COLING 82, J. Horecký (ed.)
North-Holland Publishing Company
© Academia, 1982

NATURAL-LANGUAGE-ACCESS SYSTEMS AND THE ORGANIZATION AND USE OF INFORMATION

Donald E. Walker

Artificial Intelligence Center
SRI International
Menlo Park, California 94025
U.S.A.

This paper describes a program of research whose objectives are to (1) develop systems that provide users with access to both data and text files through natural language dialogues; (2) study how people actually use the information to test hypotheses and solve problems; (3) modify the system designs on the basis of the results of the studies so that the systems more effectively support such uses and increasingly come to model the behavior of the users. Two of the systems are in the medical domain: the first provides physicians with formatted information derived from patient medical records; the second responds to requests by eliciting relevant passages from a medical monograph. The third system is a more general information retrieval facility that will support interactions among system users and enable their successive experiences to be accumulated within the system database.

OVERALL RESEARCH OBJECTIVES

This paper describes a program of research intended to clarify how people, working as scientists and professionals on problems in their respective areas of expertise, actually use information to solve those problems.[1] The strategy we are pursuing entails the construction of computer-based systems in which the users can access different kinds of information through dialogue interactions in ordinary conversational language (see Walker, 1981; 1982). To carry out this program requires an extension of computational capabilities for processing and understanding natural-language requests, for representing the information content of data and text files, and for relating the analyzed requests to the representations. Studying how people use these systems to formulate and test hypotheses and to make decisions can serve as a guide to system modification, leading not only to improvements in performance but also to effective techniques for organizing knowledge about the problem domain and incorporating it within the system. To the extent that this iterative process is successful, the systems should increasingly come to model the behavior of the users.

The systems we are developing fall somewhere in the middle of a continuum of facilities that are relevant to the organization and use of information. The two ends of the continuum are the information retrieval systems of information science and the knowledge-based expert systems of artificial intelligence. Considered in idealized form, both ends represent static states. The information retrieval systems provide access to factual data, the raw materials of a technical domain. The expert systems embody digested knowledge that is consensually validated as germane to that area of inquiry. In contrast, the "systems for experts" that we are creating constitute a milieu in which specialists in a particular field can

----

[1] The preparation of this paper was supported in part by grants from the National Cancer Institute (No. 1 R01 CA26655), the National Library of Medicine (No. 1 R01 LM03611), and SRI International Internal Research and Development funds.

explore the range of information available in order to test hypotheses or develop new insights, the results of which will eventually become part of that field's knowledge base. In that respect, our systems reflect the dynamic instabilities and uncertainties of the continuing search for new ideas and new answers--the core of the knowledge-synthesis-and-interpretation process.

It is important that systems for experts actually be used by people who are specialists in an area. The capabilities we are developing are predicated on the expectation that only the person who actually needs the information is able to evaluate its adequacy. Two considerations are worth noting here. First, we recognize that these needs may not be well formulated at the beginning. A person may recognize that he or she needs to know something about an area but not be able to specify it precisely.[2] It is for this reason that our systems provide for dialogue interactions; they make possible, on the basis of an assessment of the retrieved information, a progressive refinement of the search specification or the hypothesis to be tested. Only if the user is a person able to evaluate the adequacy of the results can such a dialogue be sustained.

Second, because of these differences in needs, the materials in the database will be interpreted in different ways. That is, the information is not intrinsic in the data; for example, a document's relevance for a user does not have to bear any necessary relation to the relevance its contents had for the person who generated it. Actually, scientists and professionals typically have complex problems for which there are no ready-made answers. Both the problems and the information necessary for their solution are dynamic. To be useful, a system must support the reorganization and reinterpretation of its "facts." It must also allow the results of these actions to be added to the database because of their value for subsequent users. As noted above, the incorporation of consensually validated information as knowledge is a key element in expert systems. The systems for experts that we are discussing here certainly must contribute to this process. However, it is essential to recognize that, even in expert systems, the underlying knowledge structures are subject to change.

The following sections describe our current efforts, which explore some of the capabilities required to achieve these objectives. The first, Providing Natural-Language Access to Data, focuses on the utility of a natural-language interface for retrieving formatted information from a database by a person familiar with the subject matter, but not the structure of the file itself. The second, Representing the Information Content of Texts, concentrates on the development of procedures for analyzing propositional content so that natural-language requests can effect selective retrieval of relevant passages. The third, Facilitating Generalized Access to Information, is directed toward establishing a more general system structure within which a group of people working in a related area can annotate and evaluate information sources in relation to their needs, store the commentaries so that they can be accessed by others, and communicate the results of their research.

PROVIDING NATURAL-LANGUAGE ACCESS TO DATA

The MEDINQUIRY project (being conducted cooperatively with research groups at the University of California at San Francisco, the University of Pennsylvania, and the National Library of Medicine) is concerned with providing natural-language access to clinical databases.[3] The MEDINQUIRY system (Epstein and Walker, 1978; Epstein,

----

[2] The characterization by Belkin and his colleagues of an information need as an "anomalous state of knowledge" reflects this insight (Belkin, 1978; 1980).

[3] This work is supported by a grant from the National Cancer Institute. I am indebted to Scott Blois and Richard Sagebiel and their colleagues at UCSF, Wallace Clark and his colleagues at the University of Pennsylvania, and Martin Epstein at the NLM, in addition to Robert Amsler and Bil Lewis at SRI for their contributions to this project.

1980) is designed to support both clinical research and patient management by physicians studying the prognosis for chronic diseases.[4] The project database, currently being established at SRI, will eventually store information on over 150 attributes from approximately 1500 records of patients with malignant melanoma, a skin cancer with an unusually high mortality rate.

MEDINQUIRY enables the physician to enter requests in English that retrieve specified data for particular patients or for groups of patients who share certain characteristics, that induce a variety of calculations, that enable browsing through the database, that support identifying and exploring relationships among patient attributes, and that relate information in the database to prognosis and outcome. The system consists of a natural language processor based on LIFER (Hendrix, 1977), a database access module, the database itself, which contains information from patient records, and a response generator. When the user types in a request, the natural-language processor attempts to analyze it using general knowledge about English and knowledge specific to melanoma, that are contained in a set of grammatical rules defining the language accepted by the system. The grammar was developed on the basis of a comprehensive review of the literature on melanoma, an analysis of the database, and discussions with melanoma experts. For requests that are analyzed successfully, the user is presented with a paraphrase to show how the system has interpreted it. For requests that cannot be analyzed, an attempt is made to explain the difficulty. Once the request has been analyzed grammatically, a set of functions is applied to create a logical statement of its content in a formal query language. That query is applied to the database, and the requested data are retrieved and returned to the natural-language processor. There, they are reorganized in a form that corresponds to the language of the original request, and the result is presented to the user.

MEDINQUIRY supports dialogue interactions; the user can follow a line of inquiry to test a particular hypothesis by entering a series of requests that are sequentially interdependent. Phrases can be used as well as complete sentences, the meaning of a given phrase being established on the basis of the analysis of the prior request. The user can actually define new constructs at word, phrase, and sentence levels that generalize to allow interpreting a set of related constructions. For every user session MEDINQUIRY automatically records a transcript that provides a complete record of requests entered and responses made. This facility proved to be extremely helpful for evaluating problems encountered during system development; we plan to use it extensively in our studies of hypothesis formation and testing by physicians.

The primary medical objectives of the project include investigating the natural history of melanoma, studying differences between the patient populations in California and Pennsylvania, and developing individual risk-prediction methods. These goals entail the acquisition and management of large volumes of data. To study a particular aspect of the disease--and exclude the effects of others--it is necessary to stratify and form arbitrary classes of data elements and then examine their interactions. The development and testing of hypotheses entail several different levels of analysis. Material from the medical records of patients constitutes the basic data. Included are the primary clinical observations and the results of laboratory tests and histopathological studies. The physician needs to identify and aggregate the critical variables and to determine how they relate to high level concepts like "stage of disease" and "high risk primary." The judgments of a particular physician, so labeled, can be entered into the database so that others can assess their utility.

MEDINQUIRY is operational, and we are in the process of entering data from patient records. When a sufficient amount of material is available, the physicians on the project will begin to access the database systematically. Then, we will begin the

[4] MEDINQUIRY, written in INTERLISP, is installed on DEC 2060 computers at SRI and at the National Library of Medicine in Bethesda, Maryland.

real process of evaluation, using those observations to guide refinements in the system design.

REPRESENTING THE INFORMATION CONTENT OF TEXTS

Increasing amounts of medical information in text form are becoming available for computer-based search and retrieval. However, the existing keyword-based procedures for locating a particular passage in a document are both awkward to use and grossly insensitive. To enable more efficient access by physicians and other health professionals, we are developing capabilities that allow a person to search a textual data base more effectively through natural-language dialogues.[5]

The initial database for our research is a computerized monograph containing current knowledge about hepatitis, the Hepatitis Knowledge Base being developed at the National Library of Medicine (Bernstein et al., 1980). We are encoding (primarily by hand but with computer assistance) a "text structure" for the document that consists of logical representations summarizing the information content of individual passages together with a specification of the hierarchical relationships among the passages. The logical representations are expressed in a formal language in which canonical predicates are used.

In the text access system we are developing, a request is analyzed in two major phases (1) A grammatical analysis determines the structure of the sentence, which is then translated into its logical form. For this purpose, we are using DIALOGIC, a natural-language-understanding system developed at SRI.[6] (2) Inferences drawn from a knowledge store are used to solve discourse problems posed by the request and to translate it into the canonical predicates in which the text structure is expressed. The result is then matched against the text structure to identify relevant passages for retrieval.[7]

The knowledge store is of particular interest, because it allows the text access system to deal with requests that go beyond its canonical vocabulary. It is not enough to represent just the vocabulary in the monograph itself, for a physician cannot be expected to be restricted to that set of terms. The user will generally be approaching the document from the broader point of view of medicine as a whole and without knowing precisely what is included in the text. Consequently, we need to incorporate knowledge about the larger domain within which a request is being formulated. Here, too, it is not sufficient simply to relate the medical knowledge of the user to the actual contents of the monograph, because requests will often concern aspects of the disease that are not mentioned in the particular text, but would be in a more comprehensive "possible" hepatitis knowledge base. For example, requests can concern aspects of hepatitis that are not yet known or are no longer believed, and have therefore been deleted from subsequent versions of the HKB. Requests may use a vocabulary that is not in the document, and may mention events, specific interactions, and exceptions for which the existing monograph has only indirect information. The existing texts on hepatitis are really only one part of a larger set of texts that did or could exist; the requester is, in effect, addressing a request to this larger set rather than to the actual document.

----

[5] This work is supported by a grant from the National Library of Medicine. It is reported in Walker and Hobbs (1981) and in Hobbs, Walker, and Amsler (1982), a paper that is included in the proceedings of this conference.

[6] DIALOGIC is described in Grosz et al. (1982), which is included in the proceedings of this conference; the grammar for the system is discussed in Robinson (1982).

[7] Hobbs (1980) provides a more detailed description of the DIANA system, which is used to perform the inferencing.

Our work on representing the information content of texts is still at an early stage of development. However, the problems we are addressing are critical for the development of capabilities that can support people who organize and use information.

FACILITATING GENERALIZED ACCESS TO INFORMATION

Polytext is a new system concept for text retrieval being developed in cooperation with Hans Karlgren and the Kval Institute for Information Science in Stockholm, Sweden (Karlgren and Walker, 1980).[8] Responding to inadequacies in current information retrieval technology, Polytext provides the following capabilities: dialogue facilities aid the user in formulating and refining requests and in evaluating the relevance of both intermediate and final results; the successive experiences users have with the data are accumulated in the database and available to others; alternative algorithms and strategies (human as well as computer) for processing texts and representing the information they contain are accommodated--and the metatextual commentary they provide is explicitly identified as to source.

The central feature of the system is the notion of "messaging": all data elements in the system are considered to be messages--as are the alternative representations of the content of each data element, the requests addressed to the system, the evaluations of the relevance of each request to the data retrieved, and other communications among users. The structural features of each such message--in particular, a topic/comment relation patterned after contemporary linguistic usage (Kiefer, 1980)--provide the basis for linking it appropriately to other messages. When a user's request is processed, pointers are provided both to relevant items in the primary source text and to the results of previous requests that appear related. It is possible to examine the rationale for the relationships adduced and to identify their origin.

After establishing the design features for Polytext, we recognized the need for a project of this magnitude to proceed by well-defined steps. Accordingly, the next phase of our research was the production of a demonstration model to verify some of the basic concepts (Loef 1980). For this initial work, we selected a short legal document containing rules for arbitrating disputes that arise in connection with contracts. We developed three ways of providing access to the text, using, respectively, (1) index terms, (2) the hierarchical structure of the text, and (3) an analysis of the predicate-argument, or propositional, structure of the text to derive a more detailed model of the information it contains. For each approach, we provided the appropriate interface to a LIFER grammar, so that it was actually possible to enter English queries and to retrieve the appropriate passage as a response.

We are attempting to keep the basic Polytext software as simple as possible. Therefore, intelligent and short-lived modules are kept outside as programs using the system rather than as parts of it. Thus text analyzers (machine or human) may take one message at a time, interpret it, and report the result as a new message, which has the analyzed message as its topic and the recoding in some meta-language as its comment. The lexicon for the system would itself be stored in message form, and the programs could use other messages for information in the course of their analyses.

In the context of the research program, Polytext constitutes an environment in which the range of issues associated with the organization and use of information can begin to be evaluated. It will be essential to incorporate sophisticated capabilities for dialogue interaction and content representation--of the kind being developed in the other two projects--but we have at least begun to establish a flexible system structure that can accommodate many users and accumulate their collective experiences.

[8] This work is supported by SRI Independent Research and Development funds and by several Swedish sources. I am deeply indebted to Hans Karlgren for his inspiration.

REFERENCES

Belkin NJ. 1978. "Information Concepts for Information Science." *Journal of Documentation* 34:55-85.

Belkin NJ. 1980. "Anomalous States of Knowledge as a Basis for Information Retrieval." *Canadian Journal of Information Science* 5:133-143.

Bernstein LM; Siegel ER; Ford WH. 1980. "The Hepatitis Knowledge Base: A Prototype Information Transfer System." *Annals of Internal Medicine* 93:165-222.

Epstein MN. 1980. Natural Language Access to Clinical Data Bases. Ph.D. Dissertation, Medical Information Science, University of California, San Francisco.

Epstein MN; Walker DE. 1978. "Natural Language Access to a Melanoma Data Base." *Proceedings of The Second Annual Symposium on Computer Applications in Medical Care* pp 320-325. New York: IEEE.

Grosz B; Haas N; Hobbs J; Martin P; Moore R; Robinson J; Rosenschein S. 1982. "DIALOGIC, A Core-Natural-Language Processing System." *COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics*, Prague, Czechoslovakia.

Hendrix GG. 1977. "The LIFER Manual: A Guide to Building Practical Natural Language Interfaces." Technical Note 138, Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California (February 1977).

Hobbs JR. 1980. "Selective Inferencing." *Proceedings of the Canadian Society for Computational Studies in Intelligence* pp 101-114. Victoria, B.C.

Hobbs JR; Walker DE; Amsler RA. 1982. "Natural Language Access to Structured Text. *COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics*, Prague, Czechoslovakia.

Karlgren H; Walker DE. 1980. "The POLYTEXT System - A New Design for a Text Retrieval System." To be published in the Proceedings of a Conference on Questions and Answers held in Visegrad, Hungary (4-6 May 1980).

Kiefer F. 1980. "Topic-Comment Structure of Texts (and Its Contribution to the Automatic Processing of Texts." *COLING 80: Proceedings of the 8th International Conference on Computational Linguistics* pp 240-241. Tokyo, Japan.

Loef S. 1980. "The POLYTEXT/ARBIT Demonstration System." FOA Report C40121-M7, Swedish National Defence Research Institute, Umea, Sweden (September 1980).

Robinson JJ. 1982. "DIAGRAM: A Grammar for Dialogues." *Communications of the ACM* 25:27-47.

Walker DE. 1981. "The Organization and Use of Information: Contributions of Information Science, Computational Linguistics and Artificial Intelligence." *Journal of the American Society for Information Science* 32:347-363.

Walker DE. 1982. "Computational Strategies for Analyzing the Organization and Use of Information." In *Knowledge Structure and Use: Perspectives on Synthesis and Interpretation*. Edited by S Ward and L Reed. National Institute of Education, Washington, D.C., in cooperation with CEMREL, Inc., St. Louis, Missouri (in press).

Walker DE; Hobbs JR. 1981. "Natural Language Access to Medical Text." *Proceedings of the Fifth Annual Symposium on Computer Applications in Medical Care* pp 269-273. New York: IEEE.