Vocabulary Decomposition for Estonian Open Vocabulary Speech Recognition

Antti Puurula and Mikko Kurimo
Adaptive Informatics Research Centre
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Finland
{puurula, mikkok}@cis.hut.fi

Abstract

Speech recognition in many morphologically rich languages suffers from a very high out-of-vocabulary (OOV) ratio. Earlier work has shown that vocabulary decomposition methods can practically solve this problem for a subset of these languages. This paper compares various vocabulary decomposition approaches to open vocabulary speech recognition, using Estonian speech recognition as a benchmark. Comparisons are performed utilizing large models of 60000 lexical items and smaller vocabularies of 5000 items. A large vocabulary model based on a manually constructed morphological tagger is shown to give the lowest word error rate, while the unsupervised morphology discovery method Morfessor Baseline gives marginally weaker results. Only the Morfessor-based approach is shown to adequately scale to smaller vocabulary sizes.

1 Introduction

1.1 OOV problem

Open vocabulary speech recognition refers to automatic speech recognition (ASR) of continuous speech, or "speech-to-text" of spoken language, where the recognizer is expected to recognize any word spoken in that language. This capability is a recent development in ASR, and is required or beneficial in many of the current applications of ASR technology. Moreover, large vocabulary speech recognition is not possible in most languages of the world without first developing the tools needed for open vocabulary speech recognition. This is due to a fundamental obstacle in current ASR called the out-of-vocabulary (OOV) problem.

The OOV problem refers to the existence of words encountered that a speech recognizer is unable to recognize, as they are not covered in the vocabulary. The OOV problem is caused by three intertwined issues. Firstly, the language model training data and the test data always come from different samplings of the language, and the mismatch between test and training data introduces some OOV words, the amount depending on the difference between the data sets. Secondly, ASR systems always use finite and preferably small sized vocabularies, since the speed of decoding rapidly slows down as the vocabulary size is increased. Vocabulary sizes depend on the application domain, sizes larger than 60000 being very rare. As some of the words encountered in the training data are left out of the vocabulary, there will be OOV words during recognition. The third and final issue is the fundamental one; languages form novel sentences not only by combining words, but also by combining sub-word items called morphs to make up the words themselves. These morphs in turn correspond to abstract grammatical items called morphemes, and morphs of the same morpheme are called allomorphs of that morpheme. The study of these facets of language is aptly called morphology, and has been largely neglected in modern ASR technology. This is due to

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 89–95,
Prague, Czech Republic, June 2007.

ASR having been developed primarily for English, where the OOV problem is not as severe as in other languages of the world.

1.2 Relevance of morphology for ASR

Morphologies in natural languages are characterized typologically using two parameters, called indexes of synthesis and fusion. Index of synthesis has been loosely defined as the ratio of morphs per word forms in the language(Comrie, 1989), while index of fusion refers to the ratio of morphs per morpheme. High frequency of verb paradigms such as "hear, hear + d, hear + d" would result in a high synthesis, low fusion language, whereas high frequency of paradigms such as "sing, sang, sung" would result in almost the opposite. Counting distinct item types and not instances of the types, the first example would have 2 word forms, 2 morphs and 2 morphemes, the second 3 word forms, 3 morphs and 1 morpheme. Although in the first example, there are 3 word instances of the 2 word forms, the latter word form being an ambiguous one referring to two distinct grammatical constructions. It should also be noted that the first morph of the first example has 2 pronunciations. Pronunciational boundaries do not always follow morphological ones, and a morph may and will have several pronunciations that depend on context, if the language in question has significant orthographic irregularity.

As can be seen, both types of morphological complexity increase the amount of distinct word forms, resulting in an increase in the OOV rate of any finite sized vocabulary for that language. In practice, the OOV increase caused by synthesis is much larger, as languages can have thousands of different word forms per word that are caused by addition of processes of word formation followed by inflections. Thus the OOV problem in ASR has been most pronounced in languages with much synthesis, regardless of the amount of fusion. The morpheme-based modeling approaches evaluated in this work are primarily intended for fixing the problem caused by synthesis, and should work less well or even adversely when attempted with low synthesis, high fusion languages. It should be noted that models based on finite state transducers have been shown to be adequate for describing fusion as well(Koskenniemi, 1983), and further work should evaluate these types of models in ASR of languages with higher indexes of fusion.

1.3 Approaches for solving the OOV problem

The traditional method for reducing OOV would be to simply increase the vocabulary size so that the rate of OOV words becomes sufficiently low. Naturally this method fails when the words are derived, compounded or inflected forms of rarer words. While this approach might still be practical in languages with a low index of synthesis such as English, it fails with most languages in the world. For example, in English with language models (LM) of 60k words trained from the Gigaword Corpus V.2(Graff et al., 2005), and testing on a very similar Voice of America -portion of TDT4 speech corpora(Kong and Graff, 2005), this gives a OOV rate of 1.5%. It should be noted that every OOV causes roughly two errors in recognition, and vocabulary decomposition approaches such as the ones evaluated here give some benefits to word error rate (WER) even in recognizing languages such as English(Bisani and Ney, 2005).

Four different approaches to lexical unit selection are evaluated in this work, all of which have been presented previously. These are hence called "word", "hybrid", "morph" and "grammar". The word approach is the default approach to lexical item selection, and is provided here as a baseline for the alternative approaches. The alternatives tested here are all based on decomposing the in-vocabulary words, OOV words, or both, in LM training data into sequences of sub-word fragments. During recognition the decoder can then construct the OOV words encountered as combinations of these fragments. Word boundaries are marked in LMs with tokens so that the words can be reconstructed from the sub-word fragments after decoding simply by removing spaces between fragments, and changing the word boundaries tokens to spaces. As splitting to sub-word items makes the span of LM histories shorter, higher order n-grams must be used to correct this. Varigrams(Siivola and Pellom, 2005) are used in this work, and to make LMs trained with each approach comparable, the varigrams have been grown to roughly sizes of 5 million counts. It should be noted that the names for the approaches here are somewhat arbitrary, as from a theoretical perspective both morph- and grammar-based approaches try to model the grammatical morph set of a language, difference being that “morph” does this with an unsupervised data-driven machine learning algorithm, whereas “grammar” does this using segmentations from a manually constructed rule-based morphological tagger.

2 Modeling approaches

2.1 Word approach

The first approach evaluated in this work is the traditional word based LM, where items are simply the most frequent words in the language model training data. OOV words are simply treated as unknown words in language model training. This has been the default approach to selection of lexical items in speech recognition for several decades, and as it has been sufficient in English ASR, there has been limited interest in any alternatives.

2.2 Hybrid approach

The second approach is a recent refinement of the traditional word-based approach. This is similar to what was introduced as “flat hybrid model”(Bisani and Ney, 2005), and it tries to model OOV-words as sequences of words and fragments. “Hybrid” refers to the LM histories being composed of hybrids of words and fragments, while “flat” refers to the model being composed of one n-gram model instead of several models for the different item types. The models tested in this work differ in that since Estonian has a very regular phonemic orthography, grapheme sequences can be directly used instead of more complex pronunciation modeling. Subsequently the fragments used are just one grapheme in length.

2.3 Morph approach

The morph-based approach has shown superior results to word-based models in languages of high synthesis and low fusion, including Estonian. This approach, called “Morfessor Baseline” is described in detail in (Creutz et al., 2007). An unsupervised machine learning algorithm is used to discover the morph set of the language in question, using minimum description length (MDL) as an optimization criterion. The algorithm is given a word list of the language, usually pruned to about 100 000 words, that it proceeds to recursively split to smaller items, using gains in MDL to optimize the item set. The resulting set of morphs models the morph set well in languages of high synthesis, but as it does not take fusion into account any manner, it should not work in languages of high fusion. It neither preserves information about pronunciations, and as these do not follow morph boundaries, the approach is unsuitable in its basic form to languages of high orthographic irregularity.

2.4 Grammar approach

The final approach applies a manually constructed rule-based morphological tagger(Alumäe, 2006). This approach is expected to give the best results, as the tagger should give the ideal segmentation along the grammatical morphs that the unsupervised and language-independent morph approach tries to find. To make this approach more comparable to the morph models, OOV morphs are modeled as sequences of graphemes similar to the hybrid approach. Small changes to the original approach were also made to make the model comparable to the other models presented here, such as using the tagger segmentations as such and not using pseudo-morphemes, as well as not tagging the items in any manner. This approach suffers from the same handicaps as the morph approach, as well as from some additional ones: morphological analyzers are not readily available for most languages, they must be tailored by linguists for new datasets, and it is an open problem as to how pronunciation dictionaries should be written for grammatical morphs in languages with significant orthographic irregularity.

2.5 Text segmentation and language modeling

For training the LMs, a subset of 43 million words from the Estonian Segakorpus was used(Segakorpus, 2005), preprocessed with a morphological analyzer(Alumäe, 2006). After selecting the item types, segmenting the training corpora and generation of a pronunciation dictionary, LMs were trained for each lexical item type. Table 1 shows the text format for LM training data after segmentation with each model. As can be seen, the word-based approach doesn’t use word boundary tokens. To keep the LMs comparable between each model-

model	text segmentation
word 5k	voodis reeglina loeme
word 60k	voodis reeglina loeme
hybrid 5k	v o o d i s <w> reeglina <w> l o e m e
hybrid 60k	voodis <w> reeglina <w> loeme
morph 5k	voodi s <w> re e g lina <w> loe me
morph 60k	voodi s <w> reegli na <w> loe me
grammar 5k	voodi s <w> reegli na <w> loe me
grammar 60k	voodi s <w> reegli na <w> loe me

Table 1. Sample segmented texts for each model.

ing approach, growing varigram models(Siivola and Pellom, 2005) were used with no limits as to the order of n-grams, but limiting the number of counts to 4.8 and 5 million counts. In some models this growing method resulted in the inclusion of very frequent long item sequences to the varigram, up to a 28-gram. Models of both 5000 and 60000 lexical items were trained in order to test if and how the modeling approaches would scale to smaller and therefore much faster vocabularies. Distribution of counts in n-gram orders can be seen in figure 1.

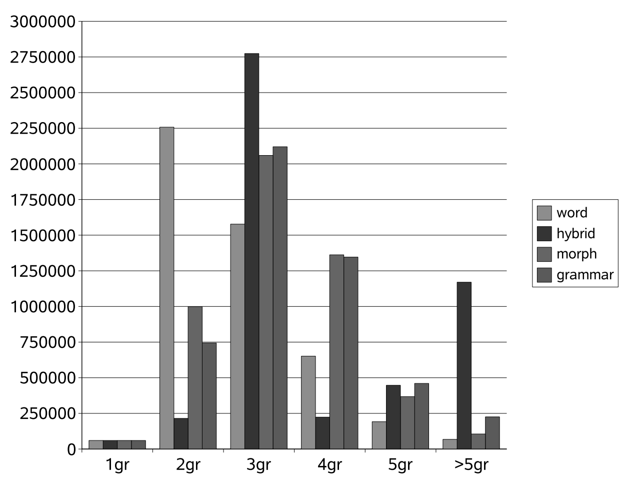

Figure 1. Number of counts included for each n-gram order in the 60k varigram models.

The performance of the statistical language models is often evaluated by perplexity or cross-entropy. However, we decided to only report the real ASR performance, because perplexity does not suit well to the comparison of models that use different lexica, have different OOV rates and have lexical units of different lengths.

3 Experimental setup

3.1 Evaluation set

Acoustic models for Estonian ASR were trained on the Estonian Speechdat-like corpus(Meister et al., 2002). This consists of spoken newspaper sentences and shorter utterances, read over a telephone by 1332 different speakers. The data therefore was quite clearly articulated, but suffered from 8kHz sample rate, different microphones, channel noises and occasional background noises. On top of this the speakers were selected to give a very broad coverage of different dialectal varieties of Estonian and were of different age groups. For these reasons, in spite of consisting of relatively common word forms from newspaper sentences, the database can be considered challenging for ASR.

Held-out sentences were from the same corpus used as development and evaluation set. 8 different sentences from 50 speakers each were used for evaluation, while sentences from 15 speakers were used for development. LM scaling factor was optimized for each model separately on the development set. On total over 200 hours of data from the database was used for acoustic model training, of which less than half was speech.

3.2 Decoding

The acoustic models were Hidden Markov Models (HMM) with Gaussian Mixture Models (GMM) for state modeling based on 39-dimensional MFCC+P+D+DD features, with windowed cepstral mean subtraction (CMS) of 1.25 second window. Maximum likelihood linear transformation (MLLT) was used during training. State-tied cross-word triphones and 3 left-to-right states were used, state durations were modeled using gamma distributions. On total 3103 tied states and 16 Gaussians per state were used.

Decoding was done with the decoder developed at TKK(Pylkkönen, 2005), which is based on a one-pass Viterbi beam search with token passing on a lexical prefix tree. The lexical prefix tree included a cross-word network for modeling triphone contexts, and the nodes in the prefix tree were tied at the triphone state level. Bigram look-ahead models were used in speeding up decoding, in addition to pruning with global beam, history, histogram and word end pruning. Due to the properties of the decoder and varigram models, very high order n-grams could be used without significant degradation in decoding speed.

As the decoder was run with only one pass, adaptation was not used in this work. In preliminary experiments simple adaptation with just constrained maximum likelihood linear regression (CMLLR) was shown to give as much as 20 % relative word error rate reductions (RWERR) with this dataset. Adaptation was not used, since it interacts with the model types, as well as with the WER from the first round of decoding, providing larger RWERR for the better models. With high WER models, adaptation matrices are less accurate, and it is also probable that the decomposition methods yield more accurate matrices, as they produce results where fewer HMM-states are misrecognized. These issues should be investigated in future research.

After decoding, the results were post-processed by removing words that seemed to be sequences of junk fragments: consonant-only sequences and 1-phoneme words. This treatment should give very significant improvements with noisy data, but in preliminary experiments it was noted that the use of sentence boundaries resulted in almost 10% RWERR weaker results for the approaches using fragments, as that almost negates the gains achieved from this post-processing. Since sentence boundary forcing is done prior to junk removal, it seems to work erroneously when it is forced to operate on noisy data. Sentence boundaries were nevertheless used, as in the same experiments the word-based models gained significantly from their use, most likely because they cannot use the fragment items for detection of acoustic junk, as the models with fragments can.

4 Results

Results of the experiments were consistent with earlier findings(Hirsimäki et al., 2006; Kurimo et al., 2006). Traditional word based LMs showed the worst performance, with all of the recently proposed alternatives giving better results. Hybrid LMs consistently outperformed traditional word-based LMs in both large and small vocabulary conditions. The two morphology-driven approaches gave similar and clearly superior results. Only the morph approach seems to scale down well to smaller vocabulary sizes, as the WER for the grammar approach increased rapidly as size of the vocabulary was decreased.

size	word	hybrid	morph	grammar
60000	53.1	47.1	39.4	38.7
5000	82.0	63.0	43.5	47.6

Table 2. Word error rates for the models (WER %).

Table 2 shows the WER for the large (60000) and small (5000) vocabulary sizes and different modeling approaches. Table 3 shows the corresponding letter error rates (LER). LERs are more comparable across some languages than WERs, as WER depends more on factors such as length, morphological complexity, and OOV of the words. However, for within-language and between-model comparisons, the RWERR should still be a valid metric, and is also usable in languages that do not use a phonemic writing system. The RWERRs of different novel methods seems to be comparable between different languages as well. Both WER and LER are high considering the task. However, standard methods such as adaptation were not used, as the intention was only to study the RWERR of the different approaches.

size	word	hybrid	morph	grammar
60000	17.8	15.8	12.4	12.3
5000	35.5	20.8	14.4	15.4

Table 3. Letter error rates for the models (LER %).

5 Discussion

Four different approaches to lexical item selection for large and open vocabulary ASR in Estonian were evaluated. It was shown that the three approaches utilizing vocabulary decomposition give substantial improvements over the traditional word based approach, and make large vocabulary ASR technology possible for languages similar to Estonian, where the traditional approach fails due to very high OOV rates. These include memetic relatives Finnish and Turkish, among other languages that have morphologies of high fusion, low synthesis and low orthographic irregularity.

5.1 Performance of the approaches

The morpheme-based approaches outperformed the word- and hybrid-based approaches clearly. The results for "hybrid" are in in the range suggested by earlier work(Bisani and Ney, 2005). One possible explanation for the discrepancy between the hybrid and morpheme-based approaches would be that the morpheme-based approaches capture items that make sense in n-gram modeling, as morphs are items that the system of language naturally operates on. These items would then be of more use when trying to predict unseen data(Creutz et al., 2007). As modeling pronunciations is much more straightforward in Estonian, the morpheme-based approaches do not suffer from erroneous pronunciations, resulting in clearly superior performance.

As for the superiority of the "grammar" over the unsupervised "morph", the difference is marginal in terms of RWERR. The grammatical tagger was tailored by hand for that particular language, whereas Morfessor method is meant to be unsupervised and language independent. There are further arguments that would suggest that the unsupervised approach is one that should be followed; only "morph" scaled well to smaller vocabulary sizes, the usual practice of pruning the word list to produce smaller morph sets gives better results than here and most importantly, it is questionable if "grammar" can be taken to languages with high indexes of fusion and orthographic irregularity, as the models have to take these into account as well.

5.2 Comparison to previous results

There are few previous results published on Estonian open vocabulary ASR. In (Alumäe, 2006) a WER of 44.5% was obtained with word-based trigrams and a WER of 37.2% with items similar to ones from "grammar" using the same speech corpus as in this work. Compared to the present work, the WER for the morpheme-based models was measured with compound words split in both hypothesis and reference texts, making the task slightly easier than here. In (Kurimo et al., 2006) a WER of 57.6% was achieved with word-based varigrams and a WER of 49.0% with morphs-based ones. This used the same evaluation set as this work, but had slightly different LMs and different acoustic modelling which is the main reason for the higher WER levels. In summary, morpheme-based approaches seem to consistently outperform the traditional word based one in Estonian ASR, regardless of the specifics of the recognition system, test set and models.

In (Hirsimäki et al., 2006) a corresponding comparison of unsupervised and grammar-based morphs was presented in Finnish, and the grammar-based model gave a significantly higher WER in one of the tasks. This result is interesting, and may stem from a number of factors, among them the different decoder and acoustic models, 4-grams versus varigrams, as well as differences in post-processing. Most likely the difference is due to lack of coverage for domain-specific words in the Finnish tagger, as it has a 4.2% OOV rate on the training data. On top of this the OOV words are modeled simply as grapheme sequences, instead of modeling only OOV morphs in that manner, as is done in this work.

5.3 Open problems in vocabulary decomposition

As stated in the introduction, modeling languages with high indexes of fusion such as Arabic will require more complex vocabulary decomposition approaches. This is verified by recent empirical results, where gains obtained from simple morphological decomposition seem to be marginal(Kirchhoff et al., 2006; Creutz et al., 2007). These languages would possibly need novel LM inference algorithms and decoder architectures. Current research seems to be heading in this direction, with weighted finite state transducers becoming standard representations for the vocabulary instead of the lexical prefix tree.

Another issue in vocabulary decomposition is orthographic irregularity, as the items resulting from decomposition do not necessarily have unambiguous pronunciations. As most modern recognizers use the Viterbi approximation with vocabularies of one pronunciation per item, this is problematic. One solution to this is expanding the different items with tags according to pronunciation, shifting the problem to language modeling(Creutz et al., 2007). For example, English plural "s" would expand to "s#1" with pronunciation "/s/", and "s#2" with pronunciation "/z/", and so on. In this case the vocabulary size increases by the amount of different pronunciations added. The new items will have pronunciations that depend on their language model context, enabling the prediction of pronunciations with language model probabilities. The only downside to this is complicating the search for optimal vocabulary decomposition, as the items should make sense in both pronunciational and morphological terms.

One can consider the originally presented hybrid approach as an approach to vocabulary decomposition that tries to keep the pronunciations of the items as good as possible, whereas the morph approach tries to find items that make sense in terms of morphology. This is obviously due to the methods having been developed on very different types of languages. The morph approach was developed for the needs of Finnish speech recognition, which is a high synthesis, moderate fusion and very low orthographic irregularity language, whereas the hybrid approach in (Bisani and Ney, 2005) was developed for English, which has low synthesis, moderate fusion, and very high orthographic irregularity. A universal approach to vocabulary decomposition would have to take all of these factors into account.

Acknowledgements

The authors would like to thank Dr. Tanel Alumäe from Tallinn University of Technology for help in performing experiments with Estonian speech and text databases. This work was supported by the Academy of Finland in the project: *New adaptive and learning methods in speech recognition*.

References

Bernard Comrie. 1972. *Language Universals and Linguistic Typology*, Second Edition. Athenæum Press Ltd, Gateshead, UK.

Kimmo Koskenniemi. 1983. *Two-level Morphology: a General Computational Model for Word-Form Recognition and Production*. University of Helsinki, Helsinki, Finland.

Tanel Alumäe. 2006. Methods for Estonian Large Vocabulary Speech Recognition. *PhD Thesis.* Tallinn University of Technology. Tallinn, Estonia.

Maximilian Bisani, Hermann Ney. 2005. Open Vocabulary Speech Recognition with Flat Hybrid Models. *INTERSPEECH-2005*, 725–728.

Janne Pylkkönen. 2005. An Efficient One-pass Decoder for Finnish Large Vocabulary Continuous Speech Recognition. *Proceedings of The 2nd Baltic Conference on Human Language Technologies*, 167–172. HLT'2005. Tallinn, Estonia.

Vesa Siivola, Bryan L. Pellom. 2005. Growing an n-Gram Language Model. *INTERSPEECH-2005*, 1309–1312.

David Graff, Junbo Kong, Ke Chen and Kazuaki Maeda. 2005. LDC Gigaword Corpora: English Gigaword Second Edition. In *LDC link: http://www.ldc.upenn.edu/Catalog/index.jsp*.

Junbo Kong and David Graff. 2005. TDT4 Multilingual Broadcast News Speech Corpus. In *LDC link: http://www.ldc.upenn.edu/Catalog/index.jsp*.

Segakorpus. 2005. Segakorpus - Mixed Corpus of Estonian. Tartu University. *http://test.cl.ut.ee/korpused/*.

Einar Meister, Jürgen Lasn and Lya Meister 2002. Estonian SpeechDat: a project in progress. In *Proceedings of the Fonetiikan Päivät - Phonetics Symposium 2002 in Finland*, 21–26.

Katrin Kirchhoff, Dimitra Vergyri, Jeff Bilmes, Kevin Duh and Andreas Stolcke 2006. Morphology-based language modeling for conversational Arabic speech recognition. *Computer Speech & Language* 20(4):589–608.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraclar and Andreas Stolcke 2007. Analysis of Morph-Based Speech Recognition and the Modeling of Out-of-Vocabulary Words Across Languages To appear in *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL-HLT 2007, Rochester, NY, USA

Mikko Kurimo, Antti Puurula, Ebru Arisoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumae and Murat Saraclar 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics*. HLT-NAACL 2006. New York, USA

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja and Janne Pylkkönen 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language* 20(4):515–541.