OVERVIEW OF AN INTELLIGENT SYSTEM FOR INFORMATION RETRIEVAL BASED ON HUMAN-MACHINE DIALOGUE THROUGH SPOKEN LANGUAGE

Hiroya Fujisaki [1], *Katsuhiko Shirai* [2], *Shuji Doshita* [3], *Seiichi Nakagawa* [4],
Keikichi Hirose [5], *Shuichi Itahashi* [6], *Tatsuya Kawahara* [7], *Sumio Ohno* [8],
Hideaki Kikuchi [9], *Kenji Abe* [1] *and Shinya Kiriyama* [5]

[1] Science University of Tokyo [2] Waseda University [3] Ryukoku University
[4] Toyohashi University of Technology [5] University of Tokyo [6] University of Tsukuba [7] Kyoto University
[8] Tokyo University of Technology [9] National Language Research Institute

ABSTRACT

This paper presents an intelligent system for information retrieval based on human-machine dialogue through spoken language with novel features such as use of key concepts, unknown word processing, dialogue management through user and system modeling, and automatic acquisition of knowledge to adapt the system to individual users. It then describes an experimental system constructed to implement these features and to demonstrate their feasibility.

1. INTRODUCTION

With the rapid and widespread use of information networks in our society, it has become increasingly important to search and retrieve information that is truly relevant. Existing search engines are designed to be easy to use for inexperienced users, but are too simple and too inefficient. On the other hand, conventional systems for information retrieval using more sophisticated procedures are not easy for inexperienced users since they presuppose certain amount of knowledge on the structure of databases and on the method of constructing the search formula. Furthermore, it is difficult for the user to express his/her intention precisely, and also for the system to infer the user's intention correctly. These difficulties can be greatly reduced by introducing spoken dialogue between the user and the system.

From this point of view, an intelligent system for information retrieval has been proposed using spoken dialogue as the main medium for user, and the research was started in August 1996 as one of the large-scale projects on 'Research for the Future' supported by the Japan Society for the Promotion of Science, and has been conducted by a team of 17 members from 10 universities and five private organizations. The present paper gives an overview of the experimental system constructed by the concerted efforts of these members and their collaborators.

2. BASIC PRINCIPLES

2.1. Spoken Dialogue Between User and System

In many cases, a user is not fully aware, nor has sufficient knowledge, of the information which he/she wishes to retrieve. In traditional information retrieval services, it is often the case that the user's query becomes definite only after he/she gets some knowledge through interview with the searcher. In the present system, therefore, spoken dialogue is introduced between the user and the system to facilitate the process of formation and expression of query on the part of the user, and also the process of its clarification on the part of the system.

2.2. Use of Key Concepts

In conventional systems for information retrieval using keywords, the user's query is represented by keywords, and the search is based solely on transcriptions, but not on concepts. In the case of polysemy/homonymy, this leads to retrieval of irrelevant items, since the system is not capable of separating the intended keyword from unintended keywords. In the case of synonymy, a conventional system will retrieve only those items that contain the keyword given by the user, and will fail to retrieve items containing its synonyms. These difficulties are avoided in the present system by adopting key concepts rather than keywords.

2.3. Processing of Unknown Keywords

Since the user's query is represented only by the surface form, it is necessary to refer to a lexicon (*i.e.*, an ordered collection of transcription-concept pairs) to infer the intended key concept, and to resolve ambiguity through dialogue with the user, if necessary. Since the lexicon in the present system can be accessed also by concepts, it is used to find synonyms of the keyword. When the keyword given by the user is 'unknown', the system obtains the underlying key concept through dialogue with the user. If, however, a keyword found in a document in the database is unknown to the system, the system infers the corresponding concept from its surface form and context.

2.4. Knowledge Acquisition

Processing of unknown keywords requires a large amount of knowledge, both linguistic and non-linguistic, that cannot be given to the system in advance. Furthermore, accurate and efficient information retrieval requires knowledge concerning the characteristics of both individual users and individual databases. Thus the system has the capability of acquiring automatically all kinds of knowledge that are useful for improving its performance.

3. AN IMPLEMENTATION OF THE SYSTEM

3.1. System Configuration

The experimental system consists of (a) speech recognizer, (b) dialogue manager, (c) database search engine, (d) speech synthesizer, and (e) system controller as shown in Fig. 1. The speech recognizer, the dialogue manager, and the speech synthesizer receive/send necessary information from/to other components through the system controller. The data exchange between these components and the system controller is executed by using two common modules: one is the communicator module which communicates the information on TCP/IP network, the other is the filter module, which converts the data format. This setup makes it easy to implement alternative elements. The database search engine module communicates directly with the dialogue management module.

In the experimental system, the speech recognizer and other components run separately on two notebook-type personal computers which are connected through a switching-hub. These components communicate with each other on the Internet and operate almost real-time, so that the entire system performs also almost real-time.

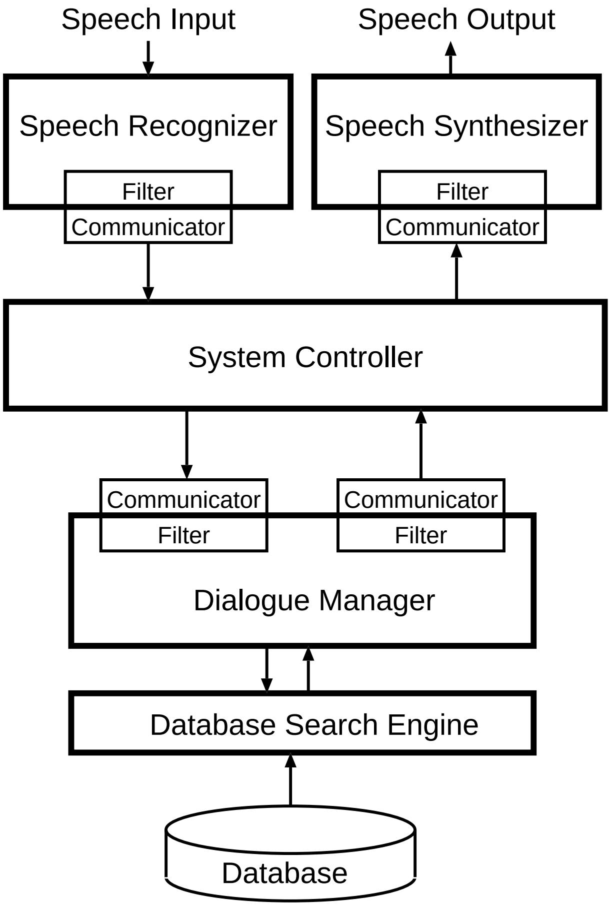

Figure 1: System Configuration.

3.2. Speech Recognizer

The JULIAN [4], developed at Kyoto University, was adopted as the speech recognition engine. It is a core component of a large vocabulary continuous speech recognition system designed as a baseline platform for research. It can deal with context-free grammar as the language model. The lexicon and the grammar for the language model were prepared for the academic information retrieval task. The vocabulary size is about 5,000 words found in 500 academic articles in the database. The number of grammar rules is 125 in the form of context-free rules obtained from transcription of simulated dialogues concerning academic information retrieval. Tables 1 and 2 illustrate a part of the lexicon and a part of the grammar rules, respectively.

As for the acoustic model, continuous density HMM with a mixture of 16 PDFs trained for male speaker was adopted. This model is a component of the Japanese dictation toolkit provided by IPA [4]. The number of states is 2,000 after clustering.

This recognition engine performs almost real-time by using a PC with the 700 MHz Pentium III processor and a 512 Mbyte memory.

Table 1: A part of the lexicon.

Word	Pronunciation in Japanese
% _TITL	
‘title’+301	t a i t o r u
% _AUTH	
‘author name’+302	ch o sh a m e i
% _ABST	
‘abstract’+303	a b u s u t o r a k u t o
% _KYWD	
‘keyword’+304	k i: w a: d o
% _YEAR	
‘published year’+305	h a q k o: n e N
% _JNAL	
‘journal name’+306	z a q sh i m e i
% _KWD_N	
‘first order Markov process’+200008	i ch i j i m a r u k o f u k a t e i
‘DP matching’+200241	d i: p i: m a q ch i N g u
‘HMM’+200338	e: ch i e m u e m u
‘hidden Markov model’+200339	k a k u r e m a r u k o f u m o d e r u
‘MAP estimation’+200439	m a q p u s u i t e i
‘array signal processing’+200897	a r e: sh i N g o: sh o r i
‘image matching’+200897	i m e: j i m a q ch i N g u
‘internet’+200904	i N t a: n e q t o
‘interface’+200913	i N t a f e: s u
:	

Table 2: A part of the grammar rules.

KWDS	:	KWD
KWDS	:	KWD KWD_V
KWDS	:	TWO_KWD
KWDS	:	TWO_KWD KWD_V
KWDS	:	KWD_FACT KWD_FACT_V
KWD_FACT	:	KWD_FACT _DE NOISE KWD_FACT2
KWD_FACT	:	_AUTH _GA NOISE _AUTH_N
KWD_FACT	:	_JNAL _GA NOISE _JNAL_N
KWD_FACT	:	_YEAR _GA NOISE _YEAR_N
KWD_FACT2	:	_AUTH _GA NOISE _AUTH_N
KWD_FACT2	:	_JNAL _GA NOISE _JNAL_N
KWD_FACT2	:	_YEAR _GA NOISE _YEAR_N
KWD_V	:	_DESU
KWD_V	:	_NITUITE _NANDESUGA
KWD_V	:	_NITUITE _NO _RONBUN _NANDESUGA
KWD_V	:	_NITUITE _NO _RONBUN _WO _SAGASITEIRUNDESUGA
KWD_FACT_V	:	_DESU
KWD_FACT_V	:	_NANDESUGA
KWD_FACT_V	:	_NO _RONBUN _NANDESUGA
:		

3.3. Dialogue Manager and Database Search Engine

The dialogue manager was developed at Science University of Tokyo [5]. One of its essential features is dialogue management through user and system modeling. Models of the user and the system were constructed as separate finite-state automata which exchange information through dialogue. The states in the user model are meant to represent the internal states of the user's intention. It should be noted, however, that the user model is *not* the user's mind itself, but is only an approximate representation to help the system to make inference on the true internal state of the user's mind and to predict the user's next utterance.

In order to construct user and system models, a large number of simulated dialogues on information retrieval were collected, by restricting the task domain to academic information retrieval, especially of articles in academic journals. Ten speakers played the role of users and seven speakers played the role of the system. The speaker playing the role of a user was assumed to have some prior knowledge of the document to be retrieved, but was allowed to speak without any constraint. The speaker playing the role of the system was assumed to have knowledge of functions and limitations of the system, and the utterances were chosen from a pre-determined set. Out of this corpus, 100 dialogues containing 3417 utterances were analyzed to study and classify the internal states of the user and the system.

A state in the user model represents a stage of information processing on the part of the user. Transition to another state is elicited by the system's response in the form of a spoken message with or without being accompanied by further information on a graphical user interface, and is usually accompanied by a user's utterance. The user model thus constructed does not represent any particular user but represents the average of a number of users. However, this average model can be adapted automatically to a specific user by using the record of his/her interactions with the system.

In the present system, the user's internal states are identified on the basis of recognition of the user's current utterance and dialogue history. The system's action is then determined by referring to the system information (current state and past history of information retrieval) and the user's state. Separate modeling of user and system not only reduces the complexity of the entire model, but also facilitates its adaptation to individual users as well as to new dialogue domains.

The database search engine module, which communicates directly with the dialogue management module, has various indices in the database. Author names, keywords, journal names, etc. are available as retrieving keys. The database consists of about 500 academic articles in the current study.

3.4. Speech Synthesizer

The speech synthesizer module was developed at University of Tokyo [6]. This module deals with message generation and speech synthesis.

With the aim of realizing speech response that is easy to be understood by users, special emphasis was placed on the following two points in developing the system.

1. To optimize the degree of redundancy in message generation.

In order to see how message comprehension is affected by the degree of redundancy, the following three levels of redundancy was adopted in message generation in speech synthesis, and tested for comprehension:

(1) messages from which the information already known to the listener (user) is completely eliminated,

(2) messages from which only the information contained in the user's immediately preceding utterance is eliminated,

(3) messages in which no information known to the user is eliminated.

The results of preliminary experiments using the system indicated that (3) was not appropriate, since the user finds it difficult to know whether their queries were correctly recognized by the system or not. Also, (1) was not appropriate since it occasionally causes misunderstanding by the user. Thus controlling the degree of redundancy to a certain degree is quite important. Elliptic expressions in the user's utterances also need to be processed. Namely, when the information necessary to access a database is not found in the user's utterance, the system has to go through past history of the dialect. If it is not found there, the system has to ask the user for the missing information.

2. To synthesize prosodic features that properly indicate the focus of an utterance.

For this purpose, a scheme for concept-to-system conversion, rather than text-to-speech conversion, was found to be necessary. Namely, in the current system, a conceptual representation for the system's utterance is first generated by referring to the user's utterance and the dialogue flow. This conceptual representation is then converted into a sequence of prosodic phrases and prosodic words, and further converted into a sequence of segmental and prosodic symbols. These symbols can then be used to control the speech synthesizer. During this conversion process, words that constitute the focus of an utterance is determined and prosodically emphasized. Basically, emphasis is placed on words containing information crucial to answer the user's query. Since the system's response starts with a conceptual representation and prosodic phrases are generated during the conversion process, focus assignment is quite straightforward.

Although a number of speech synthesis systems exist that can generate speech with a reading-style prosody, they are not suited for dialogue applications. Therefore, rules for prosodic control of dialogue-style speech were introduced. These rules generate F_0 contours using the command-response model, whose phrase and accent command parameters are determined in advance on the basis of multiple regression analysis. Other rules were also introduced to control the speech rate: slow at the beginning and the end, and fast at the middle of an utterance.

The speech synthesizer adopted for the experimental system is a formant synthesizer consisting of four sets of cascaded connections of pole/zero filters corresponding to vowels, nasals, fricatives and plosives, respectively. The formant synthesizer was adopted because it is more suitable than PSOLA-type synthesizers in realizing wider ranges of variations in both F_0 and speech rate.

3.5. System Controller

The system controller module was developed at Waseda University. It is a part of the general purpose platform for a spoken dialogue system [7]. This module communicates with other system components with standardized messages which contain the type and the content of information. The controller receives messages from other components through communication modules. It has rules of processing these messages to decide where the destination of a message is and what the required next action is. Table 3 shows the rules for message handling. The messages “`RecogRes`” and “`DlgRes`” in this table represent the results of speech recognizer and dialogue management, respectively. The message “`SttofUtt`” occurs when the onset of user’s utterance is found in the speech recognizer, while the message “`EndofUtt`” occurs when the offset of user’s utterance is found. In the system controller, pause duration between user’s utterances is measured by watching the last `EndofUtt` message and the next `SttofUtt` message. When pause duration exceeds a given threshold, the system controller produces an “`Event`” message and sends it to the dialogue management module. This `Event` message is used to initiate the inference on the internal state of the user’s model in the dialogue management module. Table 4 shows an example of the dialogue between the system and a user.

Table 3: Rules for message handling.

Message Type	Source	Destination	Action
`SttofUtt`	Speech Recognizer		Stop clocking utterance length
`EndofUtt`	Speech Recognizer		Start clocking utterance length
`RecogRes`	Speech Recognizer	Dialogue Manager	Judge priority, Send
`DlgRes`	Dialogue Manager	Speech Synthesizer	Send
`Event`		Dialogue Manager	Send

Table 4: An example of the dialogue between the system and a user.

U1:	Hello.
S1:	What kind of articles do you want?
U2:	I want to find articles on speaker adaptation.
S2:	There are 21 articles.
U3:	I want to change the condition of search.
S3:	What kind of articles do you want?
U4:	Articles on speaker adaptation without supervisor.
S4:	There are 3 articles.
U5:	Please show me their abstracts.
S5:	Sure.
U6:	Thank you.

4. SUMMARY AND CONCLUSION

This paper has presented an overview of an intelligent system for information retrieval based on human-machine dialogue through spoken language. From the point of view of information retrieval, it has several novel features such as use of key concepts and processing of unknown words. From the point of view of human-machine dialogue, it uses a new method of dialogue management based on modeling the user and the system as separate finite-state automata, and automatic acquisition of knowledge to adapt the system to individual users. These features have been implemented in an experimental system.

ACKNOWLEDGMENT

The current work was supported by the Japan Society for the Promotion of Science as a Project on ‘Research for the Future’ (Project No. JSPS-RFTF-96R15201). The authors wish to thank all the members of the project for their valuable contributions and helpful discussions.

REFERENCES

[1] Fujisaki, H., Kameda, H., Ohno, S., Ito, T., Tajima, K. and Abe, K.: “An intelligent system for information retrieval over the Internet through spoken dialogue,” *Proceedings of Eurospeech’97*, vol. 3, pp. 1675–1678, 1997.

[2] Fujisaki, H., Kameda, H., Ohno, S., Abe, K., Iijima, M., Suzuki, M. and Taketa, K.: “Principles and design of an intelligent system for information retrieval over the Internet with a multimodal dialogue interface,” *Proceedings of Eurospeech’99*, vol. 6, pp. 2467–2470, 1999.

[3] Fujisaki, H.: “Issues in realization of intelligent systems for information retrieval through human-machine dialogue,” *Proceedings of IRAL ’99*, pp. KN2-1–8, 1999.

[4] Kawahara, T., Kobayashi, T., Takeda, K., Minematsu, N., Itou, K., Yamamoto, M., Yamada, A., Utsuro, T., and Shikano, K.: “Japanese dictation toolkit — plug-and-play framework for speech recognition R&D —,” *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 393–396, 1999.

[5] Abe, K., Kuorkawa, K., Taketa, K., Ohno, S. and Fujisaki, H.: “A new method for dialogue management in an intelligent system for information retrieval,” To appear in *Proceedings of ICSLP 2000*, October 2000.

[6] Hirose, K. and Kiriyama, S.: “Generation of speech reply in a spoken dialogue system for literature retrieval,” *Proceedings of ESCA Tutorial and Research Workshop on Interactive Dialogue in Multi-Modal Systems*, pp.29–32, 1999.

[7] Aoyama, K., Hirano, I., Kikuchi, H. and Shirai, K.: “Designing a domain independent platform of spoken dialogue,” To appear in *Proceedings of ICSLP 2000*, October 2000.