Spoken Interactive ODQA System: SPIQA

Chiori Hori, Takaaki Hori, Hajime Tsukada,
Hideki Isozaki, Yutaka Sasaki and **Eisaku Maeda**
NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan

Abstract

We have been investigating an interactive approach for Open-domain QA (ODQA) and have constructed a spoken interactive ODQA system, **SPIQA**. The system derives disambiguating queries (DQs) that draw out additional information. To test the efficiency of additional information requested by the DQs, the system reconstructs the user's initial question by combining the addition information with question. The combination is then used for answer extraction. Experimental results revealed the potential of the generated DQs.

1 Introduction

Open-domain QA (ODQA), which extracts answers from large text corpora, such as newspaper texts, has been intensively investigated in the Text REtrieval Conference (TREC). ODQA systems return an actual answer in response to a question written in a natural language. However, the information in the first question input by a user is not usually sufficient to yield the desired answer. Interactions for collecting additional information to accomplish QA are needed. To construct more precise and user-friendly ODQA systems, a speech interface is used for the interaction between human beings and machines.

Our goal is to construct a spoken interactive ODQA system that includes an automatic speech recognition (ASR) system and an ODQA system. To clarify the problems presented in building such a system, the QA systems 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, as is shown in Table 1. In this table, text and speech denote text input and speech input, respectively. The term "*addition*" represents additional information queried by the QA systems. This additional information is separate to that derived from the user's initial questions.

Table 1: Domain and data structure for QA systems

target domain		specific	open
data structure		knowledge DB	unstructured text
text	without *addition*	CHAT-80	SAIQA
	with *addition*	MYCIN	(**SPIQA***)
speech	without *addition*	Harpy	VAQA
	with *addition*	JUPITER	(**SPIQA***)

$*$ **SPIQA** is our system.

To construct spoken interactive ODQA systems, the following problems must be overcome: 1. System queries for additional information to extract answers and effective interaction strategies using such queries cannot be prepared before the user inputs the question. 2. Recognition errors degrade the performance of QA systems. Some information indispensable for extracting answers is deleted or substituted with other words.

Our spoken interactive ODQA system, **SPIQA**, copes with the first problem by adopting disambiguating users' questions using system queries. In addition, a speech summarization technique is applied to handle recognition errors.

2 Spoken Interactive QA system: SPIQA

Figure 1 shows the components of our system, and the data that flows through it. This system comprises an ASR system (**SOLON**), a screening filter that uses a summarization method, and ODQA engine (**SAIQA**) for a Japanese newspaper text corpus, a Deriving Disambiguating Queries (**DDQ**) module, and a Text-to-Speech Synthesis (TTS) engine (**FinalFluet**).

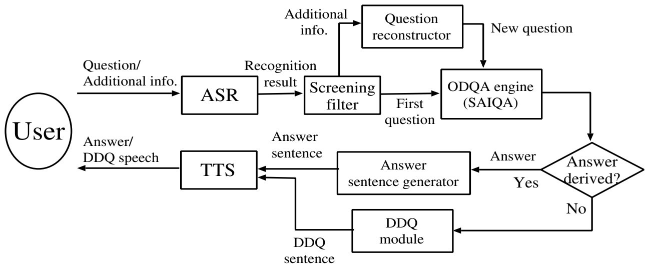

Figure 1: Components and data flow in SPIQA.

ASR system

Our ASR system is based on the Weighted Finite-State Transducers (WFST) approach that is becoming a promising alternative formulation for the traditional decoding approach. The WFST approach offers a unified framework representing various knowledge sources in addition to producing an optimized search network of HMM states. We combined cross-word triphones and trigrams into a single WFST and applied a one-pass search algorithm to it.

Screening filter

To alleviate degradation of the QA's performance by recognition errors, fillers, word fragments, and other distractors in the transcribed question, a screening filter that removes these redundant and irrelevant information and extracts meaningful information is required. The speech summarization approach (C. Hori et. al., 2003) is applied to the screening process, wherein a set of words maximizing a summarization score that indicates the appropriateness of summarization is extracted automatically from a transcribed question, and these words are then concatenated together. The extraction process is performed using a Dynamic Programming (DP) technique.

ODQA engine

The ODQA engine, **SAIQA**, has four components: question analysis, text retrieval, answer hypothesis extraction, and answer selection.

DDQ module

When the ODQA engine cannot extract an appropriate answer to a user's question, the question is considered to be "ambiguous." To disambiguate the initial questions, the DDQ module automatically derives disambiguating queries (DQs) that require information indispensable for answer extraction. The situations in which a question is considered ambiguous are those when users' questions exclude indispensable information or indispensable information is lost through ASR errors. These instances of missing information should be compensated for by the users.

To disambiguate a question, ambiguous phrases within it should be identified. The ambiguity of each phrase can be measured by using the structural ambiguity and generality score for the phrase. The structural ambiguity is based on the dependency structure of the sentence; 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. In this example, "the World Cup" has no

Figure 2: Example of dependency structure.

modifiers and needs more information to be identified. "Southeast Asia" also has no modifiers. However, since "the World Cup"appears more frequently than "Southeast Asia" in the retrieved corpus, "the World Cup" is more difficult to identify. In other words, words that frequently occur in a corpus rarely help to extract answers in ODQA systems. Therefore, it is adequate for the DDQ module to generate questions relating to "World Cup" in this example, such as "*What kind of World Cup?*" , "*What year was the World Cup held?*".

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) (C. Hori et. al., 2003).

Using this SDCFG, only the number of non-terminal symbols is determined and all combinations of rules are applied recursively. The non-terminal symbol has no specific function, such as a noun phrase. All the probabilities of rules are stochastically estimated based on data. Probabilities for frequently used rules become greater, and those for rarely used rules become smaller. Even though transcription results given by a speech recognizer are ill-formed, the dependency structure can be robustly estimated by our SDCFG.

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. Thus, "w = cont" means that w is a content word such as a noun, verb or adjective.

We generate the DQs using templates of interrogative sentences. These templates contain an interrogative and a phrase taken from the user's question, i.e., "What kind of * ?", "What year was * held?" and "Where is * ?".

The DDQ module selects the best DQ based on its linguistic appropriateness and the ambiguity of the phrase. The linguistic appropriateness of DQs can be measured by using a language model, N-gram. 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, and λ_L, λ_D, and λ_G are weighting factors to balance the scores.

Hence, the module can generate a sentence that is linguistically appropriate and asks the user to disambiguate the most ambiguous phrase in his or her question.

3 Evaluation Experiments

Questions consisting of 69 sentences read aloud by seven male speakers were transcribed by our ASR system. The question transcriptions were processed with a screening filter and input into the ODQA engine. 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. The mean word recognition accuracy obtained by the ASR system was 76%.

3.1 Screening filter

Screening was performed 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, the word significance and linguistic score for summarization were calculated using text from Mainichi newspapers published from 1994 to 2001, comprising 13.6M sentences with 232M words. The SDCFG for the word concatenation score was calculated using the manually parsed corpus of Mainichi newspapers published from 1996 to 1998, consisting of approximately 4M sentences with 68M words. The number of non-terminal symbols was 100. The posterior probability of each transcribed word in a word graph obtained by ASR was used as the confidence score.

3.2 DDQ module

The word generality score A_G was computed using the same Mainichi newspaper text described above, while the SDCFG for the dependency ambiguity score A_D for each phrase was the same as that used in (C. Hori et. al., 2003). Eighty-two types of interrogative sentences were created as disambiguating queries for each noun and noun-phrase in each question and evaluated by 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.

3.3 Evaluation method

The DQs generated by the DDQ module were evaluated in comparison with manual disambiguation queries. Although the questions read by the seven speakers had sufficient information to extract exact answers, some recognition errors resulted in a loss of information that was indispensable for obtaining the correct answers. The manual DQs were made by five subjects based on a comparison of the original written questions and the transcription results given 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. The QA performance in using recognized (REC) and screened questions (SCRN) were evaluated by MRR (Mean Reciprocal Rank) (http://trec.nist.gov/data/qa.html). SCRN was compared with the transcribed question that just had recognition errors removed (DEL). In addition, the questions reconstructed manually by merging these questions and additional information requested the DQs generated by using SCRN, (DQ) were also evaluated. The additional information was extracted from the original users' question without recognition errors. In this study, adding information by using the DQs was performed only once.

3.4 Evaluation results

Table 2 shows the evaluation results in terms of the appropriateness of the DQs and the QA-system MRRs. The results indicate that roughly 50% of the DQs generated by the DDQ module based on the screened results were APPROPRIATE. The MRR for manual transcription (TRS) with no recognition errors was 0.43. In addition, we could improve the MRR from 0.25 (REC) to 0.28 (DQ) by using the DQs only once. Experimental results revealed the potential of the generated DQs in compensating for the degradation of the QA performance due to recognition errors.

4 Conclusion

The proposed spoken interactive ODQA system, **SPIQA** copes with missing information by adopting disambiguation of users' questions by system queries. In addition, a speech summarization technique was applied for handling recognition errors. Although adding information was performed using DQs only once, experimental results revealed the potential of the generated DQs to acquire indispensable information that was lacking for extracting answers. In addition, the screening filter helped to generate the appropriate DQs. Future research will include an evaluation of the appropriateness of DQs derived repeatedly to obtain the final answers. In addition, the interaction strategy automatically generated by the DDQ module should be evaluated in terms of how much the DQs improve QA's total performance.

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

SPK	Word acc.	MRR				w/o errors	APP	IN-APP
		REC	DEL	SCRN	DQ			
A	70%	0.19	0.16	0.17	0.23	4	32	33
B	76%	0.31	0.24	0.29	0.31	8	36	25
C	79%	0.26	0.18	0.26	0.30	10	34	25
D	73%	0.27	0.21	0.24	0.30	4	35	30
E	78%	0.24	0.21	0.24	0.27	7	31	31
F	80%	0.28	0.25	0.30	0.33	8	34	27
G	74%	0.22	0.19	0.19	0.22	3	35	31
AVG	76%	0.25	0.21	0.24	0.28	9%	49%	42%

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

References

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.

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

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

L. D. Erman et. al., "The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty," *ACM computing Survays*, Vol. 12, No. 2, pp. 213 – 253, 1980.

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.

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

C. Hori et. al., "A Statistical Approach for Automatic Speech Summarization," *EURASIP Journal on Applied Signal Processing (EURASIP)*, pp128–139, 2003.

Y. Sasaki et. al., "NTT's QA Systems for NTCIR QAC-1," *Working Notes of the Third NTCIR Workshop Meeting*, pp.63–70, 2002.