Addressing Challenges of Indigenous Languages through Neural Machine Translation: The case of Inuktitut-English

Ngoc Tan Le
Université du Québec à Montréal
201, avenue du Président-Kennedy
Montréal, QC, Canada
le.ngoc_tan@courrier.uqam.ca

Fatiha Sadat
Université du Québec à Montréal
201, avenue du Président-Kennedy
Montréal, QC, Canada
sadat.fatiha@uqam.ca

Abstract

There is a growing amount of research interests towards Indigenous languages, realities and challenges within the NLP international community. Up do date, these Indigenous languages have been very challenging when dealing with many NLP tasks and applications because of multiple features such as polysynthesis, morphological complexity, dialectal variation with rich morpho-phonemics, spelling with noisy data and low resource scenarios. In this research, we systematically review the literature related to Machine Translation systems for Indigenous languages. Through this review, we make our focus on Inuktitut, one of the Indigenous polysynthetic languages spoken in Northern Canada. Experiments and evaluations on our first Inuktitut-English neural machine translation are conducted considering additional features extracted from the source-target alignment information and other bilingual lexicons.

1 Introduction

There is a great diversity of Indigenous languages, in America, about 900 different Indigenous languages spoken (Mager et al., 2018). Particularly, in Canada, Indigenous languages, which are identified in 12 linguistic families (Rice, 2011), have been at the heart of the history of First Nations, Métis and Indigenous community-oriented cultures and continue to play a vital role (Oster et al., 2014; Whalen et al., 2016; Coronel-Molina and McCarty, 2016). However, the development of Indigenous language technology faces many challenges such as polysynthetic with a high rate of morpheme per word, lack of orthographic normalization, dialectal variation and low resource (Littell et al., 2018). Inspired by statistical machine translation systems and their benefits (Koehn, 2009), with a relation to the source-target alignment information and the involvement of bilingual lexicons, this research aims at investigating our first Inuktitut-English neural machine translation. Experiments are conducted with additional features during the training and the decoding steps.

The balance of the paper is as follows: Section 2 presents the state-of-the-art on Machine Translation approaches for Indigenous languages. Section 3 presents our approach that is related to our first neural MT for one of these Indigenous languages. Section 4 highlights the significant results and section 5 concludes this research and provides some directions for future research.

2 Related work

The development of Machine Translation systems for Indigenous languages have followed the trends in the field, with (1) rule-based, (2) statistical-based and (3) neural network-based approaches. (1) Rule-Based Machine Translation (RBMT) approaches are popular and suitable for low resource languages. However, the main drawback is the training requires a lot of linguistic knowledge related to Indigenous languages (Mager et al., 2018). (2) In Statistical-based Machine Translation (SMT) approaches, for translating to and from morphologically complex languages, researchers have proposed treating words as sentences or subword units. Micher (2018) applied the Byte Pair Encoding (BPE) algorithm (Sennrich

et al., 2016) preprocessing both the English and Inuktitut sides of the Nunavut Hansard corpus, in the Inuktitut to English direction, reported a BLEU score of 30.35. (3) Neural network-based Machine Translation (NMT) approaches use neural networks architectures that are fed with very big amounts of parallel texts. However, these resources are currently unavailable in most Indigenous languages, except Inuktitut-English (Joanis et al., 2020).

3 Our proposed approach

The purpose of this paper aims to investigate our first Inuktitut-English neural machine translation system based on Transformer-based encoder-decoder architecture (Vaswani et al., 2017). Our approach consists of three main parts. First, in the preprocessing of training datasets, we deal with morphology complexity by applying unsupervised tokenization, such as Byte Pair Encoding (BPE) (Sennrich et al., 2016), for both source and target languages. Second, we incorporate source-target alignment information in the training step. We apply an unsupervised word aligner, fast_align (Dyer et al., 2013) to generate symmetrized source-target alignments, trained on BPE preprocessed data. Third, we inject, in the decoding, source-target morphological information, such as bilingual lexicon. We apply lexicon extractor from Moses (Koehn et al., 2007) to prepare a bilingual lexical shortlist which is passed to the decoder.

4 Experiments and Evaluation

In our experimental part, we train our NMT model by using the Nunavut Hansard for Inuktitut-English bilingual corpus (3rd edition) as described in Joanis et al. (2020). This corpus contains 1,293,348 sentences, 5,433 sentences and 6,139 sentences for training set, validation set and testing set, respectively.

In the preprocessing step, we use the Moses (Koehn et al., 2007) tokenizer to get tokens and apply subword-nmt (Sennrich et al., 2016) toolkit to create a BPE vocabulary with dimension of 30,000. We use Marian-nmt (Junczys-Dowmunt et al., 2018) to train our Transformer-based NMT with following hyper-parameters settings: 6-layer depth for both encoder and decoder, embedding dimension of 512, 2048 units in hidden layers in the feed-forward networks, optimizer with SGD, an initial learning rate of 0.0003. We run 50 iterations (#max_epochs) with an early stopping based on the cross-entropy scores for the validation set every 5,000 updates. We use 6-GPUs of NVIDIA GeForce GTX 2080 Ti 12Gb.

We evaluate our models by using the BLEU metric (Papineni et al., 2002) with lowercase and v13a tokenization, similare to (Joanis et al., 2020). The baseline represents the Transformer-based NMT with only BPE-preprocessed data. Then we experiment three others systems 1, 2 and 3, to optimize our results, by adding a source-target alignment information, a source-target bilingual lexicon and all additional features, respectively (Table 1). We observed that the lexical and alignment information brought more knowledge in the training and decoding phase. The performance of our NMT model can be improved by +1,03 of BLEU scores with all additional features, comparing the baseline. Our proposed approach had a positive impact on the performance of the NMT model.

Inuktitut → English	**Test set**	**Inuktitut → English**	**Test set**
SMT (Joanis et al., 2020)	27.80	Baseline + align information (System 1)	35.71
Baseline NMT (Joanis et al., 2020)	35.00	Baseline + lex.s2t (System 2)	35.93
		All (System 3)	**36.03**

Table 1: Comparison of our proposed approach for Transformer-based NMT, in terms of BLEU scores

5 Conclusion

In this paper, we presented our ideas on creating an Inuktitut-English NMT system thanks to the fast growing interest on Indigenous languages within the NLP community. Our review on NLP researches helped us to conclude that MT can be one of the best-known language technology to revitalize and preserve endangered languages. In the future, for scientific progress, we expect to conduct more research to enhance Indigenous language technology development. We also suggest to build a close collaboration with Indigenous community-driven organizations and Indigenous communities across Canada.

References

Serafín M Coronel-Molina and Teresa L McCarty. 2016. *Indigenous language revitalization in the Americas*. Routledge.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The nunavut hansard inuktitutâ"english parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France, May. European Language Resources Association.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Philipp Koehn, 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 association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Jeffrey Micher. 2018. Using the nunavut hansard data for experiments in morphological analysis and machine translation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 65–72.

Richard T Oster, Angela Grier, Rick Lightning, Maria J Mayan, and Ellen L Toth. 2014. Cultural continuity, traditional indigenous language, and diabetes in alberta first nations: a mixed methods study. *International journal for equity in health*, 13(1):92.

Kishore Papineni, 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.

Keren Rice. 2011. Documentary linguistics and community relations. *Language Documentation & Conservation*, 5:187–207.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Douglas H Whalen, Margaret Moss, and Daryl Baldwin. 2016. Healing through language: Positive physical health effects of indigenous language use. *F1000Research*, 5(852):852.