Towards a Low-Resource Neural Machine Translation for Indigenous Languages in Canada

Ngoc Tan Le* — **Fatiha Sadat***

* *Department of Computer Science, University of Quebec in Montreal, Canada*

ABSTRACT. *The Natural Language Processing research community is increasingly interested in less-resourced languages and linguistic diversity through technology. Translation to and from low-resource polysynthetic languages has, in particular, always faced numerous challenges, such as morphological complexity, dialectal variations, noisy data due to different spellings and low-resource scenarios. Moreover, the morphological segmentation for indigenous polysynthetic languages is particularly challenging with multiple individual morphemes by word and several meanings per morpheme. The present research focuses on Inuktitut and Inuinnaqtun, indigenous polysynthetic languages spoken in Northern Canada. We then build a morphological segmenter and a NMT system for these indigenous languages. Our proposed NMT model outperformed the state-of-the-art in the context of low-resource Inuktitut-English Neural Machine Translation.*

RÉSUMÉ. *La communauté de recherche sur le traitement des langues naturelles porte un intérêt croissant aux langues peu dotées et à la diversité linguistique grâce à la technologie. La traduction vers et depuis les langues polysynthétiques s'est régulièrement heurtée à de nombreux défis comme la complexité morphologique, les variants dialectiques, les données bruitées, les différentes orthographes, et les scénarios d'entraînement avec peu de données. Par ailleurs, la segmentation morphologique des langues polysynthétiques autochtones est rendue particulièrement difficile en raison de multiple morphèmes par mot et de plusieurs sens par morphème. La présente recherche se concentre sur l'inuktitut et l'inuinnaqtun, langues polysynthétiques autochtones parlées dans le nord du Canada. Nous construisons un segmenteur et un système de traduction automatique neuronale pour langues autochtones du Canada. Notre modèle de traduction automatique a surpassé l'état de l'art dans le contexte de la traduction automatique neuronale inuktitut-anglais.*

KEYWORDS: *Polysynthetic languages, Inuktitut, Inuinnaqtun, NMT, Low-resource.*

MOTS-CLÉS : *Langues polysynthétiques, Inuktitut, Inuinnaqtun, TAN, peu dotée.*

1. Introduction

According to Mager *et al.* (2018), the Americas have a diverse range of linguistic families, with approximately 900 different indigenous languages spoken. More specifically, Canada's wide range of indigenous languages, grouped into 12 language families, has played an important role in the history of First Nations, Métis, and Inuit, and continues to do so today (Rice, 2011). Due to a variety of factors, there has been very little research on indigenous languages in recent years. Natural Language Processing (NLP) researchers working with indigenous languages encounter numerous obstacles, including polysynthesis, with a high rate of morphemes per word, lack of orthographic normalization, dialectal variances, and a lack of linguistic resources and tools (Littell *et al.*, 2018; Schwartz *et al.*, 2020).

This study focuses on two indigenous polysynthetic languages spoken in Northern Canada, particularly Inuktitut and Inuinnaqtun, as well as the development of an Inuktitut-English Neural Machine Translation (NMT).

In the Northwest Territories, Inuktitut and Inuinnaqtun (a related dialect group) are recognized as official indigenous languages. They belong to the language family of Esquimo-Aleut, including the Inuit language. The Inuit language, or Inuktut, is a continuum of dialects that are spoken in the North American Arctic: in northern Alaska, in the Northwest Territories, in Nunavut, in Nunavik (northern Quebec), Nunatsiavut (in Labrador), and Greenland. Inuktitut [1] is an indigenous North American language spoken in the Canadian Arctic. Inuktitut is part of the vast Inuit language continuum (set of dialects) stretching from Alaska to Greenland. Inuktitut has official language status in Nunavut, like English and French. According to the 2016 census [2], it has approximately 39,770 speakers, 65% of whom live in Nunavut and 30.8% in Quebec. Inuinnaqtun belongs to the Western Canadian Inuktun family of languages, including two other dialects, Siglitun and Natsilingmiutut. According to Statistics Canada, in 2016, Inuinnaqtun is the mother tongue of 675 people in Canada and 1,310 people can speak this language.

This first step towards a multilingual NMT framework, which will include several endangered indigenous languages of Canada, is critical, the Nunavut-English Hansard corpus being the only parallel corpus freely available for research (Joanis *et al.*, 2020). Haddow *et al.* (2021) considered many features between high- (*e.g.* 280M parallel sentences), medium- (*e.g.* 0.7M parallel sentences), and low-resource (*e.g.* 0.035M parallel sentences) language pairs based on the number of native speakers and the quantity of parallel sentences. Joshi *et al.* (2020) presented the relationships between the types of languages, resources, and their representation in NLP conferences to understand the trajectory that different languages have followed over time. They highlighted, via a quantitative investigation, the disparity between languages, especially

1. Source: Compton, Richard . "Inuktitut". L'Encyclopédie Canadienne, 20 novembre 2019, Historica Canada. `www.thecanadianencyclopedia.ca/fr/article/inuktitut`.

2. Source: `https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016022/98-200-x2016022-fra.cfm`.

in terms of their resources. Although there is a correlation between them, there are many outliers where either widely spoken languages have a minimal parallel corpus, or languages with a limited number of speakers are resource-rich in terms of corpora. We also observed that the concept of low-resource might change over time. We could crawl additional parallel data, or use related language data or monolingual data. Several language pairings are no longer considered low-resource. Thanks to the crawled parallel sentences size, Inuktitut is currently rated a medium-resource level. However, with only a few comparable phrases, such as the Bible, Inuinnaqtun remains highly under-resourced.

The primary goal and motivation for this research project aim to revitalize and to preserve Canadian indigenous languages and cultural heritage through major NLP tasks. Our research is divided into two stages: (1) building a morphological segmenter for indigenous languages, to be integrated into (2) the framework of a Neural Machine Translation system for indigenous languages.

Inspired by the work of Farley (2012), related to the creation of the first Inuktitut finite-state transducer-based morphological analyzer, we propose a novel technique based on deep neural networks to create a word segmenter for indigenous languages. First, we investigate several methods empirically, including supervised, semi-supervised and non-supervised approaches to word segmentation task. In the supervised approach, the task is considered as a sequence labelling task. We apply the sequence-to-sequence architecture (Sutskever *et al.*, 2014) with the encoder-decoder layers (see Section 3.1). In the semi-supervised and non-supervised approaches, we adopt an Adaptor Grammars (Johnson, 2008), fine-tuning the word segmentation model using a deep learning-based architecture for indigenous languages (see Section 3.2). Second, we construct a framework for a low-resource Neural Machine Translation system by incorporating our word segmenter, during the source-side language preprocessing step (see Section 3.5).

Our contributions to the current research are as follows:

(1) to perform empirical research on several word segmentation approaches to indigenous languages, particularly Inuktitut and Inuinnaqtun;

(2) to propose a neural network-based word segmenter for indigenous languages;

(3) to enhance low-resource NMT via extensive morphological word segmentation;

(4) to empirically compare our proposed NMT technique with different designs such as Sequence-to-Sequence (Sutskever *et al.*, 2014), Transformer (Vaswani *et al.*, 2017), and multilingual NMT architecture.

The following is a description of the article's structure: The section 2 highlights the most recent advances in morphological analyzers and Machine Translation concerning indigenous languages. Our technique is described in Section 3. Section 4 provides our experiments and results. Finally, Section 5 offers our conclusion as well as potential future research directions.

2. Related work

2.1. *Morphological complexity of indigenous languages*

Most of the indigenous languages, particularly in the Americas, belong to either the polysynthetic language group or the agglutinative language group. They have a complex, rich morphology that plays an important role in human learning versus machine learning (Gasser, 2011; Littell *et al.*, 2018). Much of the research on their morphological analysis has focused only on linguistic aspects.

From a linguistic point of view, a word is formed from a word base, followed by multiple suffixing morphemes, to become a long sentence word. The suffixing morphology phenomenon makes a word very long and potentially unique (Mithun, 2015). Moreover, the composition of verbs and the incorporation of noun in indigenous languages are much more complex than in European languages, such as English or French. Indigenous languages of North America are highly inflected.

Our current research investigates Inuktitut and Inuinnaqtun, polysynthetic languages in the Inuit language family spoken in Northern Canada. Inuinnaqtun is endangered and has been estimated to disappear in less than two generations[3].

Comparing word composition in English, the word structure in Inuit languages is variable in its surface form. Words can be very short, composed of three formative features such as word base, lexical suffixes, and grammatical ending suffixes. They can be very long up to ten or even fifteen formative morphemes as features depending on the regional dialect (Lowe, 1985; Kudlak and Compton, 2018).

Composition of Inuit word = **Word base** + Lexical suffixes + *Grammatical ending suffixes*

For instance, in the Inuktitut language of Nunavut[4], Canada, a sentence word example is given below:

– (Inuktitut script) ᑐᓵᑦᓯᐊᕈᓐᓇᖕᒋᑦᑐᐊᓘᔪᖕᒐ

– (Romanized script) **tusaa**-tsia-runna-nngit-tu-alu-u-*junga*

– (Meaning) **hear**-well-be.able-NEG-DOER-very-*BE-PART.1.S*, where NEG-DOER means the negation, and PART.1.S means participle first singular.

– (English) I can't hear very well

In this example, this long word is composed of the morpheme as a root word **tusaa** (meaning: **to hear**), followed by seven morphemes: six lexical suffixes (**tsia, runna, nngit, tu, alu, u**) and one grammatical ending suffix (**junga**) that are separated by hyphens. All segmented morphemes are synthetically composed in only one unit.

3. Source: `https://www.kitikmeotheritage.ca/language`.
4. Source: `https://en.wikipedia.org/wiki/Inuit_grammar`.

In Inuinnaqtun, another sentence word example is depicted from the grammar book of Lowe (1985), illustrating the same phenomenon of word composition of an Inuinnaqtun sentence word *umingmakhiuriaqtuqatigitqilimaiqtara* that can be segmented into a word base and several suffixes as follows:

– (Inuinnaqtun script) umingmakhiuriaqtuqatigitqilimaiqtara

– (Morpheme segmentation) **umingmak**-hiu-riaqtu-qati-gi-tqi-limaiq-*ta-ra*

– (Meaning) **muskox** - hunt - go in order to - partner - have as - again - will no more - *I-him*

– (English) I will no more again have him as a partner to go hunting **muskox**

In this example, the first morpheme as a root word **umingmak** (meaning: **muskox**) is followed by six morphemes as lexical suffixes (**hiu, riaqtu, qati, gi, tqi, limaiq**) and two grammatical ending suffixes (**ta, ra**).

A single word can be used to express what would be a whole sentence in English. We note that the word composition tends generally to augment the lexical constituents with multiple formative suffixing morphemes added to a word base. Full sentences are commonly made up of only one word. Moreover, the morphology is highly inflected with a variety of lexical suffixes and grammatical ending suffixes. All these linguistic aspects make the morphological segmentation task for polysynthetic languages more challenging. One of the challenges consists in determining the word that is the basic unit, then the sub-word units (Arppe *et al.*, 2017).

2.2. *Morphological segmentation of indigenous languages*

The development of a morphological segmenter for indigenous languages was not well supported due to several challenges, as indicated above. Unsurprisingly, owing to the lack of annotated data, we used an unsupervised approach, as well as the rule-based approach used numerous works. Creutz and Lagus (2007) proposed the statistical morphological segmentation method, named Morfessor, based on the Hidden Markov Model for learning unsupervised morphology, and using a hierarchical morpheme structure.

Another method shown to be successful for unsupervised morphological segmentation is the Adaptor Grammars (AG) approach, based on non-parametric Bayesian models generalizing probabilistic context-free grammar (PCFG) (Johnson, 2008). By defining a set of morphological grammar patterns, including zero or more prefixes, stems, and suffixes, the AG models are able to induce segmentation at the morpheme level. Several studies have been conducted based on extending this approach, such as those of Botha and Blunsom (2013) for learning non-concatenative morphology, Sirts and Goldwater (2013) for minimally supervised morphological segmentation, and Eskander *et al.* (2018) for unseen languages. Godard *et al.* (2018) used this approach to experiment with the word segmentation task in very low-resource African languages. Eskander *et al.* (2019) also used this approach to deal with Mexican low-resource

polysynthetic languages such as Mexicanero, Nahuatl, Yorem Nokki and Wixarika. In the current work, we also examine the efficiency of the AG-based approach on the Inuktitut language, a polysynthetic low-resource language without annotated segmented resources.

In terms of the Inuktitut language, we noted only a few studies on morphological segmentation task. Johnson and Martin (2003) proposed an unsupervised technique, with the hubs concept in a finite-state automaton. The hubs mark the boundary between root and suffix. Concretely, Inuktitut words are segmented into morphemes and merged hubs in a finite-state automaton. They reported good performance for English morphological analysis, using the text of *Tom Sawyer*, for which they obtained 92.15% in terms of precision. However, for Inuktitut morphological analysis, they reported 31.80% precision and a low recall of 8.10%. They argued the poor performance for Inuktitut roots was due to the difficulty of identifying word-internal hubs. Farley (2012) proposed hand-crafted grammar rules and a finite-state transducer to build a morphological analysis for Inuktitut, called *Uqailaut* (pronounced Uqa·Ila·Ut). This Uqailaut project is a rule-based system based on regular morphological variations of about 3,200 head (or prefix), 350 lexical, and 1,500 grammatical morphemes, with heuristics for ranking the various readings. Nicholson *et al.* (2012) used a word alignment error rate with the dataset of English-Inuktitut Nunavut parallel corpora to evaluate the morphological analyzer for Inuktitut. They reported their best experimental results, in terms of the head (or prefix) approach, which, in Inuktitut, corresponds to the first one or two syllables of a token, with 79.70% precision and 92.20% recall. They reported that the analyzer was able to provide at least a single analysis for approximately 218k Inuktitut types (65%) from the Nunavut Hansard corpus. In addition, Micher (2017), inspired by Farley (2012) Uqailaut project, used a segmental recurrent neural network approach based on the output of this morphological analyzer for Inuktitut. The models were trained with approximately 23k types having a single analysis from the Uqailaut analyzer, with 85.07% in terms of F-measure.

2.3. *Machine translation for indigenous languages*

Machine translation (MT) is well known in language technologies. Building a reliable, high quality MT system is still a significant challenge for indigenous languages. Mager *et al.* (2018) reported an interesting and ongoing research problem in the MT task of low-resource languages, especially indigenous languages. We reviewed the development of MT systems for indigenous languages based on the following fundamental approaches: (1) rule-based, (2) statistics-based, and (3) neural network-based approaches.

(1) Rule-Based Machine Translation (RBMT) approaches are usually applied in the low-resource languages scenario. The RBMT systems do not require aligned parallel corpora. However, they require language-dependent knowledge. They have several drawbacks, mostly pertaining to translating complex structures and to building

complex rules. Apertium [5] is a free and open-source platform for developing rule-based machine translation systems. Recently, research on data-driven approaches has improved to deal with data scarcity and data sparsity (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. The performance of SMT systems is highly dependent on the quantity of training data, which represents a challenge when dealing with low-resource conditions. In the case of the native languages of the Americas, the SMT systems were challenged by the rich and complex morphology and the data sparseness (Micher, 2017) of the languages. We examined a variety of applications of this research and its foundation in the SMT line of research. Sennrich *et al.* (2016) proposed using byte pair encoding (BPE) to segment words into subword units and showed improvement in machine translation on an English to German and English to Russian task of up to 1.1 and 1.3 BLEU, respectively. Micher (2018) reported 30.04 BLEU in the English to Inuktitut direction, and 30.35 BLEU in the Inuktitut to English direction, using the BPE-preprocessed the Nunavut Hansard Inuktitut-English parallel corpora.

(3) Neural network-based Machine Translation (NMT) approaches use neural networks architectures trained with vast amounts of parallel texts. In this approach, the NMT systems are applied in several neural networks architectures such as Seq2Seq (Sutskever *et al.*, 2014), Transformer with Encoder-Decoder and Attention (Vaswani *et al.*, 2017). These systems work well when dealing with resource-rich language pairs because the training requires a significant quantity of parallel texts.

In Machine Translation task for indigenous languages, several NMT systems were presented at the WMT 2020 workshop [6] for multiple languages pairs, including Inuktitut-English. We compare our NMT model against some of them in the sub-Section 4.3.

Building an MT system for indigenous languages is considered a low-resourced scenario (Schwartz *et al.*, 2020). For many low-resourced language pairs, the corpora are derived from religious sources (*e.g. the Bible or Koran*) or technical documents (*e.g. Opus* (Tiedemann, 2012)), or from IT data localization (e.g. from open-source projects such as *GNOME or Kubuntu*) (Haddow *et al.*, 2021). Recently, Nicolai *et al.* (2021) built the JHU Bible corpus for MT of the indigenous languages of North America, with $26k$ verses in the Inuktitut family language, and it achieved only 11.8 in terms of BLEU score. Joanis *et al.* (2020) constructed $1.29M$ bilingual sentences in the Nunavut Hansard for Inuktitut-English (third edition), available for research purposes. Research on NMT with low-resource language pairs still face multiple compounding major challenges, such as lack of NLP tools, lack of parallel corpora, out-of-domain data, and noisy data (Littell *et al.*, 2018; Mager *et al.*, 2018; Joanis *et al.*, 2020; Le and Sadat, 2020; Mager *et al.*, 2021). Aside from data problems, indigenous languages are

5. Apertium: `https://wiki.apertium.org/wiki/Main_Page`.
6. Source: `http://www.statmt.org/wmt20/translation-task.html`.

frequently understudied languages, in which access to local speakers and specialists is difficult, and even fundamental toolkits such as language identifiers or morphological analyzers do not exist or are not trustworthy (Haddow *et al.*, 2021).

3. Proposed methodology

3.1. *Supervised approach for the morphological segmentation*

The neural network-based approach can be efficiently applied on word segmentation using pretrained embeddings and several deep learning techniques. Furthermore, using additional linguistic factors helps the neural model perform better, especially when dealing with data sparseness or language ambiguity in the context of indigenous languages (Kann *et al.*, 2018).

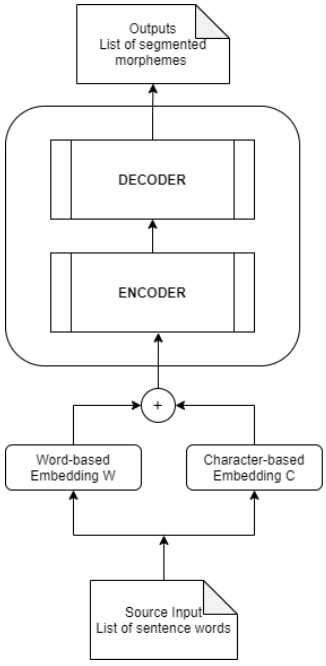

Figure 1. *Architecture of our framework: Morphological segmentation for indigenous language based on the encoder-decoder architecture.*

The goal of morphological segmentation is to divide words into morphemes. This task may be thought of as a structured classification problem, with each character being allocated to one of many predefined classes. These classes are denoted as follows: (B) represents the beginning of a multi-character morpheme, (M) the middle of a multi-character morpheme, and (E) the end of a multi-character morpheme,

and (S) denotes a single character morpheme. Other schemes are also conceivable such as IOB format (short for inside, outside, beginning) or IO or BMEO or BM (Carpenter, 2009; Ruokolainen *et al.*, 2013; Wang *et al.*, 2016).

For instance, for the Inuktitut word "*tusaattialaurit*" (meaning: to listen), the corresponding morphological segmentation should be:

tusaa+tti+ala+u+rit

By adding the two extra symbols <w> and </w> to indicate the start and the end of a word, respectively, the above segmentation form is represented as follow:

<w> tusaa tti ala u rit </w>
START BMMME BME BME S BME STOP

In this research, the morphological segmentation task is considered as a sequence labeling task, with the goal of classifying each character in a word into the appropriate class. Given an input sequence, $W = [w_0, w_1, ..., w_m]$ and $C = [c_0, c_1, ..., c_n]$ contain all the input words and the input characters. The architecture is based on Sequence-to-Sequence model (Sutskever *et al.*, 2014) with the encoder-decoder layers as shown in Figure 1. The encoder layer contains the input word sequence transformation by concatenating pretrained character-based and word-based embeddings, with the state $S = \langle W, C \rangle$. We apply the attention mechanism that allows the model to focus, in the context, on a set of characters and to learn the important letters to better predict whether a character forms a boundary. We introduce an attention vector, a_i, used to measure the weight of the sentence words in the context. The resulting context embedding, v_i, jointly learned during the training phase, helps to capture the relevant information from the context.

$$\alpha_t = \frac{exp(score(h_t, h_{s}))}{\sum_{s'=1}^{S} exp(score(h_t, h_{s}))} \quad [1]$$

$$c_t = \sum_s \alpha_t h_s \quad [2]$$

$$a_t = f(c_t, h_t) = tanh(W_c[c_t; h_t]) \quad [3]$$

where α_t is the attention weight of the target words in the context, h_s and h_t are the weight of the hidden layer for source and target words, respectively, c_t is the context vector and a_i is the attention vector.

The decoder layer $P(y|s_i)$ calculates the activation function θ as an output function and displays the output hypothesis, where y is the output prediction, s_i is a word sentence, and b_i is a bias.

$$P(y|s_i) = \theta(W_h \cdot a_i + b_i) \quad [4]$$

3.2. *Unsupervised approach for the morphological segmentation*

Inspired by the work of Eskander *et al.* (2019), we adapted an unsupervised approach in learning all possible morphological patterns using Adaptor Grammars (Johnson, 2008) and fine-tuned the outputs of the first stage by building a recurrent neural network-based architecture for Inuktitut. This approach is, basically, based on the grammar containing production rules, non-terminal and terminal symbols, and a lexicon. The deep learning can successfully handle data sparsity or language ambiguities thanks to additional linguistic factors related to a set of rich, high-quality features, such as semantic distribution information, and contextual meanings that are extracted from pretrained embeddings at character level and at word level, learned from monolingual large-scale raw corpora (Kann *et al.*, 2018).

The first phase in Adaptor Grammars-based learning consists of defining the grammar, including non-terminals, terminals, and production rules. As explained in Eskander *et al.* (2019), the grammar construction relies on three main dimensions: word modeling, abstraction level, and segmentation boundaries. The grammar patterns specify the word structures where a word is considered a sequence of prefixes, a stem, and a sequence of suffixes. Moreover, each production rule has two parameters to configure, a and b, in the Pitman-Yor process (Pitman and Yor, 1997). Setting $a = 1$ and $b = 1$ indicates to the running learner that the current non-terminals are not adapted and sampled by the general Pitman-Yor process. Otherwise, the current non-terminals are adapted and expanded as in a regular probabilistic context-free grammar. The standard grammar setting (Table 1) is language-independent, and contains all possible generic patterns, whereas the scholar-seeded grammar setting (Table 2) combines all standard grammar patterns and additional language-dependent knowledge, in this case a list of affixes. By using the list of affixes and roots, called Scholar-seeded setting, we inject linguistic knowledge into the training phase. Then, we still apply the probabilistic context-free grammar (PCFG). The model is, therefore, able to learn more patterns, with non-concatenative morphology, and to induce segmentation at the morpheme level.

We fine-tuned the outputs of the first stage through a Recurrent Neural Network-based (RNN) architecture. These outputs are fed into a bidirectional Long-Short Term Memory (Hochreiter and Schmidhuber, 1997). Formally, these input sequences are numerically vectorized using pretrained embeddings, at word-level W and at character-level C representations. The hidden feature layer then merges all input features X_W, X_C in a single vector with a k-dimension, $\langle W, C \rangle$. The output layer calculates an activation function θ, where W_o is the output weight, h is the hidden layer, and b_o is its bias.

$$h = tanh(W_{hW} \cdot X_W + W_{hC} \cdot X_C) \quad [5]$$

$$output = \theta(W_o \cdot h + b_o) \quad [6]$$

1 1 Word –>Prefix Stem Suffix	Suffix –> $$$ Suffix –>SuffixMorphs $$$
Prefix –>^^^	1 1 SuffixMorphs –>SuffixMorph SuffixMorphs
Prefix –>^^^PrefixMorphs	1 1 SuffixMorphs –>SuffixMorph
1 1 PrefixMorphs –>PrefixMorph PrefixMorphs	SuffixMorph –>SubMorphs
1 1 PrefixMorphs –>PrefixMorph	1 1 SubMorphs –>SubMorph SubMorphs
PrefixMorph –>SubMorphs	1 1 SubMorphs –>SubMorph
	SubMorph –>Chars
Stem –>SubMorphs	1 1 Chars –>Char
	1 1 Chars –>Char Chars

Table 1. *The standard PrefixStemSuffix+SuffixMorph grammar for Inuktitut. The symbols ^^^and $$$ mean the beginning and the end of the word sequence, respectively. Source: Eskander* et al. *(2019).*

[All standard setting grammar in Table 1]
1 1 PrefixMorph –> (a) (u) (l) (l) (a)
1 1 PrefixMorph –> (i) (g) (l) (u)
1 1 PrefixMorph –> (q) (i) (n) (m) (i)
1 1 PrefixMorph –> (u) (t) (a) (q) (q) (i)
[...]
1 1 SuffixMorph –> (a) (n) (n) (i) (n)
1 1 SuffixMorph –> (f) (f) (a) (a) (n) (g) (m) (i)
1 1 SuffixMorph –> (g) (i) (a) (q) (t) (u) (q)
1 1 SuffixMorph –> (m) (i) (u) (t) (a) (t)
1 1 SuffixMorph –> (n) (') (n) (g) (u) (l) (i) (q)
1 1 SuffixMorph –> (y) (u) (t)
[...]
1 1 Char –> (q)
1 1 Char –> (k)
[...]
1 1 Char –> (p)
1 1 Char –> (t)

Table 2. *The scholar-seeded PrefixStemSuffix+SuffixMorph grammar for Inuktitut, with prefixes, suffixes, and characters.*

3.3. *Uqailaut morphological analyzer for Inuktitut*

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. Heuristics make it possible to choose the shortest path for the morphological analysis. For example, *tusaattialau-*

rit is segmented as ***tusaa*** *tti ala u rit* or ***tusaa*** *ttia lau rit* or ***tusaa*** *ttia la u rit* (Table 3). The root ***tusaa*** means *to listen*, *tti*, *ala*, *u* are lexical suffixes, and *rit* is a grammatical suffix.

Morphological Segmentation	**Output**
Raw text	tusaattialaurit
Reference	**tusaa** tti ala u *rit*
First best prediction	**tusaa** tti ala u *rit*
Second best prediction	**tusaa** ttia lau *rit*
Third best prediction	**tusaa** ttia la u *rit*

Table 3. *Predictions of the Inuktitut morphological segmentation by the Uqailaut analyzer (Meaning: please listen).*

3.4. *Byte-Pair Encoding segmentation*

Sennrich *et al.* (2016) proposed the Byte-Pair Encoding (BPE) method for the word segmentation task. This method consists of unsupervised word segmentation that tries to break words into subword units, which aids in dealing with unusual and unfamiliar terms.

BPE uses the minimum entropy on subword units, often known as tokens, with a given vocabulary size. Although these tokens resemble morphemes, the BPE segmentation model is based on training data rather than linguistic knowledge bases.

For example, in Inuktitut, *'tusaattialaurit*' (meaning: *please listen* in English) may be segmented as *'tusaa@@ tti@@ alau@@ rit'* (Table 4). This word should be correctly segmented *'tusaa@@ tti@@ ala@@ u@@ rit'*, in the case of large-scale training data.

Method	**Morphological Segmentation Output**
Raw text	tusaattialaurit
Reference	tusaa tti ala u rit
Uqailaut analyzer (Farley, 2012)	tusaa tti ala u rit
BPE (Sennrich *et al.*, 2016)	tusaa@@ tti@@ ala@@ u@@ rit
Our proposed approach	tusaa tti ala u rit

Table 4. *Illustration of several Inuktitut word segmentation methods. The symbol @@ represents an in-word morpheme boundary (Meaning: please listen).*

3.5. *Polysynthetic indigenous language NMT*

The second phase of our framework consists of building an NMT for indigenous language to English based on the Transformer encoder-decoder architecture (Vaswani *et al.*, 2017). We apply our morphological segmentation method to preprocess the source indigenous language in the context of an Inuktitut-English NMT system.

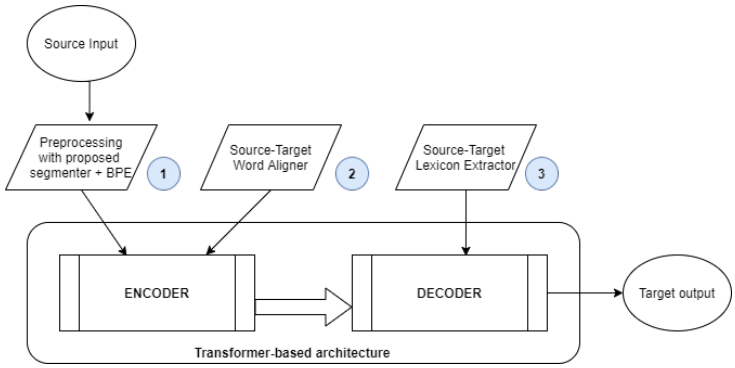

Figure 2. *Architecture of our Inuktitut-English NMT with three main parts: (1) Preprocessing with our proposed morphological segmenter and Byte Pair Encoding (BPE) for both source and target languages, respectively, (2) Building a source-target word aligner and (3) Building a source-target lexicon extractor.*

3.5.1. *Word Aligner and Lexicon Extractor for NMT*

This architecture aims to investigate our Inuktitut-English neural machine translation system while using the word alignment information and the source-target lexicon. Our approach consists of three main parts. First, the source input is preprocessed by applying our proposed morphological segmenter and Byte Pair Encoding (BPE) (Sennrich *et al.*, 2016) for both source and target languages, respectively. Second, the word alignment information is extracted from the bilingual parallel corpus and is fed into the encoder. Third, we prepare a bilingual source-target lexical shortlist. This bilingual lexicon is then used during the decoding.

3.5.2. *Multilingual NMT architecture*

Adding data from multiple languages, *i.e.* multilingual NMT, can enhance the performance of NMT systems (Aharoni *et al.*, 2019). We adapt this approach in the context of low-resource indigenous languages by using several closely-related languages (Figure 3).

For each language pair, a BPE-based model is learned jointly from the source-target sides of the parallel corpora using *subword-nmt* (Sennrich *et al.*, 2016). In addition, the source-side indigenous languages, here Inuktitut and Inuinnaqtun, are segmented by our proposed word segmenter. Then the joint BPE model is applied to all the training datasets. Moreover, we apply the BPE drop-out (Provilkov *et al.*, 2019) to deal with data sparsity and morphological complexity, such as orthographic variation or spelling errors.

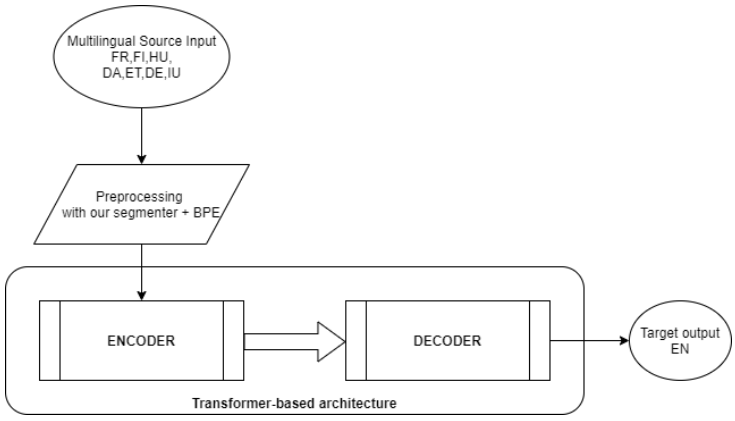

Figure 3. *Architecture of our multilingual NMT system. Here, multilingual source input is composed of multiple related languages such as French-fr, Finnish-fi, Danish-da, German-de, Hungarian-hu, Estonian-et and Inuktitut-iu, and the target output is English-en.*

4. Experiments

4.1. *Data preparation*

In our experiments and evaluations, we used the third edition of the Inuktitut-English Nunavut Hansard (Joanis *et al.*, 2020) to train our models. This parallel corpus contains 1,293,348 training sentences, 5,433 development sentences and 6,139 testing sentences, respectively. Furthermore, in order to develop our multilingual NMT model, we used several parallel corpora, including multiple language sources with an English target, provided from the shared task of WMT 2020. Tables 5 and 6 describe the statistics of the training corpora.

	#tokens	**#train**	**#dev**	**#test**
IU	20,657,477	1,293,348	5,433	6,139
EN	10,962,904	1,293,348	5,433	6,139

Table 5. *Statistics of the Nunavut Hansard for Inuktitut-English.*

Inuktitut corpus is transformed from syllabic to roman using the *uniconv* toolkit [7]. We then apply consistent preprocessing with English defaults on both source and target languages of the parallel corpora using Moses (Koehn *et al.*, 2007) scripts such as punctuation normalization, tokenization, cleaning the training corpus and truecasing on the training datasets.

7. Toolkit *uniconv* with Yudit: `www.yudit.org`.

Source-Target [English]	#train	#dev	#test
Finnish (fi-en)	1,918,232	1,000	–
French (fr-en)	2,002,165	1,000	–
Hungarian (hu-en)	623,448	1,000	–
Danish (da-en)	1,949,393	1,000	–
Estonian (et-en)	651,746	1,000	–
German (de-en)	1,920,209	1,000	–
Inuktitut (iu-en)	1,293,348	5,433	6,139
Inuinnaqtun (ikt-en)	3,511	–	–

Table 6. *Statistics of all corpora for multilingual NMT model training (Source: WMT 2020). Inuinnaqtun corpus is extracted from the Nunavut Government Website: https://www.gov.nu.ca/in/.*

For the Inuinnaqtun dataset, we manually collected a small corpus from several resources such as the Nunavut Website [8] government, open source dictionaries and grammar books (Lowe, 1985; Kudlak and Compton, 2018). The experimental corpus contains 190 word bases and 571 affixes. A small golden testing set was manually crafted containing 1,055 unique segmented words.

4.2. *Training settings*

To train the supervised morphological segmentation model, we adapted the *RichWordSegmenter* toolkit (Yang *et al.*, 2017). We chose Inuktitut source from the Nunavut Hansard to perform experiments. Then, using the Uqailaut toolkit (Farley, 2012), we annotated $11k$ training sentences, 250 development sentences, and 250 testing sentences. To pre-train the character-based and word-based embeddings, we used the Nunavut Hansard Inuktitut corpus 3.0 and the *Gensim* [9] library to train all embeddings with a dimension of 30 and 50, respectively. We found that there are only 97,785 unique terms for the word-based vocabulary, 102 unique terms for the character-based vocabulary and 1,406 unique terms for the character-based vocabulary (Table 7).

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

Table 7. *Statistics of word-based and (bi)character-based embeddings training using Nunavut Hansard Inuktitut-English parallel corpus 3.0 for Inuktitut.*

8. Nunavut government Website in Inuinnaqtun: `https://www.gov.nu.ca/in/cgs-in`.
9. Gensim library: `https://radimrehurek.com/gensim/models/word2vec.html`.

The two principal inputs, used to train the Adaptor Grammars-based unsupervised morphological segmentation model, consist of the grammar and the lexicon of the language. The learning hyperparameters are configured as in Eskander *et al.* (2019) according to the best standard PrefixStemSuffix+SuffixMorph grammar (Table 1) and the best scholar-seeded grammar (Table 2). Next, we fine-tuned the outputs of the first stage with an RNN-based architecture consisting of bi-directional Long Short-term Memory, with 200 neurons in the hidden layer. We evaluated both supervised and unsupervised proposed morphological segmentation models versus the baseline, for example Morfessor 2.0 (Virpioja *et al.*, 2013).

To train the baseline morphological segmenter, we used Morfessor 2.0 toolkit [10] with Python interpreter. The training, development and testing datasets for Morfessor are the same as the datasets used to train our proposed segmenter for both Inuktitut and Inuinnaqtun, respectively. We filtered out all tokens of the corpus which are not included in the corresponding word list. These smaller datasets were also used in the semi-supervised training experiments. The F1 scores converged after 5 iterations for all runs. As the evaluation metric, we used the micro-average segmentation boundary F1-score. The scores were calculated based on the word types in the testing sets.

To train our NMT model, we first 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 hyperparameter settings: 6-layer depth for both encoder and decoder, 8-layer multi-heads, embedding dimension of 512, hidden layers of 2,048 units in the feed-forward networks, with optimizer Adam and an initial learning rate of 0.0003. For the architecture type, we could choose either the Seq2Seq (Sutskever *et al.*, 2014) or the Transformer (Vaswani *et al.*, 2017) inside the toolkit. We performed multiple NMT experiments as follows:

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

– System 1 + align information: We incorporated source-target word alignment information in the training step. We applied an unsupervised word aligner, *fast_align* (Dyer *et al.*, 2013) to generate symmetrized source-target alignments, trained on BPE preprocessed data;

– System 1 + lex.s2t: We combined the source-target bilingual lexicon, during the decoding phase, in the baseline system. We applied the lexicon extractor from *Moses* (Koehn *et al.*, 2007) to prepare a bilingual lexical shortlist;

– System 1 + align information + lex.s2t: We combined both word alignment information and the source-target bilingual lexicon in the baseline system;

– Systems 2, 3, 4, 5: We configured the proposed morphological segmentation using the standard or scholar-seeded settings combined with the sequence-to-

10. Morfessor 2.0 toolkit: `https://morfessor.readthedocs.io/en/latest/index.html`.

sequence based or the Transformer-based architectures for our NMT model, named AG-Standard+s2s, AG-Scholar+s2s, AG-Standard+TF, AG-Scholar+TF, respectively;

– Multilingual NMT system (multiNMT): We performed the following experiments applying the word segmentation for the source-side indigenous languages, *e.g.* Inuktitut, Inuinnaqtun, within different multilingual NMT systems:

- (multiNMT) We chose, for this baseline, the same configuration as described in Joanis *et al.* (2020), with only BPE-preprocessed data, with all source-target language pairs and the test set on Inuktitut-English only,

- (multiNMT-1) The training datasets are without segmenting any indigenous language (Inuktitut, Inuinnaqtun),

- (multiNMT-2) The source-side training datasets are segmented only for Inuktitut but not for Inuinnaqtun,

- (multiNMT-3) The source-side training datasets are segmented for both Inuktitut and Inuinnaqtun,

- (multiNMT-4) The source-side training datasets are segmented for Inuinnaqtun but not for Inuktitut.

4.3. *Evaluations and discussion*

4.3.1. *Morphological segmentation task*

We evaluated the morphological segmentation system using the automatic metrics: *Precision (P)*, *Recall (R)*, and *F1 score*.

For the supervised morphological segmentation, we evaluated only the Inuktitut data source. As described in Table 8, our proposed system, with all pretrained embeddings, showed a good performance with 75.33% in terms of F1 score. However, the Morfessor system outperformed our proposed system, with a gain of +4.37 points in terms of F1 score. Using the additional golden annotated data with the training data, the Morfessor model obtained better precision and recall than our proposed model, with a gain of +1.36 points and +6.83 points. In addition, the Morfessor model used an n-best Viterbi algorithm that allows extraction of all posible segmentations for a compound and the probabilities of the segmentation.

	Precision	**Recall**	**F1 score**
Morfessor	82.15	77.40	**79.70**
supervised_Inuktitut_WS	80.79	70.57	75.33

Table 8. *Results for Inuktitut supervised morphological segmentation.*

For the unsupervised morphological segmentation, we evaluated both Inuktitut and Inuinnaqtun data sources. Table 9 shows the performance and results of our models versus Morfessor for the polysynthetic language on the Inuktitut test set. The standard setting is better than the baseline, with a gain of +8.30 points in terms of precision,

on the test set, compared with Morfessor. Moreover, we also observed large gains of +8.92 points in terms of precision, on the test set, when using the scholar-seeded setting compared with Morfessor. All models obtained low recall between 77.40% and 82.33%, including Morfessor, due to the under-segmentation.

Inuktitut	Precision	Recall	F1 score
Morfessor	82.15	77.40	79.70
AG-Standard	90.45	81.51	85.75
AG-Scholar	**91.07**	**82.33**	**86.48**

Table 9. *Morphological segmentation task: Results for the Inuktitut test set using the Standard setting (AG-Standard), Scholar seeded setting (AG-Scholar), and Morfessor toolkit. Values in bold refer to the best performances.*

Table 10 shows the performance results of our models versus Morfessor for the polysynthetic language using the Inuinnaqtun test set. Both AG-based models outperformed the baseline, with gains of +3.33%, +17.62% in terms of F1 score, for the AG-standard setting and AG-Scholar setting, respectively. The recall of all three models is good enough to recognize all possible patterns, with 75.40%, 80.30%, and 82.83% for the baseline, AG-standard setting and AG-Scholar setting, respectively. However, the baseline and the AG-Standard model obtained low precision with 48.29% and 50.76%, respectively, compared with the AG-scholar-seeded model, which obtained 71.06%. We observed an under-segmentation in these models.

Inuinnaqtun	Precision	Recall	F1 score
Morfessor	48.29	75.40	58.87
AG-Standard	50.76	80.30	62.20
AG-Scholar	**71.06**	**82.83**	**76.49**

Table 10. *Morphological segmentation task: Results for the Inuinnaqtun test set using the Standard setting (AG-Standard), Scholar seeded setting (AG-Scholar), and Morfessor toolkit. Values in bold refer to the best performances.*

We noted that our proposed models could not correctly recognize more complex morphemes due to the languages' linguistic irregularities and rich morphophonemics. In particular, they were unable to detect common affixes such as *ag*, *ik*, *iq*, *mi*, *ti* or *ut* in Inuktitut and common lexical suffixes such as *at*, *aq*, *iq*, *na*, *ng* or grammatical ending suffixes such as *a*, *k*, *q*, *t*, *n*, *it*, *mi* or *uk* in Inuinnaqtun.

Tables 11 and 12 illustrate some predictions by all the models and the performance of our models on Inuktitut and Inuinnaqtun, respectively.

4.3.2. *Machine translation task*

We conducted additional evaluations of Machine Translation task based on the BLEU scores (Papineni *et al.*, 2002) which were computed with lowercase and *v13a* tokenization, using *sacrebleu* (Post, 2018). We also used chrF++ (Popović, 2015) to

Segmentation	Sentence Example
Raw text	niqtunaqtuq piku tusaattialaurit
Reference	niqtu naq tuq piku tusaa tti ala u rit
Uqailaut analyzer (Farley, 2012)	niqtu naq tuq piku tusaa ttia lau rit
BPE (Sennrich *et al.*, 2016)	niqtunaqtuq piku tusa@@ atti@@ alaur@@ it
Our proposed approach	niqtu naq tuq piku tusaa tti ala u rit

Table 11. *Illustrations of the Inuktitut word segmentation (Meaning: Mr. Picco, please listen).*

Word	Ground Truth	Morfessor	AG-Standard	AG-Scholar
aullarnanga	aullar na nga	aulla rn anga	aulla rna nga	aullar na nga
aullaqtinnatin	aullaq tinna tin	aulla q ti nna t in	aulla q tinna tin	aullaq tinna tin
igluptun	iglu ptun	iglu p tun	iglu ptun	iglu ptun
nattiqhiuqtuq	nattiq hiuq tuq	nattiq hi uq tu q	nattiq hi uq tuq	nattiq hiuq tuq
kitkungnin	kitku ngnin	kitku ng nin	kitku ng ni n	kitku ngnin
tupaktuhi	tupak tu hi	tupa k tu h i	tupak tu hi	tupak tu hi
iqaluktinnagu	iqaluk tinna gu	iqaluk ti nna gu	iqaluk tin na gu	iqaluk tin na gu

Table 12. *Illustrations of Inuinnaqtun morpheme segmentation predictions on the test set using the different settings such as Standard (AG-Standard), Scholar seeded (AG-Scholar), and Morfessor. Red text indicates deviations in segmentation from the Ground Truth.*

calculate the F1-score averaged on character n-gram precision and recall enhanced with word n-grams for the translation references and their hypotheses.

Experiment	BLEU (dev set)	BLEU (test set)	chrF++
System 1 (*Baseline*)	41.40	35.00	65.40
System 1 + align information	41.45	35.71	65.59
System 1 + lex.s2t	41.66	35.93	65.97
System 1 + align information + lex.s2t	41.78	**36.03**	66.30

Table 13. *Performance on Inuktitut-English NMT in terms of lowercase word BLEU score, using only the BPE subword segmentation method.*

We observed that combining the word alignment information and the source-target bilingual lexicon had a positive impact on the performance of the NMT model. Compared to the baseline, with all the additional features, the NMT system obtained a gain of +1.03 points in terms of BLEU score (Table 13). However, using only the BPE subword segmentation method, the multilingual NMT system outperformed the system 1 and all variants, with a gain of +3.06 points in terms of BLEU score (Table 14).

To go further, we performed multiple variants of the multilingual NMT systems, with and without applying our proposed word segmenter to the source-side indigenous languages (Table 14). We noted that the multiNMT-1 system obtained the worst

Experiment	BLEU (test set)	chrF++
multiNMT (baseline, only BPE method)	**38.06**	**68.15**
multiNMT-1 (-INU_segmented, -IKT_segmented)	8.11	14.52
multiNMT-2 (+INU_segmented, -IKT_segmented)	40.91	73.25
multiNMT-3 (+INU_segmented, +IKT_segmented)	**41.40**	**74.13**
multiNMT-4 (-INU_segmented, +IKT_segmented)	38.08	68.19

Table 14. *Performance of all the multilingual NMT for Inuktitut-English, with and without applying our proposed segmenter, where INU and IKT refers to Inuktitut and Inuinnaqtun, respectively.*

performance, only 8.11% in terms of BLEU score or 14.52% in terms of chrF++, due to without any indigenous languages. The best performance was obtained by the multiNMT-3, 41.40% in terms of BLEU score or 74.13% in terms of chrF++. We observed that the translation quality were significantly improved as we segmented indigenous languages in the source side rather than other related languages, up to +3.34 points versus the multiNMT baseline (Table 14). It means that is sufficient to segment the source-side indigenous languages to have a better performance. The related languages are not necessarily required for word segmentation.

Moreover, we tested other NMT systems with our proposed morphological segmentation based on Adaptor Grammars. All our systems 2, 3, 4 and 5 outperformed the baseline with gains of up to +2.98 points and +3.41 points in terms of BLEU score, on the development set and the test set, respectively (Table 15), compared to the baseline.

Experiment	dev	test	chrF++
System 1-*Baseline*-(Joanis *et al.*, 2020)	41.40	35.00	65.40
System 2 (AG-Standard+s2s)	43.93	37.78	66.43
System 3 (AG-Scholar+s2s)	**44.38**	**38.41**	**68.71**
System 4 (AG-Standard+TF)	44.18	38.28	68.41
System 5 (AG-Scholar+TF)	44.41	38.32	68.61

Table 15. *Performance on Inuktitut-English NMT in terms of lowercase word BLEU score, with our proposed morpheme segmenter.*

We compared our best system against other NMT systems from WMT 2020 using morphological segmentation methods, such as Roest *et al.* (2020), and Knowles *et al.* (2020). Our best system outperformed all the NMT systems from WMT 2020 in terms of BLEU score on the third version of the Nunavut Hansard test set, with 38.41% versus 30.05% and 29.90% (Table 16).

Roest *et al.* (2020) reported their best NMT system results due to multiple reasons. First, they applied three methods of segmentation: unsupervised such as BPE, LMVR (Ataman *et al.*, 2017) and 3-step segmentation. They varied the value of the decoder's penalty length based on results on the development set with 0.8 for news and 1.4

Experiment	dev	test
System (Knowles *et al.*, 2020)	–	29.90
System (Roest *et al.*, 2020)	–	30.05
Our best system 3 (AG-Scholar+s2s)	**44.38**	**38.41**

Table 16. *Comparison of performance results of our best system on Inuktitut-English NMT in terms of lowercase word BLEU score with other best systems of WMT 2020.*

System	Sentence Example
Raw	apiqqutiga turaaqtittumajara aanniaqarnangittulirijikkut ministangannut.
Reference	I would like to direct my question to the Minister of Health.
Baseline	This is a question for the Minister of Health.
System 1	My question is for the Minister responsible for Health.
System 2	My question is directed for the Minister of Health.
System 3	I would like to direct my question to the Minister of Health.
System 4	My question is directed for the Minister of Health.
System 5	I would like to ask my question for the Minister of Health.
MultiNMT	This is my question directed for the Minister responsible for Health.

Table 17. *Illustrations of some translation predictions using different NMT systems, from Inuktitut to English.*

for Hansards, respectively. Furthermore. there was a mixture of in-domain and out-domain training data. Finally, the use of ensembling and fine-tuning on all NMT systems helped to improve the BLEU performance.

In the case of Knowles *et al.* (2020), the final systems were trained on a mix of news and Hansard data, using joint BPE, BPE-dropout, tagged back-translation for Inuktitut-English, fine-tuning, ensembling, and the use of domain-specific models.

We assume the preprocessing as word segmentation helped to solve the complex morphology of Inuktitut at source-side. Our proposed NMT model outperformed the state-of-the-art, as presented in Joanis *et al.* (2020), using only BPE-preprocessed training data, with the best performance of 44.38% and 38.41% in terms of BLEU on the development set and the test set, respectively.

5. Conclusion and Perspectives

In this paper, we empirically explored different word segmentation techniques on both Inuktitut and Inuinnaqtun. We then proposed a novel morphological segmentation technique that may be applied to any indigenous language.

In the supervised approach, the neural networks-based word segmentation model showed promising results, but not good enough, due to multiple factors, such as the quantity of the annotated data and the quality of the pretrained embedding models.

In the semi-supervised and non-supervised approaches, the Adaptor Grammars-based word segmentation models yielded better results, employing a collection of grammatical rules from grammar books, and a lexicon from relatively little data. We applied our word segmenter to preprocess the Inuktitut source-side language before implementing an Inuktitut-English NMT system. We empirically evaluated our proposed NMT method against several baseline NMT architectures. Our proposed NMT system outperformed the state-of-the-art, as described in Joanis *et al.* (2020), with just BPE-preprocessed training data.

Our study makes an important contribution by focusing on morpheme segmentation in the source-side indigenous language. This significantly enhances the MT performance in the low-resource scenario. Furthermore, the NLP community is becoming increasingly interested in indigenous languages. Indigenous language research might lead to a more thorough knowledge of human languages and the development of universal NLP models.

In the future, we plan to add more annotated data and study other domain-specific characteristics to increase the segmentation model's accuracy. Moreover, we are developing a multilingual NMT framework in order to include more indigenous languages, particularly endangered ones, with the goal of preserving and revitalizing endangered and indigenous languages, as well as their legacy and culture. Globally, our research interest focuses on an inclusive, fairer and more equitable and responsible Artificial Intelligence, while emphasizing on the revitalization and preservation of indigenous languages. We have encountered a variety of challenges, including language skills, data gathering, and validation, to name a few. Thus, we seek to conduct research "by and with" indigenous peoples, which will help validate the results and construct more reliable linguistic resources that will be, we hope, of great help to the indigenous communities.

6. References

Aharoni R., Johnson M., Firat O., "Massively multilingual neural machine translation", *arXiv preprint arXiv:1903.00089*, 2019.

Arppe A., Schmirler K., Harrigan A. G., Wolvengrey A., "A Morphosyntactically Tagged Corpus for Plains Cree", *49th Algonquian Conference, Montreal, Quebec*, p. 27-29, 2017.

Ataman D., Negri M., Turchi M., Federico M., "Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English", *arXiv preprint arXiv:1707.09879*, 2017.

Botha J. A., Blunsom P., "Adaptor Grammars for Learning Non- Concatenative Morphology", Association for Computational Linguistics, 2013.

Carpenter B., "Coding chunkers as taggers: Io, bio, bmewo, and bmewo+", *LingPipe Blog*p. 14, 2009.

Creutz M., Lagus K., "Unsupervised models for morpheme segmentation and morphology learning", *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 4, n° 1, p. 1-34, 2007.

Dyer C., Chahuneau V., Smith N. A., "A simple, fast, and effective reparameterization of ibm model 2", *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 644-648, 2013.

Eskander R., Klavans J. L., Muresan S., "Unsupervised Morphological Segmentation for Low-Resource Polysynthetic Languages", *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, p. 189-195, 2019.

Eskander R., Rambow O., Muresan S., "Automatically tailoring unsupervised morphological segmentation to the language", *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, p. 78-83, 2018.

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

Gasser M., "Computational morphology and the teaching of indigenous languages", *Indigenous Languages of Latin America Actas del Primer Simposio sobre Enseñanza de Lenguas Indígenas de América Latina*, p. 52, 2011.

Godard P., Besacier L., Yvon F., Adda-Decker M., Adda G., Maynard H., Rialland A., "Adaptor grammars for the linguist: Word segmentation experiments for very low-resource languages", *Workshop on Computational Research in Phonetics, Phonology, and Morphology*, p. 32-42, 2018.

Haddow B., Bawden R., Barone A. V. M., Helcl J., Birch A., "Survey of Low-Resource Machine Translation", *arXiv preprint arXiv:2109.00486*, 2021.

Hochreiter S., Schmidhuber J., "Long Short-Term Memory", *Neural Comput.*, vol. 9, nº 8, p. 1735-1780, November, 1997.

Joanis E., Knowles R., Kuhn R., Larkin S., Littell P., Lo C.-k., Stewart D., Micher J., "The Nunavut Hansard Inuktitut English Parallel Corpus 3.0 with Preliminary Machine Translation Results", *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 2562-2572, May, 2020.

Johnson H., Martin J., "Unsupervised learning of morphology for English and Inuktitut", *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*, Association for Computational Linguistics, p. 43-45, 2003.

Johnson M., "Unsupervised word segmentation for Sesotho using adaptor grammars", *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, p. 20-27, 2008.

Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 6282-6293, July, 2020.

Junczys-Dowmunt M., Grundkiewicz R., Dwojak T., Hoang H., Heafield K., Neckermann T., Seide F., Germann U., Aji A. F., Bogoychev N. *et al.*, "Marian: Fast neural machine translation in C++", *arXiv preprint arXiv:1804.00344*, 2018.

Kann K., Mager M., Meza-Ruiz I., Schütze H., "Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages", *arXiv preprint arXiv:1804.06024*, 2018.

Knowles R., Stewart D., Larkin S., Littell P., "NRC Systems for the 2020 Inuktitut-English News Translation Task", *Proceedings of the Fifth Conference on Machine Translation*, p. 156-170, 2020.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R. *et al.*, "Moses: Open source toolkit for statistical machine translation", *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, p. 177-180, 2007.

Kudlak E., Compton R., *Kangiryuarmiut Inuinnaqtun Uqauhiitaa Numiktitirutait — Kangiryuarmiut Inuinnaqtun Dictionary*, vol. 1, Nunavut Arctic College: Iqaluit, Nunavut, 2018.

Le T. N., Sadat F., "Revitalization of Indigenous Languages through Pre-processing and Neural Machine Translation: The case of Inuktitut", *Proceedings of the 28th International Conference on Computational Linguistics*, p. 4661-4666, 2020.

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

Lowe R., *Basic Siglit Inuvialuit Eskimo Grammar*, vol. 6, Inuvik, NWT: Committee for Original Peoples Entitlement, 1985.

Mager M., Gutierrez-Vasques X., Sierra G., Meza-Ruiz I., "Challenges of language technologies for the indigenous languages of the Americas", *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, p. 55-69, August, 2018.

Mager M., Oncevay A., Ebrahimi A., Ortega J., Gonzales A. R., Fan A., Gutierrez-Vasques X., Chiruzzo L., Lugo G. G., Ramos R. *et al.*, "Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas", *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, p. 202-217, 2021.

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

Micher J., "Using the Nunavut Hansard data for experiments in morphological analysis and machine translation", *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, p. 65-72, 2018.

Mithun M., "Morphological complexity and language contact in languages indigenous to North America", *Linguistic Discovery*, vol. 13, n° 2, p. 37-59, 2015.

Nicholson J., Cohn T., Baldwin T., "Evaluating a morphological analyser of Inuktitut", *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, p. 372-376, 2012.

Nicolai G., Coates E., Zhang M., Silfverberg M., "Expanding the JHU Bible Corpus for Machine Translation of the Indigenous Languages of North America", *Proceedings of the Workshop on Computational Methods for Endangered Languages*, vol. 1, p. 1-5, 2021.

Papineni K., Roukos S., Ward T., Zhu W.-J., "BLEU: a method for automatic evaluation of machine translation", *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 311-318, 2002.

Pitman J., Yor M., "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator", The Annals of Probability – JSTOR, p. 855-900, 1997.

Popović M., "chrF: character n-gram F-score for automatic MT evaluation", *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 392-395, 2015.

Post M., "A Call for Clarity in Reporting BLEU Scores", *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, Brussels, Belgium, p. 186-191, October, 2018.

Provilkov I., Emelianenko D., Voita E., "Bpe-dropout: Simple and effective subword regularization", *arXiv preprint arXiv:1910.13267*, 2019.

Rice K., "Documentary linguistics and community relations", *Language Documentation & Conservation*, vol. 5, p. 187-207, 2011.

Roest C., Edman L., Minnema G., Kelly K., Spenader J., Toral A., "Machine Translation for English–Inuktitut with Segmentation, Data Acquisition and Pre-Training", *Proceedings of the Fifth Conference on Machine Translation*, p. 274-281, 2020.

Ruokolainen T., Kohonen O., Virpioja S., Kurimo M., "Supervised morphological segmentation in a low-resource learning setting using conditional random fields", *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, p. 29-37, 2013.

Schwartz L., Tyers F., Levin L., Kirov C., Littell P., Lo C.-k., Prud'hommeaux E., Park H. H., Steimel K., Knowles R. *et al.*, "Neural polysynthetic language modelling", *arXiv preprint arXiv:2005.05477*, 2020.

Sennrich R., Haddow B., Birch A., "Neural Machine Translation of Rare Words with Subword Units", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, p. 1715-1725, August, 2016.

Sirts K., Goldwater S., "Minimally-supervised morphological segmentation using adaptor grammars", *Transactions of the Association for Computational Linguistics*, vol. 1, p. 255-266, 2013.

Sutskever I., Vinyals O., Le Q. V., "Sequence to sequence learning with neural networks", *Advances in neural information processing systems*, p. 3104-3112, 2014.

Tiedemann J., "Parallel data, tools and interfaces in OPUS.", *Lrec*, vol. 2012, Citeseer, p. 2214-2218, 2012.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I., "Attention is all you need", *Advances in neural information processing systems*, p. 5998-6008, 2017.

Virpioja S., Smit P., Grönroos S.-A., Kurimo M. *et al.*, "Morfessor 2.0: Python implementation and extensions for Morfessor Baseline", 2013.

Wang L., Cao Z., Xia Y., De Melo G., "Morphological segmentation with window LSTM neural networks", *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Yang J., Zhang Y., Dong F., "Neural Word Segmentation with Rich Pretraining", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, p. 839-849, July, 2017.