# Exploiting Dynamic Passage Retrieval for Spoken Question Recognition and Context Processing towards Speech-driven Information Access Dialogue

**Tomoyosi Akiba**

Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi,
Aichi, 441-8580, JAPAN
akiba@cl.ics.tut.ac.jp

### Abstract

Speech interfaces and dialogue processing abilities have promise for improving the utility of open-domain question answering (QA). We propose a novel method of resolving disambiguation problems arisen in those speech and dialogue enhanced QA tasks. The proposed method exploits passage retrieval, which is one of main components common in many QA systems. The basic idea of the method is that the similarity with some passage in the target documents can be used to select the appropriate question from the candidates. In this paper, we applied the method to solve two subtasks of QA, which are (1) N-best rescoring of LVCSR outputs, which selects a most appropriate candidate as a question sentence, in speech-driven QA (SDQA) task and (2) context processing, which compose a complete question sentence from a submitted incomplete one by using the elements appeared in the dialogue context, in information access dialogue (IAD) task. For both tasks, a dynamic passage retrieval is introduced to further improve the performance. The experimental results showed that the proposed method is quite effective in order to improve the performance of QA in both two tasks.

## 1. Introduction

Open-domain Question Answering (QA) was first evaluated extensively at TREC-8 held in 1999. From 2001, QA in Japanese have been evaluated in NTCIR Question Answering Challenge (QAC). The goal in the QA task is to extract words or phrases as the answer to a question from an unorganized document collection, rather than the document lists obtained by traditional information retrieval (IR) systems.

Speech interfaces using large vocabulary continuous speech recognition (LVCSR) decoders have promise for improving the utility of QA systems, in which natural language questions are used as inputs. We refer to the QA enhanced by the speech interface as Speech-driven Question Answering (SDQA). One of the most common problems faced with when we enhance the text-based QA system to accept spoken questions, arises from the recognition errors found in the transcription obtained by using LVCSR. The information loss caused by it gives a serious degradation to the total performance of question answering. Because LVCSR decoders can often outputs N-best list of transcriptions as the recognition candidates, this problem can be seen as resolving the ambiguity caused on the speech recognition results. It can be resolved by selecting the most appropriate question from the N-best list.

On the other hand, Information Access Dialogue (IAD) task have been evaluated in the recent NTCIR QAC series. IAD task assumes the situation in which users interactively collect information using a QA system. The QA Systems aiming at the task need the abilities of context processing. In IAD task, the systems must accept a contextual question, which has reference expressions and ellipses that refer to the entities appeared in the previous questions and answers. This incomplete question has to be completed by selecting appropriate entities from the context. Therefore, this also can be seen as a problem of ambiguity resolution, and can be resolved by selecting the entities in order to compose the most appropriate question from the history.

In this work, we propose the method of resolving those disambiguation problems by exploiting passage retrieval. The basic idea of the proposed method is as follows. Suppose an input question has at least one correct answer in the target document collection, there must be at least one similar passage in it. Therefore, the similarity with some passage in the target documents can be used to select the appropriate question from the candidates in both SDQA and IAD task situation.

The rest of the paper is organized as follows. Section 2. describes our passage retrieval method that selects the size of the passage dynamically according to the similarity with the query. It will be applied to the two disambiguation problems arisen in the QA tasks in the following two sections. Section 3. describes the N-best rescoring of candidate spoken questions hypothesized by a LVCSR system. Section 4. describes the context processing for the information access dialogue task. In Section 5., we will give the conclusion.

## 2. Dynamic Passage Retrieval

A passage, i.e. a text fragment in target documents, is used to calculate the similarity against the question. Some systems use a sentence as a passage, while other systems use a paragraph. The longer the size of a passage is selected, the more candidates of the answer can be picked up. It raises the recall of the answer, while it reduces the precision because the more incorrect candidates are also picked up. Developing a good passage retrieval method is one of the common research topics for question answering (Tellex et al., 2003).

We have proposed a dynamic passage retrieval method (Akiba et al., 2004b; Murata et al., 2005). The method selects an appropriate size of the passage on the fly by using F-measure based similarity with the question.

Let $C(s)$ be a set of passage candidates with respect to a sentence $s$ in the target documents. [1] Here we assume that the target documents are newspaper articles. Though $C(s)$ can include any size of text fragments surrounding $s$ theoretically, only the following sentences are considered in our implementation whether each of them should be included in the passage.

$s_{-1}$: the sentence immediately before $s$.

$s_{+1}$: the sentence immediately after $s$.

$h_A$: the headline of the article $A$ that $s$ belongs.

$d_A$: the date string of the article $A$ that $s$ belongs.

Therefore, we adopted the following candidate $C(s)$ in practice.

$$C(s) = \{\{s\} \cup E | E \in 2^{\{s_{-1}\ s_{+1}\ h_A\ d_A\}}\}$$

The proposed method selects a best passage $\hat{c}$ from $C(s)$ by using following F-measure based similarity $F(q, c)$ with a question $q$,

$$\hat{c} = \underset{c \in C(s)}{\operatorname{argmax}} F(q, c) \tag{1}$$

$$F(q, c) = \frac{(1 + \beta^2) P(q, c) R(q, c)}{\beta^2 P(q, c) + R(q, c)} \tag{2}$$

$$\begin{aligned} P(q, c) &= \frac{\sum_{t \in T(q) \cap T(c)} \mathrm{idf}(t)}{\sum_{t \in T(c)} \mathrm{idf}(t)}, \\ R(q, c) &= \frac{\sum_{t \in T(q) \cap T(c)} \mathrm{idf}(t)}{\sum_{t \in T(q)} \mathrm{idf}(t)} \end{aligned}$$

where $T(c)$ is a set of terms included in $c$ and $idf(t)$ is the inverse document frequency (IDF) of a term $t$.

We chose $\beta = 2$ to emphasize the recall for the N-best rescoring of spoken questions (Section 3.) and for the question answering itself, while $\beta = 1$ for the context processing (Section 4.1.).

The passage retrieval score $S_{\text{passage}}(q)$ is defined as the max value of $F(q, c)$ with respect to the target document collection $D$.

$$S_{\text{passage}}(q) = \max_{s \in D} \max_{c \in C(s)} F(q, c) \tag{3}$$

We cannot examine all of sentences in $D$ because of the computational cost. Therefore, only the sentences included in the documents that a document retrieval engine returns by submitting $q$ are examined to calculate the equation (3) for an approximation.

## 3. N-best Rescoring of Recognition Candidates

One of the most common problems faced with when we enhance the text-based QA system to accept spoken questions, arises from the uncertainty of speech recognition results. The information loss caused by it gives a serious degradation to the total performance of question answering. The task specific language modeling can improve the accuracy of speech recognition and, consequently, the total performance of question answering (Akiba et al., 2004a). However, as the n-gram language model can only model the short-term constraint of word sequence, it fails to capture the semantic consistency of sentence level.

Suppose an input question has at least one correct answer in the target document collection, there must be at least one similar passage in it. Thus the similarity with some passage in the target documents can be used to reduce the uncertainty in speech recognition process. For example, suppose two candidate sentences *"What was the name of the spacecraft landed safely on March in 1976?"* and *"What was the name of the spacecraft landed safely on Mars in 1976?"* are obtained by the speech recognition process and there found a passage "The first U.S. spacecraft to land on Mars was a spacecraft called Viking 1 which occurred on July 20, 1976." in the target documents. Because the latter candidate has more common words and therefore is more similar to the passage, it is more likely to be the correct question sentence.

The similarity to a passage appeared in an actual document expresses that the candidate word sequence is more or less semantically consistent as a whole. As a N-best list of candidates obtained by an existing LVCSR decoder often includes a lot of meaningless sentences in practice, the similarity, or the *passage retrieval score* in other words, can be used to filter out such sentences. From language modeling perspective, this process can be seen to capture the semantic consistency of the candidate in sentence level, which conventional n-gram language model fails to capture.

In (Akiba and Abe, 2005), the passage retrieval method with fixed size passage (one sentence, three sentences, or a document) was applied for N-best rescoring. In this paper, the dynamic passage retrieval described in 2. is applied.

### 3.1. Combining Speech Recognition Likelihood and Passage Retrieval Score

Using the passage retrieval score solely for rescoring does not take the plausibility of the candidate itself, measured by the speech recognition process, into consideration. Simply, the likelihood of speech recognition $P(q_i|a)$, where $q_i$ is the $i$-th best recognized sentence and $a$ is the observed acoustic signal, can be used as representative of the plausibility. Its log likelihood $\log P(q_i|a) \propto \log P(a|q_i) + \beta \log P(q_i) + \gamma |q_i|$, where $P(a|q_i)$, $P(q_i)$, $\beta$, and $\gamma$ are the acoustic model, the language model, the language model weight, and the insertion penalty respectively, is also known as recognition score and is used to guide the search in the recognition process. It is automatically obtained with each recognized sentence as the result of speech recognition.

The final rescoring measure $S_{\text{rescore}}$ is obtained by interpolating the likelihood $P(q_i|a)$ and the passage retrieval score $S_{\text{passage}}(q_i)$,

$$S_{\text{rescore}}(q_i) = P(q_i|a)^{\alpha} \cdot S_{\text{passage}}(q_i, s_{q_i}) \tag{4}$$

where $\alpha$ is the interpolation weight.

[1] More specifically, the candidates should be considered with respect to an answer candidate $a$. However we approximate $a$ to be identical with $s$ that includes $a$.

### 3.2. Test Data

The test collection constructed in the first evaluation of Question Answering Challenge (QAC-1) (Fukumoto et al., 2003), which was carried out as a task of NTCIR Workshop 3, was used to produce the test data of spoken questions for our evaluation. The task definition of QAC-1 (subtask 1 [2]) 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. 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, where MRR was used as an evaluation metric in TREC-8, 9, and 10, and the exact answer extraction is evaluated since TREC-11. The 200 questions were used for the formal evaluation, in which no answer was found for four questions in the target documents that consisted of newspaper articles in 2 years.

The 200 questions were read by four females and four males in order to produce the test speech data for our evaluation.

An existing LVCSR system (Lee et al., 2001) was used for the purpose of transcription. The language model is constructed from the 12 years newspaper articles and the vocabulary size is about 60,000 words. The resulting N-best candidates $q_1 q_2 \cdots q_N$ are rescored by $S_{\text{rescore}}(q_i)$ of the equation (4), then the top ranked sentence was selected to investigate its performance by using the evaluation metric described below.

### 3.3. Evaluation Metrics

We used three evaluation metrics for our experiments. First of all, the word error rate (WER) of the resulting sentence was investigated in order to see how our method works as a language model for speech recognition. The average WER for all 200 questions was used as the first evaluation metric The top ranked sentence $q$ after rescoring was submitted to our question answering system (Akiba et al., 2004b). The system outputs five answers $a_1..a_5$ for an input question $q$. The answers are ordered by the system from 1st to 5th positions according to their confidence about the correctness. Each answer is scored on the inverse number of its order, called Reciprocal Rank (RR). The score of the question $q$, $RR(q)$, is the highest score of its five answers.

$$rr(a_i) = \begin{cases} 1/i & \text{if } a_i \text{ is a correct answer} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$RR(q) = \max_{a_i} rr(a_i) \tag{6}$$

The mean RR (MRR) for all 196 questions that have at least one correct answer was used as the evaluation metric for question answering. Additionally, the rate of the questions in 196 that the system correctly answered within five outputs per question (% correct) was also used as the evaluation metric.

| method | WER (%) | MRR | %correct (%) |
|---|---|---|---|
| *BASELINE* | 24.8 | 0.240 | 28.2 |
| *static passage* | 23.6 | 0.284 | 36.8 |
| *dynamic passage* | 23.8 | 0.291 | 38.3 |
| *ORACLE* | 20.3 | 0.279 | 35.0 |
| *TEXT input* | 0 | 0.516 | 66.5 |

Table 1: Experimental Result of Word error rate (WER), mean reciprocal rank (MRR) and the rate of the questions correctly answered (%correct), averaged over eight speakers.

### 3.4. Results

The experimental evaluation was taken place by comparing the results obtained by rescoring methods. The baseline method (referred as *BASELINE*) simply selects a most likely candidate (with largest likelihood score) from the results of speech recognition.

The proposed methods select a candidate from 10-best list by exploiting the passage retrieval score and by rescoring them. Two passage retrieval methods are investigated. The first method selects the fixed size of passage defined in advance (referred as *static passage*). The previous experiment (Akiba and Abe, 2005) revealed that the passage size of 3 sentences window was best performed among the several sizes. The similarity measure between a question and a passage used in the method is *TF-IDF* with pivoted document length normalization (Singhal et al., 1996). The second method selects the size of passage on the fly by using the dynamic passage retrieval described in Section 2. (referred as *dynamic passage*). In both methods, the weight of the interpolation $\alpha$ in the equations (4) was estimated by using 10-fold cross validation.

As a reference, the oracle method selects the best result, which has the smallest word errors, from 10-best recognition candidates (referred as *ORACLE*).

Table 1 shows the results of WER, MRR and %correct averaged over the eight speakers.

As for speech recognition performance (WER), both the proposed methods (*static passage* and *dynamic passage*) improved the baseline about 4.0-4.8 % relative. We used the paired t-test for statistical testing, which investigates whether the improvements in performance is meaningful or simply due to chance. We found that the WER values for *BASELINE* and both the proposed methods were significantly different (at the 0.005% level). This improvement might be further increased in some way, because the ideal method (*ORACLE*) achieved the better results (the relative improvement was about 18.1 %).

The notable results were obtained for question answering. The proposed methods showed a considerable improvement in the performance compared with the baseline. The paired t-test revealed that the MRR values for *BASELINE* and the other methods were significantly different (at 0.1% level) while those between the proposed methods and *OR-*

[2] Three subtasks were performed in QAC1. See (Fukumoto et al., 2003).

*ACLE* were not.

Comparing among the passage retrieval methods, the dynamic passage retrieval did not improve the WER obtained by the static passage retrieval. However, the dynamic passage retrieval did improve the QA performance further, while the difference was not statistically significant (the p-value is 0.093).

The reason why the improvement in question answering was more remarkable than in speech recognition seems to be explained as follows: The semantic consistency within a question is crucial for question answering, where the question analysis plays an important role for collecting the information about the correct answer and it requires more precise information about words and their ordering than bag of words, while the WER metric and the document retrieval require less precise information about a question, e.g. appearance of individual words or bag of words, no matter how they relate each other.

## 4. Context Processing for IAD

Information Access Dialogue (IAD) task have been evaluated in the recent NTCIR QAC series, specifically QAC2 subtask3 (Kato et al., 2004) and QAC3 (Kato et al., 2005). IAD task assumes the situation in which users interactively collect information using a QA system. The QA systems aiming at the task need the abilities of context processing. Suppose the following series of questions

***Q1*** *"Whose monument was displayed at Yankees Stadium in 1999?"*

***Q2*** *"When did he come to Japan on honeymoon?"*

***Q3*** *"Who was the bride at that time?"*

The second question *Q2* can be answered by selecting the fragments *"Joe DiMaggio"* that is the answer to the first question and composing the complete question *"When did Joe DiMaggio come to Japan on honeymoon?"* Similarly, the third question *Q3* can be answered by selecting appropriate fragments from the previous questions and their answers (*"Joe DiMaggio"* and *"come to Japan on honeymoon"*) and composing the complete question. If the fragments is selected incorrectly, e.g. *"Yankees Stadium"* and *"1954"*(the answer of the second question), the resulting complete question is useless, rather harmful, to find the correct answer. Therefore, this can be seen as a problem of ambiguity resolution, and can be resolved by selecting the fragments from the history in order to compose the most appropriate question.

The proposed method of exploiting passage retrieval can also be applied to this disambiguation problem related to the context processing for IAD task. The similarity with some passage in the target documents can be used to select the appropriate context from the history of the questions.

### 4.1. Formulation of Context Processing for IAD task

The third question *Q3* of the last example of IAD task can be combined with any set of the text fragments extracted from the history of the series of questions and their answers, and formed a candidate of the appropriate question, e.g. *"Joe DiMaggio, come to Japan on honeymoon, Who was the bride at that time?"* is one of the candidates, while *"Yankees Stadium, 1954, Who was the bride at that time?"* is another. Suppose the passage *"Joe DiMaggio and Marilyn Monroe went to Japan for their honeymoon."* is found, the first candidate is more likely to be the appropriate question because of the higher similarity between the candidate and the passage.

This context processing problem is formulated as follows. Let $H(q_i)$ be a history of a question $q_i$, which is a set of text fragments appeared in either a previous questions $q_1 \cdots q_{i-1}$ or their answers $a_1 \cdots a_{i-1}$. Any unit can be used for the text fragment that corresponds to an element of $H(q)$: it can be a word $w \in q_1 \cup \cdots \cup q_{i-1} \cup a_1 \cup \cdots \cup a_{i-1}$, or a sentence $s \in \{q_1 \cdots q_{i-1}\ a_1 \cdots a_{i-1}\}$. In the following, we use a sentence $s$ as the unit.

Giving a question $q$ and its history $H(q)$, A candidate of the complete question of $q$ is composed by adding a set of text fragments in the history $h \in 2^{H(q)}$ to $q$, i.e. $h \cup q$. Now, the problem of context processing is defined as selecting the best context $\hat{h} \in 2^{H(q)}$ that compose the best complete question $\hat{h} \cup q$. The proposed method try to solve this problem by maximizing the passage retrieval score $S_{\text{passage}}(h \cup q)$ as follows.

$$\hat{h} = \operatorname*{argmax}_{h \in 2^{H(q)}} S_{\text{passage}}(h \cup q) \qquad (7)$$

The computational cost of calculating the equation (7) exactly gets higher with the size of $H(q)$, because all of the combinations of the elements in $H(q)$ must be compared. Therefore, we introduced the approximation to (7): we restricted the context to $\tilde{H}_{QA}(q_i) = \{q_1\ q_{i-1}\ a_1\ a_{i-1}\}$. We also exclude the case with no context. Those result in the following equation (referred as *HQA* in our experiment).

$$\hat{h} \approx \operatorname*{argmax}_{h \in 2^{\tilde{H}_{QA}(q)} - \{\phi\}} S_{\text{passage}}(h \cup q) \qquad (8)$$

Including the answers $a_1 \cdots a_{i-1}$, which are returned by the system, in $H(q)$ seems harmful, because they may be incorrect. Usually in many QA systems including ours, the string exactly appears in the question is not considered as an answer candidate. Therefore, if the system outputs an incorrect answer that is accidentally same with a future question in the same series, it will not be possible to return the correct answer to the future question. For this reason, we restricted the context to $\tilde{H}_{Q}(q_i) = \{q_1 q_{i-1}\}$ and introduced the following equation for context selection, (referred as *HQ*)

$$\begin{aligned} \hat{h} &\approx \operatorname*{argmax}_{h \in 2^{\tilde{H}_{Q}(q)} - \{\phi\}} S_{\text{passage}}(h \cup q) && (9) \\ &= \operatorname*{argmax}_{h \in \{\{q_1\}\{q_{i-1}\}\{q_1 q_{i-1}\}\}} S_{\text{passage}}(h \cup q) && (10) \end{aligned}$$

As baseline, the method using the fixed context $\tilde{h} = \{q_1\ a_1\ q_{i-1}\ a_{i-1}\}$ (referred as *baseQA*) and $\tilde{h} = \{q_1\ q_{i-1}\}$ (referred as *baseQ*) were investigated.

As reference, we also investigated the maximum performances of *baseQA* and *HQA* when the correct answers were always obtained in the previous series of questions (referred as *baseQA with CA* and *HQA with CA*, respectively).

*What genre does the "Harry Potter" series belong to?*

*Who is the author?*

*Who are the main characters in that series?*

*When was the first volume published?*

*What title does it have?*

*How many volumes were published by 2001?*

*How many languages has it been translated into?*

*How many copies have been sold in Japan?*

Figure 1: An example of the gathering type of series.

*Where was Universal Studio Japan constructed?*

*Which train station is the nearest?*

*Who is the actor who attended the ribbon-cutting ceremony on the opening day?*

*What is the movie he was featured in that was released in the New Year season of 2001?*

*What is the movie starring Kevin Costner released in the same season?*

*What was the subject matter of that movie?*

*What role did Costner play in that movie?*

Figure 2: An example of the browsing type of series.

### 4.2. Test Data

The experiment was performed by using QAC3 test collection (Kato et al., 2005). The QAC3 test collection contains 50 series and 360 questions. The number of questions in one series ranges from 5 to 10, and the average is 7.2. The target document set, where answers are intended to be extracted from a question, consists of two years of articles from two newspapers.

The test collection consists of two types of series of questions: a gathering type and a browsing type. In the gathering type, the user has a concrete objective such as writing a report and summary on a specific topic, and asks a system a series of questions related to that topic. In the browsing type, the user does not have any fixed topic of interest, which therefore varies as the dialogue progresses. Therefore, the context processing for IAD task is more critical for the browsing type than for the gathering type. The test collection contains 35 series of the gathering type and 15 series of the browsing type. Figure 1 and 2 show examples of series of those two types from (Kato et al., 2005).

### 4.3. Experimental Results

The performances of the four methods, i.e. *baseQA*, *baseQ*, *HQA*, and *HQ*, were compared by the evaluation measure MMF1 (modified F measure averaged over the questions) (Kato et al., 2005). The results were shown in Table 2.

The difference of the performance according to the types of the series was investigated. The label **All**, **Gather** and **Browse** correspond to all the series, the series of the gathering type, and the series of the browsing type, respectively. The result showed that the proposed method did not improve the baseline method with respect to entire test set (**All**): the performance of *HQA* was less than *baseline*, while the performance of *HQ* is almost equal to *baseline*.[3] However, the performances of them were quite different according to the type of series. *HQ* outperformed *baseline* with respect to the browsing type of series (**Browse**). This result indicated that the method was effective for the browsing type, in which the context processing plays much more critical role than in the gathering type.

| Method | All | Gather | Browse |
|---|---|---|---|
| #series | 50 | 35 | 15 |
| #questions | 360 | 253 | 107 |
| *baseQA* | 0.146 | 0.171 | 0.084 |
| *baseQ* | 0.193 | 0.222 | 0.125 |
| *HQA* | 0.169 | 0.188 | 0.125 |
| *HQ* | 0.194 | 0.216 | **0.143** |
| *baseQA with CA* | 0.146 | 0.157 | 0.120 |
| *HQA with CA* | 0.180 | 0.183 | **0.174** |

Table 2: QA performance differences according to the context processing methods (MMF1).

The most interesting results were those with correct answers (*baseQA with CA* and *HQA with CA*). With respect to *baseQA with CA*, using the correct answers still did not improve the performance. This indicates that the unnecessary terms for the question, whether or not they are correct, degrades the QA performance. The result of *HQA with CA* showed almost same tendency with *HQA*; though it did not improve the performance in total, it did improved the performance for browsing type. Furthermore, the improvement for browsing type of *HQA with CA* was much greater than that of *HQ*. This indicate that the proposed method selected the context appropriately and that giving the correct answers it further improved the performance.

### 4.4. Discussion

We formulated the context processing in IAD task as a problem of context selection from previous questions and answers to compose an appropriate complete question, and proposed a novel method for the problem exploiting passage retrieval. The method uses only term statistics for context processing instead of conventional NLP such as anaphora resolution. Since the current implementation of the method is naive, we think some refined implementation can improve the performance further. The combination of our method and the conventional NLP method will be also hopeful.

## 5. Conclusion

In this paper, a novel method of resolving disambiguation problems in QA by using dynamic passage retrieval was proposed. We applied the method to two subtasks of QA;

[3] Note that this is partly because the questions of the gathering type are about 2.5 times as much as that of browsing type in the QAC3 test collection.

N-best rescoring of spoken question in SDQA task and context processing in IAD task. In SDQA task, the experimental results showed that the proposed method achieved considerable improvement on both the word error rate and the QA performance. In IAD task, the experimental results showed that the method improved the QA performance when it applied to the browsing type of series of questions. Because the proposed method gives a general framework for resolving disambiguation problems arisen in open-domain QA task, it will be applied to other problems than described here, including query term expansion, etc.

## 6. Acknowledgement

This work was partly supported by Grant-in-Aid for Scientific Research (KAKENHI) (C) 17500092 from Japan Society for the Promotion of Science.

## 7. References

Tomoyosi Akiba and Hiroyuki Abe. 2005. Exploiting passage retrieval for n-best rescoring of spoken questions. In *Proceedings of International Conference on Speech Communication and Technology (Eurospeech)*, pages 65–68.

Tomoyosi Akiba, Atsushi Fujii, and Katunobu Itou. 2004a. Effects of language modeling on speech-driven question answering. In *Proceedings of International Conference on Spoken Language Processing*, pages 1053–1056.

Tomoyosi Akiba, Atsushi Fujii, and Katunobu Itou. 2004b. Question answering using "common sense" and utility maximization principle. In *Proceedings of The Fourth NTCIR Workshop*. `http://research.nii.go.jp/ntcir/workshop/OnlineProceedings4/QAC/NTCIR4-QAC-AkibaT.pdf`.

Jun'ichi Fukumoto, Tsuneaki Kato, and Fumito Masui. 2003. Question answering challenge (QAC-1) question answering evaluation at NTCIR workshop 3. In *Proceedings of The third NTCIR Workshop*.

Tsuneaki Kato, Jun'ichi Fukumoto, and Fumito Masui. 2004. Question answering challenge for information access dialogue — overview of NTCIR4 QAC2 subtask 3. In *Proceedings of The Fourth NTCIR Workshop*, pages 361–372. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/QAC/NTCIR4-QAC-KatoT.pdf`.

Tsuneaki Kato, Jun'ichi Fukumoto, and Fumito Masui. 2005. An overview of NTCIR-5 QAC3. In *Proceedings of The Fifth NTCIR Workshop*. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/QAC/NTCIR5-OV-QAC-KatoT.pdf`.

Akinobu Lee, Tatsuya Kawahara, and K. Shikano. 2001. Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1691–1694, Sept.

Yuichi Murata, Tomoyosi Akiba, Atsushi Fujii, and Katunobu Itou. 2005. Question answering experiments at NTCIR-5: Qcquisition of answer evaluation patterns and context processing using passage retrieval. In *Proceedings of The Fifth NTCIR Workshop*, pages 394–401. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/QAC/NTCIR5-QAC-MurataY.pdf`.

Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In *Proceedings of ACM SIGIR*, pages 21–29.

Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of ACM SIGIR*, pages 41–47.