagger was developed in Python3 using the ML scikit-learn library. The model was trained on 96.6 % of ARCOSG. One sentence in 20 was randomly picked for evaluation to ensure that all of the genres present in ARCOSG appear in the evaluation set.

For each word-form, the following features were included in the conditional random field model: 1) the original word-form and the lowercase one; 2) the prefix and suffix up to three letters; 3) information about symbols used in the word-form (e.g. capitals, numbers, hyphens, non-Gaelic letters); 4) the position in the sentence (initial, final, intermediate); and 5) the two previous and following word-forms in the sentence.

The accuracy is currently 0.907 (cf. 0.84 in [5]). Following the experience of [5], we retrained the tagger on the same corpus, but with a restricted set of tags (41 tags versus 246 in the full tagset), and achieved an accuracy of 0.947 (cf. 0.92 in [5]). The simplified tagger avoids the features that are difficult to grasp from the context in Gaelic, such as gender and case, for users who require less morphological granularity and desire a higher accuracy.

Except for the size of the training corpus, tagging mistakes come from different sources. As specific issues for Gaelic, we should mention the high frequency of English words mixed in Gaelic sentences (cf. [6]). We should also point inconsistencies in Gaelic orthography in the training data (cf. [7]). Among technical issues, a discrepancy exists between the segmentation used by ARCOSG, with many multiword tokens, and the general tokeniser used for segmenting input text, which is based mainly on orthographic words. We intend to address this issue in future works.

3. The Lemmatiser

The GLA lemmatiser is the first publicly presented for Gaelic. During the development stage, we instantiated two versions: a rule-based one and a lexicon-based one. Evaluation work is ongoing, but the lexicon-based one has two benefits at present: 1) the mistakes seem more acceptable, since it avoids impossible word forms and 2) the lexicon (courtesy of Michael Bauer and Will Robertson of www.faclair.com) was almost ready-to-use; it contains about 177,000 word-forms associated with their lemmas and parts of speech.

The main challenge was to locate an efficient searching algorithm. We began with a standard dictionary (or 'letter') tree – an alphabetically sorted binary tree – where node values are letters and lemmas at the current position in the tree. Nonetheless, the time for loading the tree was unacceptably high because of its recursive structure. It was possible to correct the problem with suspending the garbage collector, but this approach was sub-optimal. Therefore, we decided to apply the dictionary tree principle to a standard Python dictionary, using letters as sorted key, as in the following example:

```
{ 'a': { 'b':  { 'a': {...} },
                 ...
                 'lemmas': [('aba', 'N')]
```

```
            },
        'c':  { 'a': {...} },
              ...
            },
        ...
        'lemmas': [('a', 'P')]
      },
  'à': {...},
  'b': {...},
  ...
}
```

In the dictionary, letters at the same depth are considered alternatives, e.g. the first letter of a word, which appears in the first column of letters, can be *a*, *à*, *b*, and so on. Then the value of the key is to be understood as the group of possible letters at the next position (here, the second position), e.g. an initial *a* can be followed by *b* or *c* (for *ab...* or *ac...*), then *a* and *b* can be followed by *a*. The resulting string (*aba*) is a word-form that corresponds to the (unchanged) noun lemma *aba*, which means 'abbot'. The one-letter string *a* is also a word-form: it corresponds to the third person possessive pronoun (lemma is identical *a*). As shown in the example, lemmas are grouped under the dictionary key 'lemmas'. The value is a list, since a word-form can be related to several lemmas of one or several parts of speech. In the current version of the lemmatiser, we return the first lemma matching the requested word-form with its assigned part of speech; this provides simple results with one lemma for each analysed word. In the future, we may return all possible lemmas or start disambiguating lemmas via context.

The letter dictionary is about 50 % slower than the dictionary tree in the worst cases, when searched items are at the end of the alphabet, but loading the structure is faster and does not require suspending the garbage collector.

The letter dictionary is the core of the lemmatisation process. Searching involves the word-form and the part of speech previously indicated by the tagger. It is completed by a lower-case search: if the original word-form is not found in the dictionary, a second attempt is performed with the lowercase word-form. Fused words, such as prepositional pronouns, are treated somewhat differently, in that a lemma is provided for each fused element, e.g., *annad* 'in you' is lemmatised as *ann* ('in') + *thu* ('you').

4. The Parser

The GLA parser is based on the ready-to-use Python UDPipe library [8]. The syntactic model was trained using UDpipe executable on the Gaelic UD treebank made by Colin Bachelor [9]. The parser accuracy was evaluated with the same tool while selecting different transition systems: with the link2 parser option – UAS: 97.11 %, LAS: 96.40 %, with the swap option – UAS: 97.10 %, LAS: 96.35 %, with the projective option – UAS: 92.95 %, LAS: 91.33 %. The link2 model, which gave the best accuracy, is the one used by the parser.

These evaluations are meant to be compared together and not to other tools or data sets, since they are measured on the training set. This non-standard decision was motivated by the choice to keep all the limited data for training, because of their relative scarcity. We are aware that it will be necessary in the future to set aside control data for a proper evaluation.

5. The Web Portal

The GLA is the first component of the Scottish Gaelic Toolkit (SGT), which is accessible at the following adress: [7]. The website, which is fully bilingual in Gaelic and English, is based on a Python server solution that relies on Flask[8] and Gunicorn [9]. 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.

6. Final Remarks

Our objective with the GLA was to provide useful NLP tools online and demonstrate that these tools can run efficiently, if basic resources such as lexicons and annotated corpora are shared. A more comprehensive evaluation of the different tools, especially the lemmatizer and the parser is a future desideratum.

To increase functionality, soon we hope to provide an option to manually correct analyses. This should increase available training data via crowd-sourcing, although the pipeline required for integrating the extra data requires additional work.

References

[1] Ó Giollagáin C., Camshron, G., Moireach, P., Ó Curnáin, B., Caimbeul, I., MacDonald, B. and Péterváry, T. The Gaelic Crisis in the Vernacular Community. 2020. Aberdeen: Aberdeen University Press.

[2] Lynn T, Scannell K, Maguire E. Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets. In: Xu W, Han B, Ritter A, editors. Proc. of the First Workshop on Noisy User-generated Text; 2015 July 31, Beijing, China. European Association for Machine Translation; c2015. p 1-8

[3] Lynn T, Scannell K. Code-switching in Irish Tweets: A Preliminary Analysis. 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. 32-9.

[4] 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.

[5] 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.

[6] Smith-Christmas C. Stance and Code-Switching: Gaelic-English Bilinguals on the Isles of Skye and Harris, Scotland. In: Auer P, Caro Reina J, Kaufmann G, editors. Language Variation – European Perspectives IV. 2011 June; Freiburg. Amsterdam: John Benjamins; c2013. 229–17

[7] Ross S. The standardisation of Scottish Gaelic orthography 1750-2007: a corpus approach [PhD thesis]. [Glasgow]: University of Glasgow; 2016. 265 p.

[8] 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.

[9] 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.

[7] https://klc.vdu.lt/sgtoolkit/

[8] https://flask.palletsprojects.com

[9] https://gunicorn.org/