RC25063 (C1010-004) October 20, 2010
Computer Science

IBM Research Report

Generating Compound Words with High Order n-Gram Information in Large Vocabulary Speech Recognition Systems

Jie Zhou, Qin Shi, Yong Qin
IBM Research Division
China Research Laboratory
Building 19, Zhouguancun Software Park
8 Dongbeiwang West Road, Haidian District
Beijing, 100193
P.R.China

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

GENERATING COMPOUND WORDS WITH HIGH ORDER N-GRAM INFORMATION IN LARGE VOCABULARY SPEECH RECOGNITION SYSTEMS

Jie Zhou, Qin Shi, Yong Qin

IBM Research - China

ABSTRACT

In this work we concentrate on generating the compound words with high order n-gram information for speech recognition. It is reported that the long phrases in vocabulary are more probable to appear during the decoding task. In most existed methods, only bi-gram information is under the consideration within the constraint of the computational resources and the much longer compound words are generated in an iterative way. However, many long phrases can not be iteratively built and the bi-gram information can only provide very limited help when 4-gram Language model is used during decoding. Here we present a new form of generation criterion and separate it into prediction part and history part. This largely saves the computational cost and can be extended to any higher order cases. In our experiment on mandarin Open Voice Search (OVS) work we make 0.62% percents absolute improvement and outperform the traditional mutual information based methods.

Index Terms— speech recognition, compound words, high order, vocabulary

1. INTRODUCTION

Long words are more easily to be correctly recognized in large vocabulary speech recognition systems. One of the example is depicted in [1] the result in which are obtained from voice mail corpus. In their work, the Word Error Rate (WER) of 2-word phrases is 45% while the 6-word or even longer phrases can be recognized with only 25% WER. This can be explained from two aspects. In acoustic point of view, long phrases are able to provide more information for the decoding system. In context point of view, generating compound words is equivalent to extending the context information which can also improve the prediction ability[2].

Many efforts have been devoted into generating the compound words [1, 3, 4, 5, 6, 7]. Previous attempts to automatically obtain the multi-word phrases are successful for extending the bi-gram models to improve the performance of language model in speech recognition task. However, for 3-gram models or even higher order cases, the benefits provided by these techniques are not exciting. The reason is that most of the above efforts focus on a series of empirical variations of Mutual Information (MI) criterion which originates from the uni-gram based calculation [1, 3, 4], but nowadays 3-gram and 4-gram language models are commonly used. Although generating the compound word with 3-gram information has been briefly analyzed in [1], the frequency information is not included and most of their computation is still based on MI.

By maximizing the prediction probability for a given corpus, in this work we propose a new algorithm which is rigorously deduced without empirically setting adjustable parameters. This new criterion mainly measures the prediction difference between n-gram information and (n-1)-gram information and tells us that only those compound word candidates resulting in large prediction difference are needed to be combined. By separating this criterion into history part and prediction part, the computational cost is largely decreased. Meanwhile, we find that the extending effect from history criterion plays more important role than prediction effect from prediction criterion. Finally, the symmetric form of our criterions shows that it can be easily extended to the higher order case.

2. THEORY

Here is a sequence of symbols from a given text:

$$\dots w_{i-2}, w_{i-1}, w_i = x, w_{i+1} = y, w_{i+2}, w_{i+3} \dots \quad (1)$$

In the uni-gram case, after having the word statistics, the probability of the text is given by:

$$\begin{aligned} P_t &= \prod^{i} P(w_i) \\ &= \prod^{w_i} P(w_i)^{n(w_i)} \end{aligned} \quad (2)$$

where i is the word sequence index and $n(w_i)$ is the count of word w_i in the given text. After extracting the information related with (x, y) pair which happens $n(x, y)$ times, this can be re-expressed as:

$$\begin{aligned} P_t^{(o)} &= [\prod_{(w_i, w_{i+1}) \neq (x,y)}^{i} P(w_i)] \times [\prod_{pair=(x,y)} P(x)P(y)] \\ &= P_{others} \times [P(x)^{n(x,y)} P(y)^{n(x,y)}] \end{aligned} \quad (3)$$

P_{others} refers to the part irrelative to word pair (x, y). If we combine x and y into a compound word, P_t will be updated as the follows:

$$\begin{aligned} P_t^{(n)} &= P_{others} \times [\prod_{pair=(x,y)} P(x+y)] \\ &= P_{others} \times [P(x)^{n(x,y)} P(y|x)^{n(x,y)}] \end{aligned} \quad (4)$$

then the different of these two probabilities has the form as the so called Mutual Information:

$$\begin{aligned} I_{mut} &= \log \frac{P_t^{(n)}}{P_t^{(o)}} \\ &= \log \frac{P_{others} \times [P(x)^{n(x,y)} P(y|x)^{n(x,y)}]}{P_{others} \times [P(x)^{n(x,y)} P(y)^{n(x,y)}]} \\ &= n(x,y) \log \frac{P(x+y)}{P(x)P(y)} \end{aligned} \quad (5)$$

Only (x, y) pair is involved in the 1-gram computation. However, in 3-gram computation all these 6 words are involved. For the sake of convenience, we define this 6 consecutive words as a phrase block denoted as $B_{w_{i-2}}^{w_{i+3}}$ referring to all blocks with the same content as it. The counting times of this block is $n_{w_{i-2}}^{w_{i+3}}$. Thus the probability of the text with 3-gram information can be written as:

$$\begin{aligned} P_t^{(o,3)} &= \prod^{i} P(w_i|w_{i-1}, w_{i-2}) \\ &= [P(x|w_{i-1}, w_{i-2}) P(y|x, w_{i-1}) \\ &\quad \times P(w_{i+2}|y, x) P(w_{i+3}|w_{i+2}, y)]^{n_{w_{i-2}}^{w_{i+3}}} \\ &\quad \times P_{others}^{(3)} \end{aligned} \quad (6)$$

Similarly $P_{others}^{(3)}$ denotes the piece of the text unrelated with all phrase blocks $B_{w_{i-2}}^{w_{i+3}}$. While after the combination of x and y, the new probability can be shown as:

$$\begin{aligned} P_t^{(n,3)} &= \prod^{i} P(w_i|w_{i-1}, w_{i-2}) \\ &= [P(x+y|w_{i-1}, w_{i-2}) P(w_{i+2}|x+y, w_{i-1}) \\ &\quad \times P(w_{i+3}|w_{i+2}, x+y)]^{n_{w_{i-2}}^{w_{i+3}}} \times P_{others}^{(3)} \end{aligned} \quad (7)$$

Since all above computation are concerned with blocks $B_{w_{i-2}}^{w_{i+3}}$, we add this information for clarification. Thus the difference between these two computation is:

$$\begin{aligned} I_{tri}(B_{w_{i-2}}^{w_{i+3}}) &= \log \frac{P_t^{(n,3)}(B_{w_{i-2}}^{w_{i+3}})}{P_t^{(o,3)}(B_{w_{i-2}}^{w_{i+3}})} \\ &= n_{w_{i-2}}^{w_{i+3}} \times \big[\log \frac{P(y|x, w_{i-1}, w_{i-2})}{P(y|x, w_{i-1})} \\ &\quad + \log \frac{P(w_{i+2}|y, x, w_{i-1})}{P(w_{i+2}|y, x)} \\ &\quad + \log \frac{P(w_{i+3}|w_{i+2}, y, x)}{P(w_{i+3}|w_{i+2}, y)} \big] \end{aligned} \quad (8)$$

This is exactly what we obtain after combining x and y. In some cases, increasing the n-gram order will helps a lot for computation. For example, after having known the first two word "G." and "D.", it is very probable to guess the whole phrase must be "G.D.P.". But if we only know one pre-word "D.", it is hardly to obtain the right prediction "P.". On the contrary, for the word sequence "A.B.C.", the difference between 3-gram prediction and 2-gram prediction is not so large as that of "G.D.P.". The Eq. (8) tells us we only need to concentrate on those phrases in which a large prediction gap exists when using different order language models. For those word sequence always happen together and the order of n-gram prediction brings no difference, we don't need to consider them as new words.

Now we find all 6 variables are involved in computing the affection of combining x and y, and moreover, 4-gram knowledge should be used if we want to analysis 3-gram case. This make the computational cost appear too heavy. Fortunately, after some algebra transformations the probability difference is written in three symmetric terms with very straightforward meanings which will help us for further simplifications and saving the computational resources.

The last line of the Eq. (8) will be separated into three part and each part denotes the knowledge difference after increasing the n-gram order. The first term in the product refers to the effect of predicting the next word, and the last two terms in the product refer to the effect after elongating the historical information. Notice that although we have six variables in total, but in each separated term, only 4 or them appears. So this step not only elucidates the meaning of our computation, but also provide a chance for the further mathematical simplification to decrease the computation cost.

This computation only includes the blocks $B_{w_{i-2}}^{w_{i+3}}$ and in order to obtain the total effect for the given text after adding new phrase (x, y) we should sum over the other parameters within this block.

$$\begin{aligned} I_{tri} &= \sum_{w_{i-2}, w_{i-1}, w_{i+2}, w_{i+3}} I_{tri}(B_{w_{i-2}}^{w_{i+3}}) \\ &= \sum \log \frac{P_t^{(n,3)}(B_{w_{i-2}}^{w_{i+3}})}{P_t^{(o,3)}(B_{w_{i-2}}^{w_{i+3}})} \\ &= \sum_{w_{i-1}, w_{i-2}} P(y, x, w_{i-1}, w_{i-2}) \log \frac{P(y|x, w_{i-1}, w_{i-2})}{P(y|x, w_{i-1})} \\ &\quad + \sum_{w_{i-1}, w_{i+2}} P(w_{i+2}, y, x, w_{i-1}) \log \frac{P(w_{i+2}|y, x, w_{i-1})}{P(w_{i+2}|y, x)} \\ &\quad + \sum_{w_{i+2}, w_{i+3}} P(w_{i+3}, w_{i+2}, y, x) \log \frac{P(w_{i+3}|w_{i+2}, y, x)}{P(w_{i+3}|w_{i+2}, y)} \\ &= I_{tri}^{pre} + I_{tri}^{his} \end{aligned} \quad (9)$$

Then we have two criterions, I_{tri}^{pre} denotes prediction criterion

and I_{tri}^{his} denotes the history criterion:

$$I_{tri}^{pre} = \sum_{w_{i-1},w_{i-2}} P(y,x,w_{i-1},w_{i-2}) \log \frac{P(y|x,w_{i-1},w_{i-2})}{P(y|x,w_{i-1})}$$

$$I_{tri}^{his} = \sum_{w_{i-1},w_{i+2}} P(w_{i+2},y,x,w_{i-1}) \log \frac{P(w_{i+2}|y,x,w_{i-1})}{P(w_{i+2}|y,x)} + \sum_{w_{i+2},w_{i+3}} P(w_{i+3},w_{i+2},y,x) \log \frac{P(w_{i+3}|w_{i+2},y,x)}{P(w_{i+3}|w_{i+2},y)} \quad (10)$$

Although we start from a 6 word block, now we have restrict all computations within 4 parameters and no approximations are made here. And we can also separate the historical part into two criterions. Moreover, the symmetric property shown here can help us to extend the criterions to even higher order case.

3. EXPERIMENTAL RESULTS

The experiments are carried out on mandarin Open Voice Search(OVS) task concerned with a free style queries and messages. We have in total 2092 testing sentences collected from 20 different real speakers. The queries cover the economics, entertainment, sports news and other areas in our daily life. 10% of the data are hold out as smoothing data. The language model is a conventional linearly interpolated 4-gram model[8]. The total training corpus size is around $1G$ Bytes and the lexicon size is $107k$. The total perplexity and Char Error Rate (CER) are computed to verify our method. We also analysis the effect of prediction criterion and history criterion for the further understanding of our method. We didn't test the 3-gram decoding task since as shown below the performance on 4-gram model is good enough.

We start from the baseline vocabulary and compute the 4-gram and 3-gram information as needed in Eq. (10). Next the score of prediction criterion I_{tri}^{pre} and history criterion I_{tri}^{his} are obtained and the candidates in both groups are sorted accordingly. At last we sum up the contributions over both groups for each new word and then obtained the final ranking list. All these are exact computation and no adjustable parameters exists here. For very large corpus, we can separate all 4-grams and 3-grams information into several parts and perform the calculation one by one. So we needn't worry about the memory cost here. We can also set a threshold for 4-grams that only terms larger than this threshold will be evolved into the computation. A low threshold will not affect the final result but can speed up the computation dramatically.

In segmentation step for mandarin training corpus, we made three cycles of iteration in accumulating the statistics for the lexicon list. Then we made one iteration in adding new words and the results are shown in the following (Fig. 1). In the baseline the char error rate (CER) is 16.44% as shown by the blue dashed line. With new words added, the CER drops down gradually to the lowest point 15.82% at which we have $N(w) = 20k$ new words, and rise back later. This process is depicted in Fig. 1 by the solid circles. This is a reasonable evolution because each new word has two-side effect on the decoding system. On one hand, the new word extend the n-gram ability while on the other hand, it brings phonetic confusions into the system. After low ranked new words added, the confusion effect is more serious than the extending effect, so the CER starts to get worse. When $N(w) = 30k$, the CER rise to 15.85%, but still much better than the baseline. This proves the new word selection strategy of our algorithm a lot.

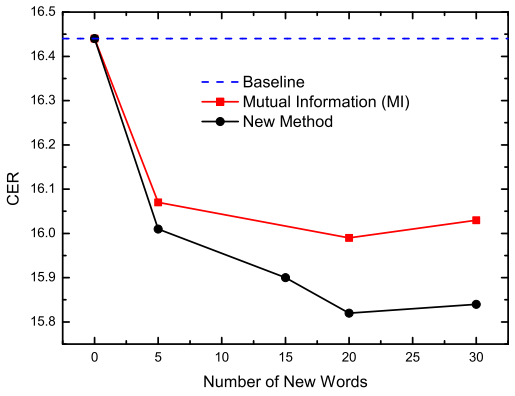

Fig. 1. Char Error Rate (CER) $vs.$ Number of New Words.

In this graph we also have our method compared with the Mutual Information(MI) method which based on the bi-gram computation as stated in Eq. (5). The largest gain from MI method also happens at $20k$ new words point and exhibit 0.45% absolute improvement, but less than our new method. The reason that why this bi-gram based computation can also give help to the 4-gram language model will be demonstrated later. All the data are listed in Table. 1.

In Fig. 2 the history criterion and prediction criterion are depicted respectively. This result is a little bit out of our expectations. We add new words into the lexicon with the hope of increasing the recognition performance of these new phrases in the testing set. However, we find that the history criterion plays a more important role than the prediction criterion and almost exhibit the equivalent performance as the total criterion. We attribute this to the high order n-gram model that we used in language modeling. Under this condition, after knowing the first several segments of a certain new phrase, it is very easy to predict the last segment of this new phrase, no matter how this new phrase is considered in our

CER(%)	$0k$	$5k$	20k	30k
MI	16.44	16.07	15.99	16.03
New Method	16.44	16.01	15.82	15.84
New Method (H)	16.44	16.03	15.82	15.86
New Method (P)	16.44	16.11	15.94	16.01

Table 1. Char Error Rate (CER) of different criterions. (H) denotes the history criterion and (P) denotes the prediction criterion

.

lexicon. On the contrary, before combination, the first several segments of this new phrase occupy the whole historical information in n-gram model and can only provide the knowledge equivalent to bi-gram model after combination for prediction. As reported before, the difference between bi-gram performance and 3-gram or 4-gram is large. From this we can also understand that the most influence of mutual information on our decoding task is the extending effect, rather than prediction effect.

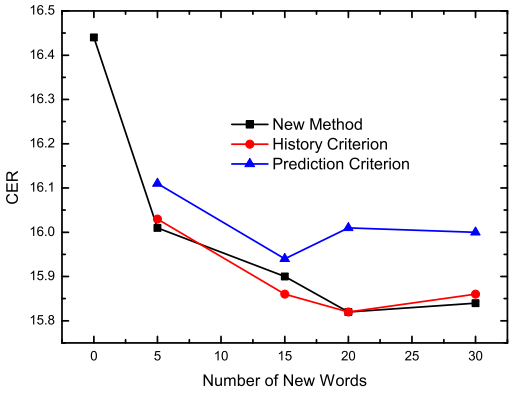

Fig. 2. Char error rate $vs.$ number of new words. The effect of history criterion and prediction criterion are depicted respectively and compared with the final criterion.

4. DISCUSSION AND CONCLUSIONS

In this work we proposed a rigorous deduction without any adjustable or fitting parameters of utilizing the high order n-gram information to extract out the new compound words which facilitates the large vocabulary speech recognition task. In our method, we try to maximize the text prediction probability with traditional n-gram language model, according to which the affection of the compound words are computed. Further we simplify the the obtained criterion by separating it into two parts, prediction part and history part. This step decreases the number of variables involved in each calculation and meanwhile exhibit a straightforward meaning to our computation. Moreover, the symmetric form of the criterion show that it can be easily extended to higher order cases.

In the OVS task our method gives 0.62% improvement as the best performance with $20k$ new words added to the baseline lexicon and also outperforms the mutual information method. History criterion and prediction criterion are meanwhile tested, the result of which shows that the extending effect brought by history criterion has more impacts on the decoding performance than the prediction effect in high order n-gram language model.

Actually every words have two side effects on the original system. On one hand it can extend the n-gram ability and on the other hand it brings the acoustic confusion effect into the system. Our work only focus on how to maximize the text prediction probability with n-gram language model while doesn't take any acoustic information into account, which turns to be serious after adding a large number of new words. This is what we are going to focus on in the next.

5. REFERENCES

[1] G. Saon and M. Padmanabhan, "Data-driven approach to designing compound words for continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 327, 2001.

[2] R. Kneser, "Statistical language modeling using a variable context length," in *ICSLP*, 1996, p. 494C497.

[3] E.P.Giachin, "Phrase bigrams for continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, 1995, pp. 225 –228 vol.1.

[4] H. K. J. Kuo and W. Reichl, "Phrase-based language models for speech recognition,," in *Proc. Eurospeech 99, Budapest, Hungary*, 1999.

[5] A. Berton, P. Fetter, and P. Regel-Brietzmann, "Compound words in large-vocabulary german speech recognition systems," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, vol. 2, pp. 1165–1168.

[6] D. Klakow, "Language-model optimization by mapping of corpora," in *ICASSP*, 1996, p. 701C704.

[7] K. Ries, F. D. Buo, and A. Waibel, "Class phrase models for language modeling," in *ICASSP*, 1996, p. 398C401.

[8] F. Jelinek, "statistical methods for speech recognition," in *in Language, Speech and Communication Series. Cambridge, MA: MIT Press*, 1999.