Morpheme Concatenation Approach in Language Modeling for Large-Vocabulary Uyghur Speech Recognition

Mijit Ablimit*, Askar Hamdulla†, Tatsuya Kawahara*

* School of Informatics, Kyoto University, Kyoto, Japan
†Institute of Information Engineering, Xinjiang University, Urumqi, China
E-mail: mijit@ar.media.kyoto-u.ac.jp

ABSTRACT

For large-vocabulary continuous speech recognition (LVCSR) of highly-inflected languages, selection of an appropriate recognition unit is the first important step. The morpheme-based approach is often adopted because of its high coverage and linguistic properties. But morpheme units are short, often consisting of one or two phonemes, thus they are more likely to be confused in ASR than word units. Generally, word units provide better linguistic constraint, but increases the vocabulary size explosively, causing OOV (out-of-vocabulary) and data sparseness problems in language modeling. In this research, we investigate approaches of selecting word entries by concatenating morpheme sequences, which would reduce word error rate (WER). Specifically, we compare the ASR results of the word-based model and those of the morpheme-based model, and extract typical patterns which would reduce the WER. This method has been successfully applied to an Uyghur LVCSR system, resulting in a significant reduction of WER without a drastic increase of the vocabulary size.

Index Terms— Speech recognition, language model, morpheme, Uyghur

1. INTRODUCTION

The recognition unit affects the vocabulary size and performance of automatic speech recognition (ASR) systems. Words are naturally selected to be the recognition unit in many languages like English. In agglutinative languages such as Japanese and Uyghur, selection of the lexical unit is not obvious, and the word vocabulary size of these languages is huge; therefore, the morpheme unit is conventionally adopted. However, there is a trade-off between the word unit and the morpheme unit; generally the word unit provides better linguistic constraint, but increases the vocabulary size explosively, causing OOV (out-of-vocabulary) and data sparseness problems in language modeling. On the other hand, morphemes are short, often consisting of one or two phonemes, thus increase acoustic confusability in ASR than word unit. The goal of this study is to incorporate effective word entries selectively while maintaining the high coverage of the morpheme unit.

Some data-driven approaches have been investigated with the word frequency basis or likelihood criterion [3][7][8][9]. However, these criteria are not directly related to WER (word error rate). In this work, we extract useful patterns by aligning and comparing the ASR results by the morpheme-based model with those by the word-based model. These patterns are classified according to length, error frequency, and attribute of the units, and individually assessed in terms of their contribution for reducing WER. Specifically, we extract frequently misrecognized morpheme sequences which are correctly recognized by merging them to words; we then identify short and frequently misrecognized morphemes by separating them to stem and word-ending, and recombine them in all possible ways. These methods are applied to and evaluated in a large-vocabulary Uyghur ASR system.

2. CORPUS AND BASELINE SYSTEMS

We have developed an Uyghur-language large-vocabulary continuous speech recognition (LVCSR) system [1]. Uyghur belongs to the Turkish language family of the Altaic language system. The morpheme structure of Uyghur words is ***" prefix + stem + suffix1 + suffix2 + … "***. A root (or stem) is followed by zero to many (at longest 10 or more) suffixes. In this work, 108 suffix types are defined according to their semantic and syntactic functions, and 305 surface forms are extracted. A few words have a prefix (only one) preceding a stem, and seven kinds of prefixes are considered.

For language modeling, a text corpus of 630k sentences is collected from general topics like newspaper articles, novels, and science textbooks. They are segmented to syllable, morpheme, and word units by our morphological analyzer [1]. Morphemes are defined according to their linguistic roles.

Table 1. Statistics of speech corpora.

Corpus	Sentences	Persons	Total utterances	Time (hour)
training	13.7K	353	62k	158.6
test	550	23	1468	2.4

A speech corpus of general topics is prepared to build an acoustic model of Uyghur. This corpus is also used as the training data set for lexical optimization addressed in this work. A test data set is also prepared from newspaper articles. Specifications of the data sets are summarized in Table 1.

Julius [1] is used to build an ASR system. Julius is an open-source LVCSR platform for researchers and developers. The acoustic models and language models are easily pluggable, and you can build various kinds of ASR systems by preparing your own models suitable for the task.

In Uyghur language, surface forms of morphemes transform as the result of phonetic harmony when they are concatenated. We keep the surface forms identical both in the word and morpheme sequences, thus the words can be recovered simply by connecting morphemes without any changes. These may cause some ambiguity in morphemes, but does not degrade segmentation accuracy. A word boundary symbol is inserted to the morpheme sequence for recovering the words from morphemes by simply reconnecting them.

Three different lexical units are used to build n-gram language models, and ASR performance is compared.

①Word-based language model.

②Morpheme-based language models.

③FMS (Frequent Morpheme Sequence) based language model. The FMS unit is built by combining neighboring morpheme sequences which frequently appeared in the training corpus. The optimal frequency threshold to produce the best result was 500 in our text corpus.

④Stem & word endings based language modeling. Suffix sequences observed in the corpus are merged into lexical entries to form word-endings.

The ASR results by these various unit-based language models are summarized in Table 2. The word boundary symbol was added to all units other than the word unit to compare the WER.

The word-based model outperforms the morpheme-based models with a much bigger vocabulary size. However, note that to have low OOV and a reliable language model with the word unit, a very large training data set is needed. Otherwise, the ASR performance would be degraded very much. This property is not good for applying ASR to various domains.

Two kinds of morpheme concatenation are also tested: FMS and stem & word endings, both based on co-occurrence statistics. They made a modest improvement, but still far from the word-based model, and they also increased the vocabulary size very much.

Table 2. ASR error rates for different units.

LM names	WER (%)	Vocabulary size
Word-3gram	25.72	227k
Morph-3gram	28.96	55.2k
Morph-4gram	27.92	55.2k
Morph-5gram	29.31	55.2k
FMS-500	28.14	274.9k
Stem & word endings	28.13	74.5k

3. RECOGNITION UNIT OPTIMIZATION

The goal of this study is to make compatible the vocabulary size of the morpheme unit and the accuracy of the word unit. The objective is to find word entries which reduce the WER with a minimum increase of the vocabulary size. This can be realized by comparing the ASR results by the morpheme-based model and those by the word-based model. As shown previously, merging morpheme sequences with simple co-occurrence statistics has a little effect for reducing WER.

We can classify the patterns related with the confusion into two categories: lexical properties and acoustic properties. The lexical properties related with the lexical unit selection include length (number of syllables) and attribute (stem or word-ending). The acoustic properties can be attributed to coarticulation effects [3]. These properties can be systematically analyzed with linguistic and phonological knowledge. However, instead of speculating the patterns of unknown results, it is convenient to directly observe the ASR results, and enumerate the problematic patterns. Thus, we identify major reasons for confusion for morpheme sequences in comparison with word sequences, as follows.

- Phonetic harmony or coarticulation. (E.g.) yigirmə-yigirmi, vottura-votturi;
- Confusion in frequent short stems with many derivatives. (E.g.) biz, vu, bash, yər;
- Phonetic similarity. (E.g.) həmmə-əmma;
- Ambiguity. (E.g.) uni-u+ni;
- Too many suffix insertions.

Among these patterns, the coarticulation problem caused by phonetic changes can be solved by recovering the morpheme sequences into word units. Similarly, we can extract other problematic morpheme sequences. A simple solution would be to extract all the problematic morpheme sequences and merge them into words, and add them to the lexicon. Our preliminary study showed that this approach works well, but it is difficult to cover all the erroneous words in the open test data. Therefore, we also explore more generalized methods.

The cause of the confusion by the morpheme-based model can be attributed to three types of features: error frequency, length, and attribute (stem or word-ending).

3.1 Error frequency feature

First, as a simple method, we investigate the frequency of misrecognized morpheme sequences. We collect all the words which are misrecognized by the morpheme-based model, but correctly recognized by the word-based model. They are added to the lexicon with a threshold of the frequency of misrecognition. From the ASR results of training data, we collect the candidates of word entries whose error frequency is higher than twice.

$$\mathrm{F(freq)} = \begin{cases} \text{true} & if\ word\ misrecognized\ more\ than\ twice \\ \text{false} & otherwise \end{cases}$$

This method can be iterated to select more candidates. On each iteration, new candidates are extracted and added to the vocabulary, until few new candidates are found.

3.2 Length feature

Short units are easily confused in ASR and usually they are very frequent. Confusion in short morphemes can be reduced by merging and making them longer. There are many single phoneme suffixes produced from our morpheme segmenter. To make them longer, all the short morphemes are merged to each other when they are neighbors or to the neighboring morphemes. Below is an example of the length feature.

$$\mathrm{F(length)} = \begin{cases} \text{true} & if\ morpheme\ length\ is\ less\ than\ 2 \\ \text{false} & otherwise \end{cases}$$

While FMS and stem & word-ending models described in Section 2 are based on statistical co-occurrence of morphemes, the proposed method directly considers their effect on the ASR performance.

3.3 Attribute (stem or word-ending)

We also conduct a simple morphological analysis to find generalized features. From the aligned ASR results, we separate the morpheme sequence into two parts within the word unit boundary: stem and word-ending. Then we separately collect all misrecognized stems and word-endings based on their error frequency. As a result, short and frequent stems are typically extracted. These short stems have many derivatives, and are easily confused. The word-endings are also collected according to their error frequency when they are connected with these short stems. A brief feature description is as follows.

$$\mathrm{F(stem)} = \begin{cases} \text{true} & if\ stem\ misrecognized\ more\ than\ 10\ times, \\ & and\ length\ is\ less\ than\ 4\ syllables \\ \text{false} & otherwise \end{cases}$$

$$\mathrm{F(word\ ending)} = \begin{cases} \text{true} & if\ word\ ending\ misrecognized \\ & when\ connected\ with\ a\ short\ stem \\ \text{false} & otherwise \end{cases}$$

These features are generalized by merging all possible combinations of stems and word-endings into words when both features are observed in the training corpus.

3.4 Combination of features and language modeling

In the above-mentioned methods, effective features are identified separately. On this basis, we can design a combined model by applying all the features.

The newly selected word entries are added to the lexicon of the baseline morpheme-based unit. N-gram language models are built with the new lexicon using a certain cutoff threshold. In this work we set the cutoff threshold to 5.

4. EXPERIMENTAL EVALUATION

The proposed methods are evaluated by applying to the Uyghur LVCSR task. The morpheme-based n-gram model is benefited from a much smaller vocabulary size, thus 4-gram language model is used and compared with the 3-gram word-based model.

The first method is based on the error frequency. From the training data, the words misrecognized more than twice are extracted, and added to the vocabulary. In Table 3, WER for the training and test data after two iterations are listed. When we extract misrecognized words from the test set, we found that only 50% of them are covered by the training data set. This simple method does not have generality; it cannot include entries that are not in the training data set. However, the method is still effective.

In the second method, the morphemes consisting of single phoneme are merged to each other or to the previous morpheme. This simple method made 0.92% reduction of WER, as shown in Table 4.

In the third method, we separate the morpheme sequences into stem and word-ending, and merge them in all possible ways. This method made 1.36% decrease in WER from the baseline model.

Finally, we combine the above proposed method as shown in Table 4. The error frequency feature is taken after the second round. We confirm an accumulative effect. The final result here outperforms the word-based model result in Table 1, with a much smaller vocabulary size.

Table 3. Result of word selection based on error frequency

Iterations	Baseline	First round	Second round
WER(%) on training data	31.95	28.62	27.01
WER(%) on test data	28.11	26.11	25.82
Vocabulary size	27066	40376	46033

Table 4. WER reduction by the proposed methods

Models	WER(%)	ΔWER(%)	Vocabulary size
Morpheme-based baseline	28.11	0	27357
Error frequency feature	26.11	2.00	40376
Length feature	27.19	0.92	32881
Attribute feature	26.74	1.36	36333
Attribute feature+ Length feature	25.80	2.31	41257
Attribute feature+ Length feature+ Error freq feature	24.89	3.22	56718

5. CONCLUSION

We have proposed methods for morpheme concatenation for effective LVCSR. Instead of analyzing linguistic or statistical property from text data, we analyze the ASR results and identify useful patterns based on the error frequency, length, and attributes. The concatenation methods based on these features significantly reduced WER from the morpheme-based baseline model without a drastic increase of the lexicon size compared with the word-based model.

6. REFERENCES

[1] Mijit Ablimit, Graham Neubig, Masato Mimura, Shinsuke Mori, Tatsuya Kawahara, Askar Hamdulla. Uyghur *Morpheme-based* Language Models and *ASR.* IEEE 10th International Conference on Signal Processing (ICSP). Beijing. October 2010.

[2] M.Ablimit, M.Eli, and T.Kawahara. Partly supervised Uyghur morpheme segmentation. In Proc. Oriental-COCOSDA Workshop, 2008, pp.71—76.

[3] G. Saon, M. Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech recognition," IEEE Transactions on Speech and Audio Processing, Vol.9, No.4, May 2001

[4] Ruhi Sarikaya and Mohamed Afify and Yuqing Gao. Joint morphological-lexical language modeling (JMLLM) for Arabic. In proceedings ICASSP 2007-4-1031.

[5] Hasim Sak, Murat Saraclar and Tunga Gungor. Morphology-based and Sub-word Language Modeling for Turkish Speech Recognition. Proceedings SAK:Turkish. 2010.

[6] O.-W. Kwon and J. Park, Korean large vocabulary continuous speech recognition with morpheme-based recognition units. Speech Communication, vol. 39, pp. 287–300, 2003.

[7] Oh-Wook Kwon, "Performance of LVCSR with morpheme-based and syllable-based recognition units" icassp, vol. 3, pp.1567-1570, Acoustics, Speech, and Signal Processing, 2000 Vol 3. 2000 IEEE International Conference on, 2000

[8] Markpong Jongtaveesataporn, Issara Thienlikit, Chai Wutiwiwatchai, Sadaoki Furui. Lexical units for Thai LVCSR. Speech Communication, 2009: 379~389

[9] Kadri Hacioglu, Bryan Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo, Mathias Creutz. *On Lexicon Creation for Turkish LVCSR*. Eighth European Conference on Speech Communication and Technology, 2003.

[10] Michael Collins. Discriminative training methods for HMMs: Theory and experiments with perceptron algorithm. AT&T Labs-Research. EMNLP 2002.

[11] Michael Collins, Brian Roark, Murat Saraclar. Discriminative syntactic language modeling for speech recognition. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), pages 507-514

[12] Zheng Chen, Kai-Fu Lee, Ming-jing Li. Discriminative Training on Language Model. in Proc. ICSLP, 2000.