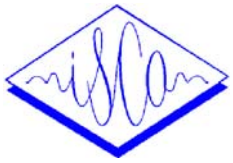

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PRINCIPLES AND DESIGN OF AN INTELLIGENT SYSTEM FOR INFORMATION RETRIEVAL OVER THE INTERNET WITH A MULTIMODAL DIALOGUE INTERFACE

Hiroya Fujisaki [1], *Hiroyuki Kameda* [2], *Sumio Ohno* [1], *Kenji Abe* [1], *Michio Iijima* [1], *Masayoshi Suzuki* [1] *and Kazunari Taketa* [1]

[1] Science University of Tokyo, Noda, 278-8510 Japan
[2] Tokyo Engineering University, 1404-1 Katakura, Hachiouji, 192-8580 Japan

ABSTRACT

In the information society of the next millenium, information retrieval over the Internet will be indispensable for everyday life, and spoken language will be an essential medium for human-machine dialogue. This paper presents an overview of an intelligent system for information retrieval based on spoken dialogue, use of key concepts, processing of unknown words, knowledge acquisition, and agent technology. Three agents are introduced to be respectively responsible for user interface, information retrieval, and information acquisition. Dialogue management through user and system modeling, implementation of information retrieval through key concepts, and inference on concepts of unknown words based on syntactic and semantic analyses of their structures, are then briefly described as the innovative features of the system.

1. INTRODUCTION

With the rapid progress of computer technology and worldwide development of information networks, a vast amount of information is now being generated, published, and stored at a number of sites distributed all over the world. Such an affluence of information, however, is useless or may even become harmful unless one has a means for rapidly retrieving the information that is truly necessary and appropriate. In this respect, conventional systems for information retrieval are far from being satisfactory, and tend to collect irrelevant information as well as to miss relevant information. These situations can be ascribed, partly to the difficulty for the user to identify and express his/her intention precisely, and partly to the difficulty 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.

While keyword search is suited for retrieving information from databases not necessarily designed by a common principle, both the accuracy and efficiency tend to be low because of polysemy and synonymity of keywords. These difficulties can be overcome by using 'key concepts' [1] rather than keywords. In this case, information retrieval is based, not on the surface forms of keywords, but on their semantic contents intended by the user. Difficulties arising from polysemy of keywords can be solved by spoken dialogue if all the keywords are 'known' to the system (*i.e.*, already registered in the lexicon of the system).

In actual information retrieval situations, where new words or new compound words made by combining known morphemes commonly occur, it is impossible that all the keywords are registered in the lexicon. Thus the system must have the ability to infer the meaning of new keywords that are 'unknown' (*i.e.*, not registered in the lexicon)[2]. In other words, the system has to possess the ability of knowledge acquisition. This is also necessary if one aims at an intelligent system which will automatically improve its performance.

The basic principles of such an intelligent system for information retrieval has been presented in a previous paper [3]. The present paper shows an overview of the prototype system design, and describes some of its essential features.

2. OVERVIEW OF THE PROTOTYPE SYSTEM

As described in a previous paper, the system is designed on the basis of the following principles:

- Spoken dialogue between user and system
- Use of key concepts
- Processing of unknown words
- Knowledge acquisition
- Use of agent technology

Figure 1 shows an overview of the prototype system designed and being constructed. The main components of the system are:

(1) User Interface Agent (UIA)

The user interface agent helps the user to clarify and state his/her intention mainly through spoken dialogue, supplemented by a graphic display. Its detailed structure and function will be described in connection with the dialogue management.

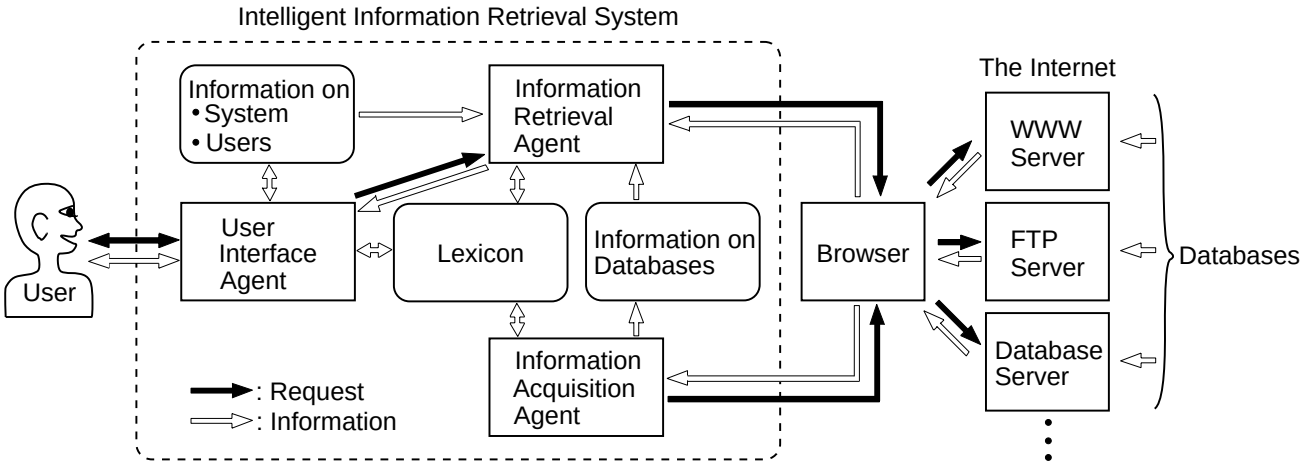

Fig. 1. Overview of the prototype system

(2) Information Retrieval Agent (IRA)

The information retrieval agent conducts the retrieval of information requested by the user. On the basis of keywords which UIA obtains from the user, IRA constructs a search formula that maximally covers the user's intended key concepts. This is accomplished by utilizing a keyword/key-concept lexicon. Failure to retrieve relevant data due to synonymity of keywords are minimized by adopting all the synonyms of the keywords supplied by the user, while irrelevant information is removed by using the collocation information. Each retrieved item is given a relevance score, which is then used to determine the order of presentation to the user if the number of retrieved items is beyond a certain threshold. Detailed descriptions of key-concept based retrieval will be given in Section 4.

(3) Information Acquisition Agent (IAA)

The information acquisition agent acquires knowledge on various databases available on the Internet, and stores it as link data concerning the address, relevant key concepts and keywords representing each database. It also acquires new keywords through unknown word processing, *viz.*, it detects and infers the concepts of unknown words found in databases and registers them as new keyword/key-concepts in the lexicon.

3. DIALOGUE MANAGEMENT THROUGH USER AND SYSTEM MODELING

Figure 2 shows a schematic diagram of UIA. One of its essential features is dialogue management through user and system modeling. In conventional dialogue systems, dialogue management is performed by modeling the dialogue itself. This is done by analyzing actual dialogues and constructing a state-transition diagram for representing possible exchanges between a user and the system. Such an approach is not ideal since it does not describe the user and the system separately, and therefore leads to complexity and inflexibility. In the present

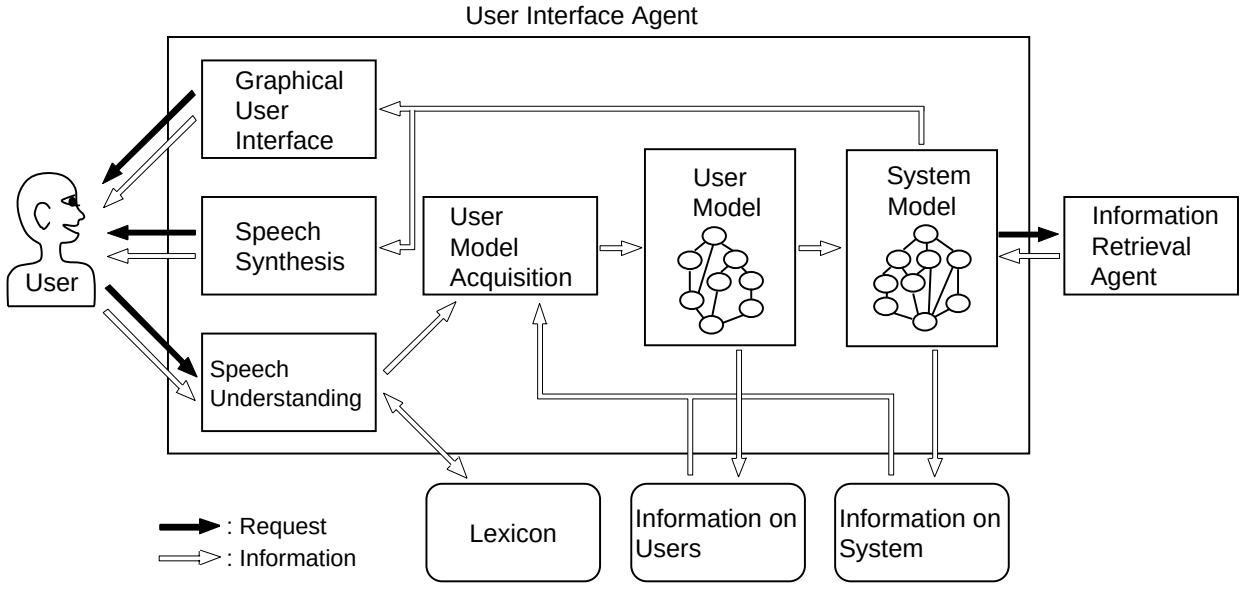

Fig. 2. Schematic diagram of UIA

system, we construct models of the user and the system as separate finite-state automata which exchange information through dialogue. In particular, the states in the user model represent the internal states of the user's knowledge and intention.

This approach has the following advantages over the conventional approach:

(1) Clearer and simpler description of the dialogue
(2) Possibility of separately modifying models of the user and the system

The first point is obvious. The second point is important since each user, exactly speaking, is different in his/her background, knowledge, interest, intention, *etc.*, and adapting the dialogue to suit each user will require adjustment or modification of the user model. It is also easier to modify the system for a new or an improved service if we have a separate model for the system.

4. IMPLEMENTATION OF INFORMATION RETRIEVAL THROUGH KEY CONCEPTS

When the user's intention is extracted by UIA as a set of keywords, they are sent to IRA which constructs a search formula. In order to minimize the failure to retrieve relevant data, all the synonyms of the original keywords are assembled from the lexicon (*i.e.*, the original set of keywords are expanded to include their synonyms), and a search formula is constructed, simply by taking their logical sum. This, however, increases the number of irrelevant data among those retrieved, and hence the cost for retrieval.

One way to find a compromise between the miss (*i.e.*, failure to retrieve relevant data) and the false alarm (*i.e.*, retrieval of irrelevant data) is to introduce a certain measure of relevance for each of the available data, and a threshold for its selection. As a tentative measure for the relevance, the following score is defined for a given data (*i.e.*, document) d:

$$S_d = \sum_{i=1}^{N_d} v_d(i),$$

$$v_d(i) \begin{cases} = c \log n_i + 1, & \text{for } n_i \geq 1, \\ = 0, & \text{for } n_i = 1, \end{cases}$$

where

n_i: number of times that a keyword w_i appears in document d
$v_d(i)$: contribution of keyword w_i to the score for document d
N_d: total number of keywords for document d

By normalizing the total score of each data by the maximum score found for a set of data, and by introducing a relative threshold for selection, the number of misses and false alarms can be experimentally obtained. The relevance of each data was based on human judgment. At first, each data was given a 5-point rating (4, 3, 2, 1, 0) of relevance for the specific request, and those data with the highest rating were treated as relevant.

Figures 3 and 4 compare the proposed scoring method with the conventional *tf·idf* method in an experiment on information retrieval, and the curves indicate the averages of 30 trials. Figure 3 shows the miss rates (to be denoted by M) as a function of the relative threshold θ of the relevance score. It is clear that the proposed scoring method consistently gives lower miss rates. Figure 4 shows the false alarm rates (to be denoted by F). Again the proposed scoring method is found to give lower alarm rates in almost all the cases of interest.

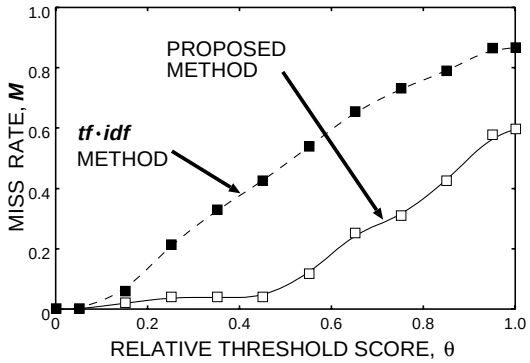

Fig. 3. Comparison of miss rates obtained by the proposed method and by the *tf·idf* method.

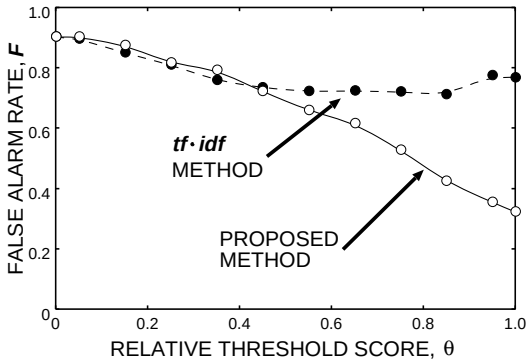

Fig. 4. Comparison of false alarm rates obtained by the proposed method and by the *tf·idf* method.

The optimum value for the threshold θ can be found by minimizing the combined loss which is a weighted sum of the losses due to misses and false alarms,

$$L(\theta, \alpha) = M(\theta) + \alpha F(\theta).$$

In the above example, $L(\theta)$ is lowest at $\theta = 0.5$ for $\alpha = 1$.

In some cases, the loss should be expressed, not by percentages, but by the actual numbers of the data missed or falsely retrieved. In still other cases, the loss (cost)

may best be expressed as the cost of retrieval for one data correctly retrieved. All these cases can be easily accommodated on the basis of the above formulation. Figure 5 shows the actual numbers of misses and false alarms of the proposed scoring method.

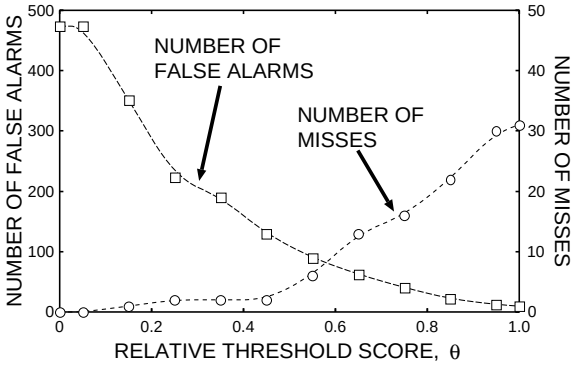

Fig. 5. Numbers of misses and false alarms

5. PROCESSING OF UNKNOWN WORDS

As already mentioned, processing of unknown words is an important task for IAA. Unknown words can be classified into the four categories:

• Category 1 - A word for which only the transcription is registered in the lexicon. This occurs when two or more keywords with the identical surface form exist, but not all of them are registered in the lexicon.

• Category 2 - A word for which only the concept is registered in the lexicon. This occurs often in Japanese, which allows more than one transcriptions because of lack of strict orthographic rules, and lexicons usually do not list all the variations of transcription.

• Category 3 - A word whose constituent morphemes are known, but is not registered as a whole. This is also quite common in Japanese, in which two or more morphemes can be combined to form a compound word.

• Category 4 - A word whose constituent morphemes are partially or totally unknown.

A survey of more than 10,000 keywords indicated that more than 58% of them are unknown words to the expanded EDR dictionary, which is assumed to be the system's lexicon. Table 1 shows the number of unknown keywords in each of the four categories. Because they are largest in number, processing of unknown words of Category 3 is especially important. Analysis of their structures shows that about 70% of them consist of two morphemes.

Inference on the concept of an unknown compound word of Category 3 consist of two stages. The first stage is the syntactic analysis to identify the surface structure of the compound word in terms of five syntactic elements: noun element (N), verb element (V), adjective element (ADJ), adverb element (ADV) and affix element (AFF). The second stage is the semantic analysis to identify the deep structure of the compound word in terms of 23 case frames for each constituent morpheme.

Table 1. Classification of unknown words.

Category	No. of words	
1	25	(0.5%)
2	12	(0.2%)
3	3,951	(81.9%)
4	842	(17.4%)

6. SUMMARY

This paper has presented an overview of an intelligent system for information retrieval over the Internet based on human-machine dialogue. The system has been designed on the basis of spoken dialogue between human and machine, use of key concepts for information retrieval, processing of unknown words, knowledge acquisition, and agent technology. The roles of three agents, respectively responsible for user interface, information retrieval, and information acquisition, have then been discussed. Finally, dialogue management through user and system modeling, implementation of information retrieval through key concepts, and inference on concepts of unknown words based on syntactic and semantic analyses of their structures, have been briefly described as the innovative features of the system.

ACKNOWLEDGMENT

The current work was supported by Japan Society for the Promotion of Science as a Project on 'Research for the Future' (Project No. JSPS-RFTF-96R15201).

REFERENCES

[1] Fujisaki, H. Kameda, H. and Kawai, H. "A system for information retrieval of newspaper articles based on key concepts," *Transactions on Natural Language Processing, Information Processing Society of Japan*, 44-4 (1984).

[2] Kameda, H., Fujisaki, H., Morita, T. and Kurashima, A. "Classification and processing of unknown words," *Transactions of National Convention, Information Processing Society of Japan*, pp. 1195–1196 (1988).

[3] Fujisaki, H., Kameda, H., Ohno, S., Ito, T., Tajima, K. and Abe, A. "An intelligent system for information retrieval over the internet through spoken dialogue," *Proceedings of Eurospeech'97*, vol. 3, pp. 1675-1678 (1997).