Improving Translation Quality for Low-Resource Inuktitut with Various Preprocessing Techniques

Mathias Hans Erik Stenlund, Matilde Nanni, Micaella Bruton, Meriem Beloucif
Uppsala University
meriem.beloucif@lingfil.uu.se

Abstract

Neural machine translation has been shown to outperform all other machine translation paradigms when trained in a high-resource setting. However, it still performs poorly when dealing with low-resource languages, for which parallel data for training is scarce. This is especially the case for morphologically complex languages such as Turkish, Tamil, Uyghur, etc. In this paper, we investigate various preprocessing methods for Inuktitut, a low-resource indigenous language from North America, without a morphological analyzer. On both the original and romanized scripts, we test various preprocessing techniques such as Byte-Pair Encoding, random stemming, and data augmentation using Hungarian for the Inuktitut-to-English translation task. We found that there are benefits to retaining the original script as it helps to achieve higher BLEU scores than the romanized models.

1 Introduction

While state-of-the-art Machine Translation (MT) systems are achieving close to human-like translations on a restricted set of highly researched languages (Luong et al., 2015; Sennrich et al., 2015; Luong and Manning, 2016; Neubig, 2015; Cho et al., 2014; Luong et al., 2017; Vaswani et al., 2017), they fail to obtain equally good results on languages for which there is a lack of resources (Haddow et al., 2022). In fact, these end-to-end neural encoder-decoder MT systems are quite data hungry, requiring parallel datasets in the tens or even hundreds of millions of sentences to outperform statistical models; datasets which are only available for a few of the spoken languages of the world (Ranathunga et al., 2021). The unavailability of parallel data for most world languages is only the tip of the iceberg because, even when there is data available, the data can be very domain-specific and contain a lot of noise (Haddow et al., 2022). Ranathunga et al. (2021) and Haddow et al. (2022) provide an overview of current research methods tackling low-resource MT, by addressing different aspects and problems. The data and tools scarcity problem in NLP creates the need to simulate low-resource scenarios by taking a small sample of data from a high-resource language so that currently existing tools can be easily tested in low-resource settings (Haddow et al., 2022). The lack of suitable preprocessing tools hinders research on these languages (Haddow et al., 2022). When available, linguistic tools, such as morphological segmentation, are paramount for preprocessing the data and obtaining subword segmentation, to better deal with out-of-vocabulary words; the most common strategies include BPE and SentencePiece (Haddow et al., 2022).

In this paper, we tackle the issue of preprocessing and its effect on translation quality when dealing with a highly agglutinative and morphologically complex low-resource language, Inuktitut. Our goal is to test several preprocessing techniques to determine which yields the best MT results for Inuktitut-English. We experiment with Byte-Pair Encoding (BPE) and Random Stemming, on both the romanized and the original Inuktitut scripts. We also incorporate Hungarian data into training, to determine if additional in-domain data from another language would help increase the translation quality.

2 Related Works

2.1 The Inuktitut Language

One of the many indigenous languages spoken throughout North America, Inuktitut has 33,790 speakers according to the 2021 Canadian census (Government of Canada, 2022). It is one of the official languages of the Canadian province Nunavut, where it is spoken by nearly 60% of the population and used in an official capacity, both in schools and legislative assemblies (Tulloch et al., 2017; Govern-

Proceedings of Recent Advances in Natural Language Processing, pages 475–479
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_053

ment of Nunavut). It is an agglutinative language with a rich morphological system.

A single Inuktitut word could be translated into an entire English sentence as in the case of “ᖃᖓᑕᓲᒃᑯᕕᒻᒨᕆᐊᖃᓛᖅᑐᖓ” [romanized: qangatasuukkuvimmuuriaqalaaqtunga, morphological breakdown: qangata-suu-kku-vim-mu-u-ria-qa-laaq-tunga] meaning “I’ll have to go to the airport” (Dench et al., 2011). Written Inuktitut utilizes an adapted version of the Cree syllabary known as Inuktitut Syllabics, an abugida writing system where consonant-vowel pairs are written as a collective unit, with the main consonant letter adapting to the currently attached vowel through movement or additional notation (Government of Nunavut). Romanized orthography of Inuktitut is also available through the use of Qaliujaaqpait (Government of Nunavut).

2.2 NLP for the Inuktitut Language

Low-resource MT of Inuktitut saw a rise in popularity in 2020 when the 3.0 version of the Nunavut Hansard Inuktitut-English parallel Corpus was released (Joanis et al., 2020). Most studies opt for the transliteration of the original Inuktitut script and apply morphological preprocessing on the romanized version of the dataset; one exception is the work by Joanis et al. (2020), who conduct experiments on both the original and the romanized script.

When it comes to MT translation of Inuktitut, the main issue is breaking down words into morphemes. Micher (2018) proposes to combine the UQA·ILA·UT analyzer developed at the Institute for Information Technology within the National Research Council of Canada (Farley), with a segmental recurrent neural network (SRNN) to expand morphological preprocessing coverage of the corpus. They point out that the UQA·ILA·UT analyzer cannot analyze 30% of the types from the corpus so he trains an SRNN model to identify the unrecognized morphemes (Micher, 2018). Joanis et al. (2020) use the same morphological preprocessing as Micher (2018) but they also take an alternative approach and simulate stemming, by choosing prefixes of three characters for Inuktitut words and five characters for English words (Joanis et al., 2020).Ngoc Le and Sadat (2020) build a deep learning-based word segmentation tool for Inuktitut, using a bidirectional long short-term memory neural network for word segmentation. Hernandez and Nguyen (2020) suggest a multilingual approach and train a transformer model on two additional agglutinative languages, Finnish and Estonian. Roest et al. (2020) test eight different segmentation techniques, including Rule-Based with UQA·ILA·UT, but they use a neural segmentation method built on a Transformer architecture instead of RNN. They also employ back-translated Inuktitut data, and additional data from a related language, Greenlandic (Roest et al., 2020), which had no positive effect.

3 Method

3.1 Corpora

In this paper, We use the Nunavut Hansard Inuktitut-English Parallel Corpus 3.0, as described in Joanis et al. (2020). The data consists of aligned sentences from proceedings of the Legislative Assembly of Nunavut. The Inuktitut syllabic data was romanized using a syllabic converter[1] to create a parallel romanized Inuktitut set. All in all, this amounts to a total of around 1.3 million aligned sentence pairs. The data, as provided by the National Research Council of Canada[2], is already divided into a train, dev, and test set for each language.

Although being a relatively large parallel corpus, the language follows mostly legislative assembly debates, which creates a lot of redundancies. For instance, the sentence “Thank you, Mr. Speaker” is found around 17,000 times throughout the entire corpora. Another sign of the debate-style type of language found in the corpus becomes evident as many sentences are very long, presumably due to the turn-taking nature of debates. The provided train, dev and test sets are also very messy in terms of special characters, such as parentheses, full stops etc., that appear in the middle of sentences, seemingly put there by the transcriber to clarify who’s talking or where interpretations start and end. There are also many empty lines that divide the different speaker turns as well as many very short lines (1-3 tokens) of audio interpretation.

We use Hungarian data for data augmentation. The Hungarian data is taken from the Hungarian to English EUROPARL parallel corpus v.7[3]. The main advantage of using this data in combination with the aforementioned Inuktitut data is that it is also derived from a similar domain, namely proceedings, and hence follows a similar debate-like

[1] https://www.syllabics.net/convert/inuktitut
[2] https://nrc-digital-repository.canada.ca
[3] https://www.statmt.org/europarl/

type of language. The data also happens to be very clean in terms of special characters littering the sentences and it is free of empty lines.

3.2 General Preprocessing

The data was stripped of special characters as the sheer number of them and their appearances in many sentences were deemed too noisy for training. A selected few sentences and phrases that were very common were also removed. Post-preprocessing the total number of lines in the Inuktitut-English corpus had been reduced to 661,263, which is approximately 26% of the original 2,575,449 lines. Many of these lines were, however, completely empty in the beginning. The full data split post-cleaning is presented in Table 1.

	train	**dev**	**test**	**total**
iu-en	655 765	2 422	3 076	661 263
hun-en	525 725	N/A	N/A	525 725

Table 1: Data split in sentence pairs.

We then used both Byte-Pair Encoding encoding and stemming simulation as segmentation tools. All the experiments were run using default OpenNMT-py parameters to create the vocabularies and to train the model.

3.3 Random stemming

Random stemming is a technique employed to approximate the retrieval of word stems, or root forms, by eliminating part of a word (Dolamic and Savoy, 2008). Stemming can be systematic when consisting of removing inflectional and derivative suffixes, or random, in the event that the suffixes are unknown (Dolamic and Savoy, 2008). In the latter case, one can decide on a set number of characters to approximate stems, 3 and 5 for Inuktitut and English respectively, in Joanis et al. (2020).

4 Experiments

4.1 Baseline

Our core baseline model in the experiments below is based on the Transformer architecture (Vaswani et al., 2017) trained on the iu-en parallel data. The latter, currently the *de facto* standard baseline in NMT, relies on the concept of self-attention, i.e., the ability to learn attending to different positions of the input sequence to compute a representation of that sequence. Another experiment was conducted using OpenNMT-py BPE-tokenizer with 12,000 merge operations, following the preprocessing steps taken by Hernandez and Nguyen (2020) of the same data. They mention that using a fewer number of merge operations for agglutinative languages might be beneficial for MT. For the BPE + Hu experiment, Hungarian data was added when training the model, using the OpenNMT weighting mechanism, to train on batches of training data from different languages. The Inuktitut corpus was given the weight of 8, while the Hungarian corpus was given the weight of 2.

4.2 Random Stemming Experiments

As an alternative to BPE encoding, stemming simulation was also applied, based on previous experiments by Joanis et al. (2020). We start by simulating Inuktitut prefixes, by truncating words at the third character, and English prefixes, by truncating words at the fifth character. Subsequently, a second experiment was conducted where only Inuktitut was preprocessed to simulate stemming and English was left untouched. Inuktitut words were stemmed randomly so that in the end the corpus was composed of stems ranging from two to six characters.

5 Results

We use BLEU (Papineni et al., 2002) for evaluating our models. All the results from the experiments are presented in Table 2.

	Inuktitut Script	Romanized
Baseline	11.3	13.0
Rand: Inuk	14.9	14.5
Rand: Iu_3, En_5	19.4	17.3
BPE	20.6	**20.3**
BPE + Hu	**21.0**	20.2

Table 2: BLEU scores of all experiments

The baseline model achieved a BLEU score of 11.3 on the Inuktitut script and 13.0 on the romanized script. The BPE-only model achieved the best BLEU score of all of the romanized experiments, with a score of 20.3, but was still outperformed by the model trained on the Inuktitut script, which achieved a BLEU score of 20.6. The BPE + Hungarian model achieved the best BLEU score overall, scoring 21.0 on the Inuktitut script. For both initial random stemming and Iu_3 , En_5 stemming experiments, models using the Inuktitut script per-

Model Output:	"This year ■'■ s young people will be graduating in Sanikiluaq and I would like to congratulate them in their future future"
Reference:	"This year saw nine students graduate from Sanikiluaq's high school which is a good sign for our future"
Model Output:	"Mr Speaker students graduating from high school will be graduating in the future"
Reference:	"Mr Speaker a High School Diploma is a stepping stone to future learning In achieving this goal our young graduates can look forward to greater opportunities in life"

Figure 1: Model predictions with semantic variation

Model Output:	"Member ■'■ s Statement – <unk> <unk> <unk> <unk>"
Reference:	"Member's Statement – Responsible Internet Use Mikkungwak"
Model Output:	"I encourage Nunavummiut to be safe and safe and safe and safe"
Reference:	"I also encourage all Nunavummiut to remain vigilant in keeping our communities safe in both the physical and virtual worlds"

Figure 2: Model predictions with repetitions

formed best, achieving a BLEU score of 14.9 and 19.4 respectively.

6 Discussion

The overall lack of parallel data to be used during training led to a lack of varied language, resulting both in an abundance of unknown tokens and the repeated use of simplified words in final translations, as displayed in 1. For some translations, it seems as if the model is taking liberties with the original intent of the speaker. See Figure 1.

There are a few cases where the translation does not match the reference sentence, but it still infers a similar meaning, for instance, by congratulating the students in the first example and linking "graduating in the future" to "can look forward to greater opportunities" in the second. For this reason, having fluent human speakers rate final translations may be helpful for future experiments to determine the semantic intent of the original Inuktitut sentence and the differences in speaking these in the original language. The inclusion of the original script during training showed better results in certain contexts, which has often been ignored in other research.

6.1 BPE and Random Stemming

Though the BPE+Hu model outperformed all other models, it is unclear if this is due to the Hungarian data specifically, or more generally having more data due to the overall lack of Inuktitut-English parallel data. For future experiments, it is recommended that additional languages are researched to determine their effects, as well as the inclusion of additional Inuktitut data to provide a clear decision on this matter.

Though performance does not quite match BPE experiments, the Iu_3, En_5 model appears to have potential as a preprocessing method. Further research should be performed using varying stemming configurations to determine the full potential of the effects of random stemming, especially on non-romanized Inuktitut script. Also, stemming the romanized equivalent of the Inuktitut script at the third character might not be the best idea since each Inuktitut syllabic character is transcribed into either two, or even three romanized characters.

7 Conclusion

We show that using BPE and (random) stemming as preprocessing techniques improves the translation quality for Inuktitut when no morphological analyzer is available for the original Inuktitut script, which has not received much attention thus far. We also experiment with data augmentation using Hungarian, which yielded better translation quality on the Inuktitut-English translation task.

References

Statistics Canada Government of Canada. 2022. 2021 census.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics.

C. Dench, Patricia Cleave, J. Tagak, and J. Beddard. 2011. The development of an inuktitut and english language screening tool in nunavut. *Canadian Journal of Speech-Language Pathology and Audiology*, 35:168–176.

Ljiljana Dolamic and Jacques Savoy. 2008. Stemming approaches for east european languages. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers 8*, pages 37–44. Springer.

B Farley. The uqailaut project.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

François Hernandez and Vincent Nguyen. 2020. The ubiqus English-Inuktitut system for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 213–217, Online. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural machine translation (seq2seq) tutorial. *https://github.com/tensorflow/nmt*.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. 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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Graham Neubig. 2015. lamtram: A toolkit for language and translation modeling using neural networks. http://www.github.com/neubig/lamtram.

Tan Ngoc Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nunavut Government of Nunavut. Inuktitut tusaalanga.

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 of the Association for Computational Linguistics*.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey.

Christian Roest, Lukas Edman, Gosse Minnema, Kevin Kelly, Jennifer Spenader, and Antonio Toral. 2020. Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 274–281, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Shelley Tulloch, Lena Metuq, Jukeepa Hainnu, Saa Pitsiulak, E E Flaherty, Cathy Yeonchoo Lee, and Fiona Walton. 2017. Inuit principals and the changing context of bilingual education in nunavut.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.