Lattice Extension and Rescoring Based Approaches for LVCSR of Turkish

Ebru Arısoy and Murat Saraçlar

Electrical and Electronic Engineering Department
Boğaziçi University, 34342 Bebek, Istanbul, Turkey
{arisoyeb,murat.saraclar}@boun.edu.tr

Abstract

In this paper, we present some techniques to solve the problems of Turkish Large Vocabulary Continuous Speech Recognition (LVCSR). Its agglutinative nature makes Turkish a challenging language in terms of speech recognition since it is impossible to include all possible words in the recognition lexicon. Therefore, data-driven sub-word recognition units, in addition to words, are used in a newspaper content transcription task. We obtain Word Error Rates (WER) of 38.8% for the baseline word model and 33.9% for the baseline sub-word model. In addition, some new methods are investigated. Baseline lattice outputs of each model are rescored with the root and root-class language models for words and first-sub-word language model for sub-words. The word-root interpolation achieves 0.5% decrease in the WER. Other two approaches fail due to the non-robust estimates over the baseline models. Moreover, we have tried dynamic vocabulary extension techniques to handle the Out-of-Vocabulary (OOV) problem in the word model and to remove non-word items in the sub-word model. Applying this method to the 50K baseline word model, in the best situation, we obtain an error rate of 36.2%. In average, the lexicon size of this method is around 188K. However, the error rate is approximately same as the 120K lexicon recognizer. For sub-words, 1.1% absolute improvement is achieved with the vocabulary extension technique giving us our best result.

Index Terms: speech recognition, language modelling, vocabulary extension, agglutinative languages.

1. Introduction

Turkish is a challenging language for LVCSR. The agglutinative nature of the language causes the vocabulary to expand significantly which is problematic for speech recognition. It is not possible to add all the words to the lexicon to handle the OOV problem. Also, huge lexicon sizes may result in confusion of acoustically similar words and require a huge amount of text data for robust language model estimates. The most common method proposed to handle OOV words and non robust language model estimates is to use sub-word recognition units instead of words [1, 2, 3].

There has not been many Turkish speech recognition studies until recently. In terms of sub-word approaches, morpheme-based models [4], stem-ending based models [5], and a unified model using all the previous methods together [6] were investigated. Post-processing of the sub-word recognition output using vowel harmony rules gave slight improvements [7]. Unsupervised segmentation of words using Minimum Description Length (MDL) principle, first applied to Finnish [2], was also applied to Turkish [8]. A detailed comparison of these unsupervised segments with word based recognition units for Turkish as well as for other agglutinative languages, Finnish and Estonian, showed promising results [9].

This study is an attempt to solve the main problems of both word and sub-word approaches for Turkish. The high number of OOV words and the non-robust language model estimates due to the vocabulary explosion are the main problems of the word-based language model. The agglutinative morphology gives rise to thousands of new words from the same root. Therefore, the main idea to handle data sparseness in word-based modelling is to use the roots in addition to words as language modeling units. Root-based and word-based language models are interpolated to obtain better language model estimates. To alleviate the OOV problem, we investigate using vocabulary extension [10] in a lattice rescoring framework. We chose to extend the vocabulary by adding words that share the same root as the words in the lattice. Both techniques result in a decrease in WER compared to the baseline model. Since sub-word based models result in better performance than word-based models, we adapted the same strategies to sub-words. Instead of roots we use first sub-word of the whole word for the interpolation. One drawback of the sub-word approach is that it can generate any combination of sub-word units which include non-word sequences. When we join the sub-word units with the help of a word boundary symbol, for the test set we obtain 6759 words, 159 of which do not occur in the full 683K training text vocabulary. Out of these non-vocabulary items, only 19 are correct Turkish words. Other kinds of errors are wrong insertion of word boundary, incorrect morphotactics and meaningless sequences. Simply getting rid of these non-words by pruning them from the lattice increases the WER. Therefore, we adapt the vocabulary extension strategy mentioned above by once again using the first sub-words instead of roots. This ensures that the output will contain valid words.

This paper is organized as follows. In the next section details of the language modelling units with a comparison in terms of coverage and recognition performance are given. Section 3 explains the methods used to modify the word-based and sub-word based baseline recognizers. Section 4 explains the experiments and discusses the results. Finally Section 5 concludes the paper.

2. Statistical Language Modelling Units

In this research, words and statistical morph models are used as language modelling units. The morphological productivity of Turkish makes it difficult to construct a robust word-based language model. With a dictionary size of a few hundred thousand words, we can still have out of vocabulary words, which are constructed through legal morphological rules. The morph model is a sub-word approach where a recursive Minimum Description

Table 1: WER and LER for the baseline recognizers for word-based and morph-based models

Experiments	Lexicon	Coverage Test (%)	WER (%)	LER (%)
Baseline-word	50K	88.2	38.8	15.2
Baseline-word*	50.7K	100	30.0	11.9
Baseline-word	120K	94.4	36.0	14.1
Baseline-morph	34.3K	100	33.9	12.4

Length (MDL) algorithm learns a sub-word lexicon in an unsupervised manner from a training lexicon of words [11]. In order to recover the word sequences from morph sequences, a word boundary morph '#' is added between each morph in the language model.

Morphs: istekler imizi # el de # etti k # de di

Words: isteklerimizi elde ettik dedi

Using a 26.6M words text corpus, 683K word and 34.3K morph types are generated. For word based-model, 90% and 95% self-coverages are achieved using the most frequent 50K and 120K words respectively. Morph-based models handle the OOV problem with a smaller vocabulary size using units that are still meaningful for language modelling. Comparison of these two units in terms of test coverage, Word Error Rate (WER) and Letter Error Rate (LER) is given in Table 1[1]. Best results are obtained using 3-gram word and 5-gram morph language models. In the text corpus, the ratio of morph tokens to word tokens are calculated as 2.43 including the '#' symbol. This suggests that 6-grams for morphs is comparable to 3-grams for words. In addition, Baseline-word* is a cheating experiment where all the OOV words in the test data are added to the lexicon to give a lower WER bound for our experiments.

3. Application of Different Language Modelling Approaches

In this section, we apply some language modelling techniques to Turkish for both the word and the morph-based models.

3.1. Modifications to Word-based Model

The main drawbacks of the word-based model are i) non-robust n-gram estimates, ii) high number of OOV rates, which are due to the agglutinative nature of the language. Our first two approaches target to generate more robust estimates for the word-based language model and the third approach targets to overcome the OOV problem by dynamically extending vocabularies.

3.1.1. Root-based Language Models

For agglutinative languages like Turkish, thousands of new word forms can be generated from a single stem. Therefore, robust n-gram language model estimates for these languages need a huge size of text data. In root-based language modelling, the standpoint is that the roots can capture the regularities better than word-based models, since word-based model results in more sparse data compared to roots (See Figure 1). The average ratio of words to roots is 3.7. In this technique, roots are extracted using a simple stemmer and they are considered as functions of words denoted by $r(w)$. N-gram estimates are generated by considering the roots as words. In this approach, the word trigram probabilities are approximated by root trigram probabilities.

$$P(w_3|w_2, w_1) = P(r(w_3)|r(w_2), r(w_1))$$

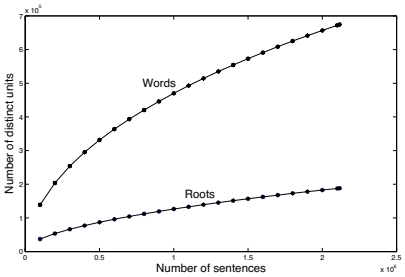

Figure 1: Vocabulary growth for words and roots

3.1.2. Class-based Language Models

Class-based language modelling is an other way to handle the data sparseness problem. The aim in this model is to group the words that have similar grammatical or semantic properties. We use the root of each word as the class of that word. Word trigram probabilities are calculated as:

$$P(w_3|w_2, w_1) = P(w_3|r(w_3)) * P(r(w_3)|r(w_2), r(w_1))$$

This model corrects the approximation used in the root-based model by considering the effect of the probability of word w_i given its root $r(w_i)$, $P(w_i|r(w_i))$, during the calculation of n-gram probabilities.

3.1.3. Vocabulary Extension

Existence of OOV words is a significant source of recognition errors in speech recognition since if the word is not in the recognition lexicon the recognizer has no chance to recognize it correctly. As shown in Table 1, we achieve 2.8% absolute improvement using a larger lexicon. Although, it is practically and theoretically impossible to add all the available words to the lexicon, a cheating experiment where all OOV words are added to the lexicon gets the WER down to 30%.

The main idea in vocabulary extension is to dynamically add similar words to the lexicon and extend the utterance lattice to decrease the errors due to OOV words, then perform second pass recognition. In addition to morphology-based similarity [10], phonetic distance-based similarity [12] has been suggested. In our experiments, the similarity criteria between words is having the same roots. Instead of building a language model for each utterance, we use a single language model built using a closed vocabulary including all 683K words seen in the training corpus. We also investigated using language models with topic dependent vocabularies (188K words in average). The vocabularies were determined from the first pass lattice outputs.

3.2. Modifications to Morph-based Language Modelling

As was shown in Section 2 in Table 1, the sub-word approach outperforms the word-based model. Then, we assume that the same methods, which we tried for words, may give better results on the morph-based model and as the counterpart of roots, we decide to use the first morph of each word.

3.2.1. First-morph-based Language Models

Similar to the root-based modeling approach for words, we generate a new language model using only the first morph of each word, instead of a language model with all the morph units. Figure 2

[1] Results for baseline word and morph models were previously reported in [9].

illustrates this. The dashed lines show the baseline morph model and the bold lines show the new model (first-morph-based) dependencies for m_{31}. As seen from the figure, the effect of the new model is only on the first morph of the word and there is no contribution from the new model for m_{32}. The n-gram probabilities for m_{31} can be calculated as:

First-morph model :$P(m_{31}|m_{21}, m_{11})$

Morph model :$P(m_{31}|\#, m_{23}, m_{22}, m_{21})$

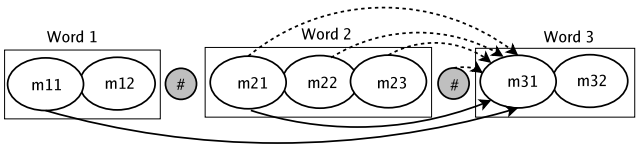

Figure 2: Dependencies for morph-based and first-morph-based models

3.2.2. Vocabulary Extension

Although the morph-model performs better than the word-model, 2.35% of the letter sequences in the best recognition path do not occur in the closed training vocabulary. As mentioned before, removing the arcs containing non-vocabulary items increases the WER to 35.7%. Therefore, in order to deal with these non-vocabulary items we decided to do a mapping from the morphs to words in a way similar to vocabulary extension. Although this method brings back the OOV problem, the output of the recognizer will be morphologically correct Turkish words. The mapping function converts a sequence of morphs to a sequence of words having the same sequence of first-morphs. All unique words are split into morphs, and sets of words sharing the same first-morph are determined. The morph lattice is converted into an extended word lattice using these sets. Then, second pass recognition is performed over the extended word lattice.

4. Experiments

4.1. Experimental Setup

In this research, we have used 2 different text corpora for statistical language modelling. One of them contains 11.6M words from various domains and other one is specific to sports news with 15M words. Statistical language models are generated using SRI Language Modelling toolkit [13] with interpolated modified Kneser-Ney smoothing for each corpus and these two models are interpolated to minimize the test set perplexity. The recognition tasks are performed using AT&T Decoder [14]. We use decision-tree state clustered cross-word triphone models with approximately 5000 HMM states. Instead of using letter to phoneme rules, the acoustic models are based directly on letters. Each state of the speaker independent HMMs has a GMM with 6 mixture components. The training data contains 17 hours of speech from over 250 speakers. The test material consisted of approximately one hour (6989 words) of newspaper sentences read by one female speaker.

4.2. Results

In all of the new language modelling experiments, except for the vocabulary extension, log probability of the original and the new language models are interpolated with an interpolation constant of α, where $0 < \alpha < 1$. Then lattice rescoring strategy is applied to evaluate the results in terms of WER.

Figure 3 shows the effect of the α constant on the WER for three different experiments. $\alpha = 0$ means acoustic lattice is rescored using only the new language model and $\alpha = 1$ means acoustic lattice is rescored using only the original language model. For root-based approach, $\alpha = 0.6$ gives the lowest WER, 38.3%. We obtain an absolute reduction of 0.5% in the WER (See left most plot of Figure 3). Using the 50K lexicon, self coverage is calculated as 90%, however in terms of roots this coverage increases to 97.3%. So better estimates for language model probabilities can be achieved when roots are considered as functions of words.

For the class-based model using the root classes, interpolated class-based language model does not give any improvement over the baseline model (See middle plot of Figure 3). This might be because $P(w_i|c_i)$ probabilities are not robustly estimated from the available training data.

For the morph-based model, the original 5-gram language model is interpolated with 2-gram and 3-gram first-morph language models since the ratio of $\frac{morphs}{words}$ including # symbols in training data is 2.43. However, no improvement is achieved (See right most plot of Figure 3).

For vocabulary extension, we perform three experiments: i) for word-based model using closed vocabulary, ii) for morph-based model using closed vocabulary, iii) for word-based model using topic specific vocabularies. In each experiment we use the lattice output of the baseline recognizer. In the first experiment, all possible words are added to the lattice using the root similarity. Morph-based vocabulary extension is performed in a similar manner using first-morph similarity. In the third experiment, utterances are divided into 40 topics/stories and a different topic vocabulary specific language model is estimated for each topic. The lattices are extended and a second pass recognition is performed. The recognition results are given in Table 2 for WER, LER and Lattice Word Error Rate (LWER). For 40 stories extended lattice experiment, the LWER is given in terms of mean and standard deviation (std), since each story has its own language model. We obtain 2.2% absolute improvement using vocabulary extension on word lattices. Also there is a significant reduction in LWER. Vocabulary extension using topic adaptive vocabularies shows slight improvement over using a single closed vocabulary. The vocabulary for the topic dependent approach is determined from the lattices. Therefore, each topic has different lexicons. Although, the average lexicon size for this experiment is 188K, the WER of the word model with 120K vocabulary, 36.0% is slightly better.

Table 2: WER, LER and LWER for the baseline recognizers and the extended lattices

	WER	LER	LWER
Words			
Baseline (50K)	38.8	15.2	15.5
Extended Lattice	36.6	14.3	9.6
Extended Lattice (40 stories)	36.2	14.0	10.445(mean) 5.76 (std)
Morphs			
Baseline (34.3K)	33.9	12.4	14.7
Extended Lattice	32.8	12.2	6.0

For the morph-based vocabulary extension, extended lattice reduces the WER to 32.8%. To be able to calculate the LWER for baseline morph model, we generate 10,000 best paths from the recognition lattice and remove the arcs that contain non-vocabulary items. Then the remaining paths are converted to a lattice. The WER for vocabulary extended morph model is 1.1% lower than the original morph-based model.

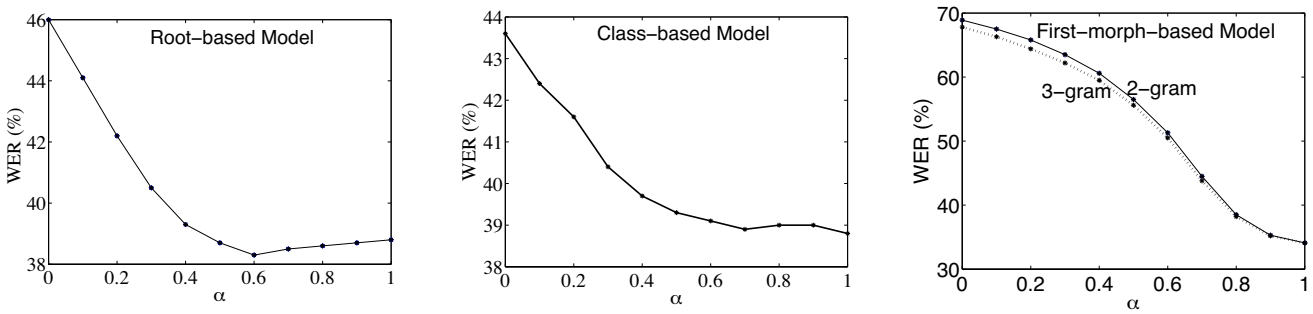

Figure 3: Interpolation parameter versus WER for the lattice rescoring experiments. The plots from left to right are the interpolation of root-based, class-based and first-morph-based language models with the baseline language model respectively.

5. Conclusion

In this paper, word-based and morph-based recognition units are used for the transcription of newspaper content and WERs of 38.8% for 50K word lexicon and 33.9% for 34.3K morph lexicon are obtained recpectively. To handle the non-robust language model estimates due to the data sparseness for word-based approach, root-based and class-based models are tried and only 0.5% absolute improvement is achieved with the root model. To handle OOV problem vocabulary extension method is applied. However, the best result obtained with this method is similar to using a 120K lexicon and we are still far away from the lower bound of 30% obtained by a cheating experiment where all OOV words seen in the test set are included in the vocabulary. For morph-based units, as a counterpart of roots, we try first-morph based language modelling and vocabulary extension. No improvement is achieved in first-morph-based language modelling. Vocabulary extension using the first-morphs brings 1.1% absolute improvement giving us our best result.

6. Acknowledgements

The authors would like to thank Sabanci and ODTU universities for the Turkish acoustic and text data and AT&T Labs – Research for the software. This research is partially supported by SIMILAR Network of Excellence, TÜBİTAK BDP (Unified Doctorate Program of the Scientific and Technological Research Council of Turkey) and Boğaziçi University Research Fund (Project code: 05HA202).

7. References

[1] Oh-Wook Kwon and Jun Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," *Speech Communication*, vol. 39, pp. 287–300, 2003.

[2] Vesa Siivola, Teemu Hirsimäki, Mathias Creutz, and Mikko Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003, pp. 2293–2296.

[3] Jan Kneissler and Dietrich Klakow, "Speech recognition for huge vocabularies by using optimized sub-word units," in *Proc. EUROSPEECH 2001*, Aalborg, Denmark, 2001, pp. 69–73.

[4] Kenan Carkı, Petra Geutner, and Tanja Schultz, "Turkish LVCSR: Towards better speech recognition for agglutinative languages," in *Proc. ICASSP 2000*, Istanbul, Turkey, 2000, vol. 3, pp. 1563–1566.

[5] Erhan Mengusoglu and Olivier Deroo, "Turkish LVCSR: Database preparation and language modeling for an agglutinative language," in *Proc. ICASSP 2001, Student Forum*, Salt-Lake City, 2001.

[6] Ebru Arısoy and Levent M. Arslan, "Turkish dictation system for broadcast news applications," in *Proc. EUSIPCO 2005*, Antalya, Turkey, 2005.

[7] Hakan Erdogan, Osman Buyuk, and Kemal Oflazer, "Incorporating language constraints in sub-word based speech recognition," in *Proc. ASRU 2005*, Cancun, Mexico, 2005.

[8] Kadri Hacioglu, Brian Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo, and Mathias Creutz, "On lexicon creation for Turkish LVCSR," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003, pp. 1165–1168.

[9] Mikko Kurimo, Antti Puurula, Ebru Arısoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumäe, and Murat Saraçlar, "Unlimited vocabulary speech recognition for agglutinative languages," in *Proc. HLT-NAACL 2006*, New York, USA, 2006.

[10] Petra Geutner, Michael Finke, Peter Scheytt, Alex Waibel, and Howard Wactlar, "Transcribing multilingual broadcast news using hypothesis driven lexical adaptation," in *DARPA Broadcast News Workshop*, Herndon, USA, 1998.

[11] Mathias Creutz and Krista Lagus, "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, March.," 2005.

[12] Petra Geutner, Michael Finke, and Alex Waibel, "Phonetic-distance-based hypothesis driven lexical adaptation for transcribing multilingual broadcast news," in *ICSLP 1998*, Sydney, Australia, 1998.

[13] Andreas Stolcke, "Srilm – An extensible language modeling toolkit," in *Proc. ICSLP 2002*, Denver, 2002, vol. 2, pp. 901–904.

[14] Mehryar Mohri and Michael D. Riley, "Dcd library, speech recognition decoder library, AT&T Labs - Research. http://www.research.att.com/sw/tools/dcd/," 2002.