Natural Language Access to Frequently Requested Information

Michael P. Jones, 1 Thomas K. Landauer, 2 and James H. Martin 1

1 Institute of Cognitive Science and Department of Computer Science
2 Institute of Cognitive Science and Department of Psychology
University of Colorado, Boulder, CO 80309

The quantity of on-line information is growing and changing at a startling rate. The number of users seeking this information is also growing quickly. Many users are computer novices who are likely to encounter great difficulty locating the information they seek within a complex information space. This difficulty arises from two principal sources: complex, unintuitive user interfaces and the ill-structured nature of most information spaces. Our current research addresses these difficulties through the use of spoken and written natural language interfaces coupled with automated indexing techniques designed to extract semantic sense from ill-structured data.

Our interface work is centered around the belief that users should be able to ask for information in the most natural way possible: asking for it in their own words. We have collected a corpus of information requests from various environments. Some of the corpus resulted from Wizard of Oz experiments in which users are given several tasks to perform. Two data sets were collected with a tape recorder at the reference desk of a public library and at the help desk in a campus computer lab. Our analysis of this corpus shows that the variability and complexity of the language used makes simple keyword matching an insufficient approach to the task of information indexing and retrieval[1]. On the other hand, our analysis of this corpus also shows a marked bias on the part of the speakers towards simple, frequently asked, and easily answered questions. These observations led us to examine existing knowledge sources that might contain the kind of information needed to answer the type of questions prevalent in our corpus. The primary body of knowledge we are currently using is the collection of Frequently Asked Questions files (or FAQs) associated with the Usenet newsgroups. Using FAQ files as a primary information source is warranted since their existence comes about in an effort to answer the most common requests for information in a given domain. Consequently, we expect this source of information to contain the information sought for a large percentage of users. However, the FAQs are a very dynamic information source. They are augmented and modified continually; therefore, our system demands an indexing scheme which requires little human knowledge engineering.

We are using Latent Semantic Indexing (LSI)[2] to automate the information indexing process. LSI was designed to accept natural language queries and to overcome the keyword match problem. It "deduces" semantic word similarities from the indexed information such that relevant documents are retrieved even if a query uses words not found in those documents. Our initial efforts have provided evidence for this ability. For example, the question "How do I select text?" posed to an LSI index of the Emacs FAQ returns the correct document even though the word "highlight" rather than "select" is used in the document. LSI is able to generalize or "learn" sufficiently well the semantic relationship between highlighting and selecting text such that the appropriate response from the FAQ file is returned. LSI does this automatically from the information source without the need for additional human knowledge engineering.

In our ongoing efforts, we are exploring how we can enhance the performance of LSI using various syntactic and semantic information from both the information source and the user's query. Additionally, we are looking at methods to increase our confidence that the information presented to the user addresses the question asked. Finally, an important part of the user interface is the presentation of information to the user and the mechanism for the user to provide relevance feedback. We are experimenting with these facets of the interface.

References