INFERENCING IN INFORMATION RETRIEVAL

Alexa T. McCray

National Library of Medicine
Bethesda, Maryland 20894

ABSTRACT

We have been addressing the problem of providing access to the free text in biomedical databases. The focus of our work is the development of SPECIALIST, an experimental NLP system for the biomedical domain. The system includes a broad coverage parser supported by a large lexicon, a module that provides access to extensive biomedical knowledge sources, and a retrieval module that allows us to carry out experiments in information retrieval. We have recently conducted experiments with a test collection of queries and documents retrieved for those queries. The purpose of the investigation has been to determine the type of information that is required in order to effect a map between the language of queries and the language of relevant documents.

1. INTRODUCTION

Retrieval of information from computerized databases is a complex process whose success depends heavily on the user's knowledge of the structure and logic of the particular database being searched. Many databases have associated with them a controlled indexing vocabulary, or thesaurus, which is the primary access point to the material at search time. For example, the National of Library of Medicine's MeSH® thesaurus includes some 16,000 headings that are available for indexing and searching the biomedical literature stored in MEDLINE®, NLM's bibliographic database. The major retrieval strategy is to coordinate MeSH terms with boolean operators, although limited text word searching of titles and abstracts is also possible.

Several years ago NLM launched its Unified Medical Language System™ (UMLS™) project. This is a major research initiative whose goal it is to facilitate retrieval and integration of information from multiple disparate biomedical databases. NLM itself has developed and maintains over 40 databases, and there are many other sources of computerized information in the biomedical sciences. These include factual databases of various kinds, diagnostic expert systems, clinical information systems, as well as bibliographic databases. The UMLS project is attempting to develop methods whereby access is provided to these different systems with their different vocabularies in a way which allows the user to navigate among them with relative ease. Recent results of the project have been the development of an Information Sources Map of biomedical databases, a Metathesaurus™ of biomedical vocabularies and a Semantic Network of high-level biomedical concepts[1,2]. The first release of the Information Sources Map contains a description of the scope, content, and access conditions for approximately fifty biomedical databases. The Metathesaurus includes over 67,000 biomedical concepts from a variety of controlled vocabularies. Definitions, lexical category information, hierarchical contexts, and interrelationships among many of the terms found in its constituent vocabularies are provided. Each concept in the Metathesaurus has been assigned to at least one of the 131 semantic types in the Semantic Network. The Network has top level nodes for organisms, anatomical structures, biologic function and dysfunction, chemicals, events, and concepts. The Network defines these types and establishes a set of 35 potential relationships between them. These include physical, temporal, functional, and conceptual links, e.g., *part of, co-occurs with, causes, measures*. The Network and the Metathesaurus together form a rich knowledge source of biomedical concepts. The knowledge sources will continue to be augmented and refined based on experimentation in a variety of applications, including our own.

Our work is motivated by an interest in the development and testing of natural language processing techniques for improved methods of information retrieval. Document retrieval systems, in particular, are "language-rich" and afford the opportunity to conduct basic research in processing complex natural language text. The focus of our work is the development of SPECIALIST, an experimental NLP system for the biomedical domain[3,4,5]. The system includes a broad coverage parser[1] supported by a large lexicon, a module that accesses the UMLS knowledge sources, and a retrieval module. SPECIALIST runs on Sun workstations and is implemented in Quintus Prolog, with some support modules written in C.

We have recently conducted experiments using a test collection of user queries and MEDLINE citation records retrieved for those queries. The data for the test collection were se-

[1]During the academic year 1988-1989 we awarded a research contract to the Paoli Research Center of the Unisys Corporation. As a result of this successful collaboration between our two research groups, the syntactic component of the system is extremely robust. See[6,7] for a description of the Paoli system.

lected from 2,000 search request forms submitted by health professionals to the NIH and NLM libraries. 155 queries were chosen, approximately 50 each in the three major areas covered by MEDLINE - clinical medicine research, basic science research, and health services research. Searches were conducted by an expert NLM searcher, and the approximately 3,000 citations retrieved were evaluated for relevancy by a subject matter expert[8]. Each citation record in the collection includes a title and an author-prepared abstract.

We parsed the queries, titles and selected portions of abstracts in the test collection. For all successful parses, noun phrases were extracted and whatever synonyms could be found in the Metathesaurus and in our online version of the *Dorland's Illustrated Medical Dictionary*[5] were added to the noun phrases to form a concept group. We then attempted to effect a match between the concepts in the queries and the concepts in relevant citations. We found that the mapping involves a wide range of inferences. It is only in rare cases that concepts map directly from queries to documents. More commonly, several inferences are necessary in order to determine that a citation is in fact relevant to a request.

The remainder of this paper begins with a discussion of some of the salient issues in information retrieval. This is followed by a brief description of the major components of the SPECIALIST system, and the paper ends with an account of our recent investigations in mapping queries to relevant documents.

2. THE INFORMATION RETRIEVAL PROBLEM

One of the essential characteristics of document storage and retrieval is the parallelism between the indexing and searching processes. Text is subjected to either manual or automatic indexing. If it is manual, there will generally be indexing rules. For example, in the case of NLM's MEDLINE database, one rule says that articles should be indexed with the most specific terms available in NLM's MeSH vocabulary[9]. Thus, if an article is about aplastic anemia it should be indexed under that term (which is a bona fide MeSH term) and not under either the more general term "anemia" or the even more general term "hematologic diseases". At search time, the user (or program acting on the user's behalf) needs to take this into account when formulating a search strategy and statement. Even if the indexing is automatic, and it may be as simple as creating an inverted index for all the words in the document, the user (or program) needs to be aware of the conventions for creating that index. This includes recognition of the fact that text words are generally run against a stopword list of function words and other highly frequent words before they are entered in the database. For example, if the user wants to query the MEDLINE database on "the effects of acidosis on ATP" and uses text words only, the two words "acidosis" and "ATP" will individually yield many postings and their coordination will yield another, smaller, set. However, adding in the word "effects" will not make the search results any more precise, since, as a highly frequent word in biomedical documents, this word has been placed on the stopword list. Without some knowledge of these conventions, the results can be confusing to end users.

Feedback of various kinds allows the user to negotiate with the retrieval system. This may involve refining a search statement based on viewing the set of titles or documents initially retrieved, or finding that because the number of postings for a search statement is unacceptably large or small that the search strategy has been too broad, or too narrow, or misformulated in some other way. It may also involve accessing information about the indexing rules or controlled vocabulary used in the system. The effect of this feedback is that it makes the user more aware of both the potential of the retrieval system as well as its limitations. Most researchers in intelligent interface design assume to one degree or another that the user will be "left in the loop" to negotiate with the system, resolving ambiguities, making relevancy judgements, and revising searches based on (user-independent) information supplied by the system. (See[10] for a strong statement about the desirabilty of giving the user maximum control over the entire search interaction.)

Many of the attempts that have been made to apply NLP to information retrieval have involved the search interface; others have involved the indexing process. See[11] for a review of some of the more recent research efforts. The results of applying NLP to the information retrieval problem have not always been encouraging[12]. It is important to recognize why this might be so. First, retrieval experiments have been carried out that use partially developed parsing systems and then compare these results with other non-NLP methods. The results of these comparisons should, therefore, be viewed with caution. In some cases, so-called stemming procedures have been used which embody some linguistic sophistication, but, again, are not fully motivated or developed. The results of these experiments again underscore the limitations of the incomplete methods used. Second, given that the indexing and retrieval processes are so closely related, a successful application of NLP will need to be fully integrated with both processes. Some of the inconclusive results in[13], for example, may derive from a decision to ignore this point.

3. THE SPECIALIST SYSTEM

3.1 Lexicon and Parser

Lexical information is central to our parsing system. The lexicon currently contains some 51,000 lexical items, with over 88,000 lexical forms. It includes both general English lexical items as well as items specific to the domain of

biomedicine. Lexical entries are created using our lexicon building tool called Lextool. Lextool is a menu-based system which accepts as input either a file of lexical items or lexical items typed in from the keyboard. With the interactive aid of the user, it generates fully specified lexical frames. Lextool incorporates rules that dictate which slots are permissible for the syntactic category in question. The coding system is closely tied to the codes given in the first edition of the *Longman Dictionary of Contemporary English*[14], although we have modified this scheme somewhat, and we have added additional codes, for example, those for logical interpretation, such as subject control, object raising, etc. We do not have the Longman dictionary in machine readable form, but other online information sources are available to lexical coders in the Lextool environment. These include the *Dorland's Illustrated Medical Dictionary*[5]; Meshtool, our MeSH vocabulary browser; Meta, the Metathesaurus retrieval system; and access to sample sentences from MEDLINE citations which contain the lexical items in question. The two sample records shown below illustrate the type of information that is encoded for lexical items[2].

base=sad
entry=1
  cat=adj
  variants=regd
  position=attrib(1)
  position=pred
  compl=fincomp(t):subj
  compl=fincomp(t)
  nominalization=sadness

base=aim
entry=1
  cat=noun
  variants=uncount
  variants=reg
entry=2
  cat=verb
  variants=reg
  intran
  tran=infcomp:subjc
  tran=np
  tran=pphr(at,np)
  tran=pphr(at,ingcomp:subjc)
  tran=pphr(for,np)
  tran=pphr(for,ingcomp:subjc)
  ditran=np,pphr(at,np)

[2]Semantic and pragmatic information is not stored directly with lexical entries. We are, however, currently considering a variety of approaches to semantics and our future work in this area may well have an impact on the structure of the lexical entries.

The record for "sad" illustrates the sort of information we encode for adjectives. Included is variant information (i.e., whether the adjective forms regular comparative and superlatives); positional information, e.g., whether the adjective is predicative, attributive, or both; adjective type (e.g., the "1" in "attrib(1)" indicates that this is an adjective of quality); information about possible complements (e.g, finite, infinitival complements); and information about any nominalizations.

The record for "aim" illustrates some of the information we encode for nouns and verbs. Noun frames include variant information and information about possible complements and nominalizations, if relevant. Verbs are most extensively coded. While any particular complement slot of a verb is optional, at least one from the set "intran, "tran", "ditran", or "cplxtran", must be chosen. In addition, the particular type of object is encoded. For example, *aim* as a verb may be transitive, and if so, it can take a single np as an object or one of a variety of prepositional phrase complements (e.g., "aim at the target", "aim at winning", "aim for the best", etc.).

The grammar includes context-free BNF rules together with context-sensitive restrictions. It is based heavily on the Pundit grammar, but we continue to refine and modify it so that it can handle new constructions and additional lexical attributes. A slightly simplified sample parse is shown below.

S = Rifampin is administered in the treatment of tuberculosis.

OPS: present, passive
VERB: administer
SUBJ: null
OBJ: rifampin (sing,(Pharmacologic Substance))
PP: in
  treatment (sing,(Therapeutic Procedure))
  RMOD: of
    tuberculosis (sing,(Disease or Syndrome))

We have investigated the possiblity of using the UMLS semantic types for expressing selectional restrictions. Our initial assessment is that they may be profitably used, but since we are currently developing a general approach to semantics, we have not yet implemented any restrictions of this sort. In the meantime, we report semantic types in the output parse. The semantic types are not directly encoded in lexical entries, but are looked up at parse time in our Metathesaurus retrieval application.

3.2 Access to Knowledge Sources

The Metathesaurus application allows users (or programs) to search for Metathesaurus terminology, reporting the term and its source vocabulary; its definition, synonyms, related or associated terms; its semantic types; its lexical tags and variants; or its contexts, e.g., its ancestors or descendants.

Simplified sample output for some queries for "Gierke's disease" are shown below. Note that "Gierke's disease" is a synonym of "Glycogen Storage Disease Type I", and is, therefore, mapped to this term throughout.

[CN = concept name, DEF = definition, VOC = source vocabulary (MSH = MeSH, SNOMED = Systematized Nomenclature of Medicine), STY = semantic type, SY = synonym].

Concept Definition [return to quit]: Gierke's disease

CN: Glycogen Storage Disease Type I
DEF: An autosomal recessive disease in which gene expression of glucose-6-phosphatase is absent, resulting in hypoglycemia due to lack of glucose production. Accumulation of glycogen in liver and kidney leads to organomegaly, particularly massive hepatomegaly. Increased concentrations of lactic acid and hyperlipidemia appear in the plasma. Clinical gout often appears in early childhood.
VOC: MSH

Semantic Type: [return to quit]: Gierke's disease

CN: Glycogen Storage Disease Type I
STY: Disease or Syndrome

Synonyms [return to quit]: Gierke's disease

CN: Glycogen Storage Disease Type I
SY: Gierke's Disease
SY: Glucose-6-Phosphatase Deficiency
SY: Glucosephosphatase Deficiency
SY: Glycogenosis 1
SY: Hepatorenal Glycogen Storage Disease
SY: Hepatorenal glycogenosis
SY: Von Gierke Disease

Ancestors [return to quit]: Gierke's disease

CN: Glycogen Storage Disease Type I
VOC: MeSH

Diseases
  Nutritional and Metabolic Diseases
    Metabolic Diseases
      Metabolism, Inborn Errors
        Carbohydrate Metabolism, Inborn Errors
          Glycogen Storage Disease
            Glycogen Storage Disease Type I

CN: Glycogen Storage Disease Type I
VOC: SNOMED

Disease Axis
  Metabolic and Nutritional Diseases and Syndromes
    Diseases of Carbohydrate Metabolism
      Glycogen storage disease, type I

3.3 Retrieval Module

As noted above, we have developed a retrieval module in order to test the extent to which NLP techniques may improve information retrieval. The current implementation of the module processes files such as MEDLINE citation records, creates an index for the items in all relevant fields, including MeSH terminology and text words, and provides for boolean retrieval of these items. In addition to retrieval based on the MeSH vocabulary and text words, the retrieval module also provides for noun phrase extraction, indexing, and retrieval. A noun phrase index is created by parsing the textual fields of input records, generating several variants of each noun phrase and computing synonyms of each variant. During retrieval, noun phrases are similarly extracted from a parse of the user's query and processed against the noun phrase index.

The retrieval module gives us direct access to the test collection of queries and citation records and was heavily used in the experiment reported below.

4. MAPPING QUERIES TO DOCUMENTS

Locating relevant documents in a bibliographic database is a complex process that involves users - their knowledge of the subject matter, their understanding of the conventions of the database, their familiarity with the interface to that database - and it involves the relationship between the meaning of a query and the meaning of a relevant document.

A query generally is directed to just one, or perhaps a few aspects, of a full document. The relationship between the query and document may be direct, or it may be quite indirect. The following examples from the test collection illustrate[3]. A query in the clinical medicine research portion of the collection is, "Causes, treatment, signs and symptoms of depression specifically in the post partum period (i.e., first year after childbirth or traceable to the event of childbirth). To include mild depression (also known as 'baby blues') to post partum psychosis." The title of a relevant citation is, "A prospective study of postpartum psychoses in a high-risk group. Clinical characteristics of the current postpartum episodes." Here the title clearly answers at least part of the

[3]As noted above, the queries were collected from two medical libraries. They consist primarily of search request forms filled in by users of these libraries. The language is, therefore, the natural language of the user, and it is directed to a human search specialist rather than to a computer interface.

query directly and is, thus, deemed relevant.

A somewhat less direct correspondence between the query and document is shown by an example from the health services research portion of the collection. The query is, "Attitudes of health personnel as it relates to neoplasms, AIDS, and ALS." The title of one of the documents retrieved for this query is, "The impact of a program to enhance the competencies of primary care physicians in caring for patients with AIDS." The abstract, while not directly discussing attitudes of physicians treating AIDS patients does indicate that of 635 physicians interviewed, only 30 percent "demonstrated adequate knowledge of practices necessary to deal with patients' AIDS-related symptoms and concerns."

Our recent investigations have looked at the degree of similarity between the language of a query and the language of a relevant document[4]. Our experiments involved parsing the query and document texts, extracting the constituent noun phrases, augmenting these with synonyms and other variants, and then attempting to map queries to relevant documents. We found that the mappings are almost never straightforward and almost always involve multiple inferences.

Our current parsing system was able to handle about 45 percent of the 155 queries and about 55 percent of the 3,078 titles in the collection. As we analyzed particular phenomena, we parsed selected portions of some of the abstracts. Both the queries and the titles are generally complex noun phrases, but queries tend to be more elliptical and much less well-formed than titles. Abstracts consist of well-formed English sentences, but some of the structures found there are highly specialized. The following sentence from one of the abstracts illustrates: "At 55-57 days of age, the animals were divided into the following dietary treatment groups: A) 4.5 % fat [control fat (CF)]; B) CF + 1.0 MMOL ROA/kg diet (CF + ROA); C) 20.0 % fat [high fat (HF)]; D) HF + ROA."

Our investigations have indicated that the mapping between queries and documents involves a range of phenomena. When concepts do not map directly to each other, it is often the case that various types of relations between them are the key to a successful mapping. The synonymy relation is clearly of great importance to robust retrieval systems. The more synonyms or closely related terms there are available at search time, the more likely it is that a user will find the desired documents. (For example, see[15] for the view that traditional retrieval systems would be greatly improved by the addition of huge numbers of synonyms, or "aliases"). The synonymy must, however, go beyond the word-level to the phrase-level. An example from our experiment illustrates. The fairly simple query is, "**Vitamin C** and immunity". The title of a relevant citation is "Effect of **ascorbic acid** on humoral and other factors of immunity in coal-tar exposed workers." Both the Metathesaurus and the Dorland dictionary list "vitamin C" and "ascorbic acid" as synonyms, so, in this case, parsing the query and title, together with a look-up in our online resources has the desired effect.

Another example illustrates some of the more complex relations that may exist between concepts in queries and documents. The query is, "Hematoporphyrin derivative treatment of tumors using a laser." The first sentence of a relevant citation is, "Photoradiation with photosensitizing porphyrins offers a potentially useful approach to the diagnosis and treatment of certain human cancers." The system must recognize that hematoporphyrin is a kind of porphyrin, that tumors are related to cancer, and that the use of a laser is implied by photoradiation. Access to the knowledge contained in the Metathesaurus does, in fact, allow these inferences to be made. A sub-tree in the MeSH hierarchy, one of the constituent vocabularies in the Metathesaurus, is shown below. Hematoporphyrin is shown to be a narrower term than porphyrin and the isa link is implied:

Chemicals and Drugs
  Growth Substances, Pigments, Vitamins
    Pigments
      Porphyrins
        Hematoporphyrins

Tumor is listed as a synonym of neoplasm which is itself a broader term than cancer in the Metathesaurus, and photoradiation is listed as a synonym of light which is broader than lasers:

Physical Sciences
  Physics
    Optics
      Light
        Lasers

By navigating through the interrelationships expressed in the Metathesaurus structure, the system is able to draw the appropriate inferences.

Another example illustrates a somewhat more complex case. The query is, "Ocular complications of Myasthenia Gravis". A relevant title is, "Myasthenia gravis and recurrent retrobulbar optic neuritis: an unusual combination of diseases". Myasthenia gravis is a neuromuscular disorder and is generally associated with ocular complications of a muscular nature, such as ptosis, diplopia, and ophthalmoplegia. The optic neuritis mentioned in the title is, however, an inflammatory disorder. The correct inference can be made by referring to the Semantic Network which has established the potential relation "complicates" between any two co-occurring diseases. In this case, then, the literature has actually instanti-

[4]For our purposes a document consists of a title and an abstract.

ated the "complicates" relationship between the two normally unrelated disorders mentioned in the title.

It is clear that while identifying noun phrases in queries and documents will improve the mapping capabilities of a retrieval system, it will not be capable of drawing many of the deeper inferences that are required. A fairly simple example makes the point. The query is, "Thermography for indications other than breast." An obviously relevant title is, "Use of thermogram in detection of meningitis." Here a system needs to know that "breast" actually refers to "breast disorders" and that "other than" is a negative operator. As we incorporate more semantics into our parser, some of these inferences should fall out.

Most often the process of locating a relevant document involves mapping sets of concepts and their interrelationships in queries onto similar sets of concepts and interrelationships in documents. These interrelationships between major concepts may be explicit or they may be implicit. An example of an explicit relation is shown in the following query, "Transillumination light scanning for use in the **detection** of diseases of the breast.". A relevant title for this query is "The value of diaphanography as an adjunct to mammography in breast **diagnostics**." Here the notion of using a particular technique to *detect*, or *diagnose*, the disorder is of paramount importance.

An example of an implicit relationship is shown in the query, "Neoplasia in kidney, heart, and liver transplant recipients." The user is probably interested in articles that discuss neoplasia arising as a **result of** of the transplant (or more likely the immunosuppressive therapy associated with the transplant), but this is not directly stated. A relevant title for this query is, in fact, "Development of incidence of cancer **following** cyclosporine therapy."

In many cases, it will not be possible for a system to draw the appropriate inferences without the interactive aid of the user. This is most likely if only noun phrases are presented as a search statement. For example, if a query consists simply of the two terms "rifampin" and "tuberculosis", multiple interpretations of the relationship between these terms are possible. The Semantic Network, for example, provides the following potential relationships between drugs and diseases: *affects, prevents, complicates, treats, diagnoses*, and *causes*. If the user is presented with the set of possible relations between drugs and diseases, a choice can be made and the query can be further refined.

Our work to date has revealed a variety of inferences that must be made if the attempt to map a query to a relevant document is to be successful. We intend to continue our explorations of these phenomena, and we have begun to develop an approach to handling some of them. Our online sources of biomedical information have already proven to be of direct use in making some of the appropriate inferences.

REFERENCES

1. Lindberg, D.A.B. and Humphreys, B.L. "The UMLS Knowledge Sources: Tools for Building Better User Interfaces," *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, R.A. Miller (ed.), IEEE Computer Society Press, 1990, pp. 121-125.
2. McCray, A.T. and Hole, W.T. "The Scope and Structure of the First Version of the UMLS Semantic Network," *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, R.A. Miller (ed.), IEEE Computer Society Press, 1990, pp. 126-130.
3. McCray, A.T. "Natural Language Processing for Intelligent Information Retrieval," *Proceedings of the Annual Conference of the IEEE Engineering in Medicine and Biology Society*, Volume 13, 1991, pp. 1160-1161.
4. McCray, A.T. "Extending a Natural Language Parser with UMLS Knowledge," *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care*, P.D. Clayton (ed.), McGraw-Hill, Inc. 1991, pp. 194-198.
5. McCray, A.T. and Srinivasan, S. "Automated Access to a Large Medical Dictionary: Online Assistance for Research and Application in Natural Language Processing", *Computers and Biomedical Research*, Vol. 23, 1990, pp. 179-198.
6. Hirschman, L., Palmer,M. Dowding, J., Dahl, D. Linebarger, M. Passonneau, R., Lang, F.-M., Ball, C. and Weir, C., "The PUNDIT Natural-Language Processing System", *AI Systems in Government Conference*, Computer Society of the IEEE, March 1989.
7. Hirschman, L. and Dowding, J., "Restriction Grammar: A Logic Grammar", *Logic and Logic Grammars for Language Processing*, Saint-Dizier, P., and S. Szpakowicz (eds.), Ellis Horwood, 1990, 141-167.
8. Schuyler, P.L., McCray, A.T., and Schoolman, H.M., "A Test Collection for Experimentation in Bibliographic Retrieval, *MEDINFO89* North-Holland, 1989, pp. 910-912.
9. Charen, T. , "Medlars Indexing Manual", Technical Report NLM-MED-83-06, National Technical Information Service, 1983.
10. Bates, M.J., "Where should the person stop and the information interface start?", *Information Processing & Management*, Vol. 26, No. 5, 1990, pp. 575-591.
11. Smeaton, A.F., "Information Retrieval and Natural Language Processing", *Prospects for Intelligent Retrieval*, Informatics 10, Jones, K. (ed.), 1990, pp. 1-14. 1990, pp. 575-591.
12. Salton, G., Buckley, C., and Smith, M., "On the Application of Syntactic Methodologies in Automatic Text Analysis.", *Information Processing & Management*, Vol. 26, No. 1, 1990, pp. 73-92.
13. Sparck Jones, K. and Tait, J.I., "Automatic Search Term Variant Generation", *Journal of Documentation*, Vol. 40, No. 1, March 1984, pp. 50-66.
14. Procter, P. (ed) *Longman Dictionary of Contemporary English*. Longman Group Limited 1978.
15. Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T., "The Vocabulary Problem in Human-System Communication"., *Communications of the ACM*, vol. 30, No. 11, 1987, pp. 964-971.