10

Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages

Ebru Arısoy[1], Mikko Kurimo[2], Murat Saraçlar[1], Teemu Hirsimäki[2], Janne Pylkkönen[2], Tanel Alumäe[3] and Haşim Sak[1]
[1]*Boğaziçi University,*
[2]*Helsinki University of Technology,*
[3]*Tallinn University of Technology,*
[1]*Turkey*
[2]*Finland*
[3]*Estonia*

1. Introduction

Automatic Speech Recognition (ASR) systems utilize statistical acoustic and language models to find the most probable word sequence when the speech signal is given. Hidden Markov Models (HMMs) are used as acoustic models and language model probabilities are approximated using *n*-grams where the probability of a word is conditioned on *n-1* previous words. The *n*-gram probabilities are estimated by Maximum Likelihood Estimation. One of the problems in *n*-gram language modeling is the data sparseness that results in non-robust probability estimates especially for rare and unseen *n*-grams. Therefore, smoothing is applied to produce better estimates for these *n*-grams.

The traditional *n*-gram word language models are commonly used in state-of-the-art Large Vocabulary Continuous Speech Recognition (LVSCR) systems. These systems result in reasonable recognition performances for languages such as English and French. For instance, broadcast news (BN) in English can now be recognized with about ten percent word error rate (WER) (NIST, 2000) which results in mostly quite understandable text. Some rare and new words may be missing in the vocabulary but the result has proven to be sufficient for many important applications, such as browsing and retrieval of recorded speech and information retrieval from the speech (Garofolo et al., 2000). However, LVCSR attempts with similar systems in agglutinative languages, such as Finnish, Estonian, Hungarian and Turkish so far have not resulted in comparable performance to the English systems. The main reason of this performance deterioration in those languages is their rich morphological structure. In agglutinative languages, words are formed mainly by concatenation of several suffixes to the roots and together with compounding and inflections this leads to millions of different, but still frequent word forms. Therefore, it is practically impossible to build a word-based vocabulary for speech recognition in agglutinative languages that would cover all the relevant words. If words are used as language modeling units, there will be many out-of-vocabulary (OOV) words due to using limited vocabulary sizes in ASR systems. It was shown that with an optimized 60K lexicon

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

the OOV rate is less than 1% for North American Business news (Rosenfeld, 1995). Highly inflectional and agglutinative languages suffer from high number of OOV words with similar size vocabularies. In our Turkish BN transcription system, the OOV rate is 9.3% for a 50K lexicon. For other agglutinative languages like Finnish and Estonian, OOV rates are around 15% for a 69K lexicon (Hirsimäki et al., 2006) and 10% for a 60K lexicon respectively and 8.27% for Czech, a highly inflectional language, with a 60K lexicon (Podvesky & Machek, 2005). As a rule of thumb an OOV word brings up on average 1.5 recognition errors (Hetherington, 1995). Therefore solving the OOV problem is crucial for obtaining better accuracies in the ASR of agglutinative languages. OOV rate can be decreased to an extent by increasing the vocabulary size. However, even doubling the vocabulary is not a sufficient solution, because a vocabulary twice as large (120K) would only reduce the OOV rate to 6% in Estonian and 4.6% in Turkish. In Finnish even a 500K vocabulary of the most common words still gives 5.4% OOV in the language model training material. In addition, huge lexicon sizes may result in confusion of acoustically similar words and require a huge amount of text data for robust language model estimates. Therefore, sub-words are proposed as language modeling units to alleviate the OOV and data sparseness problems that plague systems based on word-based recognition units in agglutinative languages.

In sub-word-based ASR; (i) words are decomposed into meaningful units in terms of speech recognition, (ii) these units are used as vocabulary items in *n*-gram language models, (iii) decoding is performed with these *n*-gram models and sub-word sequences are obtained, (iv) word-like units are generated from sub-word sequences as the final ASR output.

In this chapter, we mainly focus on the decomposition of words into sub-words for LVCSR of agglutinative languages. Due to inflections, ambiguity and other phenomena, it is not trivial to automatically split the words into meaningful parts. Therefore, this splitting can be performed by using rule-based morphological analyzers or by some statistical techniques. The sub-words learned with morphological analyzers and statistical techniques are called grammatical and statistical sub-words respectively. Morphemes and stem-endings can be used as the grammatical sub-words. The statistical sub-word approach presented in this chapter relies on a data-driven algorithm called Morfessor Baseline (Creutz & Lagus, 2002; Creutz & Lagus, 2005) which is a language independent unsupervised machine learning method to find morpheme-like units (called *statistical morphs*) from a large text corpus.

After generating the sub-word units, *n*-gram models are trained with sub-words similarly as if the language modeling units were words. In order to facilitate converting sub-word sequences into word sequences after decoding, word break symbols can be added as additional units or special markers can be attached to non-initial sub-words in language modeling. ASR systems that successfully utilize the n-gram language models trained for sub-word units are used in the decoding task. Finally, word-like ASR output is obtained from sub-word sequences by concatenating the sub-words between consecutive word breaks or by gluing marked non-initial sub-words to initial ones. The performance of words and sub-words are evaluated for three agglutinative languages, Finnish, Estonian and Turkish.

This chapter is organized as follow: In Section 2, our statistical language modeling approaches are explained in detail. Section 3 contains the experimental setup for each language. Experimental results are given in Section 4. Finally, this chapter is concluded with a detailed comparison of the proposed approaches for agglutinative languages.

2. Statistical language modeling approaches

The morphological productivity of agglutinative languages makes it difficult to construct robust and effective word-based language models. With a dictionary size of a few hundred thousand words, we can still have OOV words, which are constructed through legal morphological rules. Therefore, in addition to words, sub-word units are utilized in LVCSR tasks for Finnish, Estonian and Turkish. Fig. 1 shows a phrase in each language segmented into proposed grammatical and statistical sub-word units. The details of these units will be explained thoroughly in this section.

Finnish example: Words: pekingissä vieraileville suomalaisille kansanedustajille
Grammatical sub-words:
Morphemes: pekingi ssä # vieraile v i lle # suomalais i lle # kansa n edusta j i lle
Statistical sub-words:
Morphs: peking issä # vieraile ville # suomalaisille # kansanedustaj ille

Estonian example: Words: teede ja sideministeerium on teinud ettepaneku
Grammatical sub-words:
Morphemes: tee de # ja # side ministeerium # on # tei nud # ette paneku
Statistical sub-words:
Morphs: teede # ja # sideministeerium # on # te i nud # ettepaneku

Turkish example: Words: tüketici derneklerinin öncülüğünde
Grammatical sub-words:
Morphemes: tüketici # dernek leri nin # öncü lüğ ü nde
Stem-endings: tüketici # dernek lerinin # öncü lüğünde
Statistical sub-words:
Morphs: tüketici # dernek lerinin # öncü lüğü nde

Fig. 1. Finnish, Estonian and Turkish phrases segmented into statistical and grammatical sub-words

2.1 Word-based model

Using words as recognition units is a classical approach employed in most state-of-the-art recognition systems. The word model has the advantage of having longer recognition units which results in better acoustic discrimination among vocabulary items. However the vocabulary growth for words is almost unlimited for agglutinative languages and this leads to high number of OOV words with moderate size vocabularies in ASR systems. It has been reported that the same size text corpora (40M words) result in less than 200K word types for English and 1.8M and 1.5M word types for Finnish and Estonian respectively (Creutz et al., 2007a). The number of word types is 735K for the same size Turkish corpus.

2.2 Sub-word-based models

Large number of OOV words and data sparseness are the main drawbacks of the word-based language modeling units in ASR of agglutinative and highly inflectional languages. Therefore, several sub-word units were explored for those languages to handle these drawbacks. Naturally, there are many ways to split the words into smaller units to reduce a lexicon to a tractable size. However, for a sub-word lexicon suitable for language modeling applications such as speech recognition, several properties are desirable:

i. The size of the lexicon should be small enough that the n-gram modeling becomes more feasible than the conventional word based modeling.
ii. The coverage of the target language by words that can be built by concatenating the units should be high enough to avoid the OOV problem.
iii. The units should be somehow meaningful, so that the previously observed units can help in predicting the next one.
iv. For speech recognition one should be able to determine the pronunciation for each unit.

A common approach to find the sub-word units is to program the language-dependent grammatical rules into a morphological analyzer and utilize it to split the text corpus into morphemes. As an alternative approach, sub-word units that meet the above desirable properties can be learned with unsupervised machine learning algorithms. In this section, we investigated both of the approaches.

2.2.1 Grammatical sub-words; morphemes and stem-endings

Using morphemes and stem-endings as recognition units is becoming a common approach in statistical language modeling of morphologically rich languages. Morphemes were utilized as language modeling units in agglutinative languages such as Finnish (Hirsimäki et al., 2006), Estonian (Alumäe, 2005) and Turkish (Hacioglu et al., 2003) as well as in Czech (Byrne et al., 2001) which is a highly inflectional language. Merged morphemes were proposed instead of word phrases for Korean (Kwon and Park, 2003). In Kanevsky and Roukos (1998) stem-ending based modeling was proposed for agglutinative languages and it is used in ASR of Turkish with both surface form (Mengüşoğlu & Deroo, 2001; Bayer et al., 2006) and lexical form (Arısoy et al., 2007) representations of endings. In addition, a unified model using both words, stem-endings and morphemes was proposed for Turkish (Arısoy et al., 2006).

A morphological analyzer is required to obtain morphemes, stems and endings. However, due to the handcrafted rules, morphological analyzers may suffer from an OOV problem, since in addition to morphotactic and morphophonemic rules, a limited root vocabulary is also compiled in the morphological analyzer. For instance, a Turkish morphological parser (Sak et al., 2008) with 54,267 roots can analyze 96.7% of the word tokens and 52.2% of the word types in a text corpus of 212M words with 2.2M unique words. An example output from this parser for Turkish word *alın* is given in Fig. 2. The English glosses are given in parenthesis for convenience. The inflectional morphemes start with a + sign and the derivational morphemes start with a - sign. Part-of-speech tags are attached to roots in brackets and lexical morphemes are followed by nominal and verbal morphological features in brackets. As was shown in Fig. 2, the morphological parsing of a word may result in multiple interpretations of that word due to complex morphology. This ambiguity can be resolved using morphological disambiguation tools for Turkish (Sak et al., 2007).

alın[Noun]+[A3sg]+[Pnon]+[Nom] *(forehead)*
al[Noun]+[A3sg]+Hn[P2sg]+[Nom] *(your red)*
al[Adj]-[Noun]+[A3sg]+Hn[P2sg]+[Nom] *(your red)*
al[Noun]+[A3sg]+[Pnon]+NHn[Gen] *(of red)*
al[Adj]-[Noun]+[A3sg]+[Pnon]+NHn[Gen] *(of red)*
alın[Verb]+[Pos]+[Imp]+[A2sg] *((you) be offended)*
al[Verb]+[Pos]+[Imp]+YHn[A2pl] *((you) take)*
al[Verb]-Hn[Verb+Pass]+[Pos]+[Imp]+[A2sg] *((you) be taken)*

Fig. 2. Output of the Turkish morphological parser (Sak et al., 2008) with English glosses.

To obtain a morpheme-based language model, all the words in the training text corpus are decomposed into their morphemes using a morphological analyzer. Then a morphological disambiguation tool is required to choose the correct analysis among all the possible candidates using the given context. In Arısoy et al. (2007) the parse with the minimum number of morphemes is chosen as the correct parse since the output of the morphological parser used in the experiments was not compatible with the available disambiguation tools. Also, a morphophonemic transducer is required to obtain the surface form representations of the morphemes if the morphological parser output is in the lexical form as in Fig. 2.
In statistical language modeling, there is a trade-off between using short and long units. When grammatical morphemes are used for language modeling, there can be some problems related to the pronunciations of very short inflection-type units. Stem-endings are a compromise between words and morphemes. They provide better OOV rate than words, and they lead to more robust language models than morphemes which require longer *n*-grams. The stems and endings are also obtained from the morphological analyzer. Endings are generated by concatenating the consecutive morphemes.
Even though morphemes and stem-endings are logical sub-word choices in ASR, they require some language dependent tools such as morphological analyzers and disambiguators. The lack of successful morphological disambiguation tools may result in ambiguous splits and the limited root vocabulary compiled in the morphological parsers may result in poor coverage, especially for many names and foreign words which mostly occur in news texts.
One way to extend the rule-based grammatical morpheme analysis to new words that inevitably occur in large corpora, is to split the words using a similar maximum likelihood word segmentation by Viterbi search as in the unsupervised word segmentation (statistical morphs in section 2.2.2), but here using the lexicon of grammatical morphs. This drops the OOV rate significantly and helps to choose the segmentation using the most common units where alternative morphological segmentations are available.

2.2.2 Statistical sub-words; morphs

Statistical morphs are morpheme-like units obtained by a data driven approach based on the Minimum Description Length (MDL) principle which learns a sub-word lexicon in an unsupervised manner from a training lexicon of words (Creutz & Lagus, 2005). The main idea is to find an optimal encoding of the data with a concise lexicon and a concise representation of the corpus.
In this chapter, we have adopted a similar approach as Hirsimäki et al. (2006). The Morfessor Baseline algorithm (Creutz & Lagus, 2005) is used to automatically segment the word types seen in the training text corpus. In the Morfessor Baseline algorithm the minimized cost is the coding length of the lexicon and the words in the corpus represented by the units of the lexicon. This MDL based cost function is especially appealing, because it tends to give units that are both as frequent and as long as possible to suit well for both training the language models and also decoding of the speech. Full coverage of the language is also guaranteed by splitting the rare words into very short units, even to single phonemes if necessary. For language models utilized in speech recognition, the lexicon of the statistical morphs can be further reduced by omitting the rare words from the input of the Morfessor Baseline algorithm. This operation does not reduce the coverage of the lexicon, because it just splits the rare words then into smaller units, but the smaller lexicon may offer a remarkable speed up of the recognition. The pronunciation of, especially, the short units may be ambiguous and may cause severe problems in languages like English, in which the

pronunciations can not be adequately determined from the orthography. In most agglutinative languages, such as Finnish, Estonian and Turkish, rather simple letter-to-phoneme rules are, however, sufficient for most cases.

The steps in the process of estimating a language model based on statistical morphs from a text corpus is shown in Fig. 3. First word types are extracted from a text corpus. Rare words are removed from the word types by setting a frequency cut-off. Elimination of the rare words is required to reduce the morph lexicon size. Then the remaining word types are passed through a word splitting transformation. Based on the learned morph lexicon, the best split for each word is determined by performing a Viterbi search using within-word n-gram probabilities of the units. At this point the word break symbols, # (See Fig. 1), are added between each word in order to incorporate that information in the statistical language models, as well. We prefer to use additional word break symbols in morph-based language modeling since unlike stems, a statistical morph can occur at any position in a word and marking the non-initial morphs increases the vocabulary size.

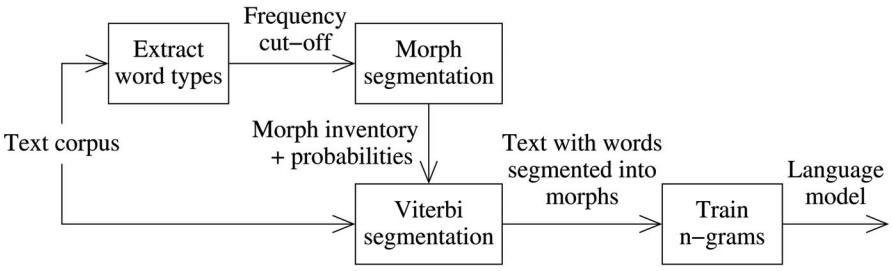

Fig. 3. The steps in the process of estimating a language model based on statistical morphs from a text corpus (Hirsimäki et al., 2006).

The statistical morph model has several advantages over the rule-based grammatical morphemes, e.g. that no hand-crafted rules are needed and all words can be processed, even the foreign ones. Even if good grammatical morphemes are available for Finnish, it has been shown that the language modeling results by the statistical morphs seem to be at least as good, if not better (Hirsimäki et al., 2006; Creutz et al., 2007b).

3. Experimental setups

Statistical and grammatical units are used as the sub-word approaches in the Finnish, Estonian and Turkish LVCSR experiments. For language model training in Finnish and Estonian experiments we used the growing n-gram training algorithm (Siivola & Pellom, 2005). In this algorithm, the n-grams that increase the training set likelihood enough with respect to the corresponding increase in the model size are accepted into the model (as in the MDL principle). After the growing process the model is further pruned with entropy based pruning. The method allows us to train compact and properly smoothed models using high order n-grams, since only the necessary high-order statistics are collected and stored (Siivola et al., 2007). Using the variable order n-grams we can also effectively control the size of the models to make all compared language models equally large. In this way the n-grams using shorter units do not suffer from a restricted span length which is the case when only 3-grams or 4-grams are available. For language model training in Turkish, n-gram language models were built with SRILM toolkit (Stolcke, 2002). To be able to handle computational

limitations, entropy-based pruning (Stolcke, 1998) is applied. In this pruning, the *n*-grams that change the model entropy less than a given threshold are discarded from the model.
The recognition tasks are speaker independent fluent dictation of sentences taken from newspapers and books for Finnish and Estonian. BN transcription system is used for Turkish experiments.

3.1 Finnish

Finnish is a highly inflected language, in which words are formed mainly by agglutination and compounding. Finnish is also the language for which the algorithm for the unsupervised morpheme discovery (Creutz & Lagus, 2002) was originally developed. The units of the morph lexicon for the experiments in this paper were learned from a joint corpus containing newspapers, books and newswire stories of totally about 150 million words (CSC, 2001). We obtained a lexicon of 50K statistical morphs by feeding the learning algorithm with the word list containing the 390K most common words. The average length of a morph was 3.4 letters including a word break symbol whereas the average word length was 7.9 letters. For comparison we also created a lexicon of 69K grammatical morphs based on rule-based morphological analysis of the words. For language model training we used the same text corpus and the growing n-gram training algorithm (Siivola & Pellom, 2005) and limited the language model size to approximately 40M n-grams for both statistical and grammatical morphs and words.
The speech recognition task was speaker independent reading of full sentences recorded over fixed telephone line. Cross-word triphone models were trained using 39 hours from 3838 speakers. The development set was 46 minutes from 79 new speakers and the evaluation set was another corresponding set. The models included tied state hidden HMMs of totally 1918 different states and 76046 Gaussian mixture (GMM) components, short-time mel-cepstral features (MFCCs), maximum likelihood linear transformation (MLLT) and explicit phone duration models (Pylkkönen & Kurimo, 2004). No speaker or telephone call specific adaptation was performed. Real-time factor of recognition speed was about 10 xRT.

3.2 Estonian

Estonian is closely related to Finnish and a similar language modeling approach was directly applied to the Estonian recognition task. The text corpus used to learn the morph units and train the statistical language model consisted of newspapers and books, altogether about 127 million words (Segakorpus, 2005). As in the Finnish experiments, a lexicon of 50K statistical morphs was created using the Morfessor Baseline algorithm as well as a word lexicon with a vocabulary of 500K most common words in the corpus. The average length of a morph was 2.9 letters including a word break symbol whereas the average word length was 6.6 letters. The available grammatical morphs in Estonian were, in fact, closer to the stem-ending models, for which a vocabulary of 500K most common units was chosen. Corresponding growing n-gram language models (approximately 40M n-grams) as in Finnish were trained from the Estonian corpus.
The speech recognition task in Estonian consisted of long sentences read by 50 randomly picked held-out test speakers, 8 sentences each (a part of (Meister et al., 2002)). The training data consisted of 110 hours from 1266 speakers recorded over fixed telephone line as well as cellular network. This task was more difficult than the Finnish one, one reason being the more diverse noise and recording conditions. The acoustic models were rather similar cross-

word triphone GMM-HMMs with MFCC features, MLLT transformation and the explicit phone duration modeling than the Finnish one, except larger: 3101 different states and 49648 GMMs (fixed 16 Gaussians per state). Thus, the recognition speed is also slower than in Finnish, about 30 xRT. No speaker or telephone call specific adaptation was performed.

3.3 Turkish

Turkish is another agglutinative language with relatively free word order. The same Morfessor Baseline algorithm (Creutz & Lagus, 2005) as in Finnish and Estonian was applied to Turkish texts as well. Using the 394K most common words from the training corpus, 34.7K morph units were obtained. The training corpus consists of 96.4M words taken from various sources: online books, newspapers, journals, magazines, etc. In average, there were 2.38 morphs per word including the word break symbol. Therefore, n-gram orders higher than words are required to track the n-gram word statistics and this results in more complicated language models. The average length of a morph was 3.1 letters including a word break symbol whereas the average word length was 6.4 letters. As a reference model for grammatical sub-words, we also performed experiments with stem-endings. The reason for not using grammatical morphemes is that they introduced several very short recognition units. In the stem-ending model, we selected the most frequent 50K units from the corpus. This corresponds to the most frequent 40.4K roots and 9.6K endings. The word OOV rate with this lexicon was 2.5% for the test data. The advantage of these units compared to the other sub-words is that we have longer recognition units with an acceptable OOV rate. In the stem-ending model, the root of each word was marked instead of using word break symbols to locate the word boundaries easily after recognition. In addition, a simple restriction was applied to enforce the decoder not to generate consecutive ending sequences.
For the acoustic data, we used the Turkish Broadcast News database collected at Boğaziçi University (Arısoy et et al., 2007). This data was partitioned into training (68.6 hours) and test (2.5 hours) sets. The training and test data were disjoint in terms of the selected dates.
N-gram language models for different orders with interpolated Kneser-Ney smoothing were built for the sub-word lexicons using the SRILM toolkit (Stolcke, 2002) with entropy-based pruning. In order to eliminate the effect of language model pruning in sub-words, lattice output of the recognizer was re-scored with the same order *n*-gram language model pruned with a smaller pruning constant. The transcriptions of acoustic training data were used in addition to the text corpus in order to reduce the effect of out-of-domain data in language modeling. A simple linear interpolation approach was applied for domain adaptation.
The recognition tasks were performed using the AT&T Decoder (Mohri & Riley, 2002). We used decision-tree state clustered cross-word triphone models with approximately 7500 HMM states. Instead of using letter to phoneme rules, the acoustic models were based directly on letters. Each state of the speaker independent HMMs had a GMM with 11 mixture components. The HTK front-end (Young et al., 2002) was used to get the MFCC based acoustic features. The baseline acoustic models were adapted to each TV/Radio channel using supervised MAP adaptation on the training data, giving us the channel adapted acoustic models.

4. Experimental results

The recognition results for the three different tasks: Finnish, Estonian and Turkish, are provided in Tables 1-3. In addition to sub-word language models, large vocabulary word-

based language models were built as the reference systems with similar OOV rates for each language. The word-based reference language models were trained as much as possible in the same way as the corresponding morph language models. For Finnish and Estonian the growing *n*-grams (Siivola & Pellom, 2005) were used. For Turkish a conventional n-gram with entropy-based pruning was built by using SRILM toolkit similarly as for the morphs.
For Finnish, Estonian and Turkish experiments, the LVCSR systems described in Section 3 are utilized. In each task the word error rate (WER) and letter error rate (LER) statistics for the morph-based system is compared to corresponding grammatical sub-word-based and word-based systems. The resulting sub-word strings are glued to form the word-like units according to the word break symbols included in the language model (see Fig. 1) and the markers attached to the units. The WER is computed as the sum of substituted, inserted and deleted words divided by the correct number of words. In agglutinative languages the words are long and contain a variable amount of morphemes. Thus, any incorrect prefix or suffix would make the whole word incorrect. Therefore, in addition to WER, LER is included here as well.

Finnish	Lexicon	OOV (%)	WER (%)	LER (%)
Words	500 K	5.4	26.8	7.7
Statistical morphs	50 K	0	21.7	6.8
Grammatical morphemes	69 K	0*	21.6	6.9

Table 1. The LVCSR performance for the Finnish telephone speech task (see Section 3.1). The words in (*) were segmented into grammatical morphs using a maximum likelihood segmentation by Viterbi search.

Estonian	Lexicon	OOV (%)	WER (%)	LER (%)
Words	500 K	5.6	34.0	12.3
Statistical morphs	50 K	0	33.9	12.2
Grammatical morphemes	500 K	0.5*	33.5	12.4

Table 2. The LVCSR performance for the Estonian telephone speech (see Section 3.2). The words in (*) were segmented into grammatical morphs using a maximum likelihood segmentation by Viterbi search.

Turkish	Lexicon	OOV (%)	WER (%)	LER (%)
Words	100K	5.3	37.0	19.3
Statistical morphs	37.4K	0	35.4	18.5
Grammatical stem-endings	50K	2.5	36.5	18.3

Table 3. Turkish BN transcription performance with channel adapted acoustic models (see Section 3.3). Best results are obtained with 3-gram word, 5-gram morph and 4-gram stem-ending language models. Note that roots are marked in stem-endings instead of using word break symbols.

In all three languages statistical morphs perform almost the same or better than the large vocabulary word reference models with smaller vocabulary sizes. The performance of the morph model is more pronounced in the Finnish system where the Morfessor algorithm was

originally proposed. In addition, grammatical morphemes achieve similar performances with their statistical counterparts. Even though grammatical stem-endings in the Turkish system attain almost the same LER with the statistical morphs, statistical morphs perform better than stem-endings in terms of the WER.

5. Conclusion

This work presents statistical language models trained on different agglutinative languages utilizing a lexicon based on the recently proposed unsupervised statistical morphs. The significance of this work is that similarly generated sub-word unit lexica are developed and successfully evaluated in three different LVCSR systems in different languages. In each case the morph-based approach is at least as good or better than a very large vocabulary word-based LVCSR language model. Even though using sub-words alleviates the OOV problem and performs better than word language models, concatenation of sub-words may result in over-generated items. It has been shown that with sub-words recognition accuracy can be further improved with post processing of the decoder output (Erdoğan et al., 2005; Arısoy & Saraçlar, 2006).

The key result of this chapter is that we can successfully apply the unsupervised statistical morphs in large vocabulary language models in all the three experimented agglutinative languages. Furthermore, the results show that in all the different LVCSR tasks, the morph-based language models perform very well compared to the reference language model based on very large vocabulary of words. The way that the lexicon is built from the word fragments allows the construction of statistical language models, in practice, for almost an unlimited vocabulary by a lexicon that still has a convenient size. The recognition was here restricted to agglutinative languages and tasks in which the language used is both rather general and matches fairly well with the available training texts. Significant performance variation in different languages can be observed here, because of the different tasks and the fact that comparable recognition conditions and training resources have not been possible to arrange. However, we believe that the tasks are still both difficult and realistic enough to illustrate the difference of performance when using language models based on a lexicon of morphs vs. words in each task. There are no directly comparable previous LVCSR results on the same tasks and data, but the closest ones which can be found are around 15% WER for a Finnish microphone speech task (Siivola et al., 2007), around 40% WER for the same Estonian task (Alumäe, 2005; Puurula & Kurimo, 2007) and slightly over 30% WER for a Turkish task (Erdoğan et al., 2005).

Future work will be the mixing of the grammatical and statistical sub-word-based language models, as well as extending this evaluation work to new languages.

6. Acknowledgments

The authors would like to thank Sabancı and ODTÜ universities for the Turkish text data and AT&T Labs – Research for the software. This research is partially supported by TÜBİTAK (The Scientific and Technological Research Council of Turkey) BDP (Unified Doctorate Program), TÜBİTAK Project No: 105E102, Boğaziçi University Research Fund Project No: 05HA202 and the Academy of Finland in the projects *Adaptive Informatics* and *New adaptive and learning methods in speech recognition*.

7. References

Alumäe, T. (2005). Phonological and morphological modeling in large vocabulary continuous Estonian speech recognition system, *Proceedings of Second Baltic Conference on Human Language Technologies*, pages 89–94.

Arısoy, E.; Dutağacı, H. & Arslan, L. M. (2006). A unified language model for large vocabulary continuous speech recognition of Turkish. Signal Processing, vol. 86, pp. 2844–2862.

Arısoy, E. & Saraçlar, M. (2006). Lattice extension and rescoring based approaches for LVCSR of Turkish, *Proceedings of Interspeech*, Pittsburgh, PA, USA.

Arısoy, E.; Sak, H. & Saraçlar, M. (2007). Language modeling for automatic Turkish broadcast news transcription, Proceedings of Interspeech, Antwerp, Belgium.

Bayer, A. O.; Çiloğlu, T & Yöndem, M. T. (2006). Investigation of different language models for Turkish speech recognition, *Proceedings of 14th IEEE Signal Processing and Communications Applications,* pp. 1–4, Antalya, Turkey.

Byrne, W.; Hajic, J.; Ircing, P.; Jelinek, F.; Khudanpur, S.; Krbec, P. & Psutka, J. (2001). On large vocabulary continuous speech recognition of highly inflectional language - Czech, *Proceedings of Eurospeech 2001*, pp. 487–490, Aalborg, Denmark.

Creutz, M. & Lagus, K. (2002). Unsupervised discovery of morphemes, *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30.

Creutz, M. & Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology. URL: http://www.cis.hut.fi/projects/morpho/.

Creutz, M.; Hirsimaki, T.; Kurimo, M.; Puurula, A.; Pylkkönen, J.; Siivola, V.; Varjokallio, M.; Arısoy, E.; Saraçlar, M. & Stolcke, A. (2007a). Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages. *Proceedings of HLT-NAACL 2007*, pp. 380–387, Rochester, NY, USA.

Creutz, M.; Hirsimäki, T.; Kurimo, M.; Puurula, A.; Pylkkönen, J.; Siivola, V.; Varjokallio, M.; Arısoy, E.; Saraçlar, M. & Stolcke, A. (2007b). Morph-Based Speech Recognition and Modeling of Out-of-Vocabulary Words Across Languages. *ACM Transactions on Speech and Language Processing*, Vol. 5, No. 1, Article 3.

Erdoğan, H.; Büyük, O. & Oflazer, K. (2005). Incorporating language constraints in sub-word based speech recognition. *Proceedings of IEEE ASRU*, San Juan, Puerto Rico

Hacıoğlu, K.; Pellom, B.; Çiloğlu, T.; Öztürk, Ö; Kurimo, M. & Creutz, M. (2003). On lexicon creation for Turkish LVCSR, *Proceedings of Eurospeech,* Geneva, Switzerland.

Hetherington, I. L. (1995). A characterization of the problem of new, out-of-vocabulary words in continuous-speech recognition and understanding. *Ph.D. dissertation*, Massachusetts Institute of Technology.

Hirsimäki T.; Creutz, M.; Siivola, V.; Kurimo, M.; Virpioja, S. & J. Pylkkönen. (2006). Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer, Speech and Language*, vol. 20, no. 4, pp. 515–541.

Garofolo, J.; Auzanne, G. & Voorhees, E. (2000). The TREC spoken document retrieval track: A success story, *Proceedings of Content Based Multimedia Information Access Conference*, April 12-14.

Kanevsky, D.; Roukos, S.; & Sedivy, J. (1998). Statistical language model for inflected languages. US patent No: 5,835,888.

Kwon, O.-W. & Park, J. (2003). Korean large vocabulary continuous speech recognition with morpheme-based recognition units. *Speech Communication*, vol. 39, pp. 287–300.

Meister, E.; Lasn, J. & Meister, L. (2002). Estonian SpeechDat: a project in progress, *Proceedings of the Fonetiikan Päivät–Phonetics Symposium 2002 in Finland*, pages 21–26.

Mengüşoğlu, E. & Deroo, O. (2001). Turkish LVCSR: Database preparation and language modeling for an agglutinative language. *Proceedings of ICASSP 2001, Student Forum*, Salt-Lake City.

Mohri, M & Riley, M. D. DCD Library – Speech Recognition Decoder Library. AT&T Labs – Research. http://www.research.att.com/sw/tools/dcd/.

NIST. (2000). *Proceedings of DARPA workshop on Automatic Transcription of Broadcast News*, NIST, Washington DC, May.

Podvesky, P. & Machek, P. (2005). Speech recognition of Czech - inclusion of rare words helps, *Proceedings of the ACL SRW*, pp. 121–126, Ann Arbor, Michigan, USA.

Puurula, A. & Kurimo M. (2007). Vocabulary Decomposition for Estonian Open Vocabulary Speech Recognition. *Proceedings of the* ACL 2007.

Pylkkönen, J. & Kurimo, M. (2004). Duration modeling techniques for continuous speech recognition, *Proceedings of the International Conference on Spoken Language Processing*.

Pylkkönen, J. (2005). New pruning criteria for efficient decoding, *Proceedings of 9th European Conference on Speech Communication and Technology*.

Rosenfeld, R. (1995). Optimizing lexical and n-gram coverage via judicious use of linguistic data, *Proceedings of Eurospeech*, pp. 1763–1766.

Sak, H.; Güngör, T. & Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus, *Proceedings of 6th International Conference on Natural Language Processing, GoTAL* 2008*, LNAI 5221*, pp. 417–427..

Sak, H.; Güngör, T. & Saraçlar, M. (2007). Morphological disambiguation of Turkish text with perceptron algorithm, Proceedings of *CICLing 2007, LNCS 4394*, pp. 107–118.

Segakorpus–Mixed Corpus of Estonian. Tartu University. http://test.cl.ut.ee/korpused/segakorpus/.

Siivola, V. & Pellom, B. (2005). Growing an n-gram language model, *Proceedings of 9th European Conference on Speech Communication and Technology*.

Siivola, V.; Hirsimäki, T. & Virpioja, S. (2007). On Growing and Pruning Kneser-Ney Smoothed N-Gram Models. *IEEE Transactions on Audio, Speech and Language Processing*, Volume 15, Number 5, pp. 1617-1624.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit, *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904.

Stolcke, A. (1998). Entropy-based pruning of back-off language models, *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274.