*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 2521–2527
Marseille, 11–16 May 2020


# Automatic Transcription Challenges for Inuktitut, a Low-Resource Polysynthetic Language

**Vishwa Gupta, Gilles Boulianne**
Centre de recherche informatique de Montréal (CRIM)
{vishwa.gupta, gilles.boulianne}@crim.ca

**Abstract**

We introduce the first attempt at automatic speech recognition (ASR) in Inuktitut, as a representative for polysynthetic, low-resource languages, like many of the 900 Indigenous languages spoken in the Americas. As most previous work on Inuktitut, we use texts from parliament proceedings, but in addition we have access to 23 hours of transcribed oral stories. With this corpus, we show that Inuktitut displays a much higher degree of polysynthesis than other agglutinative languages usually considered in ASR, such as Finnish or Turkish. Even with a vocabulary of 1.3 million words derived from proceedings and stories, held-out stories have more than 60% of words out-of-vocabulary. We train bi-directional LSTM acoustic models, then investigate word and subword units, morphemes and syllables, and a deep neural network that finds word boundaries in subword sequences. We show that acoustic decoding using syllables marked with word boundary markers results in the lowest word error rate.

**Keywords:** DNN, deep neural networks, automated transcription of Inuktitut, polysynthetic language, low resource language

## 1. Introduction

Inuktitut is one of the 60 Indigenous languages from 12 distinct language families currently spoken in Canada, and approximately 36,000 people declare Inuktitut as their mother tongue[1]. There is a growing interest shown by Indigenous communities, and federal and provincial governments, in the development of language technology (Littell et al., 2018), as it could help revitalize historical recorded archives, provide language course materials, and create apps to promote language dissemination among the young. Inuktitut is described as a highly polysynthetic language: its words are composed of many morphemes, such that single words can express what usually requires a whole sentence in other languages. Polysynthetic languages are often termed agglutinative when their morphemes have clear boundaries and thus are easily segmentable (Klavans, 2018). It is not entirely the case in Inuktitut, where morphemes combine in rich and complex ways that affect their pronunciation, giving rise to a complicated mapping between surface segmentation and underlying morphemes (Micher, 2017).

Previous work in speech recognition have investigated several polysynthetic agglutinative, resource-rich languages such as Finnish, Estonian, Turkish, Korean, or Hungarian (Mihajlik et al., 2007) (Kurimo et al., 2007) (Erdoğan et al., 2005) (Kwon and Hwang, 1999). Speech recognition systems usually rely on a large enough word lexicon to cover most words in an unseen text, as each occurrence of an out-of-vocabulary word will cause at least one recognition error. Polysynthetic languages challenge this approach since they easily generate very large number of distinct words through combination of morphemes.

Many studies resort to subword units in order to increase the coverage with a reasonable size subword lexicon. Several possible subword units have been studied, including syllables (He et al., 2016)(Enarvi et al., 2017)(Smit et al., 2017), byte-pair encoding (Smit et al., 2017), or graphemes (Mihajlik et al., 2007). Most studies focus on morphological subword units, using Morfessor (Virpioja et al., 2013) to automatically generate a morpheme dictionary that yields a high word coverage, while keeping the total number of morphemes to a reasonable level. For example in (Kurimo et al., 2007), the authors get significant reduction in word error rates (WER) for Finnish, Estonian and Turkish using these morphemes compared to using a word dictionary. Even in a low-resource situation, simulated using Finnish and Estonian, subword units are advantageous (Kurimo et al., 2017).

Our contributions in this paper are fourfold. First, we use transcribed oral stories, not just parliament proceedings. Stories are more diverse than parliament proceedings that contain lots of repetitive and stereotyped content. Morphological analysis of oral stories turns out to be more difficult both for rule-based and automatically trained analyzers. Second, we show that a highly polysynthetic language like Inuktitut is much more challenging for conventional ASR than agglutinative languages usually studied, with OOV rates at least 4 times larger than Finnish, Estonian and Turkish, for comparable vocabulary sizes. Third, we set the first published baselines for speech recognition in Inuktitut, to our knowledge. Fourth, we reformulate the problem of segmenting subword sequences into words as a classification problem, and train a deep neural network to this end.

In the following sections we provide more details about the data we used, and describe our acoustic model training. We have also tried Morfessor with Inuktitut to generate morphemes and compared these morphemes with words and syllables as units for recognition. In syllable recognition, we tried two different variants to convert decoded syllable sequences to word sequences. In one variation, we distinguish between syllables at the start of word, within word and at the end of the word in the dictionary. The resulting language model contains word boundary information. The decoded sequence of syllables are then converted into word sequence with the help of the syllables marked with begin

[1]Statistics Canada, 2016 Census of Population.

or end of word. Another variation for converting syllable sequences to word sequences is to train a DNN that takes syllable sequences as input and generates word boundary markers (whether the current syllable is at the end of the word or not). These word boundary markers are then used to generate word sequences from syllable sequences. We show that syllables result in the lowest perplexity and also the lowest word error rate.

## 2. Acoustic and language model training data

Inuktitut recordings and transcriptions were provided by the Pirurvik Centre[2]. Recordings are stories told by renowned elders, and they sometimes include singing without instrumental accompaniment. There were 64 transcribed wave files for a total of 25.92 hours of audio and 63419 transcribed words. The recordings contain a total of 15 male and 8 female speakers. One male and one female speaker (7 files and 7673 words) were used for development, and the 57 remaining audio files (with 53299 words) were used for training. Because the available dataset is so modest, and has few speakers, we opted for a small development set, and no separate test set; this should be taken into account when interpreting the results presented here.
The Inuktitut text is in syllabics, a writing system where each consonant-vowel pair is represented by single symbol. Mappings from syllabics to International Phonetic Alphabet were taken from the Wikipedia Inuktitut pronunciation key[3] with some minor hand corrections.
Possible syllables in Inuktitut are: V, CV and CV+final consonant, V+final consonant. Total number of syllables is 2,304. We had to provide special handling for:

- Characters that can be present in borrowed words (b and H) are mapped to a matching IPA symbol.
- Isolated diacritics, which are not allowed but are found in practice, are assigned to preceding vowel.
- For syllabics-to-roman conversion, post-processing is applied to convert roman sequence 'qk' to 'qq' (and the reverse in roman-to-syllabic conversion).

We convert text in syllabics to roman in order to make it easier to visualize.
For language modeling, we have Nunavut Parliament Hansards[4] for the years 2006–2016, containing 6.7 M words of text (1.34 M distinct words). We added 53 k words in our acoustic training set to this data. For a 300 k word lexicon generated from this combined data, the weighted out-of-vocabulary (OOV) rate is 61.5% for the text in the development set, so the coverage is very low. One of the reasons could be that the contents of parliament proceedings are very different from those of the development set oral stories. The total breakdown for the total number of word tokens (words), number of distinct words (vocab), number of syllable tokens (syll), and number of distinct syllables (syll voc) for various data partitions is shown in Table 1.

[2]https://www.pirurvik.ca/
[3]https://en.wikipedia.org/wiki/Help:IPA/Inuktitut
[4]Kindly provided by Marc Tessier from NRC.

| Source | Words | Vocab | Syll | Syll voc |
|---|---|---|---|---|
| Hansards, train | 6.5 M | 1.32 M | 28.1 M | 3633 |
| Hansards, dev | 148 k | 47.0 k | 775 k | 2075 |
| Acoust. train | 53.3 k | 31.9 k | 301 k | 2009 |
| Acoust. dev | 7.6 k | 4.90 k | 42.7 k | 1185 |

Table 1: Training and development text sources.

## 3. Experiments with transcribed data

In order to get the lowest possible word error rate (WER), we tried four different strategies: recognition using a large word-based dictionary, morpheme-based subword units generated using Morfessor (Virpioja et al., 2013) and syllable-based subword units with/without word boundary markers. For syllables without word boundary marker, we trained a DNN that provided word markers for syllable sequences.

### 3.1. Training acoustic models

All the above recognition experiments do not affect the training, since all the above multiple recognition scenarios use the same set of phonemes. For acoustic training, the dictionary contains all the words in the acoustic training set and they are transcribed using an X-SAMPA[5] phoneme set. The acoustic models are trained with roughly 23 hours of audio in Inuktitut (14 male and 7 female speakers).
Since the Inuktitut training dataset is small, we trained it together with about 4,000 hours of English audio from LibriSpeech and LDC datasets, including Hub4, RT03, RT04, Market, WSJ, switchboard and Fisher. The following steps outline the complete training:

1. Train bi-directional long-short-term-memory deep neural network (BLSTM) acoustic models using English + Inuktitut audio. Using Inuktitut audio is important as some phones in the Inuktitut dictionary are not found in the English dictionary. Both the English and Inuktitut dictionaries use X-SAMPA phones. We use 40-dim MFCC features together with 100 dimensional i-vectors (Gupta et al., 2014) (Saon et al., 2013) (Senior and Lopez-Moreno, 2014) as input features to the BLSTM acoustic models. We use the Kaldi toolkit (Povey et al., 2011) for training the BLSTM acoustic models.
2. Speed perturb Inuktitut data with speeds of 0.9 and 1.1 (Ko et al., 2015) to generate additional acoustic data for Inuktitut.
3. Starting with models trained in the previous step, train new BLSTM models with just the Inuktitut speed perturbed data for 6 epochs (adapt to Inuktitut data).
4. Create a new alignment with models trained in the previous step, then train new BLSTM models with just

[5]https://fr.wikipedia.org/wiki/X-SAMPA

the Inuktitut speed perturbed data for 6 more epochs (2nd adaptation).

5. There were two files in the Inuktitut training set that did not align, so we assumed that there was something wrong in the transcript. So we purified the two transcripts (Manohar et al., 2017) by generating another transcript through recognition, then comparing the two transcripts, and only using segments of the transcripts that match well. Using these two audio files with the purified transcripts and doing another adaptation iteration (6 epochs) with speed perturbed Inuktitut data reduced the WER by a small margin. The resulting models are the final BLSTM models we used for all the recognition experiments.

We also tried TDNN-F models, but we got significantly worse results (20% relative degradation of WER), which is consistent with what we observed on another Indigenous language Cree.

### 3.2. Word-based lexicon

The language model and the lexicon were created from 6.7 million words of Nunavut Hansards and from 51.2 k words of acoustic training text. The 6.7 M words of Nunavut Hansards text was divided into 6.5M for training and 148 k for validation (lm dev). The Nunavut Hansards training text together with the acoustic training text have a total of 1.3 M distinct words.

Figure 1 shows the out-of-vocabulary rate (weighted by word frequency) on held-out development text from the acoustic (ac dev) and Hansards (lm dev) sources. Although the Hansards rate goes below 20% for vocabulary size greater than 300 k words, the acoustic development set rate stays above 60% even when all words found in Hansards and acoustic training sets are combined in a 1.3 M word vocabulary. Not surprisingly, a 4-gram language model trained on the Hansards has a perplexity of 252 on the Hansards development set but 32,000 on the oral stories development set.

In contrast a 69 k word lexicon in Finnish has an OOV (out-of-vocabulary) rate of 15%, a 60 k word lexicon in Estonian has an OOV rate of 10% and a 50 k word lexicon in Turkish has an OOV rate of 9% (Kurimo et al., 2007). So at more than 60% OOV, a lexicon of 70 k words in Inuktitut has more than 4 times the OOV rate of these languages.

The selected vocabulary for Inuktitut contains the most frequent 100 k words from language model training set (or equivalently all words that appear at least 3 times) plus all the words in acoustic training, for a total of 129,330 distinct words. This vocabulary has a weighted OOV rate of 62.6% on acoustic development set and 26.3% on LM development set.

Since the OOV rate is very high, perplexity depends very much on how you treat the OOV words. With a single placeholder word for all OOV words, the acoustic dev set perplexity is around 99. But this figure is misleading, since it only tells how easy it is to predict that a word is out-of-vocabulary. During recognition only in-vocabulary words can be considered, so a more realistic perplexity measure is

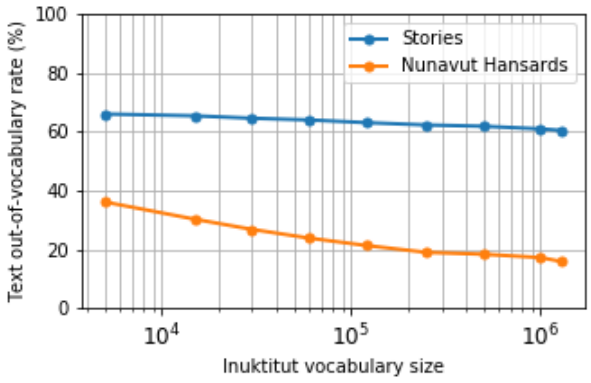


Figure 1: Inuktitut words out-of-vocabulary rate as a function of vocabulary size.

obtained by ignoring OOV words, and that value is around 11,000.

We interpolate between a language model trained on the LM training set and a language model trained on the acoustic training set, optimizing the weight of the language model trained on acoustic training set for best perplexity (of the interpolated language model) on the acoustic development set. The best interpolated model (with a weight of 0.25 on the LM trained on the acoustic training set) has a perplexity of 1251, ignoring OOV words. Both 3-gram and 4-gram LM have very similar perplexities (1250.86 for 3-grams, 1250.9 for 4-grams).

Despite the low coverage of words in the acoustic development set, we did speech recognition using the lexicon of 129 k words in order to compare with alternative strategies. With our best acoustic models, the word error rate (WER) we achieve on the dev set is 108.7%. There are a total of 7,673 words in the dev set and the error breakdown is as follows: 2,470 insertions, 108 deletions, and 5,762 substitutions. Since 4,803 words are OOV, at best the recognizer could have recognized 2,870 words. So the percentage correct for these words is (7673-5762-108)/2870 or 62.8%. The large insertion rate is due to many long words broken down into shorter words. It is probably easier to match the sequence of syllables using shorter words than longer words.

### 3.3. Morpheme subword units

Most previous work on agglutinative languages have shown that morpheme-based subword units can provide significant reduction in word error rate (Kurimo et al., 2017)(Mihajlik et al., 2007)(Erdoğan et al., 2005). We used Morfessor (Virpioja et al., 2013) to derive a set of morpheme-like units that provide good coverage of Inuktitut text. A Morfessor model can be trained in an unsupervised way, from unannotated raw text, or semi-supervised if reference morphemes are available for part of the words in the training set.

To provide reference morphemes for supervision and for measuring Morfessor model accuracy, we use the rule-based morphological analyzer Uqailaut[6], a well-known

[6]Freely available on www.inuktitutcomputing.ca/Uqailaut

computational model of the rich morphology of Inuktitut, developed at the National Research Council of Canada by Benoît Farley using Nunavut Hansards available prior to 2014. It finds at least one decomposition for around 65% (Nicholson et al., 2012) to 70% (Micher, 2017) of the vocabulary in the Nunavut Hansards. Recent work uses recurrent neural networks to extend the coverage (Micher, 2018).

In our case, Uqailaut was able to successfully decompose only 46% of the word vocabulary found in the Hansards training set of 6.5 M words from Table 1, possibly because it contains text posterior to Uqailaut development. However, Uqailaut found a decomposition for only 23% of the vocabulary from the acoustic training set, showing how much words in oral stories differ from words in parliament proceedings.

We used Uqailaut surface form decompositions of the acoustic development set as the reference analysis. For each word, we check if the Morfessor decomposition is amongst one of the possible Uqailaut decompositions (when a decomposition exists), and express the accuracy as the number of words with a correct decomposition, relative to number of reference words that have a decomposition.

We tried combining the acoustic training text with various amounts of Hansards text for training Morfessor models. The best unsupervised model was obtained with 53 k words from acoustic training and 515 k words from Hansards, and its accuracy was 41% on the held-out acoustic development set. Semi-supervised training of Morfessor models improves the accuracy to 61%, using a mix of 53 k words from acoustic training and 109 k words from Hansards.

We use the surface segmentation provided by the trained Morfessor model, and add word begin and end markers B_ and _E to morphemes to get morpheme subword units with word boundary markers.

The vocabulary is selected to contain morpheme units from LM training set that appear at least 3 times, plus all the morphemes in the acoustic training set, for a total of 35,057 distinct morphemes in the unsupervised training case, and 23,159 in the semisupervised case, as shown in Table 2. This vocabulary has a weighted OOV rate of 0.83% on acoustic development set for unsupervised morphemes, and 0.40% for the semisupervised ones.

LMs were trained on the LM training set which contains 14.7 M morpheme tokens, and the acoustic training set which contains 127 K morpheme tokens. They were interpolated with a weight of 0.25 for the acoustic training set (which provided the best development set perplexity).

The morpheme units LM models selected for recognition are the interpolated 4-grams, with probabilities renormalized after removing the OOV placeholder. Their perplexities on the acoustic development set are shown in Table 2.

Perplexity is usually reported as word perplexity, and considering each subword unit (morphemes here) as a word yields a unit-based perplexity (U-ppl) shown in Table 2. However, this measure is not comparable across different subword units, with varying vocabulary sizes and token lengths. Here we also use character-based perplexity (C-ppl), based on counting roman characters rather than words or units. It is directly related to the bits-per-character measure (Narasimhan et al., 2015), and is less dependent on the subword inventory.

| Model | Voc | U-ppl | C-ppl | OOV |
|---|---|---|---|---|
| Unsupervised | 35,057 | 2212 | 4.312 | 0.83% |
| Semisupervised | 23,159 | 778 | 4.312 | 0.40% |

Table 2: Morpheme unit language models evaluated on acoust dev set. U-ppl is unit-based perplexity (morphemes here), C-ppl is character-based perplexity.

We used morpheme units and 4-gram LMs to recognize the Inuktitut development set. Note that the morphemes are augmented with word markers so that the sequence of decoded morphemes can be combined into a sequence of words. Unsupervised morphemes result in a WER of 80.7%, and semisupervised morphemes in a WER of 79.4%.

### 3.4. Syllable subword units

Although perplexity (U-ppl) figures for morphemes cannot be directly compared with perplexity for words, they still look large for an inventory of a few tens of thousand units. As an alternative, we tried syllables as units. Syllable units are based on actual syllables and are different from syllabic characters. The following is an example of 3 sentences decomposed into syllable units (sentence breaks are not realistic but were added for illustration). Note that B_ and _E are added to syllable begin or end to represent start or end of word:

```
taanna maligaksaq kiinaujait atuqtuksat
akitujuqturutiksait
maligaq pingajuannik uqalimaaqtauqullugu

<s> B_taan na_E B_ma li gak saq_E
B_kii na u ja it_E B_a tuq tuk sat_E </s>
<s> B_a ki tu juq tu ru tik sa it_E </s>
<s> B_ma li gaq_E B_pi nga ju an nik_E
B_u qa li maaq ta u qul lu gu_E </s>
```

The number of such units in the Hansards and acoustic data is shown in Table 1, 4th and 5th columns. The following Table 3 gives perplexities on the acoustic development set of various syllable unit language models. OOV rate is 0.1%. All the LMs are interpolation of an LM trained on the Hansards and one trained on the acoustic training set text. For reference, we also give perplexities for byte-pair encodings (BPE) (Sennrich et al., 2015) and SentencePiece units (Kudo and Richardson, 2018), trained on same texts as syllables units, and both with a vocabulary of 3,000 units. So it turns out that we can get lower perplexity with syllables than with words or morphemes. Also, the number of syllables are much smaller than the number of morphemes or words. So the real question is how well can we perform with syllable recognition, and how well the syllable sequences translate into word sequences.

### 3.5. Subword sequence segmentation

There are two ways we can do word recognition for Inuktitut using syllables. One is to use word boundary markers

| Model | Order | U-ppl | C-ppl |
|---|---|---|---|
| Syll. B_ _E marks | 4-gram | 31.0 | 4.093 |
| Syll. no marks | 4-gram | 31.6 | 4.132 |
| BPE | 4-gram | 72.5 | 4.083 |
| SentencePiece | 4-gram | 37.1 | 4.038 |
| Syll. B_ _E marks | 5-gram | 30.8 | 4.085 |
| Syll. no marks | 5-gram | 31.4 | 4.120 |
| BPE | 5-gram | 72.1 | 4.075 |
| SentencePiece | 5-gram | 36.6 | 4.018 |

Table 3: Interpolated language model perplexities for various subword units.

| Model size | Input syllables | % WER |
|---|---|---|
| 100/20 | +/- 4 | 76.5 |
| 250/50 | +/- 5 | **76.3** |
| 250/50 | +/- 6 | 76.6 |
| 250/50 | +/- 7 | 76.5 |
| 100/20 | +/- 5 | 76.6 |
| 250/50 | +/- 10 | 77.0 |

Table 4: % WER with varying number of syllables as input to the feed forward DNN.

(B_ and _E) to represent begin and end of syllables, and just as in the case of morphemes, these markers will mark the word boundaries during decoding. We generated the syllabic dictionary from the LM training and acoustic training set to represent the most frequent syllables. The dictionary has a total of 3,158 syllables, and the syllable OOV rate for the dev set is only 0.1%. With our best acoustic models, the word error rate (note that syllables have word end markers) is 74.3%. The corresponding syllable error rate is 34.9%. Out of 44,747 syllables, there are 481 insertions, 5,109 deletions and 9,737 substitutions. So percent correct syllables is (44747-5109-9737)/44747 or 66.8%.

Another way of converting syllable sequences to word sequences is to recognize syllables without word boundary markers, and then use a DNN to mark the word boundaries based on syllable sequences input to the DNN. In other words, we train a DNN that outputs word boundary markers. In an oracle experiment using the reference syllable labels, we found that if the DNN was able to perfectly identify word boundaries, word error rate would drop to 70.4%. The input to the DNN are syllables and the output is a marker that tells whether the input syllable corresponds to last syllable in the word or not. We trained two different DNNs: one DNN is bidirectional LSTM (BLSTM), and the other one is a simple feedforward neural net.

The BLSTM has only one syllable as input and two softmax outputs (0 = no word boundary, 1 = word boundary). The BLSTM has two layers with cell dimension of 512 and recurrent projection dimension of 128. The BLSTM is trained with the syllable sequences from the Nunavut Hansards training text and from the acoustic training text. The input to the BLSTM is a one hot vector corresponding to the syllable. The best result we obtained with BLSTM model is a WER of 80.16% after training for 6 epochs.

We also experimented with simple feed forward neural net with 5 hidden layers. Each hidden layer has 250 outputs that go into a p-norm component (Zhang et al., 2014) (p = 2). Each p-norm component has 50 outputs. The final layer is a sigmoid with 2 outputs. We varied the number of syllables input to this DNN from +/- 4 syllables to +/- 10 syllables. Each syllable is input as a one hot vector. Results are shown in Table 4. The best results are with +/- 5 syllables as input, but all the results are quite close.

The above results are comparable since we used the same syllable sequence to transform to word sequence. Since the input is the same, we added the posterior log likelihoods to see if we can reduce WER by averaging the log likelihoods. When we add the posterior log likelihoods of many DNNs including the LSTM, the best result we get is 75.6% WER, so the WER does go down by 0.9% absolute.

In the above scenario, we are training the DNN with correct syllable sequences to label syllables with word boundary. So the DNN only sees the correct syllable sequences during training. However, when we label decoded syllable sequences, the sequences have many insertions and deletions which were not seen in training. In order to compensate for that, we recognized the acoustic training set, aligned the resulting syllable sequences with the reference syllable sequences, and marked the recognized syllable as a word boundary or not based on the reference syllable marker. We then used this recognized and marked syllable sequence to train the DNN further for a few iterations with small learning rate. The resulting DNN reduced WER from 76.3% to 76.1%. The effect is small because the acoustic training set is small.

Marking word boundaries by using syllables with word boundary markers during decoding gives lower WER than marking word boundary by using a DNN as in the previous two paragraphs. Another thing we can try is to input the word boundary marker information from decoded syllable sequence that uses syllables with word boundary marks, to the DNN trained without any syllable markers. So the additional input is 0 or 1 based on whether the central syllable is a word boundary or not. This should give additional information to the DNN whether the syllable is a likely word boundary or not. However, the training set is significantly reduced since the number of syllables in the LM training set (from Hansards) is much larger than in the acoustic training set. Probably due to this reason, the WER gets worse (79.5%). The problem with using acoustic cues for word boundary detection is that the acoustic cues have to be derived from the acoustic training set, and we have a severly limited acoustic training set currently.

Table 5 summarizes our WER results for the various subword units. The 4-gram syllable language model obtains an actual WER of 74.3%, compared to 70.3% with oracle word boundaries, while using word boundaries from the DNN yield a 75.6% WER (last line). So why does 4-gram syllable language model work so well (there is only 3.9% absolute difference between oracle and actual WER), and why doesn't the DNN give better word boundary detection than the 4-gram LM? The syllable unit dictionary and the

| Units | N. of units | OOV rate | WER |
|---|---|---|---|
| Words | 129 k | 62.6% | 108.7% |
| Unsup. morph. | 35.1 k | 0.8% | 80.7% |
| Semisup. morph. | 23.2 k | 0.4% | 79.4% |
| Syll. + B_, E_ | 3.2 k | 0.1% | **74.3%** |
| Syll. + DNN bound. | 3.2 k | 0.1% | 75.6% |

Table 5: Summary of weighted OOV rate and WER on the development set, for various subword units.

4-gram language model are created from 6.5 million words from Hansard and 53k words from acoustic training set. Table 6 shows part of this syllable dictionary. The syllables are shown in roman characters. Note that syllable *aaj* can only be in the beginning or middle of the word, syllable *kaal* can be in begin, middle or end of the word, while syllable *aak* can even be a single word. The limited set of possible syllable positions, together with the 4-gram language model, impose strong syllabic constraints and yield correct word boundary markers after decoding. The low development set perplexity and OOV means that these syllabic constraints are equally valid for the development set. They may also be independent of the context, for example parliament proceedings versus stories. So syllables seem to be good subword units for language modeling and decoding for Inuktitut, since a strong LM would provide fairly accurate word boundary markers.

| Syllable | Pronunciation |
|---|---|
| B_aaj | a: j |
| aaj | a: j |
| B_kaal | k a: l |
| kaal | k a: l |
| kaal_E | k a: l |
| B_aak | a: k |
| aak | a: k |
| aak_E | a: k |
| B_aak_E | a: k |

Table 6: Some dictionary entries for syllables.

So why word markers using DNNs do not give better results than the 4-gram LM? The reason seems quite simple. The 4-gram LM is used jointly with acoustic model for decoding, whereas the DNN takes the best syllable sequence from 4-gram LM decoding and looks for word boundary marks. We need to jointly decode a DNN-based LM with the acoustic model in order to take advantage of the better prediction capability of DNN models. Usually what is done is to generate multiple choices from an N-gram LM and then to rescore with a recursive DNN-based LM. This modification should give us better results.

## 4. Conclusion

In this work, we used transcribed Inuktitut oral stories in addition to parliament proceedings, in contrast to most previous work. We found that the existing rule-based analyzer can decompose only a small fraction of the words in these stories, and that a trained Morfessor model can predict morphemes with at most 60% accuracy.

We also found that Inuktitut's highly polysynthetic morphology is a tough challenge for conventional word-based approaches. The out-of-vocabulary (OOV) rate is more than four times higher than that of other, previously studied agglutinative languages, for a similar large vocabulary size. For that reason, the word error rate (WER) with a word-based dictionary is over 100%. On the development set, over 62% words were OOV, and out of the remaining words, the correct recognition rate was 63%. Long OOV words were transcribed as a sequence of short words resulting in significant insertion rate.

We tried two sub-word units in order to reduce WER. One subword unit is morphemes. We used Morfessor to generate an optimized list of morphemes. Using these morphemes we were able to reduce the WER to 79.4%.

Another sub-word unit we tried are syllables. Compared to morphemes, the syllables significantly reduced the size of the dictionary from 23.2 k morphemes to less than 3,600 syllables. Also the perplexity went down from 778 to less than 40. The WER goes down from 79.4% to 74.3%.

We also tried different ways of associating word boundaries with a sequence of syllables. In one case the begin or end of word is associated with the syllable itself, so the decoded syllable sequence contains word boundary markers. This method gave 74.3% WER, while training a DNN to find word boundaries in a sequence of syllables gave 75.6% WER. We tried to associate acoustic cues with syllables to improve word boundary detection, but the results are poor because the acoustic data is much smaller than the language modeling data.

## 5. Acknowledgements

This work was funded in part by the National Research Council of Canada (NRC) and the Ministère de l'économie, innovation et exportation (MEIE) of Gouvernement du Québec.

## 6. Bibliographical References

Enarvi, S., Smit, P., Virpioja, S., and Kurimo, M. (2017). Automatic Speech Recognition with Very Large Conversational Finnish and Estonian Vocabularies. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 25(11):2085–2097.

Erdoğan, H., Büyük, O., and Oflazer, K. (2005). Incorporating language constraints in sub-word based speech recognition. In *Proc. ASRU*, volume 2005, pages 281–286.

Gupta, V., Kenny, P., Ouellet, P., and Stafylakis, T. (2014). I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In *Proc. ICASSP*, pages 6334–6338.

He, Y., Baumann, P., Fang, H., Hutchinson, B., Jaech, A., Ostendorf, M., Fosler-Lussier, E., and Pierrehumbert, J. (2016). Using pronunciation-based morphological subword units to improve OOV handling in keyword search. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 24(1):79–92.

Klavans, J. L. (2018). Computational Modeling of Polysynthetic Languages. In *Proc. Workshop on Polysynthetic Languages*, pages 1–11.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio Augmentation for Speech Recognition. In *Proc. Interspeech*, pages 1–4.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proc. EMNLP*, pages 66–71.

Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimäki, T., Pylkkönen, J., Alumäe, T., and Saraclar, M. (2007). Unlimited vocabulary speech recognition for agglutinative languages. In *Proc. NAACL HLT*, pages 487–494.

Kurimo, M., Enarvi, S., Tilk, O., Varjokallio, M., Mansikkaniemi, A., and Alumäe, T. (2017). Modeling under-resourced languages for speech recognition. *Language Resources and Evaluation*, 51(4):961–987.

Kwon, O.-w. and Hwang, K. (1999). Korean large vocabulary continuous speech recognition with morpheme-based recognition units. *Speech Communication*, 39(2003):10–13.

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

Manohar, V., Povey, D., and Khudanpur, S. (2017). JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning. In *Proc. ASRU*, pages 346–352.

Micher, J. (2017). Improving Coverage of an Inuktitut Morphological Analyzer Using a Segmental Recurrent Neural Network. In *Proc. 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106.

Micher, J. C. (2018). Using the Nunavut Hansard Data for Experiments in Morphological Analysis and Machine Translation. In *Proc. of Workshop on Polysynthetic Languages*, pages 65–72.

Mihajlik, P., Fegyó, T., Tüske, Z., and Ircing, P. (2007). A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian. In *Proc. Interspeech*, pages 1497–1500.

Narasimhan, K., Karakos, D., Schwartz, R., Tsakalidis, S., and Barzilay, R. (2015). Morphological Segmentation for Keyword Spotting. In *Proc. EMNLP*, pages 880–885.

Nicholson, J., Cohn, T., and Baldwin, T. (2012). Evaluating a Morphological Analyser of Inuktitut. In *Proc. NAACL HLT*, pages 372–376.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proc. ASRU*.

Saon, G., Soltau, H., Nahamoo, D., and Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *Proc. ASRU*, pages 55–59.

Senior, A. and Lopez-Moreno, I. (2014). Improving DNN speaker independence with I-vector inputs. In *Proc. ICASSP*, pages 225–229.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural Machine Translation of Rare Words with Subword Units. http://arxiv.org/abs/1508.07909.

Smit, P., Virpioja, S., and Kurimo, M. (2017). Improved subword modeling for WFST-based speech recognition. In *Proc. Interspeech*, pages 2551–2555.

Virpioja, S., Smit, P., Grönroos, S.-A., and Kurimo, M. (2013). Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. Technical report, Aalto University.

Zhang, X., Trmal, J., Povey, D., and Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. In *Proc. ICASSP*, pages 215–219.