Towards a First Automatic Unsupervised Morphological Segmentation for Inuinnaqtun

Tan Le Ngoc and **Fatiha Sadat**
Université du Québec à Montréal / Montreal, Quebec, Canada
201, avenue du Président-Kennedy, H2X 3Y7 Montréal
le.ngoc_tan@courrier.uqam.ca, sadat.fatiha@uqam.ca

Abstract

Low-resource polysynthetic languages pose many challenges in NLP tasks, such as morphological analysis and Machine Translation, due to available resources and tools, and the morphologically complex languages. This research focuses on the morphological segmentation while adapting an unsupervised approach based on Adaptor Grammars in low-resource setting. Experiments and evaluations on Inuinnaqtun, one of Inuit language family in Northern Canada, considered a language that will be extinct in less than two generations, have shown promising results.

1 Introduction

NLP has significant achievements when dealing with different types of languages, such as isolating, inflectional or agglutinative language families. However, Indigenous polysynthetic languages still pose several challenges within NLP tasks and applications, such as morphological analysis or machine translation, due to their complex linguistic particularities and due to the scarcity of linguistic resources and reliable tools (Littell et al., 2018; Mager et al., 2018; Micher, 2019; Le Ngoc and Sadat, 2020).

Herein, we propose an unsupervised morphological segmentation approach, which is primarily based on the grammar containing production rules, non-terminal and terminal symbols, and a lexicon using Adaptor Grammars (Johnson, 2008). Our current research investigates Inuinnaqtun - a polysynthetic language spoken in Northern Canada, in the Inuit language family. Inuinnaqtun is considered as a language that will be extinct in less than two generations[1].

Regarding the Eskimo-Aleut language family including the Inuit, unlike words in English, the word structure of Eskimo are very variable in their form (Lowe, 1985; Kudlak and Compton, 2018). Words may be very short, built up of three formative elements such as word base, lexical suffixes, and grammatical ending suffixes, or very long, with up to ten or even fifteen formative morphemes depending on the dialect.

- Eskimo word structure = **Word base** + Lexical suffixes + *Grammatical ending suffixes*

A single word can be used to express a whole sentence in English. The following example, extracted from (Lowe, 1985), illustrates the polysynthesis effect of *umingmakhiuriaqtuqatigitqilimaiqtara*, an Inuinnaqtun sentence-word, split up into several morphemes:

umingmak-hiu-riaqtu-qati-gi-tqi-limaiq-*ta-ra*

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

(*Meaning: I* will no more again have *him* as a partner to go hunting **muskox**.)

We observe there is a general tendency to increase the lexical constituents with a word-base by adding more formative elements. A single word can express the meaning of a whole sentence. Moreover, morphology is highly developed and has extensive use of lexical and grammatical ending suffixes. All these linguistic aspects make the morphological segmentation task for polysynthetic languages more challenging. On the other hand, the benefit of this work helps to identify more unknown word bases by deducting from the known affixes, which in turn helps to enrich the Inuinnaqtun lexicon. The global contribution consists of helping to revitalize and preserve low-resource Indigenous languages and the transmission of the related ancestral knowledge and culture.

The structure of this paper is described as follows: Section 2 presents relevant works. Section 3 describes our proposed approach. Then, Section 4 presents experiments and evaluations. Finally,

[1] https://www.kitikmeotheritage.ca/language

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 159–162
June 11, 2021. ©2021 Association for Computational Linguistics

Section 5 gives some conclusions and perspectives for future research.

2 Related work

Creutz and Lagus (2007) proposed the Morfessor, for the unsupervised discovery of morphemes. This work was based on Hidden Markov Model for learning the unsupervised morphological segmentation, and by using the hierarchical structure of the morphemes. This framework became a benchmark in unsupervised morphological analysis, such as Morfessor 2.0 (Virpioja et al., 2013).

Johnson (2008) proposed Adaptor Grammars approach that was successful for the unsupervised morphological segmentation. This approach used non-parametric Bayesian models generalizing probabilistic context-free grammar (PCFG). In this approach, a PCFG is considered as a morphological grammar of word structures. Then the AG models can be able to induce the segmentation at the morpheme level.

This approach has been extended in several studies (Botha and Blunsom, 2013; Sirts and Goldwater, 2013; Eskander et al., 2018) for learning non-concatenative morphology, or for unsupervised morphological segmentation of unseen languages. Recently, Godard et al. (2018) applied AG approach for the linguists with word segmentation experiments for very low-resource African languages. Eskander et al. (2019) has applied the AG approach in an unsupervised morphological segmentation of the low-resource polysynthetic languages such as Mexicanero, Nahuatl, Yorem Nokki and Wixarika. Their evaluations have shown a significant improvement up to 87.90% in terms of F1-score, compared to the supervised approaches (Kann et al., 2018). Our work examines the efficiency of the AG-based approach on Inuinnaqtun, a polysynthetic low-resource Inuit language.

3 Our approach

Inspired by the work of Eskander et al. (2019), we adapt an unsupervised morphological segmentation with the Adaptor Grammars (AG) approach for the Inuit language family, by completing an empirical study on Inuinnaqtun.

The main process consists of defining (1) the grammar including non-terminal, terminal symbols, a set of production rules, and (2) collecting a large amount of unsegmented word list in order to discover and to learn all possible morphological patterns.

In our work, we consider that word structures are specified in the grammar patterns where a word is constituted as one word base, a sequence of possible lexical suffixes and grammatical ending suffixes (see Table 1). In contrast, as explained in (Eskander et al., 2019), the word structure is composed of a sequence of prefixes, a stem and a sequence of suffixes. Then, in each production rule, a and b are two parameters of Pitman-Yor process (Pitman and Yor, 1997). Setting $a = 1$ and $b = 1$ indicate, to the running learner, 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.

In order to adapt the AG scholar-seeded setting with linguistic knowledge, we have collected a list of affixes from dictionaries and Websites in the appropriate language.

4 Experiments

4.1 Data Preparation

In order to train the Adaptor Grammars-based unsupervised morphological segmentation model, the two principal inputs consists of the grammar and the lexicon of the language. The lexicon consists of a unique list of unsegmented words, more than $50K$ words, with the sequence length between three letters and 30 letters.

We collected manually a small corpus from several resources such as the Website of Nunavut[2] government for Inuinnaqtun, 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 is manually crafted containing 1,055 unique segmented words.

4.2 Training Settings

We used the MorphAGram toolkit (Eskander et al., 2020) to train our unsupervised morphological segmentation model. Following (Eskander et al., 2019), we set up the same configuration with adaptation of the best learning settings: the best standard *PrefixStemSuffix+SuffixMorph* grammar and the best scholar-seeded grammar, that become here an adaptation of the standard grammar *WordBase+LexicalSuffix+GrammaticalSuffix* pattern for

[2] https://www.gov.nu.ca/in/cgs-in

1 1 Word –>WordBase LexicalSuffix GrammaticalSuffix WordBase –> ^^^ WordBase –> ^^^ WordBaseMorphs 1 1 WordBaseMorphs –> WordBaseMorph WordBaseMorph –> SubMorphs LexicalSuffix –> SubMorphs LexicalSuffix –> SuffixMorphs $$$ LexicalSuffix –> $$$	GrammaticalSuffix –> SuffixMorphs $$$ 1 1 SuffixMorphs –> SuffixMorph SuffixMorphs 1 1 SuffixMorphs –> SuffixMorph 1 1 SubMorphs –> SubMorph SubMorphs 1 1 SubMorphs –> SubMorph SubMorph –> Chars 1 1 Chars –> Char 1 1 Chars –> Char Chars

Table 1: Adaptation of the standard grammar WordBase+LexicalSuffix+GrammaticalSuffix pattern for Inuinnaqtun. The symbols ^^^ and $$$ mean the beginning and the end of the word sequence, respectively. Source: see the standard PrefixStemSuffix+SuffixMorph grammar pattern (Eskander et al., 2019).

Word	Ground Truth	Morfessor	AG-Standard	AG-Scholar
aullarnatin	aullar na tin	aulla rn at in	a ulla rna tin	aullar nati n
havangnatik	havang na tik	hav ang na tik	hav a ngna tik	havang na tik
iaqluktinnagu	iqaluk tinna gu	iqalu k ti nna gu	iqa luk tinna gu	iaqluk tinna gu
nirihuiqtunga	niri huiq tunga	niri huiq tu ng a	niri huiq tu ng a	niri huiq tunga
niritinnagit	niri tinna git	niri ti nna gi t	niri tinna git	niri tinna git
umiarmi	umiar mi	umi a rmi	umi armi	umia r mi
umiaq	umiaq	umi aq	u mi aq	umiaq
tikinnanuk	tikin na nuk	tikinnanuk	t iki nna nuk	tikin na nuk

Table 2: Illustrations of morpheme segmentation predictions on the test set using the different settings such as Standard (AG-Standard), Scholar seeded (AG-Scholar) and Morfessor.

Inuinnaqtun (see Table 1). We evaluate our different models against the baseline, based on Morfessor (Virpioja et al., 2013).

4.3 Evaluations

All the model performances are calculated using common evaluation metrics, such as Precision (P), Recall (R) and F1 score.

$$P = \frac{|\{relevant\ tokens\} \cap \{found\ tokens\}|}{\{found\ tokens\}} \quad (1)$$

$$R = \frac{|\{relevant\ tokens\} \cap \{found\ tokens\}|}{\{relevant\ tokens\}} \quad (2)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (3)$$

where $\{found\ tokens\}$ means the amount of predicted *tokens*; and $\{relevant\ tokens\}$ indicates the amount of *tokens* which are correctly segmented.

Tables 2 and 3 show some illustrations of prediction by all the models and the performance of our models versus Morfessor as baseline on the test set. The AG-standard model is better than the baseline, with a gain of +2.47%, +4.9% in terms of precision and recall, on the test set, respectively. Both baseline and AG-Standard models obtained low precision between 48.29% and 50.76%. We observed an over-segmentation in both models.Furthermore, we noticed that the scholar-seeded learning outperformed all the baseline and the standard setting, with performances of 71.06%, 82.83%, 76.49% in terms of Precision, Recall and F1 score, respectively. Our models tend to over-segment more complex morphemes due to the linguistic irregularities and the morphophonological phenomena, to detect 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*.

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

Table 3: The results on the test set using the different settings such as Standard (AG-Standard), Scholar seeded (AG-Scholar) and Morfessor.

5 Conclusion

In this research paper, we presented how to build the unsupervised morphological segmentation with Adaptor Grammars approach for Inuinnaqtun, an Inuit language, considered as an extremely low-resource polysynthetic language, that will be extinct in less than two generations, as described and referenced above. This Adaptor Grammars-based approach showed promising results, when using a set of grammar rules, that can be collected from grammar books; and a lexicon extracted from very little data. As a perspective, we intend to develop more efficient unsupervised morphological segmentation methods and to extend our research to other Indigenous languages and dialects, especially the very endangered ones; with applications on Machine Translation and Information Retrieval.

Acknowledgments

The authors wish to express their thanks for professor Richard Compton, Canada Research Chair in Knowledge and Transmission of the Inuit Language, and the reviewers, for their constructive feedbacks.

References

Jan A Botha and Phil Blunsom. 2013. Adaptor grammars for learning non- concatenative morphology. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

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.

Ramy Eskander, Owen Rambow, and Smaranda Muresan. 2018. Automatically tailoring unsupervised morphological segmentation to the language. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 78–83.

Pierre Godard, Laurent Besacier, François Yvon, Martine Adda-Decker, Gilles Adda, Hélène Maynard, and Annie Rialland. 2018. Adaptor grammars for the linguist: Word segmentation experiments for very low-resource languages. In *Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–42.

Mark Johnson. 2008. Unsupervised word segmentation for sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27.

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*.

Emily Kudlak and Richard Compton. 2018. *Kangiryuarmiut Inuinnaqtun Uqauhiitaa Numiktitirutait — Kangiryuarmiut Inuinnaqtun Dictionary*, volume 1. Nunavut Arctic College: Iqaluit, Nunavut.

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

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.

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

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. 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.

Jim Pitman and Marc Yor. 1997. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266.

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