SAPHIRE International: A Tool for Cross-Language Information Retrieval

William R. Hersh, M.D., Laurence C. Donohoe, M.L.I.S.

Division of Medical Informatics and Outcomes Research, School of Medicine
Oregon Health Sciences University
Portland, OR, USA

The world's foremost medical literature is written in English, yet much of the world does not speak English as a primary language. This has led to increasing research interest in cross-language information retrieval, where textual databases are queried in languages other than the one in which they are written. We describe enhancements to the SAPHIRE concept-retrieval system, which maps free-text documents and queries to concepts in the UMLS Metathesaurus, that allow it to accept text input and provide Metathesaurus concept output in any of six languages: English, German, French, Russian, Spanish, and Portuguese. An example of the use of SAPHIRE International is shown in the CliniWeb catalogue of clinically-oriented Web pages. A formative evaluation of German terms shows that additional work is required in handling plural and other suffix variants as well as expanding the breadth of synonyms in the UMLS Metathesaurus.

Introudction

A growing area of research in the information retrieval (IR) field concerns cross-language retrieval. That is, there is increasing interest in the ability to enter queries in one language and retrieve documents in another. In particular, while the majority of the world does not speak English, most scientific, especially medical, literature is written in English. The capability for users to enter queries in their native language and retrieve documents in English is an example where cross-language IR might be of benefit. In this paper, we describe the internationalization of the SAPHIRE concept-matching system, based largely on the multi-lingual aspects of the Unified Medical Language System (UMLS) Metathesaurus.

A variety of techniques have been developed to enable users to query a database in a language different than that in which its text is written. Oard presents a classification of approaches, defining two broad types of cross-language IR, which are based on controlled vocabulary and free text (1). *Controlled vocabulary* approaches rely on human-constructed thesauri, whereas *free text* methods utilize resources and algorithms derived from actual texts. Techniques based on free text methods can be *corpus-based*, where word translations are derived from parallel or comparable document collections in multiple languages, or *knowledge-based*, where translations are handled by dictionaries or more complex natural language processing tools.

Many cross-language techniques, particularly those based on free text, have been utilized outside the medical domain. The earliest work in cross-language IR was performed by Salton, who coupled word-based thesauri with his general vector-space approach (2). One of the advantages of non-linguistic approaches like Salton's is that the IR is essentially language-independent, since the text words are just "tokens" with no semantic meaning. A similar approach undertaken more recently has been the adaptation of latent semantic indexing techniques on parallel or similar corpora (3). The bulk of recent work, however, has focused on the use of cross-language dictionaries in the context of the National Institute for Standard and Technology (NIST) Text Retrieval Conference (TREC) (4-6). Most of the TREC-related work has been evaluated based on the level of recall and precision that can be achieved with multi-lingual queries versus mono-lingual queries. In general, multi-lingual systems have been able to achieve 65-75% of the performance of mono-lingual systems.

We have used a dictionary approach to cross-language IR in internationalizing the SAPHIRE system, taking advantage of a widely-available resource unique to medicine: the UMLS Metathesaurus. In this paper, we first briefly SAPHIRE and its use of the Metathesaurus. We then describe the methods used to internationalize it. This is followed by a description of applications that use multi-lingual SAPHIRE. Finally, we report a formative evaluation of German terms that identifies what additional work will be necessary to improve the system before undertaking a large-scale evaluation.

SAPHIRE

The goal of SAPHIRE is to extract concepts in controlled vocabularies from free text (7, 8). The text can be a medical document or user query to a retrieval system. SAPHIRE is built to explicitly utilize the UMLS Metathesaurus, a rich source of clinical concepts with a great deal of synonym terms from multiple medical vocabularies (9). SAPHIRE is one of several systems that have been developed to extract concepts, recognized by their varying synonyms, from free text (10-15).

SAPHIRE uses minimal amounts of syntactic and semantic information that characterize advanced natural language processing (NLP) systems. While this reduces the complexity of medical phrases that SAPHIRE can recognize, it minimizes dependence on part-of-speech taggers, parsers, complex lexicons, and other linguistic tools, resulting in an algorithm that is relatively fast and does not require the complex maintenance usually associated with these resources. In fact, the major maintenance required for the data used by SAPHIRE is done at the National Library of Medicine (NLM) in maintaining the Metathesaurus.

Before describing SAPHIRE algorithm, one must understand the structure of the UMLS Metathesaurus (9). The Metathesaurus is organized into *concepts*, which have a unique identifier (the CUI). Each major synonym form that is not just a simple lexical variant (i.e., plural or word order change) is a *term*, each of which also has a unique identifier (the LUI). There can be one or more LUI's for each CUI. Each lexical variant of each term is a *string* (with a unique identifier SUI), and there can be more than one SUI for each LUI. As an example, consider the concept *atrial fibrillation*, which has terms *atrial fibrillation* and *auricular fibrillation*. The former term has the lexical variants *fibrillation*, *atrial* and *atrial fibrillations*.

The algorithm begins by breaking the input string (which can be a sentence or phrase from a document or a user's query) into individual words. Words are designated as *common* if they occur with a frequency above a specified cut-off in the Metathesaurus. The purpose of designating words as common is to reduce the computational overload for words which are occasionally important in some terms but occur frequently in others, such as the word *A* in *Vitamin A* or *acute* in *acute abdomen*. Since the words *A* and *acute* occur commonly in many other terms, calculating weights for these additional terms adds a large and unnecessary computational burden.

For each word in the input string, a list of Metathesaurus terms in which the word occurs is constructed. The Metathesaurus term lists for common words contain only those terms that also occur in one or more of the non-common words in the input string. Using one of the above examples, if the string were *acute abdomen*, the common word *acute* would only contain the term *acute abdomen* and not the term *acute leukemia*.

Once the term lists for each word are created, a master term list is created that contains any term which occurs in one or more individual word lists. Terms in which less than half of the words occur in the input string are discarded. (Thus, a partial match must have half or more of the words from the term in the input string.) The terms are then weighted based on formula that gives weight to terms that are longest, have the highest proportion of words from the term in the string, and have the words of the term occurring in close proximity to each other. Terms that match all the words in the input string exactly are given additional weight.

Internationalization Of SAPHIRE

The internationalization of SAPHIRE is made possible by the foreign language terms in the Metathesaurus. The unique aspects of this work are the use of a widely available vocabulary resource and its integration into a concept-matching system. SAPHIRE International is not a mere term translator but actually recognizes many varying expressions of the same underlying concept.

The 1998 Metathesaurus contains terms in five languages in addition to English: German, French, Russian, Spanish, and Portuguese. Terms from these languages are represented as “synonyms” to their equivalent English concepts. At the present time, all of the foreign language terms in the Metathesaurus derive from translations of the Medical Subject Headings (MeSH) vocabulary. (MeSH has actually been translated into 23 languages, but only five are present in the 1998 Metathesaurus.)

Before now, SAPHIRE has not used the non-English terms in the Metathesaurus. However, since the SAPHIRE algorithm is based on lexical matching, internationalization is a straight-forward process. It can be based on the original algorithm described above, with foreign language words comprising the terms that are derived and weighted in the output. The major addition required is the means to handle non-English characters. Most non-English European

languages use diacritical characters, such as umlauts and accents, that are not part of the "7-bit" English ASCII code upon which the Metathesaurus is based. Many of these characters are represented in the upper half of "8-bit" ASCII, as designated by an international standard, the ISO Latin-1 Character Set (ISO 8879). A translation table has been added to SAPHIRE that converts 8-bit ASCII codes into their 7-bit transliterations so they can be used to retrieve words from the Metathesaurus.

SAPHIRE's output is return in the 7-bit ASCII format in which the foreign-language terms have been transliterated in the Metathesaurus. An additional change was necessary in weighting the output, since some foreign languages, particularly German, combine multiple word phrases into single words. This causes the original algorithm to inappropriately downweight such terms, since part of the weighting algorithm is based on the proportion of words common to the input and matched term. Some example queries are shown in Figure 1.

Applications Of SAPHIRE International

SAPHIRE has been used in a variety of applications. Its initial application was for automated indexing in IR applications (7). It has also been used to identify concepts in electronic medical records (16). But its main use recently has been to provide access to index terms in the CliniWeb catalog of clinically-oriented pages on the World Wide Web (17). ClinWeb catalogs over 10,000 Web pages that are oriented to health care professionals and students; home pages, advertisements, and consumer-oriented pages are not included in its database. Each page is indexed with terms from the Medical Subject Headings (MeSH) disease and anatomy trees.

The main value of SAPHIRE International is to allow users to enter query terms in multiple language to find clinical Web pages in CliniWeb. The benefit is that users can enter terms in their native language to retrieve English documents from the Web. There is no reason why documents in other languages could not be retrieved, although there are at present no non-English documents in the CliniWeb database. Figure 2 shows the results of a German query entered into CliniWeb.

SAPHIRE International also enhances another recent addition to CliniWeb, which is the addition of links to the PubMed MEDLINE system (http://www.ncbi.nlm.nih.gov/), a free MEDLINE search facility on the NLM Web site. By using SAPHIRE International, foreign-language entry into PubMed is facilitated. CliniWeb provides generic PubMed searches for all MeSH disease terms. Four different searches are available for each term: reviews, treatment, diagnosis, and the three of these combined. Review articles are found using the MeSH publication type, Review Article, while diagnosis and therapy articles are retrieved using the optimal strategies for best evidence developed by Haynes et al. (18).

Formative Evaluation

There are an increasing number of evaluations of cross-language retrieval tools, mostly based on how well documents can be retrieved from "parallel" document collections that contain nearly identical documents in more than one language. For the initial development of SAPHIRE, such a collection was not available. Furthermore, we decided to focus our initial evaluation on performance of direct language translation capability. Our initial efforts were limited to the German language. An American librarian fluent in German and a German medical documentation specialist were asked to type a variety of German medical terms they encountered in their everyday work into the system.

Both subjects entered about two dozen queries each. A number of generalizations about the system were observed:

1. Unlike their English counterparts in the Metathesaurus, many terms did not have both plural and singular forms present. Until the Metathesaurus increases richness of lexical variants in foreign terms, a stemming algorithm to handle common suffix variants will be necessary.
2. Many German terms in the Metathesaurus are in their Latin as opposed to German form. For example, the Latin *Myokardinfarkt* is present as opposed to the German *Herzinfarkt*. Clearly an increased number of synonyms will be necessary as well.
3. In the German language, individual words that come together in phrases form single words that are generally not present in the Metathesaurus, e.g., *Oesophagusvarizenblutung* or *esophogeal bleeding*. This may represent a necessary enhancement for future versions of the Metathesaurus if more comprehensive foreign-language coverage is desired.

Future Directions

The preliminary version of SAPHIRE International shows promise as a tool for cross-language IR that leverages the foreign-language terms in the UMLS Metathesaurus. Further research will focus on the accuracy of its algorithm as well as its optimal use in

actual IR systems. Improving the accuracy will require enhancements to the algorithm for stemming and term weighting as well as increasing the coverage of foreign-language synonyms in the Metathesaurus.

SAPHIRE International can be accessed at: http://www.ohsu.edu/cliniweb/saphint/
Due to vocabulary copyright restrictions, only the Russian, Spanish, and Portuguese versions are publicly accessible. CliniWeb International can be accessed at: http://www.ohsu.edu/cliniweb/

Acknowledgements

We thank Fredereike Rollman and Patty Davies for carrying out the formative evaluation. This work was supported by cooperative agreement LM05879 from the National Library of Medicine and grant DE-FG03-94ER61918 from the Department of Energy.

References

1. Oard D. Alternative approaches for cross-language text retrieval. In: Hull D, Oard D, eds. 1997 AAAI Symposium on Cross-Language Text and Speech Retrieval, 1997.
2. Salton G. Experiments in multi-lingual information retrieval. Cornell University, 1972.
3. Landauer T, Littman M. Fully automatic cross-language document retrieval using latent semantic indexing. Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, 1990:31-38.
4. Hull D, Greffenstette G. Querying across languages: a dictionary-based approach to multilingual information retrieval. In: Croft B, vanRijsbergen C, eds. Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval, 1996:49-57.
5. Davis M. New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In: Harman D, ed. The Fifth Text REtrieval Conference, 1996:447-454.
6. Ballesteros L, Croft W. Phrasal translation and query expansion techniques for cross-language information retrieval. In: Belkin N, Narasimhalu A, Willett P, eds. Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval, 1997:84-91.
7. Hersh W, Hickam D. Information retrieval in medicine: the SAPHIRE experience. Journal of the American Society for Information Science 1995;46:743-747.
8. Hersh W, Leone T. The SAPHIRE server: a new algorithm and implementation. In: Gardner R, ed. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1995:858-862.
9. Humphreys B, Lindberg D, Schoolman H, Barnett G. The Unified Medical Language System: an informatics research collaboration. Journal of the American Medical Informatics Association 1998;5:1-11.
10. Vries J, Shoval P, Evans D, Moossy J, Banks G, Latchaw R. An Expert System for Indexing and Retrieving Medical Information. Technical Report, University of Pittsburgh School of Medicine 1986.
11. Yang Y, Chute C. An application of linear least squares fit mapping to clinical classification. In: Frisse M, ed. Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care, 1992:460-464.
12. Sager N, Lyman M, Nhan N, Tick L. Automatic encoding into SNOMED III: a preliminary investigation. In: Ozbolt J, ed. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994:230-234.
13. Friedman C, Alderson P, Austin J, Cimino J, Johnson S. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association 1994;1:161-174.
14. Haug P, Koehler S, Lau L, Wang P, Rocha R, Huff S. A natural language understanding system combining syntactic and semantic techniques. In: Ozbolt J, ed. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994:247-251.
15. Cole W, Sherertz D, Tuttle M, Keck K, Olson N. Metaphrase: achieving formalized EMR problem lists from informal input. In: Masys D, ed. Proceedings of the 1997 Annual AMIA Fall Symposium, 1997:931.
16. Lowe H, Hersh W, Smith C. The multimedia medical record as a virtual library: a multidimensional model for indexing the content of medical images using the Unified Medical Language System. In: Campbell J, ed. Final Program and Abstract Book, AMIA Spring Congress, 1998: 83.
17. Hersh W, Brown K, Donohoe L, Campbell E, Horacek A. CliniWeb: managing clinical information on the World Wide Web. Journal of the American Medical Informatics Association 1996;3:273-280.
18. Haynes R, Wilczynski N, McKibbon K, Walker C, Sinclair J. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. Journal of the American Medical Informatics Association 1994;1:447-458.

German to English: *natriumbicarbonat für diabetische ketoazidose*

CUI	String	Vocabularies
C0011880	Diabetic Ketoacidosis	MSH97
C0074722	Sodium Bicarbonate	MSH97
C0022638	Ketosis	DOR27 ICD91 MSH97 MTH

Spanish to English: *la náusea y el vómito*

CUI	String	Vocabularies
C0042963	Vomiting	ICD91 LCH90 MSH97 PSY94 RCD95 RCDSY SNM2
C0027497	Nausea	DOR27 ICD91 LCH90 MSH97 PSY94 RCD95 SNM2 SNMI95

English to French: *surgery for appendicitis*

CUI	String	Vocabulary
C0038894	CHIRURGIE	INS97
C0003615	APPENDICITE	INS97

Figure 1: Cross-lingual queries – results of queries with Metathesaurus concept identifier (CUI), preferred form string, and vocabularies in which terms occur.

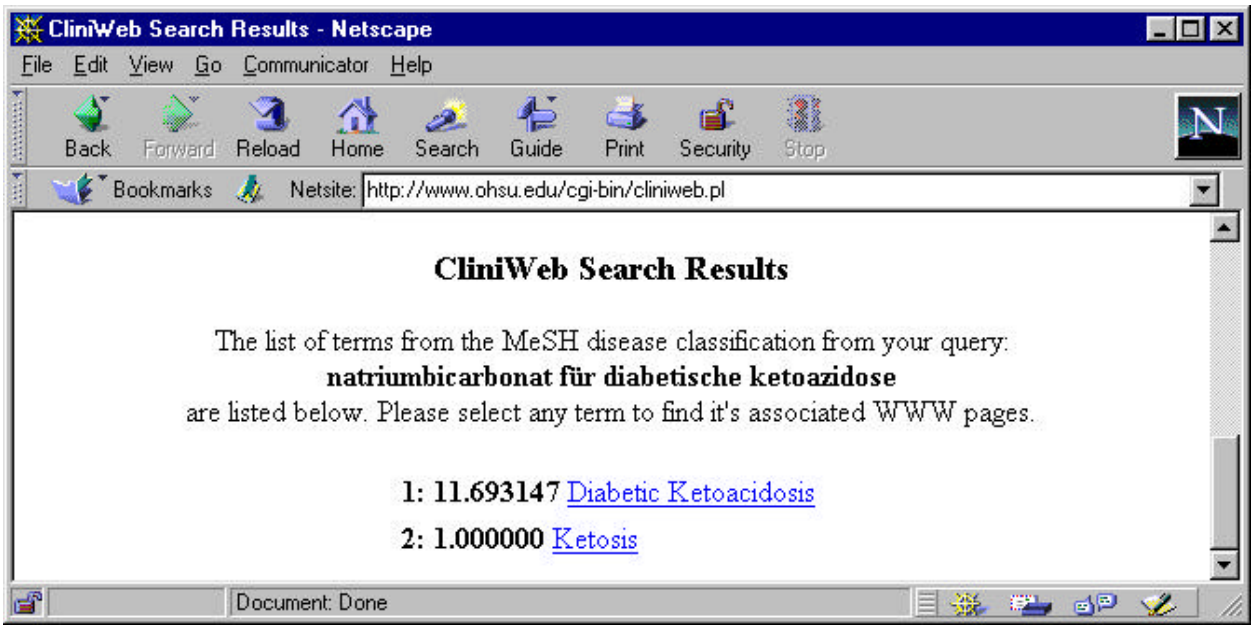

Figure 2: Results of CliniWeb search on a German query, *natriumbicarbonat für diabetische ketoazidose*. Sodium bicarbonate is not returned since CliniWeb only contains MeSH disease and anatomy terms.