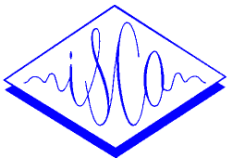

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

STUDY ON SPOKEN INTERACTIVE OPEN DOMAIN QUESTION ANSWERING

Chiori Hori, Takaaki Hori, Hideki Isozaki,
Eisaku Maeda and Shigeru Katagiri

NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
{chiori,hori,isozaki}@cslab.kecl.ntt.co.jp

Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology
furui@cs.titech.ac.jp

ABSTRACT

This paper proposes an interactive approach to spoken interactive open-domain question answering (ODQA) systems. The goal of ODQA systems is to extract an exact answer to user's question from unstructured information sources such as large text corpora. When the reliabilities for answer hypotheses obtained by an ODQA system are low, systems need more information to effectively distinguish the exact answer required by users. In our spoken interactive ODQA system, **SPIQA**, spoken questions are recognized by an automatic speech recognition (ASR) system and disambiguous queries (DQs) are automatically generated to disambiguate transcribed questions. To derive appropriate DQs, ambiguous information is detected based on recognition reliability, dependency structures between phrases in the users' questions, and features of word occurrence in the retrieved corpus. We confirmed the appropriateness of the derived DQs by comparing them with manually prepared ones. We also reconstructed the questions manually using additional information that was required by the DQs. We then tested the effect of the additional information on the performance of our ODQA system.

1. INTRODUCTION

Human and machine dialog systems using a speech interface have been intensively researched in the field of spoken language processing (SLP). Such conversational dialogs to exchange information through question answering (QA) are a natural communication modality. However, state-of-the-art dialog systems only operate for specific-domain question answering (SDQA) dialogs. To achieve more natural communication between human beings and machines, spoken dialog systems for open domains are necessary. Specifically open-domain question answering (ODQA) is an important function in natural communication. Our goal is to construct a spoken interactive ODQA system, which includes an ASR system and an ODQA system. To clarify the problems presented in building such a system, the QA systems that have been constructed so far have been classified into a number of groups depending on their target domains, interfaces, and interactions to draw out additional information from users to accomplish set tasks shown in Table 1. In this table, text and speech denote text input and speech input. The term "*addition*" represents additional information queried by QA systems. This additional information is information other than that derived from the user's initial questions.

The ODQA systems that have been researched in the field of natural language processing (NLP) [1] return an actual answer rather than a ranked list of documents in response to the question written in a natural language. However, SDQA has been studied in the area of Artificial Intelligence (AI). The differences between SDQA and ODQA systems are in their data structure and the design of dialog scenarios. Since information in a specific domain can be arranged in a table, the SDQA systems can accomplish QA by table lookup techniques [2]. In interactive SDQA systems that require sufficient information to yield the desired answer through queries, all solutions to extract the answer have been designed through dialog scenarios using a knowledge database and IF-THEN rules [3]. However, various spoken QA systems incorporating interactions through speech have been investigated to achieve more natural communication between humans and machines [4] [5]. Currently, a spoken ODQA system including an ODQA system [6] using a speech interface instead of text input are being constructed [7].

Table 1. Dialog domain and data structure for QA systems

target domain		specific	open
data structure		knowledge DB	unstructured text
text	without *addition*	CHAT-80 [2]	FALCON [6]
	with *addition*	MYCIN [3]	(**SPIQA***)
speech	without *addition*	Harpy [4]	VAQA [7]
	with *addition*	JUPITER [5]	(**SPIQA***)

∗ **SPIQA** is our system.

To construct more precise and user-friendly ODQA systems, this paper proposes an interactive approach to spoken ODQA systems. Three main issues that need to be addressed to construct spoken interactive ODQA systems are:

1. The ODQA problems:
   Answers are not in a table and are scattered throughout unstructured text.
2. The interactive ODQA problems:
   Since user's questions are not restricted, system queries for additional information to extract answers and effective interaction strategies using such queries cannot be prepared before the user inputs the question.
3. The spoken QA problems:
   Recognition errors degrade the performance of QA systems. Some indispensable information to extract answers is deleted or substituted by other words.

This paper proposes an interactive approach that is based on the disambiguation of users' questions in interactive ODQA systems.

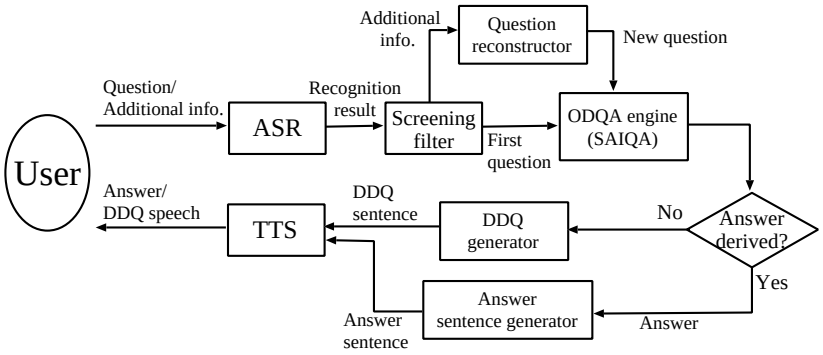

Fig. 1. Components and data flow in SPIQA.

In addition, we introduce our spoken interactive ODQA system, i.e., **SPIQA**. Since questions input through text are more formal than transcribed speech, the interactive approach applied to SPIQA can also deal with text interaction.

2. SPOKEN INTERACTIVE QA SYSTEM: SPIQA

Figure 1 shows the components of our system, and the data that flows through it. This system is comprised of an ASR system [8], a screening filter that uses a summarization method [9], an ODQA engine (**SAIQA**) [10] for a Japanese newspaper text corpus, and a Deriving Disambiguous Queries (**DDQ**) module.

ASR system

Our ASR system is based on the WFST approach , which offers a unified framework representing various knowledge sources and it produces an optimized search network of HMM states [8]. We combined cross-word triphones and trigrams into a single WFST and applied a one-pass search algorithm to it. A confidence measure for each word is calculated by post-processing.

Screening filter

The question transcribed by an ASR system sometimes incorporates not only redundant information caused by the spontaneity of human speech but also irrelevant information due to recognition errors. Recognition errors, fillers, word fragments, and other distractors are removed from the transcribed question by a screening filter that extracts meaningful information. the summarization method [9] is applied to the screening process.

Since recognition errors in the recognition results directly degrade QA performance, the screening filter should remove them. In this study, the screening process was done in two steps. The first step was to remove acoustically and linguistically unreliable words based on the threshold for the confidence measure. The second step was to construct a meaningful sentence from the results after removing unreliable words using the speech summarization technique through word extraction [9]. Hence, the screened results excluded large recognition errors and made the sentence understandable. Finally, the screened results were input into the ODQA engine.

ODQA engine

The ODQA engine has four components: question analysis, text retrieval, answer hypothesis extraction, and answer selection. Nouns/noun-phrases are classified into category classes such as ORGANIZATION or PERSON. A given question sentence is analyzed to determine the type of expected answer and keywords using the question analysis module. Paragraphs/documents that match the keywords are then extracted by the text retrieval module. The nouns/noun-phrases in the retrieved relevant documents that belong to the expected category class are extracted and used to output answers.

DDQ module

When the ODQA engine cannot extract an appropriate answer to a user's question, the question is considered "ambiguous." There are two situations where the question is considered ambiguous. The first is when the user does not supply sufficient information in his/her question. The other is when some necessary information to extract the answer is lost through ASR. Since all information in a users' question is not always useful to extract answers, only indispensable information to do this should be compensated by additional information that is inputs by users. The DDQ module derives disambiguous queries (DQs) that require such indispensable information.

The DQs are generated using templates of interrogative sentences, each of which contains an interrogative and a phrase taken from the user's question after speech recognition and screening. The DDQ module selects the best DQ based on its linguistic appropriateness and the ambiguity of the phrase. Hence, the module can generate a sentence that is linguistically appropriate and asks the user to disambiguate the most ambiguous phrase in his/her question.

Suppose the DDQ module is posed with this question:

Which country in South America won the World Cup?

If the phrase "the World Cup" is considered ambiguous, it is necessary to ask the user to supplement information corresponding to "the World Cup" such as the name of the sport (i.e. soccer, volleyball), the venue, the season, and other characteristics. For example, the following DQs can be hypothesized by inserting an ambiguous

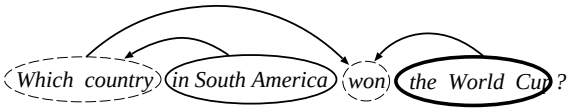

Fig. 2. Example of dependency structure.

phrase into the templates.

What kind of World Cup?
What year was the World Cup held?
Where is South America?

The linguistic appropriateness of DQs can be measured by using a language model such as a trigram. The ambiguity of each phrase is measured by using the structural ambiguity and generality score for the phrase.

The structural ambiguity is based on the dependency structure of the sentence. A phrase that is not modified by other phrases is considered to be highly ambiguous. Figure 2 has an example of a dependency structure, where the question is separated into phrases. Each arrow represents the dependency between two phrases. Here, no phrases modify "the World Cup." We assume that ambiguity for such a phrase would be higher than for others. The structural ambiguity of the n-th phrase is defined as

$$A_D(P_n) = \log \left\{ 1 - \sum_{i=1:i\neq n}^{N} D(P_i, P_n) \right\},$$

where the complete question is separated into N phrases, and $D(P_i, P_n)$ is the probability that phrase P_n will be modified by phrase P_i, which can be calculated using Stochastic Dependency Context Free Grammar (SDCFG) [11].

In addition, the generality score of a phrase is also incorporated into measuring the ambiguity of noun/noun-phrases. Nouns/noun-phrases that frequently occur in a corpus rarely help to extract answers. We assume that such a phrase is ambiguous and should be modified by additional information. The generality score is defined as

$$A_G(P_n) = \sum_{w\in P_n : w=\text{cont}} \log P(w),$$

where $P(w)$ is the unigram probability of w based on the corpus to be retrieved. "$w = $ cont" means that w is a content word such as a noun, verb or adjective.

Let S_{mn} be a DQ generated by inserting the n-th phrase into the m-th template. The DDQ module selects the DQ that maximizes the DQ score:

$$H(S_{mn}) = \lambda_L L(S_{mn}) + \lambda_D A_D(P_n) + \lambda_G A_G(P_n),$$

where $L(\cdot)$ is a linguistic score such as the logarithm for trigram probability. λ_L, λ_D, and λ_G are weighting factors to balance the scores.

Our system is actually built for Japanese speech. Japanese sentences can be divided into phrase-like units (*bunsetsu*). The phrase-like unit *bunsetsu* is denoted by 'phrase'. Since a new phrase always starts from a content word, a sentence is split into a phrase sequence based on the first content word. Each phrase is made up of a content word followed by zero or more function words, and each word modifies succeeding words within the phrase. In addition, since Japanese sentences have only "right-headed" dependency, the dependency probability $D(P_k, P_l)$ is 0 if $k \geq l$.

Question reconstructor

Additional information drawn out by DQs is incorporated into screened questions to reconstruct original questions. Suppose the answer for the DQ, "In what year was the World Cup held?", is "2002", the reconstructed question is "Which country in South America won the World Cup in 2002?".

3. EVALUATION EXPERIMENTS

Questions consisting of 69 sentences read aloud by seven male speakers were transcribed by our ASR system [8]. These questions were prepared to test the performance of our ODQA engine [10]. Each question consisted of about 19 morphemes on average. The sentences were grammatically correct, formally structured, and had enough information for the ODQA engine to extract the correct answers. Therefore, transcription results with 100% word accuracy could extract answers accurately. In contrast, transcription results with recognition errors failed to extract correct answers. The mean word recognition accuracy of 69 questions was 76%. The question transcriptions were processed with a screening filter and input into the ODQA engine. The DDQ module generated DQs based on the screened results. The questions were also manually reconstructed by combining additional information required by the DQs with the screened questions. To test the effect of the additional information had, answers to the reconstructed questions were extracted through the ODQA system.

3.1. ASR system

The speech signal was sampled at 16 kHz with 16 bit quantization. Feature vectors had 25 elements consisting of 12 MFCC, their delta, and delta log energies. Tied-state triphone HMMs with 3000 states and 16 Gaussians per state were prepared by using 338 spontaneous presentations uttered by male speakers (approximately 59 hours). Decoding was done with a one-pass Viterbi search using WFST, integrating cross-word triphone HMMs and trigrams [8].

3.2. Screening filter

Screening was done by removing recognition errors using a confidence measure as a threshold and then summarizing it within an 80% to 100% compaction ratio. In this summarization technique [9], the word significance and linguistic score for summarization were calculated using text from the Mainichi newspaper published from 1994 to 2001, comprised of 13.6M sentences with 232M words. The SDCFG for the word concatenation score was the same as that used in [9]. The posterior probability of each transcribed word in a word graph obtained by ASR was used as the confidence score.

3.3. DDQ module

The word generality score A_G was computed using the same Mainichi newspaper text that was used for screening. Eighty-two kinds of interrogative sentences were created as disambiguous queries for each noun/noun-phrase in each question and evaluated in the DDQ module. The linguistic score L indicating the appropriateness of interrogative sentences was calculated using 1000 questions and newspaper text extracted for three years. The structural ambiguity score A_D was calculated based on the SDCFG which was used for the screening filter.

3.4. Evaluation method

The questions read by the seven speakers had enough information to supply exact answers. These read questions included enough information to extract answers but they also included redundancy. Not all recognition errors resulted in loss of information that was indispensable to obtain the correct answers. Therefore, we tested the performance of the DDQ module based on the degree recognition errors were compensated and to what extent the compensated information was indispensable for QA.

To test compensation for recognition errors, the DQs generated by the DDQ module were evaluated by comparing them with manual disambiguation queries. These manual queries were presented by five human subjects based on a comparison of the original written questions and the transcription results provided by the ASR system. The automatic DQs were categorized into two classes: APPROPRIATE when they had the same meaning as at least one of the five manual DQs and InAPPROPRIATE when there was no match.

We tested the questions that were manually reconstructed using screened questions and additional information required by the DQs and the effect this additional information had on the performance of the ODQA. QA performance using recognition results was evaluated through the MRR (Mean Reciprocal Rank) [12].When the correct answer for each question was included with the top five answers given by the ODQA system, the answer was judged to be correct, and its reciprocal rank was accumulated. When QA systems outputted perfect answers, the MRRs was 1.0. The higher MRRs indicated that QA performance was higher.

3.5. Evaluation results

Table 2 shows the evaluation results in terms of the appropriateness of DQs and the QA-system MRR. The results indicate that 49% of the DQs generated by the DDQ module based on recognition results were APPROPRIATE. The mean MRRs for manual transcription (TRS), recognition questions (REC), screened questions (SCRN), and reconstructed questions using supplementary information determined by DQs (DQ) were 0.43, 0.25, 0.25 and 0.28, respectively. These MRRs demonstrate the potential of generated DQs in requiring indispensable information that is lacking to extract answers.

Table 2. Evaluation results of disambiguous queries generated by the DDQ module.

SPK	Word acc.	MRR			w/o errors	APP	In-APP
		REC	SCRN	DQ			
A	70%	0.19	0.20	0.23	4	32	33
B	76%	0.31	0.28	0.31	8	36	25
C	79%	0.25	0.26	0.30	10	34	25
D	73%	0.28	0.27	0.30	4	35	30
E	78%	0.26	0.24	0.27	7	31	31
F	80%	0.30	0.29	0.33	8	34	27
G	74%	0.19	0.19	0.22	3	35	31
AVG	76%	0.25	0.25	0.28	9%	49%	42%

An integer without a % other than MRRs indicates number of sentences. Word acc.:word accuracy, SPK:speaker, REC: transcribed questions, SCRN: screened questions, DQ: compensation by the DQs, AVG: averaged values, w/o errors: transcribed sentences without recognition errors, APP: appropriate DQs and InAPP: inappropriate DQs

4. CONCLUSION

This paper proposed a new strategy for spoken interactive ODQA (open-domain question answering) systems. In this strategy, when a user's question is ambiguous, additional information indispensable to extract the exact answer is automatically queried by the DDQ (deriving disambiguous queries) module. The DDQ module generates a DQ (disambiguous query) using an ambiguous phrase in the user's question that was extracted based on the structural ambiguity of the question and the generality of the phrase. Experimental results revealed the potential of the generated DQs in not requiring indispensable information that was lacking to extract answers. Future research will include an evaluation of the proposed strategy in a total spoken interactive ODQA system to assess how much total performance is improved by using repeated DQs.

5. REFERENCES

[1] http://trec.nist.gov

[2] F. Pereira et. al., "Definite Clause Grammars for Language Analysis –a Survey of the Formalism and a Comparison with Augmented Transition Networks," *Artificial Intelligence*, 13:231-278, 1980.

[3] E. H. Shortliffe, "Computer-Based Medical Consultations: MYCIN," *Elsevier/North Holland*, New York NY, 1976.

[4] T. Lowerre et. al., "The Harpy speech understanding system," W. A. Lea (Ed.), *Trends in Speech recognition*, pp. 340, Prentice Hall.

[5] V. Zue, et al., "JUPITER: A Telephone-Based Conversational Interface for Weather Information," *IEEE Transactions on Speech and Audio Processing*, Vol. 8 , No. 1, 2000.

[6] S. Harabagiu et. al., "Experiments with Open-Domain Textual Question Answering," *COLING-2000*, pp. 292-298, Saarbrüken Germany, August 2000.

[7] S. Harabagiu et. al., "Open-Domain Voice-Activated Question Answering," *COLING2002*, Vol.I, pp. 321–327, Taipei, 2002.

[8] D. Willett et. al., "Time and memory efficient Viterbi decoding for LVCSR using a precompiled search network," *Proc. of Eurospeech 2001*, Vol. 2, pp. 847–850, 2001.

[9] C. Hori et.al., "A New Approach to Automatic Speech Summarization," To appear in the *IEEE Transactions on Multimedia*, 2003.

[10] Y. Sasaki et. al., "NTT's QA Systems for NTCIR QAC-1," *Proc. of NTCIR Workshop Meeting*, pp. 63–70, 2000.

[11] C. Hori et.al., "A Statistical Approach for Automatic Speech Summarization," To appear in the *EURASIP Journal on Applied Signal Processing*, 2003.

[12] http://trec.nist.gov/data/qa.html