Improving Coverage of an Inuktitut Morphological Analyzer Using a Segmental Recurrent Neural Network

Jeffrey C. Micher
US Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783
jeffrey.c.micher.civ@mail.mil

Abstract

Languages such as Inuktitut are particularly challenging for natural language processing because of polysynthesis, abundance of grammatical features represented via morphology, morphophonemics, dialect variation, and noisy data. We make use of an existing morphological analyzer, the Uqailaut analyzer, and a dataset, the Nunavut Hansards, and experiment with improving the analyzer via bootstrapping of a segmental recurrent neural network onto it. We present results of the accuracy of this approach which works better for a coarse-grained analysis than a fine-grained analysis. We also report on accuracy of just the "closed-class" suffix parts of the Inuktitut words, which are better than the overall accuracy on the full words.

1 Introduction

Inuktitut is a polysynthetic language, and is part of the Inuit language dialect continuum spoken in Arctic North America[1]. It is one of the official languages of the territory of Nunavut, Canada and is used widely in government and educational documents there. Despite its elevated status of official language, very little research on natural language processing of Inuktitut has been carried out. On the one hand, the complexity of Inuktitut makes it challenging for typical natural language processing applications. On the other hand, Inuktitut may only be spoken by around 23,000 people as a first language,[2] and as a result, receives minimal commercial attention. The National Research Council of Canada has produced a very important morphological analyzer for Inuktitut. We make use of this analyzer and propose a neural network approach to enhancing its output. This paper is organized as follows: first we talk about the nature of Inuktitut, focusing on polysynthesis, abundance of grammatical suffixes, morphophonemics, and spelling. Second, we describe the Uqailaut morphological analyzer for Inuktitut. Third, we describe the Nunavut Hansard dataset used in the current experiments. Forth, we describe an approach to enhancing the morphological analysis based on a segmental recurrent neural network architecture, with character sequences as input. Finally, we present ongoing experimental research results.

2 Nature of Inuktitut

2.1 Polysynthesis

The Eskimo-Aleut language family, to which Inuktitut belongs, is the canonical language family for exemplifying what is meant by polysynthesis. Peter Stephen DuPonceau coined the term in 1819 to describe the structural characteristics of languages in the Americas, and it further became part of Edward Sapir's classic linguistic typology distinctions.[3] Polysynthetic

[1] https://en.wikipedia.org/wiki/Inuit_languages

[2] http://nunavuttourism.com/about-nunavut/welcome-to-nunavut

[3] https://en.wikipedia.org/wiki/Polysynthetic_language

languages show a high degree of synthesis, more so than other synthetic languages, in that single words in a polysynthetic language can express what is usually expressed in full clauses in other languages. Not only are these languages highly inflected, but they show a high degree of incorporation as well. In Figure 1 we see an example of a sentence in Inuktitut, *Qanniqlaunngikkalauqtuqlu aninngittunga.,* consisting of two words, and we break those words down into their component morphemes.

Qanniqlaunngikkalauqtuqlu
qanniq-lak-uq-nngit-galauq-tuq-lu
snow-a_little-frequently-NOT-although-3.IND.S-and
"*And even though it's not snowing a great deal,"*

aninngittunga
ani-nngit-junga
go_out-NOT-1.IND.S
"I'm not going out"

Figure 1: Inuktitut Words Analyzed

In this example, two Inuktitut words express what is expressed by two complete clauses in English. The first Inuktitut word shows how many morphemes representing a variety of grammatical and semantic functions (quantity, "a_little", frequency "frequently", negation, and concession) as well as grammatical inflection (3rd person indicative singular), can be added onto a root ("snow"), in addition to a clitic ("and"), and the second word shows the same, but to a lesser degree.

2.2 Abundance of grammatical suffixes

Morphology in Inuktitut is used to express a variety of abundant grammatical features. Among those features expressed are: eight verbs "moods" (declarative, gerundive, interrogative, imperative, causative, conditional, frequentative, dubitative) ; two distinct sets of subject and subject-object markers, *per mood*; four persons; three numbers, (singular, dual, plural); eight "cases" on nouns: nominative (absolutive), accusative, genitive (ergative), dative ("to"), ablative ("from"), locative ("at, on, in"), similaris ("as"), and vialis ("through"), and noun possessors (with number and person variations). In addition, demonstratives show a greater variety of dimensions than most languages, including location, directionality, specificity, and previous mention. The result of a language that uses morphology (usually suffixes) to express such a great number of variation is an abundance of word types and very sparse data sets.

2.3 Morphophonemics

In addition to the abundance of morphological suffixes that Inuktitut roots can take on, the morphophonemics of Inuktitut are quite complex. Each morpheme in Inuktitut dictates the possible sound changes that can occur to its left, and/or to itself. As a result, morphological analysis cannot proceed as mere segmentation, but rather, each surface segmentation must map back to an underlying morpheme. In this paper, we refer to these different morpheme forms as 'surface' morphemes and 'deep' morphemes. The example below demonstrates some of the typical morphophonemic alternations that can occur in an Inuktitut word, using the word *mivviliarumalauqturuuq*, 'he said he wanted to go to the landing strip':

Romanized Inuktitut word:	mivviliarumalauqturuuq
Surface segmentation	miv-vi-lia-ruma-lauq-tu-ruuq
Deep form segmentation	mit-vik-liaq-juma-lauq-juq-guuq
Gloss	land-place-go_to-want-PAST-3.IND.S-he_says

Figure 2: Morphophonemic Example

We proceed from the end to the beginning to explain the morphophonemic rules. 'guuq' is a regressive assimilator, acting on the point of articulation, and it also deletes. So 'guuq' changes to 'ruuq' and deletes the preceding consonant 'q' of 'juq.' 'juq shows an alternation in its first consonant, which appears as 't' after a consonant, and 'j' otherwise. 'lauq' is neutral after a vowel, so there is no change. 'juma' is like 'guuq', a regressive assimilator on the point of articulation, and it deletes. So 'juma' becomes 'ruma,' and the 'q' of the preceding morpheme is deleted. 'liaq' is a deleter, so the preceding 'vik' becomes 'vi.' Finally, 'vik' regressively assimilates a preceding't' completely, so 'mit' becomes 'miv.'[4]

2.4 Dialect differences / spelling variation

The fourth aspect of Inuktitut which contributes to the challenge of processing it with a computer is the abundance of spelling variation seen in the electronically available texts. Three aspects of spelling variation must be taken into account. First, Inuktitut, like all languages, can be divided into a number of different dialects. Dorais, (1990) lists ten: Uummarmiutun, Siglitun, Inuinnaqtun, Natsilik, Kivallirmiutun,

[4]http://www.inuktitutcomputing.ca/Technocrats/ILFT.php#morphology

Aivilik, North Baffin, South Baffin, Arctic Quebec, and Laborador. The primary distinction between the dialects is phonological, which is reflected in spelling. See Dorais (1990) for a discussion of dialect variation. Second, a notable error on the part of the designers of the Romanized transcription system has produced a confusion between 'r's and 'q's. It is best explained in a quote by Mick Mallon:[5]

> *It's a long story, but I'll shorten it. Back in 1976, at the ICI standardization conference, because of my belief that it was a good idea to mirror the Assimilation of Manner in the orthography, it was decided to use **q** for the first consonant in voiceless clusters, and **r** for the first consonant in voiced and nasal clusters. That was a mistake. That particular distinction does not come natural to Inuit writers, (possibly because of the non-phonemic status of [ɴ].) Public signs, newspaper articles, government publications, children's literature produced by the Department of Education, all are littered with **q**s where there should be **r**s, and **r**s where there should be **q**s. Kativik did the right thing in switching to the use of **r**s medially, with **q**s left for word initial and word final. When things settle down, maybe Nunavut will make that change. It won't affect the keyboard or the fonts, but it will reduce spelling errors among the otherwise literate by about 30%.*

Third, an inspection of the word types that cannot be analyzed by the Uqailaut analyzer reveals that transcribers and translators do not adhere to a single standard of spelling. As an example, the root for 'hamlet', borrowed from English, appears in a variety of spelling variations in the Nunavut Hansard dataset. The Uqailaut root unique ID is "Haammalat/1n", mapped to the surface form "Haammalat". However, in the dataset, surface forms abound: *Haamalaujunut, Haamlaujunut, Hamalakkunnit, Hammakkut, Hammalakkunnut, Hammalat,* and *Hmlatni*.

In sum, the combination of polysynthesis, abundance of grammatical morphemes, morphophonemics, and spelling variation, make Inuktitut a particularly challenging language for natural language processing. In this work, we hope to begin to develop methods to overcome these challenges.

[5]http://www.inuktitutcomputing.ca/Technocrats/ILFT.php#morphology

3 Related work

The work on morphological processing is abundant, so here we simply present a selected subset. Creutz and Lagus (2006) present an unsupervised approach to learning morphological segmentation, relying on a minimum description length principle (Morfessor). Kohonen et al, (2010) use a semi-supervised version of Morfessor on English and Finnish segmentation tasks to make use of unlabeled data. Narasimhan et al. (2015) use an unsupervised method to learn morphological "chains" (which extend base forms available in the lexicon) and report results on English, Arabic, and Turkish. Neural network approaches to morphological processing include a number of neural model types. Among then, Wang, et al. (2016) use Long Short Term Memory (LSTM) neural morphological analyzers to learn segmentations from unsupervised text, by exploiting windows of characters to capture context and report results on Hebrew and Arabic. Malouf (2016) uses LSTM neural networks to learn morphological paradigms for several morphologically complex languages (Finnish, Russian, Irish, Khaling, and Maltese). One recent work which may come close to the challenge of segmentation and labeling for Inuktitut is presented by Morita et al. (2014) who use a Recurrent Neural Network Language Model during morphological analysis of unsegmented, morphologically complex languages such as Japanese.

However, none of the studies to date have attempted to improve analysis for polysynthetic languages such as Inuktitut. Given the complexity of Inuktitut, we proceed initially by bootstrapping from an available analyzer.

4 The Uqailaut morphological analyzer

A morphological analyzer exists for Inuktitut called the Uqailaut analyzer. It was created by Benoît Farley at the National Research Council of Canada and is freely available for use from the Uqailaut website[6]. The analyzer was used as downloaded, with no alterations to the source code whatsoever. It is a finite state transducer, which takes a word as input and returns a set of analyses on multiple lines. Each morpheme returned in an analysis is delineated by curly braces and contains the surface form of the morpheme, followed by a colon, followed by a unique ID for that morpheme. The unique ID

[6] http://www.inuktitutcomputing.ca/Uqailaut/.

contains the deep (dictionary) form of the morpheme, followed by a forward slash, followed by any specific morphological information. For example, processing the word 'saqqitaujuq' yields:

{saqqi:saqqik/1v}{ta:jaq/1vn}{u:u/1nv}{juq:juq/1vn}
{saqqi:saqqik/1v}{ta:jaq/1vn}{u:u/1nv}{juq:juq/tv-ger-3s}

Figure 3: Sample Analyzer Output

See the README file provided with the Uqailaut analyzer for specifics about the codes used in the unique ID for a morpheme.[7]

5 Nunavut Hansard dataset

A parallel Inuktitut-English dataset originated during the ACL 2003 Workshop entitled "Building and Using Parallel Texts: Data-driven Machine Translation and Beyond[8]" and was made available during this workshop. The data was subsequently used for a shared task on alignment that took place in the same workshop in 2005[9]. The dataset was assembled and aligned, and is described in Martin et al., (2003). The dataset is available from the web[10], and was downloaded from there.

6 Morphological processing of the Nunavut Hansard dataset

Since the morphological analyzer is not sensitive to context, all of the unique types from the Inuktitut side of the Nunavut Hansard were collected and prepared for analysis. Punctuation was tokenized beyond the initial tokenization provided in the dataset to facilitate processing. Types that contained an alphabetic string concatenated with a numeral (either an unwanted processing error or a numeral with grammatical inflexion) which would have failed in morphological analysis, were filtered out. A total of 287,858 types were successfully processed by the analyzer, leaving 125,695 types unprocessed, of which 124,189 were truly not processed by the analyzer (the remaining 1506 types yielded various processing errors). The number of types not processed was around 30% of the total types collected from the corpus.

[7] http://www.inuktitutcomputing.ca/Uqailaut/
[8] http://web.eecs.umich.edu/~mihalcea/wpt/
[9] http://www.statmt.org/wpt05/
[10] http://www.inuktitutcomputing.ca/NunavutHansard/ info.php

7 Enhancing the output of the analyzer

The goal of the current research is to provide an analysis for the 30% of word types that the analyzer fails on without completely rebuilding the underlying analyzer. The purpose of this approach is to determine whether morphological analysis for such a complex, polysynthetic language can be learned from annotated data when only limited annotated data or a less-than-optimal, incomplete analyzer is available. We make use of the word types that were successfully analyzed into one single, unambiguous analysis as training data for a neural network morphological analyzer for the remaining unanalyzed word types. Following Kong et. al (2015), we make use of a segmental recurrent neural network (SRNN) that can be used to map character input to analysis output. The input to the network is the individual characters of the word type to be analyzed. Output from the network is a sequence of labels annotated with the number of characters that each label covers. For example, the word type, "qauqujaujunu", is analyzed to "ROOT:3 LEX:2 LEX:2 LEX:1 LEX:2 GRAM:2" which then can be converted to qau/ROOT qu/LEX ja/LEX u/LEX ju/LEX nu/GRAM, to provide a surface segmentation and morphological analysis information in the form of a coarse-grained label. Thus the SRNN can be used to perform segment labeling in addition to simple segmentation.

We train two separate neural analyzers. The first outputs a coarse-grained label indicating the type of each analyzed morpheme, as in the example in the previous paragraph. The second outputs the unique ID of each morpheme (formatted slightly differently from what the Uqailaut analyzer produces due to processing constraints). The output from analyzing the word "qauqujaujunu" would be "qau_1v:3 qu_2vv:2 jaq_1vn:2 u_1nv:1 juq_1vn:2 nut_tn-dat-p:2." The fine-grained labels consist of various bits of grammatical information, such as root type, lexical postbase type, or grammatical ending type with noun case or verb mood indicators. Over the dataset used in the experiments here, 1691 fine-grained labels were counted.

Training data for the neural network is provided by the Inuktitut side of the Nunavut Hansard corpus. All word types were analyzed by the Uqailaut morphological analyzer. Word types having a single analysis were used in the training set, since this subset of word types can be argued to be the most correctly analyzed set out of what

is available. (Further experiments could make use of the types that have ambiguous, i.e. two or more analyses, as the training data, for example.) This subset consists of approximately 26K word types. For the "Coarse-Grained" model, 25,897 types were used. The Uqailaut analysis for each type was converted into a simple "surface-form/label," for each morpheme. A total of sixteen labels resulted from this conversion (listed in Table 1). A development (dev) set and a test set of 1000 items each were randomly held out from training.

ROOT	root (noun or verb)
LEX	lexical postbase
GRAM	grammatical suffix
CL	clitic
ADV	adverb
CONJ	conjunction
EXCLAM	exclamation
PR, RP, PRP	pronoun
AD	demonstrative adverb
PD	demonstrative pronoun
RAD	demonstrative adverb root
RPD	demonstrative pronoun root
TAD	demonstrative adverb suffix
TPD	demonstrative pronoun suffix

Table 1: List of Coarse-Grained Labels

For the "Fine-Grained" model, the full unique ID was used, with 1691 labels present in the set of analyses. Because the SRNN program did not allow for unseen labels when running in test mode, selection of the dev and test sets was not random and proceeded as follows: First, under the assumption that the greatest variation of labels would occur in the roots of the word types, (the "open-class" morphemes, versus the "closed-class" lexical post-base, grammatical endings, and clitics), the selection proceeded based on root labels. Of the 1,198 unique root labels, 898 occurred in two or more word types. For example, the root label "qauq_1v" occurs in six types, "qaurniq," "qaunimautilik," "qauqujaujut," "qauqujaulluni," "qauqujaujunu" and "qauvitaq." At least 1 of each of these types per root label was placed in the dev/test pool, with the remaining types containing that root label being assigned to the train set. To select which of the two or more types to put into each set, the longest (in terms of number of morphemes in the type) was selected for the dev/test pool, with the remaining going into the train set. Then, the dev/test pool was split into two sets of 449 items each.

8 Results

The program implementing the SRNN reports precision, recall, and f-measure calculated over segmentations alone (seg), or segmentations with tagging (tag). Table 2 shows the results on the two held-out sets for the two different models.

Model	Set	seg/tag	Precision	Recall	F-measure
Coarse-Grained	dev	seg	0.9627	0.9554	0.9591
		tag	0.9602	0.9529	0.9565
	test	seg	0.9463	0.9456	0.9460
		tag	0.9430	0.9424	0.9427
Fine-Grained	dev	seg	0.8640	0.8647	0.8644
		tag	0.7351	0.7357	0.7354
	test	seg	0.8291	0.8450	0.8369
		tag	0.7099	0.7235	0.7166

Table 2: Accuracy scores on full words

As would be expected, the model producing a coarse-grained output performs better than the model producing a fine-grained output. The model only has to decide between 16 labels in the former, versus 1691 labels in the latter. In addition, it should be kept in mind that the dev and test sets for the two models are not the same, due to the processing constraints of the program implementing the SRNN.

8.1 Accuracy on non-root morphemes

Most of the mislabeling errors occurred in the root morphemes of the analyzed words. As was mentioned above, the set of root morphemes can be likened to the set of "open-class" vocabulary, whereas the remaining morphemes (suffixes) of words are "closed-class." In order to attempt to filter out the randomness effect of trying to identify "open-class" root morphemes, scores were calculated over the output of the Fine-Grained model leaving out the roots. Table 3 displays these results.

Model	Set	seg/tag	Precision	Recall	F-measure
Fine-Grained roots absent in scoring	dev	seg	0.8838	0. 8860	0.8849
		tag	0.8178	0.8199	0.8188
	test	seg	0.8560	0.8807	0.8682
		tag	0.7922	0.8151	0.8035

Table 3: Accuracy scores on suffixes only

As expected, these scores (suffixes only) are higher than those measured on the full words

(root+suffixes), yet still fall below 90% accuracy.

9 Conclusion

In this work in progress, we have proposed a method of improving the coverage of an existing morphological analyzer for the Inuktitut language using an SRNN bootstrapped from a portion of the analyses of words in a corpus. We show accuracy scores on two models, yielding coarse-, or fine-grained labels. Accuracy scores are also calculated on just the "closed-class" suffix strings (removing the "open-class" roots) which show a slight improvement over accuracy on the full words. Further experimentation is needed to refine the accuracy of this approach and to begin to develop methods of processing this highly complex, polysynthetic language.

References

Creutz, M. and Lagus, K., 2006. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*.

Dorais, L., 1990. "The Canadian Inuit and their Language," in *Arctic Languages, An Awakening,* Ed. Dirmid R. F. Collins, Unesco, pp. 185-289.

Kohonen, O., Virpioja, S., and Lagus, K., 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pp. 78-86. Association for Computational Linguistics.

Kong, L., Dyer, C., and Smith, N., 2015. Segmental recurrent neural networks, *arXiv preprint arXiv:1511.06018.*

Malouf, R., 2016. Generating morphological paradigms with a recurrent neural network, *San Diego Linguistic Papers 6*, 122-129.

Martin, J., Johnson, H., Farley, B., and Maclachlan, A., 2003, Aligning and Using an English-Inuktitut Parallel Corpus, HLT-NAACL Workshop: Building and Using Parallel Texts Data Driven Machine Translation and Beyond, pp. 115-118, Edmonton, May-June 2003.

Morita, H., Kawahara, D., and Kurohashi, S., 2015. Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model, Association for Computational Linguistics, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2297, Lisbon, Portugal.

Narasimhan, K., Barzilay, R., and Jaakkola, T., 2015. An Unsupervised Method for Uncovering Morphological Chains. *Transactions of the Association for Computational Linguistics*, [S.l.], v. 3, pp. 157-167.

Wang, L., Cao, Z., Xia, Y., and de Melo, G., 2016. Morphological segmentation with window LSTM neural networks. In Proceedings of AAAI.