INTELLIGENT RETRIEVAL
Open Journal

Review

Design and Principal Information Recovery System using the Internet and a Multimodal Dialogue Interface

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

[1]*Science University of Tokyo, Noda, 278-8510, Japan*

[2]*Department of Computer Engineering and Computer Science University of Missouri-Columbia Columbia, MO 65211, USA*

Corresponding author
Hiroya Fujisaki
Science University of Tokyo, Noda, 278-8510, Japan

Article information
Received: January 13th, 2024; **Revised:** February 21st, 2024; **Accepted:** March 1st, 2024; **Published:** March 28th, 2024

Cite this article
Fujisaki H, Kameda H, Ohno S, Abe K, Iijima M, Suzuki M, Taketa K. Design and principal information recovery system using the internet and a multimodal dialogue interface. 2024; 2(1). doi: https://doi.org/10.70705/ppp.ir.2024.v02.i01.pp21-23

ABSTRACT

Everyday life in the information society of the future century will revolve on information retrieval via the Internet, and spoken language will play a crucial role in human-machine dialogue. This article provides a high-level description of an intelligent information retrieval system that makes use of agent technology, processes unfamiliar words, uses important ideas, and has conversations. In this model, three agents are defined to handle the user interface, data retrieval, and data gathering, in that order. The novel aspects of the system are thereafter summarized as follows: dialogue management by use of user and system modeling; information retrieval by means of key concepts; and inference on the meaning of unfamiliar terms by means of structural analyses of their syntax and semantics.

INTRODUCTION

A tremendous quantity of data is currently created, shared, and kept at various locations throughout the globe as a result of the fast advancement of computer technology and the global expansion of information networks. But without a way to quickly retrieve the information that is authentically important and relevant, such a deluge of data is either worthless or even detrimental. This is one area where traditional methods of information retrieval fall woefully short of expectations; they often gather unnecessary data while ignoring crucial details. Reasons for these occurrences include the user's inability to clearly define and communicate their goal and the system's inability to accurately infer their intention. An improvement over these problems would be to allow the user and the machine to communicate verbally.

ADue to the polysemy and synonymy of terms, keyword search tends to be inefficient and inaccurate, but it is suitable for accessing information from databases that are not necessarily structured by a common premise. Using "key concepts" [1] instead of keywords might help you overcome these challenges. When this occurs, data recovery is depending, not on the literal meanings of keywords but on the user-intentioned semantic content. If all the keywords are "known" to the system—that is, previously recorded in the system's lexicon—then spoken conversation may overcome difficulties stemming from polysemy of keywords. Having every potential keyword recorded in the lexicon is obviously not feasible in real-world information retrieval scenarios due to the high frequency of newly-formed words or compound words formed by mixing known morphemes. Therefore, it is essential that the system can deduce the meaning of novel, "unknown" (i.e., not yet included to the lexicon) terms [2]. What this means is that the system has to be able to learn. If the goal is to create an intelligent system that can enhance its performance on its own, this is also essential. An earlier research [3] laid down the groundwork for such an intelligent information retrieval system. This document gives a high-level description of the prototype system and its design, including some of its key characteristics.

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 in-

tention mainly through spoken dialogue, supplemented by a graphic display. Its detailed structure and function will be described in connection with the dialogue management.

Intelligent Information Retrieval System
The Internet
Databases

(2) Information Retrieval Agent (IRA)
(3) The user makes a request, and the information retrieval agent goes out and gets the desired information. After UIA collects user-inputted keywords, IRA builds a search formula to cover the user's intended core ideas as thoroughly as possible. A keyword/key-concept lexicon is used to do this. If we utilize all the synonyms of the user-supplied keywords, we can reduce the likelihood that we won't be able to get relevant data, and if we use collocation information, we can eliminate irrelevant data. If the amount of things retrieved exceeds a certain threshold, the relevance score of each retrieved item is used to decide the order in which they are presented to the user. In Section 4, we will get into the specifics of key-concept based retrieval. Agent for Acquiring Information (IAA) Link data pertaining to the address, pertinent key ideas, and knowledge acquired from numerous online sources are stored by the information acquisition agent. as well as the terms that characterize each database. In addition, it finds and infers the ideas of unknown words discovered in databases, adding them as new keywords or key-concepts to the lexicon, as part of its unknown word processing.

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

User Interface Agent

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

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.

RELATIVE THRESHOLD SCORE, □

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:

SUMMARY

This work has provided a high-level outline of a human-machine dialogue-based intelligent system for online information retrieval. Conceptually, the system is based on human-computer conversation.

5. PROCESSING OF UNKNOWN WORDS

One of IAA's primary functions is, as previously shown, the processing of unfamiliar words. There are four groups of unfamiliar terms:
Words in Category 1 have just their transcriptions recorded in the lexicon. This happens when the lexicon does not include all of the keywords that have the same surface shape, even if there may be more than one.
Class 2: A word whose lexical record consists only of its notion. The absence of rigid orthographic norms in Japanese makes it possible

for there to be more than one transcription, and lexicons often do not include all of the possible transcription variants.

A word falls into Category 3 if its individual morphemes are known but the term itself is not. Combining two or more morphemes to create a compound word is also extremely prevalent in Japanese.

The fourth category includes words where the exact composition of the morphemes is uncertain.

An analysis of over 10,000 keywords revealed that the extended EDR dictionary—presumed to serve as the system's lexicon—is unfamiliar with over 58% of them. The amount of unknown terms in each of the four categories is shown in Table 1. It is particularly crucial to digest unfamiliar terms of Category 3 since they are the most numerous. According to structural analysis, two morphemes make up around 70% of them.

There are two steps to drawing an inference about the meaning of a Category 3 compound term whose notion is unknown. Finding the surface structure of using syntactic analysis is the first step. device, application of fundamental ideas for data retrieval, handling of unfamiliar terms, information gathering, and agent technology. Next, we covered the responsibilities of three agents: one for the user interface, one for retrieving information, and one for acquiring information. The novel aspects of the system have been briefly discussed, including its ability to manage discourse via user and system modeling, retrieve information using key ideas, and infer the meaning of new words using syntactic and semantic analyses of their structures.

REFERENCES

[1. Fujisaki, H. Paper presented at the 1984 Transactions on Natural Language Processing by the Information Processing Society of Japan, titled "A system for information retrieval of newspaper articles based on key concepts," was written by Kameda and Kawai.

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

In a study conducted by Fujisaki, Kameda, Ohno, Ito, and Tajima

3. [3], In the Proceedings of Eurospeech'97, volume 3, pages 1675-1678, K. and Abe, A. presented "An intelligent system for information retrieval over the internet through spoken dialogue" (1997).