ENHANCED MAP ADAPTATION OF N-GRAM LANGUAGE MODELS USING INDIRECT CORRELATION OF DISTANT WORDS

Takaaki Moriya†, *Keikichi Hirose*†, *Nobuaki Minematsu*‡ *and Hui Jiang*††

†Graduate School of Frontier Sciences, University of Tokyo,
‡Graduate School of Information Science and Technology, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, JAPAN

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

E-mail: {moriya,hirose,mine}@gavo.t.u-tokyo.ac.jp hui@research.bell-labs.com

ABSTRACT

A novel and effective method to adapt n-gram language models to a new domain has been developed. We propose a heuristic method of language model adaptation using indirect correlation between words which are distant from each other, in addition to the conventional n-gram correlation, which represents only superficial and direct information of adjacent words. By adding the correlation of distant words, the adapted models come to include more information on co-occurrence of words of a target domain and improve their performance as perplexity reduction. Furthermore, since the new correlation covers indirect one not appearing in surface sentences, the adapted models still work well in domains somewhat different from the target domain. Experiments show that, in comparison with well-known MAP-based adaptation, the proposed method improves the performance of perplexity reduction by approximately 10% in the target domain and also in another domain.

1. INTRODUCTION

In large vocabulary speech recognition systems, statistical language modeling usually requires a huge amount of text corpus. Since recognition task is often specific to a domain, high performance of the recognition can be realized with domain-dependent models. However, collecting sufficient text data on that domain is usually a tough and time-consuming task. To solve this problem, adaptation of the language models is widely conducted. One of the most effective methods is MAP-based adaptation[1], which mixes domain-independent (DI) text and domain-dependent (DD) text with an adequate weight before calculating n-gram probabilities.

In conventional training/adaptation methods of n-gram models, only direct correlation of adjacent words was used. If it is possible to capture the correlation between words distant from each other and integrate it into MAP-based adaptation adequately, the adapted models should characterize better the target domain. However, directly calculating long-distance correlation in n-gram strategy is impossible because it requires unreasonably large amount of text data in the target domain. In this paper, indirect correlation between words is used, which was introduced in our previous study[2] and is explained in detail in section 3. By using this correlation, word sequences which are not *actually* seen in adaptation data but *potentially* found in the data are obtained. When adapting n-gram models, the word sequences are counted in addition to those actually observed in adaptation data. Although this technique was already reported in the previous study[2], the experiments in the work were rather preliminary. Therefore, in this paper, its effectiveness is examined with a larger text corpus. The definition of indirect correlation is also refined from the previous work.

The use of indirect correlation lets us expect an interesting effect of the proposed method. Since it enables n-gram models to capture potential, but not superficial, relation between words, n-gram models adapted to a target domain with the proposed method are expected to work well still in domains somewhat different from the target domain. This expectation is verified experimentally in the paper.

In section 2, the conventional MAP adaptation of n-gram models is described. Indirect correlation between words and the algorithm of integrating it into MAP-based adaptation are explained in section 3. After that, two experiments are shown in sections 4 and 5. In section 4, the effectiveness of the proposed method in terms of perplexity reduction is shown both for the target domain and for another domain. In section 5, the proposed method is evaluated in terms of speech recognition performance. Finally, the paper is summarized in section 6.

2. MAP ADAPTATION OF N-GRAM MODEL

In this work, MAP-based adaptation method[3, 4] is adopted as the baseline of the proposed method, where DI and DD texts are mixed with an adequate weight. This framework is formulated as follows:

Let y and θ denote observations and estimation parameters, respectively. Using Bayes' theorem, *a posteriori* distribution $p(\theta|y)$ is re-written as

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta) \quad (1)$$

In the MAP adaptation framework, θ is estimated by maximizing

$p(\theta|y)$.

$$\hat{\theta} = \arg\max_{\theta} p(\theta|y) = \arg\max_{\theta} p(y|\theta)p(\theta) \quad (2)$$

Here, y means occurrence of words and θ indicates n-gram probabilities. While $p(\theta)$ is *a priori* distribution of n-gram probabilities, which is calculated with DI text, $p(y|\theta)$ is likelihood distribution of words y being observed in DD adaptation text. Assuming that *a priori* distribution $p(\theta)$ follows Dirichlet or beta distribution with a hyper-parameter, $p(y|\theta)$ and $p(\theta)$ are obtained as the following forms with weight ω to the word frequency in adaptation data.

$$p(y|\theta) \propto \theta^{\omega N^A_{hw}}(1-\theta)^{\omega(N^A_h - N^A_{hw})} \quad (3)$$

$$p(\theta) \propto \theta^{N^I_{hw}}(1-\theta)^{N^I_h - N^I_{hw}} \quad (4)$$

where N^I_x and N^A_x denotes frequency of words x in DI and DD (adaptation) text, respectively and h describes history, which is a sequence of $n-1$ words immediately before a word w. From Eq.(3) and Eq.(4), the MAP-based adapted n-gram model is derived as

$$P(w|h) = \arg\max_{\theta} p(y|\theta)p(\theta) = \frac{N^I_{hw} + \omega N^A_{hw}}{\sum_w (N^I_{hw} + \omega N^A_{hw})} \quad (5)$$

3. MAP ADAPTATION USING INDIRECT CORRELATION OF DISTANT WORDS

3.1. Algorithm

In this section, indirect correlation between a word and a word sequence is first introduced, and then it is integrated into MAP adaptation heuristically. In this scheme, since indirect correlation is calculated not only between a word and its adjacent word sequence but also between a word and its distant word sequences, more information on co-occurrence of words should be introduced into the MAP adaptation.

1. Generate a common word list (CWL) using DI text. CWL includes all the words frequently observed irrespective of domains, e.g. prepositions and auxiliary verbs.
2. Divide DI text ($= T^I$) into a set of consecutive segments.

$$T^I = T^I_1 T^I_2 \cdots T^I_{n_I} \quad (6)$$

   Although the division is possible into any types of segments such as sentences, paragraphs and so on, a segment should relate only to one domain.
3. For word sequence hw, calculate its binary correlation to word v. In the case of bi-gram model adaptation (experiments of the current paper), h is composed of a single word.

$$q^k_{v[hw]} = \begin{cases} 1 & \text{if } v \text{ and } hw \text{ are found in } T^I_k, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

   Next, summarize $q^k_{v[hw]}$ over all the segments T^I_k and divide by the total number of segments n_I;

$$q^I_{v[hw]} = \frac{1}{n_I}\sum_{k=1}^{n_I} q^k_{v[hw]} \qquad (0 \le q^I_{v[hw]} \le 1) \quad (8)$$

   $q^I_{v[hw]}$ means indirect correlation between v and hw in DI text. As for DD adaptation text ($= T^A = T^A_1 \cdots T^A_{n_A}$), the correlation can be calculated as $q^A_{v[hw]}$. It should be noted that $q^k_{v[hw]}$ is not calculated between hw of a domain and v of another because the correlation across domains will degrade adaptation performance.
4. By using $q^X_{v[hw]}$, compute the following quantity Q^X_{hw}, where v^A does not belong to CWL.

$$Q^X_{hw} = \frac{1}{\sum_{v^A} C(v^A)} \sum_{v^A \in T^A} C(v^A) q^X_{v^A[hw]} \quad (9)$$

$$(X = I, A, \quad 0 \le Q^X_{hw} \le 1, \quad v^A, hw \ne UNK)$$

   where $C(v^A)$ denotes frequency of v^A in T^A. Q^X_{hw} can be considered as indirect correlation of word sequence hw to T^A, calculated by referring to $q^X_{v^A[hw]}$ of T^X.
5. Mix Q^I_{hw} and Q^A_{hw} with weight λ.

$$Q_{hw} = (1-\lambda)Q^I_{hw} + \lambda Q^A_{hw} \qquad (0 \le Q_{hw} \le 1) \quad (10)$$

   As in the previous step, Q_{hw} means indirect correlation of word sequence hw to T^A, obtained by considering both T^I and T^A. Using Q_{hw}, we hypothesize the following criterion; in the case that Q_{hw} is quite high for a particular hw, even if the hw is *not* actually seen in T^A, it should be treated as if it were found in T^A.
6. Finally, modify the conventional method of MAP adaptation in Eq.(5) as follows:

$$P(w|h) = \frac{N^I_{hw} + \omega N^A_{hw} + \alpha Q_{hw}}{\sum_w (N^I_{hw} + \omega N^A_{hw} + \alpha Q_{hw})} \quad (11)$$

   where α is weight to control the contribution of Q_{hw}. In this equation, αQ_{hw} is considered to be frequency of potentially observable word sequence hw.

3.2. Expected effects of the proposed method

As mentioned above, the proposed method can handle unseen word sequence hw as if it were observed in the DD text. Therefore, the adaptation performance will be expected to be improved in the following two cases:

A) The amount of DD text is small and the original MAP adaptation is not effective.

B) DD adaptation text and test set data belong to somewhat *different* domains.

Both of the above cases may be often found in a real world situation. Especially in case B, we expect that similar text can make up for the shortage of DD data.

4. EXPERIMENTS OF N-GRAM ADAPTATION

4.1. Conditions of experiments

In order to evaluate the effectiveness of the proposed method, experiments of adapting n-gram models were carried out for Japanese texts. As DI and DD texts, two years' Mainichi newspaper corpus and "Peter Pan" (originally written by J. M. Barrie) were adopted, respectively. We selected three sets of sentences with different sizes out of the DD text, 133, 526 and 699 sentences, to investigate the dependence of the proposed method on adaptation data size. For test set data, we used not only "Peter Pan" but also "The Little Match Girl" (originally written by H. C. Andersen), which is close to but somewhat different from "Peter Pan" in terms of vocabulary and sentence styles. The quantitative difference between the two novels is described in section 4.3.

In the experiments, we used CMU-Cambridge statistical language modeling toolkit to build bi-gram models and to calculate perplexity. Vocabulary size of the bi-gram models was set to 20K, and unknown word rate was about 8% for each test set data. The size of CWL was set to 8,000 words, which were selected from the newspaper corpus by a mutual information technique[2]. In step 2 and 3 of the algorithm, we divided text data sentence by sentence because preliminary experiments showed us that sentence-based segmentation gave the best performance compared to article- or paragraph-based segmentation.

4.2. Experiments with small adaptation data

Adaptation experiments were performed, where DD text and test set data belong to the same domain but the size of DD text is small. Some sentences in "Peter Pan" were used as DD adaptation data and other ones in the same novel were utilized as test set data. Before the adaptation experiments, the optimal value of ω in the baseline MAP adaptation (Eq.(5)) was first obtained. Then, the proposed method was compared to the optimized MAP adaptation in terms of perplexity reduction.

4.2.1. Optimal weight of the baseline MAP adaptation

The optimal weight ω of the MAP adaptation was experimentally calculated by using "Peter Pan" as DD adaptation data and test set data. Fig.1 shows the reduction of adjusted perplexity (APP) as a function of MAP weight ω separately for each case of DD adaptation text size. It is found that the APP reduction strongly depends upon the DD text size but that the optimal weight can be found around 3000 in every case. In the following experiments, 3000, 2500 and 2500 is assigned to ω in the case of 133, 526 and 699 DD sentences, respectively.

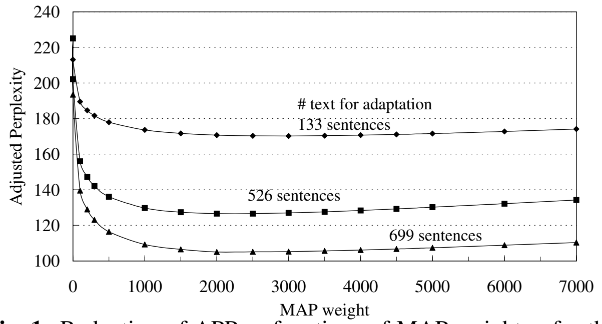

Fig. 1. Reduction of APP as functions of MAP weight ω for the three cases of amount of DD adaptation sentences. Both of adaptation data and test set data are selected from "Peter Pan."

4.2.2. Comparison between the proposed method and the optimized MAP adaptation

The performance of the proposed method was compared to that of the optimized MAP adaptation in terms of APP reduction. APP of test set data is shown in Figs. 2 to 4 as a function of α and λ for the three cases of 133, 526 and 699 DD sentences. In the figures, $\alpha = 0$ corresponds to the optimized MAP adaptation, which is the baseline of the experiments. In every figure, $(\alpha, \lambda) = (10^{10}, 0.01)$ clearly gives us maximal decrease of APP approximately by 10%. If it is allowed to roughly estimate the number of additionally required sentences to realize the improved performance only with the baseline MAP adaptation, about 150, 70 and 40 sentences must be added to DD text in the order of Figs. 2 to 4. Further, it should be noted that the rate of the improvement is constant among the three cases, which implies that the improvement of adaptation performance by using the proposed method is independent of the amount of DD text.

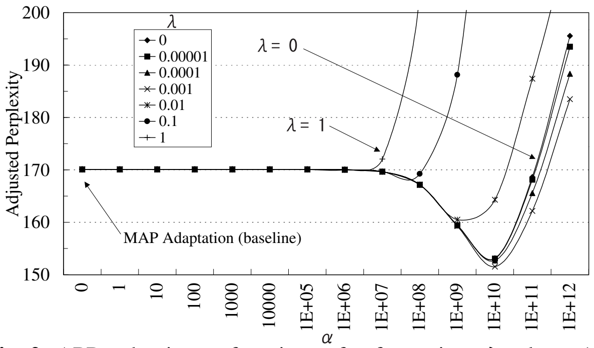

Fig. 2. APP reduction as functions of α for various λ values. 133 DD sentences. ("Peter Pan")

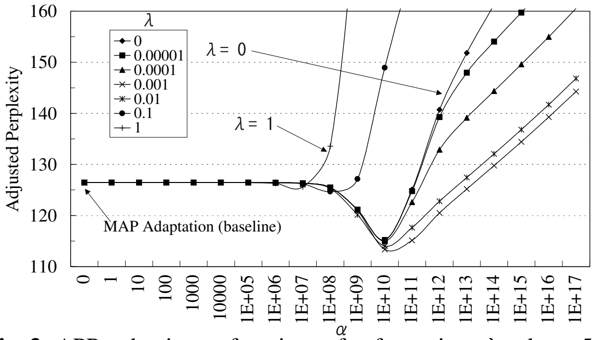

Fig. 3. APP reduction as functions of α for various λ values. 526 DD sentences. ("Peter Pan")

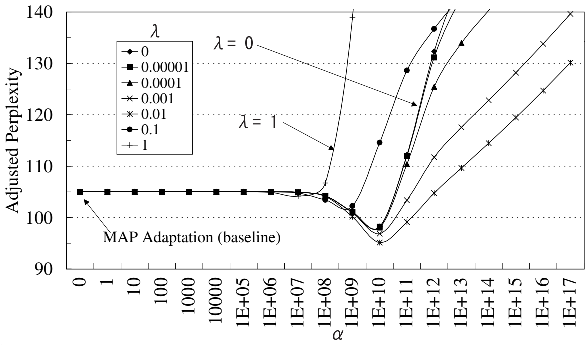

Fig. 4. APP reduction as functions of α for various λ values. 699 DD sentences. ("Peter Pan")

4.3. Experiments with adaptation and testing data of different domains

In this experiment, test set data was selected so that its linguistic content was somewhat different from that of DD adaptation text. This is because we expected that indirect correlation between words would enlarge the generality or robustness of adapted models compared to the model adapted only with MAP. Here, "The Little Match Girl" was used as test set data, while "Peter Pan" was adopted as DD adaptation data.

4.3.1. Optimal weight of the baseline MAP adaptation

As in section 4.2, weight ω in Eq.(5) was optimized, and APP of "The Little Match Girl" sentences was examined by bi-gram models adapted to "Peter Pan" sentences. Results are shown in Fig.5. We find that the optimal value of ω is ranged from 1,000 to 2,500 in every case. However, APP reduction is much smaller than that shown in section 4.2.1. Especially, larger ω clearly degrades the performance, which is definitely interpreted as the difference in linguistic content between the two novels; while "Peter Pan" is a story with a number of conversations, "The Little Match Girl" is a children's fairy tale. Further it should be noted that the figure also shows the difficulty of setting ω optimally when DD text and test set data belong to different domains.

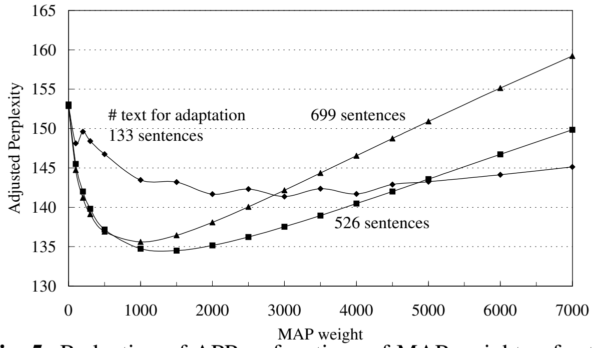

Fig. 5. Reduction of APP as functions of MAP weight ω for the three cases of amount of DD adaptation sentences. DD text was selected from "Peter Pan" and test set was selected from "The Little Match Girl."

4.3.2. Comparison between the proposed method and the baseline MAP adaptation

Experimental results are shown in Fig.6, where ω was set to 2,500 and the number of DD sentences was fixed to 526. The figure shows that the proposed method can reduce APP approximately by 10%, similar to the case of Fig.3. The proposed method can still effectively reduce APP even when DD text and test set data belong to different domains. It is quite surprising that Fig.6 shows the new n-gram models can better deal with test set data of another domain compared to the original MAP adapted models. This is because the proposed method *indirectly* treats unseen word sequence hw with high Q_{hw} as if it were actually observed.

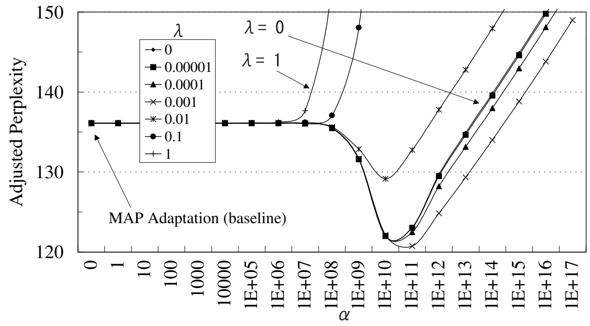

Fig. 6. APP reduction as functions of α for various λ values when DD adaptation text was selected from "Peter Pan" and test set was selected from "The Little Match Girl."

5. EXPERIMENTS OF SPEECH RECOGNITION

Bi-gram models adapted with the proposed method were examined in large vocabulary speech recognition. Japanese tied-state triphone set (totally 3K distinct states) and JULIUS v3.1 were used as speaker-independent acoustic models and a decoder, both of which are provided by CSR consortium[5]. First, bi-gram models were trained with two years' newspaper corpus as the initial language models. Then, two types of optimally adapted bi-gram models were built; models of the baseline MAP adaptation and those of the proposed method. Both sets of models were adapted to "Peter Pan". Since the experiment was designed only to evaluate the performance of language modeling, the output of the fist pass decoding was compared between the two methods. This is why we adopted bi-gram models. As for test set data, we collected speech samples of the two different domains, "Peter Pan" and "The Little Match Girl", which were spoken by one male speaker. Speech recognition results are shown in Table.1, where the number of DD adaptation sentences was fixed to 526. The table shows that word recognition rate and word accuracy are improved by a few percentages in both cases.

Table 1. Speech Recognition Results

	Peter Pan		The Little Match Girl	
	baseline	proposed	baseline	proposed
Correct (%)	73.74	75.05	67.87	69.76
Accuracy (%)	68.44	69.94	62.94	65.09
substitutions (%)	22.88	21.77	27.32	25.82
deletions (%)	3.38	3.19	4.81	4.41
insertions (%)	5.31	5.11	4.94	4.68

6. CONCLUSIONS

In this paper, we have proposed a heuristic method to improve MAP adaptation of n-gram language model by using indirect correlation of distant words. Experimental results show that the proposed method is more effective than conventional MAP adaptation, not only when text data size for adaptation is very small, but also when only text data close to but strictly different from the target domain is available. Besides, the relative improvement of the performance from the baseline method is shown to be independent of the amount of adaptation data. We can conclude that this method has significant robustness in terms of both quantity and quality of available data for adaptation.

7. REFERENCES

[1] M.Federico, "Bayesian Estimation Methods for N-gram Language Model Adaptation," *Proc.ICSLP-96*, vol. 1, pp. 240–243, 1996.

[2] K.Sasaki et al., "Rapid Adaptation of N-gram Language Models Using Inter-word Correlation for Speech Recognition," *Proc.ICSLP-2000*, vol. 4, pp. 508–511, 2000.

[3] H.Masataki et al., "Task Adaptation Using MAP Estimation in N-gram Language Modeling," *Proc.ICASSP-97*, vol. 2, pp. 783–786, 1997.

[4] A.Ito et al., "Evaluation of Task Adaptation Using N-gram Count Mixture," *The Transactions of the IEICE*, vol. J83-D-II, no. 11, pp. 2418–2427, 2000, (in Japanese).

[5] http://www.lang.astem.or.jp/CSRC/ .