Developing An Auto-Glosser for Scottish Gaelic Using a Corpus of Interlinear Glossed Text

1 Introduction

Interlinear Glossed Text (IGT) is widely used in linguistic studies. (1) is an example of Scottish Gaelic IGT.

(1) Tha a athair nas sine na a mhàthair.
be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother
'His father is older than his mother.'

IGT is essential in linguistic research and analysis of a language's grammar. However, building large IGT corpora of under-resourced languages is expensive and time consuming. Endeavors to develop resources and tools for under-resourced languages are ongoing (Raghallaigh and Měchura, 2014; Lamb and Sinclair, 2016; Lamb and Danso, 2014).

In this paper, we present the first auto-glosser for Scottish Gaelic, which automatically generates the gloss line. We also introduce a corpus of Scottish Gaelic IGT from which the auto-glosser is trained. Additionally, we show that this glossing data significantly improves the performance of the machine translation systems that we are currently developing.

2 The Auto-Glosser

The Hidden Markov Model (HMM) is used widely and successfully in part of speech tagging tasks (Kupiec, 1992). We treat the glosses as special part of speech tags, and build the auto-glosser using HMM. Specifically, the gloss of a target Gaelic word is inferred by considering the relationship of the target Gaelic word to the predicted glosses of the two preceding Gaelic words. Consider the following example:

(2) ... word_1 word_2 word_3 ...
... gloss_1 gloss_2 ??? ...

To determine the gloss of word_3, the glosser selects the gloss that is most likely in the given context of word_3, gloss_1 and gloss_2. The accuracy rate of this auto-glosser is 65.8%, with 3986 possible glosses.

The primary goal of the auto-glosser is to facilitate the glossing process. We continue to expand the IGT corpus by incorporating data collected during a language documentation project. The auto-glosser aids and expedites the glossing task by generating a draft of Gaelic IGT, which a team member then corrects and verifies.

To understand the utility of the auto-glosser, we consider the IGT in (1) again. Specifically, only the first line in (1) is independent; the second line is determined by the first line. If the line one in (1) is our newly collected data, it will be glossed by the auto-glosser first, and the output is (3). Our team member then reviews and corrects (3) into (4), and then provides a free-translation of the Gaelic sentence into English.

(3) Tha **a** athair nas sine na a mhàthair.
be.pres **det** father comp old.cmpr comp 3sm.poss mother

(4) Tha **a** athair nas sine na a mhàthair.
be.pres **3sm.poss** father comp old.cmpr comp 3sm.poss mother
'His father is older than his mother.'

In this manner, the task is machine-aided, and we do not need to gloss from scratch.

3 Description of the Corpus

The essential component for any machine learning systems is a sizable and accurate training data. The key of the auto-glosser is our corpus of Scottish Gaelic IGT. The corpus has 8,367 Gaelic sentences, and in term of words, it has 52,778 Gaelic words/glosses. The data of the corpus is from two different sources: fieldwork and data elicitation. IGT in our corpus is treated as parallel text, a format is commonly used in machine translation. Specifically, IGT is stored in three related plain text files: 1) language text in orthography, 2) sequences of glosses, and 3) free English translations.

4 Other Application of the IGT Corpus: Machine Translation System

Given that the IGT data can be viewed as parallel texts, they can be the used directly as training data for machine translation systems. Using our data and OpenNMT (Klein et al., 2017), we built neural network machine translation systems and ran several experiments comparing two types of systems. Both types of systems have the Gaelic transcription and English translation data; The critical difference is whether the gloss data is incorporated. To the best of our knowledge, no published machine translation system exists that incorporates gloss data. The performances of the systems with gloss data incorporated are a lot better than those built only with Gaelic and English data. By including the gloss data, the BLEU score is more than doubled (without gloss: 0.165; with gloss: 0.357). Moreover, the gloss-incorporated systems also outperform Google translation (a BLEU of 0.248).

The important implication of this application is that linguistics can be really relevant and beneficial for NLP.

5 Conclusion

We introduce a IGT corpus and its applications. The corpus and the source codes of the applications will be made freely available. These endeavors are meant to provide open, useful and usable tools and resources for Scottish Gaelic. Moreover, the applications of IGT demonstrate that natural language processing techniques and linguistics research can be mutually beneficial and informative.

References

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*, 2017.

J. Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

W. Lamb and S. Danso. Developing an automatic part-of-speech tagger for scottish gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, pages 1–5, 2014.

W. Lamb and M. Sinclair. Developing word embedding models for scottish gaelic. *PARIS Inalco du 4 au 8 juillet 2016*, page 31, 2016.

B. O. Raghallaigh and M. B. Měchura. Developing high-end reusable tools and resources for irish-language terminology, lexicography, onomastics (toponymy), folkloristics, and more, using modern web and database technologies. *CLTW 2014*, page 66, 2014.