UMLS® knowledge for biomedical language processing

By *Alexa T. McCray, Ph.D.*
Deputy Chief, Computer Science Branch

Alan R. Aronson, Ph.D.
Computer Scientist

Allen C. Browne
Computational Linguist

Thomas C. Rindflesch, Ph.D.
Computational Linguist

National Library of Medicine
8600 Rockville Pike
Bethesda, Maryland 20894

Amir Razi, Ph.D.
Computer Scientist

Suresh Srinivasan
Computer Scientist

Management Systems Designers, Inc.
131 Park Street
Vienna, Virginia 22180

This paper describes efforts to provide access to the free text in biomedical databases. The focus of the effort is the development of SPECIALIST, an experimental natural language processing system for the biomedical domain. The system includes a broad coverage parser supported by a large lexicon, modules that provide access to the extensive Unified Medical Language System® (UMLS®) Knowledge Sources, and a retrieval module that permits experiments in information retrieval. The UMLS Metathesaurus® and Semantic Network provide a rich source of biomedical concepts and their interrelationships. Investigations have been conducted to determine the type of information required to effect a map between the language of queries and the language of relevant documents. Mappings are never straightforward and often involve multiple inferences.

INTRODUCTION

Retrieval of information from computerized databases is a complex process. Success depends heavily on the user's knowledge of the structure and logic of the particular database being searched. The Unified Medical Language System® (UMLS®) project addresses the problem of assisting users as they attempt to interact with databases of varying degrees of complexity with sometimes widely differing access methods. A significant aspect of the interaction between computers and humans involves questions of language. The UMLS Metathesaurus® and UMLS Semantic Network represent and link a number of biomedical vocabularies with the goal of using this knowledge to help users retrieve information from a wide variety of biomedical information sources.

The authors' research in natural language processing (NLP) addresses the contributions NLP techniques can make to this complex task of mediating between the language of users and the language of the databases they attempt to access. The focus of the effort is the development of SPECIALIST, an experimental NLP system for the biomedical domain [1–3]. The system includes a broad coverage parser supported by a large lexicon, a module that accesses the UMLS Knowledge Sources, and a retrieval module. SPECIALIST runs on Sun Sparcstations and is implemented in Quintus Prolog, with some support modules written in C.

An investigation was recently conducted using the UMLS test collection of user queries and MEDLINE® citation records retrieved for those queries [4]. The queries, titles, and selected portions of abstracts were parsed. For all successful parses, noun phrases were extracted, and synonyms from both the Metathesaurus and an online version of the *Dorland's Illustrated Medical Dictionary* [5] were added to the noun phrase to form a concept group. Then an attempt was made to match concepts in the queries to concepts in relevant citations. It is only in rare cases that concepts map directly from queries to documents. More commonly, several inferences are necessary to determine that a citation is relevant to a request. The UMLS Metathesaurus and Semantic Network are valuable knowledge sources in making the appropriate inferences.

MAPPING QUERIES TO DOCUMENTS

Mapping queries to relevant documents involves a range of phenomena. As the experimental system is developed further, the understanding of these phenomena will continue to be refined, as the following examples from the UMLS test collection illustrate. One query in the clinical medicine research section of the collection is "Causes, treatment, signs, and symptoms of depression specifically in the postpartum period (i.e., first year after childbirth or traceable to the event of childbirth). To include mild depression (also known as baby blues) to postpartum 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 query directly and thus is deemed relevant.

A somewhat less direct correspondence between the query and document is shown by an example from the health services research section 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." While not directly discussing the attitudes of physicians treating AIDS patients, the abstract does indicate that of 635 physicians interviewed, only 30% "demonstrated adequate knowledge of practices necessary to deal with patients' AIDS-related symptoms and concerns."

When concepts do not map directly to each other, the various types of relations between them often 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, the more likely it is that a user will find the desired documents. The synonymy must, however, go beyond the word level to the phrase level, as an example from the test collection 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 parsing the query and title, together with a look-up in these 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 hematophorphyrin 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 subtree in MeSH®, showing that *hematophorphyrin* is a narrower term than *porphyrin*, is shown below:

```
Chemicals and Drugs
  Growth Substances, Pigments, Vitamins
    Pigments
      Porphyrins
        Hematoporphyrin
```

"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 a broader term 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 generally is 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 UMLS Semantic Network, which has established the potential relation "complicates" between any two co-occurring diseases. In this case, then, the literature actually has represented the "complicates" relationship between the two normally unrelated disorders mentioned in the title.

It is clear that although identifying noun phrases in queries and documents improves mapping capabilities, the retrieval system still 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.

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 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 is "The Value of Diaphanograpy 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 probably is interested in articles that discuss neoplasia arising as a result of the transplant (or more likely the immunosuppressive therapy associated with the transplant), but this is not stated directly. A relevant title for this query is, in fact, "Development of Incidence of Cancer Following Cyclosporine Therapy."

In many cases, a system cannot draw the appropriate inferences without the 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 these options, a choice can be made and the query can be refined further.

The 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 successful. The authors intend to continue explorations of these phenomena and have begun to develop an approach to handle some of them. The UMLS Knowledge Sources have been shown to be of direct use in making some of the appropriate inferences.

UMLS KNOWLEDGE

Metathesaurus

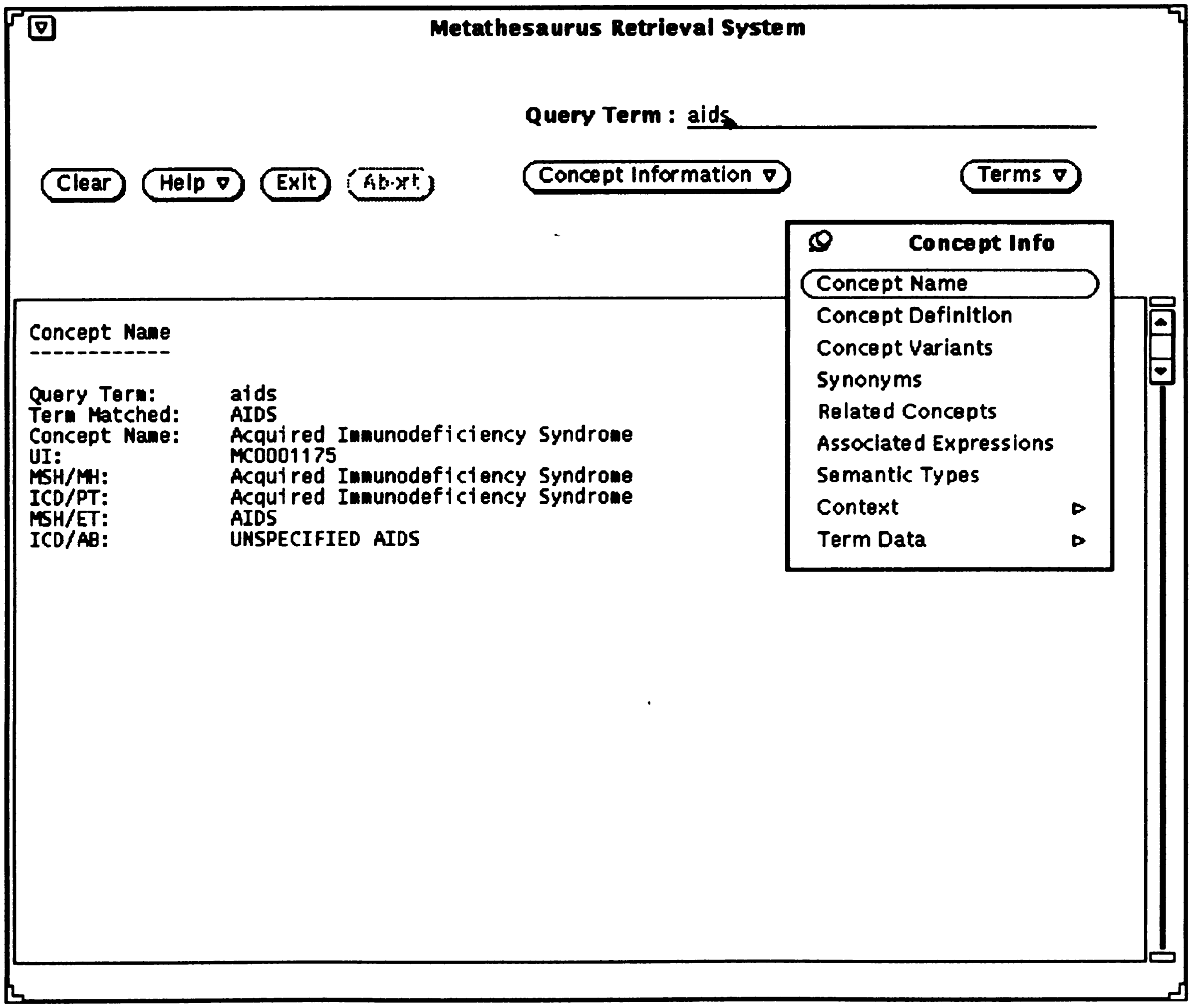
The continually evolving Metathesaurus is a rich source of biomedical vocabulary (see Humphreys et al. [6] and Schuyler et al. [7]). The 1992 release includes over 130,000 concepts with more than 240,000 strings. The authors have developed a UNIX-based retrieval system called "Meta" for browsing, navigating, and extracting information from the Metathesaurus. Meta can be used for batch processing, or it can be used interactively to answer individual user questions. The system can be used to query any of the information in the Metathesaurus for a particular concept. In addition, global searches can be executed that report all concepts with a particular feature or set of features. Figures 1 and 2 are sample screens from the interactive version of Meta.

Figure 1 contains the type of information provided for individual concepts. Note that the query term *AIDS* matches the concept *acquired immunodeficiency syndrome*. The display reveals that this concept also appears in the International Classification of Diseases (ICD) vocabulary. The pop-up menu allows the user to ask for the concept definition, variants, synonyms, related concepts, and semantic types. Choosing the "context" option produces another menu, which allows the user to ask for the ancestors, descendants, parents, and siblings of the current concept. Choosing the "term data" option produces yet another menu, which allows the user to ask questions about the lexical variants (e.g., singular and plural); lexical tags (e.g., acronym, eponym, trade name); and syntactic category (e.g., noun, verb, adjective) assigned to the terms that make up the concept.

Figure 2 depicts a global search for all concepts assigned the semantic types "Injury or Poisoning" and "Disease or Syndrome." The scrolling window shows the top of the list of concepts that fit this description.

The Metathesaurus contains a range of information that is useful for NLP systems [8]. It is an extensive resource of biomedical terminology that can be integrated into and exploited by natural language sys-

Figure 1
Information for individual concepts provided by Meta

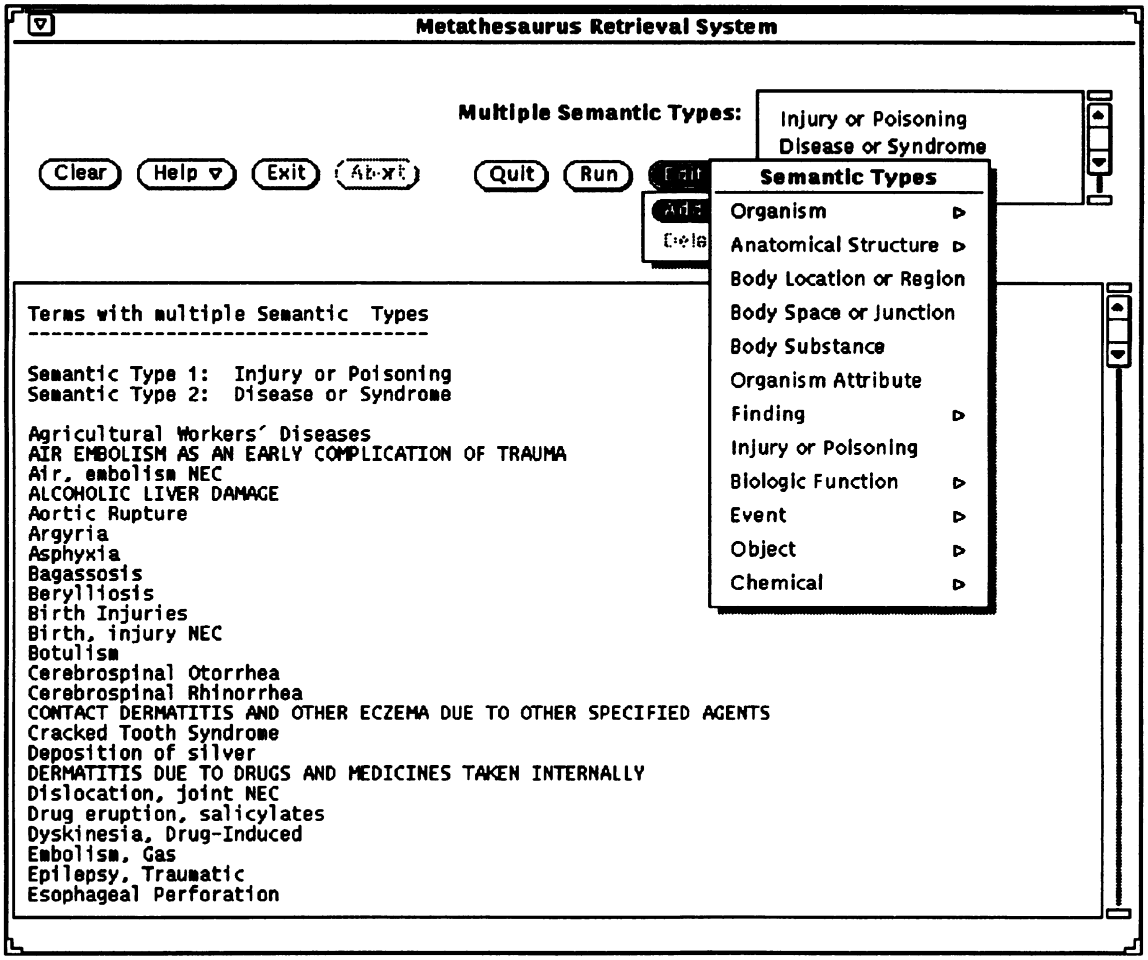


tems. The terms, together with their lexical attributes, can be used as a base in building a lexicon for the biomedical domain. Perhaps the greatest value of the Metathesaurus, however, lies in the multitude of connections it provides between and among biomedical concepts. These relationships become especially important in the context of matching requests to information. As illustrated above, the language of a request and the language in which the information is expressed are almost never identical. Therefore it becomes necessary to calculate, or infer, the relationship between the two. An important aspect of this inferring involves the relationships between lexical items. Many relationships of this type are represented explicitly in the UMLS Metathesaurus.

Semantic Network

The purpose of the UMLS Semantic Network is to provide a consistent categorization of all concepts in the Metathesaurus and to provide useful links between these concepts at the level of the semantic types. The network consists of a set of semantic types and relationships [9–11]. The current set of 134 types includes types denoting physical objects, ideas, activities, biologic functions, anatomical structures, and chemicals. The primary relation is the "isa" link. This link establishes the hierarchy of types within the Network. By traversing the links, an interpretation can be computed for any given node in the network. Figure 3 depicts the top-level hierarchies as they are

Figure 2
Global search for all concepts by Meta

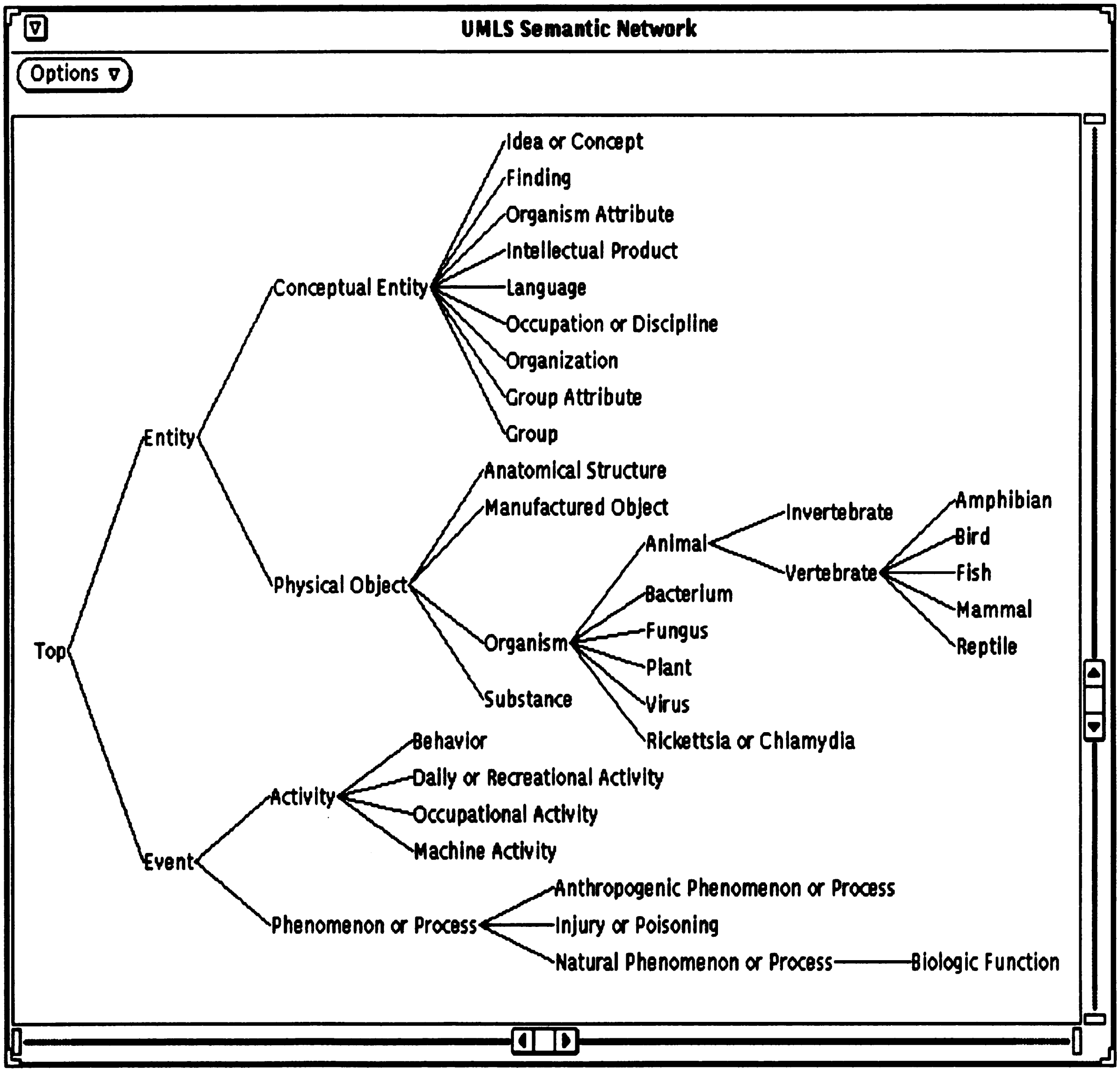


displayed in the browsing and retrieval system developed for the Semantic Network.

By traversing the isa links, starting with "Animal," the following interpretation can be built: an animal is an organism, which is a physical object, which is an entity. This type of knowledge can be used by computer programs as they carry out higher-level reasoning tasks. For example, in attempting to build an analysis for the phrase "pharmacokinetics of cefotiam during hemofiltration," the system needs to know whether the prepositional phrase modifies "pharmacokinetics" or "cefotiam." Prepositional phrases introduced by "during" modify only phrases that refer to events, not phrases that refer to entities. In the Metathesaurus, "pharmacokinetics" has been assigned to the semantic type "Physiologic Function." By traversing the network, it becomes clear a physiologic function is, in fact, an "Event" (Figure 4). "Cefotiam," on the other hand, is a substance (a "Pharmacologic Substance"). Figure 3 demonstrates that substances are entities.

In addition to the isa link, other nonhierarchical links are included in the Network. These links are grouped into four broad categories: temporal, physical, functional, and conceptual relations. The relationships are binary; that is, they link two semantic types in a particular way. For example, the semantic type "Biologic Function" is linked to the semantic type "Organism" by the relationship "process of." Relationships may be stated at any level in the Network, but they generally are stated as high as possible in a type hierarchy. All descendants of the linked types inherit the relationship, but none of their ancestors do. In the example just given, this means that

Figure 3
Hierarchies developed for the Semantic Network



biologic function is a process of plants and animals and therefore also of vertebrates and invertebrates.

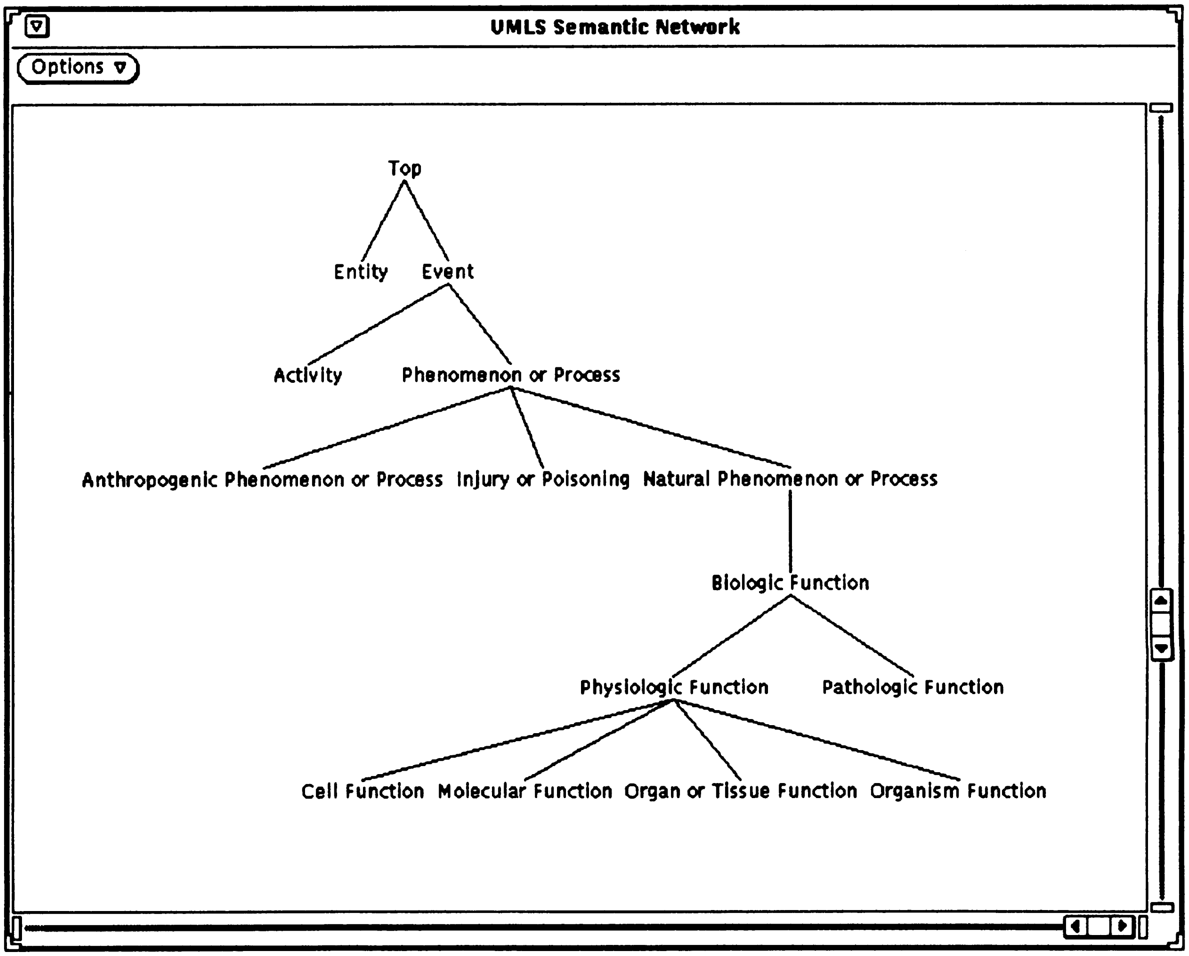
The semantic types cover the full scope of biomedicine, even at a fairly coarse level. It is a testable hypothesis that even this level can be useful for a variety of fairly sophisticated applications. Initial investigation of the use of the network as a domain model for NLP indicates that the semantic types can be used profitably to establish conceptual structures for phrases and sentences in biomedical texts.

It is the authors' view that retrieval of information from textual databases essentially involves mapping one conceptual structure to another. The development of a system that will generate structures of this type is one of the goals of the project. This is described in greater detail in the next section.

THE SPECIALIST SYSTEM

The SPECIALIST system includes several modules based on the major components of natural language [12–15]. The morphological component is concerned

Figure 4
More hierarchies developed for the Semantic Network



with the structure of words and the rules of word formation. The syntactic component treats the constituent structure of phrases and sentences. The semantic component is concerned with the meaning of words, sentences, and discourses. All three rely heavily on the lexical component, which encodes the information specific to the words in the language.

SPECIALIST LEXICON

The lexicon currently contains more than 50,000 lexical entries, with more than 88,000 lexical forms. It includes both general English lexical items as well as items specific to biomedicine. Lexical entries are created using a lexicon-building tool called Lextool, a menu-based system that accepts as input either a file of lexical items or lexical items entered at the keyboard. With the interactive aid of the user, the system generates fully specified lexical frames. Lextool incorporates rules that dictate which slots are permissible for different syntactic categories. These rules have been formalized in a grammar that includes all the allowable slots and values. The grammar serves to constrain the possible choices that must be considered when entering an item, and it also serves as an automatic check of the correct form of completed lexical records. A variety of online information sources are available to lexical coders in the Lextool environment. These include Dorland's Illustrated Medical Dictionary; Meta, the Metathesaurus browser and retrieval system; and the UMLS test collection, which has sentences containing the lexical items in question.

Recent work has focused on the incorporation of semantic information in the lexicon. This task involves identifying the various senses a single lexical item may have and assigning semantic roles and a semantic structure to its complements.* It also involves developing semantic rules to interpret the lexical structures that have been built. The following example will illustrate (numbers in angled brackets have been included for expository purposes, and "etc." has been added to indicate that there are more senses for this verb):

```
{base=treat
entry=1
  cat=verb
  variants=reg
  sense=1 (medical)
⟨1⟩ tran=np[B] | np[A]
⟨2⟩ tran=np[C] | np[A]
⟨3⟩ tran=np[C] | np[D]
⟨4⟩ ditran=np[B],pphr(for,np[C]) | np[A]
⟨5⟩ ditran=np[B],pphr(with,np[D]) | np[A]
⟨6⟩ ditran=np[C],pphr(with,np[D]) | np[A]
    nominalization=treatment
  sense=2 (act upon with some agent)
    tran=np[F] | np[E]
    ditran=np[F],pphr(with,np[G]) | np[E]
    nominalization=treatment
  sense=3 (act toward in a manner)
    cplxtran=np[I],advb[J] | np[H]
  etc.
{semantic_structure_of:treat
cat=verb
  sense=1
    AGENT=A; OBJECT1=B; OBJECT2=C;
        INSTRUMENT=D
    ss= treat1(A,B,C,D),human(A),human(B),
        disorder(C),therapy(D);
  sense=2
    AGENT=E; OBJECT=F; INSTRUMENT=G
    ss= treat2(E,F,G),not_human(F),chemical(G);
  sense=3
    AGENT=H; OBJECT=I; MANNER=J
    ss= treat3(H,I,J),human(H);
etc.
```

The first part of the lexical entry for "treat" gives syntactic and morphological information and indicates the senses this verb may have. For ease of development, informal mnemonic definitions are given in parentheses following each sense. This entry has the syntactic category "verb" and its variants follow a regular inflectional pattern ("treats, treating, treated"). The verb can have a simple noun phrase subject; the possible complements of the verb also are listed. The first sense, for example, can have either one or two objects, marked as "tran" (transitive) or "ditran" (ditransitive), respectively. Roles are assigned to the noun phrases for subsequent use in the semantic portion of the entry where the interpretation for each role is encoded. Thus, the medical sense of "treat" potentially can involve four roles: an "agent" who does the treating, two different "objects" that can be treated, and an "instrument" that is used in the treating.

In the following, meanings are accounted for by combining information from the syntactic and semantic representations for the first sense (numbers refer to numbers in angled brackets above):
⟨1⟩ The physician (A) treated the patient (B).
⟨2⟩ The physician (A) treated the disorder (C).
⟨3⟩ This therapy (D) treats that disorder (C).
⟨4⟩ The physician (A) treated the patient (B) for the disorder (C).
⟨5⟩ The physician (A) treated the patient (B) with the therapy (D).
⟨6⟩ The physician (A) treated the disorder (C) with the therapy (D).

The roles are labeled with tags, such as "therapy" and "disorder," and these labels map directly to UMLS semantic types. For example, the label "disorder" maps to the disorders represented in the Network; these include "Pathologic Function" and all of its descendants—the abnormalities "Congenital Abnormality," "Acquired Abnormality," and "Injury or Poisoning."

SPECIALIST morphology

Inflectional morphology deals with the different lexical forms of a given base. In English, this means nouns are marked for number, verbs for tense, and adjectives and adverbs for their comparative and superlative forms. For example, the lexical item "watch" has the noun inflection "watches" and the verb inflections "watches" (present), "watched" (past), and "watching" (present participle). An adjective such as "lively" has the inflections "livelier" (comparative) and "liveliest" (superlative). To the extent that Greco-Latin inflectional variation is productive in modern English, it can be accounted for in an NLP system. Variations such as those illustrated by "bacterium" and "bacteria," "criterion" and "criteria," and "index" and "indices" are fairly productive and are captured by rule.

Derivational morphology links alternates of lexical items that are related grammatically by affixation, but these generally are not in the same word class. For example, "procedure" is a noun that is related derivationally to the adjective "procedural" by the suffix "-al." Derivational morphology is highly idiosyncratic in English and for that reason it is preferable to store these alternates directly. However, when a particular alternate is missing from the database, rules of morphology can be used heuristically to identify the grammatical relationship between pairs of lexical items.

* Complements are elements of a verb phase (predicate), such as objects, that are needed to complete the meaning of the verb.

Morphological analysis attempts to recognize derivational variants of known words. Common suffix alternations and the categories of the corresponding words are recorded as derivational rules. For example, the suffix rule "acy|noun|ate|adj" indicates that an adjective ending in "-ate" may be related morphologically to a noun ending in "-acy." Words such as "adequate" and "adequacy" are related in this way. The morphological analysis procedure is interleaved with the lexical look-up procedure to ensure that all available information about a word is retrieved. In addition, to minimize the incorrect application of the derivational rules, a list of known exceptions is maintained.

SPECIALIST parser and grammar

As part of natural language interpretation, the SPECIALIST parser assigns syntactic structures to sentences exhibiting a wide range of linguistic phenomena. The syntactic component of SPECIALIST is based on the logic grammar formalism [16–18]†.

The grammar includes context-free phrase structure rules together with context-sensitive restrictions on the structures actually built. The final output of the grammar rules is a predicate argument structure, which shows the relevant logical relations among the major constituents of each clause. Predicates and their arguments are identified, along with other syntactic information crucial to the semantic interpretation of the sentence being processed. Following is a simplified sample parse for the sentence "Rifampin is administered in the treatment of tuberculosis":

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

The parser looked up the words of the sentence in the SPECIALIST lexicon and used the linguistic information associated with each lexical item, together with the appropriate rules, to construct the structured representation shown. The operators (OPS) in this sentence are the present tense and the passive voice. The subject (SUBJ) is marked as null since it is not explicitly present in the sentence. The adverbial prepositional phrase (PP) has further internal structure. It consists of a head noun ("treatment") and a postmodifying prepositional phrase ("of tuberculosis"). Note that a look-up in the UMLS Metathesaurus reveals that the semantic type of the object (OBJ) "rifampin" is "Pharmacologic Substance," the semantic type of "treatment" is "Therapeutic or Preventive Procedure," and the semantic type of "tuberculosis" is "Disease or Syndrome."

Syntactic processing is meant to provide the basis for determining the final interpretation of linguistic structures and, as such, is but one component of a complete NLP system. The parser reliably determines the syntactic structure that supports the semantic interpretation of a particular sentence. However, it typically also produces numerous additional parses that are syntactically correct yet do not contribute to the final interpretation. The authors are exploring a method of eliminating these unwanted parses. This extension exploits the syntactic information available in the lexicon along with a much reduced parser to produce a single syntactic representation for each linguistic structure encountered. This representation depends on flexible semantic interpretation and robust domain knowledge processing to produce a final interpretation, or conceptual structure.

SPECIALIST semantics

The current work involves writing semantic rules with the goal of specifying a conceptual structure for phrases and sentences. This project distinguishes between the relational vocabulary (i.e., verbs, adjectives, relational nouns, and prepositions) and the nonrelational vocabulary. Relational nouns are nouns that take arguments. For example, in the phrase "effects of CPAP treatment on sleep pattern," the relational noun "effect" has the arguments "CPAP treatment" and "sleep pattern." An example of a nonrelational noun is "arthritis." For the most part, the nonrelational vocabulary includes terminology that maps to a representation in the UMLS Knowledge Sources. Thus, semantic interpretation of these words is the associated UMLS information.

Part of the interpretation of the relational vocabulary is the way in which these lexical items contribute to a conceptual structure. For example, these items are predicates with a certain number of arguments and these arguments have specified semantic roles. The UMLS semantic types are used to mark these roles in lexical entries. The major thrust of the interpretation is to discover the relationships among the components of underlying predicate argument structures. The rules for accomplishing this objective depend on the structure obtained from the syntactic analysis as well as the semantic information associated with items in the lexicon.

Many rules depend on the UMLS semantic types

† Due to a successful collaboration between the NLP group at the Paramax (then Unisys) research center during the academic year 1988–1989, the syntactic component of the system is extremely robust.

assigned to the roles in lexical entries. For example, the following rule imposes the same conceptual structure on the three examples below:

Semantic Rule:
⟨⟨ therapy ⟩⟩ FOR ⟨⟨ disorder ⟩⟩→
treat(object(⟨⟨ disorder ⟩⟩), instrument(⟨⟨ therapy ⟩⟩))
Examples:
Electrocoagulation for gastrointestinal hemorrhage.
Continuous positive airway pressure (CPAP) for obstructive sleep apnea.
Topical therapy for oropharyngeal symptoms of myasthenia gravis.

Each of the examples deals with a procedure used in the treatment of some disorder. The rule says that if a sentence or phrase consists of "therapy for disorder," then this is an instance of a "treat" concept. The lexical information for "treat" then is used in building the final conceptual structure. The conceptual structure for the first example is

```
conceptual_structure([
  treat([
    object([ mod( [ string(gastrointestinal) ]),
      head([ string(hemorrhage),
        semtype(disorder])]),
    instrument([ head([ string(electrocoagulation),
      semtype(therapy)])] ]) ])
```

Syntactic and semantic rules, together with information from the lexicon, have resulted in the final conceptual structure shown above. The "treat" concept takes an object and an instrument as arguments. The "object" argument consists of a head noun ("hemorrhage"), which is modified ("mod") by "gastrointestinal." The semantic type of the object is one of the disorders, and the semantic type of the instrument is one of the therapies. The authors are developing an initial set of semantic rules, such as the one shown for handling sample text in the UMLS test collection.

CONCLUSION

The problem of providing users with the information they seek can be viewed as the problem of mapping the language of the user to the language of a database. Users formulate queries in ways that reflect their knowledge and understanding of the topic and expect it will be sufficient to retrieve relevant information from the database. Because of the richness and diversity of natural language, mapping between requests and information is rarely straightforward.

The goal of this research is to establish a more precise understanding of the relationship between user queries and the information that may be relevant to those queries. The hypothesis of providing structured representations of both queries and documents is an essential first step in the mapping process. The mapping then becomes a matter of matching conceptual structures to each other. Since there is often an indirect correspondence between these structures, a variety of inferencing mechanisms must be used.

The UMLS Metathesaurus and Semantic Network provide an extensive knowledge of the biomedical domain. This knowledge is useful in building the conceptual structures that represent the phrases and sentences in biomedical text. The initial investigations have shown that UMLS knowledge is also a powerful resource in the mapping of these conceptual structures to each other for the purpose of effecting successful information retrieval. The variety and depth of the interconnections between biomedical concepts in the UMLS Knowledge Sources form a strong foundation for continued experimentation in language and information processing.

REFERENCES

1. McCray AT. Extending a natural language parser with UMLS knowledge. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care: assessing the value of medical informatics, November 17–20, 1991, Washington, DC. New York: McGraw-Hill, 1992:194–8.
2. McCray AT, Srinivasan S. Automated access to a large medical dictionary: online assistance for research and application in natural language processing. Comput Biomed Res 1990;23:179–98.
3. McCray AT, Sponsler JL, Brylawski B, Browne AC. The role of lexical knowledge in biomedical text understanding. In: Stead W, ed. Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care, November 1–4, 1987, Washington, DC. Los Alamitos, CA: IEEE Computer Society Press, 1987:103–7.
4. Schuyler PL, McCray AT, Schoolman HM. A test collection for experimentation in bibliographic retrieval In: Barber B, Cao D, Qin D, Wagner G, eds. MEDINFO 89, December 10–14, 1989, Singapore, Republic of Singapore. Amsterdam: North Holland, 1989:910–2.
5. Dorland's illustrated medical dictionary. 26th ed. Philadelphia: W. B. Saunders, 1985.
6. Humphreys BL, Lindberg DAB. The UMLS® project: making the conceptual connection between users and the information they need. Bull Med Libr Assoc 1992 Apr;81(2): 170–7.
7. Schuyler PL, Hole WT, Tuttle M, Sherertz DD. The UMLS Metathesaurus®: representing different views of biomedical concepts. Bull Med Libr Assoc 1992 Apr;81(2): 217–22.
8. McCray, Extending a natural language parser.
9. McCray AT, Hole WT. The scope and structure of the first version of the UMLS Semantic Network. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care: standards in medical informatics; November 4–7, 1990, Washington, DC. Los Alamitos, CA: IEEE Computer Society Press, 1990:126–30.

10. SOWA JF, ed. Principles of semantic networks: explorations in the representation of knowledge. San Mateo: Morgan Kaufmann, 1991.
11. LEHMANN F, ed. Semantic networks in artificial intelligence. Oxford: Pergamon Press, 1992.
12. ALLEN J. Natural language understanding. Menlo Park, CA: Benjamin/Cummings, 1987.
13. CHARNIAK E, MCDERMOTT D. Introduction to artificial intelligence. Reading, MA: Addison-Wesley, 1985.
14. LYONS J. Language and linguistics. Cambridge: Cambridge University Press, 1981.
15. CRYSTAL D. The Cambridge encyclopedia of language. Cambridge: Cambridge University Press, 1987.
16. MCCRAY, Automated access to a large medical dictionary.
17. HIRSCHMAN L, DOWDING J. Restriction grammar: a logic grammar. In: Saint-Dixior P, Szpakowicz S, eds. Logic and logic grammars for language processing. New York: Ellis Horwood, 1990:141-67.
18. HIRSCHMAN L. Conjunction in meta-restriction grammar. J Logic Progr 1986;3:299-329.

Received October 1992; accepted November 1992