Using sub-word n-gram models for dealing with OOV in large vocabulary speech recognition for Latvian

Askars Salimbajevs
Tilde, Vienibas gatve 75a, Riga, Latvia
askars.salimbajevs@tilde.lv

Jevgenijs Strigins
Tilde, Vienibas gatve 75a, Riga, Latvia
jevgenijs.strigins@tilde.lv

Abstract

In the Latvian language, one word can have tens or even hundreds of surface forms. This is a serious problem for large vocabulary speech recognition. Inclusion of every form in vocabulary will make it intractable, but, on the other hand, even with a vocabulary of 400K, the out-of-vocabulary (OOV) rate will be very high. In this paper, the authors investigate the possibility of using sub-word vocabularies where words are split into frequent and common parts. The results of our experiment show that this allows to significantly reduce the OOV rate.

1 Introduction

The Latvian language is a moderately inflected language, with complex nominal and verbal morphology. Latvian also has a selection of prefixes that can modify nouns, adjectives, adverbs, and verbs either in a qualitative or a spatial sense. There is no definite or indefinite article in Latvian, but definiteness can be indicated by the endings of adjectives.

Because of these properties, one word in Latvian can have tens or even hundreds (in the case of verbs) of surface forms. A successful large vocabulary speech recognition system must be able to recognize most (if not all) of these forms. This means that the vocabulary of the system must be really huge and contain about a million or more source forms. Speech recognition with such a vocabulary can be computationally intractable on most consumer hardware. On the other hand, reducing vocabulary size increases the OOV rate and significantly degrades the quality of recognition. For example, an out-of-vocabulary word is known to generate between 1.5 and 2 errors (Schwartz et al., 1994).

In this paper, the authors explore the sub-word approach, i.e., prefixes and endings, which are mostly common for all words, are split and treated as separate words. This splitting greatly reduces vocabulary size, but can introduce other problems.

There have been many efforts in using word decomposition and sub-word based language models (LM) for dealing with OOV in inflective languages such as Arabic (El-Desoky et al., 2013; Choueiter et al., 2006), Czech (Ircing et al., 2001), Estonian (Alumae, 2004), Finnish (Siivola et al., 2003), Russian (Oparin, 2008; Shin et al., 2013), Turkish (Yuret and Biçici, 2009), and Slovenian (Maučec et al., 2009). However, the authors could not find any reports on similar efforts for Latvian.

Significant improvement in the OOV rate was reported in all cases, but the changes in WER were not as dramatic. The exception was the Finnish language (Siivola et al., 2003), where an astonishing improvement from 56% to 31% was achieved.

Significant improvement was also observed for the Estonian language (Alumae, 2004); WER dropped by about 6.5% absolute with the morpheme language model (LM) and by more than 10% absolute when using the interpolated morpheme and class LM.

Different approaches for selecting sub-word units have been explored. These can be divided into two groups: (1) data-driven methods (Maučec et al., 2009; Siivola et al., 2003; Singh et al., 2002) and (2) supervised methods with some embedded language knowledge (e.g., morphological analyzers, stemmers) (Alumae, 2004; Choueiter et al., 2006; El-Desoky et at., 2013; Ircing et al., 2001; Shin et al., 2013). In this work the authors investigate methods from both groups.

2 Word n-gram language models

Word n-gram language models (LM) are probabilistic models that attempt to predict the next word based on the previous *n-1* words. To approximate the underlying language in this way, the assumption that each word depends only on the previous *n-1* words must be made. This assumption is very important, because it massively simplifies the estimation of such a model from the given data.

To estimate an n-gram language model, a large text corpus is used. For an estimated model, probabilities are calculated in the following way:

$$P\left(w_i \middle| w_{i-1}, \dots, w_{(i-n)+1}\right) = \frac{cnt(w_{(i-n)+1}, \dots, w_{i-1}, w_i)}{cnt(w_{(i-n)+1}, \dots, w_{i-1})}$$

where *cnt* is the count of given word sequences in a text corpus.

N-gram models do not recognize different inflected forms of the same word and treat them as separate words. For a closed vocabulary system, this means:

- If an inflected form is not presented in the training corpus, then it will not be recognized correctly.
- The full vocabulary of such a LM will contain about a million or more surface forms. The number of n-grams will be more than 200 million for 3-gram model. Because of high memory and computational resource requirements, speech recognition with such a LM will be too slow or even impossible on most consumer hardware. Therefore, vocabulary must be cut, and the model must be pruned, which will result in high OOV rates and increased perplexity (increased LM confusion on test data).
- Estimation of model of this size requires a huge amount of training data in order to get reliable probability estimates for all possible surface forms.

3 Sub-word n-gram language models

Sub-word based search vocabularies and language models can reduce the OOV rate of a speech recognition system by decomposing whole words into smaller units. These smaller units are selected to be common for a large number of words.

Using sub-word vocabulary requires the following steps to be taken:

- Decomposition: The original words need to be decomposed into smaller sub-word units. The units need to be common for many words, so that the new sub-word vocabulary size is clearly smaller than that of the whole word vocabulary.
- Pronunciation Generation: In this step sub-word unit pronunciations are being added to the speech recognition engine. In general, deducing the pronunciation of a sub-word unit from the pronunciation of a whole word is often challenging and even impossible in some cases. However, Latvian has a strong correspondence between written form and phoneme sequence, and this makes it possible to use a grapheme-based approach in this step.
- Language Model Training: A new language model needs to be trained for recognition of sub-word units. A model is usually trained on the same text corpus that was used for deriving the vocabulary.
- Word Reconstruction: After decoding, the recognized sub-words need to be recombined in order to obtain a valid word sequence.

3.1 Unsupervised word decomposition

One approach to decomposing words into sub-word units is to use probabilistic machine learning methods. In this paper, the authors use *Morfessor* 2.0 (Creutz and Lagus, 2005; Virpioja et al., 2013) – a family of methods for unsupervised learning of morphological segmentation. The *Morfessor* model is trained on a text corpus, and then this corpus is segmented using this model. The result is a corpus made from sub-word units, which can be used to train an n-gram language model and derive a vocabulary of noticeably smaller size (see Table 1).

Using this vocabulary, the output of the speech recognizer will be a sequence of sub-word units. In order to reconstruct the surface forms, a separate hidden-event language model (Stolcke et al., 1998) is used. This model is trained on a corpus in which the places where the word was divided are treated as hidden events and are marked using special connector tags. Applying it to a sequence of sub-word units produces a sequence of the most likely sub-word units and connector tags from which full words can be reconstructed.

3.2 Word decomposition using a stemmer

Another approach is to perform decomposition by separating stems and endings only. Forms with different prefixes will still be treated as separate words.

Decomposition is done using the Latvian stemmer developed by Pinnis and Skadiņš (2012) for their machine translation experiments. The stemmer outputs the stem for any given word. Endings can then be obtained by comparing the stem and the original word.

The Latvian stemmer can be run in two modes: (1) short mode, where only short basic endings are cut, and (2) full mode, where full endings are recognized.

In order to simplify word reconstruction, every ending is marked. After decoding, words can be reconstructed by simply concatenating stems and their marked endings.

4 Set-up and results

4.1 Data and experiments

In this work, the authors used a 22 million sentence text corpus, which was collected by crawling Latvian web news portals. The corpus is used for training the *Morfessor* model and extracting vocabularies.

Sub-word vocabularies are extracted by performing a word decomposition on the text corpus and taking the 100 thousand most frequent units. For comparison, the full vocabulary of this corpus contains approximately 1.5 million surface forms.

Also, vocabularies of the 100, 200, and 400 thousand most frequent surface forms were extracted as a baseline.

For evaluation, a small 23-minute long annotated speech corpus from 10 speakers was used.

4.2 OOV experiments

First, OOV rates for different methods were calculated on the evaluation corpus transcripts. As shown in Table 1, even with a 400K vocabulary, OOV is still very high – 7.15%. Both of the proposed methods, which use sub-word units instead of words, show a significant reduction of the OOV rate in the test corpus, while using a much smaller vocabulary.

Vocabulary containing sub-word units from *Morfessor* output almost completely solves the OOV problem, despite being the smallest among other vocabularies.

Method	**Size**	**OOV, %**
Baseline	100K	11.2 %
Baseline	200K	8.7 %
Baseline	400K	7.15 %
Stemmer (short)	100K	2.6 %
Stemmer (full)	100K	1.5%
Morfessor	76K	**<0.01 %**

Table 1: OOV rate comparison

4.3 Experiments with speech recognition

In order to evaluate the influence of sub-word vocabularies on the speech recognition task, the authors set up the following speech recognition system:

- The system uses the HMM-GMM (4000 senones and 90000 Gaussians) approach and is based on the *Kaldi* toolkit (Povey et al., 2011)
- The acoustic model is trained on a 100-hour long Latvian Speech Recognition Corpus (Pinnis et al., 2014)
- Grapheme-based pronunciations
- fMLLR speaker adaptation

For systems with sub-word vocabularies, training set transcripts were also split using the previously described models and tools and were used during the retraining of the acoustic models, so that the model is more adapted for recognizing sub-word units.

For language modeling, we used the same 22 million sentence text corpus. 3-gram models were used in all experiments, except for the *Morfessor*-

based system, where 6-gram models were also trained. All models are pruned with equal parameters. The results of the speech recognition are shown in Table 2.

Despite the reduction in the OOV rate, no significant improvement in WER has been achieved. On the contrary, the baseline system with a 200K vocabulary showed the best results, while the *Morfessor* based system showed only a very small improvement (1%) in comparison to the baseline 100K system. For systems using decomposition with a stemmer, an increase in the WER was observed.

Method	**WER, %**	
	Sub-word units	**Words**
Baseline, 100K	-	40.49 %
Baseline, 200K	-	**38.26 %**
Morfessor	38.79 %	39.43 %
Morfessor, 6-gram LM	39.33 %	39.60 %
Stemmer (short)	35.30 %	42.02 %
Stemmer (full)	**35.26 %**	43.11 %

Table 2: Word error rate comparison

5 Discussion

Intuitively, any OOV improvement should also result in improvement of recognition quality. For example, the same 200K baseline system shows about 27% WER on a subset of evaluation data with no OOV words. However, experiments performed in this work showed mostly negative results, despite big improvement in the OOV rate.

One possible reason for this is the fact that sub-word units are difficult to discriminate acoustically. For example, when using the stemmer (short mode) for decomposition, the WER for stems is 33% and around 37% for endings. For comparison, the stemmer in full mode produces shorter, more morphologically correct stems and longer endings, and, as a result, WER for stems increased to 35%. The same reason can also be applied to *Morfessor*-derived “morphemes”. This means that a more careful selection of sub-words is needed and that more morphologically correct decomposition will not guarantee a better result.

Another reason for such results can be the fact that the context of the units gets expanded after splitting words, i.e., 3-grams for sub-word units covers only some part of 3-grams for whole words and is more comparable to 2-gram word models. It can be concluded that more powerful language models are needed for sub-word vocabularies.

In order to test this hypothesis, the authors trained a 6-gram model for *Morfessor* sub-word units. However, only a tiny increase in WER has been achieved. This result is counterintuitive, and more careful analysis is needed in future work.

The authors also experimented with different pruning parameters, but classic word models still showed better results.

6 Conclusion

In this paper, the authors presented a report of the current research on the use of sub-word vocabularies for large vocabulary speech recognition for Latvian. This approach significantly reduces the OOV rate.

The authors explored two different methods: (1) fully unsupervised and data driven word decomposition using the *Morfessor* tool and (2) word decomposition using a stemmer.

Despite the fact that both methods have demonstrated significant reduction in OOV rates (almost 0% in the case of *Morfessor*), speech recognition results can be described as negative, because the best results were obtained using the baseline word based system.

In future work, the authors plan to investigate better ways for selecting acoustically distinguishable sub-word units and to explore methods that would compensate for a weaker LM when using sub-word units.

Acknowledgments

The research leading to these results has received funding from the research project “Information and Communication Technology Competence Centre” of EU Structural funds, contract nr. L-KC-11-0003 signed between the ICT Competence Centre (www.itkc.lv) and the Investment and Development Agency of Latvia, research No. 2.4 “Speech recognition technologies”.

References

Alumäe, T. (2004). Large Vocabulary Continuous Speech Recognition for Estonian Using Morphemes and Classes. In K. Sojka, Petr and Kopeček, Ivan and Pala (Ed.), *Text, Speech and Dialogue. Lecture Notes in Computer Science* (pp. 245–252). Springer Berlin Heidelberg. doi:10.1007/978-3-540-30120-2_31

Choueiter, G., Povey, D., Chen, S. F., & Zweig, G. (2006). Morpheme-Based Language Modeling for Arabic Lvcsr. *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. doi:10.1109/ICASSP.2006.1660205

Creutz, M., & Lagus, K. (2005). *Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. Publications in Computer and Information Science, Report A81, Helsinki University of Technology* (pp. 1–27).

El-Desoky Mousa, A., Kuo, H. K. J., Mangu, L., & Soltau, H. (2013). Morpheme-based feature-rich language models using Deep Neural Networks for LVCSR of Egyptian Arabic. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (pp. 8435–8439). doi:10.1109/ICASSP.2013.6639311

Ircing, P., Krbec, P., Hajic, J., Psutka, J., Khudanpur, S., Jelinek, F., & Byrne, W. (2001). On large vocabulary continuous speech recognition of highly inflectional language czech. In *European Conference on Speech Communication and Technology (EUROSPEECH)*.

Maučec, M. S., Rotovnik, T., Kačič, Z., & Brest, J. (2009). Using data-driven subword units in language model of highly inflective Slovenian language. International Journal of Pattern Recognition and Artificial Intelligence. doi:10.1142/S0218001409007119

Oparin, I. (2008). *Language models for automatic speech recognition of inflectional languages*. University of West Bohemia.

Pinnis, M., Auziņa, I., & Goba, K. (2014). Designing the Latvian Speech Recognition Corpus. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)* (pp. 1547–1553).

Pinnis, M., & Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains – What Works and What Not. In Human Language Technologies – The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012 (Vol. 247, pp. 176–184). Tartu, Estonia, Estonia: IOS Press.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Schwartz, R., Nguyen, L., Kubala, F., Chou, G., Zavaliagkos, G., & Makhoul, J. (1994). On Using Written Language Training Data for Spoken Language Modeling. In *Proceedings of the Workshop on Human Language Technology* (pp. 94–98). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1075812.1075830

Shin, E., Stüker, S., Kilgour, K., Fügen, C., & Waibel, A. (2013). Maximum Entropy Language Modeling for Russian ASR. In *Proceedings of the International Workshop for Spoken Language Translation (IWSLT 2013)*. Heidelberg.

Siivola, V., Hirsimäki, T., Creutz, M., & Kurimo, M. (2003). Unlimited Vocabulary Speech Recognition Based on Morphs Discovered in an Unsupervised Manner. In *European Conference on Speech Communication and Technology (EUROSPEECH)* (pp. 2293–2296).

Singh, R., Raj, B., & Stern, R. M. (2002). Automatic generation of subword units for speech recognition systems. *IEEE Transactions on Speech and Audio Processing*, *10*, 89–99. doi:10.1109/89.985546

Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plache, M., … Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. *ICSLP*. doi:10.1.1.53.7127

Virpioja, S., Smit, P., Grönroos, S.-A., & Kurimo, M. (2013). Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. *Aalto University Publication Series SCIENCE + TECHNOLOGY, 25/2013*.

Yuret, D., & Biçici, E. (2009). Modeling Morphologically Rich Languages Using Split Words and Unstructured Dependencies. In *ACL-IJCNLP 2009*.