Low-Resource NMT: an Empirical Study on the Effect of Rich Morphological Word Segmentation on Inuktitut

Ngoc Tan Le le.ngoc_tan@courrier.uqam.ca
Fatiha Sadat sadat.fatiha@uqam.ca
Department of Computer Science, Universite du Quebec a Montreal
201, avenue du President-Kennedy, Montreal, QC, Canada

Abstract

Nowadays, very little research for Indigenous languages has been studied. Indigenous languages bring significant challenges for Natural Language Processing approaches because of multiple features such as polysynthesis, morphological complexity, dialectal variations with rich morpho-phonemics, spelling with noisy data and low resource scenarios. Particularly, morphological segmentation for polysynthetic languages is challenging because a word may contain many individual morphemes and training data can be extremely scarce. The current research paper focuses on Inuktitut, one of the Indigenous polysynthetic language spoken in Northern Canada. First, a rich word segmentation for Inuktitut is studied using a set of rich features and by leveraging (bi-)character-based and word-based pretrained embeddings from large-scale raw corpora. Second, we incorporated this pre-processing step into our first Neural Machine Translation system. Our evaluations showed promising results and performance improvements in the context of low-resource Inuktitut-English neural machine translation.

1 Introduction

In the Americas, there are a wide range of linguistic families about 140 linguistic families in the world. About 900 different Indigenous languages spoken in the Americas approximately are reported in (Mager et al., 2018). In Canada, there is a great diversity of Indigenous languages, grouped into 12 language families, that have been central to the history of First Nations people, Métis and Inuit in Canada and continue to play a vital role to this day (Rice, 2011).

This research paper focuses on Inuktitut, one of the Indigenous polysynthetic languages spoken in Northern Canada and the development of a Neural Machine Translation (NMT) for Inuktitut-English. The main objective and motivation of this project is the revitalization and preservation of Indigenous languages and cultural heritage through major tasks in NLP.

However, the development of Indigenous language technology faces many challenges such as polysynthesis with a high rate of morphemes per word, lack of orthographic normalization, dialectal variations and lack of linguistic resources and tool such as corpora (Littell et al., 2018). This first step towards a multilingual NMT framework that will involve several endangered Indigenous languages of Canada, is essential as the only parallel corpus freely available for research is the Nunavut-English Hansard corpus (Joanis et al., 2020).

Inspired by the work of Farley (2012), related on the creation of the first Inuktitut morphological analyzer, we build a neural network-based word segmenter for Inuktitut. The main goal of this research is two folds: (1) to investigate empirically several word segmentation methods on Inuktitut; and (2) to improve low-resource NMT that involves Inuktitut through rich morphological word segmentation.

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 methodology. Then, in section 4, we present our experiments and evaluations. Finally, in section 5, we present our conclusion and some perspectives for future research.

2 Related work

Farley (2012) developed a morphological analyzer for Inuktitut, which makes use of a finite-state transducer and hand-crafted rules. This Uqa·Ila·Ut project (Uqailaut)[1] is a rule-based system that involves regular morphological variations of about 3200 head, 350 lexical, and 1500 grammatical morphemes, with heuristics for ranking the various readings. 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 development of MT systems for Indigenous languages have followed the trends in the field, with rule-based, statistical-based and neural network-based approaches. Micher (2018) applied the Byte Pair Encoding (BPE) algorithm (Sennrich et al., 2016) pre-processing both the English and Inuktitut sides of the Nunavut Hansard corpus, in the Inuktitut to English direction, reported a BLEU score of 30.35. NMT approaches use neural networks architectures that are fed with very big amounts of parallel texts. However, these resources are currently unavailable or scarce for most Indigenous languages, especially for the endangered such as Inuinnaqtun, except Inuktitut-English (Joanis et al., 2020).

3 Methodology

In this section, we present some existing word segmentation methods for Inuktitut as well as our proposed one.

3.1 Uqailaut morphological analysis

The Uqailaut project, proposed by Farley (2012), is based on a Finite-State Transducers (FST), while applying several techniques and resources such as grammar rules, linguistic knowledge and heuristics. The FST-based morphological analyzer produces one or more morphological predictions for a given word. The heuristics allow to choose the shortest path of the morphological analysis. For example, *'katimajiit'* is segmented as *'kati ma ji it'* (Table 1). The root *'kati'* means *'to accumulate; to gather; to join; to unite'*. And the suffixes are *'ma'* means *'to be in a state of'*; *'ji'* means *'one whose job is; agent'*; *'it'* means *'to be such'*.

Segmentation	Sentence Example
Raw text	ilangit katimajiit : angiqpugut . (Meaning: *some members : agreed .*)
Reference	ila ngit kati ma ji it : angiq pugut .
Uqailaut project (Farley, 2012)	ila ngit kati ma ji it : angiq pugut .
BPE (Sennrich et al., 2016)	ilangit kati@@ ma@@ jiit : angiq@@ pu@@ gut .
Our proposed approach	ila ngit kati ma ji it : angiq pugut .

Table 1: Illustrations on the Inuktitut word segmentation involving our proposed approach and others

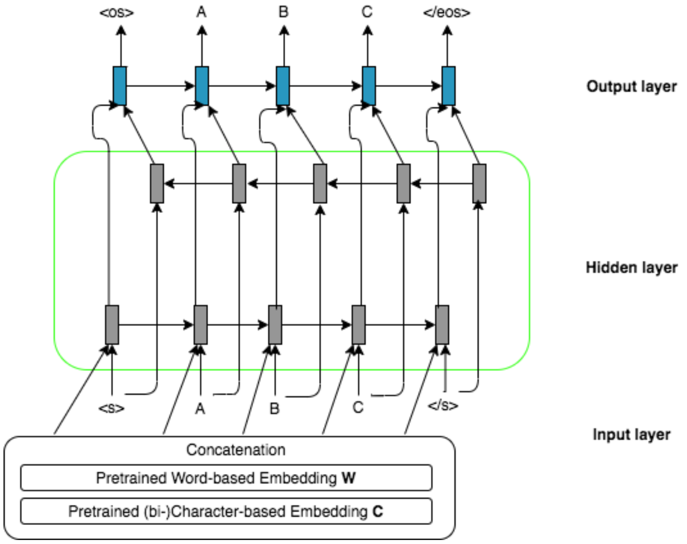
[1] http://www.inuktitutcomputing.ca/Uqailaut/info.php

3.2 Byte-Pair Encoding

The byte-pair encoding (BPE), proposed by Sennrich et al. (2016), is an unsupervised word segmentation that aims at splitting word by subword units, which helps cope with rare and unknown words. BPE applies the minimum entropy on the subword units, aka tokens, with the predefined vocabulary size. These tokens look like morphemes, although BPE segmentation model is based on training data rather than on linguistic knowledge. For example, in Inuktitut, *'katimajiit'* (meaning: *members* in English) may be segmented as *'kati@@ ma@@ jiit'* (Table 1). This word should be correctly segmented *'kati@@ ma@@ ji@@ it'*, in the case of a large-scale training data. The symbol @@ represents an in-word morpheme boundary.

3.3 Our proposed approach

Neural network-based approaches have shown their efficiency when applied on word segmentation and enhanced with large-scale raw texts to pretrain embeddings. Adding these linguistic factors allows the neural model to perform better, especially when dealing with data sparseness or language ambiguity (Kann et al., 2018).



<os>
A
B
C
</eos>
Output layer
Hidden layer
<s>
A
B
C
</s>
Input layer
Concatenation
Pretrained Word-based Embedding **W**
Pretrained (bi-)Character-based Embedding **C**

Figure 1: The system architecture of Inuktitut Rich Morphological Word Segmentation based on neural networks

In this research, the word segmentation task is considered as a sequence labeling task. Given an input sequence, W and C that contain all the input words and the input (bi-)characters. At each step, the state consists of a sequence of words, $W = [w_0, w_1, ..., w_m]$, that have been fully recognized, and a sequence of next incoming (bi-)characters $C = [c_0, c_1, ..., c_n]$, as shown in Figure 1. The architecture is composed of three main layers: the input layer, the hidden layer and the output layer. The input layer contains the input word sequence transformation by concatenating pretrained (bi-)character-based and word-based embeddings, with the state $S = \langle W, C \rangle$. On the top of the input representation layer, we use a hidden feature layer h to merge all input features X_W, X_C into a single vector. This layer is built with bidirectional LSTM (*Long-Short Term Memory*) (Hochreiter and Schmidhuber, 1997).

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

The output layer o calculates the activation function as an output function and displays the output hypothesis.

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

4 Experiments

4.1 Data Preparation

We train our NMT model by using the Nunavut Hansard for Inuktitut-English (third edition). As described in Joanis et al. (2020), this corpus contains 1,293,348 sentences, 5,433 sentences and 6,139 sentences for the training, development and testing sets, respectively (Table 2).

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 2: Statistics of Nunavut Hansard Inuktitut-English parallel corpus 3.0

In order to pre-train the (bi-)character-based and word-based embeddings for Inuktitut, these Nunavut Hansard datasets were used with the *word2vec* toolkit (Mikolov et al., 2013) with a dimension of 50 and 30 for word-based and (bi-)character-based embeddings, respectively, and option $CBOW$ (*Continuous Bag-Of-Words*) by default. 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 bicharacter-based vocabulary (Table 3). To train our rich word segmenter[2], we annotated 11K sentences, 250 sentences, 250 sentences for training, development and testing, respectively. We used Uqailaut toolkit (Farley, 2012) to annotate the training data.

Embedding type	#terms	#dimension
word-based	97,785	50
character-based	102	30
bicharacter-based	1,406	30

Table 3: Statistics of word-based, (bi)character-based embeddings training by using Nunavut Hansard Inuktitut-English parallel corpus 3.0 for Inuktitut

4.2 Training Configuration

For word segmentation, we adapted the *RichWordSegmenter* toolkit (Yang et al., 2017) to train our Inuktitut word segmenter. The model is composed of 2-layer bi-directional Long Short-term Memory (LSTM) cells, with a dimension size of 50 for the projection layer to encode the input sequences and a dimension size of 200 for the hidden layer. The *Adam* optimizer (Kingma and Ba, 2014) was used to learn the network's weights with a learning rate of 0.001. In the BPE subword segmentation, we used *subword-nmt* (Sennrich et al., 2016) toolkit to create a BPE joint source-target vocabulary with dimension of 30,000.

In the preprocessing step, we used *Moses* (Koehn et al., 2007) tokenizer in all experiments and Moses preprocessing scripts to clean the training data with a threshold of 50 words by sentences and without repetitive sentences. We used Marian-nmt (Junczys-Dowmunt et al., 2018) to train our Transformer-based NMT with the following hyper-parameters settings: 6-layer depth for both encoder and decoder, embedding dimension of 512, 2048 units in hidden

[2] Github repository: `https://github.com/NgocTanLE/Inuktitut-English-NMT`

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 used 6-GPUs of NVIDIA GeForce GTX 2080 Ti 12Gb.

Our experiments on NMT using the Transformer-based architecture (Vaswani et al., 2017) are described as follows: (1) the Baseline with only tokenized training data; (2) System 1 with only BPE-preprocessed data; (3) System 2 with our proposed Inuktitut word segmentation ; and (4) System 3 that combines both BPE-segmentation and our proposed word segmentation. In the system 3, the training data are firstly segmented by using our Inuktitut word segmentation. Then these segmented training data are secondly processed in subwords with the BPE-segmentation method.

4.3 Evaluations

Evaluations on word segmentation were performed using different automatic metrics such as *Recall*, *Accuracy* and *F-measure*. As described in Table 4, the model of EXP2, with all pre-trained embeddings, showed better performances than the model of EXP1, that uses only pre-trained word-based embedding, with F-measure of 72.21% and 75.33% (+3.21%) on the *test set* respectively (Table 4).

		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 4: Results on Inuktitut word segmentation. EXP1: word-based embedding, EXP2: all character-based, bicharacter-based, word-based embeddings

The NMT models were evaluated with the BLEU metric (Papineni et al., 2002), with lowercase and v13a tokenization, similar to Joanis et al. (2020). All the systems outperformed the baseline with gains up to +24.32 for the development set and up to +18.44 for the test set in terms of BLEU scores (Table 5). The performance of our proposed NMT models has significantly improved up to +1.03 and to +7.09 of BLEU scores with Systems 1, 2 and 3, respectively, compared to the NMT system of Joanis et al. (2020).

Experiment	**dev set**	**test set**
Baseline (tokenized)	27.98	23.65
(Joanis et al., 2020) (BPE)	41.40	35.00
System 1 (BPE)	42.62	36.03
System 2 (our Inuktitut WS)	49.12	**39.53**
System 3 (our Inuktitut WS+BPE)	52.30	**42.09**

Table 5: Performances on Inuktitut-English NMT in terms of lowercase word BLEU score

Our proposed NMT approach, that incorporates word segmentation for the Inuktitut as a source language, achieved significant improvement over the baseline system, 39.53 versus 23.65 in terms of BLEU score. Moreover, although our Inuktitut word segmentation model is trained on a small annotated corpus, the translation model can show good predictions compared to other models (Table 6). The experimental results have a positive impact on the translation model performance.

System	Sentence Example
[iu] raw	pikkugijauningit quviasuutauninganut arraagunit 15 @-@ nik uiviititut ilinniarvimmik
[en] Reference	congratulations *on celebrating 15 years* of the *francophone* school
[en] Baseline	congratulations *to 15 years* of *french* school
[en] Joanis et al. (2020)	congratulations *to the 15 years* of the *french* school
[en] System 1	congratulations *to the 15 years* of the *french* school
[en] System 2	congratulations *on celebrating 15 years* of the *french* school
[en] System 3	congratulations *on celebrating the 15 years* of the *french* school

Table 6: Illustrations on some translation predictions using different NMT systems

5 Conclusion and Perspective

In this paper, we presented an empirical study for rich morphological word segmentation on Inuktitut in Machine Translation. A neural network-based word segmenter was built for Inuktitut Indigenous language, with the use of a rich features set by leveraging (bi-)character-based and word-based pretrained embeddings from large raw texts. We applied our method to preprocess the Inuktitut source language into an Inuktitut-English NMT system. 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.

The interests for Indigenous languages are growing in NLP community. We noticed that North American languages are the most studied. We observed the past, current NLP researches have been done for morphology and machine translation. The study of Indigenous languages could lead us for a more complete understanding of human languages and advance towards universal NLP models.

Future work will focus on adding more annotated data and additional domain-specific features to improve the accuracy of our model. Moreover, we are working towards a multilingual NMT framework to incorporate more Indigenous languages, more precisely the endangered ones, with the aim of preserving and revitalizing endangered and Indigenous languages, their heritage and culture.

References

Farley, B. (2012). The uqailaut project. *URL http://www. inuktitutcomputing. ca*.

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

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k., Stewart, D., and Micher, J. (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. European Language Resources Association.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Kann, K., Mager Hois, J. M., Meza-Ruiz, I. V., and Schütze, H. (2018). Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

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

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., 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.

Kong, F., Li, S., and Zhou, G. (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.

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C., and Junker, M.-O. (2018). Indigenous language technologies in canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (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. Association for Computational Linguistics.

Micher, J. (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.

Micher, J. (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.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *hlt-Naacl*, volume 13, pages 746–751.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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

Sennrich, R., Haddow, B., and Birch, A. (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. 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. In *Advances in neural information processing systems*, pages 5998–6008.

Yang, J., Zhang, Y., and Dong, F. (2017). Neural word segmentation with rich pretraining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–849, Vancouver, Canada. Association for Computational Linguistics.