27 April 1988

MEMORANDUM FOR THE RECORD

SUBJECT: RIAO '88 Trip Report

On March 21-24, 1988 a conference was held at M.I.T. on the theme "User-oriented Content-Based Text and Image Handling." The RIAO is an international conference organized by the Centre de Hautes Etudes Internationales d'Informatique Documentaire. It was the first conference to be devoted entirely to the field of Information Retrieval and the response was overwhelming. The 700 attendees were more than double what the sponsor expected, causing a last minute shift to larger rooms for the presentations. Attached is a complete list of all the papers delivered at the conference and each attendee was given 3 bound volumes containing the complete text of all the papers that were presented. The bibliographies alone (listed for each paper) are extremely valuable sources for the history and state-of-the-art of this subfield of computer science worldwide. It would be impossible to review all of the 75 papers presented, but I have attempted to list some of the more interesting ideas given in three different categories: Methodologies that were measured and clearly shown to work, methodologies that are currently being evaluated with no measurements available yet, and miscellaneous systems that do not fit into the other two categories.

Most of the attendees were from academia, government, or from companies such as Boeing that have the same information retrieval problems as everyone else. Conspicuously absent from the conference were all the commercial marketers of text/search retrieval software that were in force at the CD-ROM conference a few weeks earlier. At the beginning of one of the sessions, someone summed up the situation quite accurately when they said that in the beginning of the field of IR (Information Retrieval), there was a lot of theory but no applications. Now there are a lot of applications (commercial) with no regard to theory. IR research has shown clearly that Boolean type retrieval of full text databases yields only 16% of relevant documents even when queried by an expert (DIALOG, SAFE, etc). This conference demonstated that there are many methodologies that can be employed to increase this rate of recall and precision. In fact, each

methodology tends to yield a different set of valid hits, thereby pointing to solutions that employ multiple, different types of approaches simultaneously. AI techniques and knowledge bases were clearly shown to be of use in textual processing especially when they employed techniques that used feedback from users in an interative fashion to refine the results.

Category 1: Proven Methodologies

a) Attribute-based model described in "A Technique to Improve the Precision of Full-Text Database Search," by David Hsiao, Naval Postgraduate School.

b) "Towards a Friendly Adaptable Information Retrieval System," by Shih-Chio Chang, GTE Laboratories. Uses PROLOG to take relevance feedback until it meets the needs of the user.

c) "Inverted Signature Trees: An Efficient Text Searching Technique for Use with CD-ROMs," by Alan Tharp, North Carolina State University. A new data structure called inverted signature trees is introduced and compared with text signatures and the B+ tree. This new data structure facilitates rapid access to all sentences in a text file that contain specific search words while still maintaining search words in alphabetical order. Using the 12,000 page World Book Encyclopedia stored on a CD-ROM, a 4 word search using text signatures would require 3.8 minutes, whereas it would only take 6 seconds using inverted signature trees. If a word is not present in the file, the hashing procedure used to form the signatures returns this result immediately since the inverted lists need not be accessed. For the B+ tree, four seconds would be required if the entire B+ tree index is stored in memory, otherwise it would take 10 seconds.

d) "Who Knows: A system Based on Automatic Representation of Semantic Structure," by Lynn Streeter, Bell Communications Research. I found this approach (which is purely mathematical as opposed to linguistic) to be one of the most interesting of the conference. The actual application, "Who Knows" was developed so that anyone calling Bell for technical information would be directed to the appropriate internal organizations. For instance, what if someone wanted to know something about Artificial Intelligence and who was doing what. In Bell, as in our own Agency, there might be many different people/organizations working in this area. Based upon an arbitrary natural language request from a user, the system returns a ranked list of all pertinent organizations containing the who, what, where and how that a user wants to know. When Bell did a controlled test of this method which uses vector mathematics to place relevant documents in the same vector space, they found it to be more than 5 times more accurate than traditional keyword matching systems, even when the keywords were weighted. "Who Knows" is written in C and runs under a standard Unix system. A typical query takes 2-3 seconds on a Sun 3/75. The system was also able to predict, based upon an analysis of 263 new technical abstracts, the actual department

that produced each one.

(e) "An Application of Artificial Intelligence Techniques to Automated Key-Wording," by James Driscoll, Univ. of Central Florida. The JCS (Joint Chiefs of Staff) uses a computer program to store, categorize and retrieve observations/lessons learned from military exercises, wargames and real-world military operations. The system requires interraction with keyword experts for the documents to be catalogued correctly. Using Artificial Intelligence techniques to automate this manual keywording process proved not only to be more accurate than any given keyword expert's performance (as measured by a team of 5 experts collaborating together), but it also cut the indexing time per document down from 7 minutes to 15 seconds.

(f) "Looking for Needles In a Haystack or Locating Collocational Expressions in Large Textual Databases," by Yaacov Choueka, Bell Communications Research. This paper was delivered under the session titled "Automatic Thesaurus Construction." It addresses the problem of collocational expressions in generating an automatic thesaurus. A collocational expression is defined as a sequence of two or more words whose exact meaning or connotation cannot be derived directly from the meaning of its components. Examples: west bank, big board, the west coast. Using ten million words of the New York Times News Service (9 1/2 months of all the wire stories) a thesaurus was generated of collocational expressions using techniques described in the paper which contained a high degree of precision and recall.

(g) "Conceptual Information Retrieval From Full-Text," by Richard Tong, Advanced Decision Systems. This paper describes RUBRIC, a system developed for an unnamed client (obviously from the intelligence community based on the "terrorism" examples used). It's goal is to "understand" the text of documents in sufficient detail that an automatic determination can be made as to whether a user should analyze it. The system replaces the parsing of text employed by other natural language efforts with an evidential accrual and reasoning mechanism. The results are impressive: 68% precision and 68% recall based upon 730 articles taken from the Reuters newservice. ADS has two other operational systems which run on a SUN. One contains 2000 rules in the knowledge base and the average retrieval speed is between 1 and 10 seconds against a 100 MB collection of mixed documents (manuals, almanacs, and short reports).

Category 2: Methodologies Currently Being Evaluated

a) "Intelligent Search of Full-Text Databases," by Susan Gauch, Univ. of North Carolina. An expert system reformulates a user's Boolean query depending upon whether whether too many or too few hits were obtained. A hierarchical thesaurus is used as one of the techniques for expanding a user's Boolean query when too few hits are obtained.

b) "About Reformulation in Full-Text IRS (Information Retrieval Systems)," by Christian Fluhr, Centre National de la recherche Scientifique (France). An expert system is used as a reformulation strategy manager to translate a request for information from very large textual databases in unrestricted natural language. Reformulation of both the query (which is often expressed in terms different from those found in the documents that should be returned) and the database texts themselves so that they converge is accomplished via an expert system which applies various reformulation techniques such as lemmatization, misprints and spelling correctors, explicitization, syntagmatic extensions, and weighted comparisons.

c) "The TINA Project: Text Content Analysis," by Dr. Schwarz, Siemens, West Germany. The U.S. Department of Commerce U.S. Patent and Trademark Office is conducting a large scale test of TINA Natural Language programs on 195,000 patent abstracts. This software combines morphological software with syntactic analysis of free-text to aid a user in modifying his search question or in indexing a given document he may be adding to the system. Future plans include the addition of semantic processing of free-text and the integration of parallel processing techniques for the retrieval and matching algorithms.

d) "Tex-Nat: A Tool for Indexing and Information Retieval," by J.M. Lancel, CAP SOGETI INNOVATION. This tool for indexing documents (not using the traditional stop list) and retrieving them in spite of spelling errors via a phonetic equivalent is available on PC-DOS and UNIX machines and will be ported to the IBM/VM and PS\OS2 environments.

e) "Conceptual Information Extraction and Retrieval From Natural Language Input," by Lisa Rau, AI Program, GE Research and Development Center, N.Y. SCISOR (System for Conceptual Information Summarization, Organization and Retrieval) is a prototype system for taking texts in constrained domains and storing them in conceptual format which can be interrogated by natural language. The system is written in Common Lisp and processes stories at the rate of a few seconds per sentence. An on-line news service is currently being connected to enable extensive testing of the system. Other potential uses of the system are structured database generation from text and automatic summaries of complex events.

f) "Automatic Thesaurus Construction by Machine Learning from Retrieval Sessions," by U. Guntzer, Technical University of Munich. TEGEN is written in PASCAL using a knowledge-based programming method running on a CYBER 990 computer. It is is presently being used to build a thesaurus from a controlled vocabulary learnt by the computer by analyzing the searches carried out by users over a period of time based upon both implicit and explicit feedback. The system is presently being used with a collection of 130,000 documents of the Faculty of

Mathematics and Computer Science. Results of a manual analysis of the thesaurus entries generated will be reported on at a later date.

g) "A Logic Programming Approach to Full-Text Database Manipulation," by R. Marshall, Loyola College. NASA's ENVIRONET database was captured in the form of facts and rules using first order predicate logic (PROLOG) to facilitate user's retrieval of information in text, mathematical formula-based numerical extraction, interpolation and extrapolation, and tables and graphs. Also, the system understands the difference between an aerodynamics expert or a computer engineer retrieving information about "RAM". The system runs on a MicroVAX II at Goddard Space Flight Center in Greenbelt, MD.

h) "Interactive Knowledge-Based Indexing: The MedIndEx System," by Susanne Humphrey, National Library of Medicine. This paper describes the MedIndEx System which uses a knowledge base to provide interactive assistance in indexing the periodical medical literature. The system is in prototype form and the results of evaluation programs are not yet available.

Category 3: Miscellaneous Topics

a) "Structure of Information in Full-Text Abstracts," by Elizabeth Liddy, Syracuse University. Proves that discourse linguistics can be used to extract a structure from free-text documents such as scientific abstracts, thereby making it possible for these documents to be handled successfully in information retrieval systems. The automatic detection and use of this structure is being implemented on the Connection Machine at Syracuse University.

b) "Using English for Indexing and Retrieving," by Boris Katz, AI Laboratory, M.I.T. START (Syntactic Analysis using Reversible Transformations) is an attempt to understand English text and index the knowledge contained within it (for use in generating an AI knowledge base) and also to generate English text from retrieved portions of the knowledge base.

c) "Information Aids for Technological Decision-Making: New Data Processing and Interrogation Techniques for Full-Text Patent Databases," by W.A. Turner, CDST/ CNRS, France. This technical alert system was developed by CNRS to provide French Industry with a competitive edge in international markets. It exploits the data in full-text patent databases. Its premise is that pre-established classification schemes are not useful for describing the scientific and technological state of the art at any given moment. Through co-word analysis techniques they have been able to detect subject areas in which a statistically significant number of patents have been filed during a given period of time. The user is able to obtain activity profiles for any research topic (i.e., conductive thin film) which shows, by country, whether or not that country's "innovative activity" is weaker or stronger than

expected.

d) "Browsing and Authoring Tools for a Unified Medical Language System," by Henry Kamorowski, National Library of Medicine. Because available commercial knowledge engineering environments have only rudimentary display and browsing tools, IntelliCorp's KEE and Interlisp-D were combined with their own set of lower level layout and display routines to provide a "fish-eye view" of a semantic net. BATS (Browsing and Authoring Tools for Semantic networks) was written so that it could easily be ported to another environment, i.e., the MAC II.

e) "The Informatics Calculus: A Graphical Functional Query Language for Information Resources," by Richard Epstein, West Chester University, PA. A doctoral thesis which describes the design of a query language which uses informatics calculus to meld traditional database and hypertext technologies.

f) "Implementing a Distributed Expert-Based Information Retrieval System," by Edward Fox, VPI, VA. For four years, Fox et al at VPI have been developing a distributed expert-based information system written in C and PROLOG called CODER. Various modules (Query Formulator, Search, Problem Description Builder, Browse, Lexical) work together under a Retrieval Strategist (Planning and Coordination Heuristics) to process a user's request at various levels simultaneously. Future plans include an evaluation of 3 different approaches to accessing data on a CD-ROM: TOPIC (based on RUBRIC) is an AI type of retrieval, Personal Librarian which is based on statistical algorithms to rank output, and Window Book, a hypertext system.

g) "Universal Multilingual Information Interchange System," by Suban Krishnamoorthy, Framingham State College, MA. Once in awhile some unknown person from some unknown place comes up with a truly revolutionary concept which largely goes unnoticed for a number of years. Such is my impression of this paper which was given by this native Indian (who had to cope with several of the fifteen different Indian languages as he moved around India as a student together with several European languages plus English as a graduate student and professor). He contends that a unified solution to handle and interchange information in any (variable) number of scripts and languages does not exist. He then describes a complete design for just such a universal multilingual communication system down to the minutest detail, including character sets, language coding, I/O devices, multilinual keyboard design, font design, etc. He concludes by stating that the design of his system is simple, practical and viable now and that a universal computer communication system (CCS) such as the one he shows (with no linguistic barriers or boundaries) will be the future CCSs of the world and have enormous social impact.

h) "Automatic Recognition of Sentence Dependency Structures," by Timothy Craven, University of Western Ontario. A system for creating automatic abstracts from documents. Based on results, it

appears to have more immediate utility as a means for extracting structured information from long texts for use in online databases.

i) "The Utah Retrieval System Architecture (URSA)," by Lee Hollaar, University of Utah. Using a distributed information retrieval system using Apollo workstations, windows, and specialized backend processors, a 40 GB database with 120 parallel searches and an aggregate search speed of 150 MB/second is currently being implemented and will cost about $2-3 million. (There were a number of other papers describing experimental hardware architecture solutions such as VLSI , neural networks and parallel processing.)