Abstract

We integrate automatic speech recognition (ASR) and question answering (QA) to realize a speech-driven QA system, and evaluate its performance. We adapt an N-gram language model to natural language questions, so that the input of our system can be recognized with a high accuracy. We target WH-questions which consist of the topic part and fixed phrase used to ask about something. We first produce a general N-gram model intended to recognize the topic and emphasize the counts of the N-grams that correspond to the fixed phrases. Given a transcription by the ASR engine, the QA engine extracts the answer candidates from target documents. We propose a passage retrieval method robust against recognition errors in the transcription. We use the QA test collection produced in NTCIR, which is a TREC-style evaluation workshop, and show the effectiveness of our method by means of experiments.

1. Introduction

Question Answering (QA) was first evaluated extensively at TREC-8 [9]. The goal in the QA task is to extract words or phrases as the answer to a question, rather than the document lists obtained by traditional information retrieval (IR) systems. Speech interfaces have promise for improving the utility of QA systems, in which natural language questions are used as inputs. We enhanced our speech-driven IR system [5] to accept spoken questions.

In this paper, we evaluate the effects of language modeling on speech-driven question answering. In past literature, language models were evaluated independent of specific tasks. Perplexity is one of the common measures to evaluate language models, irrespective of the speech recognition accuracy. Word error rate (WER) is another common measure, which directly evaluates the accuracy of speech recognition. However, it is not clear that they can evaluate the performance of specific information processing systems using speech interfaces. Because question answering is one of the well-defined tasks and has been evaluated by formal evaluation workshops, e.g., TREC and NTCIR, we can evaluate components of a system, in particular language modeling, through a rigorous method.

Section 2 describes our language modeling method for speech-driven question answering [2]. Section 3 describes our question answering engine [3]. Section 4 describes the experimental results.

2. Language Modeling for Question Answering

Question answering systems accept a question consisting of the part that conveys a topic and the part that represents a fixed phrase for question sentences. The following is an example question:

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 question, 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, and can be recognized by an N-gram model trained with target documents (e.g., newspaper articles). The latter half of the question, i.e., “naN to yu-namae desu ka (What was the name?)”, is a fixed phrase typically used in interrogative questions, which is not very frequent in newspaper articles. Thus, we need a language model adapted to both types of expressions.

Note that recognizing the fixed phrases with high accuracy is crucial in question answering, because these phrases convey clues to determine the question and answer types. For example, a fixed phrase indicates that the answer should be the name of an object as in the previous example, while another question can potentially indicate that the answer should be the date of an event (e.g., “On what date was...”).

In this paper, we use our previous method [2], in which a language model for question answering are produced from a list of the fixed phrases typically used in interrogative questions. This method emphasizes the N-gram subset corresponding to the fixed phrases. This
method can be recast as a variant of maximum a posteriori probability (MAP) estimation, in which the N-gram subset of a background corpus is used as a posterior distribution.

2.1. Language Modeling by Emphasizing N-gram Subsets

Let $S$ be a set of sentences. Let $S_{FP}$ be a subset of $S$ that consists of the sentences including the fixed phrases in a list. Let $P$ be a language model of generating sentences (i.e., $s \in S$) 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 provides higher probability scores for sentence $\hat{s} \in S_{FP}$ but maintains the order relations on sentences $s \in S - S_{FP}$ as much as possible.

The adapted model $P'$ is produced by the following two steps.

1. Revise the maximum likelihood estimates of $P$:
   \[
P_{ML(1)}(w_i), P_{ML(2)}(w_i|w_i-1), \ldots, P_{ML(N)}(w_i|w_i-1, N+1)
   \]
   which are calculated for each value of $n(1 \leq n \leq N)$.

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

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

   (1) If the postfix $w_{i-k+1} \ldots w_{i}(1 \leq k < n)$ of the word sequence $w_i \ldots w_1$ is equal to the prefix $\hat{w}_p \ldots \hat{w}_{p+k-1}$ of one of the fixed phrases $\hat{w}_p \ldots \hat{w}_q$ then emphasize the $P_{ML}$ as follows:
   \[
P'_{ML(n)}(\hat{w}_{p+k-1}|w_{p-n+k}^p) = \beta_n(w_{p-n+k}^p) \cdot \gamma P_{ML(n)}(\hat{w}_{p+k-1}|w_{p-n+k}^p)
   \]
   otherwise, go to step (2).

   (2) If the word sequence $w_{i-n+1} \ldots w_{i}$ is equal to the subsequence $\hat{w}_p \ldots \hat{w}_q$ of one of the fixed phrases $\hat{w}_p \ldots \hat{w}_q$ then emphasize only the longest N-gram probability $P_{ML(N)}$ as follows:
   \[
P'_{ML(N)}(\hat{w}_{i-N+1}) = \beta_N(\hat{w}_{i-N+1}) \cdot \gamma P_{ML(N)}(\hat{w}_{i-N+1})
   \]
   otherwise, go to step (3).

3. Question Answering Engine

3.1. Question Answering as a Search Problem

The question answering process is often seen as the sequence of the question analysis, the relevant document (or passage) retrieval, answer extraction and answer selection processes. In this paper, we recast these processes as a search problem.

**Question Answering** Given query $q$ and document set $D$, from all the substrings in $D$, $S = \{(d, p_s, p_f)|d \in D, p_s < p_f, p_s$ and $p_f$ are positions in $d\}$, by using an evaluation function $L(q|a)$ defined on $a \in S$, the greatest appropriate answer $\hat{a}$ such that
\[
\hat{a} = \arg \max_a, L(a|q).
\]

This defines the problem of finding a single best answer, which corresponds to the factoid question in TREC and the subtask I of NTCIR Question Answering Challenge (QAC).

3.2. Passage Retrieval

The evaluation function $L$ is constructed in various aspects. One of them is the similarity between the question and the context of an answer candidate. Selecting

![Fixed Phrase](image)

Figure 1: Emphasizing trigram counts.

| (3). For all $n(1 \leq n \leq N)$, the revised probability is:
\[
P'_{ML(n)}(w_i|w_i-n+1) = \alpha_n(w_i-n+1) \cdot \beta_n(w_i-n+1) \cdot \gamma P_{ML(n)}(w_i|w_i-n+1)
\]

Here, $\gamma(\geq 1)$ is a multiplier that selects the N-gra,

\[FP\text{ is equivalent to the maximum likelihood estimate calculated as follows,}
\]

\[
P'_{ML(n)}(w_i|w_i-n+1) = \frac{C_n(w_i-n+1)}{\sum w_i \cdot C_n(w_i-n+1)}
\]

where the N-gram counts $C_n$ of each value of $n(1 \leq n \leq N)$ are obtained by emphasizing the selected subset of the original N-gram counts $C$, as shown in Fig. 1.
the context, or passage retrieval, is one of the common research topics for question answering [8].

Because by definition speech-driven question answering accepts a result of speech recognition as an input, which often includes errors, the passage retrieval must be robust against those errors. We propose a dynamic passage retrieval method that can accept an input including misrecognized words.

Suppose, from given query \( q \), we select the context of an answer candidate \( a \), which belongs to sentence \( s_i \) of document \( d = s_1 s_2 \cdots s_i \cdots s_n \). Let \( s'_i = s_i - \{a\} \), \( h \) be the headline of \( d \), and \( t \) be the string “Kotoshi Kongetsu Kyou” (this year, this month, today). Given a number \( k > 0 \), let \( S_i = \{h, t, s_{i-k}, \cdots, s_{i-1}, s'_i, s_{i+1}, \cdots, s_{i+k}\} \).

The optimal context \( \hat{C}_i \) is selected from \( C_i \in 2^S_i \) by maximizing the following evaluation measure \( F(C_i) \).

\[
F(C_i) = \frac{1 + \beta^2}{R(C_i) + \frac{1}{P(C_i)}}
\]

\[
R(C_i) = \frac{\text{Score}(q \land C_i)}{\text{Score}(q)}
\]

\[
P(C_i) = \frac{\text{Score}(q \land C_i)}{\text{Score}(C_i)}
\]

Here, \( \text{Score}(A) \) is a sum of the IDFs (inverse document frequencies) of the elements in \( A \) and \( \text{Score}(A \land B) \) is a sum of the IDFs of the elements appeared commonly in \( A \) and \( B \).

We used \( k = 1 \) for our experiments. The measure \( F \) corresponds to the (weighted) F-measure often used in IR research. The recall is more influential than precision in calculating the F-measure, if the value of \( \beta \) is more than one. Because the recall is important for selecting answer candidates, we set \( \beta = 2 \).

4. Evaluation

4.1. NTCIR Question Answering Challenge

The test collection constructed in the first evaluation of Question Answering Challenge (QAC-1) [6], which was carried out as a task of NTCIR Workshop 3, was used as the test data for our evaluation. The task definition of QAC-1 is as follows.

Target documents are two years of Japanese newspaper articles, from which the answers of a given question must be extracted. The answer is a noun or a noun phrase, e.g., person names, organization names, names of various artifacts, money, size and date. Three subtasks were performed in QAC1, among which the subtask 1 is defined as follows.

System extracts at most five answers from the documents for each question. The reciprocal number of the rank is the score for the question. For example, if the second answer candidate is correct, the score is 0.5.

This definition is almost equivalent to the factoid question answering in TREC. The 200 queries were used for the formal evaluation, in which no answer was found for four questions in the target documents. Mean Reciprocal Rank (MRR) of the 196 queries was used to evaluate the performance of participant systems.

4.2. Experimental Results

The effects of language modeling on question answering were experimentally investigated. We extracted N-gram counts from newspaper articles in 111 months. The vocabulary size was 60,000. We produced a word network for the Japanese fixed phrases used for question...
investigates whether the difference in performance is meaningful or simply due to chance. We found that the MRR values for BASE and EMP were significantly different (at the 5% level).

5. Conclusion

In this paper, we proposed a speech-driven question answering (QA) system and evaluated its performance, focusing mainly on the effects of language modeling. For evaluation purposes, we used the test questions in the NTCIR collection, read by eight human subjects. The experimental results showed that our language modeling method improved the accuracy of recognizing spoken questions and consequently the accuracy of question answering. At the same time, when compared with text-based QA, the performance of speech-driven QA system was not satisfactory from a practical point of view. Future work includes improving each module through a glass-box error analysis and extending our system to spontaneously spoken questions [1].

6. References