Vol.2011-SLP-87 No.5
2011/7/21

Lexicon Optimization for Automatic Speech Recognition based on Discriminative Learning

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

In agglutinative languages such as Japanese and Uyghur, selection of lexical unit is not obvious and one of the important issues in designing language model for automatic speech recognition (ASR). In this paper, we propose a discriminative learning method to select word entries which would reduce the word error rate (WER). 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 two units (morpheme and word). Then, the weights of the features are learned by a perceptron algorithm. Finally, word entries with higher evaluation scores are selected. The discriminative method is successfully applied to an Uyghur large-vocabulary continuous speech recognition system, resulting in a significant reduction of WER without a drastic increase of the vocabulary size.

識別学習に基づく音声認識単語辞書の最適化

アブリミテ・ミジテ[†]，河原達也[†]，ハムヅラ・アスカ[††]

日本語やウイグル語のような膠着言語では、単語の単位が自明でなく、音声認識の言語モデルの設計においても重要な問題となっている。本稿では、音声認識誤り（単語誤り率）を削減するような単語エントリを識別学習により選択する方法を提案する。各単語エントリに対して素性の集合とそれらの重みからなる評価関数、及び、形態素単位のモデルと単語単位のモデルの誤り率の差による誤分類尺度を定義した上で、パーセプトロン学習によって素性の重みを学習する。本手法をウイグル語の大語彙連続音声認識システムに適用し、形態素単位のモデルに比べて語彙サイズをあまり増やすことなく、単語誤り率を大きく削減することができた。

1. Introduction

In agglutinative languages such as Japanese and Uyghur, 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, morpheme unit is conventionally adopted in agglutinative languages. But morphemes are short, often consisting of one or two phonemes, thus they are more likely to be confused 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. This approach was often investigated with the word frequency basis or likelihood criterion [7][8][9], but it does not necessarily lead to better ASR performance.

In this paper, we propose a discriminative learning method to select word entries which is likely to reduce the word error rate (WER). 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 word-based model. Then, the weights of the features are learned by a perceptron algorithm. Finally, word entries with higher evaluation scores are selected. The discriminative method is applied to an Uyghur large-vocabulary continuous speech recognition (LVCSR) system,

The remainder of the paper is organized as follows: we first review the problem of lexicon design for Uyghur speech recognition, with the description of the corpus and the baseline ASR system in Section II. Then, the proposed method is described in Section III and its evaluation is presented in Section IV.

2. Corpus and baseline systems

We have developed an Uyghur-language large-vocabulary continuous speech recognition system [1]. For language modeling, a text corpus of 630k sentences is collected from general topics like newspaper articles, novels, and general science textbooks. They are segmented to morpheme units and word units by our morphological analyzer.

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

† 京都大学情報学研究科
Kyoto University, School of Informatics
†† 新疆大学信息学院
Xinjiang University, Information Institute

test data set is also prepared from newspaper articles. Specifications of the data sets are summarized in Table 1. Comparison of the baseline morpheme and word-based models are shown in Table 2.

Table 1. Statistics of speech corpora.

Corpus	Sentences	Person	Total utterance	Time (hour)
traing	13.7K	353	62k	158.6
test	550	23	1468	2.4

Table 2. Comparison of morpheme and word units.

	Morpheme unit		Word unit	
	training set	test set	training set	test set
1-gram coverage	98.4	99.7	94.5	97.2
2-gram coverage	94.5	96.4	75.5	71.0
3-gram coverage	82.0	81.6	47.4	31.3
perplexity	82.8	52.2	2436	2497

Table 3. ASR error rates by baseline models.

LM	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

Julius is used to build an ASR system [1]. 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.

The ASR results by various morpheme and word-based models are summarized in Table3. A word boundary symbol is added to the results of the morpheme-based models to compute WER.

It is observed that 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.

In order to improve the morpheme-based model performance, rule-based morpheme concatenation and statistical concatenation methods have been investigated in Korean and Thai [7][8]. The rule-based method concatenates short and frequent morphemes. Statistical methods concatenate the morphemes based on a frequency or likelihood criterion. But these methods do not necessarily lead to ASR improvement.

3. Proposed method

In this paper, we propose a novel approach for lexicon optimization based on discriminative learning. The proposed method selects word entries which are more likely to reduce WER. This is realized by comparing the ASR results by the morpheme-based model and those by the word-based model. A naive method can be implemented by selecting entries which are mis-recognized by the morpheme-based model, but correctly recognized by the word-based model. However, this naive method does not have generality; it cannot include entries that are not in the training data set. In a preliminary experiment, this method reduced the WER dramatically for the closed (training) data set, but not so much for the test data set.

Therefore, we adopt a more generalized scheme. In this scheme, we describe each word by a set of features of the constitute morphemes $\Phi(m)$. In this work, we assume that they are binary (1 for true, 0 for otherwise). Morphological features include phonological (length, phone category), lexical (stem/suffix), or syntactic (part-of-speech). The details are discussed in Section 3.3. Then, we define an evaluation function as a linear weighted sum of the features.

$$f(w) = \sum_s \Phi_s(m)\, \alpha_s$$

Here, $\boldsymbol{\alpha}$ is a set of weights for the features. The 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.

Then, values of the weights $\boldsymbol{\alpha}$ are estimated based on discriminative learning using the training data set. In this work, we adopt a simple perceptron algorithm, since the evaluation function is linear. As a result of this learning, we can compute evaluation scores for all words. Then, we select word entries which have a larger score than a threshold.

The detailed formulation and procedures are explained in the following subsections.

3.1 Parameter estimation for discriminative learning

Morpheme-based and word-based ASR results are obtained on the training data. 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. The objective is to find word entries to reduce WER with a minimum increase of the vocabulary size.

$\Phi(m)$ is an arbitrary binary feature vector of aligned pairs $w \leftrightarrow m$ of the word-unit w and the morpheme-unit sequences ($m = m_1 m_2 \cdots$). $f(w)$ is an evaluation function from the morpheme sequence m for a word w, and $\boldsymbol{\alpha}$ is a weight vector.

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

Each weight parameter α_s is bound with its feature $\Phi_s(m)$.

The desired output $d(w)$ is defined by the difference of the two ASR results: $d(w) = 1$ if m is not matched with w. Both estimated output $f(w)$ and desired output $d(w)$ are regarded as binary numbers.

$$d(w) = \begin{cases} 1 & \text{if } w \neq m \\ 0 & \text{otherwise} \end{cases}$$

$$f(w) = \begin{cases} 1 & \text{if } f(w) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Then, the learning is conducted as follows:

●Initialize: Set $\boldsymbol{\alpha} = 0$

●**Repeat:** For every aligned pairs $w \leftrightarrow m$ and desired output $d(w)$.

✧ Calculate $f(w) = \Phi(m) \cdot \boldsymbol{\alpha}$

✧ If $d(w) \neq f(w)$ then $\boldsymbol{\alpha} = \boldsymbol{\alpha} + \beta(d(w) - f(w))\Phi(m)$

●Output: parameters $\boldsymbol{\alpha}$

A smaller step β can be used to prevent fluctuation around the converging point. This learning quickly converges in one or two iterations.

There are alternative ways of the parameter estimation such as maximum entropy and log linear model. Here we use a simple perceptron algorithm to demonstrate the feasibility of the scheme.

3.2 Lexicon and language model design

These features are then generalized to all words in the morpheme-based text corpus by setting a threshold for the evaluation function $f(w)$. If the evaluation score $f(w)$ is larger than a certain threshold, then the word entry w is added to the lexicon.

Experimental results show that a smaller threshold ($th = 0.1$) would be an optimal.

N-gram language models can be built with the new lexicon using a certain cutoff threshold. Cutoff-F means that units frequency less than F times are disregarded and treated as unknown. The parameter also controls the vocabulary size and ASR performance, which are evaluated in the experiments.

3.3 Features considered

In the proposed scheme choosing proper features is very important. The features should contribute to reducing WER and the vocabulary size. Features in consideration are morpheme length, misrecognized morpheme error frequency, and morpheme context.

①In ASR, short units are easily confused than long units. In Uyghur language, there are many suffixes consisting of only one or two phonemes. We follow this general idea and try to concatenate all single phoneme morphemes as in [6] either to each other or to neighboring morphemes.

$$\Phi_s(m) = \begin{cases} 1 & if \;\; m \text{ is of one phoneme} \\ 0 & otherwise \end{cases}$$

②We can consider a more direct feature to account for the error rate, by computing the error frequency of the morphemes. We compute this feature separately for stems and word-endings.

$$\Phi_s(m) = \begin{cases} 1 & if \;\; m \text{ is misrecognized more than 10 times} \\ 0 & otherwise \end{cases}$$

③We also consider typical patterns of sub-word morphemes and morpheme bigrams. The feature vector $\Phi(m)$ can be defined for any morpheme sequences. This makes a large number of features, but we focus on mis-recognized patterns. Below is an example of bigram morpheme feature.

$$\Phi_s(m) = \begin{cases} 1 & if\ bigram\ (m_i\ m_j)\ \text{exist in}\ m \\ 0 & otherwise \end{cases}$$

Table 4. ASR results with bigram features.

Models	WER (%)		Vocabulary size	
	Cutoff-2	Cutoff-5	Cutoff-2	Cutoff-5
Baseline morpheme (4-gram)	27.92	28.11	55.2k	27.4k
Baseline word (3-gram)	25.72	26.64	227.9k	108.1k
Bigram feature-based discriminative model (4-gram)	25.57	25.64	68.3k	40.2k

Table 5. ASR results using mutiple features.

Features	WER (%)	Vocabulary size
baseline	28.11	27357
stem & word endings	26.56	55.8k
word endings	26.36	101.6k
bigram	25.64	40.2k
all features	25.98	110.1k

4. Experimental evaluation

The proposed method is applied to the Uyghur LVCSR task. Uyghur belongs to the Turkish language family of the Altaic language system. The morpheme structure of Uyghur word 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.

Among the features, the bigram morpheme feature is found to be most effective. We present the results for different cutoff threshold values in Table 4. The proposed method outperformed both of the baseline models, with a significantly smaller vocabulary size than the word-based model. Models with high cutoff-5 did not degrade WER.

As the vocabulary size of the proposed model is comparable to the baseline morpheme model, the order of N-gram can be increased. Actually, 4-gram language model has the best ASR performance. Training with n-best list would improve the performs by around 0.3%

We also investigate several features listed in Section 3.4. Cutoff-5 parameter is used in all the experiments. Most of the features are effective. However, when the bigram feature is used, combination with other mentioned features does not bring any improvement.

5. Conclusion

We proposed a discriminative learning method for lexicon optimization by selecting word entities. The proposed method significantly reduced WER from the morpheme-based model without a drastic increase of the lexicon size compared with the word-based model.

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 supervisedUyghur 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.
13) B.Roark and M.Saraclar and M.Collins. Discriminative N-gram Language Modeling. Computer Speech & Language, Vo.21, No.3, pp.458-478, 2007