Towards Speech-Driven Question Answering: Experiments Using the NTCIR-3 Question Answering Collection

Tomoyosi Akiba[†], Katunobu Itou[†,†††], Atsushi Fujii[††,†††], Tetsuya Ishikawa[††]
[†]National Institute of Advanced Industrial Science and Technology
Tsukuba Central 2, 1-1-1, Umezono, Tsukuba, 305-8568, Japan
[††]University of Library and Information Science
1-2 Kasuga Tsukuba, 305-8568, Japan
[†††]CREST, Japan Science and Technology Corporation
t-akiba@aist.go.jp

Abstract

We developed a method for producing statistical language models for speech-driven question answering, which recognizes spoken questions with high accuracy. Our method uses a target collection (i.e., a document set from which answers are derived) to extract N-grams, and adapts them to the question-answering task by way of frozen patterns typically used in interrogative questions. In addition, our method magnifies N-gram statistics corresponding to frozen patterns in the original N-gram. For the purpose of experiments, we used dictated questions in the NTCIR-3 QAC test collection, and showed that our method outperformed a conventional language model adaptation method in terms of the speech recognition accuracy.

1 Introduction

Question answering (QA) tasks were first evaluated largely at TREC-8[10]. The goal of a QA task is to retrieve small snippets of text that contain the actual answer to a question rather than lists of documents traditionally returned by text retrieval systems. We are trying to improve question answering systems to make them accept spoken queries, as traditional text retrieval systems do[4]. In this paper, we address some of the problems in language modeling by designing speech recognition models of speech-driven question answering systems.

Question answering systems receive queries that often consist of two parts – one that conveys various query information, for example, newspaper articles, and the other that represents a frozen pattern used in 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 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 as target documents[1]. The latter half of the query, i.e., "naN to yu- namae desu ka (What was the name ?)", is a frozen pattern that is typically used in interrogative questions but that is not very frequent in newspaper articles. Thus, we need language models that can deal with both types of fragments.

There are works on language model adaptation that construct language models for a specific task from both a large amount of general-purpose text corpus material and a relatively small amount of task-specific text corpus material. 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 frozen patterns used in interrogative questions. We also make the following additional assumptions: (1) the first half of the queries can be adequately modeled only by a language model created from newspaper articles, (2) the latter half of the queries consists of words that are not specific and are generic enough to appear in newspaper articles, (3) the diversity of the latter half of the queries is small enough for all the patterns to be enumerated by hand, (4) a query as a whole includes both halves, thus a language model must adequately deal with combination of these fragments.

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 frozen patterns typically used in interrogative questions. The method magnifies N-gram counts corresponding to the frozen patterns in the original N-gram.

2 Language Model Adaptation by N-gram Counts Mixture

Methods of language model adaptation for specific task use relatively small amount of task-specific text and a large amount of generic text. Even though the amount of task-specific text is small, the cost of collecting text material for a specific task is not. Few studies have attended collecting a text corpus by using methods other than conventional ones. Galescu et al. used a context-free grammar to extract a text corpus[5]. Okato et al. used example sentences written by hand[8]. In line with assumption (3) in section 1, we enumerated frozen patterns used in interrogative questions by listing them directly by hand and by constructing a grammar for the patterns.

A simple way of language model adaptation is by mixing weighed sets of N-gram counts, each of which is obtained separately from a text corpus [3, 6]. The flow of the process is illustrated in fig.1. Such a set of N-gram counts obtained directly and only from a text corpus is characterized by consistency in the N-gram length.

Definition 1 (Consistency in the N-gram length) *A set of N-gram counts is consistent with respect to its length if and only if a shorter set of N-gram counts can be uniquely calculated from a longer one.*

Corollary 1 *Any set of N-gram counts obtained only from text corpora is consistent with respect to its length.*

> **(Example)** Consider a set of tri-gram counts $C_{(3)}$ obtained from a text corpus. The bi-gram count $C_{(2)}$ of the words $w_p w_q$ can be calculated uniquely from the tri-gram counts as follows.
>
> $$\begin{aligned} C_{(2)}(w_p w_q) &= \sum_{w_i} C_{(3)}(w_i w_p w_q) \\ &= \sum_{w_i} C_{(3)}(w_p w_q w_i) \end{aligned}$$
>
> In the same way, the uni-gram count $C_{(1)}$ of the word w_p can be calculated uniquely as follows.
>
> $$\begin{aligned} C_{(1)}(w_p) &= \sum_{w_i} C_{(2)}(w_i w_p) \\ &= \sum_{w_i} C_{(2)}(w_p w_i) \end{aligned}$$

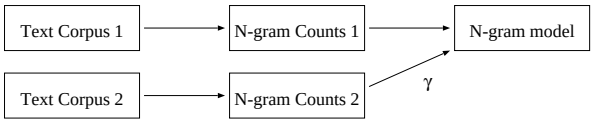

Figure 1. Adaptation using N-gram counts mixture

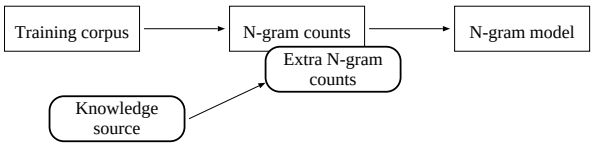

Figure 2. Our framework with extra N-gram counts

This characteristic reflects the nature of a text corpus as a single word sequence.

Using this approach and taking into account assumptions (1) and (2) in section 1, we can construct a model for question answering by mixing a set of N-grams obtained from newspaper articles for the first half of the queries, and a set of N-grams obtained from frozen patterns used in interrogative questions for the latter half of the queries. However, based on assumption (4) in section 1, the resulting model must be able to deal with a combination of these fragments. Because neither text contains a combination of these fragments, this method cannot be used to produce a language model that can deal with both types of fragments.

3 Language Model Adaptation using Extra N-gram Counts

Assumption (2) in section 1 suggests that the vocabulary of the frozen patterns used in interrogative questions is already included in the newspaper articles we use. Moreover, the distribution of N-grams corresponding to the patterns appearing in the newspaper articles can be re-used if no other information about the distribution can be used. Our first method uses the fragments of N-gram counts corresponding to the frozen patterns appearing in the newspaper articles.

The flow of this process against that of the conventional process shown in Fig. 1 is illustrated in Fig.2. The N-gram counts are directly revised by using extra N-gram counts obtained from some source of information other than the text corpus. Revising the N-gram counts directly (without using the text corpus) results in the violation of consistency of the N-gram length mentioned in section 2.

3.1 Calculating N-gram probability with Extra N-gram Counts

When N-gram counts are directly revised by using extra N-gram counts, conventional N-gram counts are no longer consistent with respect to their length. Therefore, we need to redefine the use of N-gram counts. If we use an arbitrary number of extra N-gram counts with an arbitrary length, we can no longer obtain shorter N-gram counts from the longest one. We must use N-gram counts for all length n. Moreover, we need to know what the extra N-gram counts are for because we want to introduce extra N-gram counts only to raise the probability of predicting intended words, without raising all other probabilities.

We redefine the use of N-gram counts as follows:

- N-gram counts $\mathbf{C}_1(w_i)$, $\mathbf{C}_2(w_{i-1}^i)$, $\cdots$, $\mathbf{C}_N(w_{i-N+1}^i)$ are given separately according to their length $1 \cdots N$.
- Each N-gram count $\mathbf{C}_n(w_{i-n+1}^i)$ is used only to calculate the probability of predicting the last word w_i.

We will use symbol C for the conventional N-gram counts and $\mathbf{C}_n$ for the redefined N-gram counts.

The N-gram probability calculation with the redefined N-gram counts should also be redefined. The basic formula for back-off smoothing is

$$P(w_i|w_{i-n+1}^{i-1}) = \begin{cases} d_{w_{i-n+1}^i} P_{ML}(w_i|w_{i-n+1}^{i-1}) \\ \qquad \cdots C(w_{i-n+1}^i) > 0 \\ \alpha(w_{i-n+1}^{i-1}) P(w_i|w_{i-n+2}^{i-1}) \\ \qquad \cdots C(w_{i-n+1}^i) = 0 \end{cases} \tag{1}$$

where d, P_{ML}, α are, respectively, the discount coefficient, the probability calculated by using maximum likelihood estimation, and the normalized function chosen to equalize the total probabilities to one.

In order to deal with extra N-gram counts, we must revise the above calculation. Basically, for each length n, the n-gram probability should be calculated using only the n-gram counts of length n.

The probability estimated by the maximum likelihood method (P_{ML}) can be calculated from the conventional N-gram counts (assuming their consistency in terms of the length) as follows:

$$P_{ML}(w_{i-n+1}^i) = \frac{C(w_{i-n+1}^i)}{C(w_{i-n+1}^{i-1})}$$

However the following corresponding equation with extra N-gram counts is not appropriate here because of the inconsistency between $\mathbf{C}_n$ and $\mathbf{C}_{n-1}$.

$$P_{ML}(w_{i-n+1}^i) = \frac{\mathbf{C}_n(w_{i-n+1}^i)}{\mathbf{C}_{n-1}(w_{i-n+1}^{i-1})}$$

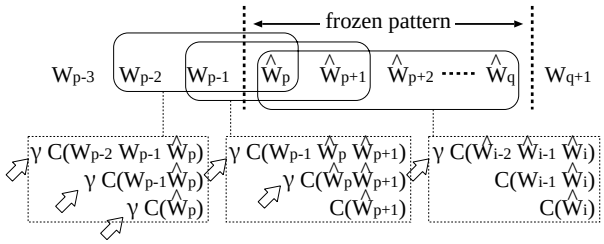

Figure 3. Providing Extra N-gram Counts (in case of tri-gram)

Instead of the above equation, we use the following equation:

$$P_{ML}(w_{i-n+1}^i) = \frac{\mathbf{C}_n(w_{i-n+1}^i)}{\sum_{w_i} \mathbf{C}_n(w_{i-n+1}^i)}$$

Several methods have been proposed for obtaining the discount coefficient d, which in turn has resulted in different back-off smoothing methods. For example, Witten-Bell smoothing [9] uses the following d_{WB} for conventional N-gram models.

$$d_{\mathrm{WB},w_{i-n+1}^i} = \frac{C(w_{i-n+1}^{i-1})}{C(w_{i-n+1}^{i-1}) + r(w_{i-n+1}^{i-1})}$$

where r is the number of different words in context w_{i-n+1}^{i-1}. In conventional N-gram calculation, the source of obtaining r is not specified, because the conventional N-gram counts are consistent in terms of their length.

In contrast with extra N-gram counts, we must obtain r from the N-gram counts of length n. We can do this as follows.

$$d_{\mathrm{WB},w_{i-n+1}^i} = \frac{\sum_{w_i} \mathbf{C}_n(w_{i-n+1}^i)}{\{\sum_{w_i} \mathbf{C}_n(w_{i-n+1}^i)\} + \mathbf{r}_n(w_{i-n+1}^{i-1})}$$

where $\mathbf{r}_n$ is the number of different words in context w_{i-n+1}^{i-1} calculated from N-gram counts $\mathbf{C}_n(w_{i-n+1}^i)$.

3.2 Emphasizing Frozen Patterns using Extra N-gram Counts

Extra N-gram counts can be used to emphasize the frozen patterns used in interrogative questions. Let $\hat{w}_p^q = \hat{w}_p \hat{w}_{p+1} \cdots \hat{w}_{q-1} \hat{w}_q$ be a frozen pattern, and let $w_{p-N+1}^{p-1} = w_{p-N+1} \cdots w_{p-1}$ be the context word sequences preceding the pattern.

The probability of sentences that include the frozen pattern should be increased and, therefore, extra N-gram counts should be given to the word sequence that is related to the pattern. For this purpose, the following counts can be increased by multiplying the original N-gram counts by $\gamma(\geq 1)$.

1. For the intermediate word sequence in the pattern, only the longest N-gram counts are increased.

$$\mathbf{C}_N(\hat{w}_{i-N+1}^{i}) = \gamma C(\hat{w}_{i-N+1}^{i}) \qquad (2)$$

For example, in a tri-gram model, only the tri-gram counts are increased.

$$\mathbf{C}_3(\hat{w}_{i-2}\hat{w}_{i-1}\hat{w}_i) = \gamma C(\hat{w}_{i-2}\hat{w}_{i-1}\hat{w}_i)$$

2. For the prefix of the pattern, the N-gram counts with a length exceeding that of the prefix are increased.

Suppose that the length of the prefix is k. Then we have:

$$\mathbf{C}_n(w_{p-n+k}^{p-1}\hat{w}_p^{p+k-1}) = \gamma C(w_{p-n+k}^{p-1}\hat{w}_p^{p+k-1}) \qquad (3)$$

for $n = k \cdots N$, for any context word sequence w_{p-n+k}^{p-1}.

In the case of a tri-gram model, we have:

$$\begin{aligned}
\mathbf{C}_3(w_{p-1}\hat{w}_p\hat{w}_{p+1}) &= \gamma C(w_{p-1}\hat{w}_p\hat{w}_{p+1}) \\
\mathbf{C}_2(\hat{w}_p\hat{w}_{p+1}) &= \gamma C(\hat{w}_p\hat{w}_{p+1}) \\
\mathbf{C}_3(w_{p-2}w_{p-1}\hat{w}_p) &= \gamma C(w_{p-2}w_{p-1}\hat{w}_p) \\
\mathbf{C}_2(w_{p-1}\hat{w}_p) &= \gamma C(w_{p-1}\hat{w}_p) \\
\mathbf{C}_1(\hat{w}_p) &= \gamma C(\hat{w}_p)
\end{aligned}$$

for any context $w_{p-2}w_{p-1}$.

3. Counts other than those mentioned above are equal to the original counts.

$$\mathbf{C}_n(w_{i-n+1}^{i}) = C(w_{i-n+1}^{i}) \qquad (4)$$

for $n = 1 \cdots N$.

This redistribution of the counts is illustrated in Fig.3.

This method can be easily extended to multiple patterns to give extra counts without duplicating the word sequences shared by two or more patterns.

The effectiveness of this adaptation method was investigated by our preliminary experiments[2]. The result showed that our method outperformed the conventional method of language model adaptation.

4 Experimental Results

We extracted N-gram counts, whose vocabulary size is 60,000 words, in newspaper articles collected over 111 months. As task-specific training data, we developed a grammar for the Japanese frozen patterns used in question answering. From the grammar, we extracted a list of all frozen patterns that were accepted by the grammar. We obtained 172 patterns.

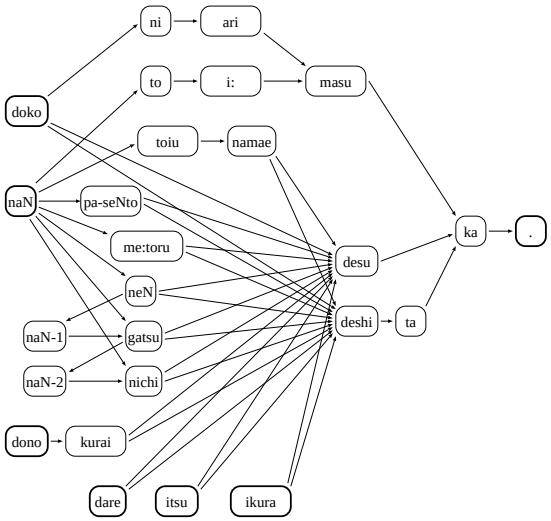

Figure 4. Word network used for the experiment

We made N-gram models using the method proposed in this paper. First, we extracted N-gram from the newspaper articles. Then, we multiplied N-gram counts (by γ) corresponding to the frozen patterns obtained from the grammar as described in section 3. We varied the multiplier γ from 0 (corresponding to the N-gram model obtained only from the newspaper articles) to 300, and obtained 31 N-gram models. We used Witten-Bell discounting method[9] for the models.

The 60 questions that used for the NTCIR-3 QAC dryrun (referred to as *DR*) were dictated by five speakers (two males and three females) and used for our evaluation. Although the grammar had been produced irrelevantly to the queries of the dry run evaluation, 43 of the 60 queries (72%) contained the patterns modeled by the grammar (referred to as *DR'*). An existing N-gram decoder [7] was used for the recognition experiments.

The results are shown in Fig.5 and Fig.6. In regard to *DR* (Fig.5), our method achieved about 10% reduction of word error rate (WER) compared with the model of no-adaptation ($\gamma = 0$). Looking at the fragments of the queries, the method gave about 50% reduction with respect to the fragments corresponding to the frozen patterns (referred to as **FP** in Fig.5), while almost no increase with respect to the other fragments of the queries (referred to as **-FP**).

In regard to *DR'* (Fig.6), no word error was achieved for **FP**. This indicates that we can improve the WER with respect to the frozen patterns by means of improving the grammar by making it model more patterns used in interrogative questions.

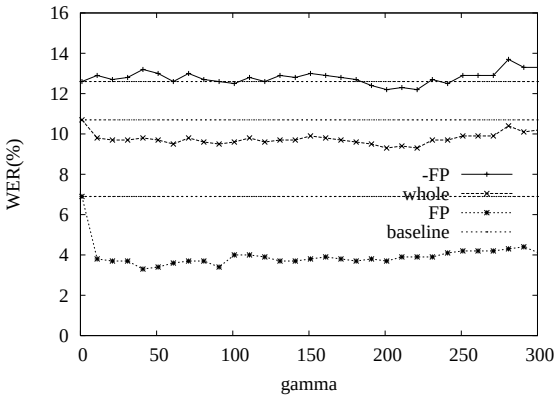

Figure 5. Word Error Rate against *DR*

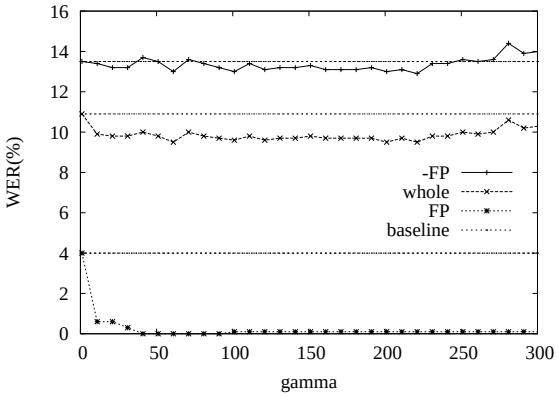

Figure 6. Word Error Rate against *DR'*

5 Conclusion

We proposed methods for language model adaptation that enable recognizing spoken questions with high accuracy. The method magnifies N-gram counts corresponding to the frozen patterns in the original N-gram. Our experiments using the NTCIR-3 QAC test collection 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. Future work would include integrating our speech recognition method and a question answering engine.

References

[1] NTCIR Workshop3 Question Answering Challenge . `http://www.nlp.cs.ritsumei.ac.jp/qac`, 2001.

[2] T. Akiba, K. Itou, A. Fujii, and T. Ishikawa. Using extra n-gram counts for statistical language model adaptation in speech-driven question answering. In *IPSJ SIGNotes of Spoken Language Processing*, volume 42, pages 31–38, 2002.

[3] M. Federico. Bayesian estimation methods for n-gram language model adaptation. In *Proceedings of International Conference on Spoken Language Processing*, pages 240–243, 1996.

[4] 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.

[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, 1998.

[6] A. Ito and M. Kohda. Evaluation of task adaptation using n-gram count mixture. *Journal of IEICE*, J83-D-II(11):2418–2427, 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.

[8] Y. Okato, J. Ishii, and T. Hanazawa. Spontaneous speech recognition using statistical language model with example sentences for spoken dialog system. In *Proceedings of the Annual Meetings of the Acoustical Society of Japan*, pages 73–74, Oct 2001.

[9] 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, 1993.

[10] E. Voorhees and D. Tice. The TREC-8 question answering track evaluation. In *Proceedings of the 8th Text Retrieval Conference*, pages 83–106, 1999.