Revitalization of Indigenous Languages through Pre-processing and Neural Machine Translation: The case of Inuktitut

Ngoc Tan Le
UQAM / Montreal, Quebec, Canada
le.ngoc_tan@uqam.ca

Fatiha Sadat
UQAM / Montreal, Quebec, Canada
sadat.fatiha@uqam.ca

Abstract

Indigenous languages have been very challenging when dealing with NLP tasks and applications because of multiple reasons. These languages, in linguistic typology, are polysynthetic and highly inflected with rich morphophonemics and variable dialectal-dependent spellings; which affected studies on any NLP task in the recent years. Moreover, Indigenous languages have been considered as low-resource and/or endangered; which poses a great challenge for research related to Artificial Intelligence and its fields, such as NLP and machine learning. In this paper, we propose a study on the Inuktitut language through pre-processing and neural machine translation, in order to revitalize the language which belongs to the Inuit family, a type of polysynthetic languages spoken in Northern Canada. Our focus is concentrated on: (1) the preprocessing phase, and (2) applications on specific NLP tasks such as morphological analysis and neural machine translation, both for Indigenous languages of Canada. Our evaluations in the context of low-resource Inuktitut-English Neural Machine Translation, showed significant improvements of the proposed approach compared to the state-of-the-art.

1 Introduction

North America and Canada represent a highly complex linguistic regions, with numerous languages and great linguistic diversity. Indigenous languages are spoken widely and are official languages in Nunavut and the Northwest Territories. Indigenous peoples are making efforts to revitalize and sustain their languages, although they face many difficulties. The case of Inuktitut language in Canada has recently been a good example to examine. This language belongs to the Inuktut - the term Inuktut is inclusive of all dialects used in Nunavut - one of the official languages of Nunavut, the largest territory of Canada, which can be written in syllabics or in roman orthography, and regional variations use different special characters and spelling conventions (Statistics-Canada, 2017). As Inuktitut is an official language in Nunavut, there exists some resources that are available in this language at a much larger scale than most other languages in the same family, notably a parallel corpus with English. The researchers have shown more interest in Inuktut languages in NLP communities to respond to an increasing demand for Indigenous language educational materials and technology.

Previous studies have demonstrated that the development of Indigenous language technology faces many challenges on one hand, due to a high morpheme per word rate and the rich morphophonemics and variable dialectal-dependent spellings, a complex polysynthesis in linguistic typology, and on the other hand, due to the lack of orthographic standardization and a large dialect variation. Digital text and voice data are limited. It poses many greater challenges for NLP to develop applications for all users (Littell et al., 2018; Mager et al., 2018).

This research paper examines the case of Inuktitut through experiments on the crucial phase as preprocessing and NLP task, to revitalize the language which belongs to the Inuit languages family - the polysynthetic languages spoken in Northern Canada. We make our focus on (1) the preprocessing phase to build (2) NLP applications for Indigenous languages, such as morphological analysis, parsing, Named Entities Recognition and Machine Translation. This first step towards a multilingual NMT framework that will involve several endangered Indigenous languages of Canada, is essential, as the only parallel

Proceedings of the 28th International Conference on Computational Linguistics, pages 4661–4666
Barcelona, Spain (Online), December 8-13, 2020

corpus freely available for research is the Nunavut-English Hansard corpus (Joanis et al., 2020). The main contribution of our research is to revitalize these Indigenous languages and help ancestral and other knowledge transmission vertically from the elders to the youth.

The structure of the paper is described as follows: Section 2 presents the state-of-the-art on MT involving Indigenous languages. In section 3, we describe our proposed approach. Then, in section 4, we present our experiments and evaluations. Finally, in section 5, we present our conclusions and some perspectives for future research.

2 Related work

Johnson and Martin (2003) proposed an unsupervised technique with the hubs concept in a finite-state automaton. Those hubs mark the boundary between root and suffix. Inuktitut words are split into morphemes and merged hubs in a finite-state automaton. In their evaluations, they reported a good performance for English morphological analysis, with the text of Tom Sawyer, with a precision of 92.15%. However, for Inuktitut morphological analysis, they reported 31.80% precision and a low recall of 8.10%. They argued that the poor performance for Inuktitut roots was due to the difficulty of identifying word-internal hubs.

Farley (2012) developed a morphological analyzer for Inuktitut, which makes use of a finite-state transducer and hand-crafted rules. Nicholson et al. (2012) presented an evaluation about the morphological analyzer for Inuktitut, proposed by Farley (2012), and about alignment error rate with the use of the English-Inuktitut Nunavut Hansard corpus. They reported the best experimental results, in terms of head approach which, in Inuktitut, corresponds to the first one or two syllables of a token, with 79.70% precision, 92.20% recall and 17.60% alignment error rate. Inspired by the Uqailaut project of Farley (2012), Micher (2017) applied a segmental recurrent neural network approach (Kong et al., 2015) from the output of this morphological analyzer for Inuktitut. The few studies that deal with Machine Translation while involving Indigenous languages are related to the lack of parallel corpora (Mager et al., 2018).

3 Our Proposed Approach

Inspired by the work of Farley (2012), on the creation of the first Inuktitut morphological analyzer based on the Finite-State Transducer method, we built a deep learning-based word segmentation tool for Inuktitut. With the emergence of deep learning and the high computation of technology, neural network-based approaches have shown their effectiveness when applied on word segmentation and enhanced with large-scale raw texts to pretrain embeddings. The neural network-based model, with these additional linguistic factors, can be able to deal with data sparseness or language ambiguity (Kann et al., 2018).

Following our previous research (Le and Sadat, 2020), we propose a two phases framework to build : (1) a bidirectional LSTM word segmentation for Inuktitut; and (2) an Inuktitut-English NMT system (Figure 1). In the first phase, the word segmentation task, considered as a sequence labeling task, is formally formulated as follows:

Given an input sequence, W and C represent all the word-based (bi-)character-based pretrained embeddings. In the input representation layer, the input sequence is vectorized based on word embeddings W. This vectorized sequence is concatenated with (bi)character-based pretrained embedding C, with the state $\langle W, C \rangle$. Then it is fed into a bidirectional LSTM (*Long-Short Term Memory*) (Hochreiter and Schmidhuber, 1997). A hidden feature layer h merges all input features X_W, X_C into a single vector with a k-dimension.

The activation function, as an output function, is calculated in the output layer o, *e.g. softmax*.

$$h = tanh(W_{hW} \cdot X_W + W_{hC} \cdot X_C) \tag{1}$$

$$o = softmax(W_o \cdot h + b_o) \tag{2}$$

The second phase consists of building a NMT for Inuktitut-English. Here, we use the Transformer-based encoder-decoder architecture (Vaswani et al., 2017). We apply our method to preprocess the Inuktitut source language within an Inuktitut-English NMT system.

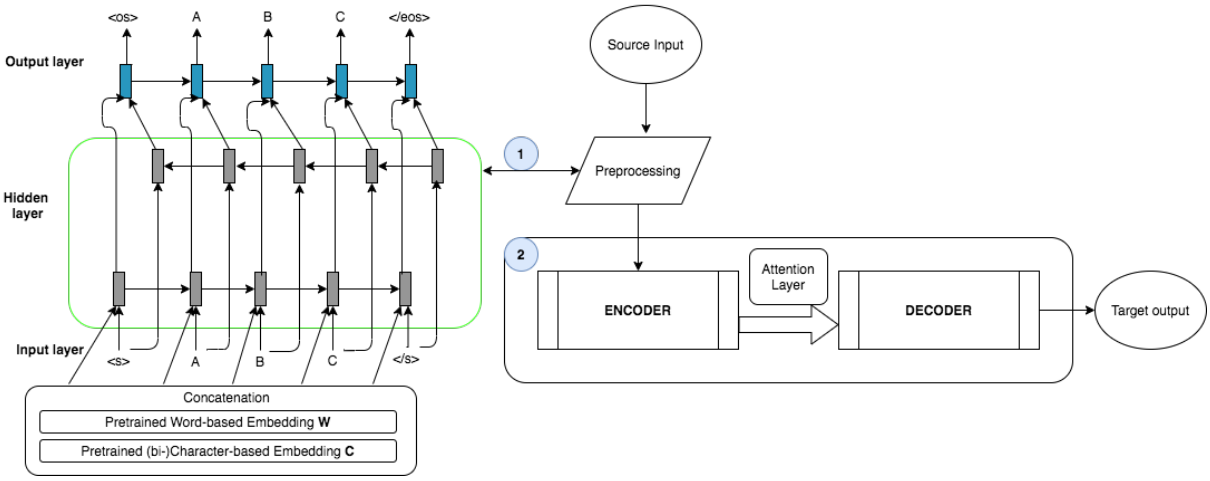

Figure 1: Architecture of our framework: (1) Building a bi-LSTM word segmentation for Inuktitut Indigenous language, (2) Building an Inuktitut-English neural machine translation system based on Transformer-based encoder-decoder architecture

The Masked Language Modeling

Methods of BERT-based pretraining (Devlin et al., 2018) for NMT are time and resource consuming because they need to pretrain large models from scratch using large-scale data. Instead of pre-training a BERT model for Inuktitut and English, we have intend to use the output of BERT as context-aware embeddings of the NMT encoder. Due to the lack of resources of Inuktitut, we apply only the Masked Language Model (MLM) concept (Lample and Conneau, 2019) to train our own monolingual BERT-like model for Inuktitut. For English, we will use pretrained models from Huggingface[1]. The main goal aims to use pretrained BERT-like model as embeddings in order to initialize the encoder of NMT model.

The BERT architecture can deal with two kinds of objective functions, namely, masked language modeling and next sentence prediction. In our context, we consider the masked language modeling objective of (Devlin et al., 2018). In the training step, 15% words in a sentence are randomly masked and the MLM model is trained to predict them with their surrounding words.

In other words, the MLM aims to reconstruct the original sentence sequence from noisy input because some words are masked with special tokens such as [MASK] or [s] or [/s] for the beginning and the end of a sequence. For example, if the original sentence is "*[/s] take a seat [/s]*" and "*[/s] [MASK] a seat [MASK]*" is given as the input word sequence, then the system predicts that the original word for the first *[MASK]* was "*take*" and the second [MASK] was "[/s]"(Figure 2).

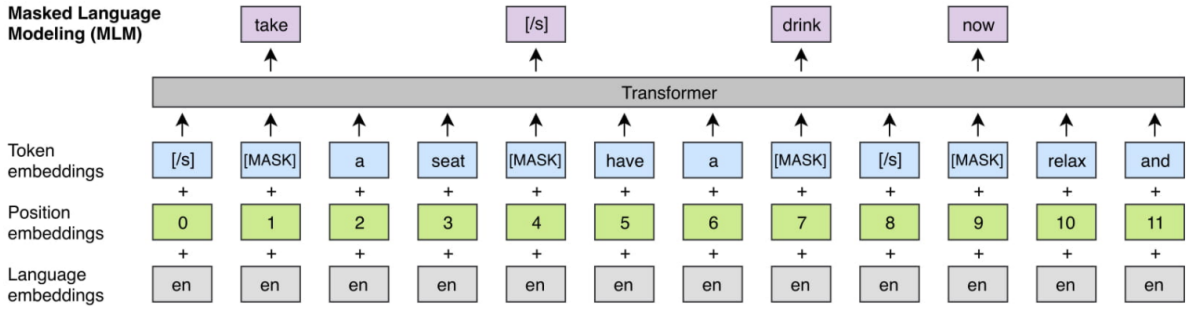

Figure 2: Structure of the Masked Language Model (MLM) objective, similar to the one of (Devlin et al., 2018). Source: (Lample and Conneau, 2019).

[1]pretrained BERT models from`https://github.com/huggingface/transformers`

4 Experiments

4.1 Data Preparation

In our evaluations, we train our NMT model by using the Nunavut Hansard for Inuktitut-English (third edition) (Joanis et al., 2020). The statistics of the training corpus are described in Table 1.

Dataset	#tokens	#train	#dev	#test
Inuktitut	20,657,477	1,293,348	5,433	6,139
English	10,962,904	1,293,348	5,433	6,139

Table 1: Statistics of Nunavut Hansard for Inuktitut-English

In order to pre-train the embeddings for Inuktitut, we used the *word2vec* toolkit (Mikolov et al., 2013) with option $CBOW$ (*Continuous Bag-Of-Words*). To train our rich word segmenter, we annotated 11K sentences, 250 sentences, 250 sentences for training, development and testing, respectively. We used the Uqailaut toolkit (Farley, 2012) to annotate the training data. We observe there are only 97,785 unique terms for word-based vocabulary, 102 unique terms for character-based vocabulary and 1,406 unique terms for bi-character-based vocabulary.

4.2 Training Configuration

The word segmentation model is composed of 2-layer bi-directional Long Short-term Memory (LSTM), with the hidden layer of 200 neurons. We performed two experiments: with only word-based pretrained embedding (EXP1) and with all (bi)character-based and word-based pretrained embeddings (EXP2).

In the preprocessing step, we used the *Moses* (Koehn et al., 2007) tokenizer in all experiments. In the BPE subword segmentation, we used the *subword-nmt* (Sennrich et al., 2016) toolkit to create a 30k BPE joint source-target vocabulary. Then to train our Transformer-based NMT models, we used the *Marian-nmt* toolkit (Junczys-Dowmunt et al., 2018) with the following hyper-parameters settings (Table 2).

Architecture Type: Transformer
Number of layers: 6-layer depth for both encoder and decoder
Number of heads: 8-layer mulit-heads
Hidden layers: 2,048 units in the feed-forward networks
Optimization: Adam
Embedding size: 512
Learning rate : 0.0003
Batch-size: 32
Number of epochs: 50 iterations
Early stop: cross-entropy scores
Validation updates: 5,000

Table 2: Hyper-parameters settings for our NMT framework

Our experiments on NMT using the Transformer-based architecture (Vaswani et al., 2017) are described as follows:

(1) System 1 as Baseline, with only BPE-preprocessed data: We choose the same configuration as described in (Joanis et al., 2020).

(2) System 2 with our proposed Inuktitut word segmentation.

(3) System 3 that combines both BPE-segmentation and our proposed word segmentation.
In the System 3, the training data are segmented by using our Inuktitut word segmentation. Then these preprocessed training data are split in subwords units with the BPE-based method.

(4) System 4 using BERT as embeddings in the encoder, with Transformer-based architecture.

4.3 Evaluations

For word segmentation, we used the automatic metrics: *Recall*, *Accuracy* and *F-measure*. We observed an improvement of +3.12% in terms of F-measure when using all (bi)character-based and word-based pretrained embeddings. The recall of EXP2 is better than one of EXP1, with a gain of +6.23% on the test set. The model is able to recognize more complex morphemes given a sentence word. However, the accuracy of EXP2 is lower than one of EXP1, with a loss of -1.49% on the test set.

		Recall	**Accuracy**	**F-measure**
EXP1	*dev*	74.52	87.68	80.57
	test	64.34	82.28	**72.21**
EXP2	*dev*	68.94	80.82	74.41
	test	70.57	80.79	**75.33**

Table 3: Evaluations of the word segmentation models for Inuktitut

Experiment	**dev set**	**test set**
System 1 - Baseline (Joanis et al., 2020) (BPE)	41.40	35.00
System 2 (our Inuktitut WS)	49.12	**39.53**
System 3 (our Inuktitut WS+BPE)	52.30	**42.09**
System 4 (BERT as embeddings)	53.93	**43.40**

Table 4: Performances on Inuktitut-English NMT in terms of lowercase word BLEU score

We conducted additional evaluations on NMT using the BLEU metric (Papineni et al., 2002), with lowercase and v13a tokenization, similar to Joanis et al. (2020). We observed a significant improvement of the performance with gains up to +4.53% for System 2, +7.09% for System 3, and +8.40% for System 4 in terms of BLEU score, compared to the System 1 as Baseline, on the test set (Table 4). We noticed that the word segmentation helped to solve the complexity of Inuktitut morphology. Our proposed NMT system showed better performance than the state-of-the-art, as presented in Joanis et al. (2020) with only BPE-preprocessed training data, thanks to the rich word segmenter.

5 Conclusion and Perspective

In this paper, we have presented how to leverage Inuktitut-English Neural Machine Translation with morphological word segmentation. We intend to apply our proposed approach in other Indigenous language families, especially related to the Inuit language family, to deal with NLP tasks. With the valuable collaboration of Indigenous communities, we will be able to collect reliable data from the speakers of these Indigenous languages. Moreover, our NLP applications could help preserve ancestral knowledge and revitalize Indigenous languages, heritage and culture with the transfer of knowledge from elders to the youth.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Benoit Farley. 2012. The uqailaut project. *URL http://www. inuktitutcomputing. ca*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

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 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France, May. European Language Resources Association.

Howard Johnson and Joel Martin. 2003. Unsupervised learning of morphology for english and inuktitut. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 43–45. Association for Computational Linguistics.

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*.

Katharina Kann, Manuel Mager, Ivan Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. *arXiv preprint arXiv:1804.06024*.

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.

Fang Kong, Sheng Li, and Guodong Zhou. 2015. The sonlp-dp system in the conll-2015 shared task. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 32–36.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Tan Ngoc Le and Fatiha Sadat. 2020. Low-resource NMT: an empirical study on the effect of rich morphological word segmentation on Inuktitut. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 165–172, Virtual, October. Association for Machine Translation in the Americas.

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. 2017. Improving coverage of an inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *hlt-Naacl*, volume 13, pages 746–751.

Jeremy Nicholson, Trevor Cohn, and Timothy Baldwin. 2012. Evaluating a morphological analyser of inuktitut. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 372–376. Association for Computational Linguistics.

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.

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.

Statistics-Canada. 2017. Census in brief: The aboriginal languages of first nations people, métis and inuit, 2017. *URL: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016022/98-200-x2016022-eng.cfm*.

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.