Adaptation of Meta-1 for SAPHIRE, A General Purpose Information Retrieval System

William R. Hersh, M.D.
Biomedical Information Communications Center
Oregon Health Sciences U.
Portland, OR

Edward Pattison-Gordon, M.S.
Robert A. Greenes, M.D., Ph.D.
Decision Systems Group
Brigham and Women's Hospital
Boston, MA

David A. Evans, Ph.D.
Laboratory for Computational Linguistics
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA

Abstract

The Unified Medical Language Systems Project (UMLS) of the National Library of Medicine (NLM) has produced Meta-1, a metathesaurus featuring over 40,000 concepts and their synonyms from several commonly-used medical vocabularies. We have adapted Meta-1 for use in SAPHIRE, an information retrieval system featuring automated indexing and probabilistic retrieval. We have also built DESYGNS, a semantic network system designed to contain Meta-1 concepts along with their semantic relationships. Future plans include improved concept matching, improved indexing capability, and the use of semantic relationships.

Introduction

One of the major obstacles to the acceptance of computerized information systems has been the existence of multiple medical vocabularies, each incompatible with the next. As a result, the vocabulary used for indexing and retrieving one facet of medical information, such as pathology reports, is incompatible with one used for indexing and retrieving other facets, such as the online literature. For example, the MeSH [1] term *Scleroderma, Systemic* is a different string form than its SNOMED [2] synonym *Generalized Scleroderma*. Some online databases which feature much useful information beyond MEDLINE (National Library of Medicine) references, such as BIOSIS (Biological Abstracts, Inc.), have vocabularies with different term names as well as instructions for use of those terms for indexing. With the proliferation of textual databases, hypermedia collections, and expert systems, this incompatibility of medical vocabularies has hindered the usability and acceptance of these systems.

This problem motivated the National Library of Medicine (NLM) to undertake the Unified Medical Language Systems (UMLS) Project in 1986 [3]. One of the goals of this project is the creation of a medical metathesaurus, allowing translation of terms between different medical vocabularies. The first version of the metathesaurus, Meta-1, will become available in 1990. Each Meta-1 concept will feature a canonical form along with synonyms from a number of different vocabularies. There will be just over 40,000 concepts in Meta-1, consisting of all MeSH terms, all DSM-III diagnoses [4], and a large number of clinical manifestations from COSTAR [5]. Meta-1 will also feature semantic typing of each term as well as a semantic network of generic relationships between the semantic types, both of which will have potential use in disambiguating queries [6].

In our work we have used a prerelease version of Meta-1 in SAPHIRE (Semantic and Probabilistic Heuristic Information Retrieval Environment), a general purpose information retrieval system which features automated recognition of concepts for indexing and retrieval [7]. The foundation of SAPHIRE is a Concept-Matching Algorithm, which processes text strings to find concepts and map them into a semantic network structure. The Concept-Matching Algorithm is used both in designation of concepts for indexing and mapping of queries to canonical concepts for retrieval. The robustness of the Concept-Matching Algorithm is dependent upon a large vocabulary with a wide variety of synonyms, which is what Meta-1 offers.

In this paper, we first describe the field of Information Retrieval (IR), the SAPHIRE Project, and DESYGNS. We next describe the adaptation of a portion of Meta-1 for use by SAPHIRE in the domain of AIDS. This is followed by a discussion of our initial results and plans for future work.

Information Retrieval

While the rapid commercial growth of conventional IR systems shows that these systems provide useful functionality to end-users, there are a number of limitations of these systems. One major problem is the large number of indexing vocabularies for medical databases, as described above. An additional limitation is the human indexing process, which can be inconsistent. Funk and Reid [8] performed a study of MEDLINE indexing consistency, noting that it varied from 61.1% in the case of central-concept main headings to 33.8% for main heading/subheading combinations. The limitations of keyword systems, particularly the expense and inconsistency of human indexing, have led to research into IR systems that perform automatic indexing and term weighting. A pioneer in this area has been Salton [9], who introduced the notion of the *term discrimination value*, a measure of how well a given word helped to discriminate documents from each other. One early quantity used for term discrimination was the *inverse document frequency*, which was a measure of how infrequently a word occurred in a document collection. The underlying assumption was that removing such a word from a document would make the documents more similar. A practical implementation of this approach is the IRX system, designed at the NLM [10].

Another approach to automated IR is through computational linguistic methods, where content is indexed and retrieved based on *concepts* instead of *terms*. The difference between these is that terms are just surface string representations of concepts. IR systems based on either controlled vocabularies or probabilistic values from word frequencies thus suffer from a deficiency in that their indexing and retrieval operations are based upon terms, which may be varying surface form representations of the same underlying

0195-4210/90/0000/0156$01.00 © 1990 SCAMC, Inc.

concept. A problem with most existing linguistic IR systems, however, is slow performance and limited domains. Evans has investigated overcoming this problem through the use of restricted linguistic methodology in the CLARIT Project [11]. Abandoning the goal of complete syntactic and semantic understanding of text, the CLARIT system aims only to use selected amounts of syntax and semantics in order to recognize concepts for indexing. The main function of the CLARIT parser is to recognize noun phrases, which represent the concepts in text.

SAPHIRE

SAPHIRE is an IR system featuring concept-based automated indexing and retrieval [7]. It shares with linguistic systems the ability to identify and search for content based on concepts, and features innovations along two lines. The first is a Concept-Matching Algorithm that processes free text to (1) convert synonyms to canonical form and (2) extract the underlying concepts. The second is the adaptation of automatic indexing and probabilistic retrieval methods to a system based on concepts rather than one based on words. In addition to concept-based searching, the user can also browse the hierarchically-organized vocabulary to add parent or child terms to the search list. The program is a prototype environment that allows indexing and retrieval of record-based information resources. These records can be hypertext nodes, article abstracts, or any other type of resource where the information is broken into records containing a body and a title. The content used for evaluation of SAPHIRE to date has been a textbook chapter on AIDS [12] from *Scientific American Medicine*, which had been converted to hypertext format for the Explorer-1 Project [13], and a collection of 5,727 MEDLINE references from 1984 to 1989 downloaded from a CD-ROM product featuring Abridged Index Medicus (Knowledge Finder, Aries Corp.).

SAPHIRE's Concept-Matching Algorithm is based on an enhanced version of an approach originated by Shoval [14], which used a *discrimination network* to find concepts in the user's query for retrieval tasks. Shoval's algorithm is extended in SAPHIRE to allow handling of a wider variety of synonyms. This is done by designating *word level synonyms*, which substitute individual words that make up a concept, and *concept level synonyms*, which substitute one surface form of a concept for another. Examples of word level synonyms are the words *cancer* and *carcinoma* in *carcinoma of the colon* and *cancer of the colon*. An example of concept level synonyms are *leukocyte* and *white blood cell*. Furthermore, a stemming algorithm is used to remove common suffixes and plurals from words [15].

For indexing, an automated approach is used whereby content for each "document" (where a document can be a journal article, abstract, or hypertext node) is processed one sentence at a time. All concepts identified are then given a weight designating their value as an indexing concept. For concept weighting in our automatic indexing scheme, we adapted the *inverse document frequency*, which assigns weights to concept inversely proportional to how many documents in which they appear [9].

We also attempt to draw on probabilistic retrieval techniques designed for automated indexing approaches, modeled after the methods used by word-based systems such as IRX. One such method is a non-Boolean method of search specification, where the user just specifies a list of search concepts. This is motivated by the observations of Borgman, who noted that Boolean query formulation tends to confuse novices and lead to searching errors [16]. Another method we use that is drawn from probabilistic IR research is a ranking algorithm for matched content. To formulate a search with SAPHIRE, the user enters a free text query and receives back a list of concepts in the query. The user can modify this list by deleting concepts or adding new concepts, either via typing them into the text box or by selecting existing concepts and having children or parent concepts from the hierarchy added. After the concepts for searching are chosen, the search is initiated. For each document in which at least one concept occurred, a score is calculated summing the term weights for each concept that occurs in both the document and the query. The scores for all documents are ranked and presented to the user. The result of this process is to give documents with the largest number of high weight concepts the topmost ranking in the retrieval.

The SAPHIRE software consists of two parts, the engine and the interface. The engine is written in ANSI C. Although the program currently runs only on the Apple Macintosh, the engine is written with the future goal of porting to other environments. The concept-matching algorithm and retrieval are very fast, with the system able to take a typical sentence-length query, find its matching concepts, and find and rank its matching documents in under three seconds on a Macintosh II.

The interface portion of the program is built with Supercard (Silicon Beach Software, Inc.). Supercard is used only as a graphical front end for the underlying C-based engine. Both programs run in tandem under MultiFinder, with communications between each handled by using text files. The time cost for the task-switching is about one second, leaving performance still reasonably fast. When System 7.0 for the Macintosh becomes available next year, the improved interprocess communications facilities may allow us to eliminate this minor overhead. Figure 1 shows a screen display of the SAPHIRE Query Window.

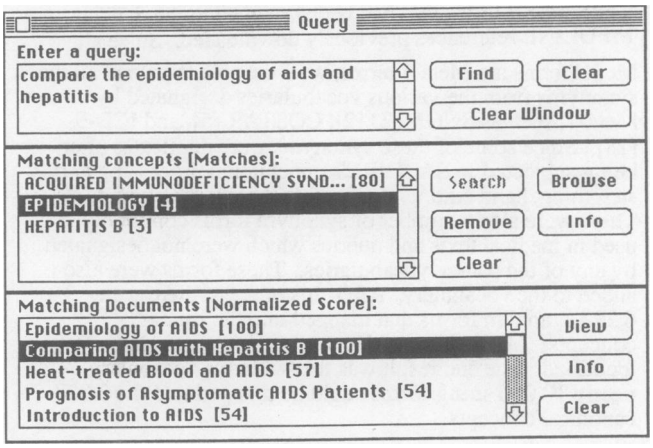

Figure 1 -- SAPHIRE Query Window

DESYGNS

DESYGNS is a library of software routines for building and accessing semantic networks used by SAPHIRE and other knowledge management tools from the Decision Systems Group, such as Explorer-2 [17]. A semantic network is a method for organizing knowledge about the terminology used in a domain of interest. The nodes of the network are concepts. The network's connections specify relationships between concepts and, because they are labeled, they also

indicate the nature of that relation.

A special relation, generally known as *IS-A*, is used to organize concepts into *classes*. A class is a group of things that are lumped together because they share some common feature or property. When a class concept is added to a semantic network, the concepts that are members of the class are connected to it with the *IS-A* relation. Concepts may not be related to one another arbitrarily. Instead, a concept's class determines which relations may be used and what concepts are available for a given relation. In DESYGNS, each relation in the network has a table that defines its usage. The first column is a list of the classes whose members can use the relation; each row in the second column lists the classes with concepts to which a concept from the class in the first column may be related. Using classes allows relations to be defined very succinctly.

A unique feature of the DESYGNS implementation is that the system is disk-based. One common weakness of many semantic network systems is that they require the entire network to reside in memory. Since most machines have more disk space than RAM, DESYGNS allows the creation of much larger networks. A B-Tree system is used for rapid retrieval of frames from disk, and a cache is used to store frequently used frames in memory.

Methods

We chose to continue to use AIDS as the domain of development for SAPHIRE. This medical topic is useful not only because of the need to disseminate continually updated knowledge to generalist clinicians, but also because its diseases, anatomy, findings, and treatment cover a broad sampling of medical areas. But despite the diversity of concepts in this area, the total number is at a reasonable level so as to enable experimentation with a prototype system.

In creating the vocabulary, our first need was for a corpus of medical concepts related to AIDS. For this we used the 3,328 MeSH terms that had been assigned to the 5,727 MEDLINE references previously downloaded. Since all MeSH terms are Meta-1 terms, we were able to extract all the synonyms from the various vocabularies designated by Meta-1, such as SNOMED [2], COSTAR [5], and ICD-9 [18]. Since some of these synonyms were duplicates after being subjected to SAPHIRE's punctuation processing and stemming, hand editing to eliminate them was required. There were also a number of synonym forms commonly used in medical texts and queries which were not designated by any of the source vocabularies. These forms were also added to the vocabulary. After this process, there were 6,293 synonym forms that mapped into the canonical concepts. About 50 word level synonyms were also identified. The end result was that with this vocabulary, nearly 10,000 string forms could be mapped into the 3,328 canonical concepts.

Once the vocabulary was built, we used SAPHIRE to index the *Scientific American Medicine* AIDS chapter, along with the titles and abstracts from the MEDLINE references in our collection. Figure 2 shows the title, abstract, and MeSH headings of one of the MEDLINE references from the AIDS collection, while Figure 3 shows the concepts and their frequencies obtained from the SAPHIRE indexing process.

Title:
Acquired immunodeficiency syndrome with severe gastrointestinal manifestations in Haiti.
Abstract:
29 patients (19 males and 10 females) in Haiti were diagnosed as having acquired immunodeficiency syndrome. Their clinical presentation was characterised by unexplained chronic diarrhoea, prolonged fever, extreme weight loss, anorexia, and severe infections. The infectious agents included: Candida albicans (27 patients), Mycobacterium tuberculosis (7 patients), Cryptosporidium (11 patients), Pneumocystis carinii (2 patients), cytomegalovirus (4 patients), and herpes virus (3 patients). In 1 woman Kaposi's sarcoma developed during the course of her disease. Immunological studies of 20 patients revealed profound cell-mediated immune deficiency with cutaneous anergy, marked decrease in the number of T helper cells, and impairment of lymphocyte proliferation. 18 patients died.
MeSH Headings:
Acquired Immunodeficiency Syndrome [immunology]; Adolescence; Adult; Body Weight; Candida Albicans [isolation & purification]; Cell Count; Chronic Disease; Coccidia [isolation & purification]; Diarrhea [etiology] [microbiology]; Esophagitis [diagnosis] [microbiology]; Fever [diagnosis]; Gastrointestinal Diseases [microbiology]; Haiti; Helper Cells; Middle Age; Pneumocystis Carinii [isolation & purification]

Figure 2 -- MEDLINE reference with title, abstract, and MeSH headings

Title: Acquired immunodeficiency syndrome with severe gastrointestinal manifestations in Haiti.
Concepts: 20
ACQUIRED IMMUNODEFICIENCY SYNDROME 2
HAITI 2
PATIENTS 9
FEVER 1
FORCED EXPIRATORY VOLUME 1
WEIGHT LOSS 1
INFECTION 1
AGED 1
CANDIDA ALBICANS 1
MYCOBACTERIUM TUBERCULOSIS 1
CRYPTOSPORIDIUM 1
PNEUMOCYSTIS CARINII 1
CYTOMEGALOVIRUSES 1
VIRUSES 1
SARCOMA 1
CELLS 1
IMMUNITY 1
DEFICIENCY 1
HELPER CELLS 1
LYMPHOCYTES 1

Figure 3 -- Concepts indexed by SAPHIRE from title and abstract of reference in Figure 1 and their frequency

Results

To evaluate indexing, we extensively looked at the indexing concepts chosen for 100 of the MEDLINE references. The most apparent observation was that references without abstracts could not be suitably indexed by SAPHIRE's method. While the titles often contained one or more important concepts, they did not contain a sufficient profile of concepts to represent the content of the article.

Another readily apparent problem was the Porter stemming algorithm [15], which is unsuitable for the medical domain. One example is the stemming of *fever* to *fev*, which is a synonym for *forced expiratory volume*, thus designating the

two as synonymous terms. Another example is the stemming of *more* to *mor*, which is also the stem form of *morals*, thus mapping all instances of *more* to the concept *morals*. Clearly a more precise method than rule-based stemming is needed to enhance the concept recognition process.

It was also seen that the automated indexing process produced a comprehensive yet different profile of indexing concepts than the human-selected MeSH designations. To determine which of these performs better can only be determined by recall and precision analysis. However, these types of studies require a test collection of known queries and matching content, which do not currently exist for the AIDS domain. A major item on our near-term research agenda is the creation of such a test collection for AIDS.

Also noted in comparing the selection of index terms was that MeSH indexers often chose a term similar to a concept in the title or abstract, but at a different level in the hierarchy. For example, a document with repeated reference to *skin lesions*, which in the patients studied happened to be mostly *neoplasms*, had the MeSH term *skin neoplasms* chosen by the indexer. The term *skin neoplasm* is, of course, a subclass of *skin lesion*. A significant proportion of MeSH terms not found by SAPHIRE fell into this category, and another future research agenda item is to develop heuristics that enable the retrieval component to automatically move up and down the hierarchy to assist the user in searching on concepts at a different level in the hierarchy.

Going beyond just the recognition of MeSH terms, we noted a number of concepts that were not part of the MeSH-based portion of Meta-1, and hence not chosen as MeSH indexing terms. This included findings such as *cotton wool spots*. One beneficial aspect of a systematic evaluation of documents for these concepts will be in the expansion of future versions of the metathesaurus to increase concept coverage.

The evaluation of indexing also identified a number of limitations in SAPHIRE's concept recognition approach. One problem noted in the indexing process was related to language syntax. SAPHIRE relies on string matching and synonym substitution in order to find concepts, and it does not currently contain any syntactic knowledge. All of SAPHIRE's concepts are nouns, but this is problematic with a word such as the noun *lead* (representing the chemical *lead*), which can also be a verb (e.g., HIV infection *leads* to the development of neurological disease.).

An additional limitation with SAPHIRE's approach is in the semantics of language. Most queries consist not only of concepts, but also semantic relationships between concepts. For example, a user searching on *quinidine* and *ventricular arrhythmias* is most likely interested in the use of quinidine to *treat* ventricular arrhythmias. However, quinidine can also *cause* ventricular arrhythmias, and if a user were searching with this relationship in mind, then he or she might end up with a number of matches on quinidine used in the treatment of ventricular arrhythmias. MEDLINE has a solution to this problem in MeSH subheadings, but as noted earlier, Funk and Reid [8] showed the highest inconsistency in MEDLINE indexing came with heading/subheading assignment. Furthermore, Sewell [19] has shown that subheadings are generally not employed by most end users of MEDLINE, who tend to use terms connected by logical AND.

SAPHIRE currently does not handle semantic relationships on this level. Miller [20] has investigated the application of these relationships to the searching process. Preliminary work shows that they can add specificity to searching, but there are major questions about the feasibility of this approach on a large scale. Not only do these relationships add to increased requirements in computational power, but it is also unclear how they can be specified in a routine fashion. Miller's current approach is through the use of rules to map a semantic-based query to a combination of MeSH headings and subheadings. For example, a query of *treatment of pneumococcal pneumonia with penicillins* would be mapped to the MeSH query *Penicillin [therapeutic use] AND Pneumonia, Pneumococcal.*

Future Plans

This preliminary investigation into the usage of Meta-1 in a general purpose automated IR system has shown the metathesaurus to be useful in building a concept recognition vocabulary. It has also highlighted a number of areas for future research. Certainly a major task is to perform a quantitative evaluation of SAPHIRE's concept-based automated approach to identify better approaches to concept selection and weighting. The use of Meta-1 has expanded our ability to recognize concepts in documents; we now need to determine optimal strategies for automated indexing. We also want to develop a test collection of queries in the AIDS domain in order to compare the recall and precision of SAPHIRE versus conventional approaches in use by clinicians.

We plan to continue to enhance SAPHIRE's concept recognition capabilities. The main problem that the current version faces is in the limitations related to the program's inability to understand word morphology, syntax, and semantics. Fortunately, these problems are addressed by the features of the CLARIT system [11], which is a computational linguistic approach to information management. In CLARIT, rule-based stemming is replaced by lexically-based morphological analysis, so that words do not merely have *suffixes* removed, but are rendered as canonical (and legitimate) lexical roots. The syntactic features of CLARIT will overcome problems with nouns and verbs of the same syntactic category (e.g., lead) and with some structural ambiguities (e.g., conjoined modifiers and heads in noun phrases). In addition, the semantic features of CLARIT will prevent words with multiple senses (e.g., *black* in black race or black discoloration of a skin lesion) from being misinterpreted.

Also planned is the investigation of semantic relationships in query processing. A semantic network browser, based on DESYGNS, will be adapted for query use in SAPHIRE, allowing users to navigate through the semantic network to add additional concepts to the query. Although it is still unclear whether a user interface can be devised to allow designation of semantic relationships between concepts without being too difficult to use, we will attempt to design and evaluate systems that feature semantic relationships.

Conclusion

We have adapted Meta-1, the first version metathesaurus of the UMLS Project, for use in the vocabulary of SAPHIRE, a computer program that features concept-based automated indexing and retrieval of medical information. Our preliminary work has shown that Meta-1 enhances SAPHIRE's concept recognition ability. Future work will focus on enhancing SAPHIRE's concept recognition

process, evaluating documents indexed in MEDLINE for improved approaches to concept selection and weighting, and exploiting relationships among concepts in the retrieval process.

Acknowledgement

This work was supported in part by Contract N01-LM-8-3513 and Grants LM07037 and LM04572 of the National Library of Medicine.

References

1. "Medical Subject Headings - Tree Structures, 1989." 1989 National Library of Medicine. Bethesda, MD.

2. Cote RA. "Systematic Nomenclature of Medicine." 1982 College of American Pathologists. Skokie, IL.

3. Humphreys BL, Lindberg DAB: Building the Unified Medical Language System. 13th SCAMC. 1989: 475-480.

4. "Diagnostic and Statistical Manual of Mental Disorders III." 1980 American Psychiatric Association. Washington, D.C.

5. Barnett GO, Justice NS, Somand ME, Adams JB, Waxman BD, Beaman PD, Parent MS, Deusen FR Van, Greenlie JK: COSTAR System. Proc. IEEE. 1979; 67: 1226-1237.

6. McCray AT: The UMLS semantic network. 13th SCAMC. 1989: 503-507.

7. Hersh WR, Greenes RA: SAPHIRE: An information retrieval environment featuring concept-matching, automatic indexing, and probabilistic retrieval. Comput Biomed Res. 1990; in press.

8. Funk ME, Reid CA: Indexing consistency in MEDLINE. Bull Med Lib Assoc. 1983; 71: 176-183.

9. Salton G. "Introduction to Modern Information Retrieval." 1983 McGraw-Hill. New York.

10. Harman D, Benson D, Fitzpatrick L, Huntzinger R, Goldstein C: IRX: An information retrieval system for experimentation and user applications. SIGIR Forum. 1988; 22: 2-10.

11. Evans DA: Notes on the CLARIT Project. Technical Report, Laboratory for Computational Linguistics, Carnegie-Mellon University. 1989.

12. Rubin RH. "Acquired Immunodeficiency Syndrome." Scientific American Medicine. Rubenstein R,Federman DD ed. 1985 Scientific American. New York.

13. Greenes RA, Tarabar DB, Krauss M, Anderson G, Wolnik WJ, Cope L, Slosser E, Hersh WR: Knowledge management as a decision support method: A diagnostic workup strategy application. Comp Biomed Res. 1989; 22: 113-135.

14. Shoval P: An Expert Consultation System for a Retrieval Database with a Semantic Network of Concepts. University of Pittsburgh Ph.D. Thesis. 1981.

15. Porter MF: An algorithm for suffix stripping. Program. 1980; 14: 130-137.

16. Borgman CL: Why are online catalogs hard to use? Lessons learned from information retrieval studies. JASIS. 1986; 37: 387-400.

17. Greenes RA, Tarabar DB, Krauss M, Cope L, Slosser E, Hersh WR, Pattison-Gordon E, Abendroth T, Rathe R, Snydr-Michal J: Explorer-2: An object-oriented framework for knowledge management. MEDINFO 89. 1989: 29-33.

18. Slee VN: The Internation Classification of Diseases: Ninth Revision (ICD-9). Ann Int Med. 1978; 88: 424-426.

19. Sewell W, Teitelbaum S: Observations of end-user online searching behavior over eleven years. JASIS. 1986; 37: 234-245.

20. Miller PL, Barwick KW, Morrow JS, Powsner SM, Riely CA: Semantic relationships and medical bibliographic retrieval: A preliminary assessment. Comp. Biomed. Res. 1988; 21: 64-77.