SOURCE-EXTENDED LANGUAGE MODEL FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Tetsunori Kobayashi, Yosuke Wada and Norihiko Kobayashi

Department of Electrical, Electronics and Computer Engineering
Waseda University
Shinjuku-ku, Tokyo 169, Japan
http://www.tk.elec.waseda.ac.jp/~ koba/

ABSTRACT

Information source extension is utilized to improve the language model for large vocabulary continuous speech recognition (LVCSR). McMillan's theory, source extension make the model entropy close to the real source entropy, implies that the better language model can be obtained by source extension (making new unit through word concatenations and using the new unit for the language modeling). In this paper, we examined the effectiveness of this source extension. Here, we tested two methods of source extension: frequency-based extension and entropy-based extension. We tested the effect in terms of perplexity and recognition accuracy using Mainichi newspaper articles and JNAS speech corpus. As the results, the bi-gram perplexity is improved from 98.6 to 70.8 and tri-gram perplexity is improved from 41.9 to 26.4. The bigram-based recognition accuracy is improved from 79.8% to 85.3%.

1. INTRODUCTION

To get language model which can provide low test-set perplexity is one of the most important issues for realizing good performance of LVCSR. Several efforts have been made to improve the language models [1],[2],[3],[4]. This paper also deal with language modeling for LVCSR, especially, the effect of the extension of lexicon by adding new words generated by concatenating some word sequences.

Information theoretically, lexicon extension has the meaning as the information source extension. When the symbols from an information source are combined with every few symbols (this process is referred as the source extension) and the entropy is measured with this new symbol sequence, the entropy per original symbol (the entropy per new symbol over the number of the combined symbols) decreases and approaches to the real entropy of the information source.

This well known theory, McMillan's theory, implies that better language model can be obtained by information source extension. That is, if we make new words by connecting several words and add them to the lexicon, better language model can be obtained.

Beside, the lexicon extension is expected to lead another merit to LVCSR, that is the recognition steadiness.

Since the Japanese language is agglutinative one, the definition of word is not clear (there is no word boundary in the sentence). For example, Japanese Roman expression of the English phrase "when I'm using telephone" is as follows:

denwawoshiteirutokiniwa ....

Many Japanese LVCSR system adopt morpheme as the word for recognition. However, there are so many very short morphemes, which consist of only one or two syllables. As for the above example, following morpheme sequence are generated:

denwa / wo / shi / te / iru / toki / ni / wa ....

This sentence includes 5 mono-syllable morphemes. Such short-syllable morphemes are easy to be deleted or added in the recognition process. This is one of major problems of the Japanese LVCSR. Since the information source extension make rather long words for the recognition, it is also expected to improve the recognition performance.

However, bad news are also expected. That is the size of vocabulary. The information source extension make the vocabulary size large. The large vocabulary make the recognition process harder.

Since the good news and bad news are conflicted, the effectiveness of source extension for the LVCSR is not clear. In this paper, we experimentally examine the effectiveness of this modeling as the function of extended vocabulary size.

2. SOURCE EXTENDED LANGUAGE MODEL (SELM)

Using the examples of "when I'm using telephone", let's see the effect of simple source extension.

The simple 2-nd extension makes following sequence:

denwa-wo / shi-te / iru-toki / ni-wa ....

The 3-rd extension makes following sequence:

denwa-wo-shi / te-iru-toki / ni-wa-. ...

As we can see in these examples, the simple n-th source extension is not realistic. Vocabulary size become extremely large. Therefore, some selection methods are required. In our experiments, two methods are tested: one is frequency-based method, and the other is entropy-based method.

2.1. Frequency-based method

In the frequency-based method, frequently occurring morpheme sequences are registered as new vocabulary. Table 1 shows the examples of frequently occurring morpheme

Table.1 Examples of frequently occurring morpheme sequence.

Order	Morpheme sequence	Count
1	shi ta	149003
2	shi te	147255
3	te iru	134317
4	de wa	70241
5	te i	63528
6	ni wa	51936
7	to shi	51849
8	sa re	47675
9	i ta	46002
10	re ta	43755
11	to shi te	42620
12	te i ta	41561
13	shi te iru	40063
:	:	:
49	nado no	14913
50	aQ ta	14750
:	:	:
99	tame ni	8084
100	neN no	8040
:	:	:

sequences. Data is from 1994 Mainichi newspaper article corpus. Morpheme analysis was done by RWCP.

Using this table, the previous example become following word sequence:

denwa / wo / shi-te-iru / toki / ni-wa ....

In this case, the number of mono-syllable words can be reduced to be only one.

2.2. Entropy-based method

In the entropy-based method, morpheme sequences contributing the reduction of entropy are registered as new vocabulary. To evaluate the degree of contribution, the loss of the bigram entropy caused by the concatenation of the moropheme sequence is utilized.

For a morpheme sequence $x_i - a - b - Y_j$, original entropy H_o is calculated as follows :

$$H_o = -Pr(x_i, a, b, y_j) \cdot \log Pr(x_i) Pr(a|x_i) Pr(b|a) Pr(y_j|b)$$

After the concatenation, entropy using SELM H_e becomes :

$$H_e = -Pr(x_i, a, b, y_j) \cdot \log Pr(x_i) Pr(a, b|x_i) Pr(y_j|a, b)$$

We define the evaluation function, $H(a, b)$, by the difference between H_o and H_e:

$$\begin{aligned} F(a,b) &= H_o - H_e \\ &= \sum_i Pr(x_i, a, b)(\log Pr(x_i|a) - \log Pr(x_i|a, b)) \\ &+ \sum_j Pr(a, b, y_j)(\log Pr(y_j|b) - \log Pr(y_j|a, b)) \end{aligned}$$

In the case of three or more word concatenation, similar evaluation functions are easily defined.

To select morpheme sequences to be concatenated, we preselect all possible morpheme sequences with length of 6 or less as first candidates. Then, the evaluation functions are calculated for these first candidates, and select second candidates using these scores. Since some sequences are included in some other sequences, the scores in this stage are not reliable. Therefore, we re-parse the training sentences with these second candidates and get new statistics for sequence observation. As the final process, we rescore the evaluation function with these new statistics and select top N candidates with high score.

This procedure to minimize entropy is not optimal but only approximation. However, we adopt it for its cheap calculation cost.

3. EXPERIMENT

The performance of SELM depends on the extended vocabulary size. Figure 1 shows the average number of morphemes in a word, when we extended the lexicon by adding top N frequently occurring morpheme sequences. In the rest of this section, several properties of the proposed model are examined as a function of the extended vocabulary size.

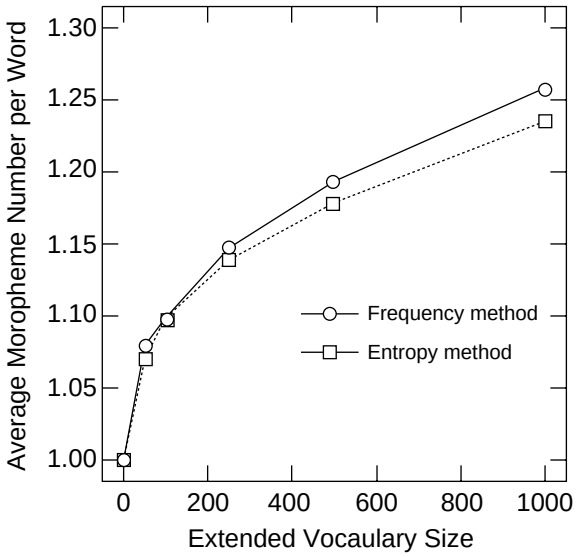

Fig.1 Number of morphemes per word as a function of extended vocabulary size.

3.1. Perplexity

In this section we examined the effectiveness of SELM in terms of perplexity.

The source extension process changes the vocabulary size. We cannot compare the values of perplexity among different language model with different vocabulary. So we introduce following normalized perplexity.

$$\tilde{H} = \frac{H}{N}$$
$$\tilde{P} = 2^{\tilde{H}},$$

where, H denotes the testset entropy and N denote the average number of morphemes in a word for the same testset.

$\tilde{H}$ and $\tilde{P}$ are the normalized entropy and the normalized perplexity, respectively.

Language models are calculated from 1994 Mainichi newspaper articles corpus. Testset consists of 550 sentences selected from 5K-set of the ASJ JNAS continuous speech recognition corpus (Japanese Newpaper Article Speech corpus) [5]. The sentences in JNAS also selected from Mainichi newspaper articles.

Figure 2 and 3 shows the normalized perplexity of SELM as a function of extended vocabulary size.

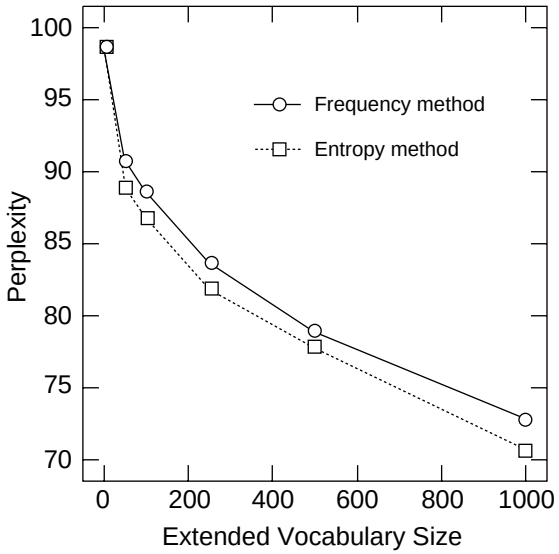

Fig.2 Perplexity of bi-gram language model for 522 sentence testset as a function of extended vocabulary size.

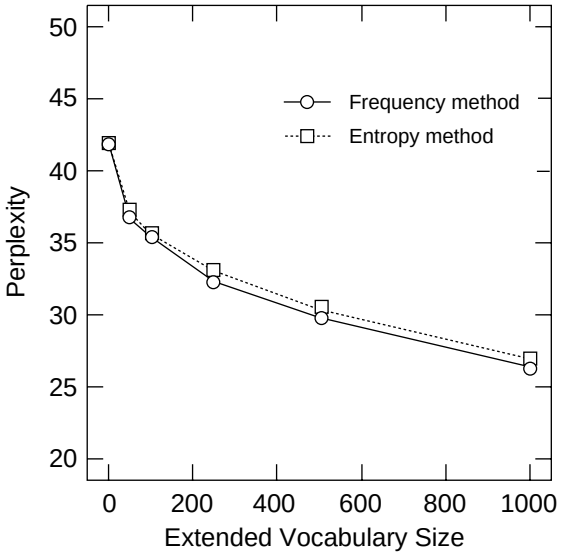

Fig.3 Perplexity of tri-gram language model for 522 sentence testset as a function of extended vocabulary size.

From these figures, we can see that the the source extension reduced the complexity of the recognition task. By adding 1000 morpheme sequences with the frequency-based method, the bigram perplexity was reduced to be 73.7% of original model. The effect of using entropy-based method is slightly better than frequency-based method: the improvement in the same condition is 71.8%.

As for the tri-gram model, the effect of source extension is higher than that of bi-gram: the improvements in the same condition are 62.9% and 64.3% in case of the frequency-based and the entropy-based methods, respectively.

3.2. Recognition performance

In this section, we examined the effectiveness of SELM in terms of recognition performance.

3.2.1. Experimental setup

Task for the evaluation is 5K word continuous speech recognition. Experimental conditions are as follows:

Test dataset: 98 sentences from JNAS continuous speech corpus uttered by 10 male speakers. Vocabulary is closed in 5K standard lexicon.

Feature parameters: 13-th order MFCC, 13-th order delta MFCC, power and delta power.

Acoustic model: a) bi-phone models (consonants depend on right contexts; vowels depend on left contexts), and b) tri-phone models. Each model has 5 states including start and end states which have no loop transition. The symbol emission probability of each state is represented by 4 mixture Gaussian distribution functions.

Language model: a) Original (5K word, no extension), b) Source extension with frequency-based method (Original+50, +100, +250, +500, +1000), c) Source extension with entropy-based method (Original+50, +100, +250, +500, +1000)

Decoder: One pass viterbi algorithm.

3.2.1. Experimental results

Figure 4 and Figure 5 show the results of the recognition experiments. Figure 4 corresponds to the results of bi-phone acoustic models and Figure 5 corresponds to that of tri-phone acoustic models.

In case of bi-phone acoustic model, the error rate (1−%accuracy) of the ordinal language model is 20.2%. By adding more words with frequency-based method, the less error rates are obtained. By adding 1000 new words, the error rate is reduced to be 14.7%. That is 27% error reduction from oridinal language model.

The entropy-based method gives better results than frequency-based method in the condition of small extended vocabulary size. For example, the 22% error reduction is realized in +250 condition. This score is 10 point better than the frequency-based method in the same condition.

However, in the condition of large extended vocabulary size, the advantage of the entropy-based method disappear. The error rate reduction is 26% in the +1000 condition that is 1 point less than the frequency-based method.

In case of tri-phone acoustic model, the advantage of source extension is rather small. The error rate of the ordinal language model is 16.8%. By adding 1000 new words, the error rates are reduced to be 14.2% and 14.5% with the frequency-based and the entropy-based method, respectively. That is 15% error reduction.

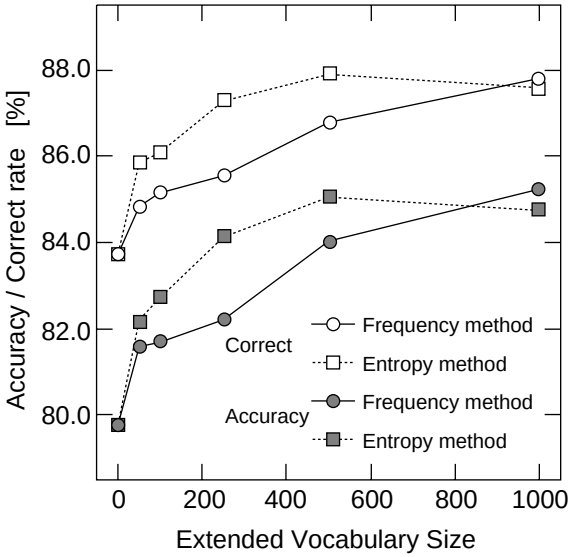

Fig.4 Performance of 5K word continuous speech recognition test using bi-phone acoustic models as a function of extended vocabulary size.

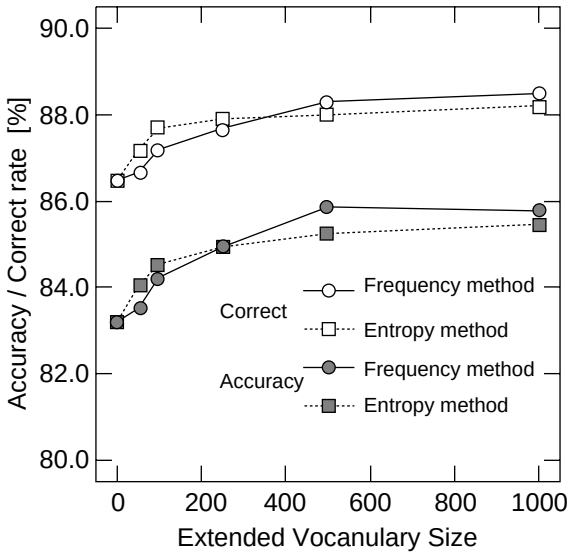

Fig.5 Performance of 5K word continuous speech recognition test using tri-phone acoustic models as a function of extended vocabulary size.

4. CONCLUSION

In this paper, we proposed SELM(Source Extended Language Model) for continuous speech recognition.

By adding 1000 frequently occurring morpheme sequences as new words, bi-gram perplexity is improved from 98.6 to 70.8 and tri-gram perplexity is improved from 41.9 to 26.4. The error rate is improved 27% at the best. Thus, the effectiveness of the proposed methods are confirmed.

By concatenating morphemes, the wider range of morpheme sequences can be used to estimate next morpheme. Moreover, the number of syllables in a word increases and occurrence of short sillable words, which is difficult to be recognized, decrease. These are the basis of the effectiveness of SELM.

We compared the two methods for the source extension: one is based on the frequency and the other is based on the entropy. The entropy-based method gives better recognition results than the frequency-based method when the extended vocabulary size is small. In other case, there are no significant difference in these two methods.

In the current stage, the entropy-based method we adopted is only approximation. Now we have a plan to develop more effective algorithm of sequence selection.

ACKNOWLEDGEMENT

In this research, we used the corpus of 1994 Mainichi newspaper articles, and its differential corpus for morpheme analysis made by RWCP (RWC-DB-TEXT-95-1) and ASJ continuous speech corpus (JNAS).

REFERENCES

1. S. Deligne and F. Bimbot, "Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams," IEEE Proc. ICASSP95, pp.169-172, May 1995.
2. H. Masataki and Y. Sagisaka, "Variable-Order N-gram Generation by Word-Class Splitting and Consecutive Word Grouping," IEEE Proc. ICASSP96, Vol.1, pp.188-191, May 1996.
3. H. Masataki, Y. Sagisaka, K. Hisaki and T. Kawahara, "Task adaptation using MAP estimation in N-gram language modeling," IEEE Proc. ICASSP97, Vol.2, pp.783-786, April 1997.
4. M. Simon, H.Ney and S. Martin, "Distant bigram language modeling using maximum entropy," IEEE Proc. ICASSP97, Vol.2, pp.787-790, April 1997.
5. K. Itou, K. Takeda, T. Takezawa, T. Matsuoka, K. Shikano, T. Kobayashi, S. Itahashi and M. Yamamoto, "Design and development of Japanese speech corpus for large vocabulary continuous speech recognition assessment," Proc. First International Workshop on East-Asian Language Resources and Evaluation, pp.98-103, May 1998.