論文 / 著書情報
Article / Book Information

Title	A New Lexicon Optimization Method for LVCSR Based on Linguistic and Acoustic Characteristics of Words
Author	Takahiro Shinozaki, Sadaoki Furui
Journal/Book name	7th International Conference on Spoken Language Processing (ICSLP-2002), Vol. , No. , pp. 717-720
発行日 / Issue date	2002, 9

A NEW LEXICON OPTIMIZATION METHOD FOR LVCSR BASED ON LINGUISTIC AND ACOUSTIC CHARACTERISTICS OF WORDS

Takahiro Shinozaki and Sadaoki Furui

Tokyo Institute of Technology
Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan.
{staka, furui}@furui.cs.titech.ac.jp

ABSTRACT

This paper proposes a new lexicon optimization method to improve recognition rate of large scale spontaneous speech recognition. Occurrence count and length of a word has strong correlation with difficulty of recognizing the word. First, we investigate the relation and make a word correctness probability model. The proposed method optimizes the lexicon by making compound words or phrases step by step based on the word correctness probability model so as to improve the estimated recognition rate of the system. The optimization method is applied to a large scale Japanese spontaneous speech corpus. Experimental results show that the language model using the optimized lexicon improves the recognition rate.

1. INTRODUCTION

Language models based on words or morphemes are widely used for large-vocabulary continuous-speech recognition (LVCSR). For languages like English, words are well defined since they are separated by space symbols in the written text. Other languages like Japanese have no spacing between words and even no clear definition of words. Therefore, for these languages, it is common to preprocess text with a morphological analysis program to automatically split sentences into morphemes to make language models. In all these languages including English and Japanese, it is not clear whether these conventional words or morphemes are optimal units for speech recognition.

From this point of view, several studies have been conducted to optimize the lexicon for improving the performance of language models[1, 2, 3, 4, 5, 6, 7, 8]. Ideas of these works include a) automatically building a lexicon based on some criteria/rules for languages having no clear word definition, b) concatenating word pairs to model longer context by N-grams without increasing N so as not to increase the parameter dimension and data sparsity, and c) concatenating words to balance the occurrences of all the word units. In these methods, basic units are concatenated based on evaluation functions such as unit pair frequency and mutual information. In [1, 2, 3, 4, 5, 7], performance was evaluated in terms of the recognition accuracy as well as the test-set perplexity. In [2], it was reported that certain word phrases were very frequent in dialogs of a limited domain. A phrase finding algorithm based on the mutual information criterion was found to improve the accuracy of a recognition system using a bigram language model. Paper [5] reported a mutual information based method using a large newspaper corpus with no improvement being achieved in the recognition accuracy. Paper [7] reported that a word pair frequency based method improved the recognition accuracy by 0.2% using a language model trained on WSJ.

One of the problems of these methods is that they are based on only linguistic aspect and no acoustic characteristics have been considered. Although perplexity is a useful measure, decrease in perplexity does not necessarily guarantee the improvement in the recognition rate.

In this paper we propose a new statistical lexicon optimization method for speech recognition based on both linguistic and acoustic features of words. This study is motivated by our recent observation that less frequent and shorter words are generally difficult to recognize [9]. In the proposed method, a word correctness probability model is estimated and used to directly estimate the recognition correctness of a system. A process of choosing and concatenating a word pair which maximizes the estimated word correctness is iterated.

This paper is organized as follows. In Section 2, an experimental set up is described. In Section 3, the relationship between word frequency, word length and recognition correctness is analyzed and modeled. The optimization method is described in Section 4. The proposed method is applied to a large scale Japanese spontaneous speech corpus in Section 5. It is shown that the method improves the recognition rate. Experimental results are discussed in Section 6. Finally in Section 7, the paper is summarized and concluded.

2. EXPERIMENTAL SET UP

2.1. Baseline recognition system

Language models are trained using transcriptions of 610 academic and non-academic lectures in the large-scale Japanese spontaneous speech corpus (CSJ)[10]. JTAG morphological analysis program is used to convert Japanese text into morpheme sequences. The training set turns out to have approximately 1.5M morphemes (morpheme will be called "word" hereafter in this paper). The most frequent 30k words are selected as the vocabulary for recognition and a trigram language model is made.

An acoustic model is made using 338 CSJ lectures presented by male speakers. The total length is approximately 59 hours. The Julius 3.1 decoder is used for speech recognition.

2.2. Recognition task

Two kinds of utterance set are used, both of which consist of academic lectures presented by male speakers in the CSJ.

Table 1. Development set

Conference name	No.lecture
Jap. Soc. Artif. Intell.	32
Acoust. Soc. Jap.	9
Soc. Jap. Linguistics	3

Table 2. Evaluation set

Lecture ID	Conference name	Length [min]
A22	Acoust. Soc. Jap.	28
A23	Acoust. Soc. Jap.	30
A97	Acoust. Soc. Jap.	12
J01	Soc. Jap. Linguistics	57
K05	National Lang. Res. Inst.	42
N07	Assoc. Natural Lang. Proc.	15
P25	Phonetics Soc. Jap.	27
S05	Assoc. Socioling. Sciences	23
Y01	Spont. Speech Corpus Meeting	14
Y05	Spont. Speech Corpus Meeting	15

Development set, consisting of 44 lectures, is used for analyzing and modeling the word correctness probability. Table 1 shows the content, in which the first 10 minutes of each lecture is used.

Evaluation set is used for evaluating our method. This consists of 10 lectures having no overlap with those in the development set. Table 2 shows its content, in which the entire length of each lecture is used. The OOV rate is 1.3% excluding word fragments.

All the speakers in these sets have no overlap with those in the training set for building acoustic and basic language models. The language model weights and the insertion penalties are adjusted to maximize the recognition accuracy for each of these sets.

3. RELATIONSHIP BETWEEN WORD OCCURRENCE COUNT, WORD LENGTH AND RECOGNITION CORRECTNESS

There exist many factors that affect the difficulty of recognizing each word. Among them, it has been observed that the number of occurrences of a word in the language model training set as well as it's length have a strong relationship with it's correctness[9]. Generally speaking, less frequent and shorter words are harder to recognize. This is probably because N-gram probability of less frequent word is more difficult to model correctly and because shorter words are acoustically more confusable in the decoding process.

To investigate these relationships, we define and calculate word attributes as follows.

Cor: Word correctness (%). (0 or 100).

WF: Number of occurrences of a word in the language model training set.

LF: Logarithmic value of the *WF*; $\log_{10}(WF + 10^{-6})$.

NP: Number of phonemes in the word.

Cor is a binary attribute, taking 100 if the word is recognized correctly and 0 if it is miss-recognized by the decoder. Insertion errors are not considered since they do not have corresponding reference words.

These attributes are computed for each word in the development set. The dotted lines in Figure 1 show the relationship between the number of occurrence of words and the recognition correctness where words are classified according to the number of phonemes they contain. The correctness is averaged over the group of words having the same range of *NP* and shown in the figure at each value of *LF*. Similarly, Figure 2 shows the relationship between the number of phonemes and the recognition correctness.

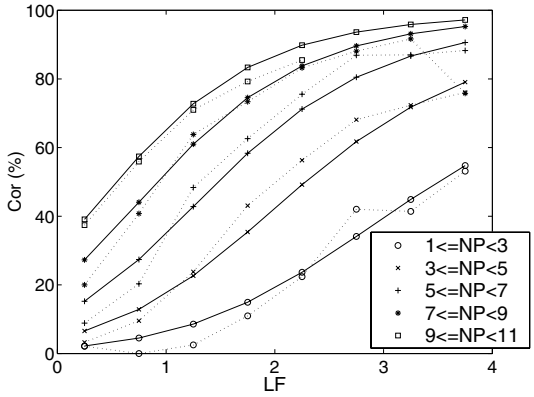

Fig. 1. Log word occurrence count and recognition correctness.

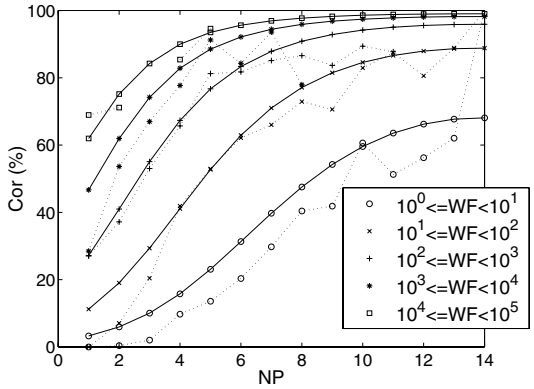

Fig. 2. Word length and recognition correctness.

From the Figures 1 and 2, it can be seen that the averaged word correctness largely changes according to the number of occurrences and the length of the word. The results are approximately continuous and monotonous.

The averaged word correctness can be regarded as an expected probability that a word is successfully recognized as a function of the attributes. We modeled the probability by a logit model having the logarithmic occurrence count (LF) and the number of phonemes (NP) as explanation variables. Second order terms of the attributes were also incorporated to improve the modeling accuracy. Equation (1) shows the obtained model in which parameters were estimated by the maximum likelihood method.

$$P\left(Cor = 100|NP, LF\right) = \Lambda\left(0.70NP - 0.03NP^2 + 1.60LF - 0.12LF^2 - 5.17\right) \quad (1)$$

where Λ is a logistic function. Λ is expressed as follows.

$$\Lambda\left(x\right) = \frac{e^x}{1 + e^x}. \quad (2)$$

Solid lines in the Figures 1 and 2 indicate the probability estimated by the logit model (1), which show that the logit model successfully indicates the global characteristics of the correctness.

4. LEXICON OPTIMIZATION METHOD

In the previous section, it was shown that the word correctness probability was effectively modeled by the logit model as a function of the word occurrence count and the word length. In this section, we propose a new recognition lexicon optimization method based on the model, in which the lexicon is optimized by concatenating basic words iteratively.

We approximate the occurrence of words in the parent set of the recognition task by the occurrence of words in the training set for building the language model. By using this assumption and the word correctness probability model, a recognition rate of the system can be estimated as an expected value of word correctness as formulated in the equation(3), in which $\alpha(w)$ is a "recognition rate normalization factor". This term is introduced to impartially compare recognition systems having different recognition units. Suppose we have two recognition systems, one has "ice" and "cream" as separate words and the other has "ice+cream" as a single word for recognition. If "peach ice cream" is spoken and recognized as "beach ice cream" and "beach ice+cream" by the two systems, these two results are basically same but their correctness values are 2/3 and 1/2 respectively. This inconsistency can be alleviated by weighting the words when calculating the recognition rate. The number of initial words before concatenation or the number of characters in the word can be assignment to $\alpha(w)$. The former assignment corresponds to using the word recognition rate based on the initial words and the latter assignment corresponds to using the character recognition rate. The character recognition rate has been sometimes used in speech recognition evaluation for the languages using Chinese characters, such as Japanese and Chinese.

Let's consider selecting a word pair $\langle w_1, w_2 \rangle$ and concatenating all the sequence of these words in this order to create a new word $w_{1,2}$ in the language model training set. New attributes after the operation can be expressed as follows.

- $NP(w_{1,2}) = NP(w_1) + NP(w_2)$
- $WF(w_{1,2})$ = The number of sequences "$w_1\ w_2$" in the original training set.

The attributes of w_1 and w_2 also need to be updated as follows.

- $NP\left(w_i^{'}\right) = NP(w_i)$
- $WF\left(w_i^{'}\right) = WF(w_i) - WF(w_{1,2})$

where $w_i^{'}$ denotes the w_i ($i = 1$ or 2) after the operation. Note that by definition $\alpha(w_{1,2}) = \alpha(w_1) + \alpha(w_2)$.

We define a delta() evaluation function of word pair concatenation as the difference of the estimated correctness of the recognition system before and after the operation as shown in the equation (4). We select a word pair among all possible word combinations which maximizes the evaluation function and concatenate those word sequences in the training set. The process is iterated by choosing a new word pair step by step. The optimization process is summarized by optlexicon() as shown below.

```
procedure optlexicon() {
 for i = 1:maxiter {
   select word pair <w1,w2>
     which maximizes delta(w1,w2);
   break if delta(w1,w2) < 0;
   merge(w1,w2);
 }
}
```

5. EXPERIMENTAL RESULTS

5.1. Application to the training set

The optimization method was applied to the training set of the baseline language model. The initial vocabulary size, the number of different words, in the training set was 34,895. $\alpha(w)$ was initialized as the number of characters in each word. After iterating the concatenation process for 500 times, the training set which had 1.5M words at the beginning was reduced to 1.3M words. The estimated α-weighted correctness gain was 1.39%, which means that by using the new language model trained using the optimized training set ideally improves the character correctness by 1.39% in the absolute value.

Figure 3 shows the concatenation evaluation score and its accumulated value. The evaluation score decreases exponentially as a function of the number of iterations.

Table 3 shows the mean and standard deviation of the attributes calculated for the training set before and after the optimization. The averaged number of phonemes increases and the averaged word frequency decreases as the result of the optimization.

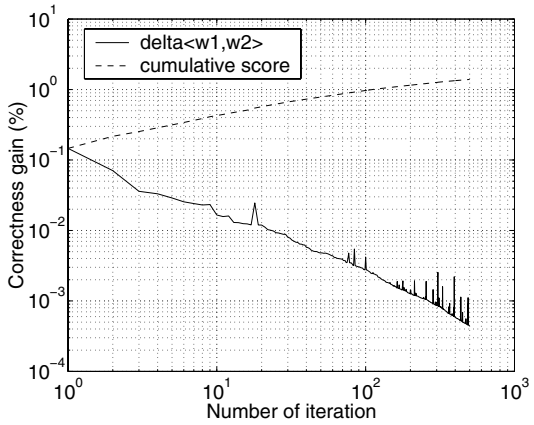

Fig. 3. Changes of the evaluation and accumulated values in 500 iterations.

5.2. Recognition results

A trigram language model having 30k words was trained using the optimized training set and used for speech recognition. Other conditions were the same as the baseline system. Recognition performance for the evaluation set using the new language model was compared with the baseline system. The language model weights and the insertion penalties were optimized for each system. The

$$E\left[Cor\right] = \frac{\sum_w P\left(Cor = 100 | NP\left(w\right), LF\left(w\right)\right) \cdot WF\left(w\right) \cdot \alpha\left(w\right)}{\sum_w WF\left(w\right) \alpha\left(w\right)}, \quad (3)$$

$$\Delta\left(w_1, w_2\right) = E_{after}\left[Cor\right] - E_{before}\left[Cor\right]. \quad (4)$$

Table 3. Mean and standard deviation of word attributes before (*base*) and after (*opt*) the optimization

	NP	*WF*	*LF*
Mean(*base*)	3.79	13541	3.23
Standard deviation(*base*)	2.14	19510	1.20
Mean(*opt*)	4.39	7829	2.94
Standard deviation(*opt*)	2.51	13637	1.13

performance measured by the character correctness and accuracy is shown in Figure 4.

The figure shows that the correctness and the accuracy were improved for nine and eight lectures, respectively, out of 10 lectures. Averaged improvements were 0.48% and 0.33% in the absolute values for the correctness and accuracy, respectively. As a supplementary experiment, we also tried lexicon optimization based on the word pair frequency criterion for comparison. The improvements in averaged accuracy was 0.11%, which was 1/3 of the improvement of our method.

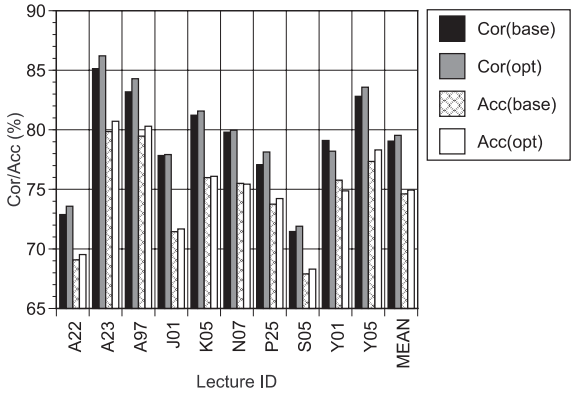

Fig. 4. Character correctness and accuracy before (*base*) and after (*opt*) the optimization.

6. DISCUSSION

Our method improved the recognition rate, but the improvement of 0.48% character correctness was much smaller than the estimated improvement of 1.39%.

This may be due to the estimation error by the word correctness probability model (1). We assumed that the model parameters are fixed during the optimization steps. If the word correctness probability change in the iteration, it may cause errors and the errors may increase as the iteration proceeds. The problem could be reduced by re-estimating the model parameters with some intervals. Another reason may be the fact that only two word attributes, the word frequency and the length, have been considered in our model. Other important attributes which contribute to improve the performance may exist. Another possible reason is that we ignored insertion errors in the analysis. Considering the insertion error in the method may also improve the recognition rate.

7. CONCLUSION

This paper first investigated the relationship between the difficulty of word recognition and two word attributes, that is the number of occurrences of each word in the language model training set and the number of phonemes in the word, using a large scale Japanese spontaneous speech corpus. It was shown that the probability of successfully recognizing each word largely varies depending on the two attributes. The relationship was then modeled using the logit model, and the model was used to build a new lexicon optimization method. The proposed method is novel in the sense that it optimizes the lexicon considering both linguistic and acoustic features of words. Recognition results showed that the trigram language model using the optimized lexicon improved the recognition rate. By improving the word correctness probability model and the concatenation algorithm, further progress is expected.

8. REFERENCES

[1] E. Giachin, P. Baggia, and G. Micca, "Language models for spontaneous speech recognition: a bootstrap method for learning phrase bigrams," *Proc. ICSLP*, vol. 2, pp. 843–846, Sept. 1994.

[2] S. Suhm and A. Waibel, "Towards better language models for spontaneous speech," *Proc. ICSLP*, vol. 2, pp. 831–834, Sept. 1994.

[3] E. Giachin, "Phrase bigrams for continuous speech recognition," *Proc. ICASSP*, vol. 1, pp. 225–228, May 1995.

[4] K. Hwang, "Vocabulary optimization based on perplexity," *Proc. ICASSP*, vol. 2, pp. 1419–1422, Apr. 1997.

[5] H. Inaba, T. Kawahara, and S. Doshita, "Reconstruction of vocabulary for large vocabulary continuous speech recognition," *Proc. JSAI*, pp. 495–498, June 1998, (in Japanese).

[6] D. Klakow, "Language-model optimization by mapping of corpora," *Proc. ICASSP*, vol. 2, pp. 701–704, May 1998.

[7] H. Kuo and W. Reichl, "Phrase-based language models for speech recognition," *Proc. EUROSPEECH*, vol. 4, pp. 1595–1598, Sept. 1999.

[8] Z. Jun, "Lexicon optimization for chinese language modeling," *Proc. ISCSLP*, Oct. 2000.

[9] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," *Proc. ASRU*, Dec. 2001.

[10] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki, and T. Ohdaira, "Toward the realization of spontaneous speech recognition," *Proc. ICSLP*, vol. 3, pp. 518–521, Oct. 2000.