AN INTELLIGENT SYSTEM FOR INFORMATION RETRIEVAL OVER THE INTERNET THROUGH SPOKEN DIALOGUE

Hiroya Fujisaki 1, Hiroyuki Kameda 2, Sumio Ohno 1, Takuya Ito 1, Ken Tajima 1, and Kenji Abe 1

1 Department of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda, 278 Japan
2 Tokyo Engineering University, 1404-1 Katakura, Hachiouji, 192 Japan

ABSTRACT

For the purpose of coping with the affluence of information available over the Internet, an efficient, accurate and user-friendly system for information retrieval is mandatory. This paper presents an intelligent system based on the use of spoken dialogue as the main channel for user-system interface, use of key concepts, processing of unknown words, automatic acquisition of various kinds of knowledge for improving the performance, and agent technologies for system realization. Details of functions required for the agents are also described.

1. INTRODUCTION

With the rapid progress of computer technology and world-wide 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 synonymy 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). 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.

Based on these considerations, the present paper first describes the basic principles for advanced information retrieval, then proposes an intelligent system based on these principles, and describes some of the key functions of the system.

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. It is often the case that the user’s intention becomes definite only after he/she gets some knowledge, through trial and error, from the system. If the user and the system can exchange information and knowledge through dialogue, especially through spoken language, it will greatly facilitate the process of formation and expression of intention 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

Here we define a `word' as the one-to-one correspondence of a surface form $T$ (i.e., a transcription/pronunciation) and a concept $C$ (i.e., a meaning). In the case of polysemy, a surface form corresponds to two or more concepts, as shown in Fig. 1(a). In the present paper they are considered to be different words sharing one surface form. In the case of synonymity, on the other hand, a concept corresponds to two or more surface forms, as shown in Fig. 1(b). In this paper as well as in conventional terminology, they are considered to be different words sharing one meaning.

In conventional systems for information retrieval using keywords, the user's intention is represented by keywords, and the search is based solely on transcriptions, but not on concepts. In the case of polysemy, this leads to the occurrence of the so-called `fallouts' (i.e., retrieval of irrelevant items), since the system is not capable of separating the intended keyword, say $T - C_1$ in Fig. 1(a), from unintended keywords such as $T - C_2$ and $T - C_3$. This can be avoided only if the system operates on the concept level, and treats the keyword $T - C_1$ as being different from other words such as $T - C_2$ and $T - C_3$ that are not the intended keywords. In this case, $C_1$ is defined as the `key concept.'

In the case of synonymity, a conventional system will retrieve only those items that contain the keyword $T_1 - C$, where $T_1$ is the transcription given by the user, and will fail to retrieve items containing synonyms such as $T_2 - C$ and $T_3 - C$ shown in Fig. 1(b). This can also be avoided if the system operates on the concept level and considers words such as $T_2 - C$ and $T_3 - C$ as equivalent to the keyword $T_1 - C$ given by the user.

2.3. Processing of Unknown Keywords

Since the user's intention 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 of a keyword, and to resolve ambiguities through dialogue with the user, if necessary. If the lexicon can be accessed also by concepts, it can be used to find synonyms of the keyword given by the user. When the keyword given by the user is `unknown', the system can obtain the underlying key concept through dialogue with the user.

If, however, a keyword attached to a document in the database is unknown to the system, the system has to infer the corresponding concept from its surface form and context. A method for the inference and acquisition of unknown words has already been presented by the authors [2].

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 an intelligent system for information retrieval should be capable of acquiring automatically all kinds of knowledge that are useful for improving its performance.

2.5. Use of Agent Technology

The entire process of information retrieval can be broadly divided into two aspects: interface with the user and interface with the database. For the ease of system design as well as for higher efficiency of processing, these two aspects ought to be handled by separate agents, viz., by an `user-interface' agent and a `search' agent. If vast differences exist in the characteristics of individual users or individual databases, they may further be dealt with individually by separate agents. In the following example, we show a system with one user-interface agent and a group of search agents dealing with different groups of databases.

3. PROPOSAL OF AN INTELLIGENT SYSTEM FOR INFORMATION RETRIEVAL [3,4,5]

Based on the principles described above, a novel, intelligent system for information retrieval over the Internet is proposed. We will first give an overview of the system, and then describe the functions required for the agents that perform the essential tasks.
3.1. Overview

Figure 2 shows the basic structure of the proposed system. The main components of the system are the spoken language interface supplemented by a display and a keyboard, a user interface agent (Agent 1), and a number of search agents (Agent 2\(^1\), Agent 2\(^2\), etc.). Each of the search agents (Agent 2\(^2\)’s) has its own specialties both in the topical area and in the type of databases that it covers.

3.2. Functions of Agent 1

3.2.1. Identification of User’s Intention and Construction of Search Formula

In order to retrieve precisely the information intended by the user, Agent 1 has to identify the user’s intention on the basis of the keywords given by the user. It often helps the user to clarify and state his intention through dialogue, mainly by spoken language supplemented by a display and a keyboard. The system always refers to the keyword/key concept lexicon, and ambiguities due to polysemy of a keyword shown by the user is dissolved through dialogue with the user. If, on the other hand, a keyword given by the user has synonyms, the system automatically adopts them also as keywords, thus ensuring completeness of search at the concept level. Through these processes Agent 1 also infers the weights which a user assigns to each of the key concepts. These weights are used later for obtaining a quantitative measure for the relevance of the information retrieved.

Once the user’s intention becomes clear as a set of key concepts, Agent 1 constructs the search formula in terms of key concepts but translates it into keywords.

3.2.2. Maintenance of Keyword/Key Concept Lexicon

A keyword/key concept lexicon which can be accessed both by transcriptions (i.e., keywords) and by meanings (i.e., key concepts) plays a crucial role in the proposed system. Since it is impossible to prepare a complete lexicon in advance, the lexicon has to be expanded while being used. Starting from an initial lexicon of a limited size and coverage, the lexicon is expanded through automatic knowledge acquisition. Agent 1 is in charge of maintenance and expansion of the lexicon.

3.2.3. Processing of Unknown Keywords Given by User

Unknown keywords given by the user are detected and clarified by Agent 1 through dialogue with the user. If the concept of an unknown keyword is found to be identical to the one already in the lexicon, the surface form is registered in the lexicon as one of the synonyms of an already registered concept. If it is not found in the lexicon, it is registered as a new keyword/key concept.

3.2.4. Acquisition of User Model

Exactly speaking, each user has his/her own idiolect as well as own interests and own way of using the system. In order to have highly accurate and efficient performance, therefore, Agent 1 has a model of a user, and automatically acquires the model for each individual user. In particular, it acquires specific patterns of search suitable for each user.

3.2.5. Request for Search by Agent 2

Based on a user’s request, Agent 1 selects appropriate search agents, i.e., Agent 2’s, for the specific search task, and requests to do the actual search.
3.2.6. Evaluation of Results Obtained by Agent 2

Using the weights mentioned in 3.2.1., Agent 1 evaluates the relevance of retrieved items, and those with higher relevance scores are presented first to the user. If they do not satisfy the user, those of lower relevance scores are shown. The final choice by the user is utilized to modify the weights and the model of the user for the purpose of improving performance. If the initial search does not satisfy the user, Agent 1 tries to obtain more key concepts through dialogue with the user. This process is repeated until the user is satisfied or the search is exhausted.

3.3. Functions of Agent 2

3.3.1. Information Retrieval from Database

Using the meta-knowledge on available databases, Agent 2 selects the databases to be accessed, and retrieves the requested information. The access is made through a web browser and relevant servers. Since the final stage of search is done through keywords, the search formula is translated into equivalent keywords, but the formula is far more precise and comprehensive than that would be obtained without going through key concepts.

3.3.2. Dealing with Ambiguities

Since the ambiguity due to polysemy of a keyword given in a data cannot be resolved through dialogue with the user, Agent 2 infers the most probable meaning on the basis of its context, i.e., in relation to other keywords/key concepts given in the data, and evaluates the relevance of the data.

3.3.3. Processing of Unknown Keywords in a Data

If a keyword given in a data is unknown, Agent 2 makes inference on its meaning by a procedure already developed [2]. It is based primarily on the structured analysis of the surface form, and the concept is inferred from those of the constituent morphemes.

3.3.4. Acquisition of Meta-Knowledge on the Databases

It is the task of Agent 2 to construct and maintain a meta-knowledge on the databases which it covers. Since the contents of many databases are frequently updated, Agent 2 conducts a regular check on each of the databases, and updates the meta-knowledge. Furthermore, it keeps record of the number of hits of each database as a meta-knowledge on its usefulness.

4. SUMMARY AND CONCLUSION

For the purpose of realizing a user-friendly and efficient system for information retrieval over the Internet, we have proposed a new system based on spoken dialogue as the main channel for user-system interface, use of key concepts, processing of unknown words, automatic acquisition of various kinds of knowledge for improving the performance, and agent technologies for system realization. Due to space limitations, many of the details have to be reported elsewhere. Work is under way to collect spoken dialogue data in various types of information retrieval problems, and to construct an efficient procedure for inferring the user’s intention mainly through spoken dialogue, supplemented occasionally by a display and a keyboard.

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