Information Retrieval Using UMLS-based Structured Queries
Lawrence M. Fagan, MD, PhD, Daniel C. Berrios*, MD, MPH, Albert Chan†, MD, Russell Cucina, MD, Anupam Datta, Maulik Shah, Sujith Surendran
Stanford University, Stanford, CA, *Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, †Family Practice Program, San Jose Medical Center, San Jose, CA

Abstract
During the last three years, we have developed and described components of ELBook, a semantically based information-retrieval system [1-4]. Using these components, domain experts can specify a query model, indexers can use the query model to index documents, and end-users can search these documents for instances of indexed queries.

Project Goals and Design
The goal of our project is to make it easier to find crucial information in online medical resources. As more full text material becomes available on the World Wide Web, there is an increased need to develop better search mechanisms. In time-critical medical situations, clinicians need to be able to pinpoint answers to diagnostic and treatment questions. We have developed novel methods for indexing and searching medical textbooks. The methods work by describing questions that the material was written to answer. An example question is “What antimicrobial agent can be used in the treatment of organism <E.Coli> in patients with infection <meningitis> with underlying disease <renal failure> or special condition <allergy to penicillin>?”. Our goal was to associate these questions with passages of textbooks (i.e., index documents with these questions), and then retrieve appropriate passages depending on specific questions that users pose to ELBook.

We first built a specialized search engine that matches the questions asked by the user to indexed questions in documents. Next, we developed ISAID, an indexing tool that is able to suggest which questions are being answered by specific portions of the text of documents. Users of this tool can fill in appropriate values for concepts such as organism or infection in such questions. The indexing process is semi-automated because of the difficulties of computer-based natural language understanding: a human is required to review automatically generated, provisional markup for correctness and completeness. We have reported the performance of knowledge-based natural language processing techniques to identify domain concepts and relations in full text sources [3,4]. These methods used the UMLS to correctly identify semantic types (e.g., amoxicillin is a drug) for about 2/3 of the medical concepts in the medical textbook chapters we have tested.

In addition to the specialized search engine and ISAID, we developed QueryEditor, a knowledge-acquisition tool that creates and maintains the description of the queries. Once the queries are described, the user can generate the templates for the indexing tool and the HTML queries for the search engine. Using the query editor, we have described a set of core medical questions that can be used as the basis for deriving question sets for more specialized areas of medicine, such as infectious disease, patient education, or military medicine.

Acknowledgements
Supported in part by Contract DA08-00-P1735 from the Center for Total Access, Fort Gordon, GA, U.S. Army. Maulik Shah and Sujith Surendran are recipients of scholarships from Harcourt, Inc.

Bibliography