Technical Report on Subword Language Modeling for Icelandic ASR

Svanhvít Lilja Ingólfsdóttir

Language and Voice Lab
Reykjavik University, Iceland
svanhviti16@ru.is

1. Summary

This report describes preliminary research on subword modeling for automatic speech recognition (ASR) when applied to Icelandic.[1] Different methods of subword splitting are described and tested out by training a speech recognizer for Icelandic with a language model based on subwords/morphemes. The subword modeling methods all result in a lower word error rate than a comparable baseline word-based model, indicating that this is a viable approach for Icelandic ASR which warrants further study.

2. Introduction

Morphological complexity is one of the factors that influence how a language can be modelled before solving different language technology (LT) tasks. Icelandic is an example of a morphologically complex language, being a highly inflected language, where each base word can have many surface forms (up to 16 variations for nouns, 120 for adjectives and over 200 for verbs, though many of these forms are very rarely used). Apart from this, Icelandic has extremely productive compounding. In fact, compound words account for around 88% of all paradigms in the Database of Icelandic Morphology, according to Bjarnadóttir's research on the issue [1]. As an example of how active this process is in modern Icelandic it suffices to note how compounds are often created on the fly for one-time use, in words such as "þriðjudagslakkrísinn", or "Védísarmegin", which might make sense in one context and then never be needed again. It is therefore an impossible task to compile a complete list of these words. Additionally, named entities (names of people, places, etc.) and misspellings appear in all languages and account for many of the rare words that further complicate language modeling.

Finding successful ways of handling both all the different grammatical surface forms of words, and the many compounds, is crucial to reducing the rate of unknown words in LT tasks. One common approach is breaking whole words into smaller units (often called *word pieces* or *subwords*) before training machine learning models. This reduces the vocabulary size of the dataset, thereby also minimizing the unknown word rate and increasing the accuracy. This approach has for example proven useful in neural machine translation (NMT), handwriting recognition (HWR) and ASR, as we will be taking a closer look at.

Subword units can range from a single character (corresponding to a single phoneme), to one or more syllables, and up to whole words. In ASR, subword extraction has been used at both the acoustic model and the language model level. In this study, our focus is on the use of subwords for language modeling, i.e. when mapping speech to words and sentences.

Language models (LMs) in traditional ASR systems are typically n-gram models or neural models which have been trained on a text corpus to model the probability of a sentence being said. The source of the corpus data can be the transcribed training data itself, or a much larger external corpus of text. Another approach is to simply use single characters as input, which eliminates the problem of unknown words altogether. Without information on how words are written, it however becomes much harder to correctly figure out what was said, so this is rarely done. In subword modeling, instead of using the whole word as the base unit (token) for training the LM, the idea is to apply some algorithm that splits words into subwords of some length (one character or more), and use those subwords as the input token for the LM. This has been a known method in ASR for about a decade, and now many recent papers address the use of subwords in end-to-end ASR systems, both attention and CTC models [2, 3]. Finding the best subword segmentation method remains an open research problem, however. The answer to that question likely depends on the language for the most part, making it all the more important to research this problem in languages other than English.

The quality of ASR systems is usually evaluated using the word error rate (WER), the measure of how many errors are reported on the test set, divided by the total number of words in the test set. Another way of gaining insight into how well the system generalizes is to look at the out-of-vocabulary (OOV) words, i.e. words in the test set that are not seen in the training set. The benefits of using subwords for reducing the OOV word rate and the WER are most evident in morphologically rich languages (e.g. an inflective language like Icelandic or German, and agglutinative languages like Turkish or Finnish) where a single base word can have many surface forms, as we have mentioned earlier. This variety of surface forms means that even though one variation of a word has been seen in the training data , such as "strútur", nominative of the masculine noun "ostrich", others may not be present, such as "strúts", genitive of "ostrich". This would mean that "strúts" would be counted as an OOV word if appearing in the test data. However, if the word is broken up into its constituents, "strút-ur" and "strút-s", the probability of having seen each of these subword units in the training data increases dramatically, since words often share a grammatical structure, such as the genitive "s" of many masculine nouns ("maður", "hestur", etc). The ASR system would simply need to put together the already seen root "strút" and the ending "s" to correctly assume that the word spoken was in fact "strúts". This method would also help in finding other surface forms of the same word, such as "strútsins", "strútarnir", "strútunum", etc.

This particular example assumes that the subwords have been extracted based on morphological rules, but there are in

[1] This work is built on coursework by Bjarni Barkarson, Steinunn Rut Friðriksdóttir and the author, in Speech Recognition, an MSc level language technology course taught in the spring of 2020 at Reykjavik University.

fact two main approaches to subword splitting: either the subword analysis is based on knowledge on the morphology of the language, or an unsupervised method learns the subword analysis from some data it is trained on. In fact, such unsupervised methods, completely agnostic of the morphology of the language, can prove just as or even more efficient, and, in the following report, we will describe and try out a few such methods, as well as an Icelandic morphology-based compound splitter.

3. Background

Subword units have been used for improving ASR in a variety of languages, such as Turkish [4], Vietnamese [5], Finnish and Estonian [6], Russian [7], Japanese and Korean [8], Amharic [9] and Uyghur [10], to name a few. Though the approaches vary, the main idea is to restrict the vocabulary size by using subword units, thereby reducing the OOV rate and WER.

As already mentioned, methods for finding subwords range from language-agnostic unsupervised methods that don't take the grammar of the language into account, to supervised methods based on language-dependent lexical databases with morphological information. Here we examine some of these methods and take a look at how they have been used in speech recognition.

3.1. Subword modeling methods in ASR

The choice of method for breaking words into subwords depends on many factors, probably most importantly the language, and the task at hand. What works best for one language may not be feasible for a language with a different grammatical structure, and a subwording method that works well for part-of-speech tagging may not work as well for ASR. By looking at the literature, especially for morphologically complex languages, we can however get a general idea of the methods that have proven useful.

Unsupervised data-driven methods for finding subwords/morphological units include algorithms based on minimum description length (MDL) [11], Lempel-Ziv-Welch (LZW) [12] and byte-pair encoding (BPE) [13]. Some of these algorithms were not designed for LT at all, but for other uses, such as data compression, but have since been adapted for use in different LT tasks, such as NMT and ASR. Both BPE and LZW are old data compression algorithms, and BPE in particular has made an impressive comeback, as it is now the basis for tokenization in almost all of the advanced models that now dominate the LT field, such as BERT [14] and GPT [15]. Supervised or rule-based methods for subword segmentation, on the other hand, are mostly language-dependent. Subword or compound splitters exist for many languages, especially ones with productive compounding, such as German [16], but cannot be applied directly to Icelandic. The only compound splitter we know for Icelandic is Kvistur, which we discuss in Section 3.1.6, but first we will take a look at some unsupervised subword segmentation methods.

3.1.1. BPE

BPE, designed for data compression and introduced in 1994 [17], is an algorithm where common pairs of bytes are replaced by a byte that does not appear in that data. A symbol table is then used to revert the single byte back to the correct byte-pair. A modified version of the algorithm was introduced in a 2016 paper [13] for an NMT task, increasing the BLEU score, over a back-off dictionary baseline, by up to 1.1 and 1.3 for English→German and English→Russian translation tasks, respectively. Since then it has gained a lot of popularity, being a simple algorithm that can be applied to massive amounts of texts.

In this version of BPE, the algorithm has been modified for the task of word segmentation, by merging characters or character sequences, instead of raw bytes. The algorithm counts the frequency of each word in the training corpus, represents the word as a sequence of characters, and adds an end-of-word symbol to each word, so that the original tokenization can be restored. Then it counts all symbol pairs and replaces each occurrence of the most frequent symbol pair (e.g. 'A', 'B') with a new symbol 'AB'. This is done iteratively until a desired subword vocabulary size is reached (this is a hyperparameter set by the user).

This variation of BPE has since been widely used in subword tokenizers such as the one used for BERT, as it seems to offer a good balance between character and word representations, and can model all rare words.

3.1.2. WordPiece

WordPiece [8] is a very similar algorithm, originally designed at Google in 2012 for voice search in Japanese and Korean. The algorithms are almost identical, the difference being that instead of selecting the subword unit pairs depending on their frequency, the WordPiece algorithm selects the subword unit which increases the likelihood on the training data the most when added to the model.

In the official BERT models provided, the WordPiece method is used for subword segmentation. Though this exact implementation has not been publicly released, a similar library, SentencePiece, is available (see Section 3.1.4).

3.1.3. Unigram (subword regularization)

Subword segmentation can be ambiguous, and multiple segmentations are possible within the same vocabulary. Kudo [18] addresses this problem with so-called subword regularization, a subword algorithm which trains the model with multiple subword segmentations which are probabilistically sampled during training. This is implemented via a subword algorithm based on a unigram LM, and is therefore often simply called Unigram. As with BPE, a fixed vocabulary size is set before training starts. This unigram model is capable of outputting multiple subword segmentations with probabilities, and has been used extensively in large LMs, especially since it is provided with the SentencePiece library.

3.1.4. SentencePiece

Subword tokenizers such as BPE and Unigram usually require pre-tokenization of the input text before they can be applied. This can mean splitting on spaces, or using more sophisticated tokenization methods, such as rule based tokenizers. Not all languages have spaces, however, such as Chinese and Japanese, and tokenizers are not available for all languages that do have spaces. For a fully unsupervised processing of large corpora of texts in any language, the SentencePiece [19] library was developed. Instead of using some external pre-tokenization method, SentencePiece processes all text as a sequence of Unicode characters, and even whitespaces are regarded as a normal symbol. This means the input is raw sentences, as opposed to individual tokens.

SentencePiece offers a choice of using either BPE or Unigram for the subword segmentation, and as with these algorithms, the final vocabulary size is set at the beginning, often at 8,000, 16,000 or 32,000 word units.

3.1.5. Morfessor

As opposed to the previously described subword algorithms, which don't really care about the morphology of a language, Morfessor is an unsupervised algorithm that in a way tries to mimic morphological splitting, by inducing a simple morphology of a natural language from a large corpus [20]. Though Morfessor is language independent and isn't provided with any information about the morphology of a language, the segmentation obtained often resembles a linguistic morpheme segmentation. This is achieved by simultaneously building a morph lexicon and representing the corpus with the induced lexicon using a probabilistic maximum a posteriori model. We have not tested this method, but it is easily accessible as a Python library, and it should be interesting to see how it compares to the Icelandic compound splitter called Kvistur.

3.1.6. Kvistur

In their 2015 paper, Daðason and Bjarnadóttir propose a way to segment Icelandic compounds into subword units [21]. Originally, Kvistur is a statistical model which determines the most likely composition of a compound word using trunk segment trees. The statistical model is trained on a large corpus of manually split compound words from the Database of Icelandic Morphology (DIM, Beygingarlýsing íslensks nútímamáls). DIM is a manually assembled dataset of analyzed compound words in Icelandic containing 2.9 million unique word forms [1]. The program assesses how likely it is that two sub-words can be used to create a compound word.

Furthermore, Kvistur has recently been enhanced with the integration of a BiLSTM model, which outperforms previous statistical methods [22]. The model predicts the best splits between constituents of the compound words by using character embeddings as input to a BiLSTM layer. The output is fed into a second layer which makes a binary prediction for each character, determining whether or not the character marks the beginning of a new constituent. By using a character-based BiLSTM, this neural version of Kvistur approaches the problem of OOV words present in the first version. Whereas Kvistur 1.0 reaches a 94.35% precision and 91.44% recall, evaluated on how correctly it splits compounds in a 6000 word sample from the Icelandic Wikipedia, Kvistur 2.0 shows precision of 98.61% and recall of 97.26%, substantially outperforming the statistical version.

4. Experiments

In our experiments, we use the Kaldi speech recognition toolkit for training a baseline word-based ASR model as well as several subword-based ASR models, and comparing the results.

4.1. Data

The data used is the Málrómur open source corpus for Icelandic voice samples [23, 24]. This dataset was collected by researchers at Reykjavik University and The Icelandic Centre for Language Technology in collaboration with Google during the years 2011 to 2017, in a few phases. This corpus includes 119,090 validated voice examples, from about 560 individuals, out of which 108,568 examples are labelled as "correct". The material ranges from news report headlines to rare triphones. Out of the voice examples from the Málrómur corpus, we used the ones labelled "correct", a total of 136 hours, with a 90/10% train/test split. We have also made use of the General Pronunciation Dictionary for ASR [25], which contains around 134,000 unique words, learned from the Málrómur data and the Leipzig Corpora Collection project, and transcribed with the IPA. We used the words from this lexicon, but were inspired by [6], who instead of using a phonetic pronunciation lexicon, simply used the graphemes as phonemes. This allowed for simpler processing when creating a subword version of the lexicon.

4.2. GALE-arabic ASR recipe

The only ready-made subword-based recipe provided in Kaldi is found within the GALE-arabic recipe, which as the name suggest was created for the Arabic language [26]. The subword recipe (found in the `gale-arabic/s5c` directory in Kaldi) is based on the original recipe for Arabic, and uses the same configurations, apart from applying BPE to the language model data and lexicon, to create the subword training data and subword lexicon, and to train a subword LM. To achieve this, the recipe includes subword versions of the conventional data preparation scripts.

We have adapted this recipe to Icelandic, for use with the Málrómur data. Some text normalization for Arabic is applied in the original recipe, which we have removed, but other than that, the recipe could be used without much modification for training an Icelandic subword ASR model based on BPE.

In order to try out some other methods for subword segmentation, we also modified the scripts to bypass the BPE algorithm, and instead use data already prepared using other subword-level LMs. These LMs were applied to the training and test texts, as well as the pronunciation lexicon.

4.3. Acoustic modeling

The recipe applies standard acoustic modeling methods available in Kaldi. First, Mel-frequency cepstral coefficient (MFCC) features are extracted from the frames, Linear Discriminative Analysis (LDA) transformation projects the concatenated frames to 40 dimensions, and then Maximum Likelihood Linear Transform (MLLT) is applied. Speaker adaptation is applied by using feature-space Maximum Likelihood Linear Regression (fMLLR). The MFCC+LDA+MLLT combination is applied for training two different Gaussian Mixture Models (GMM) models; a diagonal GMM model and a subspace GMM (SGMM). The recipe also includes a DNN-HMM based model, which we couldn't use, due to technical and computational limitations. All of the models are standard 3-states context-dependent triphone models.

4.4. Language modeling and lexicon

The LM tool used for both the baseline and subword implementations is the SRI Language Modeling Toolkit (SRILM). The LMs were built using a standard n-gram setup with Kneser-Ney smoothing, and trained only on the Málrómur training transcripts. The subword splitting methods tested out in our implementation were BPE, SentencePiece (with Unigram), and Kvistur. Subword versions of the data were created for both the training corpus and the lexicon. For the BPE implementation available in the GALE-arabic recipe, the number of merge operations was set at 1,000, and the tokenization method is a simple whitespace tokenization. For the Unigram model trained in

SentencePiece, the vocabulary size was set at 8,000. These values were mostly selected based on suggestions given for each method, and are quite low, due to the relatively small corpus size. Selecting the best vocabulary size for these algorithms is however an open research topic, as discussed in [27], and should be further experimented with for Icelandic.

In the baseline implementation, the LM is a 3-gram word-level model trained on 379,544 tokens. In the subword implementations, the LM is a 6-gram model trained with the subwords as tokens, which means that the number of tokens increases in these models, depending on how aggressively the algorithm splits the text into subwords. This can be observed in Table 1 (for comparison, we tried training the baseline word-level model using a 6-gram order, with no change in results). No special text normalization methods were applied since the Málrómur text is already in a mostly normalized form, though some types of utterances, such as web pages, have not been fully normalized (ja.is, and not ja punktur is).

In most subword algorithms, a separator symbol is applied to the data to denote where a word is split into sub-units – we use “`@@`”. This way it is clear how the word has been split, and the original tokenization can be restored. There are different ways of applying these markings: one approach is a prefix/postfix method: applying the markings at the end or at the beginning of each word (`Ice landic@@` or `@@Ice landic`). Another is to apply them between word units, and there are in fact three variations of this latter approach: to the right of the subword unit, `Ice@@ landic`, to denote that it is followed by a subword, or to the left, `Ice @@landic`, to denote the beginning of a subword, or on both sides (`Ice@@ @@landic`). In [6], experiments with different positions are done on Finnish and Estonian, where the last marking method proves to give somewhat better results than the others, but the authors report that other languages have proven to behave differently in this regard. We have not experimented with this part, and have used the method used in the BPE recipe: marking on the right side of the word. This called for some pre-processing of the Unigram output, which uses the prefix method. We created scripts that take care of this and other necessary pre-processing.

4.4.1. *Vocabulary size*

By applying the different subword algorithms, the number of word units is reduced from the whole word version of the data. For the pronunciation lexicon, for example, we reduce the vocabulary size from 32,429 word units when using whole words down to only 1,098 units when BPE has been applied, as seen in Table 1. Unigram also achieves a considerable reduction in vocabulary size, down to 11,890 unique word units. Kvistur reduces the vocabulary size down to 20,490 units. See for example how the different algorithms split the following sentence:

- Word-level: `stúlkan kinkaði kolli og litla fólkið í stúdíóinu klappaði fyrir sveinbirni`
- BPE: `stúl@@ k@@ an k@@ in@@ kaði kol@@ li og lit@@ la fól@@ kið í stú@@ dí@@ ó@@ inu k@@ la@@ pp@@ aði fyrir sv@@ ein@@ bir@@ ni`
- Unigram: `stúlka@@ n kin@@ kaði kolli og litla fólk@@ ið í stúd@@ íó@@ inu klappa@@ ði fyrir svein@@ birni`
- Kvistur: `stúlkan kinkaði kolli og litla fólkið í stúdíóinu klappaði fyrir svein@@ birni`

Table 1: *Lexicon size and the size of the training text vocabulary for the different methods.*

	Lexicon size	Vocabulary size
Baseline (word-level)	134,866	32,429
Kvistur	49,151	20,490
Unigram	12,252	11,890
BPE	1,098	1,098

We can see that no grammatical aspects of Icelandic are preserved in the subword units produced by BPE. They are also quite short and don’t bear much discernible meaning, but on the other hand, they can be used to construct all words that may appear in the language (using the given characters). The Unigram subword units seem to mimic Icelandic morphology in parts, while in Kvistur, all subwords produced are morphologically derived (though in this particular sentence there is only one split).

The lexicon size is also reduced drastically by applying the different algorithms. We see that for both Unigram and BPE, the size of the vocabulary is equal or more or less equal to the lexicon size, showing us how most or all words can be produced using only the subword units. It should however be noted that the lexicon is learned from the Málrómur data, among other texts, which means that these text sources share most of their vocabulary already.

4.5. Decoding and evaluation

To evaluate the different models, we report the WER on the test set. In the GALE-arabic subword recipe, we noticed that the WER is calculated from the total number of subwords, not the total number of whole words, as in the baseline. This makes comparison between the baseline and different subwording methods less reliable, as the number of words/subwords differs for every method tested. This was easily remedied, however, by making use of scripts already available in the recipe, which concatenate the subwords back into whole words before calculating the WER. This way we had a common denominator for comparing different methods.

5. Results and discussion

Table 2 shows the WER scores obtained for the baseline and the subword models. The results show that each of the subword models have a lower WER than the baseline model, with BPE showing the lowest WER of 9.94%. Unigram scores 18.47% at the same decoding step, while Kvistur scores 16.37%. Scoring is performed on the test set once the word pieces have been concatenated back into whole words (detokenized).

The gain from using BPE was larger than anticipated, going from 23.61% WER down to 9.94%. It seems these small word pieces do a good job of modeling words, even though we were worried that such small units would decrease the constraints of the model too much, as discussed in [28]. Kvistur is the second best model tested here, scoring higher than Unigram, even with its less constrained vocabulary. This may be an indicator that accurate morphological information is a useful factor in subword modeling for Icelandic ASR, though it cannot be confirmed from a single experiment.

One possible caveat to mention: When training the Unigram and the Kvistur models, many warnings appeared regarding missing tree stats, which may be caused by misalignment in the data. For the Kvistur data, for example, warnings about missing stats are given for 23 out of 61 phones, indicating that these phones are not present in the data, which is unlikely, and might indicate some error. Though we have not figured out why these warnings appear, we must deduce that these results should be taken with a grain of salt, and it is possible they could improve if these issues were found and fixed. We were concerned this was due to issues because we had modified the training process to fit Unigram and Kvistur, so we tried using the already pre-processed BPE data as input into that same process before training again. This produced no such warnings, and the WER results for this BPE model were the same as for the original BPE training, so the problem, if there is one, might rather lie in the subword data preparation.

Notwithstanding these issues, the main takeaway is that these subword LMs all seem to give better results than a LM trained on whole words, indicating that this is a viable approach for Icelandic ASR which warrants to be studied further.

Table 2: *Comparison of the best WER (%) for each model, at different steps of the training process.*

Model	WER tri1	WER tri2	WER tri3
Baseline	23.03	24.08	23.61
BPE	**12.95**	**11.69**	**9.94**
Unigram	21.74	20.85	18.47
Kvistur	18.05	17.55	16.37

6. Suggestions for future work

There are many possibilities for different steps to take, going forward. In these preliminary experiments, we have only used one source of data, one LM setup (standard n-gram with Kneser-Ney smoothing), and a single Kaldi recipe. The idea was that by establishing a simple baseline model using these basic settings, we could use it for comparison with the different subword models. This however means that there are many aspects yet to be explored. First of all, the results of the DNN training, which we have not completed for these models, may tell us more about the models. Also, in our setup, the corpus data used for training the LM is simply the training data transcript. It is an obvious next step to train the LM on a much larger training corpus, and to use more sophisticated methods for training, such as BiLSTM networks.

Another thing to keep in mind is that the lexicon was not transcribed phonetically; we simply used the characters of the words as phonemes. This did not seem to be detrimental for the baseline model, because training using the original IPA-transcribed pronunciation lexicon did not show any improvements in WER as compared to a model trained using the grapheme-based one. We expect that this works because Icelandic orthography is fairly close to the pronunciation of the language, though it is by no means a 1:1 mapping between grapheme and phoneme. Using the graphemes as phonemes helped to simplify preprocessing, since we needed subword versions of the lexicon for each subword method, which would call for too much manual work or some very clever solutions.

The literature on subwords in LT at the moment is very centered around BPE/Unigram and the many variations of these methods, as used in the large self-attention models, such as BERT. Finding a way of leveraging these models for Icelandic ASR is one way of going forward, but a more straightforward solution to begin with is probably to develop a subword-based Kaldi recipe for Icelandic, which can be adapted to different methods.

What we propose for future work is continuing work with SentencePiece (both BPE and Unigram) and Kvistur, and possibly Morfessor, and train them on more data, try different vocabulary sizes and maybe experiment with different ways of handling the lexicon. We especially believe that Kvistur is a promising resource, since it is learned from a large supervised database with information on Icelandic morphology, which is hard to come by using other means. The drawback of Kvistur may be that its subword units are long. Subword length is a topic that should be researced further in Icelandic, especially seeing how well BPE did in comparison, with its very short subword units.

In this work we have adapted the subword GALE-arabic recipe for Kaldi and combined it with the Málrómur data to suit our needs, as it was handy in order to reduce work[2], but other approaches are of course available, such as composing a new Kaldi recipe from scratch. Other data sources could also be explored, such as the Samrómur data that is being collected. One drawback of using the Málrómur data was that utterances are often repeated, i.e. many speakers read out the same sentences, and so there is a lot of overlap in the training and test data. This makes it harder to speculate on OOV words, since most of the words in the test data also appear in the training data. This can of course be remedied using different pre-processing methods.

Even though the state-of-the-art in subword modeling for ASR is all about attention models, it is helpful to do the groundwork by testing out different subword segmentation methods in a more controlled environment. This is especially true for Icelandic, a morphologically complex language where (to our knowledge) no previous research exists on using subwords for ASR. This is what we have tried to do in this preparatory work, in order to gain some insight for future work in this area.

7. References

[1] K. Bjarnadóttir, "Phrasal compounds in Modern Icelandic with reference to Icelandic word formation in general," *Further investigations into the nature of phrasal compounding*, vol. 1, p. 13, 2017.

[2] J. Drexler and J. R. Glass, "Subword regularization and beam search decoding for end-to-end automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019, pp. 6266–6270. [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8683531

[3] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," *arXiv e-prints*, p. arXiv:1805.03294, May 2018.

[4] H. Sak, M. Saraclar, and T. Güngör, "Morphology-based and subword language modeling for Turkish speech recognition," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 5402–5405.

[5] V.-B. Le, S. Seng, L. Besacier, and B. Bigi, "Word/sub-word lattices decomposition and combination for speech recognition," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4321–4324.

[2] The pre-processing scripts and subword data used in the experiments presented here are available upon request, along with the modified GALE-arabic Kaldi subword recipe for Icelandic.

[6] P. Smit, S. Virpioja, M. Kurimo *et al.*, "Improved Subword Modeling for WFST-Based Speech Recognition," in *INTERSPEECH*, 2017, pp. 2551–2555.

[7] E. Shin, S. Stüker, K. Kilgour, C. Fügen, and A. Waibel, "Maximum entropy language modeling for Russian ASR," in *Proceedings of the International Workshop for Spoken Language Translation (IWSLT 2013)*, 2013.

[8] M. Schuster and K. Nakajima, "Japanese and Korean voice search," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 03 2012, pp. 5149–5152.

[9] T. Pellegrini and L. Lamel, "Automatic word decompounding for ASR in a morphologically rich language: Application to Amharic," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 5, pp. 863–873, 2009.

[10] X. Li, S. Cai, J. Pan, Y. Yan, and Y. Yang, "Large vocabulary Uyghur continuous speech recognition based on stems and suffixes," in *2010 7th International Symposium on Chinese Spoken Language Processing*, 2010, pp. 220–223.

[11] M. Creutz and K. Lagus, "Unsupervised discovery of morphemes," in *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*. Association for Computational Linguistics, Jul. 2002, pp. 21–30. [Online]. Available: https://www.aclweb.org/anthology/W02-0603

[12] A. Hagen and B. L. Pellom, "Data driven subword unit modeling for speech recognition and its application to interactive reading tutors," in *INTERSPEECH*, 2005.

[13] R. Sennrich, B. Haddow, and A. Birch, "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)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://www.aclweb.org/anthology/P16-1162

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, ser. NAACL, Minneapolis, MN, USA, 2019.

[15] A. Radford and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," in *arXiv*, 2018.

[16] C. P. Escartín, "Chasing the perfect splitter: A comparison of different compound splitting tools," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3340–3347. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/909_Paper.pdf

[17] P. Gage, "A New Algorithm for Data Compression," *C Users J.*, vol. 12, no. 2, p. 23–38, Feb. 1994.

[18] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 66–75. [Online]. Available: https://www.aclweb.org/anthology/P18-1007

[19] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[20] M. Creutz, K. Lagus, and S. Virpioja, "Unsupervised morphology induction using morfessor," in *Finite-State Methods and Natural Language Processing*, A. Yli-Jyrä, L. Karttunen, and J. Karhumäki, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 300–301.

[21] J. F. Daðason and K. Bjarnadóttir, "Kvistur: Vélræn stofnhlutagreining samsettra orða," *Orð og tunga*, vol. 17, pp. 115–132, 2015.

[22] J. F. Daðason, D. Mollberg, H. Loftsson, and K. Bjarnadóttir, "Kvistur: a BiLSTM Compound Splitter for Icelandic," in *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, 2020.

[23] J. Guðnason, O. Kjartansson, J. Jóhannsson, E. Carstensdóttir, H. H. Vilhjálmsson, H. Loftsson, S. Helgadóttir, K. M. Jóhannsdóttir, and E. Rögnvaldsson, "Almannarómur: An open Icelandic speech corpus," in *Spoken Language Technologies for Under-Resourced Languages*, 2012.

[24] S. Steingrímsson, J. Guðnason, S. Helgadóttir, and E. Rögnvaldsson, "Málrómur: A manually verified corpus of recorded Icelandic speech," in *Proceedings of the 21st Nordic Conference on Computational Linguistics*, 2017, pp. 237–240.

[25] A. B. Nikulásdóttir, I. R. Helgadóttir, M. Pétursson, and J. Guðnason, "Open ASR for Icelandic: Resources and a baseline system," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[26] A. Ali, Y. Zhang, P. Cardinal, N. Dahak, S. Vogel, and J. Glass, "A complete KALDI recipe for building Arabic speech recognition systems," *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, pp. 525–529, 04 2015.

[27] S. Ding, A. Renduchintala, and K. Duh, "A call for prudent choice of subword merge operations in neural machine translation," in *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. Dublin, Ireland: European Association for Machine Translation, Aug. 2019, pp. 204–213. [Online]. Available: https://www.aclweb.org/anthology/W19-6620

[28] M. Larson, "Sub-word-based language models for speech recognition: implications for spoken document retrieval," *Workshop on Language Modeling and Information Retrieval*, 2001.