Semantic Processing for Enhanced Access to Biomedical Knowledge

Thomas C. Rindflesch, Ph.D.
Alan R. Aronson, Ph.D.
National Library of Medicine, Bethesda, Maryland 20894

Abstract. The Semantic Knowledge Representation (SKR) project at the National Library of Medicine (NLM) develops programs that extract usable semantic information from biomedical text by building on resources currently available at NLM. Two programs in particular, MetaMap and SemRep, are being applied to a variety of problems in biomedical informatics. Both programs depend on the biomedical domain knowledge available in the Unified Medical Language System® (UMLS®). In representing concepts and relationships extracted from text, the semantic predications produced by these programs support a variety of applications in biomedical information management, including automatic indexing of MEDLINE® citations, concept-based query expansion, accurate identification of anatomical terminology and relationships in clinical records, and the mining of biomedical text for drug-disease relations and molecular biology information.

1. Introduction

An overwhelming amount of human knowledge is encoded in natural language texts (as opposed to databases), and the grand challenge in information technology is to provide reliable and effective access to this knowledge. For significant advances to be achieved, a richer representation of text is required than is currently available. The Semantic Knowledge Representation (SKR) project at the National Library of Medicine (NLM) develops programs that extract usable semantic information from biomedical text by building on resources currently available at NLM.

The Unified Medical Language System (UMLS) knowledge sources and the natural language processing (NLP) tools provided by the SPECIALIST system are especially relevant. The components of the UMLS provide structured representation of concepts and relationships in the biomedical domain. The Metathesaurus and the SPECIALIST Lexicon taken together represent names for concepts, while the Metathesaurus and the Semantic Network represent relationships between concepts. Natural language processing techniques are then called upon to provide a link between this domain knowledge and text.

Two programs in particular, MetaMap and SemRep, are being applied to a variety of problems in biomedical informatics. MetaMap maps noun phrases in free text to concepts in the UMLS Metathesaurus, while SemRep uses the Semantic Network to determine the relationship asserted between those concepts. As an example of the type of enhanced representation of text we are developing, (2) contains the semantic predication which represent some of the information contained in the text in (1).

(1) We used hemofiltration to treat a patient with digoxin overdose, which was complicated by refractory hyperkalemia.

(2) Digoxin overdose-OCCURS_IN-Patients
Hemofiltration-TREATS-Patients
Hemofiltration-TREATS-Digoxin overdose
Hyperkalemia-COMPLICATES-Digoxin overdose

Each of the predications in (2) is a proposition whose predicate (in upper case) is a relation from the UMLS Semantic Network. Each of the arguments is a concept from the UMLS Metathesaurus. The set of propositions in (2) considered as the semantic representation of (1) is not complete; however, it represents the major relationships and concepts contained in the text.

This approach to NLP, with its heavy dependence on the use of domain knowledge, follows in the tradition of semantics-oriented analysis. Classic work of this type includes that of Wilks [1], Schank [2], Riesbeck [3], and Hahn [4]. Maida and Shapiro [5][6] provide one viewpoint for representing the relationship between the assertions made in text and the interaction of entities expressed in a domain model of some possible world.

Bates and Weischedel [7] emphasize the importance of domain knowledge as a basis for significant progress in natural language processing effectiveness, while Saint-Dizier and Viegas [8] concentrate on lexical semantics in this regard. Sowa [9] discusses a number of important aspects of the interaction of domain knowledge and linguistic analysis.

There is currently a considerable amount of interest in natural language processing of biomedical text. Several approaches are being explored to provide reliable automatic analyses which can support practical applications. See, for example, Haug et al. [10], Hripcsak et al. [11], Friedman et al. [12], Rassinoux et al. [13], and Zweigenbaum et al. [14]. Hahn et al. [15] discuss the design of a semantic interpretation system that relies crucially on domain knowledge resources.

In the remainder of this paper we introduce the UMLS knowledge sources and then provide an overview of the NLP system we are developing in the SKR project. Finally, we describe some examples of applications (both completed and ongoing) that draw on the SKR system.

2. Unified Medical Language System (UMLS)

The UMLS is a compilation of more than 60 controlled vocabularies in the biomedical domain and is being constructed by the National Library of Medicine under an ongoing research initiative (Lindberg et al. [16], Humphreys et al. [17]) that supports applications in processing, retrieving, and managing biomedical text. The original hierarchical structure and relationships from each vocabulary are maintained, while synonymous terms across vocabularies are grouped into concepts. Information beyond that found in the constituent vocabularies is added by the UMLS editors, including semantic categories.

Component terminologies that provide broad coverage of the domain include Medical Subject Headings (MeSH®), Systematized Nomenclature of Medicine (SNOMED), International Statistical Classification of Diseases and Related Health Problems (ICD), Physicians' Current Procedural Terminology (CPT), and Clinical Terms Version 3 (Read Codes). Information focused in subdomains of medicine can be found in vocabularies such as Diagnostic and Statistical Manual of Mental Disorders (DSM), Classification of Nursing Diagnoses (NAN), WHO Adverse Drug Reaction Terminology (WHOART), and the University of Washington Digital Anatomist Symbolic Knowledge Base (UWDA). The Metathesaurus also contains a number of terminologies and vocabularies in languages other than English.

The UMLS is structured around three separate components: The Metathesaurus, the SPECIALIST Lexicon, and the Semantic Network. At the core is the Metathesaurus, which contains semantic information about more than 800,000 biomedical concepts, each of which has variant terms with synonymous meaning. Figure 1 shows some of the Metathesaurus information for Gaucher's Disease. Hierarchical structure from constituent vocabularies

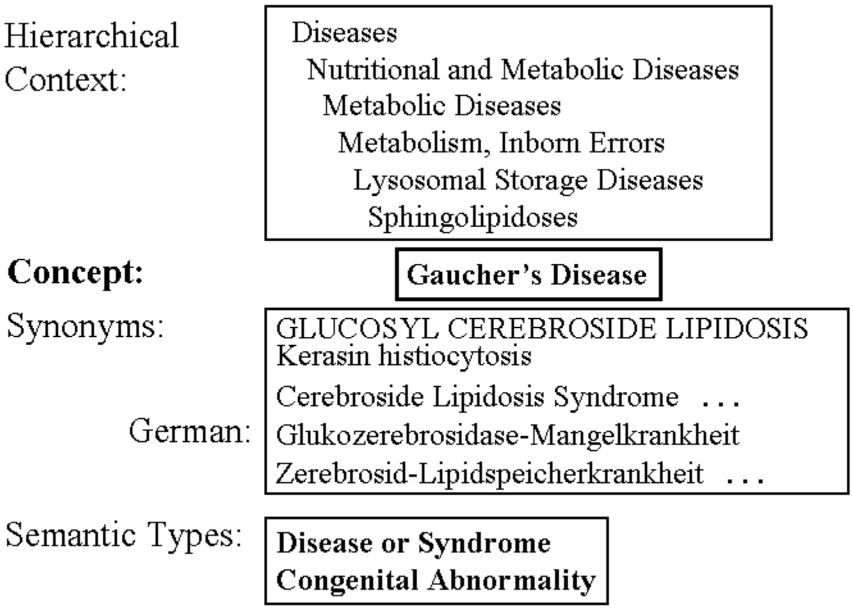
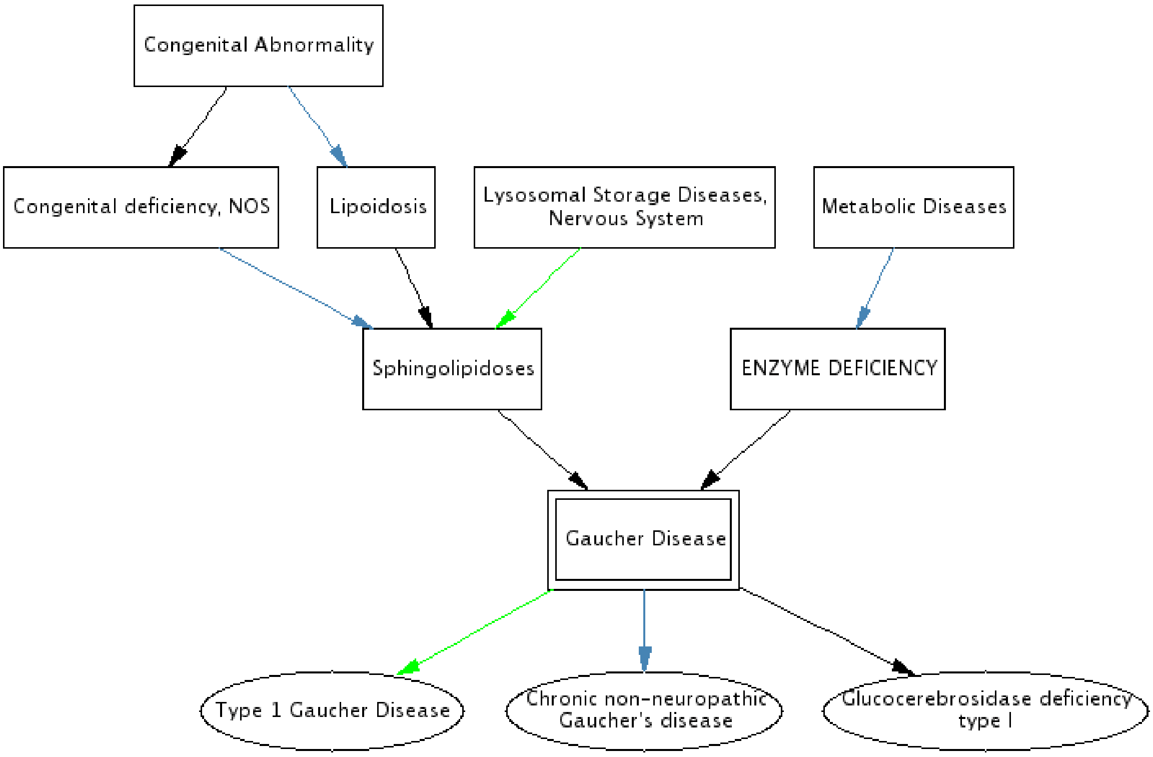
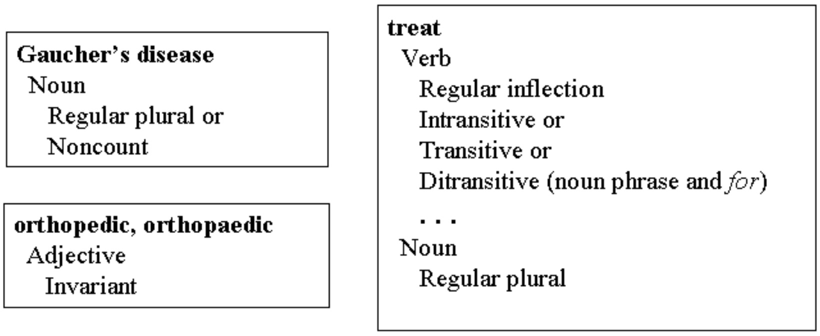

Hierarchical Context:
Diseases
Nutritional and Metabolic Diseases
Metabolic Diseases
Metabolism, Inborn Errors
Lysosomal Storage Diseases
Sphingolipidoses
Concept:
Gaucher's Disease
Synonyms:
GLUCOSYL CEREBROSIDE LIPIDOSIS
Kerasin histiocytosis
Cerebroside Lipidosis Syndrome ...
German:
Glukozerebrosidase-Mangelkrankheit
Zerebrosid-Lipidspeicherkrankheit ...
Semantic Types:
Disease or Syndrome
Congenital Abnormality

Figure 1. Part of the Metathesaurus concept for "Gaucher's Disease"


Congenital Abnormality
Congenital deficiency, NOS
Lipoidosis
Lysosomal Storage Diseases, Nervous System
Metabolic Diseases
Sphingolipidoses
ENZYME DEFICIENCY
Gaucher Disease
Type 1 Gaucher Disease
Chronic non-neuropathic Gaucher's disease
Glucocerebrosidase deficiency type I

forms the basis for relationships among concepts seen in the Metathesaurus. Using Gaucher's Disease as an example again, Figure 2 displays some of its relationships to other Metathesaurus concepts. English terms from the Metathesaurus are included in the SPECIALIST Lexicon, which contains more than 140,000 entries of general and medical terms and stipulates morphological and syntactic facts about English verbs, nouns, adjectives and adverbs (see Figure 3). Each concept in the Metathesaurus is also assigned a semantic category (or type), which appears in the Semantic Network, in which 134 semantic types interact with 54 relationships. Some of these semantic types and their relationships are shown in Figure 4.



Gaucher's disease
Noun
Regular plural or
Noncount

orthopedic, orthopaedic
Adjective
Invariant

treat
Verb
Regular inflection
Intransitive or
Transitive or
Ditransitive (noun phrase and *for*)
. . .
Noun
Regular plural

Figure 3. Examples of entries in the SPECIALIST Lexicon

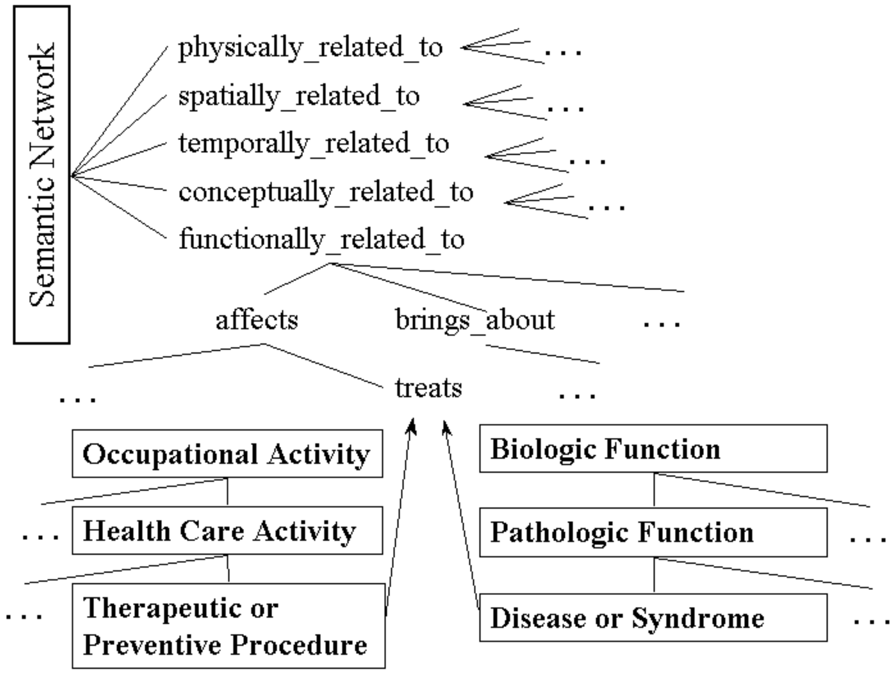


Figure 4. Part of the Semantic Network

3. Semantic Interpretation of Biomedical Text

Semantic processing in the SKR project draws on the resources being developed in the SPECIALIST NLP system (McCray et al. [18]; McCray [19]), which provides a framework for exploiting the resources of the UMLS in processing biomedical text. In addition to the Metathesaurus and Semantic Network, the SPECIALIST Lexicon and associated lexical variant programs (McCray, Srinivasan, and Browne [20]) as well as the Knowledge Source Server (McCray et al. [21]) support syntactic analysis and semantic interpretation of free text in the biomedical domain. At the core of the SKR effort are two programs, MetaMap (Aronson, Rindflesch, and Browne [22]; Rindflesch and Aronson [23]; Aronson [24]; Aronson [25]) and SemRep (Rindflesch and Aronson [26]; Rindflesch [27]; Rindflesch, Rajan, and Hunter [28]), which work in concert to provide the semantic representation given in example (2) above. An overview of NLP in the SPECIALIST system is given in Figure 5.

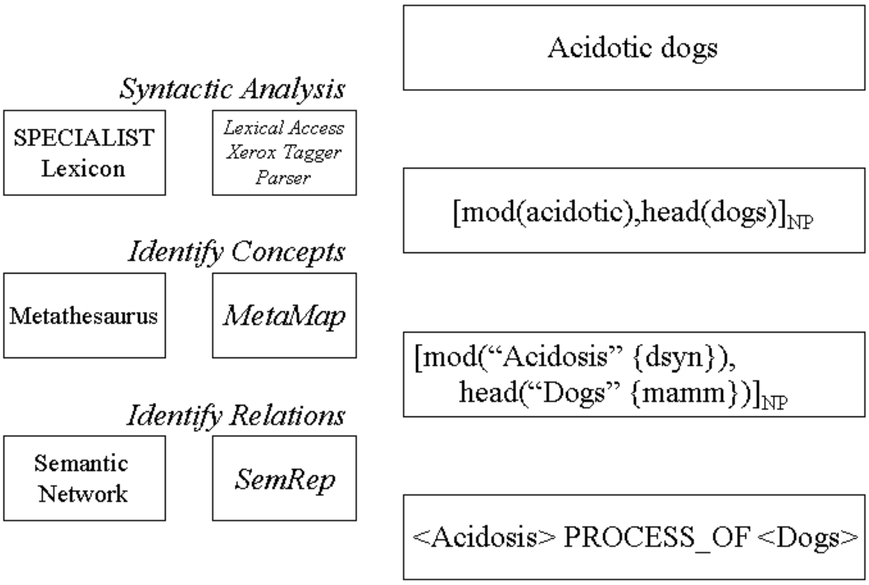

Acidotic dogs
Syntactic Analysis
SPECIALIST Lexicon
Lexical Access Xerox Tagger Parser
[mod(acidotic),head(dogs)]$_{NP}$
Identify Concepts
Metathesaurus
MetaMap
[mod("Acidosis" {dsyn}), head("Dogs" {mamm})]$_{NP}$
Identify Relations
Semantic Network
SemRep
<Acidosis> PROCESS_OF <Dogs>

Figure 5. Overview of NLP in the SPECIALIST system

The SPECIALIST system begins analysis of biomedical text by consulting the Lexicon to determine syntactic information for each lexical entry in the input. A stochastic tagger (Cutting et al. [29]) is called to resolve part-of-speech ambiguities, and an underspecified syntactic analysis is produced as the basis for further processing. For example, input text *ablation of pituitary gland* is given the following analysis:

(3) [[head(ablation)] [prep(of), head(pituitary gland)]]

Although noun phrases are correctly identified, this analysis is underspecified in the sense that overall structure is not provided. That is, no commitment has been made to the exact relationship between the two constituent phrases produced. A further example of the characteristics of this type of analysis is given in (4), which is the underspecified analysis of the input text *pancreatic secretory trypsin inhibitor.*

(4) [[mod(pancreatic), mod(secretory), mod(trypsin), head(inhibitor)]]

In particular, note that, although the head of the noun phrase and its modifiers have been identified, no indication is given of the internal syntactic structure of such phrases. It is our hypothesis that this attenuated analysis is sufficient to serve as the basis for usable semantic interpretation.

The next step in processing calls MetaMap to get concepts from the Metathesaurus. This program takes advantage of syntactic analysis and considers each noun phrase individually as it proceeds. For example, it takes as input the underspecified syntactic analysis of *ablation of pituitary gland* and finds the following Metathesaurus concepts:

(5) Excision, NOS (‘Therapeutic or Preventive Procedure’, ‘Research Activity’)
Pituitary Gland (‘Body Part, Organ, or Organ Component’)

MetaMap accomplishes its task in four steps:

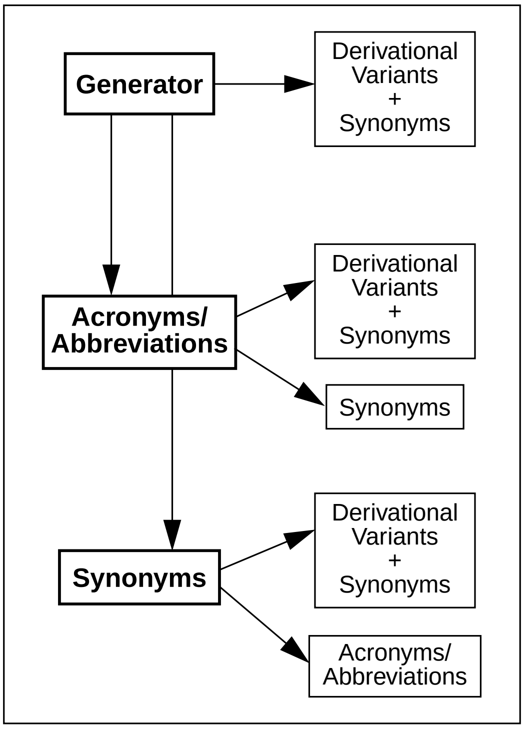

Generator
Derivational Variants + Synonyms
Acronyms/ Abbreviations
Derivational Variants + Synonyms
Synonyms
Synonyms
Derivational Variants + Synonyms
Acronyms/ Abbreviations

Figure 6. MetaMap variant generation (before inflections and spelling variants are computed)

- Variant generation: Each input word (or multi-word item, such as *wood alcohol*) generates a list of morphological variants, synonyms, and (optionally) acronyms/abbreviations plus meaningful combinations of these variants (Figure 6). For example, *aortic* and *arteria aorta* are variants of *aorta*;
- Candidate retrieval: Metathesaurus strings containing one or more of the input words are retrieved as candidates for a mapping. Some candidates for *arteriosclerosis* (in browse mode) are “Arteriosclerotic” and “Vascular Sclerosis”;
- Candidate evaluation: Each candidate is evaluated for how closely it matches the input text according to a function with four components, centrality, variation, coverage and cohesiveness; and
- Mapping formation: Finally, candidates matching different parts of the input text are combined into a single mapping and re-evaluated to compute a total mapping score.

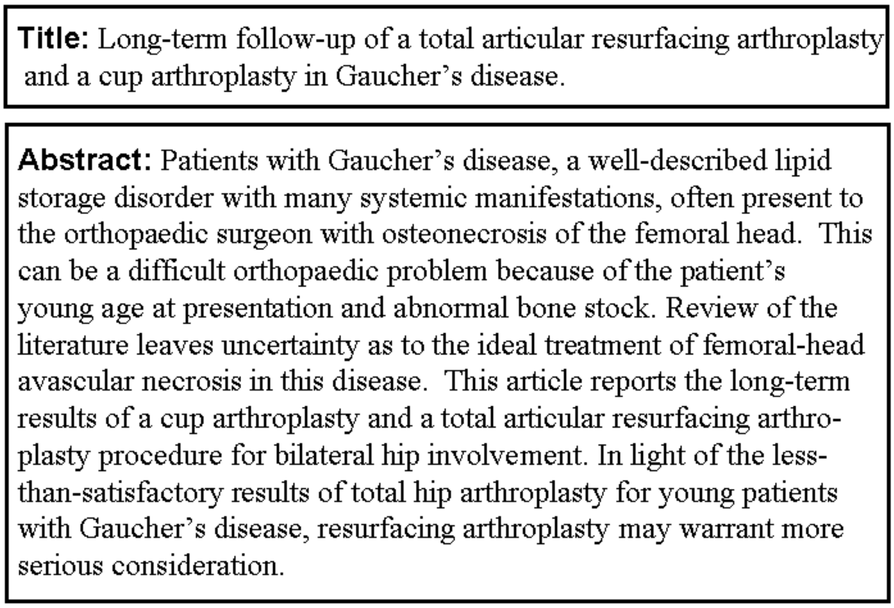
SemRep is called next and depends on both syntactic analysis and the Metathesaurus concepts provided by MetaMap. In addition, it consults the Semantic Network as part of the process of producing a final semantic interpretation. For example, in assigning an interpretation to *ablation of pituitary gland,* SemRep notes the syntactic analysis given for this input and then consults a rule which states that the preposition *of* corresponds to the Semantic Network relation LOCATION_OF, and further notes that one of the relationships in the Semantic Network with this predicate is

(6) Semantic Type 1: ‘Body Part, Organ, or Organ Component’
Relation: LOCATION_OF
Semantic Type 2: ‘Therapeutic or Preventive Procedure’

The MetaMap output for this input is then consulted, and it is noted that the Metathesaurus concept for the text phrase *ablation* is “Excision, NOS.” The semantic type for this concept is ‘Therapeutic or Preventive Procedure,’ while the type for “Pituitary Gland” is ‘Body Part, Organ, or Organ Component.’ Since these semantic types match those found in the relationship indicated by the preposition *of* (LOCATION_OF) and since the relevant noun phrases are allowable arguments of the preposition *of*, (7) is produced as the semantic interpretation for this phrase, where the corresponding Metathesaurus concepts are substituted for the semantic types in the Semantic Network relationship.

(7) Pituitary Gland-LOCATION_OF-Excision, NOS

A final example illustrates the application of MetaMap and SemRep to an entire MEDLINE citation (Figure 7). The output (Figure 8) provides a structured semantic overview of the contents of this citation. Metathesaurus terms found by MetaMap in the text ("Patients," "Arthroplasty," "Cups," "Gaucher's Disease," "Osteonecrosis," and "Femur Head") represent the concepts central to this discourse. (Three concepts have been assigned an "aboutness" value indicating their high saliency.) UMLS semantic types (such as 'Therapeutic or Preventive Procedure' and 'Pathologic Function') associated with each concept provide



Title: Long-term follow-up of a total articular resurfacing arthroplasty and a cup arthroplasty in Gaucher's disease.

Abstract: Patients with Gaucher's disease, a well-described lipid storage disorder with many systemic manifestations, often present to the orthopaedic surgeon with osteonecrosis of the femoral head. This can be a difficult orthopaedic problem because of the patient's young age at presentation and abnormal bone stock. Review of the literature leaves uncertainty as to the ideal treatment of femoral-head avascular necrosis in this disease. This article reports the long-term results of a cup arthroplasty and a total articular resurfacing arthroplasty procedure for bilateral hip involvement. In light of the less-than-satisfactory results of total hip arthroplasty for young patients with Gaucher's disease, resurfacing arthroplasty may warrant more serious consideration.

Figure 7. Title and abstract from a MEDLINE citation

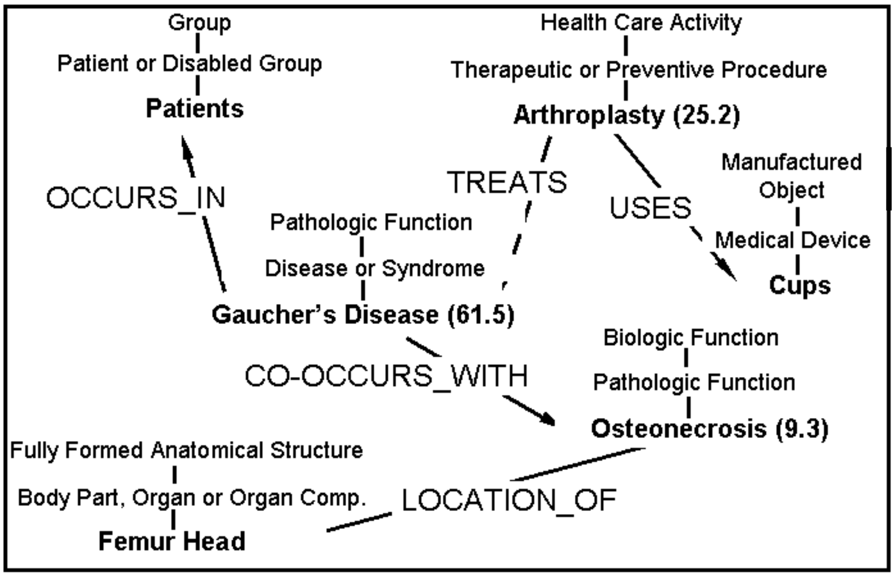


Figure 8. Semantic representation for the citation in Figure 7

more general categorization. SemRep has computed Semantic Network relations between the concepts, which further enrich the representation of the content of this text.

4. Semantic Representation in Information Management Applications

SKR processing serves as the basis for a number of research projects that investigate the use of semantic knowledge representation for enhanced management of biomedical information. These projects are being conducted in collaboration with investigators at NLM and at other institutions.

MetaMap has been used for query expansion and indexing in research on concept-based information retrieval and in support of literature-based discovery systems. Most notably, MetaMap constitutes one of the core components of the Indexing Initiative system, which suggests automatically-generated indexing terms for MEDLINE citations.

SemRep has been applied to processing the research literature as well as clinical text. One project investigates data mining of drug-treatment relations from MEDLINE citations, while another identifies clinical findings in the literature on Parkinson’s Disease. SemRep has also been applied to the task of identifying arterial branching relations in cardiac catheterization reports. Two further projects are aimed at extracting molecular biology information from text. The first of these addresses macromolecular binding relations, while the other is concerned with the interaction of genes, drugs, and cells in the research literature on molecular pharmacology for cancer treatment.

4.1 Information Retrieval Applications

Recent work (including that of Srinivasan [30][31][32][33]) has demonstrated the importance of query expansion based on retrieval feedback for improving retrieval effectiveness when applying statistically-based systems to MEDLINE citations. As an alternative method of query expansion, we have used MetaMap for associating Metathesaurus concepts with the original query. Our experiments show that this methodology compares favorably with retrieval feedback (Aronson and Rindflesch [34]).

MetaMap also served as the basis for research exploring full-text retrieval combined with techniques for hierarchical indexing (Wright, Grossetta Nardini, Aronson, and Rindflesch [35]). A subset of NLM’s Health Services/Technology Assessment Text (HSTAT) database was processed with MetaMap, and the resulting Metathesaurus concepts were used in a hierarchical indexing method supporting information retrieval from full-text sources. Informal experiments suggested the value of this approach for improving results in both source and document selection when accessing large, multiple-source full-text document collections.

MetaMap was initially developed for improved retrieval of MEDLINE citations. The methodology was tested by applying this program to the queries and citations of the NLM Test Collection, replacing text with the Metathesaurus concepts discovered by MetaMap. Retrieval experiments using SMART were performed both on the unmodified test collection and on the MetaMapped version of the collection. The result was a 4% increase in average precision (Aronson, Rindflesch, and Browne [22]).

The Medical Text Indexer (MTI) system has been developed as part of the NLM Indexing Initiative (Aronson et al. [36]). In this project, several indexing methodologies are applied to the task of automatically indexing the biomedical literature, especially MEDLINE citations. The MTI system consists of three fundamental indexing methods:

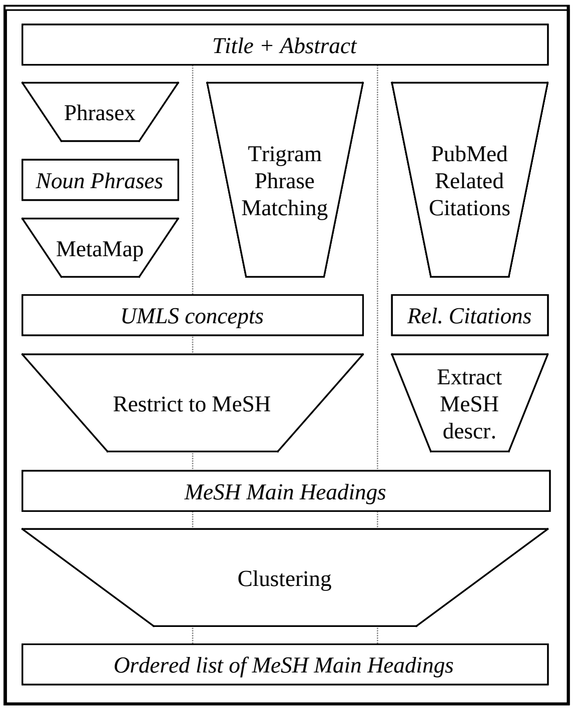


Figure 9. The Medical Text Indexer (MTI)

- MetaMap Indexing (MMI), a linguistically rigorous method in which concepts found by MetaMap are ranked emphasizing either presence in the title or frequency of occurrence and the specificity of the concepts according to their depth in the MeSH hierarchy;
- Trigram Phrase Matching, a statistical method employing character trigrams to define a notion of phrases that are matched against Metathesaurus concepts; and
- PubMed® Related Citations, a variant of the “Related Articles” feature in PubMed to find articles that are similar to a citation of interest; selected MeSH headings from the most closely matching citations are the result of this indexing method.

The first two basic methods produce UMLS Metathesaurus concepts that are then mapped to MeSH headings by the Restrict to MeSH method. This method uses relationships among Metathesaurus concepts to find the MeSH heading which is semantically closest to a given concept. The results of all the basic methods are combined using a Clustering method which generates a final ranking using several system parameters including weights for each of the basic methods.

The MTI system has been recently augmented with several postprocessing steps designed to increase the conformity of its recommendations to NLM indexing policy. MeSH headings which are never used for indexing are dropped; a choice is made between two headings, one of which is more specific than the other; modifications to the ranking of certain headings such as chemicals are performed; and several other similar changes are made. The result is a smaller number of recommendations which are generally more accurate than the original list.

The MTI system is beginning to be used both in computer-assisted and in fully automatic environments. NLM indexers can refer to the system’s recommendations as they index MEDLINE citations using NLM’s Document Control Management System (DCMS). MTI indexing is also being used in the NLM Gateway, a single retrieval system for accessing many of NLM’s information resources, to index some collections which will not receive manual indexing. The collections include meeting abstracts on AIDS/HIV, health services research, and space life sciences.

4.2 Literature-Based Discovery

The DAD system developed by Weeber [37] is a concept-based literature discovery tool that enables biomedical researchers to explore domains which are not familiar to them but which may nonetheless provide information relevant to their research. Inspired by the work of Swanson [38][39], the DAD system relates knowledge about a disease C to a therapeutic substance A via a link B, typically a physiological process. So, for example, Swanson discovered that fish oil has therapeutic value for patients with Raynaud's Disease. The link here was the effect that fish oil has on platelet aggregation. The DAD system uses concepts found by MetaMap in MEDLINE citations to emulate the search for links between diseases and therapeutic substances, for example. The algorithm is focused by concentrating on concepts with appropriate semantic types at each stage of the search process. The combination of using concepts instead of words together with restriction by semantic types produces a more accurate as well as efficient methodology. Besides replicating Swanson's discoveries on Raynaud's Disease and another one on the therapeutic effect of magnesium on migraine headaches, the DAD system has been used to generate a hypothesis regarding the therapeutic application of an existing drug to a disease unrelated to the original intended use of the drug.

4.3 Text Mining

Text mining applications seek to extract knowledge from text that was not explicitly present in the source being mined. For example, MeSHmap (Srinivasan [40]) uses MeSH heading subheading combinations (the indexing terms assigned to MEDLINE abstracts) in order to provide semantic summaries of sets of documents retrieved by a user. Such summaries allow the user to explore relationships that were not overtly asserted in the input texts.

An extension of the MeSHmap technology (Srinivasan and Rindflesch [41]) uses SemRep in cooperation with MeSH indexing terms to provide increased confidence in identifying potentially interesting semantic relationships in large sets of MEDLINE citations. An example of the methodology is discussed on the basis of a set of citations having MeSH indexing terms in which a drug concept is modified by the subheading "therapeutic use" and a disease concept is modified by "drug therapy." For all such citations, the semantic interpretation of the title was obtained from SemRep. For example, the title in (8) was given the interpretational in (9). Two indexing terms assigned to the citation having this title are shown in (10).

(8) Pentoxifylline in cerebrovascular dementia

(9) Pentoxifylline-TREATS-Dementia

(10) Dementia, Multi-Infarct/diagnosis/*drug therapy/etiology
Pentoxifylline/administration & dosage/pharmacology/*therapeutic use

Relevant MeSH indexing terms combined with SemRep predications were extracted from more than 15,000 MEDLINE citations discussing drug therapies and diseases. Of the 7,332 drug-disease pairs identified, the five most frequent are shown in Table 1.

Further research is planned on the basis of concept pairs such as those in Table 1. When the entire list of extracted pairs is examined, it can be determined that certain drugs have been discussed in disease contexts of varying diversity. For example Srinivasan and Rindflesch [41] report that pyrithrioxin appears in a rather homogeneous context (largely Alzheimer disease and dementia), while pyridazines have been associated with an array of disorders, including congestive heart failure, depressive disorders, and the common cold. It is appealing to suggest that research such as this, in computing a "diversity index" for drugs and diseases encountered in the research literature, may provide useful information to the health care practitioner as well as the researcher. Intuitively, information about drugs with a high diversity index, such as pyridazines, may stimulate discovery regarding diseases and effective therapies.

Table 1. Most-Frequent Drug-Disease Pairs ("Occurrence" indicates the number of citations in which a pair appeared.)

Drug concept	Disease Concept	Occurrence
antihypertensive agents	hypertension	66
nifedipine	angina pectoris	43
calcium channel blockers	angina pectoris	41
atenolol	hypertension	39
propanolamines	hypertension	34

4.4 Processing Clinical Data

Focused use of MetaMap along with a number of rules describing the internal structure of findings formed the core of the FINDX program, which was developed to identify findings in clinical text and MEDLINE abstracts (Sneiderman, Rindflesch, and Aronson [42]). One of the conclusions drawn from this study is that the structure of findings in the research literature and in clinical text is essentially the same: an attribute of the patient under consideration is reported along with a value for that attribute. For example, FINDX identified findings such as *elevated liver function tests* in clinical text and findings such as *eight demented cases had absent neocortical neurofibrillary tangles* in the research literature. The program proceeds by first MetaMapping the input text to Metathesaurus concepts. A set of findings rules then looks for concepts having semantic types, such as 'Physiologic Function', or 'Laboratory Procedure' occurring in close syntactic proximity to values such as *absent, elevated, normal,* etc. FINDX was evaluated on a set of MEDLINE citations discussing the diagnosis of Parkinson's disease. A potential use for this application is filtering information retrieval results on the basis of the findings observed, perhaps in support of evidence-based medicine and clinical decision making.

Rindflesch, Bean, and Sneiderman [43] report on the use of MetaMap and SemRep for processing cardiac catheterization reports. Statements in the arteriography section of these reports describe characteristics of the arteries seen during the catheterization procedure, such as branching configurations and areas of stenosis. This project focused first on identifying the names for the coronary arteries and then on retrieving the branching relations that were asserted to obtain among the arteries observed. The focused nature of this application along with the complexity of the arterial terminology provided a useful context for development of the SKR methodology. Extensive reliance on UMLS domain knowledge contributed significantly to a highly accurate semantic analysis for these reports. For example, this processing identified the branching predications given in (12) for the text in (11), and the names for the coronary arteries in the text have been normalized to the corresponding Metathesaurus concepts.

(11) The left main gives off a left circumflex and left anterior descending branches.

(12) Anterior interventricular branch of left coronary artery-BRANCH_OF-Left coronary artery
Circumflex branch of left coronary artery-BRANCH_OF-Left coronary artery

The results from this project suggest the feasibility of extending this processing to a more comprehensive normalization of the semantic content of anatomically-oriented text. Such processing could support innovative applications of information management applied to both clinical text and the research literature.

4.5 Molecular Biology Applications

Several recent SKR projects involve the adaptation and extension of MetaMap and SemRep for extracting molecular biology information from the research literature. One such program, Arbiter, identifies macromolecular binding relationships in MEDLINE citations. Arbiter operates in two phases; the first (Rindflesch, Hunter, and Aronson [44]) identifies all binding entities mentioned in the input text and addresses such phenomena as molecules, genomic structures, cells and cell components, as well as topographic aspects of molecules, cells, and cell components. In order to identify these entities, Arbiter relies on MetaMap output and UMLS semantic types, such as ‘Amino Acid, Peptide, or Protein’, ‘Nucleotide Sequence’, ‘Carbohydrate’, ‘Cell Component’. In addition, Arbiter calls on a small set of words that appear as heads of binding terms. These include words referring to various biomolecular phenomena (*box, chain, sequence, ligand,* and *motif*), molecular or cellular topography (*spike, cleft, groove, surface*), and general terms for bindable entities (*receptor, site, target*).

During the second phase of processing, Arbiter establishes binding relationships using the general SemRep machinery, focused on forms of the verb *bind* (Rindflesch, Rajan, and Hunter [28]). Binding relationships are semantically constrained to obtain only between the binding entities identified during the first phase of processing. As an example of Arbiter output, the predication in (14) was extracted from the text in (13).

(13) In support of this notion, we have found that aminohexose pyrimidine nucleoside antibiotics, which bind to the same region in the 28S rRNA that is the target site for anisomycin, are also potent activators of SAPK/JNK1.

(14) aminohexose pyrimidine nucleoside antibiotic-BINDS-28s rrna

Arbiter was evaluated on a test collection of 116 MEDLINE citations and was then run on 500,000 citations; some 350,000 binding predications were extracted and entered into a database for further analysis. Current research on Arbiter includes extending its application to protein-protein interactions in general (Sarkar and Rindflesch [45]) as a basis for investigating protein function similarities.

Molecular pharmacology for cancer therapy is characterized by the complexity involved in the interaction between drugs, genes, and cells. Genes affect drug activity and drugs affect gene expression; at the same time, both gene expression and drug activity vary across cell types. SKR and UMLS resources are being used as the basis for the development of a program called Edgar (Rindflesch, Tanabe, Weinstein, and Hunter [46]) that is being designed to address this complexity. The program is designed to first identify drugs, genes, and cells in text and then to determine interactions such as “over expresses” that involve these entities. Edgar identifies drugs, genes, and cells in MEDLINE citations using techniques similar to those used by Arbiter. Gene identification is enhanced by calling on several statistical and empirical methods devised by Tanabe and Wilbur ([47]). Although identification of semantic relationships in this domain is still under development, SemRep underpins techniques being developed to extract, for example, the predications in (16) from the text in (15).

(15) Furthermore, RA treatment enhanced the transcriptional activity of a reporter construct containing the Sry/Sox consensus sequence in TC6 cells.

(16) RA-INCREASES_EXPRESSION-Sry/Sox
Sry/Sox-IN-TC6 cells

Libbus and Rindflesch ([48]) report on a project that draws on SKR resources to construct a general tool (called GBD) intended to help researchers manage the literature in molecular biology. GBD is designed to processes MEDLINE citations returned by searches to PubMed. A pilot project seeks to identify and extract information regarding the genetic basis of disease. GBD calls on MetaMap to identify diseases and associated clinical findings in the citations retrieved, while the methods of Tanabe and Wilbur ([47]) are used to tag genomic phenomena such as genes, alleles, mutations, polymorphism, and chromosomes. Once such information has been identified in the group of citations returned by PubMed, further processing by GBD determines distributional and cooccurrence patterns for user-selected categories. For example, the user can request a list of genes that cooccur with a specified disease in the output from PubMed; these lists contain links to the citations retrieved, as illustrated in (17).

(17) 20015025|Diabetes Mellitus, Non-Insulin-Dependent|ipf-l
20015026|Diabetes Mellitus, Non-Insulin-Dependent|basal insulin promoter
20015026|Diabetes Mellitus, Non-Insulin-Dependent|insccg243
20053748|Diabetes Mellitus, Non-Insulin-Dependent|g20r
20053748|Diabetes Mellitus, Non-Insulin-Dependent|g385v

Such lists can be generated for cooccurrence of more than one gene with a disease or a combination of genes and clinical findings. The clustering of PubMed output into categories dynamically specified by the user contributes to effective management of the current research literature.

5. Summary

The Semantic Knowledge Representation project seeks to provide usable semantic representation of biomedical text by building on resources currently available at the Library, especially the UMLS knowledge sources and the natural language processing tools provided by the SPECIALIST system. Two existing programs, MetaMap and SemRep, are being evaluated, enhanced, and applied to a variety of problems in the management of biomedical information. These include automatic indexing of MEDLINE citations, concept-based query expansion, accurate identification of anatomical terminology and relationships in clinical records, and the mining of biomedical text for drug-disease relations and molecular biology information.

Current research is investigating the application of SKR resources to applications such as question answering systems, image retrieval, and structured browsing and navigation facilities. The concepts and relationships that underlie the semantic structures produced by MetaMap and SemRep are drawn largely from the domain knowledge contained in the UMLS knowledge sources. Although the UMLS has broad coverage of the biomedical domain, there are gaps, particularly in the area of molecular biology. For example, only about half of a set of disease-related gene names and gene products occurring in the National Center for Biomedical Information (NCBI) database, LocusLink, are found in the Metathesaurus. As a way of filling this terminological gap, an effort is underway to augment the domain knowledge available to MetaMap with protein and gene names from LocusLink as well as SWISS-PROT and its supplement, TrEMBL. It is expected that the increased coverage of gene terminology will result in a corresponding increase in the accuracy of the knowledge extraction systems using MetaMap as a basic component ([28][44][45][46][48]).

References

[1] Wilks, Yorick A. 1976. Parsing English II. In Eugene Charniak and Yorick Wilks (eds.) *Computational semantics: An introduction to artificial intelligence and natural language comprehension.* Amsterdam: North-Holland Publishing Co.

[2] Schank, Roger C. 1975. *Conceptual information processing.* Amsterdam: North-Holland Publishing Co.

[3] Riesbeck, Christopher K. 1981. Perspectives on parsing issues. *Proceedings of the Nineteenth Annual Meeting of the Association for Computational Linguistics*, 105-6.

[4] Hahn, Udo. 1989. Making understanders out of parsers: Semantically driven parsing as a key concept for realistic text understanding applications. *International Journal of Intelligent Systems* 4(3):345-93.

[5] Maida, Anthony S., and Stuart C. Shapiro. 1982. Intensional concepts in propositional semantic networks. *Cognitive Science* 6:291-330. [reprinted in Brachman and Levesque 1985]

[6] Brachman, Ronald J., and Hector J. Levesque (eds.) 1985. Readings in knowledge representation. Los Altos, CA: Morgan Kaufman Publishers, Inc.

[7] Bates, M., and Weischedel R. M. 1993. *Challenges in natural language processing.* Cambridge: Cambridge University Press.

[8] Saint-Dizier, Patrick, and Evelyne Viegas (eds). 1995. *Computational lexical semantics*. Cambridge: Cambridge University Press.

[9] Sowa, John F. 2000. *Knowledge representation: Logical, philosophical, and computational foundations.* Pacific Grove, CA: Brooks/Cole.

[10] Haug, Peter J.; Spence Koehler; Lee Min Lau; Ping Wang; Roberto Rocha; and Stanley M. Huff. 1995. Experience with a mixed semantic/syntactic parser. In Reed M. Gardner (ed.) *Proceedings of the 19th Annual Symposium on Computer Applications in Medical Care*, 284-8.

[11] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. 1995. Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of Internal Medicine* 122(9):681-8.

[12] Friedman, Carol; Stephen B. Johnson; Bruce Forman; and Justin Starren. 1995. Architectural requirements for a multipurpose natural language processor in the clinical environment. In Reed M. Gardner (ed.) *Proceedings of the 19th Annual Symposium on Computer Applications in Medical Care*, 347-51.

[13] Rassinoux, Anne-Marie; Judith C. Wagner; Christian Lovis; Robert H. Baud; Alan Rector; and Jean-Raoul Scherrer. 1995. Analysis of medical texts based on a sound medical model. In Reed M. Gardner (ed.) *Proceedings of the 19th Annual Symposium on Computer Applications in Medical Care*, 27-31.

[14] Zweigenbaum, P.; B. Bachimont; J. Bouaud; J. Charlet; and J. F. Boisvieuz. 1995. A multi-lingual architecture for building a normalized conceptual representation from medical language. In Reed M. Gardner (ed.) *Proceedings of the Nineteenth Annual Symposium on Computer Applications in Medical Care*, 357-61.

[15] Hahn, Udo; Martin Romacker; and Stefan Schulz. 2000. MEDSYNDIKATE--design considerations for an ontology-based medical text understanding system. J. Marc Overhage (ed.) *Proceedings of the AMIA Annual Symposium*, 330-4.

[16] Lindberg, Donald A. B.; Betsy L. Humphreys; and Alexa T. McCray. 1993. The Unified Medical Language System. *Methods of Information in Medicine* 32:281-91.

[17] Humphreys, Betsy L.: Donald A. B. Lindberg; Harold M. Schoolman; and G. Octo Barnett. 1998. The Unified Medical language System: An informatics research collaboration. *Journal of American Medical Informatics Association* 5(1):1-13.

[18] McCray, Alexa T.; Alan R. Aronson; Allen C. Browne; Thomas C. Rindflesch; Amir Razi; and Suresh Srinivasan. 1993. UMLS knowledge for biomedical language processing. *Bulletin of the Medical Library Association* 1:184-94.

[19] McCray, Alexa T. 2000. Improving access to healthcare information: the Lister Hill National Center for Biomedical Communications. *MD Computing* 17(2):29-34.

[20] McCray Alexa T.; Suresh Srinivasan; and Allen C. Browne. 1994 Lexical methods for managing variation in biomedical terminologies. In Ozbolt JG (ed.) *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, 235-9.

[21] McCray Alexa T.; Amir M. Razi, Anantha K. Bangalore; Allen C. Browne; and P. Zoe Stavri. 1996. The UMLS Knowledge Source Server: A Versatile Internet-Based Research Tool. In Cimino JJ (ed.) *Proceedings of the AMIA Annual Fall Symposium*, 164-8.

[22] Aronson, Alan R.; Thomas C. Rindflesch; and Allen C. Browne. 1994. Exploiting a large thesaurus for information retrieval. In *Proceedings of RIAO*, 197-216.

[23] Rindflesch, Thomas C., and Alan R. Aronson. 1994. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. Judy G. Ozbolt (ed.) *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care,* 240-4.

[24] Aronson, Alan R. 1996. The effect of textual variation on concept-based information retrieval. In James J. Cimino (ed.) *Proceedings of the AMIA Annual Fall Symposium,* 373-7.

[25] Aronson, Alan R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In Suzanne Bakken (ed.) *Proceedings of the AMIA Annual Symposium,* 17-21.

[26] Rindflesch, Thomas C., and Alan R. Aronson. 1993. Semantic processing in information retrieval. Charles Safran (ed.) *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, 611-15.

[27] Rindflesch, Thomas C. 1995. Integrating natural language processing and biomedical domain knowledge for increased information retrieval effectiveness. *Proceedings of the 5th Annual Dual-use Technologies and Applications Conference,* 260-5.

[28] Rindflesch, Thomas C.; Jayant Rajan; and Lawrence Hunter. 2000. Extracting molecular binding relationships from biomedical text. *Proceedings of the 6th Applied Natural Language Processing Conference*, 188-95. Association for Computational Linguistics.

[29] Cutting D.; J. Kupiec; J. Pedersen; and P. Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*.

[30] Srinivasan, Padmini. 1996a. Query expansion and MEDLINE. *Information Processing and Management* 32(4):431-43.

[31] Srinivasan, Padmini. 1996b. Optimal document-indexing vocabulary for MEDLINE. *Information Processing and Management* 32(5):503-14.

[32] Srinivasan, Padmini. 1996c. Retrieval feedback for query design in MEDLINE. A Comparison with Expert Network and LLSF Approaches. In James J. Cimino (ed.) *Proceedings of the AMIA Annual Fall Symposium,* 353-6.

[33] Srinivasan, Padmini. 1996d. Retrieval feedback in MEDLINE. *Journal of the American Medical Informatics Association* 3(2):157-67.

[34] Aronson, Alan R., and Thomas C. Rindflesch. 1997. Query expansion using the UMLS Metathesaurus. In Daniel R. Masys (ed.) *Proceedings of the AMIA Annual Fall Symposium,* 485-9.

[35] Wright, Lawrence W.; Holly K. Grossetta Nardini; Alan R. Aronson; and Thomas C. Rindflesch. 1998. Hierarchical concept indexing of full-text documents in the UMLS Information Sources Map. *Journal of the American Society for Information Science* 50(6):514-23.

[36] Aronson, Alan R.; Olivier Bodenreider; H. Florence Chang; Susanne M. Humphrey; James G. Mork; Stuart J. Nelson; Thomas C. Rindflesch; and W. John Wilbur. 2000. The NLM indexing initiative. In J. Marc Overhage (ed.) *Proceedings of the AMIA Annual Symposium,* 17-21.

[37] Weeber, Marc; Henny Klein; Alan R. Aronson; James G. Mork; et al. 2000. Text-based discovery in medicine: The architecture of the DAD-system. In J. Marc Overhage (ed.) *Proceedings of the AMIA Annual Symposium,* 903-7.

[38] Swanson, Donald R. 1986. Fish oil, Raynaud s syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30(1), 7 18.

[39] Swanson, Donald R., and Smalheiser, Neil R. 1997. An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence* 91, 183-203.

[40] Srinivasan, Padmini. 2001. A text mining tool for MEDLINE. In Suzanne Bakken (ed.) *Proceedings of the AMIA Annual Symposium,* 642-6.

[41] Srinivasan, Padmini, and Thomas C. Rindflesch. 2002. Exploring text mining from MEDLINE. *Proceedings of the AMIA Annual Symposium,* in press.

[42] Sneiderman Charles A.; Thomas C. Rindflesch; and Alan R. Aronson. 1996. Finding the findings: Identification of findings in medical literature using restricted natural language processing. In Cimino JJ (ed.) *Proceedings of the AMIA Annual Fall Symposium*, 239-43.

[43] Rindflesch, Thomas C.; Carol A. Bean; and Charles A. Sneiderman. 2000. Argument identification for arterial branching predications asserted in cardiac catheterization reports. *Proceedings of the AMIA Symposium*, 704-8.

[44] Rindflesch, Thomas C.; Lawrence Hunter; and Alan R. Aronson. 1999. Mining molecular binding terms from biomedical text. *Proceedings of the AMIA Annual Symposium*, 127-31.

[45] Sarkar, I. Neil, and Thomas C. Rindflesch. 2002. Discovering protein similarity using natural language processing. *Proceedings of the AMIA Annual Symposium,* in press.

[46] Rindflesch, Thomas C.; Lorraine Tanabe; John W. Weinstein; and Lawrence Hunter. 2000. EDGAR: Extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing* 5:514-25.

[47] Tanabe, Lorraine, and W. John Wilbur. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics*, in press.

[48] Libbus, Bisharah, and Thomas C. Rindflesch. 2002. NLP-based information extraction for managing the molecular biology literature. *Proceedings of the AMIA Annual Symposium,* in press.