DISCRIMINATIVE APPROACH TO LEXICAL ENTRY SELECTION FOR AUTOMATIC SPEECH RECOGNITION OF AGGLUTINATIVE LANGUAGE

Mijit Ablimit[*], *Tatsuya Kawahara*[*], *Askar Hamdulla*[†]

[*]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

In agglutinative languages, selection of lexical unit is not obvious. Morpheme unit is usually adopted to ensure the sufficient coverage, but many morphemes are short, resulting in weak constraints and possible confusions. In this paper, we propose a discriminative approach to select lexical entries which will directly contribute to ASR error reduction. We define an evaluation function for each word by a set of features and their weights, and the measure for optimization by the difference of WERs by the morpheme-based model and by the word-based model. Then, the weights of the features are learned by a perceptron algorithm. Finally, word (or sub-word) entries with higher evaluation scores are selected to be added to the lexicon. This method is successfully applied to an Uyghur large-vocabulary continuous speech recognition system, resulting in a significant reduction of WER and the lexicon size. Further improvement is achieved by combining with a statistical method based on mutual information criterion.

Index Terms— speech recognition, language model, discriminative learning, Uyghur, morpheme

1. INTRODUCTION

In agglutinative languages, selection of lexical unit is not obvious and one of the important issues in designing language model for automatic speech recognition (ASR). There is a trade-off between word unit and 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. Therefore, the morpheme unit is conventionally adopted in many agglutinative languages, such as Japanese [1], Korean [5], and Turkish [9]. However, most of morphemes are short, often consisting of one or two phonemes, thus they are more likely to be confused in ASR than the word unit. The goal of this study is to incorporate effective word (or sub-word) entries selectively while maintaining the high coverage of the morpheme unit.

There are a number of previous works addressed on this problem, and many of them are based on statistical measures, such as co-occurrence frequency, mutual information, and likelihood [4]-[9]. However, these criteria are not directly related to WER (word error rate).

In this paper, we propose a discriminative approach to select word (or sub-word) entries which is likely to reduce the WER. It is realized by aligning and comparing the ASR results by the morpheme-based model with those by the word-based model. We describe each word by a set of features, and define an evaluation function with their weights. Then, the weights are learned to select "critical" word entries. This learning mechanism, which leads to reduction of WER, is applicable to any unseen words, or even sub-words.

The proposed method is applied to and evaluated in a large-vocabulary Uyghur ASR system. Several features are investigated and compared in terms of WER and the lexicon size. Moreover, the method is compared and combined with a statistical method based on mutual information. Although there are a number of works on discriminative learning for language models such as n-gram [10]-[12], there is no prior work on the use of discriminative learning for lexicon optimization.

2. CORPUS AND BASELINE SYSTEM

We have developed an Uyghur-language large-vocabulary continuous speech recognition (LVCSR) system [2]. 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 syntactic and semantic functions, which have 305 surface forms. A few words have a (only one) prefix preceding a stem; seven kinds of prefixes are considered.

For language modeling, a text corpus of 630K sentences is collected over general topics from newspaper articles, novels, and science textbooks. The sentences are segmented to morpheme and word units by our morphological analyzer [3].

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 held-out test data set is prepared from reading of newspaper articles. Specifications of the data sets are summarized in Table 1.

Table 1. Statistics of speech corpus

corpus	sentences	persons	total utterances	time (hour)
training	13.7K	353	62K	158.6
test	550	23	1468	2.4

Two different lexical units (word and morpheme) are used to build n-gram (3-gram and 4-gram) language models, and their ASR performance is compared in Table 2. The cutoff threshold also controls the lexicon size and ASR performance. Cutoff-F means that units with frequency less than F times are disregarded and treated as unknown. It is observed that the word-based model outperforms the morpheme-based models with a much bigger lexicon 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.

On the other hand, the morpheme-based model is benefited from a much smaller vocabulary size, thus 4-gram language model performs better than the 3-gram model. In the following experiments, we use the morpheme 4-gram model with cutoff-5 as a baseline as the difference from the cutoff-2 case is not statistically significant.

Table 2. ASR results for different baseline units

Models	WER (%)		lexicon size	
	Cutoff-2	Cutoff-5	Cutoff-2	Cutoff-5
Morph (3-gram)	28.96	29.17	55.2K	27.4K
Morph (4-gram)	27.92	28.11		
Word (3-gram)	25.77	26.64	229.8K	108.1K
Word (4-gram)	25.93	27.05		

3. DISCRIMINATIVE LEARNING FOR LEXICON OPTIMIZATION

The proposed discriminative approach to lexicon optimization is realized by comparing the ASR results by the morpheme-based model and those by the word-based model. The results are aligned by word with corresponding morpheme sequences. We assume each word is composed of one or more morphemes, and morpheme units do not cross word boundaries.

In majority of the differences between these two ASR results, the word-based model gives correct hypotheses while the morpheme-based model does not, as suggested by the result of Table 2. A naïve method would be to pick up these "critical" word entries to be added to the lexicon. When conducted in the closed test-set, it would result in a drastic improvement in ASR. However, the method heavily depends on the training data set since it can select only entries observed there, and thus does not have a generality.

3.1. Evaluation Function of Words with Lexical Features

In this work, we formulate a generalized scheme by introducing a set of lexical features. Each word w is described by a set of features of the constitute morphemes $\Phi_s(w)$ ($w = m_1 m_2 \ldots$). We assume that they are binary (1 for true, 0 for otherwise). Then, we define an evaluation function as a linear weighted sum of the features.

$$f(w) = \sum_s \Phi_s(w)\, \alpha_s = \boldsymbol{\Phi}(w)\, \boldsymbol{\alpha}$$

Here, α_s is a weight for the feature $\Phi_s(w)$. The above function indicates the potential importance of the word to be included in the lexicon, or how likely WER will be reduced by adding this word entry. Note that this function can be computed for any words or even sub-words consisting of morpheme sequences, so that we can select effective entries which would not be correctly recognized by the morpheme-based model.

3.2. Weight Estimation with Discriminative Learning

The values of the weights $\boldsymbol{\alpha} = \{\alpha_s\}$ are estimated based on discriminative learning using the training data set. In this work, we adopt a simple perceptron algorithm [12], since the evaluation function is linear. The standard sigmoid function is introduced to map the above evaluation score to the 0-1 range.

$$g(w) = \frac{1}{1 + e^{-\mathrm{f(w)}}}$$

$$g'(w)|_{f(w)} = g(w)(1 - g(w))$$

The desired output $d(w)$ is defined as binary, corresponding to the CRITICAL_CASE in which the word-based model outputs the correct hypothesis while the morpheme-based model does not.

$$d(w) = \begin{cases} 1 & \text{if CRITICAL\_CASE is true} \\ 0 & \text{otherwise} \end{cases}$$

Then, the weight vector is updated as:

$$\boldsymbol{\alpha} = \boldsymbol{\alpha} + \eta\, g'(w)(d(w) - g(w))\boldsymbol{\Phi}(w)$$

The learning rate parameter η is adjusted at every iteration to prevent excessive fluctuation. Here we simply reduce it by a factor of 10. This learning converges in several iterations, and we terminate at the third iteration in the experiments.

3.3. Filtering Training Samples

The simple perceptron algorithm is not robust against noisy or outlier samples. Thus, we introduce filtering so that only reliable samples are fed to the training. Specifically, we selectively use the samples whose frequency of the error pattern is more than N times over the entire training data set.

4. LEXICAL FEATURES

In this Section, we list the lexical features considered in this work.

4.1. Word ID

This corresponds to a naïve method which matches only word entries. It also makes a constant feature for all word entries, i.e. always becomes 1.

$$\Phi_{\text{word}}(w)=\begin{cases} 1 & \text{if } w=word \\ 0 & \text{otherwise} \end{cases}$$

4.2. Morpheme Length

Short units are easily confused in ASR and they are very frequent. Actually, there are many suffixes consisting of only one or two phonemes. Confusion in short morphemes can be reduced by merging and making them longer. The feature counts the length of the constitute morphemes.

$$\Phi_{\text{length}}(w)=\begin{cases} 1 & \text{if } length\ of\ m_i \le 2 \\ 0 & \text{otherwise} \end{cases}$$

4.3. Morpheme N-gram

Here, we focus on typical morpheme entries and their bigram patterns. A specific weight α is estimated for every unigram or bigram entry.

$$\Phi_{\text{unigram}}(w)=\begin{cases} 1 & \text{if } morph.m_i\ exists\ in\ w \\ 0 & \text{otherwise} \end{cases}$$

$$\Phi_{\text{bigram}}(w)=\begin{cases} 1 & \text{if } morph.bigram\ (m_i\ m_j)\ exists\ in\ w \\ 0 & \text{otherwise} \end{cases}$$

4.4. Morpheme Attributes

We also categorize morphemes into stems and word-endings which are a sequence of suffixes.

$$\Phi_{\text{stem}}(w)=\begin{cases} 1 & \text{if } stem\ exist\ in\ w \\ 0 & \text{otherwise} \end{cases}$$

$$\Phi_{\text{word-ending}}(w)=\begin{cases} 1 & \text{if } word-ending\ exists\ in\ w \\ 0 & \text{otherwise} \end{cases}$$

5. LEXICON DESIGN

These features are then generalized to all words in the text corpus for language model training. We calculate the evaluation score $f(w)$ for the morpheme sequence of every word. If the value is larger than the threshold of 0.5 (or $g(w) \approx 1$), then the word entry w is added to the lexicon.

Furthermore, the method can be applied to sub-words, which is composed of morpheme sequences within a word, except for the word ID feature. Specifically, we try to search for sub-word entries that satisfy the lexical features and $f(w) > 0.5$. The search is exhaustively done from the beginning of all words by concatenating the following morphemes while the condition is met. If the condition is not met, the search is re-started there.

For comparison, we also investigate the data-driven methods based on the following statistical measures.

(1) Co-occurrence frequency

A simple model based on statistical co-occurrence is built by merging the frequent morpheme sequences (FMS). Specifically, we count the morpheme bigram co-occurrence frequency $C(m_i\, m_j)$, and concatenate them if the frequency is higher than a threshold. The concatenation process is repeated to a sequence of morphemes, just like the sub-word model described above, except that the concatenation can be made even across the word boundaries.

(2) Mutual information

Another statistical measure is mutual information (MI) [4]. It is calculated as a geometrical mean of forward and reverse bigrams as below. The concatenation process is same as the case of the co-occurrence frequency.

$$MI(m_i\, m_j)=\sqrt{\mathrm{P_f}(m_i|m_j)\mathrm{P_r}(m_j|m_i)}=\frac{\mathrm{P}(m_i\,, m_j)}{\sqrt{\mathrm{P}(m_i)\mathrm{P}(m_j)}}$$

6. EXPERIMENTAL EVALUATION

The method has been implemented and applied to our Uyghur LVCSR system described in Section 2. The data set for acoustic model training is used for the proposed discriminative learning of lexical entry selection, and the same test set as in Section 2 is used for evaluation. Once the lexicon is prepared by adding the word or sub-word entries, 4-gram language model is trained again, and the entire test data are decoded again using the new model.

First, we investigate the effect of sample filtering described in Section 3.3. The WERs obtained by changing the threshold (N) values are listed in Table 3. We can see that removing outlier samples of only one occurrence is effective, and the accuracy is stable unless we discard too many samples ($1 \le N \le 4$). In the following experiments, we use $N = 2$.

The effect of individual features listed in Section 4 in the proposed scheme is compared in Table 4. Although the length feature alone is not so effective because of its simplicity, all other features lead to significant improvement from the baseline morpheme model (WER=28.11%), and the accuracy is comparable to the best word-based model with Cutoff-2 (the WER difference among these methods are not statistically significant). Note that the lexicon size of the enhanced morpheme-based model is much smaller than the word-based model (230K with Cutoff-2), and still expected to give broad coverage. Combinations of these features are also explored, but little additional gain is obtained due to the redundancy of these features.

We also generate a sub-word lexicon by using the morpheme N-gram features. The result in Table 5 shows that this method reduces both WER and the lexicon size significantly. The bigram-based sub-word model outperforms the best word-based model in accuracy with the lexicon size of one fourth.

Then, this method is compared with the two conventional statistical models: FMS (frequent morpheme sequence) and MI (mutual information). Note that these methods including the proposed bigram-based model concatenate a sequence of morphemes, but the criterion of the concatenation is different. The results by varying respective threshold values are listed in Tables 6 and 7. It is observed that our proposed method is slightly better than the best results by these methods. Moreover, the tuning of the threshold values for these methods are not so straight-forward, depending on the task and database, while our proposed method does not have any sensitive parameters.

Finally, we investigate the combination of the proposed method with the statistical method. Here, we adopt a tandem approach; first apply the proposed discriminative method, and then apply the best MI-based method. Lexicon entries are added by each step. The results are summarized in Table 8. The simple combination results in drastic improvement in accuracy, 1% absolute compared with the best word-based model. The result shows the discriminative model has a synergetic effect with the statistical model.

Table 3. Effect of sample filtering threshold (WER %)

threshold	N=0	N=1	N=2	N=3	N=4	N=5
unigram	26.69	25.93	25.87	26.18	26.28	26.54

Table 4. Comparison of features in word selection

Feature	WER (%)	lexicon size
word	26.18	35.8K
length =1	27.07	32.4K
length ≤ 2	27.08	35.1K
unigram	25.87	74.8K
bigram	25.99	67.3K
stem	26.10	92.7K
word-ending	26.20	92.1K
stem & word-ending	25.96	82.3K

Table 5. Result of sub-word selection

feature	WER (%)	lexicon size
unigram	25.96	40.7K
bigram	25.27	49.9K

Table 6. Result of frequent morpheme sequence (FMS) method

threshold	2000	2500	3000	3300
lexicon size	57.1K	50.7K	44.8K	42.3K
WER (%)	27.02	26.63	26.68	26.76

Table 7. Result of mutual information (MI) method

threshold	0.030	0.035	0.040	0.045	0.050	0.06
lexicon size	69.1K	60.0K	53.3K	47.0K	41.9K	36.1K
WER (%)	25.83	25.61	25.60	25.79	25.80	26.07

Table 8. Result of combination

Methods	WER (%)	lexicon size
proposed bigram sub-word	25.27	49.9K
mutual information	25.60	53.3K
combined model	24.75	67.8K

7. CONCLUSION

We have proposed a novel discriminative approach to lexicon optimization for agglutinative languages. It adopts the same scheme as the conventional statistical approach which starts with the morpheme-based model and search for effective word or sub-word entries to be added. However, the proposed discriminative learning is directly linked to the improvement of ASR accuracy. In the experimental evaluations, the proposed method achieves the best accuracy in comparison, and further improvement by combining with the statistical method, resulting in a significant gain from the baseline morpheme-based model and the word-based model without a large increase in the lexicon size.

8. REFERENCES

[1] T. Kawahara et al. Free software toolkit for Japanese large vocabulary continuous speech recognition. In Proc. ICSLP, Vol.4, pp.476--479, 2000.

[2] M. Ablimit, G. Neubig, M. Mimura, S. Mori, T. Kawahara, A. Hamdulla. Uyghur Morpheme-based Language Models and ASR. In Proc. IEEE-ICSP, 2010.

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

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

[5] 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.

[6] O-W. Kwon, Performance of LVCSR with morpheme-based and syllable-based recognition units. In Proc. ICASSP, pp.1567-1570, 2000.

[7] M. Jongtaveesataporn, I. Thienlikit, C. Wutiwiwatchai, S. Furui. Lexical units for Thai LVCSR. Speech Communication, pp.379~389, 2009.

[8] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, M. Creutz. *On Lexicon Creation for Turkish LVCSR*. In Proc. Eurospeech, 2003.

[9] Ebru Arisoy, Hasim Sak, Murat Saraclar. Language Modeling for Automatic Turkish Broadcast News Transcription. Proc. Interspeech, 2007.

[10] Brian Roark, Murat Saraclar, and Michael Collins. Discriminative n-gram language modeling. *Computer Speech and Language*, 21(2):373-392, April 2007.

[11] M. Collins, B. Roark, M. Saraclar. Discriminative syntactic language modeling for speech recognition. In Proc. ACL, pages 507-514, 2005.

[12] M. Collins. Discriminative training methods for HMMs: Theory and experiments with perceptron algorithm. In Proc. EMNLP 2002.