# TASK ADAPTATION USING MAP ESTIMATION IN N-GRAM LANGUAGE MODELING

*Hirokazu Masataki*† *Yoshinori Sagisaka*† *Kazuya Hisaki*‡ *Tatsuya Kawahara*‡

†ATR Interpreting Telecommunications Research Labs.
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan

‡Department of Information Science, Kyoto University
Sakyo-ku, Kyoto 606-01, Japan

## ABSTRACT

This paper describes a method of task adaptation in N-gram language modeling, for accurately estimating the N-gram statistics from the small amount of data of the target task. Assuming a task-independent N-gram to be a-priori knowledge, the N-gram is adapted to a target task by MAP (maximum a-posteriori probability) estimation. Experimental results showed that the perplexities of the task adapted models were 15% (trigram), 24% (bigram) lower than those of the task-independent model, and that the perplexity reduction of the adaptation went up to 39 % at maximum when the amount of text data in the adapted task was very small.

## 1. INTRODUCTION

In continuous speech recognition, N-gram language models have been widely used as effective linguistic constraints to reduce search efforts [1][2]. However, large amounts of text data are needed to obtain reliable results with N-grams. In order to cope with data sparseness, some smoothing techniques [3][4][9], or techniques to reduce the number of parameters [5][6] have been proposed. However, fairly large amounts of text data are needed if these techniques are used, and language data collection is a crucial problem in the application of current speech recognition technology.

As each task has different N-gram characteristics, language data from other tasks cannot be used as same data simply to increase data numbers. To use these data properly in statistical sense, a task adaptation technique is needed to accurately estimate N-gram statistics of the current task from small data with the good use of language corpora in other tasks.

In this paper, we propose a task adaptation of N-gram language model using MAP estimation. This method employs task-independent N-grams as a-priori knowledge, and data of the target task as a-posteriori knowledge. By using MAP estimation, the a-priori knowledge and the a-posteriori knowledge are combined in proportion to the data size, and stable parameter estimation would be possible compared with maximum likelihood estimation.

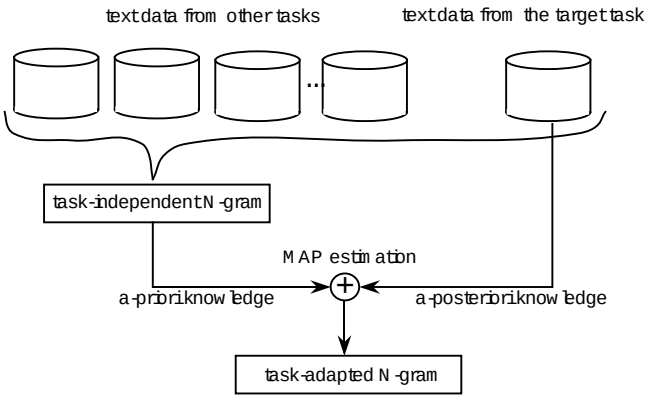


**Figure 1. Task Adaptation Using MAP Estimation**

## 2. N-GRAMS USING MAP ESTIMATION

Generally, a-priori probabilities of N-grams are calculated using ML (Maximum Likelihood) estimation. In the case of a word bigram, letting $x$ be the observed sequence, and $p$ $(= p(w_l|w_k))$ the a-priori probability of bigram, $p$ can be determined so as to maximize the likelihood function $f(x|p)$,

$$p_{ML} = \arg\max_{p} f(x|p) \tag{1}$$

When word $w_k$ occurs $N$ times and is followed by word $w_l$ $n$ times in the corpus, likelihood function $f(p)$ is described as follows,

$$f(p) = p^n(1-p)^{N-n} \tag{2}$$

By solving the maximizing condition $d\log f(p)/dp = 0$, the a-priori probability of bigram $p_{ML}$ is calculated as follows,

$$p_{ML} = n/N \tag{3}$$

Therefore, the a-priori probability of a word sequence which is not found in the sample data, is set to zero.

Using MAP(maximum a-posteriori probability) estimation, probability $p_{MAP}$ can be calculated so as to maximize the function $h(p|x)$.

$$p_{MAP} = \arg\max_{p} h(p|x) \tag{4}$$

Using Bayes' theorem, this equation can be modified as follows:

$$p_{MAP} = \arg\max_p f(x|p)g(p) \quad (5)$$

(where $g(p)$ is the a-priori distribution of the probability $p$.) Then, using MAP estimation, non-zero a-priori probability can be assigned from the a-priori distribution.

For an a-priori distribution, we adopt a beta distribution : $ap^{\alpha-1}(1-p)^{\beta-1}$ ($a$ is a coefficient for normalization). In this case, $p$ is calculated as follows using the definition of the MAP estimation,

$$p_{MAP} = \arg\max_p p^n(1-p)^{N-n} \cdot ap^{\alpha-1}(1-p)^{\beta-1} \quad (\equiv \arg\max_p L(p)) \quad (6)$$

$L(p)$ is maximized when $d\,logL(p)/dp = 0$. From this maximizing condition, the a-priori probability of bigram $p_{MAP}$ is calculated as follows:

$$p_{MAP} = \frac{n+\alpha-1}{N+\alpha+\beta-2} \quad (7)$$

The mean $\mu$ and the variance $\sigma^2$ of the beta distribution are known as [9],

$$\mu = \frac{\alpha}{\alpha+\beta} \quad , \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad (8)$$

From these equations, $\alpha$ and $\alpha+\beta$ can be expressed as,

$$\alpha = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu \quad , \quad \alpha+\beta = \frac{\mu(1-\mu)}{\sigma^2} - 1 \quad (9)$$

Therefore, a-priori probability of bigram can be solved by equation (7) and (9), from the mean and the variance of the a-priori distribution.

In these formulations, only bigram MAP-estimation is shown. It is trivial that this estimation can be applied to higher order N-grams by replacing the previous word $w_k$ by the previous (n-1) word sequence $w_1^{n-1}$.

## 3. TASK ADAPTATION USING MAP ESTIMATION

In order to apply MAP estimation to task adaptation, we regard a task-independent N-gram (generated by a large amount of text data including various tasks) as a-priori knowledge, and the data of the target task as a-posteriori knowledge.

When task-independent N-gram is assumed as a-priori knowledge, a-priori distribution is considered as the distribution of a-priori probability of N-gram of each task (Fig. 2). The a-priori probabilities of each task are calculated using maximum likelihood estimation. The values of mean

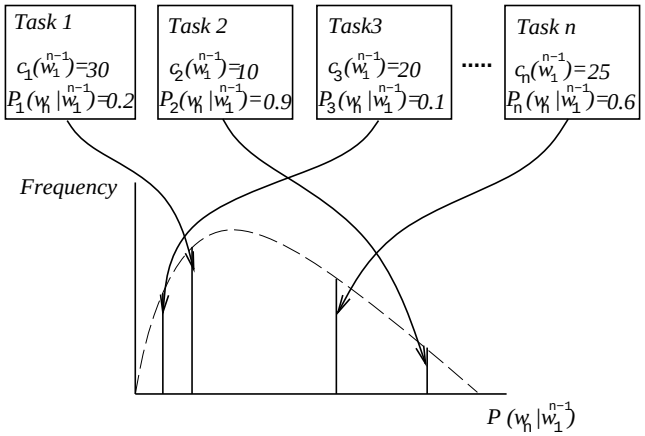


**Figure 2. A-Priori Distribution for Task Adaptation**

($\mu$) and variance ($\sigma^2$) of this distribution are calculated as follows,

$$\mu = \sum_i c_i(w_1^{n-1})p_i(w_l|w_1^{n-1}) / \sum_i c_i(w_1^{n-1}) \quad (10)$$

$$\sigma^2 = \sum_i c_i(w_1^{n-1})p_i(w_l|w_1^{n-1})^2 / \sum_i c_i(w_1^{n-1}) - \mu^2 \quad (11)$$

where $c_i(w_1^{n-1})$ is a frequency at which word sequence $w_1^{n-1}$ occurs in task $i$ and $p_i(w_l|w_1^{n-1})$ represents an a-priori probability of word sequence $w_1^{n-1}$ to $w_l$ of the N-gram of task $i$.

When the text data of the target task are assumed as the a-posteriori knowledge, the values of $n$ and $N$ in the previous section are expressed as follows,

- $N$: frequency of word sequence $w_1^{n-1}$ in the text of the target task.
- $n$: frequency of word sequence $w_1^n$ in the text of the target task.

By putting these values ($\mu$, $\sigma^2$, $N$ and $n$) into equations (7) and (9), the a-priori probabilities $p_{MAP}$ of the task-adapted N-grams can be obtained.

## 4. SMOOTHING ALGORITHM USING BACK-OFF METHOD

In the previous section, we described N-gram task adaptation using MAP estimation. There still remains two problems to use it as a language model. The first problem is that, if we use a large amount of task-independent text data, some word sequences may not be found in them. In this case, a-priori probability of N-gram is 0 even if using MAP estimation. Non-zero probabilities are to be assigned to unseen data using smoothing technique. The second problem is that, the sum of a-priori probabilities of N-grams can

not be set to 1, as each a-priori probability is calculated dependently using MAP estimation. Though the sum of a-priori probability is not to be normalized for its application to continuous speech recognition, it should be normalized to calculate perplexity score accurately. In order to solve these problems, we use the idea of the back-off smoothing method[4].

If the word sequence $w_1^n$ is found in the task-independent data, a-priori probability $p_{MAP}$ is calculated using the task adaptation method shown in the previous sections. Then a-priori probabilities are discounted using Turing's estimation. The discount coefficient is calculated using the frequency of the word sequence in the task-independent data. A surplus of probability caused by the discounting is divided to word sequences that are not found in the task-independent text data, according to the a-priori probability of the (n-1)-gram. Summarizing these procedures, the smoothing method is described as follows,

$$P_s(w_n|w_1^{n-1}) = \begin{cases} \tilde{P}(w_n|w_1^{n-1}) & (c(w_1^{n-1}) > 0) \\ \alpha(w_1^{n-1}) \cdot P_s(w_n|w_2^{n-1}) & (c(w_1^{n-1}) = 0, c(w_2^{n-1}) > 0) \\ P_s(w_n|w_2^{n-1}) & (c(w_1^{n-1}) = 0, c(w_2^{n-1}) = 0) \end{cases} \quad (12)$$

where,

$$\tilde{P}(w_n|w_1^{n-1}) = \frac{c(w_1^n)+1}{c(w_1^n)} \cdot \frac{n_{c(w_1^n)+1}}{n_{c(w_1^n)}} \cdot p_{MAP}(w_n|w_1^{n-1}) \quad (13)$$

($n_r$: number of words which occurred in the text exactly $r$ times)

$$\alpha(w_1^{n-1}) = \frac{1 - \sum_{w_n:c(w_1^n)>0} \tilde{P}(w_n|w_1^{n-1})}{1 - \sum_{w_n:c(w_1^n)>0} \tilde{P}(w_n|w_2^{n-1})} \quad (14)$$

Through out these calculations non-zero probabilities are assigned to unseen N-grams, and the sum of the a-priori probabilities is set to 1 by the normalization of the equation (14).

## 5. EXPERIMENTAL RESULTS

In order to evaluate the effectiveness of our proposed task adaptation method, an experiment was conducted using the ATR spontaneous speech database on travel arrangements [10]. This database is composed of 15 tasks (Table 1). Currently, this database consists of 1,098 dialogues with 449,070 words in total (vocabulary 6,786). We randomly selected about a quarter of the dialogues for our test set; at least one dialogue was selected from each task. The remaining data were used for training set. Three different models were compared.

**Table 1. List of Tasks**

| No. | Dialogues | Content |
|---|---|---|
| 1) | 491 | Hotel Service |
| 2) | 351 | Hotel Reservation |
| 3) | 50 | Inquiry on Sightseeing Bus Tours |
| 4) | 36 | Reservation of Meeting Room |
| 5) | 28 | Inquiry on Means of Transportation |
| 6) | 24 | Hotel Consulting |
| 7) | 22 | Airline Reservation |
| 8) | 22 | Inquiry on Bus or Train Schedule |
| 9) | 20 | Inquiry on Car Rental |
| 10) | 14 | Concert Reservation |
| 11) | 12 | Restaurant Reservation |
| 12) | 8 | Trouble and Lost Items |
| 13) | 8 | Road Guide |
| 14) | 8 | Meal Order |
| 15) | 4 | Shopping |

- Task-Independent Model:
  N-grams trained using all data of training set.
- Task-Dependent Model:
  N-grams trained using the data of target task only.
- Task-Adapted Model:
  N-grams adapted from the task-independent model to a target task using the proposed method.

Table 2 shows the perplexities of the test set for the three different models, on word bigrams and trigrams.

As shown in the table 2, the perplexities of task-adapted models are lower than those of the task-independent models and the task-dependent models, on both bigrams and trigrams. These results show that the proposed task adaptation is efficient.

As for bigrams, the perplexities of task-dependent models are lower than those of the task-independent models for most of the tasks. This is because task-dependent models express better characteristics of the target task. As for trigrams, on the other hand, the perplexities of the task-independent models are higher than those of the task-dependent models for most of the tasks. This is because the problem of sparse data is so severe that parameters of the task-dependent models can not be correctly estimated by a small amount of data. Using task adaptation, the problem of sparse data can be solved by training a large amount of task-independent training data, and the characteristics of each task can be expressed well by the task adaptation.

The effectiveness of this task adaptation is remarkable when the amount of adapted text is small. On task 12), Perplexity of task adapted model is 39% (bigram), 30% (trigram) lower than those of task-independent models.

**Table 2. Perplexity of three different models for individual tasks**

| Task | Words | | Task-Independent Models | | Task-Dependent Models | | Task-Adapted Models | |
|---|---|---|---|---|---|---|---|---|
| | Training Set | Test Set | Bigram | Trigram | Bigram | Trigram | Bigram | Trigram |
| 1) | 136,175 | 42,698 | 23.168 | 17.948 | 22.923 | 18.260 | 22.085 | 17.515 |
| 2) | 118,124 | 38,697 | 14.837 | 10.071 | 13.842 | 9.941 | 13.402 | 9.612 |
| 3) | 19,471 | 6,610 | 26.523 | 17.383 | 23.910 | 17.196 | 20.684 | 14.705 |
| 4) | 15,302 | 5,075 | 31.270 | 24.693 | 38.164 | 32.811 | 29.280 | 24.470 |
| 5) | 10,791 | 2,983 | 24.164 | 16.544 | 21.774 | 16.574 | 18.328 | 13.656 |
| 6) | 8,802 | 2,999 | 17.122 | 11.192 | 14.661 | 11.350 | 12.540 | 9.127 |
| 7) | 8,617 | 2,722 | 21.106 | 14.181 | 18.358 | 14.656 | 15.274 | 11.383 |
| 8) | 8,537 | 2,193 | 21.134 | 14.288 | 14.077 | 11.177 | 13.351 | 10.523 |
| 9) | 8,567 | 2,528 | 25.149 | 18.154 | 25.897 | 20.743 | 20.443 | 16.097 |
| 10) | 5,036 | 1,608 | 16.582 | 10.820 | 14.060 | 10.931 | 11.368 | 8.148 |
| 11) | 5,326 | 1,439 | 12.970 | 8.867 | 12.261 | 9.611 | 9.564 | 6.935 |
| 12) | 3,578 | 1,165 | 32.921 | 19.402 | 25.232 | 18.385 | 19.921 | 13.399 |
| 13) | 2,378 | 1,075 | 30.294 | 22.416 | 32.757 | 31.567 | 21.541 | 19.338 |
| 14) | 2,572 | 908 | 35.490 | 27.108 | 45.853 | 41.285 | 28.155 | 23.707 |
| 15) | 1,750 | 509 | 44.088 | 34.214 | 47.324 | 44.573 | 31.854 | 27.896 |
| total | | | 25.121 | 17.819 | 24.740 | 20.604 | 19.186 | 15.101 |

## 6. CONCLUSIONS

In this paper, we proposed a task adaptation method using MAP estimation. Experimental results showed the effectiveness of the task adaptation. This effectiveness was greater when the amount of adaptation data was small. Therefore, effective N-grams can be generated from a very small amount of data. We are planning to apply this language model to continuous speech recognition.

## REFERENCES


[1] L. R. Bahl, F. Jelinek and R.L. Mercer: "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, pp. 179-190, 1983.

[2] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga and Y. Sagisaka: "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," ICASSP'96, pp. 145-148, 1996.

[3] F. Jelinek and R. L. Mercer: "Interpolated Estimation of Markov Source Parameters from Sparse Data," Workshop Pattern Recognition in Practice, pp. 381-397, 1980.

[4] S. M. Katz: "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," IEEE Trans. on Acoustics, Speech, and Signal Processing, pp. 400-401, 1987.

[5] P. F. Brown et al.: "Class-Based n-gram Models of Natural Language," Computational Linguistics, Vol. 18, No. 4, pp. 467-479, 1992.

[6] H. Masataki and Y. Sagisaka: "Variable-order N-gram generation by word-class splitting and consecutive word grouping", ICASSP96, pp. 188-191, 1996.

[7] S. Matsunaga, T. Yamada and K. Shikano: "Task adaptation in stochastic language models for continuous speech recognition," ICASSP92, 165-168, 1992.

[8] P. S. Rao, M. D. Monkowski and S. Roukos: "Language model adaptation via minimum discrimination information," ICASSP96, 165-168, 1996.

[9] T. Kawabata and M. Tamoto: "Back-off Method for N-gram Smoothing Based on Binominal Posteriori Distribution", ICASSP96, 192-195, 1996.

[10] T. Morimoto, et al.: "A Speech and Language Database for Speech Translation Research," ICSLP94, pp. 1791-1794, 1994.