Deep Learning-Based Morphological Segmentation for Indigenous Languages: A Study Case on Innu-Aimun

Ngoc Tan Le
Université du Québec à Montréal
le.ngoc_tan@uqam.ca

Antoine Cadotte
Université du Québec à Montréal
cadotte.antoine@courrier.uqam.ca

Mathieu Boivin
Université de Montréal
mathieu.boivin.2@umontreal.ca

Fatiha Sadat
Université du Québec à Montréal
sadat.fatiha@uqam.ca

Abstract

Recent advances in the field of deep learning have led to a growing interest in the development of NLP approaches for low-resource and endangered languages. Nevertheless, relatively little research, related to NLP, has been conducted on indigenous languages. These languages are considered to be filled with complexities and challenges that make their study incredibly difficult in the NLP and AI fields. This paper focuses on the morphological segmentation of indigenous languages, an extremely challenging task because of polysynthesis, dialectal variations with rich morpho-phonemics, misspellings and resource-limited scenario issues. The proposed approach, towards a morphological segmentation of Innu-Aimun, an extremely low-resource indigenous language of Canada, is based on deep learning. Experiments and evaluations have shown promising results, compared to state-of-the-art rule-based and unsupervised approaches.

1 Introduction

Over the past decade, we have observed a successful growth in the deep learning-based approaches in several Natural Language Processing (NLP) applications. This has helped to create NLP tools and applications in resource-rich languages. On the other hand, for low-resource languages, few applications of NLP have been studied for multiple reasons (Mager et al., 2018b).

In particular, for indigenous languages, NLP applications have to deal with linguistics challenges such as polysynthesis, diversity of grammatical features of morphology, dialect variation with rich morpho-phonemics, misspellings due to noisy or scarce training data and low resource scenario challenges (Littell et al., 2018; Joanis et al., 2020). Moreover, morphological segmentation for indigenous polysynthetic languages is especially challenging because these languages have often multiple individual morphemes by word and several meanings per morpheme.

The current research focuses on the morphological segmentation task for indigenous languages, with a case study on Innu-Aimun, also called Montagnais[1]. Innu-Aimun is an Algonquian polysynthetic language spoken by over 10,000 Innu in Labrador and Quebec in Eastern Canada[2]. We choose this indigenous language for this specific NLP task because it has not yet been investigated thus far.

The main focus consists of how to develop indigenous language technology and linguistic resources, with the aim of helping the indigenous communities in the revitalization and preservation of their languages. Thus, we propose in the current study, a deep learning-based morphological segmentation for Innu-Aimun. Our contribution to the current research is twofold. Firstly, it proposes a deep learning-based word segmenter for indigenous languages. Secondly, it empirically compares the proposed approach, in a case study of Innu-Aimun, with multiple baselines such as Finite-State Transducer, Morfessor, and Adaptor Grammar-based approaches.

Overall, this study aims to serve as a benchmark for developing NLP tools and applications, which will help revitalize and preserve indigenous languages, while taking into account indigenous cultural realities and knowledge.

The paper is structured as follows: Section 2 highlights morphological analyzers for indigenous languages, with a description of Innu-Aimun. Our proposed approach is described in Section 3. Section 4 presents the experimental results, compared to other state-of-the-art approaches. Section 5 discusses our evaluations, while providing an error analysis. Finally, Section 6 presents the conclusion as well as potential future work.

[1] https://www.thecanadianencyclopedia.ca/en/article/innu-montagnais-naskapi

[2] https://en.wikipedia.org/wiki/Innu-Aimun

Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pages 146 - 151
July 14, 2022 ©2022 Association for Computational Linguistics

2 Related work

2.1 Morphological segmentation in Indigenous languages

Many indigenous languages in Canada, in the Americas and around the world have in common that they are polysynthetic. Most also share a context of extremely low or scarce resource. While morphological segmentation is highly useful—if not unavoidable—for indigenous NLP applications, data and knowledge scarcity make its development very challenging.

When there exists no language-specific tool, NLP tasks often make use of unsupervised approaches for segmentation. Byte-pair encoding (BPE) segmentation, introduced by Sennrich et al. (2016), is a common one for Neural Machine Translation. The technique has been used by (Joanis et al., 2020; Le and Sadat, 2020), for instance, to produce an Inuktitut-English NMT baseline using the Nunavut Hansard corpus.

In cases where there is a lack of annotated data, rule-based approaches, such as those based on Finite-State Transducers (FST), have been used the most. Farley (2012) proposed an FST-based morphological analyser for Inuktitut (one of Canada's most resourced and documented indigenous languages). Harrigan et al. (2017) developed an FST morphological model for Plains Cree. Arppe et al. (2017) applied the same approach partially adapted to East Cree. Mager et al. (2018a) proposed a probabilistic approach to an FST model for Wixarika (huichol).

Other proposed approaches are hybrid, adding knowledge or rules to unsupervised methods. Eskander et al. (2019) proposed an approach based on Adaptor Grammars (Johnson et al., 2006), and applied it to four Uto-Aztecan polysynthetic languages. Pan et al. (2020) combined BPE segmentation and rule-based segmentation for Uyghur, a morphologically rich language.

For deep learning-based approach, Kann et al. (2018) used the neural network-based seq2seq models for Mexican polysynthetic languages. Micher (2019) applied a recurrent neural network-based approach to deal with the word segmentation for Inuktitut.

2.2 Innu-Aimun language

Innu-Aimun is the language of the Innu, an indigenous people formerly known as the Montagnais (Mollen, 2006). This language is found in the Quebec and Labrador provinces of Canada, in a dozen communities (Baraby et al., 2017). It is a polysynthetic indigenous language, a member of the Algonquian family and is related to Cree and Naskapi with which it forms a dialectic continuum (Drapeau, 2014). Statistics Canada estimated the number of speakers at 11,360 in 2016 [3].

Although Innu-Aimun is fundamentally an oral language, its orthography was standardized in 1989 (Mollen, 2006). A first dictionnary based on the standard orthography, for Innu-French, was published in 1991 (Drapeau, 1991). There exists today a more complete, trilingual and pan-dialectal dictionnary that is being continuously updated and is available online[4]. Other online resources include a verb conjugation web application (Baraby and Junker, 2011), based on work of Baraby (1998).

The aforementioned online tools have been part of an effort by Junker et al. (2016) to develop a series of Web tools for Innu-Aimun language maintenance. This project, primarily, aimed at bilingual speakers, which also includes several primary language resources (*e.g.* lexicons, grammars, conversational guides, etc.), educational online games and a catalog of audio and written Innu-Aimun works[5].

Other than online tools, very few language technologies have been developed for Innu-Aimun, to our knowledge. A search-engine with flexible orthography has been developed by Junker and Stewart (2008) and integrated with an online dictionnary (Junker et al., 2016), in conjunction with an equivalent tool for East Cree. Other research projects have targeted the construction of Innu-Aimun corpora Cadotte et al. (2022). Drapeau and Lambert-Brétière (2013) proposed an annotated, multimodal corpus with translations. An NRC Canada indigenous languages technology project (Kuhn et al., 2020) aimed to transcribe oral recordings of several indigenous languages in Canada, including Innu-Aimun.

3 Our proposed approach

3.1 Model overview

In this paper, we focus on the surface segmentation (Ruokolainen et al., 2016; Kann et al., 2018; Liu et al., 2021), where a term is segmented in a substrings sequence.

[3] Statistics Canada: The Aboriginal languages of First Nations people, Métis and Inuit

[4] `https://dictionary.Innu-Aimun.ca/`

[5] Tshakapesh Institute - Catalogue

Given an Innu-Aimun word, the segmentation process consists of breaking down the word into separate morphemes, for example, *uminushima* → *u-minush-im-a* (in English: *her/his cats*). Our model is made following these steps: (1) apply the Transformer-based encoder-decoder architecture, with a multihead self attention mechanism (Vaswani et al., 2017); (2) deal with surface segmentation, while considering the monotonic aspect of morphotactics (that is, the constraints on the ordering of morphemes) (Figure 1). We train the positional embeddings using the position of each element in a sequence.

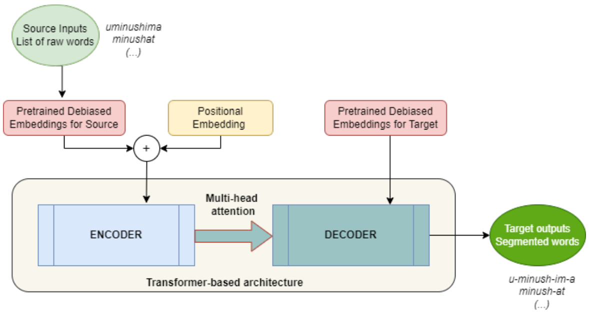
Source Inputs
List of raw words
uminushima
minushat
(...)
Pretrained Debiased Embeddings for Source
Positional Embedding
Pretrained Debiased Embeddings for Target
+
Multi-head attention
ENCODER
DECODER
Transformer-based architecture
Target outputs Segmented words
u-minush-im-a
minush-at
(...)

Figure 1: Architecture of our framework: Deep Learning-based Morphological segmentation for indigenous language, with pretrained debiased word-based embedding for source-target, and positional embedding.

3.2 Deep Learning-based morphotactics modeling

We model a deep learning-based morphological segmenter using the Transformer-based encoder-decoder architecture.

In the encoder, the input sequence is encoded at character level. Then the embedding layer is incorporated with pre-trained embeddings at multiple levels such as character, affix (prefixes and suffixes), along with multiheaded attention over the input sequence, that helps finding morpheme boundaries related to the whole word.

To ensure the monotonic aspect of morphotactics, the positional embeddings are used to encode the order of each element of a sequence in both the encoder and the decoder.

The decoder uses the same concept of multihead attention over itself and also the encoder. The attention mechanism allows to align input sequences to the correct corresponding output sequences that are segmented in individual morphemes (Figure 1).

4 Experiments and Evaluations

4.1 Data Preparation

A small corpus was manually collected from multiple resources such as the Website of Aimun-Mashinaikan-French-English dictionary Innu[6] as well as open source grammar books and the online Innu lessons platform[7] that are available at the Tshakapesh Institute (Drapeau, 2014; Mollen, 2006).

The collected experimental corpus contains 500 word bases (roots) and 500 affixes (prefixes, suffixes). A training set, crawled from the Aimun-Mashinaikan dictionary Innu, consists of 30,118 terms, used as raw word lists, non segmented, with length between 2 and 46 characters. A small golden testing set, containing 250 unique terms, was manually segmented with the help of an Innu language teacher from the Uashat Mak Mani-utenam community[8].

4.2 Training settings

We configured several baselines: (1) based on a simple weighted Finite-State Transducer (FST) to maximise the morpheme frequency (Richardson and Tyers, 2021), (2) based on Morfessor version 2.0 (Virpioja et al., 2013) to learn the morpheme boundaries using minimum description length optimization, and (3) based on the Adaptor Grammar approach. We used the MorphAGram toolkit (Eskander et al., 2020), with two settings: standard setting (AdaGra-Std) and scholar seeded setting (AdaGra-SS). We adopted the best learning settings: the best standard *PrefixStemSuffix+SuffixMorph* grammar and the best scholar-seeded grammar, as explained in (Eskander et al., 2019), for Innu-Aimun.

We configured a deep-learning based model (T-DeepLo) with an encoder-decoder Transformer model (Vaswani et al., 2017), based entirely on the multihead self-attention mechanism. For the hyperparameters, we used 4-layer both in the encoder and in the decoder. The batch size was set at 32. The initial learning rate was set to 0.0001. The hidden dimension was set at 256, and dropout with a rate of 0.2. The model is trained with 8 multihead attention in the encoder and in the decoder, using Adam optimizer (Kingma and Ba, 2014).

[6] https://dictionary.Innu-Aimun.ca/Words

[7] https://lessons.innu.atlas-ling.ca/

[8] https://www.itum.qc.ca/

4.3 Results

	Precision	Recall	F1
FST	52.71	42.96	46.11
Morfessor	43.33	38.01	40.49
AdaGra-Std	53.78	43.18	47.91
AdaGra-SS	70.45	61.36	65.60
T-DeepLo	81.27	77.15	79.16

Table 1: Evaluation on the test set using the different settings.

The performances of all the models were evaluated using the conventional automatic metrics in the field of NLP, such as Precision, Recall and F1-score.

For the unsupervised methods, we noticed that the scholar-seeded learning (AdaGra-SS) model outperformed all the other baselines, with 70.45%, 61.36%, 65.60% in terms of Precision, Recall and F1 score, respectively (Table 1). We observed both precision and recall were significantly improved while injecting a list of affixes (prefixes and suffixes) during the training. However, the Morfessor model showed the worst results, with only 40.49% in terms of F1.

The Transformer-based DeepLo model obtained the best performance across all metrics, with gains of +10.82%, +15.79%, +13.56% in terms of Precision, Recall and F1 score, respectively, compared to the AdaGra-SS model (Table 1). The T-DeepLo model showed the ability to learn and to extract more complex features, relying on the multihead self attention mechanism.

We performed an error analysis in order to shed some light on how the models were able to learn and recognize the morpheme boundary of a sequence. Table 2 shows sample prediction outputs from all the models on the test set.

5 Error analysis

Due to the complex linguistic peculiarities of Innu-Aimun and its dialectal variations, a word can be pronounced in several ways. Thus, its transcription poses more challenges in the segmentation task. Besides, a word in Innu-Aimun is always composed of a central core (root), including a verb.

With the help of an Innu language teacher, we made observations and reviewed the data and predictions to determine if the segmentation results were correct and discover the errors. Basically, our models tend to over-segment more complex morphemes due to the linguistic irregularities and the morphotactic phenomena, to detect common lexical suffixes such as *ap*, *tsh* or grammatical ending suffixes such as *at*, *eu*, *t*, *n*, *it*, *mi* or *uk*. In particular, we observed an over-segmentation in the FST and Morfessor models. These models tend to segment a term into several sub-morphemes (Table 2). The same phenomena are found in other models of AdaGra-Std and AdaGra-SS. Furthermore, the T-DeepLo model was able to better detect morpheme boundaries.

All models failed when dealing with out-of-vocabulary words. For example, here, the term *mitshuap* (meaning: *house*), which was not seen in the training, was segmented into multiple morphemes (Table 2).

Another challenge is related to the over-segmentation of all the models, down to character level, due to the length of prefixes and suffixes between one and multiple characters. For example, some models divided a term up to a character level (Table 2): (FST) **u a** pa tamu; mi **t** shu **a p**; (Morfessor) **u** apa tamu; (AdaGra-Std) minu sha **t**; (AdaGra-SS) **u** apa tamu.

6 Conclusion and Perspectives

We presented a deep learning-based method for morphological segmentation for Innu-Aimun, an indigenous language of Canada, which can be considered as a first research study on the subject, so far.

Our evaluations showed promising results. Thus, the proposed deep learning-based method, incorporating pre-trained embeddings at multiple levels, helped finding morpheme boundaries related to the whole word. This study makes an important contribution by focusing on morpheme segmentation in the low-resource indigenous language. Furthermore, through this research, we noted the importance of close collaboration and consultation with the Innu indigenous community, to ensure that language technologies are developed with respect and in accordance with the community's revitalisation objectives.

Acknowledgements

The authors are grateful to the community of Uashat Mak Mani-Utenam, Samuel Marticotte for his role in initiating this project, Mrs. Denise Jourdain for sharing her knowledge and experience,

Reference	FST	Morfessor	AdaGra-Std	AdaGra-SS	T-DeepLo
akushi ńua	**akushińua**	akushi **ń ua**	akushi ńua	akushi ńua	akushi ńua
minush at	minush at	**minu sh** at	**minu sha t**	minush at	minush at
mi tshuap	mi **t shu a p**	mi **tsh uap**	mi **tshu ap**	mi **tsh uap**	**mitshuap**
pashu e u	**pa shu eu**	**pashueu**	**pa sh** ueu	pash **u eu**	pashu **eu**
tshika papata n	**tshi ka pa pa ta n**	**tshi ka papa tan**	tshika **papatan**	tshika **papatan**	tshika **pa patan**
u minush im a	u minush **ima**	u **minu sh ima**	u minush **ima**	u minush im a	u minush im a
uapat am u	**u a pa tamu**	**u apa tamu**	**uapa tamu**	**u apa tamu**	**u apa tamu**

Table 2: Illustrations of morpheme segmentation predictions on the test set using the different settings such as Finite-State Transducer, Morfessor, Standard setting (AdaGra-Std), Scholar seeded setting (AdaGra-SS), and Deep learning-based (T-DeepLo). Strings in bold are incorrectly segmented.

Prof. Yvette Mollen and Prof. Jimena Terraza from Kiuna College, for their precious advices and feedbacks. The authors also thank the anonymous reviewers for their valuable comments.

References

Antti Arppe, Marie-Odile Junker, and Delasie Torkornoo. 2017. Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–56, Honolulu. Association for Computational Linguistics.

Anne-Marie Baraby. 1998. Guide pratique des principales conjugaisons en Montagnais. *Sept-Iles: Institut culturel et éducatif montagnais*.

Anne-Marie Baraby and Marie-Odile Junker. 2011. Conjugaisons des verbes innus.

Anne-Marie Baraby, Marie-Odile Junker, and Yvette Mollen. 2017. A 45-year old language documentation program first aimed at speakers: the case of the Innu.

Antoine Cadotte, Tan Ngoc Le, Boivin Boivin, and Fatiha Sadat. 2022. Challenges and perspectives for innu-aimun within indigenous language technologies. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 99–108, Dublin, Ireland. Association for Computational Linguistics.

Lynn Drapeau. 1991. *Dictionnaire montagnais-français*. Presses de l'Université du Québec.

Lynn Drapeau. 2014. *Grammaire de la langue innue*. Presses de l'Université du Québec.

Lynn Drapeau and Renée Lambert-Brétière. 2013. The innu language documentation project. In *Proceedings of the 17th Foundation for Endangered Languages Conference*.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith L Klavans, and Smaranda Muresan. 2020. Morphagram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7112–7122.

Ramy Eskander, Judith L Klavans, and Smaranda Muresan. 2019. Unsupervised morphological segmentation for low-resource polysynthetic languages. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195.

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

Atticus G. Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4):565–598.

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. European Language Resources Association.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Marie-Odile Junker, Yvette Mollen, Hélène St-Onge, and Delasie Torkornoo. 2016. Integrated web tools for Innu language maintenance. In *Papers of the 44th Algonquian Conference*, pages 192–210.

Marie-Odile Junker and Terry Stewart. 2008. Building search engines for Algonquian languages. *Algonquian Papers-Archive*, 39.

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*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine, Caroline Running Wolf, Eddie Santos, Darlene Stewart, Gilles Boulianne, Vishwa Gupta, Brian Maracle Owennatékha, Akwiratékha' Martin, Christopher Cox, Marie-Odile Junker, Olivia Sammons, Delasie Torkornoo, Nathan Thanyehténhas Brinklow, Sara Child, Benoît Farley, David Huggins-Daines, Daisy Rosenblum, and Heather Souter. 2020. The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5866–5878, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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 (Volume 1: Research Track)*, pages 165–172, Virtual. Association for Machine Translation in the Americas (AMTA 2020).

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.

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological segmentation for seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018a. Probabilistic Finite-State morphological segmenter for Wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087. Publisher: IOS Press.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018b. 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.

Jeffrey Micher. 2019. Bootstrapping a neural morphological generator from morphological analyzer output for inuktitut. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 2, page 7.

Yvette Mollen. 2006. Transmettre un héritage: la langue innue. *Cap-aux-Diamants: la revue d'histoire du Québec*, (85):21–25. Publisher: Les Éditions Cap-aux-Diamants inc.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Morphological word segmentation on agglutinative languages for neural machine translation.

Ivy Richardson and Francis M Tyers. 2021. A morphological analyser for k'iche'. *Procesamiento del Lenguaje Natural*, 66:99–109.

Teemu Ruokolainen, Oskar Kohonen, Kairit Sirts, Stig-Arne Grönroos, Mikko Kurimo, and Sami Virpioja. 2016. A comparative study of minimally supervised morphological segmentation. *Computational Linguistics*, 42(1):91–120.

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. Association for Computational Linguistics.

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

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.