Proceedings of the CLTW 4 @ LREC2022 , pages 94–98
Marseille, 20-25 June 2022

Use of Transformer-Based Models for Word-Level Transliteration of the Book of the Dean of Lismore

Edward Gow-Smith[1], Mark McConville[2], William Gillies[3], Jade Scott[2], Roibeard Ó Maolalaigh[2]
[1]University of Sheffield, [2]University of Glasgow, [3]University of Edinburgh
egow-smith1@sheffield.ac.uk, Mark.McConville@glasgow.ac.uk

Abstract

The Book of the Dean of Lismore (BDL) is a 16th-century Scottish Gaelic manuscript written in a non-standard orthography. In this work, we outline the problem of transliterating the text of the BDL into a standardised orthography, and perform exploratory experiments using Transformer-based models for this task. In particular, we focus on the task of word-level transliteration, and achieve a character-level BLEU score of 54.15 with our best model, a BART architecture pre-trained on the text of Scottish Gaelic Wikipedia and then fine-tuned on around 2,000 word-level parallel examples. Our initial experiments give promising results, but we highlight the shortcomings of our model, and discuss directions for future work.

Keywords: Low-Resource Neural Machine Translation, Transformer-Based Models, Scottish Gaelic, Historical Manuscript

1. Introduction

As a material object, the Book of the Dean of Lismore (henceforth BDL) is a manuscript consisting of 159 paper folios, thought to have been assembled between 1512 and 1526 in eastern Perthshire, primarily by James MacGregor (c.1480–1551), the vicar of Fortingall and titular Dean of St. Moluag's Cathedral on Lismore (Thomson, 1993, 59–60). It is believed to have been acquired by James MacPherson, the Ossian 'translator', from a Portree blacksmith around 1760, and was handed over to the Highland Society of Scotland in 1803. It is now located in the National Library of Scotland (Adv.MS.72.1.37).

As an information object, the BDL is primarily an eclectic collection of traditional Gaelic poetry, including bardic, heroic and informal verse, by diverse authors, both professional and amateur, Scottish and Irish. Perhaps the most notable feature of the manuscript is that the Gaelic verse was not written in the traditional, morphophonemic Gaelic system of orthography but rather in a heterodox, semi-phonemic system based on the one used for writing Scots at that time. Consider, for example, the following two versions of the first line of p.128:

- Ne wlli in teak mir a hest a zramm a der a weit trane
- Ní bhfuil an t-éag mar a theist, a dhream adeir a bhith tréan

The first version is essentially the one that appears in BDL itself, and the second is a reconstruction of how this would have been written in the traditional Gaelic orthography of the time. Note the seventh word *hest*:*theist*. The initial consonant in this word would have been pronounced as the voiceless glottal fricative [h] and this is clearly reflected in the Scots-based orthography. However, the reconstructed Gaelic *th* includes a representation of the underlying morphophoneme T which is associated with (at least) two different phonemes – the fortis /t/ (written as *t*) and the lenis /h/ (written as *th*). The vowel in the final word *trane*:*tréan* is another example – the vowel here is the front mid [e:], represented in Scots orthography using the discontinuous digraph *a_e* and in Gaelic as the (non-discontinuous) digraph *éa*.

Over the last 100 years, attempts have been made to **transcribe** some of the poems in BDL (i.e. decode the handwriting) and then to **transliterate** these into some version of traditional Gaelic orthography, e.g. (Quiggin, 1937; Ross, 1939; Gillies, 1977; Meek, 1982). However, until recently an internally consistent transcription and transliteration of the full manuscript had not been attempted. Since BDL is an indispensable part of the textual foundation for the *Faclair na Gàidhlig* project, which aims to create a comprehensive dictionary of Scottish Gaelic on historical principles, this has now become a priority. This paper reports on the first two phases of this work: (a) the production of a consistent transcription of the full BDL; and (b) initial experiments in constructing an automatic transliterator from the Scots-based orthography into traditional Gaelic orthography using a small amount of parallel training data.

2. Data

The work on creating a consistent digital transcription of the whole of BDL was undertaken by the third and fourth listed authors. The first phase of this project involved digitally re-transcribing the manuscript transcription of BDL produced by Rev. Walter McLeod in 1893, when the BDL folios were in better physical condition than they are nowadays (NLS MS.72.3.12). Once this had been completed, a second iteration involved comparing this digital transcription with the handwriting in the BDL itself, in order to identify and

correct any apparent errors in McLeod's manuscript. (We are grateful to NLS for providing us with high-resolution digital images of both manuscripts.) In creating the digital transcription, a standard set of Unicode character points was used to encode non-ASCII glyphs in the BDL. In general, scribal contractions were not expanded. Some light markup was included for scribal insertions and deletions, and page and line numbers.

In order to provide some training data for our automatic transliterator, the third listed author provided reconstructed 'Dean's Text' transliterations for twelve of the poems in the BDL. Due to the small amount of data available, we decided to run experiments on word-level transliteration. Thus, the original transcriptions and reconstructed transliterations were aligned, where possible, at the word level. The majority of the data is word-to-word transliterated, but there are some cases where one word in the BDL is transliterated into multiple words in Scottish Gaelic, and vice versa, making up 7.4% of the data. A discussion of the shortcomings of this approach is given in Section 5.1. In total there were 1,962 examples, and 50 examples were randomly selected to give eval and test sets.

3. Experiments

We are interested in transliterating from the BDL to Scottish Gaelic (henceforth referred to as bdl-gd) and vice versa (likewise referred to as gd-bdl), although the first direction is of greater practical importance. Character-level BLEU score (Papineni et al., 2002) is used as an evaluation metric. We ran experiments on this task using Transformer-based models, implemented in Fairseq (Ott et al., 2019)[1]. For all experiments, tokenisation was performed at the character-level. The maximum sequence length was set at 20, to cover all of the available data whilst keeping computational requirements low. We also set the batch size at 1 due to the limited size of the training data, and the known problem of poor generalisation with large batch sizes (Keskar et al., 2016). For all of our models, the best performing model (by epoch) on the eval set was taken and evaluated on the test set. Full results are shown in Table 1, and in the rest of this section we discuss the various models and approaches used.

3.1. Parallel Data Only

Our first experiments were using just the available parallel data. We trained a Transformer (Vaswani et al., 2017) architecture with 2 layers and 2 attention heads for the encoder and decoder, and an embed dimension of 64, referred to as Transformer (tiny). We experimented with larger architectures, but found they were unable to learn from the available data. Our model was trained for 100,000 updates ($\sim$52 epochs), with a linear warm-up of the learning rate for 4,000 updates to 5e-4, then a linear decay to zero. We used the Adam optimizer (Kingma and Ba, 2014) with ϵ =1e-6, $\beta = (0.9, 0.98)$. On bdl-gd, this model achieved BLEU scores of 35.32 on the eval set and 41.16 on the test set. On gd-bdl, this model achieved BLEU scores of 30.17 on the eval set and 46.26 on the test set (Table 1).

3.2. Monolingual Pre-Training

The next approach was to utilise monolingual Scottish Gaelic data for the task, so that the model would hopefully learn something of Scottish Gaelic orthography. For this, we used the text of Scottish Gaelic Wikipedia[2], split to the word level, giving $\sim$600,000 words. We then pretrained BART (Lewis et al., 2019) architectures with the denoising task on this data. We first implemented a model with 2 layers, 2 attention heads, and embed dimension of 64 (referred to as BART (tiny) in reference to the Transformer model). We trained this model for 100,000 updates ($\sim$43 epochs). This model was then fine-tuned on the parallel training data, with the same hyperparameters as for Transformer (tiny). On bdl-gd, this model achieved BLEU score of 44.93 on the eval set, performing better than Transformer (tiny), and 38.64 on the test set, performing worse than Transformer (tiny). On gd-bdl, this model achieved BLEU scores of 21.04 on the eval set and 22.18 on the test set (Table 1), performing significantly worse than Transformer (tiny). It is expected that pre-training on monolingual Scottish Gaelic data will not be of help in this direction, but the significantly worse performance is surprising (see Section 4). We next tried the default BART (base) architecture, consisting of 6 layers, 12 attention heads, and an embed dimension of 768. On bdl-gd, this model achieved BLEU scores of 58.68 on the eval set and 53.32 on the test set, significantly outperforming Transformer (tiny). On gd-bdl, this model achieved BLEU scores of 36.17 on the eval set and 30.15 on the test set. We also ran the same model with additional pretraining, up to 400,000 updates ($\sim$172 epochs), which has been shown to be of benefit to other Transformer-based models (Liu et al., 2019). On bdl-gd, this model achieved BLEU scores of 62.47 on the eval set and 53.75 on the test set, showing an increase in performance on both. On gd-bdl, this model achieved BLEU scores of 36.77 on the eval set and 38.88 on the test set, also showing an increase in performance on both (Table 1). We also experimented with finetuning for longer (also 400,000 updates compared to 100,000), but this was found to lead to a general decrease in performance in both directions, although it did improve the performance on the eval set for gd-bdl (Table 1).

3.3. Data Augmentation

Next, approaches were taken at augmenting the available training data, a common approach in low-resource

[1]We release our data and scripts for running our experiments at https://github.com/edwardgowsmith/transliteration-book-of-the-dean-of-lismore.

[2]https://gd.wikipedia.org/

Model	bdl-gd		gd-bdl	
	eval	test	eval	test
Transformer (tiny)	35.32	41.16	30.17	**46.26**
BART (tiny)	44.93	38.64	21.04	22.18
BART (base)	58.68	53.32	36.17	30.15
BART (base) + p/t longer	**62.47**	53.75	36.77	38.88
BART (base) + p/t longer + f/t longer	59.46	52.09	**36.94**	34.68
BART (base) + p/t longer + homophones	59.60	**54.15**	34.75	31.77

Table 1: Character-level BLEU scores of the models on the eval and test splits. Best results are shown in bold.

neural machine translation (Haddow et al., 2021). Since we are interested in word-level transliteration, and thus a word may be transliterated into a homophone of the provided example with a different spelling (specifically, a heterograph), we took an approach to augment the training data with homophones. We used IPA information for Scottish Gaelic provided by English Wiktionary[3] - the data was parsed in order to find homophones for words it the training data. Unfortunately, IPA information was only available for a small number of items, which increased the training data from 1,862 to 1,938 examples (an increase of ∼4%). With the addition of this augmented training data, the BLEU score of BART (base) on the eval set decreased (from 62.47 to 59.60), but the BLEU score on the test set increased (from 53.75 to 54.15), which makes sense as the introduction of heterographs should allow the model to generalise better (although we note that the increase in performance is small). Interestingly, this model performs significantly worse in the reverse direction, with BLEU scores of 34.75 and 31.77 on the eval and test sets, respectively (discussed in Section 4). It should be noted that this approach assumes that heterographs in modern Scottish Gaelic were also heterographs at the time of the BDL, which should be a valid assumption. An alternative approach to augmenting the data would be to use a rule-based approach, which we leave to future work.

4. Discussion

In this section we discuss our results. From Table 1 we can see that, in general, the performance on gd-bdl is significantly worse than that on bdl-gd. This is to be expected, since the models have access to a large amount of monolingual Scottish Gaelic (gd) data, but BDL (bdl) is effectively an unseen language, which previous work has shown results in poor performance (see e.g. Üstün et al. (2021)). What is perhaps unexpected, however, is that our best-performing model on bdl-gd, BART (base) + p/t longer + homophones, performs significantly worse than the best in the opposite direction (31.77 compared to 46.26 on the test set). In fact, our best-performing model on gd-bdl, Transformer (tiny), does not use any monolingual Scottish Gaelic data. It seems likely that our models are overfitting on the train and eval sets, as a result of their small sizes. Attempts to avoid this could be made, including using multi-fold cross-validation. Additionally, it is hoped that we will have access to more parallel data in the future which will alleviate this problem, as well as the variance of performance across the eval and test splits.

4.1. Error Analysis

In this section, we perform an error analysis by taking our best-performing model and investigating which examples in the test set this model performed worse on (by character-level BLEU score). These are shown in Table 2. We note that these examples are relatively long; for shorter examples, our model generally performs better, which is typically expected but likely exaggerated in this case due to the increasing ambiguity of a word in the BDL as length increases. We note that our model struggles with spaces: no space is added when transliterating "eflay", and a space is erroneously added when transliterating "waiwill" (although the space is correctly removed when transliterating "dwgis i"). Since examples containing spaces on either the source or target side only make up a small amount of the parallel data, and the pretraining data contains no spaces, this is an expected area of difficulty, which we discuss further in Section 5.2. We also note that, out of the seven examples here, our model appears to output only three true Scottish Gaelic words ("mha fháil" meaning "if found", "chuaiseach" meaning "cavities", and "mhíos" meaning "month"). This is not necessarily a problem, since we want our model to be able to output unseen words, for example old-fashioned spellings and proper nouns. However, contextual information may help to determine the validity of a given transliteration, though the limited data available may prove to limit the efficacy of such an approach. Interestingly, the model transliterates "di" as the "[UNK]" token, which is problematic.

4.2. Learning of Scottish Gaelic Spelling Rules

We note that all of the outputs from our best model are *plausible* words, in that they obey the spelling rules of Scottish Gaelic. This is not the case for the Transformer (tiny) model trained only on the parallel data — as an example "dwgis" is transliterated by this

[3] https://en.wiktionary.org/

model into "duigas", which is not an acceptable Scottish Gaelic word, since a medial consonant must be surrounded by vowels of the same type (Gillies, 2009). This suggests that the training on monolingual data has allowed our model to learn the rules of Scottish Gaelic spelling, which has in turn improved performance on the transliteration task.

Input	Output	Reference
eflay	e'léamh	a' phláigh
dwgis i	duise	dtugas-sa
chotly^{t}sy^{t}	chuaiseach	chodlas-sa
wawaill	mha fháil	bhfaghbha'il
deinar̄	díonar	d'éinfhear
feanē	fén	phéin
zonicht	dhuanancht	dhona
di	[UNK]	do
gawe	gáimh	gabh
weiß^{t}	mhíos	bhíos

Table 2: The ten examples that our best performing model performed worse on for the test split (from bdl-gd).

5. Future Directions

Our preliminary experiments have shown promise in the task of transliterating the BDL, however there are many areas for improvement that we hope to address in future work.

5.1. Whole Sequence Transliteration

Since our work here is on word-level transliteration, it is unclear how this will extend to longer sequences, especially in the case of many-to-one transliteration. We take an example of transliterating a whole sequence with our model, shown in Table 3.

Input	A wēni^{t} za dwgis i grawġ
Output	a bhean dhá duis a' grádh
Reference	A bhean dhá dtugas-sa grádh

Table 3: Transliterating a whole sequence with our model.

In order to transliterate this whole sequence, we split it on whitespace and then pass each word individually to the model. Since, in this case, "dwgis i" is transliterated into a single word, our model cannot capture this (although note that this model fails to correctly transliterate these two words anyway (see Table 2)). An alternative approach to transliterating multi-word sequences may therefore be needed. Currently, due to our models being set at a max sequence length of 20, longer sequences cannot be directly given to the model.

5.2. Handling of Spaces

A related problem is the tendency of the models to struggle with handling spaces, both in the case of one-to-many and many-to-one transliteration. In order to help with this problem, it is likely we will need to include examples containing spaces during pre-training, or perform oversampling on the available training data to balance the number of examples with spaces and those without.

5.3. Data for Pre-Training

As stated in Section 3.2, we used data from Scottish Gaelic Wikipedia for pretraining, which is written in standardised modern Scottish Gaelic. For the purposes of our task, we are interested in generating transliterations which are faithful to the pronunciation at the time of the BDL. Hence, other data sources may provide more relevance for pre-training, such as Corpas na Gàidhlig[4] which contains transcribed texts dating back to the 17th century, and this is a direction of future work.

6. Related Work

There is no previous work, to the best of our knowledge, that uses Transformer-based models for tasks involving Scottish Gaelic. However, such approaches have been applied to other languages in the Celtic family: multilingual BERT (Devlin et al., 2019) contains Irish, Welsh and Breton in its training data, and there is a monolingual BERT for Irish (Barry et al., 2021) which was shown to outperform multilingual BERT on a dependency parsing test. There have been previous approaches at applying Transformer-based models to the task of word-level transliteration. Wu et al. (2021) applied the vanilla Transformer to the NEWS 2015 shared task (Zhang et al., 2015), outperforming previous models. Singh and Bansal (2021) also applied various sizes of Transformer architectures to the task of transliterating Hindi and Punjabi to English.

7. Conclusion

In this paper we discuss approaches to training Transformer-based models on the task of transliterating the Book of the Dean of Lismore (BDL) from its idiosyncratic orthography into a standardised Scottish Gaelic orthography. In particular, we outline our preliminary experiments training these models for word-level transliteration using both parallel word-level transliteration data for finetuning and monolingual Scottish Gaelic data for pretraining. Our best performing model was able to achieve a character-level BLEU score of 54.15 on the test set, showing significant promise, although there are many directions for improvement and future work, including extending this work to sequence-level (multi-word) transliteration.

8. Acknowledgements

This work was supported by *Faclair na Gàidhlig* and the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

[4] https://dasg.ac.uk/corpus

9. Bibliographical References

Barry, J., Wagner, J., Cassidy, L., Cowap, A., Lynn, T., Walsh, A., Meachair, M. J. Ó., and Foster, J. (2021). gabert–an irish language model. *arXiv preprint arXiv:2107.12930*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Gillies, W. (1977). Courtly and satiric poems in the Book of the Dean of Lismore. *Scottish Studies 21*.

Gillies, W. (2009). Scottish Gaelic. In *The Celtic Languages*, pages 244–318. Routledge.

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J., and Birch, A. (2021). Survey of low-resource machine translation. *arXiv preprint arXiv:2109.00486*.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Meek, D. (1982). *The Corpus of Heroic Verse in the Book of the Dean of Lismore*. Ph.D. thesis.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

E.C. Quiggin, editor. (1937). *Poems from the Book of the Dean of Lismore*. Cambridge.

Neil Ross, editor. (1939). *Heroic Poetry from the Book of the Dean of Lismore*. Scottish Gaelic Texts Society.

Singh, A. and Bansal, J. (2021). Neural machine transliteration of indian languages. In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, pages 91–96. IEEE.

Derick S. Thomson, editor. (1993). *The Companion to Gaelic Scotland*. Blackwell.

Üstün, A., Berard, A., Besacier, L., and Gallé, M. (2021). Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6650–6662, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, S., Cotterell, R., and Hulden, M. (2021). Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online, April. Association for Computational Linguistics.

Zhang, M., Li, H., Banchs, R. E., and Kumaran, A. (2015). Whitepaper of NEWS 2015 shared task on machine transliteration. In *Proceedings of the Fifth Named Entity Workshop*, pages 1–9, Beijing, China, July. Association for Computational Linguistics.