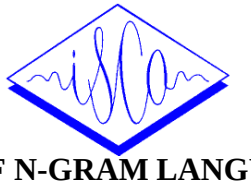

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

RAPID ADAPTATION OF N-GRAM LANGUAGE MODELS USING INTER-WORD CORRELATION FOR SPEECH RECOGNITION

Koki Sasaki†, *Hui Jiang*‡ *and Keikichi Hirose*†

† Department of Information and Communication Engineering,
University of Tokyo, Hongo 7-3-1, Bunkyu-ku, Tokyo 113-0033, Japan

‡ Dialog Systems Research, Multimedia Communication Research Lab,
Bell Labs, Lucent Technologies, Murray Hill, NJ 07974

Email: {koki,hirose}@gavo.t.u-tokyo.ac.jp hui@research.bell-labs.com

ABSTRACT

In this paper, we study the fast adaptation problem of n-gram language model under the MAP estimation framework. We have proposed a heuristic method to explore inter-word correlation to accelerate MAP adaptation of n-gram model. According to their correlations, the occurrence of one word can be used to predict all other words in adaptation text. In this way, a large n-gram model can be efficiently adapted with a small amount of adaptation data. The proposed fast adaptation approach is evaluated in a Japanese newspaper corpus. We have observed a significant perplexity reduction even when we have only several hundred adaptation sentences.

1. INTRODUCTION

It is well known that a proper n-gram language model plays an important role in large vocabulary automatic speech recognition (LVASR) system. It usually requires a huge amount of text corpus to estimate a reliable n-gram model. In reality, it is not practical to collect enough text data to train a task-dependent n-gram model for every specific task. Thus, today's LVASR system always uses a general-purpose task-independent (TI) n-gram language model for all tasks or domains. However, because of the mismatch nature among various tasks, it is strongly desirable to have a task-dependent (TD) n-gram model for each specific task in order to achieve a better recognition performance. One feasible strategy here is to adopt the adaptive learning method, i.e., we adapt a TI n-gram model to each target task by using merely a small amount of text data collected for that task. Although we have many fast adaptation techniques available for acoustic model, we have not yet found any good solution for language model adaptation because n-gram model is harder to handle in some senses.

In the paper, we study the problem to rapidly adapt n-gram model from a Bayesian viewpoint. As we will show, the MAP (maximum a posteriori) estimation of n-gram model has a straightforward form to implement, but it is too slow to converge in a new task domain. The works in the paper focuses on the rapid adaptation of n-gram model based on the Bayesian approach. Starting from the MAP formulation of n-gram model, we propose a heuristic method to investigate the correlation between all key-words to make the Bayesian adaptation of n-gram model fast, efficient and effective. Concretely, while we estimate task-independent n-gram model from a large text corpus, we also explore and record all information about the correlation between any two key-words in the corpus, i.e. the probability of co-occurrence in a certain segment of text. When it is needed to update the task-independent n-gram model to a certain target task based on a small amount of adaptation text, by using the available correlation information among key-words, the appearance of one certain key-word in adaptation text can be used to predict the occurrences of all other key-words. Then all of these predicted occurrences are added with the actual occurrence in the adaptation text. Finally, a task-adaptive n-gram model is derived under the framework of MAP estimation. In this way, a large n-gram language model can be rapidly updated based on merely a small amount of task-dependent text data.

The proposed fast adaptation method is evaluated in a Japanese Mainichi Newspaper corpus. In our experiments, totally 5000 articles in 1991 Mainichi newspaper are used to train a task-independent bi-gram model. Some dis-jointed articles on "Gulf War" are used as adaptation and evaluation data. Experimental results show that the proposed method can reduce the perplexity of bi-gram model significantly from the task-independent bi-gram by using only several hundreds of sentences as adaptation data. Some preliminary LVASR experiments also show that a speech recognizer based on the adapted bi-gram model has achieved obvious word accuracy improvement in the target task comparing with the same recognizer with a task-independent bi-gram model.

2. MAP ESTIMATION OF N-GRAM MODEL

Today the n-gram model has become the dominant language model in large vocabulary speech recognition. Generally speaking, n-gram language model Λ is composed of a set of conditional word occurrence probabilities on its corresponding history, i.e., $P(w|h)$. Depending on the case, the history h could be the previous word (bigram), the previous two words (trigram), or even longer segment. If we denote $P(w|h) \equiv \lambda_{hw}$, n-gram model Λ can be expressed as

$$\Lambda = \{ \lambda_{hw} \mid w \in W \text{ and } h \in H \}. \quad (1)$$

where W denotes the set of all possible words and H all admissible histories. Obviously, the n-gram model parameters λ_{hw} follow

the constraint

$$\sum_{w \in W} \lambda_{hw} = 1 \tag{2}$$

for every h in H.

Given any text data $\mathbf{T} = w_1 w_2 ... w_n$, the likelihood function of n-gram model Λ is computed as

$$\begin{aligned} l(\Lambda|T) &= P(w_1 w_2 ... w_n) = \prod_{i=1}^{n} P(w_i|h_i) \\ &= \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{N_{hw}} \end{aligned} \tag{3}$$

where N_{hw} denotes the occurrence number (frequency) of word sequence hw in the text T. By taking the constraints in eq.(2) into account, the maximum likelihood (ML) estimation of n-gram model is

$$\lambda_{hw}^{(ML)} = \frac{N_{hw}}{\sum_{w \in W} N_{hw}} \tag{4}$$

From eq.(3), we can see the likelihood function of n-Gram model is a multinomial distribution. It is well known that its natural conjugate prior is the so-called *Dirichlet* distribution:

$$p(\Lambda) = p(\{\lambda_{hw}\}) \propto \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{\alpha_{hw}-1} \tag{5}$$

where $\alpha_{hw} (h \in H, w \in W)$ are hyperparameters, which usually are estimated from a task-independent corpus T^i:

$$\alpha_{hw} = N_{hw}^i + 1 \ \ (w \in W, \ h \in H) \tag{6}$$

where N_{hw}^i denotes the occurrence number of hw in the corpus T^i.

According to Bayes' theorem, given an adaptation text data T^a, the posterior pdf is

$$p(\Lambda|T^a) \propto p(\Lambda) \cdot l(\Lambda|T^a) \propto \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{N_{hw}^i + N_{hw}^a} \tag{7}$$

Similarly, considering the constraints in eq.(2), the MAP (*maximum a posteriori)* estimation of n-gram model is derived as

$$\lambda_{hw}^{(MAP)} = \frac{(N_{hw}^i + N_{hw}^a)}{\sum_{w \in W} (N_{hw}^i + N_{hw}^a)} \tag{8}$$

From the above eq.(8), we note the MAP estimation of n-gram model has a very straightforward formulation. However, it converges too slow. Comparing with TI text data T^i, the adaptation text T^a usually has much less amount of data. Therefore, the values of N_{hw}^a will be much smaller than that of N_{hw}^i. A small amount of adaptation data will not change the value of λ_{hw} too much. So it usually requires a relatively large amount of adaptation data to make MAP-based adaptation effective. Because it is relatively costly to collect data in practice, it is strongly desirable to have a fast method to adapt n-gram model more efficiently.

3. RAPID MAP ADAPTATION WITH INTER-WORD CORRELATION

Starting from the MAP formulation of n-gram model, in this section, we propose a heuristic method to investigate the correlation between all key-words to make the Bayesian adaptation of n-gram model fast, efficient and effective. Concretely, while we estimate task-independent n-gram model from a large text corpus, we also explore and record all information about the correlation between any two key-words in the corpus, i.e. the probability of co-occurrence in a certain segment of text. When it is needed to update the task-independent n-gram model to a certain target task based on a small amount of adaptation text, by using the available correlation information among key-words, the appearance of one certain key-word in adaptation text can be used to predict the occurrences of all other key-words. Then all of these predicted occurrences are added with the actual occurrence in the adaptation text. Finally, the task-adaptive n-gram model are derived under the framework of MAP estimation. In this way, a large n-gram language model can be rapidly updated based on merely a small amount of task-dependent text data.

Our proposed fast adaptation algorithm is performed as follows:

I. Estimate TI n-gram model and record correlation information:

1. From T^i, we estimate a TI n-gram model $\{\lambda_{hw}\}$ and record all sufficient statistics N_{hw}^i for every n-gram hw.
2. Based on task-independent (TI) data T^i, we build a common word list (CWL) which includes all common words appearing equally everywhere in the language, such as preposition, adverb, etc. Hereafter, *key-word* is defined to be all words not included in CWL.
3. Partition TI data T^i into some consecutive segments: $T^i = T_1^i T_2^i ... T_K^i$. Here each $T_k^i (1 \le k \le K)$ can be a sentence, a paragraph, or even an article.
4. For any n-gram hw, we calculate its co-occurrence relation with every key-word w', i.e.,

$$q_{w'[hw]}^k = \begin{cases} 1 & \text{If } w' \text{ and } hw \text{ co-occur in } T_k^i, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

   Then we summarize $q_{w'[hw]}^k$ over all segments T_k^i, i.e.,

$$q_{w'[hw]} = \sum_{1 \le k \le K} q_{w'[hw]}^k \tag{10}$$

   Note that in this paper we don't consider unknown word problem in adaptation for simplicity. Thus we just set $q_{w'[hw]} = 0$ when either hw or w' contains the unknown word, denoted as *UNK*.

II. Fast adaptation

5. Given adaptation data T^a, we first collect sufficient statistics, N_{hw}^a, i.e., occurrence number in T^a, of every n-gram hw.
6. Then we compute the predicted occurrence number Q_{hw} of every n-gram hw based on all key-words v in T^a. That is,

$$Q_{hw} = \sum_{v \ne w, v \in T^a, v \overline{\in} CWL} N_v^a \cdot q_{v[hw]} \ \ (\text{For all } hw) \tag{11}$$

   where N_v^a denotes the occurrence number of word v in T^a.
7. Based on MAP formulation, we update N-Gram model as follows:

- When $w \neq UNK$ (not unknown word),

$$\lambda'_{hw} = \frac{N^i_{hw} + N^a_{hw} + \alpha \cdot Q_{hw}}{\sum_{w \in W} N^i_{hw} + \sum_{w \in W} N^a_{hw} + \alpha \sum_{w \in W} Q_{hw}} \tag{12}$$

where α is a weight to control the contribution of the predicted occurrence number.

- When $w = UNK$,[1]

$$\lambda'_{hw} = \frac{N^i_{hw} + N^a_{hw}}{\sum_{w \in W} N^i_{hw} + \sum_{w \in W} N^a_{hw}} \tag{13}$$

8. Obviously, the new n-gram model updated from eqs.(12) (13) does not satisfy the constraint $\sum_{w \in W} \lambda'_{hw} = 1$. Thus we use the following strategy to normalize λ'_{hw} and finally get the adapted n-gram model as:

$$\lambda_{hw} = \begin{cases} C(h) \cdot \lambda'_{hw} & w \neq UNK, \\ \lambda'_{hw} & w = UNK \end{cases} \tag{14}$$

where

$$C(h) = \frac{\sum_{w \in W, w \neq UNK} \lambda''_{hw}}{\sum_{w \in W, w \neq UNK} \lambda'_{hw}} \tag{15}$$

and λ''_{hw} denotes the pure MAP estimate:

$$\lambda''_{hw} = \frac{N^i_{hw} + N^a_{hw}}{\sum_{w \in W} N^i_{hw} + \sum_{w \in W} N^a_{hw}} \tag{16}$$

4. TASK ADAPTATION EXPERIMENTS

Our proposed fast adaptation method is evaluated in a Japanese Mainichi newspaper corpus. In our experiments, totally 5000 articles (approximately 1.2M words) in 1991 Mainichi newspaper are selected to train a task-independent (TI) bigram model. Some dis-jointed articles on "Gulf War" are used as adaptation and evaluation data. We use from 100 up to 1000 sentences as adaptation data for different cases and another 100 sentences (totally 3178 words) as test data. In the experiments, we use CMU-Cambridge statistical language modeling toolkit to build bigram model and calculate perplexity. Good-Turing discounting method is used in bigram model construction. And vocabulary size of bigram model is chosen to be 5k.

4.1. How to build Common Word List (CWL)

As in [4], we use mutual information to select the most common words for CWL. We use all articles in 1991 Mainichi newspaper (except those on 'Gulf War') and partition them into 10 topics, i.e., $T = \{t_1, t_2, ..., t_{10}\}$. The mutual information of each word w and the text T is calculated as

$$I(T; w) = -\sum_{i=1}^{10} P(t_i) \log P(t_i) + \sum_{i=1}^{10} P(t_i|w) \log(t_i|w) \tag{17}$$

[1] In case w is unknown word, in eq.(12), we have $Q_{hw} = 0$ in numerator but a large number $\sum_{w \in W} Q_{hw} \geq 0$ in denominator. Thus eq.(12) will underestimate the probabilities of UNK.

where $P(t_i) = 0.1$ and $P(t_i|w) = \frac{\text{frequency of } w \text{ in } t_i}{\text{frequency of } w \text{ in } T}$. Thus, $I(T; w)$ indicates nonuniformity of the frequency of the word w in various topics. We first select top 20000 words according to their frequencies in T. Then we calculate $I(T; w)$ for each word and sort them according to their $I(T; w)$ values. Finally we pick up the last N_c words which have smallest $I(T; w)$ values to build the Common Word List (CWL).

4.2. Effects of partition unit and α

In step 3, we need partition text data into small segments to calculate correlation. We have investigated our fast adaptation algorithm based on different partition unit, sentence (PS) or paragraph (PP) or article (PA). Perplexity of test set is shown as a function of α for different partition units in figure 1, We have found that PS and PP give much more perplexity reduction than PA and PS shows the best performance. In the following experiments, we will adopt PS as partition unit. As for α, it definitely is task-dependent and usually gives good performance in $[0.01, 0.1]$. In our following experiments, we fix $\alpha = 0.03$ except explicitly stated.

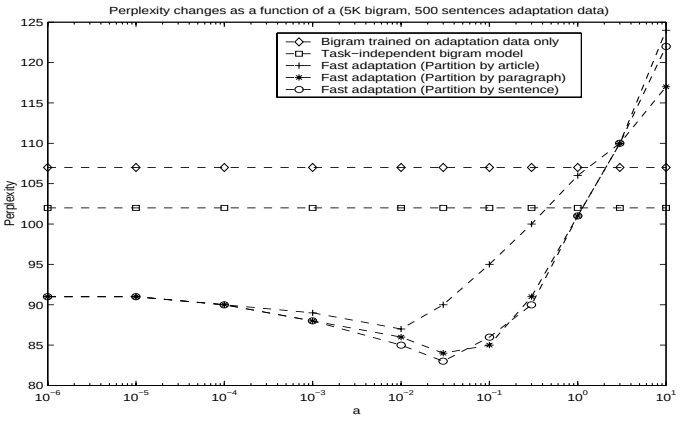

Figure 1: Perplexity changes as a function of α for different partition methods (in sentence, paragraph, or article) in case of 500 adaptation sentences.

4.3. Effect of the size of CWL (N_c)

We also test the influence of the size of CWL, N_c, on the perplexity reduction in our fast adaptation approach. From the results in Figure 2, the fast adaptation method performs well in a quite wide range $[6000, 16000]$ of N_c. Thus, we choose $N_c = 7000$ in the following experiments.

4.4. An improved fast adaptation strategy

The algorithm described in section 3 only use TI data T^i to compute predicted occurrence Q_{hw}. To further accelerate adaptation, we can also similarly compute Q_{hw} for adaptation data T^a. Moreover, an extra weight β is introduced to emphasize occurrence N^a_{hw} in adaptation data T^a. Therefore, the adaptation equation (12) is accordingly modified as

$$\lambda_{hw} = \frac{(N^i_{hw} + \beta \cdot N^a_{hw}) + \alpha(Q^i_{hw} + Q^a_{hw})}{\sum_{w \in W}(N^i_{hw} + \beta \cdot N^a_{hw}) + \alpha \sum_{w \in W}(Q^i_{hw} + Q^a_{hw})} \tag{18}$$

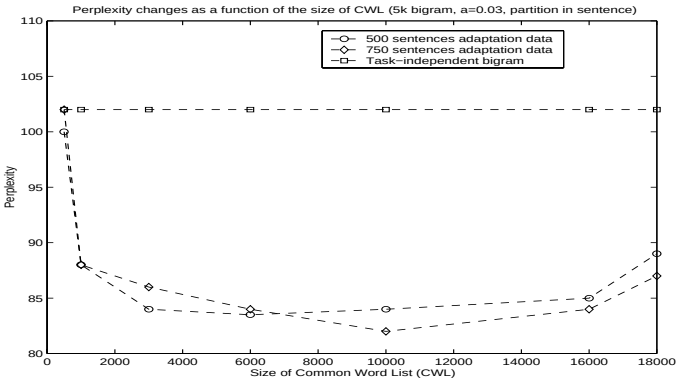

Figure 2: Perplexity changes as a function of CWL size N_c. ($\alpha = 0.03$ and partition in sentence)

where Q^i_{hw} and Q^a_{hw} are calculated from T^i and T^a respectively and β is another control parameter. We denote the adaptation method in eq.(12) as FA and eq.(18) as FA2. In Figure 3, we compare our fast adaptation methods FA, FA2 with the normal MAP estimation for various amount of adaptation text data. It is clear that both FA and FA2 converge much faster than MAP and FA2 with $\beta = 30$ gives the maximum perplexity reduction.

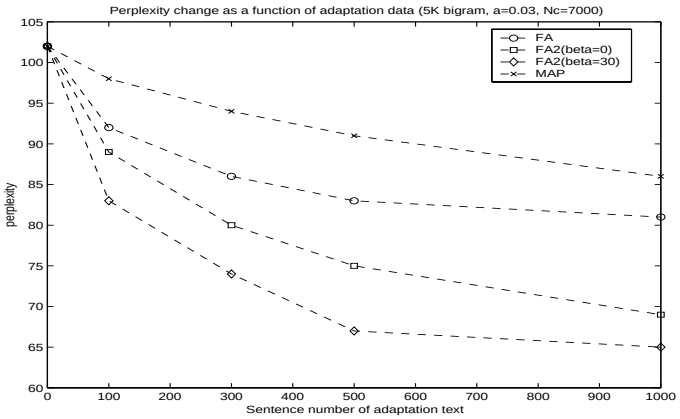

Figure 3: Comparative experimental results of FA, FA2 and MAP as a function of adaptation text data. ($\alpha = 0.03$ and $N_c = 7000$)

4.5. Preliminary speech recognition results

To perform real speech recognition experiments, we collect some speech data on the target task domain from one male speaker. The speaker is asked to read all 100 test sentences. In decoding, we employ the standard Japanese JULIUS decoding software and Japanese state-tied triphone acoustic model (totally 3k distinct states) supplied with the decoder[3]. The speech data is evaluated with bigram language models derived from different adaptation methods. Comparative results of word accuracy are shown in Figure 4. We have found that our fast language adaptation method also obviously improve speech recognition performance even when we have only 100 or 300 sentences to adapt language model.

Because of the mismatched recording conditions and other factors, the performance of this speech recognition baseline system is too low, around 40%. It will be very interesting to see how these fast language model adaptation methods work when the baseline performance is reasonably good.

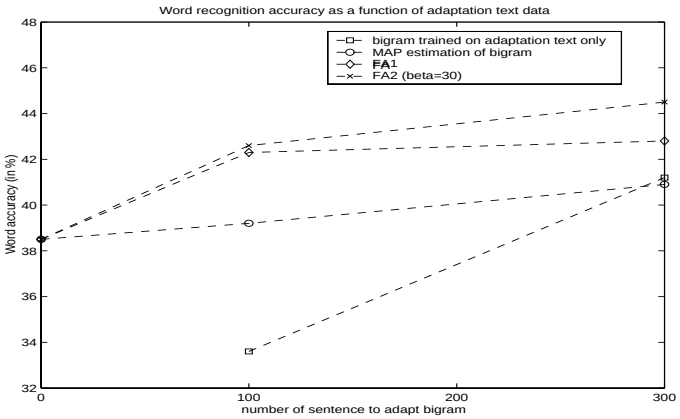

Figure 4: Comparative speech recognition results based on different language model adaptation methods.

5. CONCLUSIONS

In this paper, we have proposed a heuristic method to accelerate MAP adaptation of n-gram language model by using correlation among all key-words in the model. The fast adaptation method is evaluated in task adaptation of a Japanese newspaper corpus. Experimental results show some significant perplexity reduction of n-gram model even when we have only several hundred adaptation sentences from the target task domain. Some preliminary ASR recognition results also show the similar word recognition improvements over a task-independent language model. Fast adaptation techniques for n-gram language model is very important in many applications. The work in this paper shows that the use of correlation information among words in language model is a promising way to perform rapid n-gram language model adaptation.

6. REFERENCES

[1] P. Clarkson and R. Rosenfeld, "Statistical Language modelling using CMU-Cambridge toolkit,"*Proc. of Eurospeech'97*, Vol.5, pp.2707-2710, Sep. 1997.

[2] M. Federico, "Bayesian Estimation methods for N-gam language model adaptation," *Proc. of ICSLP'96*, pp.240-243, Oct. 1996.

[3] T. Kawahara, *et. al.*, "Evaluation of fundamental Japanese dictation software in 1998", *Technical report of IEICE*. (in Japanese)

[4] T. Kawahara and S. Doshita, "Topic independent language model for key-phrase detection and verification," *Proc. of ICASSP'99*, pp.685-688, March 1999.

[5] K. Sasaki, "Language model adaptation by using inter-word correlation for speech recognition", *Master thesis*, University of Tokyo, Feb. 2000. (in Japanese)