Leveraging backtranslation to improve machine translation for Gaelic languages

Meghan Dowling
ADAPT Centre
Dublin City University

Teresa Lynn
ADAPT Centre
Dublin City University

Andy Way
ADAPT Centre
Dublin City University

firstname.lastname@adaptcentre.ie

Abstract

Irish and Scottish Gaelic are similar but distinct languages from the Celtic language family. Both languages are under-resourced in terms of machine translation (MT), with Irish being the better resourced. In this paper, we show how backtranslation can be used to harness the resources of these similar low-resourced languages and build a Scottish-Gaelic to English MT system with little or no high-quality bilingual data.

1 Introduction

Irish (GA) and Scottish Gaelic (GD) are recognised minority languages, both in their native countries and in the EU. Both languages are minority languages, with English (EN) as the dominant language. Irish is also the first official language of Ireland and an official EU language. This benefits the Irish language in terms of MT resources because a certain amount of public information is required by law to be available in Irish, both at a national and European level[1]. Although Scottish Gaelic is recognised in the UK by the Gaelic Language Act (2005)[2], neither the UK government nor the EU are legally obliged to publish Scottish Gaelic texts. This has led to a shortage in available corpora suitable for training statistical machine translation (SMT) and neural machine translation (NMT) systems. Without the support of laws that require the output of Scottish Gaelic content, there is the risk that GD MT will not be able to reach the same status as other major languages.

As with other low-resourced and inflected languages, Gaelic languages suffer from data sparsity. While other language pairs can achieve high translation accuracy using state-of-the-art data-hungry methods, language pairs with fewer resources often have to employ creative methods to improve MT quality. One such approach is to create artificial data to boost the amount of corpora available for training. The premise of this method is that even if the data is not of a high quality, the MT system can still draw benefits from the extra data. Backtranslation is one such method for increasing the amount of creating artificial data. This paper describes our efforts to, through backtranslation, leverage the greater number of language resources available to Irish to improve MT systems for GD↔GA and GD↔EN.

This paper is laid out as follows: Section 2 and Section 3 give some background in terms of MT and linguistics. The data used in these experiments is detailed in Section 4. Section 5 describes the methodology employed in these experiments, the results of which are presented and discussed in Section 6. Finally, some avenues for future work are described in Section 7.

2 MT background

Data sparsity in low-resourced languages is exacerbated by the advent of NMT (Sutskever et al., 2014; Bahdanau et al., 2015), a data-hungry MT paradigm that requires huge amounts of parallel text to train a system of sufficient quality.[3] We be-

[1]There is currently a derogation in place within the EU which restricts the amount of content required to be translated to Irish. This is due to lift at the end of 2021.

[2]https://www.legislation.gov.uk/asp/2005/7/contents

[3]A marker of sufficient quality could be taken from Escartín and Arcedillo (2015) who indicate that a BLEU score of 45+ can increase translator productivity for EN-ES.

lieve that language technology resources are vital for the preservation and growth of every language and that it is necessary to develop methods of creating MT systems for languages without an extensive amount of language data available.

Previous experiments have shown backtranslation to be a viable method of artificial data creation (Sennrich et al., 2015; Burlot and Yvon, 2018; Poncelas et al., 2018). One possible benefit of backtranslation is that it allows the use of more than one MT paradigm (e.g. rule-based, statistical, neural) to create a MT model. In this way, the resulting model could gain benefits from each paradigm used.

Apertium (Forcada et al., 2011) is an open source machine translation platform which uses rule-based machine translation (RBMT) as the underlying MT technology. One of the benefits of RBMT is that it requires no parallel data, apart from a dictionary. There have been some efforts towards creating a RBMT system for GA↔GD. However, the GA↔GD Apertium module is listed as being in the *incubator* stage, which indicates that more work is needed before the MT system can be classed as being reliable.

There has been some previous work to create a GA→GD MT system with little or no data (Scannell, 2006). In this approach, the author builds a pipeline-style MT system which uses stages of standardisation, part-of-speech tagging, word sense disambiguation, syntactic transfer, lexical transfer and post-processing. There is also some literature surrounding the development of a SMT system for the GA–GD pair (Scannell, 2014). This approach involves training a word-based model, similar to the IBM model 1.

Research has been carried out on GD-EN NMT (Chen, 2018), in which the author uses linguistic features such as glosses to improve the system.

3 Linguistic overview

Translating between sentences with differing sentence structures can be a challenge for MT systems and can lead to poor quality MT output, particularly for longer sentences (Koehn and Knowles, 2017). Gaelic languages employ a verb-subject-object (VSO) sentence structure, different to the sentence-verb-object (SVO) structure more commonly seen in Indo-European languages. Figure 1 illustrates the similar word order of Scottish Gaelic and Irish, and how it diverges with that of English.

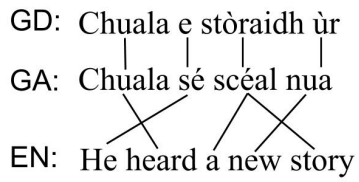

Figure 1: An example sentence highlighting the divergent word order between English and both Irish and Gaelic

Irish and Scottish Gaelic both display richer morphology than English. Example sentence 1 shows the inflection of the feminine nouns *'creag'* (GD) and *'carraig'* (GA), both meaning 'rock' or 'cliff'[4]. Inflection can have an impact on data sparsity (inflected words seen less frequently in training data) and also on automatic evaluation metrics such as BLEU (Papineni et al., 2002), which considers inflected words as being wholly different from their uninflected counterparts, and can sometimes penalise translation output too harshly as a result (Callison-Burch et al., 2006).

(1)	*creag*	rock/a rock	*carraig*
	*a' c**h**reag*	the rock	*an c**h**arraig*
	*creag**an***	rocks	carraig**eacha**
	*na cre**i**g**e***	of the rock	*na carraig**e***

4 Data

SMT and NMT, currently the two most prominent MT paradigms, require large amounts of bilingual data. Therefore, the availability of data plays a huge part in the quality of MT output. In this section we describe the GD and GA language data resources used in our experiments.

4.1 Scottish Gaelic

Wikipedia Scottish Gaelic language Wikipedia (Uicipeid[5]) contains 14,801 articles at the time of download [6]. Pre-processing including sentence tokenising, removal of wiki-text, tags and blank lines was performed, providing us with a resulting corpus of 87,788 sentences of monolingual Scottish Gaelic. This corpus can be described as being of mixed domain, with clear, formal sentences.

OPUS OPUS (Tiedemann, 2012) is a repository of language resources available for download from

[4]For clarity, the inflection markers (letters) in each example are displayed in bold

[5]`https://gd.wikipedia.org`

[6]04/04/2019

the web[7]. OPUS provides us with bilingual GA–GD and EN-GD corpora from a number of sources. Two bilingual GA–GD corpora that OPUS provides us with are the Ubuntu (655 parallel sentences) and GNOME (5,317 sentences) manuals. These are strictly within the technical domain, and often contain 'sentences' that are in fact 1-3 word phrases rich in technical jargon. Tatoeba, another OPUS source, is a corpus of short, simplified sentences for language learning purposes. While there was not a GD–GA Tatoeba corpus available, we downloaded the monolingual corpora for each language and manually aligned any matching sentences (referred to as Tatoeba-ga). OPUS also provides us with EN-GD parallel corpora from Tatoeba (Tatoeba-en), Ubuntu and GNOME.

4.2 Irish

In this work, we use the datasets described by Dowling et al. (2018). This consists of 108,000 parallel sentences from sources such as the Department of Culture, Heritage and the Gaeltacht and the Citizens Information website[8].

Corpus	# GA words	# GD words	# EN words
Uicipeid	N/A	1,449,636	N/A
Ubuntu	20,166	25,125	N/A
GNOME	14,897	19,956	N/A
Tatoeba-ga	466	489	N/A
Tatoeba-en	N/A	2,556	2,254
EN–GA	1,859,042	N/A	1,697,387
TOTAL	**1,894,571**	**1,497,762**	**1,699,641**

Table 1: Number of words in bilingual (GD-EN, GD-GA, GA-EN) and monolingual (GD only) corpora used

5 Method

In these experiments we take an approach to building an MT system using backtranslation illustrated by Figure 2. In step (1) monolingual data in language X (e.g. GA) is translated to language Y (e.g. GD) using the Apertium RBMT system. This creates an artificial parallel dataset. In step (2) this artificial dataset is then used to train a SMT system in the opposite language direction (e.g. GD→GA). (3) The resulting system can be used to translate new documents from language Y to language X.

[7] http://opus.nlpl.eu/
[8] https://www.citizensinformation.ie

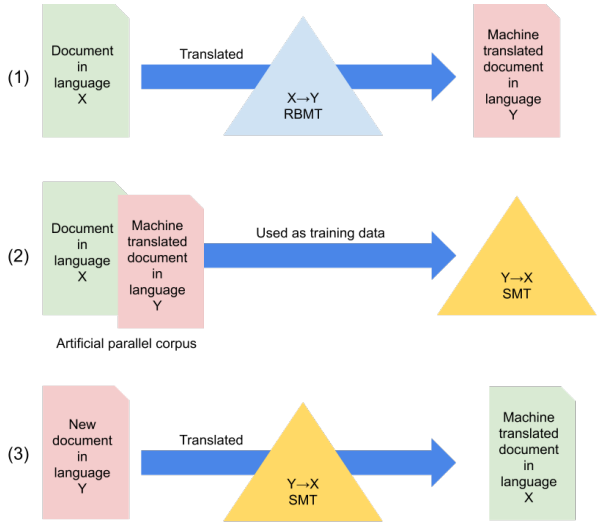

Figure 2: Simplified diagram of backtranslation method used to build SMT systems in these experiments

We carry out four sets of experiments (1, 2, 3 and 4) based on each language pair.

Experiment 1: GD→GA In these experiments (1A–C in Table 2), the Ubuntu and GNOME data sets are used as the authentic training data and the Uicipeid dataset is used as the basis of the artificial bilingual dataset (see Section 4).

Experiment 2: GA→GD To maintain consistency, the authentic dataset used in Exp. 1 is also used in these experiments (2A–C in Table 2). The bilingual artificial dataset is generated through backtranslation of the GA dataset used in previous EN-GA research, as described in Section 4.2.

Experiment 3: GD→EN With a relatively large EN-GA parallel dataset at our disposal, we chose to take this backtranslation method a step further. In these experiments (3A–C in Table 2), the GA side of the EN-GA dataset is translated to Scottish Gaelic using Apertium, as in 5. However, rather than pairing the machine translated Scottish Gaelic text with the authentic Irish text, we instead choose to train a system using the EN portion of the authentic EN-GA dataset. This results in a GD→EN SMT system.

Experiment 4: EN→GD The method of generating artificial corpora is identical to that of Exp. 3, with the exception of the change in language direction. The results for these experiments are presented as experiments 4A–C in Table 2.

5.1 Building and adding to the baseline

Each experiment contains three parts (referred to in Table 2). Part **A** involves creating a baseline by training a SMT system using only authentic data. Part **B** trains a SMT system using the artificial dataset created through backtranslation. This experiment most closely resembles Figure 2. Finally, in part **C**, the authentic and artificial datasets are combined to train a SMT system. Systems are trained using Moses (Koehn et al., 2007) with default parameters, with the exception of the GD↔EN systems which use a 6-gram language model and hierarchical reordering tables to partly address the divergent word order between the two languages.

6 Results and Conclusions

We report on BLEU (Papineni et al., 2002), an automatic metric of evaluating MT, to provide an indication of quality for the MT systems trained. For consistency in domain, the test data for all systems comes from the Tatoeba source. It should be noted that while the source is the same, Tateoba-ga and Tatoeba-en differ in both content and size (see Section 4).

Exp.	Auth.	Artif.	Lang.	BLEU
Apert.	N/A	N/A	GA→GD	8.67
1A	5,645	0	GA→GD	12.43
1B	0	87,788	GA→GD	16.63
1C	5,645	87,788	GA→GD	**25.45**
Apert.	N/A	N/A	GA→GD	13.73
2A	5,645	0	GD→GA	14.32
2B	0	108,000	GD→GA	17.46
2C	5,645	108,000	GD→GA	**22.55**
3A	18,785	0	GD→EN	3.73
3B	0	108,000	GD→EN	6.53
3C	18,785	108,000	GD→EN	**11.41**
4A	18,785	0	EN→GD	3.05
4B	0	108,000	EN→GD	7.03
4C	18,785	108,000	EN→GD	**10.59**

Table 2: BLEU scores for each experiment (Exp.), with the number of authentic (Auth.) and artificial (Artif.) sentences used to train each system. Scores are also given for the Apertium (Apert.) system used to generate the artificial data.

The results presented in Table 2 and Fig. 3 show a marked improvement in BLEU score over the baseline when backtranslated data is included as training data. We also include BLEU scores for the Apertium GA-GD module, generated through the translation of the test corpus Tatoeba-ga. Despite the low BLEU score of the Apertium GA-GD module, SMT systems trained using solely artificial data also show an increase in BLEU over the baseline. This indicates that even if the quality of the MT system used to backtranslate is poor, it may still be possible to gain benefits from the backtranslated data. The highest automatic scores from all 4 experiment series are produced when the authentic corpus is paired with the artificial data. It is interesting to note that while BLEU scores for the EN↔GD experiments (3A-4C) are substantially lower the same trend can still be seen. This could indicate that backtranslation is a usable method of artificial data creation, even with linguistically different language pairs such as EN–GD.

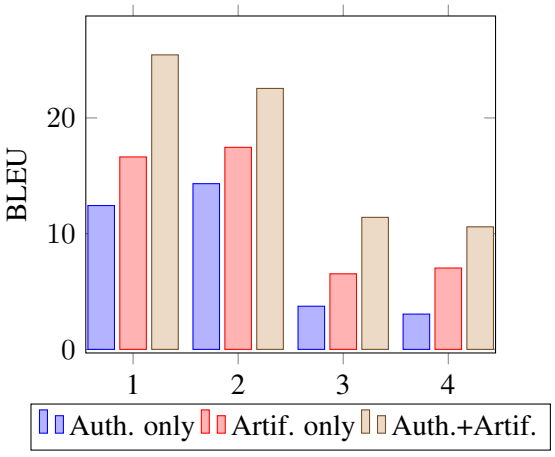

Figure 3: Bar chart of BLEU scores for each experiment. 1, 2, 3 and 4 refer to Experiments 1-4 in Section 5 and Table 2.

7 Future Work

In terms of future work, human analysis will be necessary to determine if there is an actual increase in quality (in terms of usabiltiy, fluency, etc.), rather than relying on automatic metrics.

Another possible avenue for future work is to use a system similar to that of Scannell (2006) to assess whether backtranslated data of a higher quality could be produced, presumably resulting in a more accurate MT output. Furthermore, while GA and GD are generally similar in sentence structure, there are a few cases where the two differ. It would be interesting to observe the standard of MT within these divergent situations and, if the standard is lower, investigate whether the inclusion of linguistic rules such as those used in Scannell (2006) could lead to an increase in quality.

We also note that Tatoeba is a corpus of sim-

ple, short sentences. It would be pertinent to repeat these experiments with test data from different domains to investigate if the same increase in BLEU is witnessed with other types of input.

Other monolingual data sources, such as Corpas na Gàidhlig[9] or Irish Wikipedia[10] could be used as sources for the creation of more backtranslated data. It would be interesting to view the effect of additional artificial data on the MT output. Moreover, if a large enough artificial corpus could be generated, these experiments could be repeated with NMT instead of SMT and compared to the research of Chen (2018).

It is our hope that this work could form a basis on which to extend to other Celtic languages and investigate whether it is useful for improving resources for similarly resourced languages.

Acknowledgements

This work is funded by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) at Dublin City University and partly funded by the Irish Government Department of Culture, Heritage and the Gaeltacht. We thank the reviewers for their valuable comments and advice.

References

Bahdanau, Dzmitry, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations*.

Burlot, Franck and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Chen, Yuan-Lu. 2018. *Improving Neural Net Machine Translation Systems with Linguistic Information*. Ph.D. thesis.

Dowling, Meghan, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20.

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

[10] https://ga.wikipedia.org

Escartín, Carla Parra and Manuel Arcedillo. 2015. Living on the edge: productivity gain thresholds in machine translation evaluation metrics. In *4th Workshop on Post-Editing Technology and Practice (WPTP4)*, page 46.

Forcada, Mikel L, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, page 249.

Scannell, Kevin P. 2006. Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, pages 103–109. Citeseer.

Scannell, Kevin. 2014. Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, pages 33–40.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*, volume 2012, pages 2214–2218.