

Adapting Language Models for Frequent Fixed Phrases by Emphasizing N-gram Subsets

Tomoyosi AKIBA[†], Katunobu ITOU[†], Atsushi FUJII[‡]

[†] National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, 305-8568, JAPAN, E-mail: t-akiba@aist.go.jp

[‡] University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, JAPAN

Abstract

In support of speech-driven question answering, we propose a method to construct N-gram language models for recognizing spoken questions with high accuracy. Question-answering systems receive queries that often consist of two parts: one conveys the query topic and the other is a fixed phrase used in query sentences. A language model constructed by using a target collection of QA, for example, newspaper articles, can model the former part, but cannot model the latter part appropriately. We tackle this problem as task adaptation from language models obtained from background corpora (e.g., newspaper articles) to the fixed phrases, and propose a method that does not use the task-specific corpus, which is often difficult to obtain, but instead uses only manually listed fixed phrases. The method emphasizes a subset of N-grams obtained from a background corpus that corresponds to fixed phrases specified by the list. Theoretically, this method can be regarded as maximizing a posteriori probability (MAP) estimation using the subset of the N-grams as a posteriori distribution. Some experiments show the effectiveness of our method.

1. Introduction

Question answering (QA) was first evaluated largely at TREC-8[11]. The goal in the QA task is to retrieve small snippets of text that contain the actual answer to a question rather than the document lists traditionally returned by text retrieval systems. We are trying to extend question-answering systems as traditional text retrieval systems[3] that accept spoken queries. In this paper, we address issues related to language modeling for the speech recognition subsystem of speech-driven question-answering systems.

Question-answering systems receive queries that often consist of a part that conveys various query contents about, for example, newspaper articles, and a part that represents a fixed phrase for query sentences. For example, the following query may be submitted.

seN / kyu- / hyaku / nana / ju- / roku / neN / ni / kasei / ni / naN / chakuriku / shita / taNsaki / wa / naN / to / yu- / namae / desu / ka
(What was the name of the spacecraft that landed safely on Mars in 1976?)

The first half of the query, i.e., “seN kyu- hyaku nana ju- roku neN ni kasei ni naN chakuriku shita taNsaki wa (the spacecraft that landed safely on Mars in 1976)”, conveys the topic of the retrieval, and is best dealt with by using an N-gram model trained with the target documents of QA systems. In this paper,

newspaper articles are used for the target documents[4]. On the other hand, the latter half of the query, i.e., “naN to yu- namae desu ka (What was the name?)”, is a fixed phrase typically used in interrogative questions, but is not very frequent in newspaper articles. Thus, we require language models that can deal with both types of fragments.

Note that recognizing the fixed phrases with high accuracy is crucial to success in question answering, because they convey clues to determine the query type[6]. For example, a fixed phrase might indicate that the answer should be a name of some object as in the last example, while another might indicate that the answer should be a date of some event (e.g., in English, “On what date was...”).

There has been work on language model adaptation in which language models for a specific task were constructed from both a large general-purpose corpus and a relatively small task-specific corpus. Using this approach, we can construct a language model for question answering from both a large number of generic newspaper articles and a small number of query sentences for QA.

One issue that should be considered when using this approach is how the task-specific corpus should be acquired. If the corpus does not exist already, it must be collected somehow or other, and collecting a new corpus directly from practical use is always expensive, even if the resulting corpus is small. Alternative methods have been proposed to obtain a considerable amount of task-specific corpus indirectly, including such methods as: automatically generating sentences from a hand-made task-specific grammar[5]; incorporating a task-specific grammar-based model into the background N-gram[1]; and utilizing the results of speech recognition using a general-purpose language model[8, 9].

In our case, the number of the fixed phrases used in QA is small enough for all the patterns to be enumerated by hand. Thus we can inexpensively prepare a list of phrases instead of collecting a corpus of query sentences. In this paper, we propose a method of constructing language models for question answering from a target collection (e.g., newspaper articles) and a list of the fixed phrases typically used in interrogative questions. The method emphasizes N-gram subsets corresponding to the fixed phrases and can be considered as a variant of a maximum a posteriori probability (MAP) estimation using the N-gram subsets of a background corpus as an a posteriori distribution.

2. The Method

Figure 1 illustrates our proposed method of adapting a language model for fixed phrases. The list of fixed phrases is used to select the subset of N-grams related to the phrases. Then, adding the subset to the original N-grams produces the adapted model.

The second and third authors are also members of CREST, Japan Science and Technology Corporation.

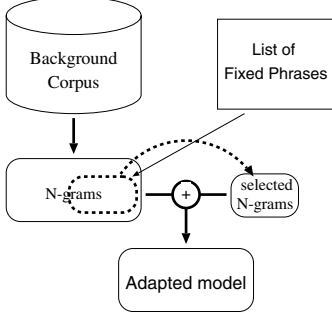


Figure 1: Language model adaptation using a list of fixed phrases

2.1. Language model adaptation for fixed phrases

Let S be a set of sentences. Let S_{FP} be a subset of S that consists only of sentences that have the fixed phrases specified by the list. Let P be a language model for generating sentences $s (\in S)$ obtained from a general-purpose background corpus. The aim of the language model adaptation for the fixed phrases is to obtain the adapted language model P' , which gives relatively high probabilities to the sentences $\hat{s} \in S_{FP}$ but preserves the order relations on the sentences $s \in S - S_{FP}$ as much as possible.

The generative probability that the sentence \hat{s} includes a fixed phrase $\hat{w}_p \dots \hat{w}_q$ is:

$$\begin{aligned}
P(\hat{s}) &= \prod_{i=1}^n P(w_i | w_1 \dots w_{i-1}) \\
&\approx P(w_1) P(w_2 | w_1) \dots P(w_i | w_{i-N+1}^{i-1}) \dots \\
&\dots P(\hat{w}_p | w_{p-N+1}^{p-1}) P(\hat{w}_{p+1} | w_{p-N+2}^{p-1} \hat{w}_{p-1}) \dots \\
&\dots P(\hat{w}_{p+N-2} | w_{p-1} \hat{w}_p^{p+N-3}) \dots \\
&\dots P(\hat{w}_{p+N-1} | \hat{w}_p^{p+N-2}) \dots \\
&\dots P(\hat{w}_q | \hat{w}_{q-N+1}^{q-1}) P(w_{q+1} | \hat{w}_{q-N+2}^q) \dots \\
&\dots P(w_m | w_{m-N+1}^{m-1})
\end{aligned}$$

The following components of the above equation are important in obtaining P' :

$$P(\hat{w}_p | w_{p-N+1}^{p-1}) \dots P(\hat{w}_{p+N-2} | w_{p-1} \hat{w}_p^{p+N-3}), \quad (a)$$

$$P(\hat{w}_{p+N-1} | \hat{w}_p^{p+N-2}) \dots P(\hat{w}_q | \hat{w}_{q-N+1}^{q-1}). \quad (b)$$

The component (a) corresponds to the generative probabilities of the prefix words of the fixed phrases, each of which, in its condition part, has one or more words other than those that consist of the fixed phrases. The component (b) corresponds to the generative probabilities of the intermediate words of the fixed phrases, each of which has only the words of the fixed phrases in its condition part.

The adapted model P' is calculated using the following two steps.

- i. Revise the maximum likelihood estimates of P :

$$P_{ML(1)}(w_i) P_{ML(2)}(w_i | w_{i-1}) \dots P_{ML(N)}(w_i | w_{i-N+1}^{i-1}),$$

which are calculated for each length $n (1 \leq n \leq N)$.

- ii. Apply back-off smoothing to integrate the revised ML estimates $P'_{ML(n)}(w_i | w_{i-n+1}^{i-1}) (1 \leq n \leq N)$.

The proposed method emphasizes only the carefully selected $P_{ML(n)}$ s that are meaningful for following back-off smoothing calculation, to make the produced model harmless to the other generative probabilities assigned to the sentences that do not have the fixed phrases.

2.2. Revision of the maximum likelihood estimate

For all lengths $n (1 \leq n \leq N)$, the maximum likelihood estimates $P_{ML(n)}(w_i | w_{i-n+1}^{i-1})$ of the N-gram probability P obtained from the background corpus are revised to P'_{ML} by the following procedure.

- (1). If the postfix $w_{i-k+1} \dots w_i (1 \leq k < n)$ of the word sequence $w_{i-n+1} \dots w_i$ is equal to the prefix $\hat{w}_p \dots \hat{w}_{p+k-1}$ of one of the fixed phrases $\hat{w}_p \dots \hat{w}_q$, then emphasize the P_{ML} as follows:

$$\begin{aligned}
P'_{ML(n)}(\hat{w}_{p+k-1} | w_{p-n+k}^{p-1} \hat{w}_p^{p+k-2}) &= \\
\beta_n(w_{p-n+k}^{p-1} \hat{w}_p^{p+k-2}) &\cdot \\
\gamma P_{ML(n)}(\hat{w}_{p+k-1} | w_{p-n+k}^{p-1} \hat{w}_p^{p+k-2}) &
\end{aligned}$$

Otherwise, go to step (2).

For example, for tri-grams, for all context word sequences $w_{p-2} w_{p-1}$, we have:

$$\begin{aligned}
P'_{ML(3)}(\hat{w}_{p+1} | w_{p-1} \hat{w}_p) &= \\
\beta_3(w_{p-1} \hat{w}_p) \cdot \gamma P_{ML(3)}(\hat{w}_{p+1} | w_{p-1} \hat{w}_p) & \\
P'_{ML(2)}(\hat{w}_{p+1} | \hat{w}_p) &= \\
\beta_2(\hat{w}_p) \cdot \gamma P_{ML(2)}(\hat{w}_{p+1} | \hat{w}_p) & \\
P'_{ML(3)}(\hat{w}_p | w_{p-2} w_{p-1}) &= \\
\beta_3(w_{p-2} w_{p-1}) \cdot \gamma P_{ML(3)}(\hat{w}_p | w_{p-2} w_{p-1}) & \\
P'_{ML(2)}(\hat{w}_p | w_{p-1}) &= \\
\beta_2(w_{p-1}) \cdot \gamma P_{ML(2)}(\hat{w}_p | w_{p-1}) & \\
P'_{ML(1)}(\hat{w}_p) &= \beta_1(\epsilon) \cdot \gamma P_{ML(1)}(\hat{w}_p)
\end{aligned}$$

- (2). If the word sequence $w_{i-n+1} \dots w_i$ is equal to the subsequence $\hat{w}_{i-n+1} \dots \hat{w}_i$ of one of the fixed phrases $\hat{w}_p \dots \hat{w}_q$ then emphasize only the longest N-gram probability $P_{ML(N)}$ as follows:

$$\begin{aligned}
P'_{ML(N)}(\hat{w}_i | \hat{w}_{i-N+1}^{i-1}) &= \\
\beta_N(\hat{w}_{i-N+1}^{i-1}) \cdot \gamma P_{ML(N)}(\hat{w}_i | \hat{w}_{i-N+1}^{i-1}) &
\end{aligned}$$

Otherwise, go to step (3).

For example, for tri-grams, only the tri-gram probability should be emphasized:

$$\begin{aligned}
P'_{ML(3)}(\hat{w}_i | \hat{w}_{i-2} \hat{w}_{i-1}) &= \\
\beta_3(\hat{w}_{i-2} \hat{w}_{i-1}) \cdot \gamma P_{ML(3)}(\hat{w}_i | \hat{w}_{i-2} \hat{w}_{i-1}) &
\end{aligned}$$

- (3). For all $n (1 \leq n \leq N)$, the revised probability is:

$$\begin{aligned}
P'_{ML(n)}(w_i | w_{i-n+1}^{i-1}) &= \\
\beta_n(w_{i-n+1}^{i-1}) \cdot P_{ML(n)}(w_i | w_{i-n+1}^{i-1}) &
\end{aligned}$$

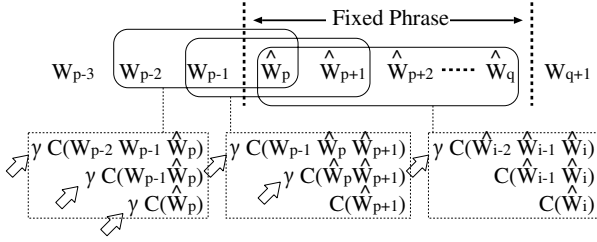


Figure 2: Emphasizing N-gram Counts (for tri-grams)

where $\gamma (\geq 1)$ is a multiplier that emphasizes the selected N-grams, and $\beta_1(\epsilon) \cdots \beta_N(w_{i-N+1}^{i-1})$ are normalized coefficients that make the probabilities sum to one.

This can be seen as the task adaptation process by maximum a posteriori probability (MAP) estimation[2] using the N-gram subsets corresponding to the fixed phrases as task specific data for adaptation, because P'_{ML} is equivalent to the maximum likelihood estimate calculated from the N-gram counts C'_n of each length $n(1 \leq n \leq N)$ obtained by emphasizing the selected subset of the original N-gram counts C , as shown in Fig. 2.

$$P'_{ML(n)}(w_i | w_{i-n+1}^{i-1}) = \frac{C'_n(w_{i-n+1}^{i-1})}{\sum_{w_i} C'_n(w_{i-n+1}^{i-1})}$$

2.3. Back-off smoothing

Back-off smoothing integrates the revised ML probabilities $P'_{ML(n)}$ of each length n to produce the final adapted language model. Any back-off smoothing method can be applied, except that the discount coefficient should be calculated using the a priori knowledge of the adaptation, i.e., the N-gram counts obtained from the background corpus.

For example, for Witten-Bell smoothing [10], the following discount coefficient $d'_{w_{i-n+1}^{i-1}}$ should be used for the proposed method.

$$d'_{w_{i-n+1}^{i-1}} P'_{ML(n)}(w_i | w_{i-n+1}^{i-1}) = \frac{C_n(w_{i-n+1}^{i-1})}{\{\sum_{w_i} C_n(w_{i-n+1}^{i-1})\} + r(w_{i-n+1}^{i-1})}$$

where r is the number of different words appearing after the word context w_{i-n+1}^{i-1} in the background corpus.

3. Experimental Results

We extracted N-gram counts of 20,000 words that were obtained from newspaper articles collected over 111 months. As task-specific training data, we developed a word network for the Japanese fixed phrases used for question answering. From the network, we extracted a list of all the 172 fixed phrases that were acceptable to the network. Then we compared several adaptation methods including that mentioned in this paper. We applied Witten-Bell discounting[10] for all methods.

We first made the N-gram model from only the newspaper articles as the baseline (referred to as the *BASE* model). As a conventional MAP adaptation method[2], we mixed two sets of N-gram counts obtained from newspaper articles and the list of fixed phrases (magnified by w), and obtained the adapted model referred to as *MIX*. As the method proposed in this paper, we magnified N-gram counts corresponding to the fixed phrases in the N-gram of newspaper articles (by γ), and obtained the adapted model referred to as *EMP*. Finally, using the method that we had previously proposed [1], we integrated the N-gram

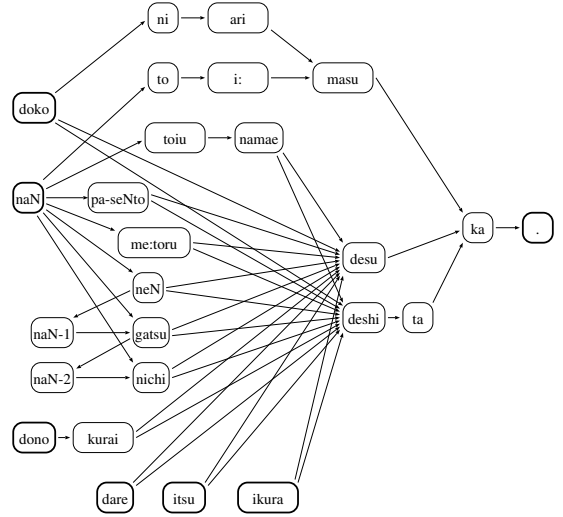


Figure 3: A word network for the Japanese fixed phrases frequently used in queries for QA

of newspaper articles and the word network for fixed phrases (magnified by γ), and obtained the adapted model referred to as *NET*.

We prepared 100 sentences from the newspaper articles (referred to as *NP*) and 50 query sentences for the QA system (referred to as *QA*), and these were recorded for four speakers (two men and two women). Though the word network was relatively small and had only 33 nodes (31 words), 36 of the 50 queries had the fixed phrases characterized in the network.

We used an existing N-gram decoder [7] for the recognition experiments. The language model weight and the insertion penalty were set to the best values for the newspaper (*BASE*) model. The results are shown in Fig. 4 and Fig. 5.

Figure 4 shows the relations between word error rate (WER) and the parameter (w or γ) with respect to both the target *QA* and *NP*. The best results with respect to *QA* of *BASE*, *MIX*, *EMP* and *NET* are obtained by adjusting the parameter to 16.9, 15.4, 13.8 and 14.7, respectively. The proposed model *EMP* outperformed the other models, while it did not worsen the WER for the sentences that did not have the fixed phrases (*NP*).

Figure 5 shows the difference between WERs for the first (referred to as *FH*) and latter (referred to as *LH*) half of the interrogative sentences (*QA*). We divided each sentence of *QA* into first and latter half by using a Japanese WH-word as the boundary (the latter half included the WH-word), and investigated the WERs of both halves separately. Note that the latter halves roughly correspond to the fixed phrases. It indicated that the proposed method (*EMP*) best reduced the WER corresponding to the fixed phrases (*LH*), while it did not worsen the WER for the other part of the input sentences (*FH*).

4. Conclusion

We have proposed methods for language model adaptation that enable recognition of spoken queries submitted to QA systems with high accuracy. The method does not require a task-specific corpus but, instead, uses a list of fixed phrases enumerated by hand. Our experiments showed that the method outperformed a conventional language model adaptation method in terms of the recognition accuracy. The proposed methods can be used for other task-adaptation problems in language modeling where the variation in expressions to be adapted is relatively small allowing for these expressions to be enumerated by hand without collecting a new text corpus.

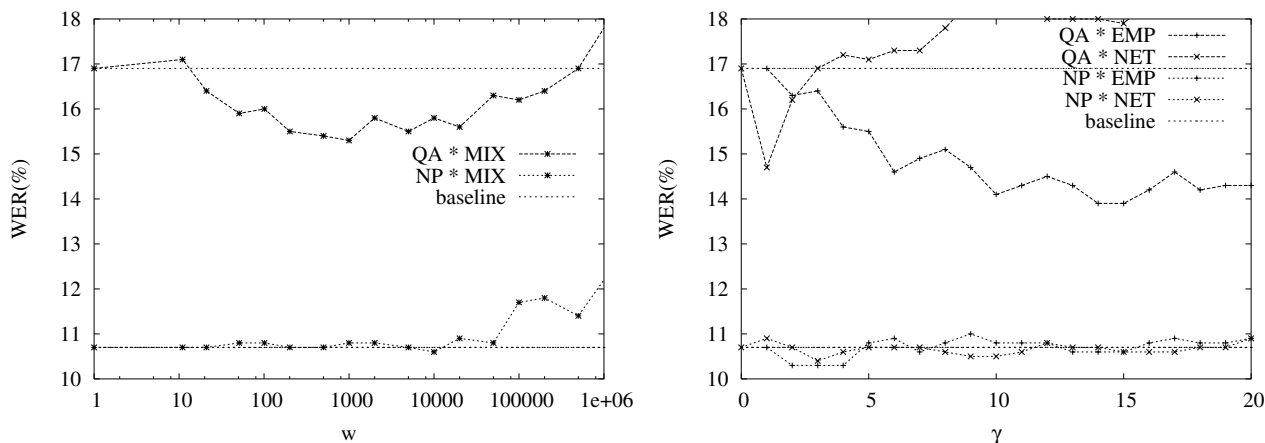


Figure 4: The relation between word error rate and the parameter.

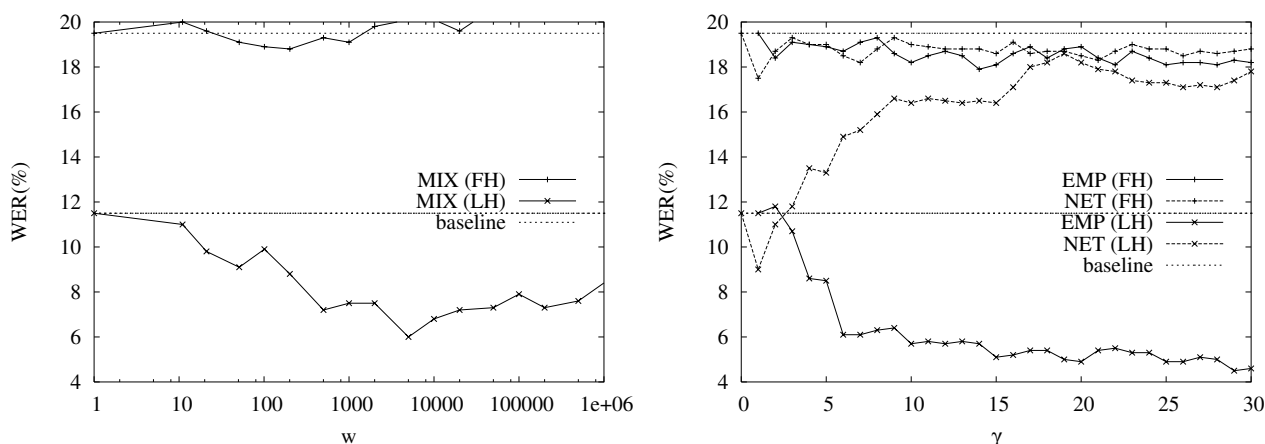


Figure 5: Word error rates for first and latter halves of sentences (QA).

5. References

- [1] T. Akiba, K. Itou, A. Fujii, and T. Ishikawa. Selective back-off smoothing for incorporating grammatical constraints into the n-gram language model. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 881–884, Denver, Colorado, Sept. 2002.
- [2] M. Federico. Bayesian estimation methods for n-gram language model adaptation. In *Proceedings of International Conference on Spoken Language Processing*, pages 240–243, Philadelphia, USA, 1996.
- [3] A. Fujii, K. Itou, and T. Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In A. R. Coden, E. W. Brown, and S. Srinivasan, editors, *Information Retrieval Techniques for Speech Applications (LNCS 2273)*, pages 94–104. Springer, 2002.
- [4] J. Fukumoto, T. Kato, and F. Masui. Question answering challenge (QAC-1) question answering evaluation at NTCIR workshop 3. In *Working Notes of the Third NTCIR Workshop Meeting*, pages 1–6, Tokyo, Japan, Oct. 2002.
- [5] L. Galescu, E. Ringger, and J. Allen. Rapid language model development for new task domains. In *Proceedings of International Conference on Language Resources and Evaluation*, pages 807–812, Granada, Spain, May 1998.
- [6] A. Ittycheriah, M. Franz, W.-J. Zhu, and A. Ratnaparkhi. IBM's statistical question answering system. In *Proceedings of the 9th Text Retrieval Conference*, pages 229–234, Maryland, 2000.
- [7] A. Lee, T. Kawahara, and K. Shikano. Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1691–1694, Aalborg, Denmark, Sept.
- [8] M. Mahajan, D. Beeferman, and X. D. Huang. Improved topic-dependent language modeling using information retrieval techniques. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, Phoenix, Arizona, March 1999.
- [9] T. Niesler and D. Willett. Unsupervised language model adaptation for lecture speech transcription. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 1413–1416, Denver, Colorado, Sept. 2002.
- [10] P. Placeway, R. Schwartz, P. Fung, and L. Nguyen. The estimation of powerful language models from small and large corpora. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, volume 2, pages 33–36, Minneapolis, USA, April 1993.
- [11] E. Voorhees and D. Tice. The TREC-8 question answering track evaluation. In *Proceedings of the 8th Text Retrieval Conference*, pages 83–106, Gaithersburg, Maryland, 1999.