A.T. McCray

Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Md, USA

Research and Education

Medical Informatics Research and Training at the Lister Hill National Center for Biomedical Communications

Introduction

The Lister Hill National Center for Biomedical Communications was established by a joint resolution of the United States Congress in 1968 as a research and development division of the National Library of Medicine (NLM). Lister Hill Center research is carried out through several major programs, all sharing the purpose of improving health-care information dissemination and use. We conduct our research by drawing on a diverse set of scientific fields and methods (see [1-25] for a sample of our recent publications.) Our researchers have backgrounds in medicine, computer science, library and information science, linguistics, engineering, and education. We conduct both basic and applied informatics research and are engaged in a number of broad research areas. Knowledge processing research includes language and information processing and expert systems research. Information systems research includes consumer health informatics, database systems, digital library research, and medical education systems. Image processing research includes image segmentation, compression, and transmission methods and algorithms. We are also the focal point for NLM's high performance computing activities, including research support for telemedicine and health applications for the Next Generation Internet. We offer a range of medical informatics training opportunities to individuals who would like to work with our research staff on projects of mutual interest. The most current information about Lister Hill Center activities can be found at our Web site http://lhncbc.nlm.nih.gov/.

Knowledge Processing

Unified Medical Language System

The Unified Medical Language System (UMLS) project conducts research to improve the retrieval and integration of biomedical information. The UMLS project continues to develop knowledge sources that can be used by a wide variety of application programs to overcome retrieval problems caused by differences in the usage of biomedical terminology across a wide range of information resources. The Metathesaurus is the largest of the UMLS knowledge sources, representing multiple biomedical vocabularies organized as concepts in a common format. The 1999 version of the Metathesaurus interrelates over 600,000 concepts from some 40 vocabularies. The SPECIALIST lexicon is a large syntactic lexicon of medical terminology, currently containing some 125,000 lexical items. The lexicon is accompanied by a set of lexical tools that are designed to help users abstract away from linguistic variations such as singular and plural forms of the same word, differences in punctuation and word order, and spelling variation. The programs offer methods that can combine these functions to produce indexes with varying degrees of aggressiveness for matching terms. The Semantic Network provides a consistent categorization of all concepts represented in the Metathe-saurus. The current network contains 134 semantic types and 54 relationships, with major groupings for organisms, anatomic structures, biologic function, chemicals, events, physical objects, and concepts or ideas. Relationships fall into several groups; physical, temporal, spatial, functional, and conceptual relationships are all included.

The UMLS Knowledge Source Server is an evolving tool for providing Internet access to the information stored in the UMLS Knowledge Sources. The purpose of the Knowledge Source Server is to make the UMLS data more accessible to users. Access to the system is provided through a command-line interface, through an Application Programming Interface (API), and through the World Wide Web. The Web interface allows users to browse and explore the data and to see how those data are organized in the UMLS. Information about Metathesaurus concepts, semantic types and relations in the Semantic Network, and lexical items in the SPECIALIST lexicon can be found by querying the system. The command-line interface is best suited for batch processing. Researchers can submit a list of terms to the server and retrieve a range of information about those terms, and they can further filter the results, limiting the result set, for example, by a particular attribute. The API allows developers at remote sites to embed calls in their application programs to the Knowledge Source Server, thereby accessing the UMLS data directly over the Internet.

Medical Language Processing

A text analysis system based on the SPECIALIST lexicon and the lexical tools is under development. The system is modular in design to allow for flexible use and continuous revision. The modules are servers which will be available to a variety of clients for a variety of uses. The system consists of the following modules: a tokenizer module to analyze text into tokens and label them; a sentence identification module to analyze text into sentences; a lexical look-up module to find lexical items in the text, a shapes module to identify items in the text that do not occur in the lexicon but have types recognizable from their form, and a parser module to assign phrase structure to the sentences of the text. The design of the parser module is strongly based on the syntactic information encoded in the SPECIALIST lexicon. After consideration of several grammar paradigms, Categorial Grammar was chosen as a grammatical system that expresses this lexically based approach to sentence structure.

Indexing Initiative

The indexing initiative project is investigating methods whereby automated indexing methods may partially or completely substitute for expert indexing of the biomedical literature by humans. The project is investigating concept-based indexing methods that go well beyond automatic word-based indexing. One method seeks to discover semantic relationships between phrases in text as a way of more accurately representing content. Another identifies Metathesaurus concepts in biomedical text and then maps these through a weighting algorithm based on linguistic and knowledge-based techniques to appropriate MeSH terms. One set of experiments has demonstrated the value of such concepts for automatic query expansion. Another investigation provides an enhanced representation of semantic content by ranking concepts assigned to MEDLINE abstracts on the basis of frequency of occurrence and specificity as measured by hierarchical depth in MeSH.

Expert Systems Research

We have developed expert systems to investigate issues in knowledge representation and knowledge base structure, knowledge acquisition and knowledge base maintenance, the evaluation of knowledge-based systems, and the creation and delivery of knowledge-based systems over the Internet. We have built a multimedia expert system shell called CTX, for Criteria Table Expert. The shell includes a knowledge base compiler, a run-time system that can access information from multiple knowledge sources, a knowledge base editor, a case editor and a suite of automated testing programs for analyzing system performance against sets of benchmark test cases.

Information Systems

Consumer Health Informatics

Consumer health informatics is an important new research area for us. Increasingly people are turning to the Internet to look for answers to their health questions. This raises a number of research questions, including the type of content that should be created and how that content can be put into the appropriate medical context. As a result of recent legislation which required that the NIH create a database of clinical trials information, our research team has initiated the development of such a system. The database will be an extensive resource that provides patients, families, and members of the public with easy Web-based access to information about clinical trials, including information about which clinical trials are currently recruiting patients, where the trials are being conducted, what the design and purpose of the research study is, and what the criteria are for participating. An important feature of the database will be to provide access to other online health resources that help place clinical trials in the context of patients' overall medical care.

Database Systems

Internet Grateful Med is an intelligent gateway system designed to provide assisted Web-based searching across multiple NLM database systems. Users can search by subject, project, institution, author name, text word in title, and other parameters appropriate to each database. Results can be downloaded to disk for later viewing, manipulation or loading into a reference manager program. The gateway architecture has proven a successful means of transparently connecting users to several different types of retrieval systems while insulating them from the specifics of differing command languages.

Our research in federated databases involves the investigation of issues in concurrent searching of multiple independent databases distributed at different locations on the World Wide Web. There are issues of retrieval performance as the amount of information on the Web scales upward. There is the issue of semantic interoperability; that is, of performing retrievals that span differences in the representation of concepts across different information collections. There are also issues inherent in creating and maintaining collections of specialized information that require substantial domain-specific knowledge.

The HSTAT system is a Web-based resource that provides access to the full-text of documents useful in health care decision making. The system includes clinical practice guidelines, quick-reference guides for clinicians, consumer brochures, technology assessment and evidence reports, consensus development conference reports, HIV and AIDS resource documents; and substance abuse treatment improvement protocols.

Digital Library Research

Digital library research involves all aspects of creating and disseminating digital collections including proposed and adopted standards, emerging technologies and formats, copyright and legal issues, protection of original materials, and permanent archival of digital surrogates. Research issues that need to be addressed are long-term preservation of digital archives, innovative methods for creating and accessing digital library collections, the development of modular and open information environments, investigation of the role of well-structured metadata, and the exploration of different "points of view" on the same underlying data set.

In the fall of 1998 the *Profiles in Science* Web site was launched. This digital library site is designed for scientists, scholars, and students, all of whom may gain an appreciation of the history of early scientific discoveries, and share in the excitement of the scientific enterprise. The collections have been donated to the NLM and contain published and unpublished materials, including books, journal volumes, pamphlets, diaries, letters, manuscripts, photographs, audio tapes and other audiovisual materials. The first collection featured on the site is a special collection of materials relating to the work of Oswald Avery (1877-1955), one of this country's first molecular biologists, whose findings proved that the genetic material is DNA. The second collection represents the papers of Joshua Lederberg (b. 1925), an American geneticist and microbiologist who received the Nobel prize in 1958 for his work in bacterial genetics.

The Medical Article Record System (MARS) is a system that achieves a degree of automation in the entry of citation data from medical journal articles for the MEDLINE database. This data entry has traditionally been done by manual keyboarding. In early 1996 the MARS team developed the first generation of a system that combines the keyboarding of citation data (journal name, date, author, title, affiliation, page numbers, etc.) with scanning and automatic text conversion by optical character recognition of abstracts. The first generation system, MARS-1, consists of about two dozen workstations of three types: scanner, reconciling, and keyboarded citation entry. While MARS-1 is in routine production, the team is conducting research toward more comprehensive automation. The design of a database-driven next generation MARS-2 has begun.

The DocView project involves research in advanced technologies for document delivery over the Internet. The documents may already exist in digital form or be scanned and sent over the Internet by an Ariel system, a system which is widely used for interlibrary loan services. While libraries and document suppliers use Ariel routinely to send documents via the Internet to similar organizations, there are few options for end users to receive them directly. The DocView 1.0 client software, which runs under any version of Microsoft Windows, enables an end user to receive documents over the Internet at the desktop, retain them in electronic form, view the images, organize the received documents into "folders" and "file cabinets", electronically bookmark selected pages, manipulate the images (zoom, pan, scroll), copy and paste images, and print them if desired. DocView also serves as a TIFF viewer for compressed images received through the Internet by other means, such as World Wide Web browsers.

Medical Education

In collaboration with other groups, we have created a variety of programs for use in educational settings. Two recent examples are the cervical cancer instructional program and the motion disorders video. The cervical cancer system is designed to promote the early detection of cancer and incorporates text, graphics, still pictures and video clips. The system was designed together with content experts from the National Cancer Institute. In conjunction with the Yale University School of Medicine, we have produced a motion disorders video for educational purposes. Patients with Parkinson's Syndrome were videotaped performing standard, diagnostic routines including hand and foot exercises, walking and reading. High-end Sony digital Betacam cameras were used to guarantee the highest quality images. The videotape was edited in the AVID nonlinear editing suite, combining the two camera video into a dual screen, real-time presentation.

Image Processing

Visible Human Project

The Visible Human Project data sets are designed to serve as a common reference point for the study of human anatomy, as a set of common public domain data for testing medical imaging algorithms, and as a test-bed and model for the construction of image libraries that can be accessed through networks. The Visible Human data includes a complete human male and female cadaver in MRI, CT and anatomical modes. There are 1871 cross-sections for each mode, CT and anatomy, obtained from the male cadaver. The data set from the female cadaver has the same characteristics as the male cadaver with one exception. The axial anatomical images were obtained at 0.33 mm intervals instead of the 1.0 mm intervals for the male, resulting in over 5,000 anatomical images. Image files are stored in a compressed format in directory structures for the male and female images. Online demand for the data has remained high since its availability. Now that the data collection phase of the Visible Human Project is completed, a second phase has begun - the segmentation, classification, and three-dimensional rendering of the data set. A new research effort is underway whose ultimate objective is to identify all anatomical structures within the Visible Human data set. As a first experiment, each object in each cross-section of the male thorax will be labeled, and the relationship of each object to the other objects in its cross-section and in the adjacent cross-sections will be catalogued. We are also investigating compression and transmission techniques to improve access to, and delivery of, data-intensive biomedical images, with specific focus on the Visible Human color image set. These datasets strain both storage and transmission resources, and research has been done toward the development of prototype lossy and lossless compression techniques. The eventual goal is to design a system combining both techniques so that storage is achieved losslessly, and data to be delivered to a user is compressed lossily, at a quality level required by the user.

Digital X-rays

In collaboration with the National Center for Health Statistics and the National Institute of Arthritis, Musculoskeletal and Skin Diseases, we are investigating fundamental questions that arise in the handling, organization, storage, access and transmission of very large digitized x-rays. The x-rays, consisting of about 17,000 cervical and lumbar spine films, were collected during the second National Health and Nutrition Examination Survey, one of a series of nationwide surveys designed to provide a snapshot of the nation's health. As films they are relatively inaccessible, a major motivation for digitizing them. The project involves the design, development and evaluation of prototype systems which serve as test-beds to investigate image compression techniques, especially high yield lossy methods, and tools to interactively select compression parameters; techniques to organize images and associated textual data for ready retrieval and use; procedures and algorithms to implement transparent hierarchical storage using heterogeneous storage systems and media to match usage patterns; and multi socket transmission methods to segment large images and to send the pieces concurrently over multiple socket pairs to overcome the inefficiencies of conventional transmission.

Advanced Medical Imaging Tools

Consistent with the increasing trend for medical information data banks to incorporate images, tools must be available to enable users to search and retrieve such data easily over the Internet, and to evaluate the returned image data against "gold standard" or reference images. Our current focus is the building of two tools, one as a medical image reference aid, and the other a multimedia database access tool. The first tool is a platform-independent digital radiological atlas of the cervical and lumbar spine building on prior work involving digitized spine x-rays. The approach is to use Java software technology to create the atlas whose images will be displayable on conventional monitors as well as high resolution Megascan monitors. The second tool is an advanced Web-enabled medical image database tool for searching and retrieving the contents of biomedical databases containing both text and images.

High-Performance Computing

Telemedicine

The growth of the Internet and the increasing access to high-speed computers and communications by consumers, health care providers, public health professionals, and basic, clinical, and health services researchers is having a fundamental effect on health and human services throughout the world. Major research issues include the impact of telemedicine on the health care system as a whole and on cost, quality, and access to care for specific populations; the benefits of integrated access to practice guidelines, expert systems, bibliographic databases, electronic publications, and other knowledge-based information from within computer-based patient record systems and other automated systems that support research and practice; the maintenance of patient confidentiality as increasing amounts of electronic health data are transmitted via telecommunications during health care and aggregated for important public health and research purposes; and the development of data standards and uniform practices for effective transmission, aggregation, and integration of health care, public health, and research data.

Next Generation Internet

A three phase effort to support test-bed projects that demonstrate the use of Next Generation Internet (NGI) capabilities by the health care community has been designed. Funded projects should improve our understanding of the impact of NGI capabilities on health care, health education, and health research systems in such areas as cost, quality, usability, efficacy and security. Phase 1, which began in late 1998, is a nine-month planning effort. Each plan must identify the relevant outcomes, processes and cost variables and present a strategy for their measurement. Phase 2, a two-year effort, will support the implementation of these plans within a limited geographic scope. Phase 3, a two-year effort, will test the scalability of phase 2 NGI projects to a national scope.

Medical Informatics Training

We established a formal program in Medical Informatics Training in 1996. The program provides trainees at various stages of their careers an opportunity to participate in collaborative research on our campus. Trainees have the opportunity to work closely with our staff in any of the research areas in which we are engaged. They have access to our on-site resources, facilities, and staff. Trainees have the opportunity to make significant contributions to our research programs, while at the same time being able to pursue their own research interests. Some recent projects of our trainees have involved image processing for extraction of quantitative information, speech recognition for image-based systems, development of a palm-top computer-based medical reference manager, a comparison of patient attribute databases using the UMLS for concept mapping, the role of negation in medical language processing, patient feedback for asthma control through data entry, and vector-based rankings of information retrieval collections.

References

1. Ackerman MJ. The Visible Human Project : A resource for anatomical visualization. J Am Int Health Council. Medical Updates on Therapy, Diagnosis, and Prevention 1997; 1:58-61.
2. Aronson AR, Rindflesch TC. Query expansion using the UMLS metathesaurus. J Am Med Inform Assoc 1997;4(suppl):485-9.
3. Athreya, BH, Cheh, ML, Kingsland, LC III. Computer-assisted diagnosis of pediatric rheumatic diseases. Pediatrics 1998;102:48.
4. Bean CA, Rindflesch TC, Sneiderman CA. Automatic semantic interpretation of anatomic spatial relationships in clinical text. J Am Med Inform Assoc 1998;5 (suppl):871-901.
5. Bodenreider O, Burgun A, Botti G, Fieschi M, Le Beux P, Kohler F. Evaluation of the Unified Medical Language System as a medical knowledge source. J Am Med Inform Assoc 1998; 5:76-87.
6. Cesnik B, McCray AT, Scherrer J-R, eds. *Proceedings Medinfo '98*. Amsterdam: IOS Press, 1998 (1359 pages).
7. Divita G, Browne AC, Rindflesch TC. Evaluating lexical variant generation to improve information retrieval. J Am Med Inform Assoc 1998;5(suppl):775-9.
8. Hauser SE, Browne AC, Thoma GR and McCray AT. Lexicon assistance reduces manual verification of OCR output. In: *Proceedings Eleventh IEEE Symposium on Computer-Based Medical Systems*. Los Alamitos, CA: IEEE Computer Society, 1998:90-5.
9. Humphrey SM. A new approach to automatic indexing using journal descriptors. In: Preston CM, ed. *Proceedings 61st ASIS Meeting*. Medford, NJ: Learned Information, Inc., 1998:496-500.
10. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: Report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc 1997;4:484-500.
11. Khan, K, Locatis, C. Searching through cyberspace: Effects of link cues and correspondence on information retrieval from hypertext on the World Wide Web. J Am Soc Inform Sci 1998;49:1248-53.
12. Locatis C., Weisberg, M. Distributed learning and the Internet. Contemp Educ 1997;68:100-3.
13. Long LR, Goh G-H, Neve L, Thoma GR. Architecture for biomedical multimedia information delivery on the World Wide Web. SPIE 1997;3229:160-9.
14. Long LR, Goh G-H, Thoma GR. Online digital X-ray atlas as a reference tool. Int J Digit Libr 1997;1:220-30.
15. McCray AT, Browne AC. Discovering the modifiers in a terminology data set. J Am Med Inform Assoc 1998;5(suppl):780-4.
16. McCray AT. Conceptual complexity in biomedical terminologies: The UMLS approach. Classification and knowledge organization. Berlin, Springer Verlag; 1997:475-89.
17. McCray AT. The nature of lexical knowledge. Meth Inform Med 1998;37:353-60.
18. Meadows S, Thoma GR, Long LR, Mitra S. Entropy encoding of difference images from adjacent Visible Human digital color photographic slices for lossless compression. SPIE 1997;3031:749-55.
19. Parascandola J. Alice Evans: An early woman scientist at NIH. Public Health Rep 1998; 113: 472-4.
20. Parascandola J. Doctors at the gate: PHS at Ellis Island. Public Health Rep 1998;113:83-86.
21. Sneiderman CA, Rindflesch TC, Bean CA. Identification of anatomical terminology in medical text. J Am Med Inform Assoc 1998;5(suppl):428-32.
22. Snyder L. Integrating American Indians and Alaska Natives into the Body Politic. Public Health Rep 1998;113:365-8.
23. Thoma GR, Long LR. Compressing and transmitting visible human images. IEEE Multimedia 1997;4:36-45.
24. Walker FL, Thoma GR. Internet document delivery: An end user survey. In: *Proceedings IOLS'97*. Medford, NJ: Information Today: 1997, 145-53.
25. Van Bemmel JH, McCray AT, eds. *Yearbook of Medical Informatics 98. Health Informatics and the Internet.* Stuttgart/New York: Schattauer Verlag, 1998 (534 pages).

Address of the author:
A.T. McCray,
Lister Hill National Center for Biomedical Communications,
National Library of Medicine,
8600 Rockville Pike
Bethesda, Md 20894, USA
e-mail: mccray@nlm.nih.gov