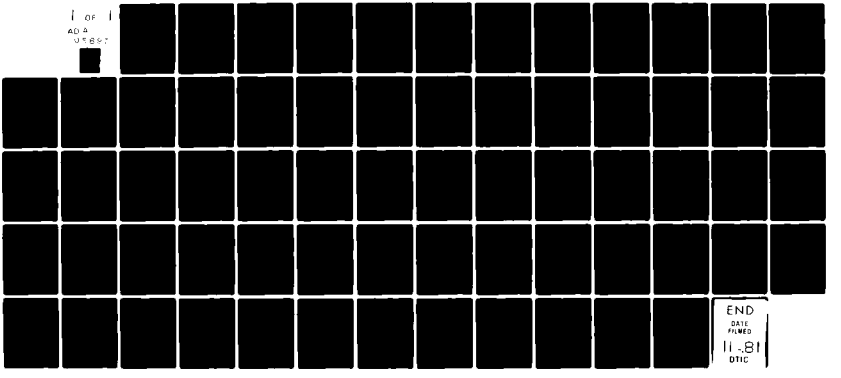
AD-A105 897 CENTER FOR POLICY RESEARCH INC BETHESDA MD F/G 5/2
A WORKSHOP ON QUALITATIVE INFORMATION RETRIEVAL, NOVEMBER 18-20--ETC(U)
SEP 81 S B WEINER N00014-80-C-0105
UNCLASSIFIED NL



AD A105897

CENTER FOR POLICY RESEARCH, INC.
NEW YORK, N.Y. BETHESDA, MD.
5272 RIVER ROAD, BETHESDA, MD. 20016
(301) 657-2277

September 28, 1981

A WORKSHOP ON QUALITATIVE INFORMATION RETRIEVAL

November 18-20, 1980

Contract N00014-80-C-0105
NR-049-473

DTIC FILE COPY

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

81 9 30 002

TABLE OF CONTENTS

Accession For	
NTIS GRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By Per Ltr. on File	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

INTRODUCTION

Work in the Office of Administration of the Carter White House on the development of a non-intermediated information retrieval system led to consideration of the problem of information retrieval when controlled vocabulary fails. Discussion of that problem led to the development of a typology of such failures and to an examination of methods of resolving the problems. While a detailed discussion of the typology and the methods of resolution is not appropriate at this point, that typology was presented as the point of departure for deliberation by the Workshop on Qualitative Information Retrieval, the results of which this report summarizes. It is, therefore, appropriate to review briefly that typology.

Retrieval problems appear to be of three types distinguished by two variables: 1) who defines and limits acceptable search terms and 2) recall and precision of retrieval. The three basic types are:

TYPE I: The system defines and limits search terms via the index/thesaurus. User needs are congruent to system capabilities using those terms and such adjunct capabilities as free text searching. Recall and precision are high and user satisfaction is concomitantly high. An example of a Type I problem is found when an on-line searcher seeks information on UNDERDEVELOPED NATIONS. By definition, the system concept of an underdeveloped nation is the same as the users. A retrieval problem occurs when the user

terminology does not conform to the controlled vocabulary, or when user terminology does not include all of the synonyms which might be encountered in a full text search. Thus if the user searches on UNDERDEVELOPED NATIONS while the controlled vocabulary prefers DEVELOPING NATIONS, a retrieval problem is encountered. Or, if the user performs a full text search on UNDERDEVELOPED NATIONS or DEVELOPING NATIONS but neglects such synonyms as EMERGING NATIONS, EXPECTANT NATIONS, or THIRD WORLD NATIONS, again, a retrieval problem develops.

Type I problems are trivial to the extent that adequate techniques exist to overcome them. Indeed, such techniques as automatic synonym control on on-line thesaurus assistance place little or no additional burden on the user. Other search features exist which also help resolve the retrieval problem and assist in obtaining high recall and precision rates.

TYPE II: The user defines the terms. If those terms can be operationally defined in such a way that the stored data can be manipulated to measure each record against that definition, recall and precision can be high. If the definition cannot be operationalized in such a way, recall and precision will be low.

The Type II problems that are most amenable to solution (i.e., most likely to yield satisfactory levels of recall and precision) are those in which the operational definition takes the form of quantifiable factors. Thus, a user of the system might want QUALITY studies on a particular drug. He can define "quality" in terms of the number of subjects, use of a control

group, level of significance reported in results, or the dosage of the drug used in the study. Assuming all of the data are available in each record and that the field in which each is stored can be searched, studies meeting the criteria established can be retrieved. Those criteria can be changed to meet the demands of each user. Recall and precision will be high.

On the other hand, a research request for a search in the psychology literature on the experience one has that he has experienced an event before would be very difficult to perform if the seracher did not know the term DE JA VU. The operational definition is not quantifiable, and any qualitative definition would be virtually impossible to search.

For the sake of completeness, it should be noted that operational definitions that can be presented in nominal terms are also Type II problems. For example, a measure of quality might be the author or the place of publication. Assuming such nominal information is in the record and the files in which they are located can be searched, recall and precision can be obtained at acceptably high levels.

TYPE III: The final retrieval problem is one in which neither the system nor the user has defined the search term. Basically, a Type III problem is one in which the user asks the system to define a term after being given an example. Or, the system might be asked to generalize from a particular. An example might be asking the system to find articles giving examples

of Murphy's Law in action in an academic setting after being primed with a description of the construction of the Rayburn House Office Building.

Placing all instances in which controlled vocabulary fails into the Procrustean bed of "retrieval problem" led to a situation in which attempting to find methods of resolving those problems was difficult because there was no clear starting point. However valid the typology might or might not be, it proved useful in aiding the start of investigations into methods of improving recall and precision when controlled vocabulary yielded unacceptably low results.

Many such methods exist; some have been mentioned above and many others should come to mind, from Boolean searches to truncation searches, from string and proximity searching to weighing of terms. In considering all of this in developing a new information retrieval system, a number of questions arise. If all features and search capabilities cannot be built into one system, how does one compare systems with different capabilities? Is it possible to develop a way to describe and measure systems' retrieval capabilities in such a way that the descriptions are uniform and, thus, comparable? Since most of the systems today are based on hardware that is several generations old, would a new system, developed for new hardware, be significantly different enough to obviate the need for many of today's search capabilities? In the past, for example, the storage capacity of many systems was limited and the expense of searching was not practical.

It is now. What difference does that make?

A number of such questions arose. Investigation and discussion both within and without the Executive Office of the President suggested that answers, if they existed could best be expected from a review of the work being done in a variety of related fields--information science, computer technology and programming, philosophy, psychology, cybernetics, linguistics, artificial intelligence, and others. The Workshop on Qualitative Information Retrieval was proposed to bring together experts in those fields and to begin to answer those questions.

The research done in the Office of Administration, Executive Office of the President was suggested by Amitai Etzioni, then Senior Consultant to the President's Special Assistant for Information Management, Richard M. Harden. Both of these men provided an environment in which such an investigation could be carried out and encouraged other members of the Executive Department to participate. In large part, because of the recognition by these two individuals of the importance of the problem, it was possible to develop a proposal which was attractive to two other individuals and two organizations that made it possible to hold the Workshop.

Professor Martha E. Williams and Dr. Donald Hillman of Lehigh University served as Co-chairmen of the Workshop. Their work during the Workshop was invaluable but was only a small fraction of their assistance. Professors Williams and Hillman provided guidance in focusing the purpose of the Workshop, suggested

participants, advised on format, and identified topics for presentations. In addition, they served as co-principal investigators for the proposal submitted to the National Science Foundation to fund this Workshop.

The Office of Naval Reserach, through Contract #N00014-80-C-0105, and the National Science Foundation, by Grant IST-8018904, provided funding to hold the Workshop on Qualitative Information Retrieval.

As the Workshop Organizer and Principal Investigator, I owe a debt to Martha E. Williams, Donald Hillman, ONR and NSF--both to the institutions and grant officers who provided so much assistance. And, I owe a great deal to the participants of the Workshop who are identified in the report.

Stephen B. Weiner, Ph.D.
Workshop Organizer
Principal Investigator

THE WORKSHOP

A Workshop on Qualitative Information Retrieval was held in Washington, D.C. November 18-20, 1980. The Workshop was jointly funded by the Office of Naval Research (ONR Contract #G00014-80-0105) and the National Science Foundation (NSF Grant #IST-8018904). The contract and grant were awarded to the D.C. campus of the Center for Policy Research. Participating in the Workshop were:

Stephen B. Weiner (Principal Investigator, Workshop Organizer)
Center for Policy Research

Martha E. Williams (Co-principal Investigator, Co-chairman)
University of Illinois at Urbana

Donald A. Hillman (Co-principal Investigator, Co-chairman)
Lehigh University

Pauline Atherton (Keynote Speaker)
Syracuse University

Tamas E. Doszkocs
National Library of Medicine

Lynette Hirschman
New York University, Linguistic String Project

Paul Kantor
Tantalus, Inc.

A. Donald Merritt
National Library of Medicine

Robert Niehoff
Battell Memorial Institute, Columbus

Gerard Salton
Cornell University

Workshop Participants (continued):

Roger Schank
Yale University

Linda Smith
University of Illinois at Urbana

Irene Travis
PRC Information Sciences Company

The Workshop was designed to bring together experts in fields germane to the problems of information retrieval. These fields included information science, psychology, linguistics, artificial intelligence, and computer technology. The purpose of bringing together these individuals was to discuss the problems involved in retrieving information when controlled vocabulary fails. The final products of the Workshop are to be:

1) A state-of-the-art description of information retrieval hardware, software, and techniques;

2) A typology of information retrieval problems; and,

3) A method by which information retrieval systems could be described in such a way that the descriptions were both objective and comparable.

A combination of presentations and discussions during the evening of the Keynote Address and the two days of Workshop proceedings led the participants to four conclusions which are listed here and elaborated upon more completely below.

1. Research on information retrieval, particulary that involving new hardware capabilities, new file structures, and new software (including artificial intelligence) does not impact on commercially available systems as much as it could. Therefore,

a state-of-the-art assessment, in and of itself, might not have immediate practical results.

2. Systems descriptions cannot be generalized, except in the most trivial cases, so as to allow comparison of different systems.

3. Measures exist for accurately describing any given system in any particular environment, which may be the most that can be hoped for.

4. An emphasis on controlled vocabulary is misplaced. Adequate methods exist or are close to operational in the laboratory setting to resolve most, if not all, of the retrieval problems which occur when controlled vocabulary fails.

STATE-OF-THE-ART

There are two arts that must be considered in information retrieval, the tools and the measures. The tools include both hardware and software; and the measures allow assessment of the effectiveness of the tools.

In order to measure and assess a system's ability to retrieve information, models of that system and the functions it is to perform need to be developed. In her keynote address, Pauline Atherton spoke to that point. Noting that she was speaking to the issue of the quality of information retrieval rather than on qualitative retrieval, Dr. Atherton suggested that we are still using models and measures which reflect the state-of-the-art

in the 1950's and 1960's. She suggested a need to review and revise the literature of that period to reflect the changes which are evident in today's information retrieval environments.

Specifically she noted that Lancaster's classic model of the major functions performed in information centers needed to be revised to take into account that most information centers now access multiple data bases, access multiple data base systems, and do so in an on-line environment. While the Lancaster model may be an accurate description of an automated system which accesses only one data base and uses only one retrieval system, it is an equally accurate model of the traditional libaray with card catalog. It does not accurately reflect the functions that most information systems must perform in the 1980's. It therefore cannot be used to develop standards against which to measure performance.

The Salton and Lesk flowchart of user-system interaction is more appropriate for describing the on-line environment than is Lancaster's model, yet it too is an incomplete description of the multiple data base, multiple system environment.

Similarly, H. P. Edmundson's model of the automatic extracting system must be revised to account for document and search requests as well as for user objectives among other factors.

This does not mean that there are no contemporary studies or models which are useful. Indeed, Martha Williams, in her opening remarks at the Workshop spoke to that point.

Professor Williams presented a list of factors which affect the quality of retrieval and noted that measures might be explicit or implicit. Among the factors she cited were the quality of the data itself, both form (citation or text) and content (acuracy, completeness, timeliness); the quality of the data delivered (recall, precision); and the quality of the process (speed, ease of use, efficiency, effectiveness, expense). And, she pointed out, we have measures for all of these.

The question of measuring information retrieval systems will be discussed at greater length below, so it is sufficient here to note that at the abstract level, the state-of-the-art in measuring information retrieval systems is sufficient for the need to measure.

The state-of-the-art in systems, hardware and software, is more difficult to describe. This is true for two reasons. First, developments are not limited to any one field or research; rather they are spread over many fields. Second, the state-of-the-art in tested hardware and software is not the same as the state-of-the-art in commercially available systems.

Advances in information retrieval systems are being made in hardware and software. They are being made by computer technologists, programmers, psychologists, cyberneticists, philosophers, linguists, and researchers in other related fields. Unfortunately, as many Workshop participants noted, even practitioners in closely allied fields rarely have a chance to interact, to exchange ideas, or to report on progress in their fields. This Workshop was an attempt

to provide such an opportunity. Several participants were asked, or volunteered, to give informal presentations of the work being done in their particular areas of interest. Understandable, most reported on their own work, both through those presentations and through the frequently lively discussions they generated, a reasonable appreciation of the state-of-the-art was presented. Since the presentors were asked to speak informally and to provide only a brief outline of their remarks, it would be an injustice to present their written documents here. Rather, a summary of the material presented in those remarks will be given here and a more lengthy paraphrase will be provided in the appendix.

Lynette Hirschman reported on the work being done on the Linguistic String Project. This project is working on developing an automated information processing system that can analyze text on a sentence-by-sentence basis using a detailed syntactic parse. The system can then store the information for fact retrieval in response to questions requiring a direct factual response. Working with limited subfields such as pharmacology and hospital discharge summaries, a system has been developed which can parse the records and describe information structures based on repetition and variation of basic patterns. While this system has immediate practical value in the limited subfields with which it has been developed, it has potentially much wider application.

The techniques developed to automatically process natural language records in one subfield can be adapted to allow the processing of records in other subfields. Irene Travis suggested

that in this instance, as in many others, the question was not one of generalizability--can the system be applied on a larger scale--but one of portability, can it be used in a different problem area.

The work on sublanguage information structures which led to the development of this system is based on the fact that language is highly structured and that information is carried in a systematic way. This allows for a system to be developed which can recognize objects of a subfield, the relationships between those objects, the quantification of the relationship, and sentential connectives (e.g., cause, be associated with, inhibits). The system can also process negatives, conjunctions, and time expressions. All of these abilities are necessary if the system is to be either generalizable or portable.

By restricting the system to a limited subfield, the problems of controlling for local synonyms and avoiding homographs become manageable. This system, then is portable, but it may not be generalizable. This does not dimish its value or importance, and the work reported on by Dr. Hirschman as well as similar work being carried out at other locations, leads the Workshop to believe that the state-of-the-art allows, within limits, the automatic processing of records into a fact retrieval system. This ability can benefit many fields other than the medical ones on which it was developed.

Dr. A. Donald Merritt of the National Library of Medicine reported on the implementation and testing of a consensus-derived knowledge-based retrieval system.

The development of this system was predicated on the problems medical doctors encounter in using a bibliographic retrieval system. In many cases, the doctors do not have the time to do an extensive search only to be given citations. The information they seek need to be available for immediate use. NLM has developed an alternative to the citation retrieval system, an alternative that allows the user to retrieve information on-line.

As a test, NLM created an knowledge-based system on viral hepatitis. This was done by first identifying a number of experts in the field of hepatitis and asking them to review the literature on viral hepatitis and to exerpt the information they felt most important. The panel of experts, using word processors electronic mail communications developed what is, in effect, a text book on viral hepatitis. The text includes virtually any aspect of the problem a doctor might need to find, from symptoms and diagnosis to vectors and treatment. Where disagreement exists, the varying viewpoints are represented. A user can go on-line and using the keywords, ask for information of the effectiveness of a particular drug in treating the disease and for contraindications for the use of that drug as well as alternate treatments if contraindication exists, and receive the basic information or the actual text of the document from which the information was taken so that he can read the original research or report. The data base is continually

up-dated as additional articles and texts appear.

NLM is in the process of preparing a similar consensus-derived knowledge retrieval system on peptic ulcers and has plans to develop one on human genetics which will incorporate visual material as well as text.

Again, we have a system limited to a very specific topic, but one which has immediate and practical use and which is portable. It should be noted that such a system is only practical because of the availability of terminals and improved computer based communications systems which allows the consulting experts to exchange information efficiently and rapidly and for the timely up-date of the knowledge base.

The utility of developing similar knowledge bases in fields other than medicine seems to be obvious. But, the practicality of doing so is still to be determined. The viral hepatitis data base cost some $500,000 to create. While some savings might be realized in future knowledge bases because much of the basic research into indexing strategies and indexing has already been done, those savings might be more than offset by increased costs incurred in adding such features as visuals, or by inflation.

Tamas E. Doszkocs.reported on recent enhancements to the MEDLARS II system. Unlike other presentations which described experimental work or the development of new systems, Dr. Doszkocs reported on adaptations to an existing system. Given the extent

of file structures and investment in the MEDLARS II system, the problem was one of incorporating advanced techniques to allow open access to the system by more people. In making such adaptations, the spectrum of users, from the trained sophisticated searcher to the totally untrained searcher had to be considered. Further it was necessary to keep in mind that while the searcher might be untrained in the use of information retrieval systems, he did have a knowledge of the topics he was searching and would be able to recognize relevant information better than a non-medical but trained search specialist could.

The basic adaptation to the MEDLARS II system was the introduction of a stand-alone application program that is used on the existing files. This program allows the user to enter a natural language query from which the system extracts keywords and performs the search of the inverted files. The user is then offered abstracts of the items retrieved and asked to indicate which are relevant. On the basis of this, the system assigns a postings value to each keyword in the index field of the items indicated as relevant and, ranks the retrieved items according to decreasing weight. The weight is assigned on the basis of the inverse collection frequency of each term. By providing a method for the searcher to indicate the relevancy of items retrieved through a search query that might not have been optimal, the system provides a method by which the user can approximate the performance of the average trained searcher.

Other, more sophisticated techniques are being considered for adaptation into MEDLARS II include automatic stemming, heuristic techniques for modifying strategy, and partial matching.

Methods of using artificial intelligence to retrieve information were reported on by Roger Schank. Dr. Schank began by describing the capabilities of the ideal information retrieval system. It should be one in which the computer "knows" what it is looking for when it searches memory. It should be able to find partial matches, communicate in natural language, be able to understand enough to give what is really wanted rather than what is asked for.

The Yale Artificial Intelligence Project is attempting to develop such a system, one which knows a great deal of information about a subject and which can be queried in a natural way. His project is developing programs which are the beginnings of solutions to some of the problems involved in such an undertaking. The requirements of that ideal system are that:

1. The system automatically understands natural language text. This involves producing a representation of the content of language input.

2. Information in the system should be automatically formatted and organized for retrieval in such a way that the conceptual content of an item can be used for retrieval rather than key words. Such automatic parsing should be integrated into the memory organization.

3. The system needs rules for accessing its memory, both for directing input of new information and for directing retrieval.

4. Rules for answering questions on the basis of intent rather than literal meaning must be developed.

A number of AI systems have been developed at Yale. Among these are:

1. FRUMP--skims stories from the UPI news wire and outputs a conceptual representation of the important events of each story that it understands. It skims, looking for what it wants to know about. New domains are added by supplying only the domain specific world knowledge; FRUMP's linguistic knowledge is independent of any particular domain. FRUMP uses a data structure called a sketchy script to organize its knowledge about the world, describing what can occur in a given situation. Understanding has two phases: (1) select a sketch script; (2) process the story using the selected sketch script to predict the concepts in the important events in the story. It is very fast and can process storeis as quickly as they come over the wire, producing summaries of the stories that it reads in a variety of languages.

2. IPP--reads stories in one domain where it has extensive knowledge which is used to understand and learn from what it reads. It can make generalizations based upon what it reads.

3. CYRUS--a memory model that organized biographical information about people and uses knowledge about its organization for retrieval and automatic updating. CYRUS has two modules--a question-answering module which answers questions put to it by a human user

and an updating module which automatically adds new information to memory after that information has been pre-processed by FRUMP. The memory is organized in episodic categories called MOPs. Similar episodes are stored in the same MPO along with the generalized knowledge built up from the similarities in the episodes (e.g., preconditions and enablement conditions for episodes, sequence of events, usual results). In retrieval CYRUS makes use of approximately 10 search strategies based on those observed in people answering questions.

4. BORIS is a prototype of the ultimate understanding program. It reads every word of a story and follows out all the inferences it can about the motivations, personal relationships and beliefs of all the characters it finds. It has a memory model within it that it exploits to help it detect patterns in the world worth noting and reasons why various actions occur.

According to Dr. Schank, many if not all, of the AI techniques developed for these systems are available for use in operational systems. That is, they are no longer to be considered simply experimental. Dr. Schank sees no virtual limit to the ability of AI to solve retrieval problems inherent in systems not using artificial intelligence.

Although he was unable to attend the Workshop, Charles T. Meadow of Drexel University has been working on a system component designed to help the searcher maximize his search strategy. It is

a program which monitors an on-line search. If it notes that search results are poor (either very few or very many retrievals) the program will come on-line and offer suggestions. For example, it might ask the searcher if he would like to see an on-line thesaurus in order to broaden his search term.

When he began the project, Dr. Meadow hopes to be able to develop a search monitor that could be used on any, or at least most, of the major data base systems. That has proven, to this point, not to be feasible. However, it should be possible to develop such a program for systems on an individual basis should such a search aid prove valuable.

Robert Niehoff described a Vocabulary Switching System (VSS) developed at the Battelle Memorial Laboratories in Columbus. The system is for use when multiple data bases are accessed. The vocabularies that Battelle has used in developing the system include those of business, psychology, biomedicine, and physical science. The VSS allows a user to define his own switching strategy or to use the default strategy provided. (Battelle is working on determining the optimal strategy). Basically, the system provides the searcher with nineteen search options including exact match, synonyms, adjacency, and BT and NT switching (if the vocabulary has a hierarchical structure). The system accepts the options in any sequence. The system will accept the output of one option as the input for the next.

The VSS will search multiple data bases for the user. Both

recall and precision range from 50-60% in evaluations studies performed on the system. The homograph problem is not difficult when similar files are searched. In addition to aiding the user in accessing multiple files, the VSS can compensate for insufficient clues to alternate search terms within a data base. It can also be used to improve searching on the target data base as well as related ones and can suggest additional files to search.

The Vocabulary Switching System is one of the few tools specifically designed for the new environmental conditions referred to by Pauline Atherton.

In addition to the progress being made in developing more powerful software to aid in information retrieval and in developing different file structures and knowledge rather than citation data bases, advances have been made in the hardware protions of information systems as well. Most current systems, and all of the major commercial and public data bases were developed for use on hardware that had limited storage ability. Even given unlimited storage ability, the ability to search very large data bases, or data bases composed of full texts of documents was too slow to be economically practical. Further, since matching records against the query was done in the CPU, and since CPU time is the most expensive component in the cost of a search, the ability to store unlimited amounts of information would not be useful.

New developments in the area of reading records and matching them against a query, however, make full text storage and retrieval

more practical. There are a number of new systems in operation or being tested. They operate on slightly different principles, but all have the characteristic of being able to scan large portions of the data base quickly and identifying relevant items without depending on the CPU. A typical system will use multiple read/write heads. At the extreme, there might be one for each groove in the memory disc. This allows the entire memory to be searched in one revolution of the disc rather than requiring a single head to read each groove independently. Between each read/write head and the CPU is a mini-processor which does the matching of the information against the query that would otherwise be done by the CPU. The combination of quickly reading the entire data base and avoiding the use of the CPU makes searches of very large data bases feasible. In theory, entire full text libraries could be stored on discs and retrieval could be made on full text searches.

Even without going to the extreme case and placing a read/write head on each groove, the ability to economically access full texts or long abstracts means that the necessity of inverted file structure may no longer exist. Information can be stored in files created to meet user needs rather than those designed to meet systems demands.

In brief, the state-of-the-art of information retrieval, both from the software and the hardware perspectives seems to be fine, at least in theory. The problem, as Pauline Atherton pointed out in a slightly different context, is that research has had little impact on application. Tamas Doszkocs pointed out

the problem in his presentation when he began with a pragmatic look at MEDLARS. A given file structure exists and that structure represents a major investment. Any new development which would necessitate a change in the file structure to implement is not likely to be adopted. Any improvement would have to be one which can be made without disrupting the basic system.

While the National Library of Medicine has taken the lead in developing such improvements, it may not have much company. Speaking of ways in which existing systems could be significantly improved in terms of their ability to yield search results with up to 60% higher recall and precision, Gerard Salton stated that tests have shown that better indexing using content analysis of documents or by using improved thesauri to allow a searcher to broaden narrow terms could provide that improvement. The use of such techniques would not require any modification of existing data bases, but, according to Salton, adminstrative, financial, and human barriers exist to the implementation of these techniques. For example, weighing retrievals as the MEDLARS II system can now do, may be an effective method of improving the quality of those retrievals. But, says Salton, users often do not understand how to use the weighings. /NLM retrieval specialist told Dr. Weiner the same thing several monts prior to the Workshop./

On the whole, the available retrieval systems meet user needs at a satisfactory level; there is, therefore, no motive for the purveyors of those systems to change or adopt the new hardware and software.

SYSTEMS MEASUREMENT AND COMPARISON

As noted above, systems can be measured on a number of variables--recall, precision, ease of use, speed, cost. . .It is indeed possible to measure every system on those criteria. It might even be possible to measure systems on their ability to retrieve when controlled vocabulary fails. But, it was the consensus of Workshop participants that comparing systems, in all but the most trivial of cases cannot be done. Irene Travis indicated the problem when she pointed out the difficulty of the interconnectivity of variables and the impossibility of generalizing beyond laboratory research which retrieval aids perform better.

Among the variables that Dr. Travis referred to were the skills of the user, the uses to which retrieval systems are put, and the environment in which they are located. Pauline Atherton referred to the enormous individual variability in searching behavior and performance of descriptors, operators, number of documents retrieved, and the number of concept groups in the data base; the variation in the complexity of searches which may be realted to the institutional setting; and the work speed which is influenced by search formulation, presearch preparation, cost conscious attitudes, and searcher familiarity with the data base and the system.

While many, and perhaps all of these factors might be controlled for in a laboratory setting, once a system is moved

into an operation setting, the experimental findings have no validity. This is not to say that measuring cannot be done. Indeed, Paul Kantor, in the paper provided in the appendix, makes some excellent suggestions for ways in which systems can be measured and, perhaps eventually, compared.

But, overall, the Workshop offers two alternatives to those interested in choosing a system for their particular retrieval needs. The first is to determine the needs which the system must meet and then to measure each system against its ability to meet those needs. Each system is different and the tradeoffs that each organization will make are different. One will trade money for speed; another recall for precision; and a third will trade ease-of-use for lower precision. There is no one best system, only a system which most suits the specific demands of an individual environment. The second option suggested is to determine the needs which the system must meet and to design a system to meet them.

CONTROLLED VOCABULARY

It was quite evident from the outset of the Workshop that participants felt the emphasis on problems of retrieval when controlled vocabulary fails was misplaced. Although participants seemed to accept the utility of creating a typology of retrievable problems that arise when the controlled vocabulary failed, it was their opinion that the ability to overcome those problems already exists. Many of the standard features of widely available

systems are desinged to meet the problems. Among them are included Boolean searches, full text searches, searches on fields other than index terms, truncated searches, searches on broader and narrower terms when the index is hierarchical, proximity searches, and string searches. Automatic synonym control is also an aid.

In addition, a number of the developments reported on at the Workshop are also useful in increasing recall and precision when problems of vocabulary arise. These include switching systems in multiple data base systems and the weighing of retrievals.

These tools are available now and can be used with existing data bases with little or no changes to the data base itself, particularly with no necessity to engage in the costly process of restructuring files.

The work reported from the Linguistic String Project and the Yale Laboratory working on artificial intelligence provide even more powerful tools for retrieving knowledge from automated data bases.

It is the consensus of the Workshop that virtually all retrieval problems are resolvable with existing or preliminarily tested hardware and software. Whether or not the available tools will be used, whether or not existing systems will adopt them and new systems will be planned to incorporate them is a different

problem, one that raises the question of cost-benefit and how much retrieval is enough, both beyond the scope of the Workshop.

CONCLUSIONS

The idea for this Workshop came as the result of trying to answer the question, "How can one compare systems on their ability to retrieve information when the controlled vocabulary fails?" The Workshop was intended to not only answer that question but to suggest a method to define and describe systems in such a way that the description would be comparable and therby allows the comparison. The Workshop did not do this.

Scientists are prone to repeat the axiom that there is no such thing as an experiment that fails. Although the Workshop did not accomplish the goal set for it at its inception, the results, and perhaps more important, the process, were valuable.

The Workshop participants suggest that in evaluating a retrieval system, the retrieval needs must first be determined. These needs are determined by the information that is required, the format is is required in, and the end user of that information. The quality of information provided by a system must then be looked at. Among other variables, the more important ones include the quality of the information itself--its accuracy, completeness, currency--, the quality of the retrieval--recall and precision--, and the quality of the process--speed, ease, cost. Once the needs

and the quality are determined, choose or design the system that can provide for it. The tools and techniques are available.

Workshop participants indicated a need to identify priorities for the implementation of and research on new procedures. There is a need to develop procedures that are adaptable, or as Irene Travis put it, procedures which can be moved from one system to another or even one data base to another with minimal cost and effort.

There is a need to develop more front end devises, preferably transparent to the user, to increase access to automated information retrieval systems. These devises should help the user choose and enter into the most appropriate data base(s) for his needs.

Perhaps most significantly, Workshop participants indicated the need for more forums at which researchers in different, but related, fields could meet to exchange information. Several participants commented on the "remarkably long-standing antipathy" between individuals engaged in work on artificial intelligence and those engaged in information retrieval. Both are working on similar problems. Similar solutions to those problems have been proposed by both groups. Yet there is little communication between the two. This Workshop was one such occasion and one which seemed to be welcome by representative from both groups. More opportunities would be beneficial to the individuals and to the fields they represent.

APPENDIX

NOTES* SUBMITTED TO WORKSHIP PARTICIPANTS

* Several Workshop participants were asked to deliver informal presentations on their research to serve as catalysts for discussions. A number of participants also provided notes or outlines of material they presented. These were not solicited as finished, polished, journal articles; they should not be judged as such. They are included here as a courtesy to the reader, to allow him to obtain a more complete sense of the material summarized in the Final Report.

The following material was submitted by Lynette Hirschman and served as the basis for her presentation.

SUBLANGUAGE INFORMATION STRUCTURES

Lynette Hirschman
New York University

1. Discovering sublanguage patterns

   A. Language highly structured, carries information in a systematic way: within a limited subfield can describe information structures, based on repetition and variation of basic patternsn, obtained by linguistic analysis.1/

   B. Can determine sublanguage informational classes:
   words which appear in same syntactic enviromment (e.g., as the subject of a particular verb+object) share an element of meaning, e.g., sodium, potassium, Na^+ can all appear as the subject of the predicate flow into the cell in pharmacology texts;
   group together words which share a number of enviornments, to get a semantic class consisting of subfield synonyms (e.g. sodium, Na^+) and words with related meanings: sodium and potassium are both ions;
   this technique can be used to determine a set of informational classes for a particular subfield; it has also been automated.2/

   C. Patterns of semantic classes indicated information structure of subfield. This can be described in terms of a hierarchy of predicates on arguments; lowest level: the objects of the subfield, e.g., for pharmacology subfield, ions, drugs, tissues, organs, etc.
   relations between objects: predicates on these classes, e.g., move class: ion flows into/accumulates in/enters/cell;
   quantification of relations operate on elementary relations; increase/reduction of ion influx into cell;
   sentential connectives: operate on other predicates: e.g., cause, be associated with, inhibit, as in: (administration of digitalis) inhibits (the influx of sodium into the cell).

   D. Why restriction to a subfield?
   Ambiguity is reduced within a subfield;
   local synonyms can be defined (e.g., film in radiology reports is a local synonym of x-ray, and never refers to a movie);
   facilitates collection of data about repeating patterns and variations; distributions too fuzzy in language as a whole.

2. Uses of sublanguage classes and patterns

A. Fact of retrieval

Use patterns to construct an information format for a subfield (a table whose columns correspond to sublanguage classes); syntactically analyzed sentences are mapped into the information format to create a tabulary structured arrangement of data that can be used for fact retrieval.3,4/

use to construct a data base schema, using conventional DBMS.5/

B. Use to improve natural language processing:

experimental application in automatic classification of new sublanguage words based on similarity in distribution to words of know class;

reduction of amibiguity and resolution of homographs based on constraints on which word classes can occur in combination with other word classes (as predicates or arguments).

quality control for processed sentences:

flag combinations that do not conform to known sublanguage patterns: these represent either incorrectly parsed sentences or new sublanguage patterns.

References

1/ Sager, N, Syntatic Formatting of Science Information, Proceedings of the 1972 Fall Joint Computer Conference, AFIPS Press, Montvale, NJ (1972), 791-800.

2/ Hirschman, L., Grishman, R., and Sager, N., Grammatically-Based Automatic Word Class Formation, Information Processing and Management 11 (1975), 39-57.

3/ Sager, N., Natural Language Information Formatting: The Automatic Conversion of Texts to a Structured Data Base. In Advances in Computers 17 (M.C. Yovits and M. Rubinoff, eds.), Academic Press, New York, 1978, 89-162.

4/ Hirschman, L. and Sager, N., Automatic Information Formatting of a Medical Sublanguage. In Sublanguage: Studies of Language in Restricted Semantic Domains (r. Kittredge and J. Lehrberger, eds.), Walter de Gruyter, Perlin (in press).

5/ Story, G. and Hirschman, L., Database Design for Natural Language Medical Data. To appear in Proceedings of the 14th Annual Hawaii International Conference on System Sciences, III.

A SUBFIELD OF PHARMACOLOGY

1) Sample paragraph

Toxic doses of digitalis consistently reduce the intracellular concentration of potassium in a wide variety of cells, including cardiac muscle cells. This results from the slowing of the influx of potassium into the cell. Concurrently, intracellular sodium and water are increased. It is not certain whether these linked changes in sodium and potassium are produced by a single effect or are separately mediated.

2) Grouping together words from the same syntactic environment

potassium	flow <into> cell	
sodium	flow <out of> cell	potassium, sodium, Na^+ similar
Na^+	flow <out of> cell	

3) Basic sublanguage entities

A pharmacological agents: (digitalis, glycoside, ...)

I ion: (sodium, potassium, calcium, cation, ...)

R reticulum: (sarcoplasmic reticulum, SR, ...)

C cell: (cell, red blood cell, ...)

T tissue: (muscle, fiber, ...)

M membrane: (membrane, ...)

O organ: (heart, kidney, ...)

...

4) Relations between basic entities

lose	R/C/T/O__I/G	lose, take up, ...
permeate	I __M	permeate, ...
move	I __C/T	move, enter, flow, ...
expel	R/C/T/O__I	expel, excrete, regain, ...
exchange with	I __I	exchange with, replace, ...

...

5) Quantifiers and Quantificational operators

Q (quantifier)	V_q (quantificational operator)
dose, amount, rate, ...	increase, change, elevate, decrease, ...

6) V_{SS}: connective verbs (connecting sentential subject and object)

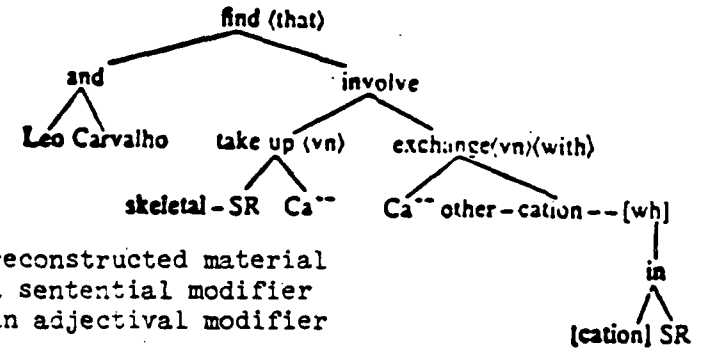
affect, trigger, cause, involve, inhibit, result, ...

7) V_S: verbs with human subject and sentential object

believe, find, publish

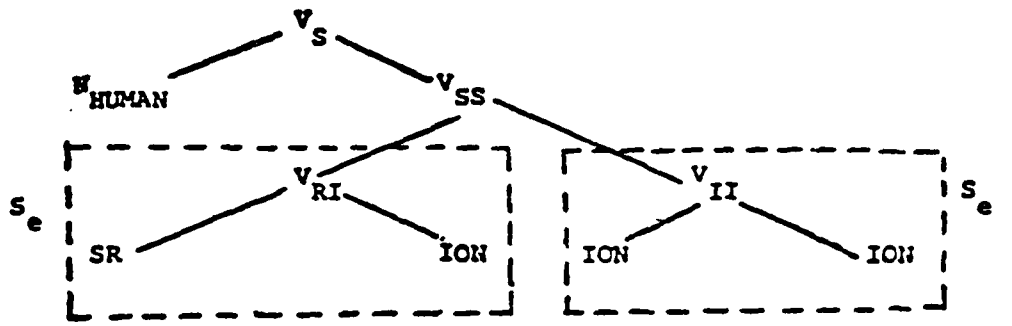
Operator-argument representation of a sublanguage sentence, showing the hierarchy of relations.

Carvalho and Leo found that the Ca^{++} uptake of skeletal SR involves the exchange of Ca^{++} with other cations in SR. (LE711 13C.5.7)



[ ] indicates reconstructed material
-- indicates a sentential modifier
- indicates an adjectival modifier

Schematic representation of operator-argument hierarchy



V_S : verb class with sentential object, human subject (e.g. find, think assume)

V_{SS}: verb class with sentential subject and object, often causal in nature (e.g. influence, affect)

S_e : elementary sentence format, specific to the subfield; here involving:

V_{RI}: verb class connecting nouns of the SR class with those of the ION class (e.g. take up, contain)

V_{II}: verb class connecting nouns of the ION class to nouns of the ION class (e.g. exchange, compete)

A MEDICAL RECORDS SUBFIELD

1) Sample paragraph

The child in hospital was seen by infectious disease people who reviewed history and physical and agreed with findings reiterated above. On 1-22 the child was doing well, was afebrile and was feeding. All cultures including urine, blood, throat, CSF were normal. It was decided on 1-29 to continue the child on chloramphenicol for a total of 14 days. Initial treatment that the child received in hospital consisted of chloramphenicol 300 mgm, intravenously Q6H. This was continued until discharge.

2) Grouping together words from the same syntactic environment

patient developed	fever cold anorexia seizures	group together into a SIGN-SYMPTOM class

3) Basic sublanguage entities:

SIGN-SYMPTOM:	(fever, cold, distress, pain, ...)
DIAGNOSIS:	(chicken-pox, meningitis, sickle cell disease, ...)
BODY-PART:	(leg, hand, heart, GI, ...)
LAB-TEST:	(x-ray, urinalysis, culture, ...)
DRUG:	(chloramphenicol, salicylate, drug, ...)
TREATMENT:	(bed rest, transfusion, ...)
PATIENT:	(patient, pt)
DOCTOR:	(MD, doctor, consultant, radiologist, ...)
INSTITUTION:	(hospital, emergency room, out-patient clinic, ...)
...	

4) Relations between basic entities:

$V_{patient}$	PATIENT ..... SIGN-SYMPTOM	(have, develop, complain of, ...)
V_{treat}	DOCTOR ..... DRUG prep PATIENT	(give, order, prescribe, ...)
V_{admin}	DOCTOR ..... PATIENT prep INST	(admit, transfer, discharge, ...)
V_{show}	TEST ..... SIGN-SYMPTOM	(show, reveal, indicate, ...)
...		

5) Quantifiers and Quantificational operators

Q (quantifier)	V_q (quantificational operator)
dose, level, ...	increase, reduce, change, ...

6) V_{SS}: connective verbs (connecting sentential subject and object)

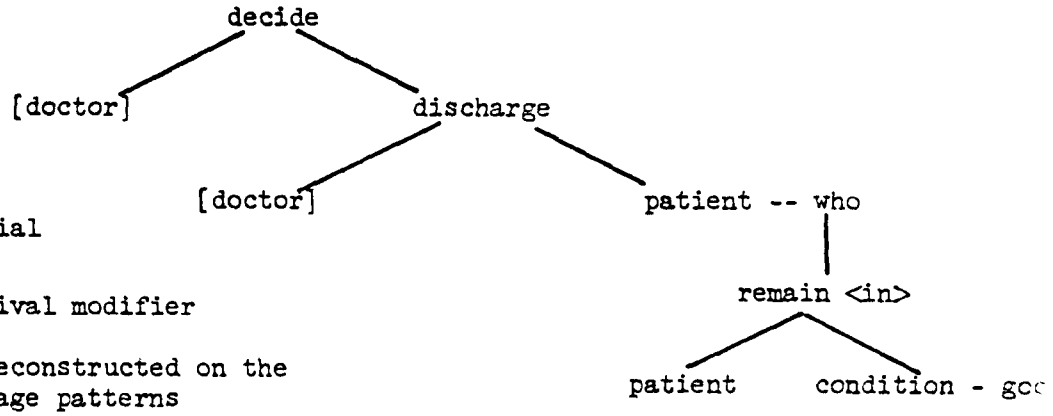
be realted to, be associated with, represent, be typical of, ...

6) V_S: verbs with human subject and sentential object

review, agree, feel, decide, ...

Operator-argument representation of a sublanguage sentence, showing hierarchy of relations.

It was decided to discharge the patient who remained in good condition.

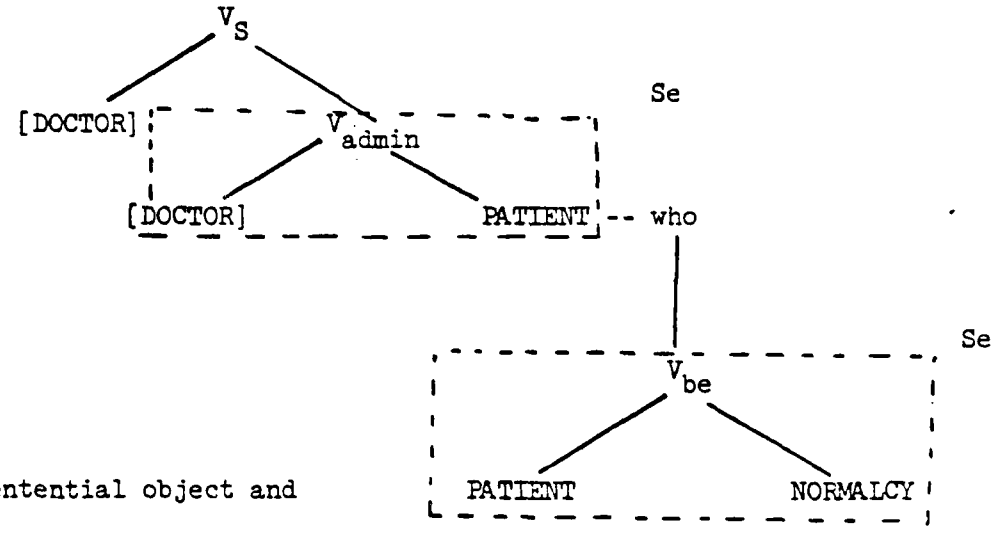


-- indicates a sentential modifier

- indicates an adjectival modifier

[ ] indicates words reconstructed on the basis of sublanguage patterns

Schematic representation of operator argument hierarchy



V_S: verb class with sentential object and human subject

Se: elementary sentence format, specific to the subfield, here involving

V_{admin}: verb class connecting DOCTOR with PATIENT, discussing medical administrative actions (hospitalize, discharge)

V_{be}: verb class connecting PATIENT class with either SIGN-SYMPTOM (e.g. PATIENT is sick) or NORMALCY class (e.g. PATIENT is well)

Format for a subfield of pharmacology

GL 641 2.2.1 MORE DETAILED STUDIES OF THE AFFECTS OF CARDIAC GLYCOSIDES ON SODIUM AND POTASSIUM MOVEMENTS IN RED CELLS HAVE BEEN MADE BY KAHN AND ACHESON (99), SOLOMON ET AL (168) AND GLYNN (67).

HUMAN	V-STUDY	DRUG	V-CAUSE	ARG1	V-PHYS	ARG2	CONJ
K AND A (99) S ET AL (168) AND G (67)	HAVE MADE MORE DETAILED STUDIES OF	{CARDIAC {GLYCOSIDES	AFFECT	SODIUM	MOVE IN	RED CELLS	AND
				POTASSIUM	[MOVE IN]	[RED CELLS]	

Format for hospital discharge summary (partial)

Sentence 1: Patient was given a penicillin injection.

Sentence 2: Tonsillitis was treated with ampicillin.

	MD	TREATMENT		PATIENT	PT-STATUS								
		V-TREAT	RX		V-PT	BODY-PART	FINDING						
							TEST		V-SHOW	NEG	RESULT		
							LAB	EXAM-TEST			NORMALCY	QUANT	QUAL
													SIGN-SYMPTOM	LAB-RES
1.		give	peni- cillin injection	patient									
2.		treat with	ampi- cillin			[tonsils]							tonsillitis	

Paul B. Kantor provided the follwoing material for inclusion in the final report. It is a more formal presentation of extemporaneous remarks he offered at the Workshop.

Remarks on the Evaluation of Retrieval Systems

Paul B. Kantor
Tantalus Inc.

My remarks are addressed to a problem which has not yet commanded a great deal of attention: the use of highly redundant data bases, for a variety of purposes, by persons who are not expert in the fields about which they inquire. I will comment upon (1) possible use of bibliometric models in the management of such bases; (2) the competing goals of "proximity" and "scope;" (3) evaluation of the usefulness to the user, of data retrieved.

It is easy to give examples of access to highly redundant files. For example, suppose a Professor of Engineering goes to her school's library and asks for an introductory textbook on Economics, explaining that her current project involves working with an economist, and she wants to know what economics is about. She is delighted with the first textbook drawn from the shelf. Then she asked for a second typical textbook. The collection is highly redundant: there are perhaps twenty nearly identical elementary textbooks. What she really wants is another example which is "rather far" from the first one, to explore the scope of the information available. Proximity is required, since the items sought must have something to do with learning about economics. But excessive proximity is redundant and wasteful.

The significance of this problem increases enormously when the data file contains, for example, a full year's newspaper clippings from a large set of

newspapers. The same item will appear in hundreds of nearly identical forms, and scope is at least as important as proximity in expanding the retrieved set. Similar remarks apply to an extended file of intelligence reports from varying sources, and in the latter case the number of times an essentailly similar report has been filed may be, itself, important. Some screening and selectivity must be built into the system, and an optimally designed base and retrieval system must accomodate a steady influx of new personnel who are not only unfamiliar with the retrieval system but are also not really "expert" in the subject area.

As long as the subject matters does not require graphics, or the inclusion of algebraic expressions, one may assume that free text searching will be possible.

Specific remarks:

1. Bibliometrics as a tool.

It is know that the distribution of use over terms is a combination of meaningful processes, related somehow to the importance of the terms for the discussion, linguistic regularities which require the presence of certain common words, and stochastic accumulation, by which more heavily used terms "attract" heavier use. This leads to a "buzz word" effect, and the specific words may change, within a given field, over a short span of time. For example, "quality of life", "impact", "enviornmental considerations" and "conservation" have all, at various times been "buzz words" for the closely related concepts to which they refer. A better understanding of the bibliometrics governing the distribution of such terms would support the development of an automated system for recognizing them as they appear. There has been some significant progress

recently, as we have solved the Cumulative Advantage Model of Simon (later, Price) exactly. The results do not yield a Bradford distribution, and are much better approximated by a geometric distribution.

2. Proximity and Scope

In the free text situation (as in the controlled vocabulary situation) one can define a measure of the proximity of two items (for example, by overlap of terms, weighted to reflect the rarity of "information value" of those terms.) The system described by Doszkocs essentially does just this. More generally, the proximity measure can be interpreted as arc length on a hyperspherical surface in a space of large enough dismension, so that the items in the file are represented by points on that surface. The arc-wise distance between two points represents their proximity, while the area of the hypersurface which several points define represented their combined scope. In a likely problem situation, the user will have specified the number of items (n) to be retrieved, and the relative importance of scope and proximity in his search.

This results in a number of interesting problems of an Operations Research nature, including: how to handle the multiple objectives of scope and proximity and how to determine the best set of n items. It is not yet known whether the problem can be solved by dynamic programming and induction on the size of n. There are also interesting problems of a more mathematical character such as fixing the dimension of the space, and the volume measure on the hypersurface. These questions must be reduced to specific algorithms, so that the system can perform the calculations in response to a given question, in real time.

3. User-Oriented Evaluation

Designers of retrieval systems must, naturally, focus attention upon the set of retrieved documents. The institution which pays the bill is concerned only with their impact on its total performance. The problem defines multiple objectives. We are currently focusing upon two measures:

A. End user estimate (subjective) of the value of the set of items retrieved, on several scales.

B. End user reporting of the total time required to sort through the retrieved set, and arrive at the judgement given in A.

Although these are competing objectives in many situations, it is quite conceivable that, in particular context, one system (or one operator) may dominate another. That is, it may produce items of greater value, and consume less of the end user's time. Where dominance does not occur, consideration must be given to the trade-off of these two goals.

References on multiple objective problems.

1. Arrow, K.J., Social Choice and Individual Values. Wiley. New York. 2nd Ed. 1963.

2. Fishburn, P.C., Decision and Value Theory. Wiley. New York. 1964.

3. --- The Theory of Social Choice. Princeton University Press. Princeton, New Jersey. 1973.

4. Grochow, J.M., On User Supplied Evaluations of Time Shared Systems. IEEE Transactions on Systems, Man and Cybernetics. SMC-3 pp. 204-205 (1973)

5. Kantor, P.B. and Nelson, R.J., Social Decision Making in the Presence of Complex Goals: Ethics and the Environment. Theory and Decision v. 10 pp.181-200 (1979). Reprinted in Key Papers in Decision Theory, Reidel Press, to be published.

6. Keeney, R.L. and Raiffa, Howard, Decisions with Multiple Objectives: Preferences and Trade-off Values. Wiley. New York. 1977.

7. Luce, R.D. and Raiffa, H., Games and Decisions. Wiley. New York. 1957.

8. Sen, A.K., Collective Choice and Social Welfare. Holden-Day. San Francisco. 1970.

9. von Neumann, J. and Morganstern, O., Theory of Games and Economic Behavior. Princeton University Press, Princeton, New Jersey. 2nd Ed. 1947.

This is a draft of the notes A. Donald Merritt used in presenting his remarks to the Workshop.

NATIONAL LIBRARY OF MEDICINE
KNOWLEDGE BASE RESEARCH PROGRAM

A. Donald Merritt, M.D., Chief
Health Professions Applications Branch
Lister Hill National Center for Biomedical Communications
National Library of Medicine

My perspective is that of a physician with concern for and interest in the transfer of published biomedical information to health care providers at all levels. My remarks are a synthesis of what I have learned from my colleagues -- from the group of researchers which has been assembled during the past year. The program has a number of facets, each of which has a major focus for the professionals involved.

Purpose

1. Describe the "problem"
2. Provide background for the research effort
   - o Published literature
   - o Costs of information transfer in biomedicine
3. Describe the viral hepatitis experiment
4. Describe the Knowledge Base Research Program

Introduction

For the past twenty years there has been emphasis on the transfer of information -- and technology to the biomedical community. President John Kennedy, in a message to Congress in 1962, stated the communication problem well, "The accumulation of knowledge is of little avail if it is not brought within reach of those who would use it. Faster and more complete communication from scientist to scientist is needed, so that their efforts reinforce and complement each other; from researcher to practicing physician, so that new knowledge can save lives as swiftly as possible; and from the health professions to the public so that people may act to protect their own health." The problem then is rapid access to the content of the published literature to facilitate biomedical information transfer from:

- o Scientist to scientist
- o Researcher to physician
- o Health professions to the public

At about the same time four specifics were addressed by the Public Health Service and the President's Science Advisory Committee (PSAC).

1. Central resources to allow browsing and retrieval of documents relieving library pressures
2. Critical reviews of timely subjects
3. Make better use of audiovisuals
4. Stress interdependence of human-mechanical aspects of communication

Let me now turn to some cost estimates for the information transfer process. Incremental data as represented by books, journals, reports, etc., continues to enlarge. Last year we attempted to quantify, with the assistance of King Research, Inc., published literature costs as well as costs in formal communical, that is excluding conferences, consultations, and the like. The number of publications (S1) is plotted from 1960 to 1985. In 1977 the total was approximately 20,000 in biomedicine alone. Their growth is steady at about 5% per year and is projected to a total of approximately 30,000 by 1985. What are the attendant costs? These are estimated with the usual qualifiers concerning sampling and variability in source material as indicated on the next slide (S2). These dollar estimates are enormous. In summary, writing and editing are approximately 400 million dollars; publishing, over 1 billion dollars; library services, approximately 2 million dollars; and an estimate of 3 billion dollars for use by the researcher, educator, and physican for a total of 4 billion dollars per year. Annual costs are plotted on the next slide (S3) from 1960 to 1977 with projections to 1985. Both current and cost of dollars are represented. Although these estimates may have large variances, establishment of research programs addressing the management of information taken against these expenditures of this order of magnitude are likely to be small. One could argue that any improvement in the dissemination of information might significantly improve information transfer without significantly increasing costs.

Even with these enormous efforts and costs the timely transfer of information to the health practitioner remains a major health policy issue today and a challenge to the information scientist. Conventional bibliographic

data bases and document dissemination services meet most information needs of the research scientist, but are generally inefficient transfer mechanisms for the busy practitioner who frequently lacks the time necessary to access the formal literature and the expertise necessary to adequately process the massive amount of information likely to be available on any given patient management problem. Not surprisingly, we find that the practitioner tends to place proportionately greater emphasis on informal communication channels as sources of information; such as personal consultation with an influential and knowledgeable colleague.

Knowledge Base Research Program

In order to better serve the information needs of the health practitioner, the National Library of Medicine has established a Knowledge Base Research Program whose objective is the development, implementation, and evaluation of consensus-derived, knowledge-based representations of selected biomedical content.

We distinguish between a data base and a knowledge base in that the concept of a knowledge base involves the synthesis of information contained in documents identified and retrieved from one or more biomedical data bases. Thus, we define the content of a knowledge base as information which is selected, reviewed, condensed, synthesized, and reorganized by medical experts for computer representation. A parallel program of intramural research and development in medical computer and information science and computer communications is investigating methods for computer acquisition of text and visual content, their computer representation and access.

As was so congently stated in the 1963 Weinberg Report on information tranfer, "The fundamental task is switching information not documents with the ultimate aim of quickly and efficiently connecting the user to the proper information." In the context of our efforts, "proper information" is construed to be knowledge--the derivative of information synthesis.

During the next 5 to 10 years we expect that most health professionals will have easy access to terminals--probably CRT's and printers. We also presume marked improvement in communications networks. With these "givens" major characteristics of a knowledge-based system are: (S5,6)

- o computer-based, condensed representation of published information
- o organized with varying levels of specificity
- o citations selected to provide authority
- o capable of presentation of illustrations complementing textual material
- o current data amenable to continuous updating
- o based on expert consensus
- o end-user convenient--simplified interactive information access through text and illustrations

Our interdisciplinary effort has three major components. Each of these integrated research efforts will be briefly described.

- o Text and visual content
- o Medical computer science
- o Biomedical communication processes

The Viral Hepatitis Experiment

The initial steps in constructing the Hepatitis Knowledge Base were: identifying syntheses published by hepatitis experts.1/ Further synthesizing and reorganizing the content by Lister Hill Center staff, and obtaining consensus review and updating by panel of ten subject matter experts.

The resulting text data base, corresponding to approximately 400 pages, is organized hierarchically by topic. For each topic heading there is an accompanying synthesis statement which represents the state of knowledge. Each heading and synthesis is followed by supporting elements derived from published source documents. Included within supporting paragraphs are citations to primary publications used to provide authority for assertions made in the synthesis statements. The system is now accessed from selected sites via a nationwide timesharing network.

Consensus updaing of the Hepatitis Knowledge Base is facilitated by use of the Electronic Information Exchange System (EIES). This computer conferencing network serves as the principal medium of communication linking the geographically dispersed experts with each other and with the National Library of Medicine.2/

Initial field testing of the Hepatitis Knowledge Base has begun at the nine sites of the collaborating hepatitis experts. Study areas of particular interest are the following: volume of use (by what numbers of what categories of users for what purpose); performance (quality, reliability and accuracy) of the system; and effectiveness (user satisfaction and user behavior). Physicians, medical students, librarians and other health professionals at these locations

have received training in the use of the Hepatitis Knowledge Base and are participating in the evaluations.

An abridge form of the Hepatitis Knowledge Base has been published as a supplement to a widely distributed medical journal for review by a general medical audience.3/

1/ "Viral Hepatitis" by J.W. Mosley and J.T. Galambos, in Diseases of the Liver, 1975; ed., L. Schiff; contents of the 1975 National Academy of Sciences Symposium on Viral Hepatitis published in the American Journal of Medical Sciences: plus selected journal articles.

2/ Elliot R. Siegel: "Use of Computer Conferencing to Validate and Update NLM's Hepatitis Data Base" in Electronic Communication Technology and Impacts," M.M. Henderson and M.J. MacNaughton, Eds. AAAS, 1980.

3/ Lionel M. Bernstein, E. R. Siegel and C.M. Goldstein: "The Hepatitis Knowledge Base: A Prototype Information Transfer System." Annals of Internal Medicine 93 (Part 2) 169-181, 1980.

How Does the System Work?

Interactive knowledge base access is supported by a Data General Eclipse C/330 minicomputer, using the MUMPS/MIIS operating system and the Hewlett-Packard display terminal.

(Demonstration with slides)

(S7) HP2648A Terminal

(S8) "Vertical transmission..." --heading, heading statement, pilot, data elements

(S9) Reference

(S10) Figure/table

(S11) *Table of contents* -- "Prevention..."

The Hepatitis Knowledge Base experience demonstrates the feasibility of constructing knowledge base content from which may be obtained in-depth synthesis on the current state-of-the-art biomedical knowledge in areas of interest to health practitioners. Yet, this is very much a "first-cut" effort, primitive in the ways in which textual and illustrative content is capable of being encoded, manipulated, restructured, and accessed.

Peptic Ulcer

A second content set in the field of gastroenterology is under study for peptic ulcer disease. There are significant new developments in diagnosing and treating this common and often serious disorder. The initial text is derived from an existing synthesis.4/ The content development is taking place in collaboration with members of the Center for Ulcer Research and Education (CURE) associated with the University of California at Los Angeles and the Wadsworth Veterans Administration Hospital. A panel of experts are developing and updating this information.

4/ Sleisenger, M.H. and Fordtran, J.S., eds. Gastrointestinal Diseases, 2nd edition, W.B. Saunders, Philadelphia, 1978.

Human Genetics

Diseases which have a significant genetic component, even the risk of these diseases to prospective offspring, result in an enormous burden on the health care system and its users. Dr. V.A. McKusick's Catalog of Mendelian Inheritance in Man 5/ is the foundation for an expanded Human Genetics Knowledge Base that covers some 3,000 monogenic traits. Through consensus, a panel of subject matter experts is expanding and updating its contents using the text editing system (WYLBUR) at the Division of Computer Research and Technology, NIH.

The capability to systematically assemble and process illustrative materials to support the Human Genetics Knowledge Base text is being developed in-house. Plans are being defined using emerging technologies for storing and retrieving both digital and visual information. For example, optical videodisc technology will permit storage of and access to visual information in the form of single illustrations or motion supplements with audio. Simultaneously accessing textual and visual materials will assist health practitioners by providing current knowledge in the important, complex, and rapidly expanding field of human genetics.

(Slides for KBRP)

(S12) Summary (Process)
(S13) Summary (Activities)
(S14) Detail
(S15) Medical text
(S16) Medical visuals
(S17) Biomedical communication processes
(S18) Medical computer science
(S19) Program summary

5/ McKusick, V.A., Mendelian Inheritance in Man: Catalog of Autosomal Dominant, Autosomal Recessive and X-linked Phenotypes, 5th Edition, The Johns Hopkins University Press, Baltimore, 1978.

Medical Computer Science

The Hepatitis Knowledge Base access module supports the storage and retrieval of specified portions of the text. Yet there are many issues that must be solved. For example, how does one update such a data base, when the material to be updated is widely scattered throughout the text? Is it possible to assemble an answer de novo, from relevant sentences in different paragraphs? Is it possible to use computer systems to assist with the process of update and review with a minimum of human intervention?

The Lister Hill Center is starting a program of basic and applied research to attempt to answer these and similar questions. Relevant research areas include indexing strategies and coding systems for better indexing and representation of medical information, and data structures for storing this information; inquiry systems that use natural language processing techniques; high level computer languages to provide a programming environment for developing methods to provide access to medical information in multiple data bases and improved communications networks to ensure reliable and efficient interactions between widely distributed users of knowledge bases and the knowledge base information support system.

At present, program activities are supported by four different computers. A single dedicated computer resource, including hardware, software and personnel is being acquired to support the above areas of research and to support a mail system, and to provide a facility for experimenting with alternative methods for representing the medical content of these knowledge bases at they are constructed.

Research Issues

- Indexing strategies and coding systems to index and represent medical information
- Data structures to store the medical information
- Inquiry systems which include natural language processing techniques to index and access medical information
- Parsers and complier generators that support coding of and access to medical information
- High level computer languages such as INTERLISP, PASCAL, SAIL, AND C that provide a programming environment for developing and information support system
- Operating systems that support knowledge-based systems effectively and efficiently
- Distributed processing methods to provide access to medical information in multiple data bases
- Computer networks to ensure reliable and efficient communications between users of knowledge bases and the knowledge base information support system

In the development of the Knowledge Base Research Program the Lister Hill Center recognizes and attempts to address the need to present advances of medicine as represented in the published literature to practitioners, researchers and educators. Certainly it is our wish that practitioners incorporate new information in their daily work. The volume of papers being published is not surmountable. In addition the practitioner has a heavy schedule in patient care leaving limited time to read the literature delivered to his office or home, much less time to search for specific information in the library. Even the problem of storing journals is difficult for the professional community. Storage of information and appropriately indexing ideas in each paper is a great problem should one attempt to address it. Much valuable time is lost in scanning articles to find hidden and usable ideas and facts. Only rarely

is sufficient effort put forth to sample the literature to update one's knowledge. Therefore, having condensed information structured and indexed for easy access could well be helpful. Many of use are years behind in our reading. It is stated that practitioners have more difficulty than their academic brethren. Therefore, obtaining pinpointed information relevant to clinical problems on demand at the terminal may well result in patients benefiting from recent medical advances.

Finally, the requirement that technology be adopted by end-users as an extension of their own capabilities has proven to be a task of formidable dimensions. The man/machine interface and all that it implies, represents a genuine challenge for those of us concerned with creating tomorrow's information services today.

Without help, practitioners can cope -- but ...(S20).

Robert Niehoff submitted a copy of The Design Evaluation of a Vocabulary Switching System for use in Multi-Base Search Environments, the final report on that project delivered to the National Science Foundation in completion of NSF Grant #IST 77-04498. The report is authored by Robert Niehoff (Principal Investigator), Stan Kwasny, Ann Walker, and Mike Wessells. It can be obtained from NSF or from the Battelle Columbus Laboratories.

In addition to the description of artificial intelligence systems Roger C. Schank presented at the Workshop, the following material was also provided.

ARTIFICIAL INTELLIGENCE AND INFORMATION RETRIEVAL
SOME QUESTIONS AND ANSWERS

Roger C. Schank
Yale University

Q1: What is it that the ideal information retrieval system would be capable of doing?

A1: Ideally, a computer should really know about what it is looking for when it searches it memory.

a - We would like our retrieval programs to come up with information that was not exactly what was asked for, but nevertheless fits the bill.

b - We would like partial matches, where half an answer is better than none.

c - We would like to be able to talk to our program in natural English.

d - We would like our program to understand what we are after wll enough for it to be capable of giving us what we really wanted, rather than what we actually asked for.

Q2: How far wawy are we from all this?

A2: Recent research at the Yale Artificial Intelligence Project has been directed towards finding out how to do some of the things mentioned above. The question we have posed for ourselves is, can we design a system that knows a great deal of information about a subject that we can query in a natural way?

It is not a question of strating out with a data base of informaton that we would like to query. Rather, we must ask how a data base must be structured so that natural language communication with it can be facilitated.

Q3: How must information be structured to facilitate natural communication?

A3: It seems obvious to say it, but we cannot reasonably expect to talk to a system about the information it contains unless that system itself understands what information it contains. In other words, data must be represented in a form that expresses the meaning of data. If the meaning of the data is accessible then English questions that relate to that meaning can be entertained.

Q4: Why hasn't this kind of thing been seriously attempted before?

A4: Most research in Information Retrieval has been concentrated on document retrieval and not on problems of organizing and retrieving the information contained in those documents. Most IR systems require the user to post his questions as a complex boolean or extended boolean expression of key words for specifying documents. He must then sort through the retrieved documents himself for the information he needs. Many times the required boolean expression is so complex that only a technician with intimate knowledge about the system, i.e., somebody familiar with both its content and idiosyncrasies, can formulate that expression. These things make it extremely hard for a novice or casual user to access the system, or for even an experienced user to get the information he needs quickly. Clearly it would be advantageous to eliminate these problems by the use of natural English. But, the natural language problem being very hard, many computer workers have chosen to get around the problem by the use of key word schemes.

Q5: Why aren't key word schemes sufficient?

A5: In a system which relies on key words and thier synonyms, relevant documents in the data base will often be missed because they do not contain specified words, and many unrelated documents might be retrieved. We would like to have books and articles indexed by their content rather than by their titles.

Q6: Are there any shortcuts available that will obviate the need for full

natural language understanding systems?

A6: Yes and No. In the long run we cannot realistically expect that we can get by without full language understanding capabilities. We do not easily cope now with all the information we would like to have available to our data bases. As computers become more ubiquitous this need will grow tremendously. Word processing systems and networks are making a tremendous amount of information available in English that is in machine-readable form. Users will demand that such information be available to them.

While the long term solution must be full comprehension abilities for machines, clearly we do not yet have those capabilities. In the mean time, we have developed programs that are prototypes of the beginning of the solutions to some of these problems.

Q7: What must we be able to do to do it right?

A7: Simply put, a data base system must understand what it has read.

A7:(continued)

This involves:

a - Automatically understanding natural language text -- both input to the data base and queries to the system. This understanding involves producing a representation to the content of the natural language input.

b - Information in the system should be formatted and organized (automatically) in such a way that the conceptual content or meaning of an item can be used for retrieval rather than its key words. The representations produced by the parser should be adequate for this task, and should be automatically integrated into the memory organization.

c - The system needs rules for accessing its memory -- both for directing input of new information into the memory and for directing retrieval. Those access rules will need to use knowledge about the memory's domains of information.

d - The most significant questions according to their intention rather than according to their literal meaning must be developed.

Q8: How far away from all this are we now?

A8: We have four computer programs currently under development at Yale that illustrate our progress on different aspects of the problems mentioned above. These are:

a - FRUMP

FRUMP is a program that skims stories from the United Press International news wire and outputs a conceptual representation of the important events of each story that it understands. It is a skimmer as opposed to a full reader. It just look for what it wants to know about. The program can process text from diverse domains such as reports of plane crashes, countries establishing diplomatic ties, forest fires, and wars. New domains are added by supplying only the domain specific world knowledge; FRUMP's linguistic knowledge is independent of any particular domain.

FRUMP uses a data structure called a sketchy script to organize its knowledge about the world. Each sketchy script the respository for the knowledge FRUMP has about what can occur in a given situation. FRUMP currently has sketchy scripts for 60 different situations.

In understanding a story, FRUMP goes through two phases. First, it must elect a sketchy script. Second, it processes the story using the selected sketchy script to predict the concepts in the important events in the story.

FRUMP has a vocabulary of approximately 2300 root words. It also has morphological rules which allow it to recognize regular plurals, past participles (ed en), gerunds (ing) and nominalizations (tion) forms from

root words which extends its vocabulary considerably. It currently has about a 75% accuracy reading storeis for which is a good depth of knowledge. It is very fast and can process stories as quickly as they come over the wire. It produces summaries of the stories that it reads in a variety of languages.

b - IPP

IPP is a program that reads stories in only one domain, terrorism. It has extensive knowledge of that domain however, and uses that knowledge to understand and learn from what it reads. It has the capability of making generalizations based upon what it reads and has come to some new and surprising conclusions about terrorism in various places around the world. IPP represents . . our first attempt at a program that self-organizes its memory and learns from its experiences how to improve that organization.

c - CYRUS

The CYRUS system is a memory model that organizes biographical information about people and uses knowledge about its organization for retrieval and automatic updating. CYRUS has to modules -- a question-answering module which answers questions put to it by a human user, and an updating module which automatically adds new information to memory after that information has been pre-processed by FRUMP. The CYRUS system contains information about former U.S. Secretary of State Cyrus Vance, who was chosen as the model for the system since he is in the news often enough to generate a large number of news updates. More recently, CYRUS has begun collecting information about U.S. Secretary of State Edmund Muskie.

Because CYRUS stores episodes, its memory is organized in episodic categories called MOPs. Similar episodes are stored in the same MOP, along with the generalized knowledge built up from the similarities in the episodes. MOPs act as event categories in memory holding episodes and knowledge about those episodes. The generalized information a MOP holds includes such things as typical pre-conditions and enablement conditions for the episodes, the typical sequence of events for episodes of that class, larger episodes they are usually part of, their usual results, typical location, duration, participants, etc.

When a new story is sent to CYRUS from FRUMP, CYRUS must add the events in that story to its memory. Its first step is to make necessary inferences to decide which MOP the new events fall into. Thus, if FRUMP send CYRUS a summary such as "Vance and Gromyko met yesterday in Moscow to talk about SALT II", CYRUS infers that the meeting was a diplomatic meeting. It also fills in contextual details based on that initial categorization. For the summary above, it infers that the meeting was part of SALT II negotiations and that Vance must have been on a diplomatic trip.

After making this initial categorization and inferring contextual details, the event is added to the chosen MOP. A MOP organizes more specific MOPs and also indexes events according to their differences and variations from the norm.

In adding an event to a MOP, then, it will either be indexed into another more specific MOP where the same indexing will happen, it will be indexed uniquely, or it will fall into an index point inhabited by another event. The event above would be indexed as a "meeting with Gromyko", "meeting in Moscow", "meeting about SALT". etc.

If the new event is indexed to a point which already holds an event, the system looks at the similarities and differences between those events, makes generalizations based on that, and indexes the two events according to their differences. In that way, new more specific memory categories or MOPs are formed.

CYRUS' other module takes care of retrieval. Organization according to differences allows CYRUS to retrieve events that differ from the norm more easily than those that are normal. It also makes it possible to answer questions about normal activities without having to retrieve all or many distinct episodes. Search strategies are applied to retrieve possible related contexts which might point to an event in question. Thus, one way to find a recent meeting between Vance and Gromyko is to search for SALT negotiations (which don't happen as often and might be easier to find), and if that is found, search its sequence of events. CYRUS makes use of approximately 10 search strategies based on those we observed people using to answer questions. If a unique event still cannot be found, generalized information can be used to infer a probably answer.

d - BORIS

BORIS is a prototype of the ultimate understanding program. It reads every word of a story and follows out all the inferences it can about the motivations, personal relationships and beliefs of all the characters it finds. It has a memory model within it that it exploits to help it detect patterms in the world worth noting and reasons why various actions occur. At the moment BORIS only understands a few stories.

BIBLIOGRAPHY

1) DeJong, G.F. (1979). Prediction and Substantiation: A New Approach to Natural Language Processing. Cognitive Science, Vol. 3, no. 3.

2) Dyer, M.G. and Lehnert, W. (1980). Memory Organization and Search Processes for Narratives. Research report #175, Dept. of Computer Science, Yale University.

3) Kolodner, J.L. (1980). Organizing Memory and Keeping It Organized. In Proceedings of the First Annual National Conference on Artificial Intelligence, American Association for Artificial Intelligence, La Canada, Calif.

4) Lebowitz, M (1980). Language and Memory: Generalization as a Part of Understanding. In Proceedings of the First Annual National Conference on Artificial Intelligence, American Association for Artificial Intelligence, La Canada, Calif.

5) Lehnert, W. (1977). Human and Computational Question Answering. Cognitive Science, Vol. 1, no. 1.

6) Schank, R.C. (1980). Language and Memory. Cognitive Science, Vol. 4, no. 3.

7) Schank, R.C., Kolodner, J., and DeJong, G. (1980). Conceptual Information Retrieval. In Proceedings of Research and Development in Information Retrieval Conference, St. John's College, Cambridge, England.

8) Schank, R.C. and Abelson, R.P. (1977). Scripts, Plans, Goals, and Understanding. Lawrence Erlbaum Press, Hillsdale, N.J.

9) Schank, R.C., Lebowitz, M., and Birnbaum, L. (1980). An Integrated Understander. American Journal of Computational Linguistics, Vol. 6, no. 1.


DATE
ILME