Chapter 14

SEMANTIC INTERPRETATION FOR THE BIOMEDICAL RESEARCH LITERATURE

Thomas C. Rindflesch[1], Marcelo Fiszman[2], and Bisharah Libbus[3]
National Library of Medicine, 8600 Rockville Pike, Bethesda, Maryland 20894

Abstract: Natural language processing is increasingly used to support biomedical applications that manipulate information rather than documents. Examples include automatic summarization, question answering, and literature-based scientific discovery. Semantic processing is a method of automatic language analysis that identifies concepts and relationships to represent document content. The identification of this information depends on structured knowledge, and in the biomedical domain, one such resource is the Unified Medical Language System. After providing some linguistic background, we discuss several semantic interpretation systems being developed in biomedicine. Finally, we briefly investigate two applications that exploit semantic information in MEDLINE citations; one focuses on automatic summarization and the other is directed at information extraction for molecular biology research.

Key words: natural language processing; semantic interpretation; information extraction; automatic summarization; UMLS

1. INTRODUCTION

Automatic access to online information is an integral part of daily life as well as academic research. In this chapter, we explore the use of natural language processing (NLP), that is, automatic analysis of online text, as a way of supporting and enhancing professional access to the biomedical research literature. We discuss a particular approach that identifies concepts and relations through the (partial) semantic interpretation of text. For example, this processing identifies the semantic proposition (2) from (1).

(1) A randomized trial of etanercept as monotherapy for psoriasis

(2) ETANERCEPT TREATS Psoriasis

Although such an interpretation does not capture the complete meaning of (1) (*randomized trial* and *monotherapy* are not addressed), it provides the basis for systems that depend on the manipulation of information rather than documents.

Information retrieval is a mature application that provides documents relevant to a user-specified topic. The information sought is presumed to be in the documents retrieved but is not made overt. Emerging applications focus on explicit manipulation of information as the basis for decision support systems (Cimino and Barnett, 1993; Mendonça and Cimino, 2000) or for connecting patient records to bibliographic resources (Cimino, 1996), for example. Others use extracted information for literature-based scientific discovery (Srinivasan and Libbus, 2004; Fuller et al., in press).

These applications often depend on MeSH indexing terms assigned (by humans) to MEDLINE citations. However, there are important reasons for supplementing MeSH resources. Reliable indexing is not always available outside MEDLINE, and the information needed by an application may not be supplied by MeSH terms. Increasingly, NLP is used to support information manipulation applications, including, in addition to those mentioned, automatic summarization (Fiszman et al., 2004), question answering (Jacquemart and Zweigenbaum, 2003), and enhanced information retrieval (Grishman et al., 2002).

2. NATURAL LANGUAGE PROCESSING

2.1 Overview

NLP methodologies in the biomedical domain can be considered from the point of view of the text they address and the NLP technology used. Two important content subdomains are clinical medicine and molecular biology. In the clinical domain, the emphasis is on disease, anatomy, etiology, and intervention, along with the interaction among these phenomena. A second important content area is molecular biology. A major challenge is recognizing entities such as genes (and other aspects of the genome) and proteins. Important relationships refer to the way these interact among themselves, as well as with genetic diseases. Below, we briefly discuss one approach to NLP in molecular biology. More extensive coverage is provided in another chapter, in Unit III (Palakal et al., in this volume).

Another way to investigate NLP systems is to consider the genre of the text being processed. Two relevant genres in biomedicine are clinical records (such as discharge summaries and imaging reports) and the research literature. Important differences in both syntactic structure and terminology distinguish the two, and in this chapter we concentrate on the literature, particularly MEDLINE citations. Semantic processing in clinical text is discussed in another chapter in this unit (Friedman, in this volume).

Various linguistic approaches have been used to process biomedical text. These can be broadly categorized as either statistical or symbolic rule-based systems. In medicine, the latter predominate; however, Taira and Soderland (1999) and Pakhomov et al. (2002) have pursued statistical approaches, which assign an analysis to input text by matching it to training text annotated (usually by hand) with target structures. Rule-based NLP systems in medicine fall into one of three categories, based on the linguistic formalism used: phrase structure grammar (Christensen et. al., 2002), which concentrates on syntactic constituents; dependency grammar (Hahn, 2002), which emphasizes relations between words; and semantic grammar (Friedman et al., 1994), which relies on distributional patterns of semantic concepts.

Due to the complexity of language, systems often focus on one aspect of linguistic structure: words, phrases, semantic concepts, or semantic relations. Words can be identified with little (or no) linguistic processing. Phrases are normally identified on the basis of at least some syntactic analysis, using part-of-speech categories and rules for defining phrase patterns in English (Leroy et al., 2003). The identification of concepts and relations constitutes semantic processing and requires that text be mapped to a knowledge structure. In the biomedical domain, the Unified Medical Language System (UMLS) provides one such resource.

2.2 Levels of Linguistic Structure

Textual information management systems based solely on words have enjoyed considerable popularity, largely because the underlying processing is relatively easy to implement. After grammatical function words such as determiners *the* and *this* and prepositions *of* and *with* are eliminated, the remaining words are taken as a surrogate representation of semantic content. In (3), for example, *arthritis*, *children*, *hexacetonide*, and *triamcinolone* represent part of the meaning of the text.

(3) The purpose of this study was to compare the efficacy and safety of intra-articular triamcinolone hexacetonide and triamcinolone acetonide in children with oligoarticular juvenile idiopathic arthritis.

However, such a representation lacks expressiveness. It does not, for example, explicitly represent the fact that the disorder discussed is juvenile idiopathic arthritis or that there are two drugs, triamcinolone hexacetonide and triamcinolone acetonide mentioned.

Phrasal processing addresses some of these deficiencies. For example, the identification of *intra articular triamcinolone hexacetonide* and *oligoarticular juvenile idiopathic arthritis* isolates the relevant strings. However, these phrases alone do not indicate that the first is a drug and the second a disease. Nor do they provide the information that childhood arthritis is another name for this disorder.

Semantic processing enhances phrasal analysis with this kind of information. For example, the phrases in the previous paragraph can be mapped to concepts in the UMLS Metathesaurus (discussed in more detail below): the first to “triamcinolone hexacetonide” and the second to “Chronic Childhood Arthritis.” From information in the Metathesaurus it is possible to determine that the first is a drug and the second a disease.

Identification of concepts provides an enriched representation of the meaning of text; however, an additional level of processing combines concepts into relationships that explicitly represent their interaction. These relationships are often called predications or propositions and are made up of arguments (concepts) and a predicate (relation). Processing to construct semantic predications (called semantic interpretation) determines in (3), for example, that “triamcinolone hexacetonide” treats (rather than causes) “Chronic Childhood Arthritis.” Since the UMLS knowledge sources serve as an enabling resource for semantic interpretation in the biomedical domain, we discuss their main characteristics.

3. DOMAIN KNOWLEDGE: THE UMLS

The UMLS (Humphreys et al., 1998) consists of three components that provide structured knowledge in the biomedical domain: the SPECIALIST Lexicon (McCray et al., 1994), the Semantic Network (McCray, 2003), and the Metathesaurus. The Lexicon supports syntactic analysis, while the Metathesaurus allows concepts to be identified in text; finally, the Semantic Network underpins the identification of semantic relationships.

3.1 SPECIALIST Lexicon

The SPECIALIST Lexicon describes syntactic characteristics of biomedical and general English terms, and this comprehensive resource provides the basis for NLP in the biomedical domain. In addition to part-of-speech labels for each entry, spelling variation when it occurs (particularly British forms) and inflection for nouns, verbs, and adjectives are included. Inflection is encoded by referring to rules for regular variants (*-s* for nouns and *-s*, *-ed*, *-ing* for verbs, for example) as well as Greco-Latin plurals. Irregular forms are listed where they apply. The variant annotation for *sarcoma* (4), for example, indicates that this form may either appear invariant (*sarcoma*), with a regular plural (*sarcomas*), or with Greco-Latin morphology (*sarcomata*).

```
(4)  sarcoma
        cat=noun
        variants=uncount
        variants=reg
        variants=glreg
```

For verbs, complement patterns and nominalizations are included. The verb *manage* (5) takes regular verbal inflection and has nominalization *management*. It may occur with no object (intran), with a noun phrase object (tran=np), or with an infinitival complement, in which case the subject of *manage* is also the subject of the infinitive (tran=infcomp:subjc), as in *she managed to win the race*.

```
(5)  manage
        cat=verb
        variants=reg
        intran
        tran=np
        tran=infcomp:subjc
        nominalization=management
```

3.2 Metathesaurus

The Metathesaurus is a compilation of more than 100 terminologies and controlled vocabularies in the biomedical domain, and includes those with comprehensive coverage, such as Medical Subject Headings (MeSH) and Systematized Nomenclature of Medicine (SNOMED), as well as those focused on subdomains such as dentistry (Current Dental Terminology) or nursing (Nursing Interventions Classification). Others provide specialized terms for components of the medical domain, such as anatomy (University of Washington Digital Anatomist) or medical devices (Universal Medical Device Nomenclature System).

Terms from the constituent vocabularies are organized into more than a million concepts (in the 2004 release) that reflect synonymous meaning. For example, the concept "Chronic Childhood Arthritis" contains synonymous terms "Arthritis, Juvenile Rheumatoid" (from MeSH and SNOMED) and "Rheumatoid arthritis in children" (Library of Congress Subject Headings), among others.

Hierarchical information inherent in component vocabularies is maintained in the Metathesaurus. For example, part of the structure for the concept "Juvenile Rheumatoid Arthritis" is given in (6).

```
(6)  Immunologic Diseases
        Autoimmune Diseases
           Arthritis, Rheumatoid
              Arthritis, Juvenile Rheumatoid
```

Each concept in the Metathesaurus is assigned at least one semantic type, selected from 135 general categories relevant to the biomedical domain. Examples include 'Pharmacological Substance', 'Disease or Syndrome', 'Therapeutic or Preventive Procedure', and 'Amino Acid, Peptide, or Protein'.

Identical concepts with different meanings reflect word sense ambiguity in English, and such terms are distinguished in the Metathesaurus. For example, "Strains <1>" (with semantic type 'Injury or Poisoning') has synonyms "Muscle strain" and "Pulled muscle" and is distinguished from "Strains <2>" (semantic type 'Intellectual Product') with synonym "Microbiology subtype strains."

3.3 Semantic Network

The UMLS Semantic Network constitutes an upper-level ontology of medicine. Its components are the 135 semantic types assigned to Metathesaurus concepts as well as 54 relationships. The semantic types are organized into two hierarchies whose roots are 'Entity' and 'Event'. The two immediate children of 'Entity' are 'Physical Object' and 'Conceptual Entity', while 'Activity' and 'Phenomenon or Process' are immediately dominated by 'Event'. The hierarchical structure of the semantic type 'Pharmacologic Substance' is given in (7).

```
(7)  Entity
        Physical Object
           Substance
              Chemical
```

Chemical Viewed Functionally
Pharmacologic Substance

Semantic types are also organized into higher level groups (McCray et al., 2001), which reflect semantic coherence among members. For example, the semantic group Disorders includes such semantic types as 'Acquired Abnormality', 'Disease or Syndrome', and 'Injury or Poisoning', while the group Procedures includes 'Diagnostic Procedure' and 'Therapeutic or Preventive Procedure'.

The 54 relationships in the Semantic Network are organized hierarchically under nodes that include PHYSICALLY_RELATED_TO (e.g. PART_OF and CONNECTED_TO), FUNCTIONALLY_RELATED_TO (e.g. DISRUPTS and TREATS), and CONCEPTUALLY_RELATED_TO (e.g. PROPERTY_OF and MEASURES). These relationships serve as the predicates of semantic predications whose arguments are semantic types. Some examples are given in (8).

(8) 'Therapeutic or Preventive Procedure' TREATS 'Injury or Poisoning'
'Organism Attribute' PROPERTY_OF 'Mammal'
'Body Space or Junction' CONNECTED_TO 'Tissue'
'Bacterium' CAUSES 'Pathologic Function'

The predications in the Semantic Network define a model of the medical domain and provide an important constraint on semantic interpretation.

4. SEMANTIC INTERPRETATION FOR THE BIOMEDICAL LITERATURE

4.1 Overview

Semantic interpretation relies on the identification of concepts in an outside knowledge structure and then determines relationships asserted between these concepts in text. We consider three approaches to semantic processing in the biomedical domain: AQUA (Johnson et al., 1993), PROTEUS-BIO (Grishman et al., 2002), and SemRep (Rindflesch and Fiszman, 2003). All three depend on biomedical knowledge sources and produce semantic predications as output. They differ primarily regarding the goals for which they were devised. They are based on varying linguistic formalisms and the particular knowledge sources used. Each system has specific strengths (and limitations). Given the challenges posed by natural language it is not possible for any system to produce a complete semantic analysis.

4.2 AQUA

AQUA (A QUery Analyzer) is an underspecified semantic interpreter that was originally devised for processing MEDLINE queries. The general approach is to identify salient medical concepts along with the syntactic phenomena that cue relations between them, without constructing a complete analysis. There are general principles for ignoring syntactic aspects of the input that are not directly concerned with key relations, such as *I am interested in articles about…*

The linguistic approach is based on operator grammar (Johnson and Gottfried, 1989), which provides rules for the ordering of operators and arguments in sentences. For example, the operator *with* occurs between its arguments in *patients with liver abscess*, while the operator *treatment* precedes its arguments in *the treatment of tuberculosis with rifampin*. Operator grammar supports a principled means of formulating generalizations that relate syntactic operator-argument patterns to underlying semantic predications.

The parsing formalism in AQUA is implemented as a definite clause grammar, which affords a flexible way of recognizing the argument-operator patterns defined by the operator grammar. This formalism allows both syntactic and semantic constraints to be included in parsing rules and also accommodates skipping part of the input. The parser depends on a lexicon that was derived from the UMLS (final editing was by hand). The AQUA lexicon contains semantic information (including semantic types) as well as part-of speech labels, and explicitly indicates whether an entry functions as an argument or an operator.

The combination of operator grammar, definite clause grammar, and semantic lexicon underpins AQUA's ability to map queries to semantic predications, which are represented as conceptual graphs, a more expressive form of the first-order predicate calculus (Sowa, 2000). For example the query (9) is interpreted as the proposition (10), which captures the key relations that infections and liver abscesses occur in patients who also have Hodgkin's disease.

(9) Request search for papers detailing infections, specifically liver abscesses, in patients with Hodgkin's disease

(10) [Pathologic Function: {infections, liver abscesses}] -
(occurs_in) → [Patient or Disabled Group: patients] -
(occurs_in) ← [Disease or Syndrome: Hodgkin's disease]

Semantic predications produced by AQUA have been validated against the UMLS Semantic Network. Recent work using AQUA focuses on semantic relations in clinical text and connecting that text with MEDLINE citations (Mendonça et al., 2002).

4.3 PROTEUS-BIO

PROTEUS-BIO is an information extraction system that depends on underspecified semantic interpretation as its core element. The system applies to Web documents on infectious disease outbreaks; it extracts semantic predications relevant to this domain and stores them in a database, which can be queried by users.

Semantic interpretation in PROTEUS-BIO identifies relationships pertinent to the domain, such as "outbreak of <disease> killed <victims>." Concepts in the entity classes in this domain, namely diseases, victims, and geographic locations, are stored in a hierarchical knowledge structure, which was specifically constructed for this application.

Initial processing concentrates on syntactic patterns to find the entities that can serve as arguments. In addition to noun phrases, verb groups such as *were killed* are identified. Noun phrases are labeled with semantic classes (such as <disease> or <victim>) during this phase and are then available to the next phase.

Processing to identify semantic predications is based on event patterns, which are defined in terms of the argument classes identified in the previous phase. For example, the pattern (11) matches the text (12).

(11) np(<disease>) vg(KILL) np(<victim>)

(12) Cholera killed 23 inhabitants

Additional patterns are defined to accommodate passive structures (based on the verb groups identified in the first phase). A metarule is designed to allow an event pattern to apply to text that includes adverbial constructions either before or after the components of the pattern. The metarule, for example, allows all the examples in (13) to match the event pattern (11), despite the occurrence of the adverbial expression *last week*.

(13) last week 23 inhabitants were killed by cholera
23 inhabitants were killed last week by cholera
23 inhabitants were killed by cholera last week

The accuracy of the semantic predications extracted by PROTEUS-BIO was evaluated on an annotated test collection of 32 documents. Precision was 79% and recall was 41%. As noted earlier, semantic processing in this system is meant to support information retrieval applications. Predications identified by PROTEUS-BIO are stored in a database and are linked to the documents from which they were extracted. It is thus possible to use this database to enhance the results of queries seeking documents in the disease outbreak domain. A task-oriented evaluation to measure effectiveness in achieving this goal was conducted, and initial results indicate that precision was notably increased using the PROTEUS-BIO system.

4.4 SemRep

SemRep is being developed to recover semantic propositions from the biomedical research literature (concentrating on MEDLINE citations) using underspecified syntactic analysis and structured domain knowledge. Processing begins with a lexical analysis based on the SPECIALIST Lexicon and a stochastic tagger. This serves as input to an underspecified parser, which provides the basis for semantic analysis (also underspecified). In analyzing (14), for example, after tokenization, the SPECIALIST Lexicon is consulted.

(14) Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes

Each lexical entry (including multiword forms like *Doppler echocardiology*) is assigned a part-of-speech label, and lexical ambiguities are assigned more than one label. For example, *used* has labels "verb" and "adj" in the lexicon, while *left* has "adj," "adv," "noun," and "verb."

A stochastic tagger (Smith et al., 2004) then resolves part-of-speech ambiguities based on common patterns seen in training data. The tagged text in (15) serves as input to the parser.

(15)

Doppler echocardiography	can	be	used	to	diagnose	left	anterior
noun	modal	aux	verb	adv	verb	noun	adj

descending	artery	stenosis	in	patients	with	type	2	diabetes
adj	noun	noun	prep	noun	prep	noun	num	noun

Note that taggers do not have 100% accuracy. For example, *left* should be tagged as an adjective in this context rather than as a noun.

The underspecified syntactic analysis is based on part-of-speech labels and segments the input into phrases that correspond to the lowest level structures in a full syntactic analysis. Segmentation is based on barrier words, which serve as boundaries between phrases. These include modals (*can* in the current example), auxiliaries (*be*), verbs (*used, diagnose*), and prepositions (*in, with*). The exploitation of these barriers in an algorithm that uses them to close one phrase and open another produces the analysis in (16). Any phrase containing a noun constitutes a (simple) noun phrase. The rightmost noun is relabeled as "head" and items to the left of the head (other than determiners and prepositions) are labeled as "mod."

```
(16) [[head(’Doppler echocardiography’) ],
     [modal(can) ],
     [aux(be) ],
     [verb(used) ],
     [adv(to) ],
     [verb(diagnose) ],
     [mod(left), mod(anterior), mod(descending), mod(artery),
        head(stenosis)],
     [prep(in), head(patients) ],
     [prep(with), head(’type 2 diabetes’) ] ]
```

Simple noun phrases constitute the referential vocabulary. The concepts they refer to in the domain model are computed by using MetaMap (Aronson, 2001) to match elements in each noun phrase to concepts in the UMLS Metathesaurus. MetaMap examines all the words in a phrase and then determines the best match with a term in the Metathesaurus, taking into account inflectional and derivational variation and allowing for partial and multiple mappings.

The phrases *Doppler echocardiography* and *patients*, for example, match exactly to concepts: “Echocardiography, Doppler” (with semantic type ‘Diagnostic Procedure’) and “Patients” (‘Patient or Disabled Group’). The phrase *left anterior descending artery stenosis* maps to two concepts: “Anterior descending branch of left coronary artery” (‘Body Part, Organ, or Organ Component’) and “Acquired stenosis” (‘Finding’ and ‘Pathologic Function’). When MetaMap has found a viable match between text words and a Metathesaurus term, it provides the preferred Metathesaurus name for that term, as in the case of the coronary artery mentioned here. Similarly, although the term *type 2 diabetes* occurs in the Metathesaurus, its preferred name is “Diabetes Mellitus, Non-Insulin-Dependent” (‘Disease or Syndrome’). Metathesaurus concepts for a noun phrase become a part of the representation of that phrase as semantic enhancement.

The interpretation of semantic predications asserted in the input text depends on the syntactic and semantic information contained in the underspecified parse structure enhanced with UMLS concepts and semantic types. Syntactic phenomena (including verbs, prepositions, nominalizations, and the head-modifier relation in noun phrases) “indicate” semantic predicates and are mapped to relations in the Semantic Network. The indicators in (14) are the verb *diagnose*, the prepositions *in* and *with*, and the modifier-head structure in the noun phrase whose head is *stenosis*.

Indicators are syntactic predicates that anchor the interpretation of syntactic structures as semantic predications, and two phenomena are involved in this process: argument identification and mapping to relations in the Semantic Network. Argument identification is controlled by a dependency grammar that establishes a syntactic relation between the indicator and the head of a simple noun phrase serving as its argument. Rules in this grammar are stated in very general terms for each class of indicator. For example, the argument identification rules for verbs stipulate that subjects occur to the left of the verb and objects to the right.

The syntactic constraint imposed by the dependency grammar serves as a necessary condition on the interpretation of a syntactic indicator and its arguments as a semantic predication. In (14), for example, the rules applied to *diagnose* limit the subject of this verb to the noun phrase *Doppler echocardiography*; the object, however, could be any of the three noun phrases to the right of *diagnose*: *left anterior descending artery stenosis*, *patients*, or *type 2 diabetes*. Further semantic conditions apply in determining which of these is the object of *diagnose* in (14).

All indicators are linked by rule to relations in the UMLS Semantic Network. The indicator rules needed to interpret (14) are given in (17); syntactic phenomena (part-of-speech or structure) occur to the left of the arrow and Semantic Network relations occur to the right.

(17) *diagnose* (verb) → DIAGNOSES
modifier-head (structure) → LOCATION_OF
in (preposition) → OCCURS_IN
with (preposition) → CO-OCCURS_WITH

The complete relationships, with semantic types as arguments, are given in (18) for the Semantic Network predicates in (17).

(18) 'Diagnostic Procedure' DIAGNOSES 'Pathologic Function'
'Body Part, Organ, or Organ Component' LOCATION_OF 'Pathologic Function'
'Pathologic Function' OCCURS_IN 'Patient or Disabled Group'
'Pathologic Function' CO-OCCURS_WITH 'Disease or Syndrome'

A metarule ensures that all semantic propositions identified by SemRep are sanctioned by a predication in the Semantic Network, and this restriction limits the identification of arguments. For example, the Semantic Network predication DIAGNOSES has the semantic type 'Pathologic Function' as one of its arguments. Therefore, any syntactic indicator linked to DIAGNOSES must have an argument whose head has been mapped to a Metathesaurus concept with the same semantic type. In (14), the only potential object of the verb *diagnose* that fulfills this requirement is the head of *left anterior descending artery stenosis* (whose semantic type is 'Pathologic Function'). *Doppler echocardiography* was identified syntactically as an argument of *diagnose*, and its semantic type, 'Diagnostic Procedure', matches the other argument of DIAGNOSES in the Semantic Network.

When these syntactic and semantic conditions are satisfied, a semantic predication can be constructed that is the interpretation of the syntactic indicator and its (syntactic) arguments. The predicate in this semantic proposition is the Semantic Network relation to the right of the arrow in the indicator rule; the arguments are the Metathesaurus concepts from the syntactic arguments of the indicator. In the case of the indicator *diagnose*, the predicate is DIAGNOSES and the arguments are the concepts "Echocardiography, Doppler" and "Acquired stenosis." The complete predication is

(19) Echocardiography, Doppler DIAGNOSES Acquired stenosis

When similar rules are applied to the other indicators in (14), namely the prepositions *in* (OCCURS_IN) and *with* (CO-OCCURS_WITH) and the head-modifier construction in the *stenosis* noun phrase (LOCATION_OF), the semantic propositions in (20) are produced.

(20) Acquired stenosis OCCURS_IN Patients
Acquired stenosis CO-OCCURS_WITH Diabetes Mellitus, Non-Insulin-Dependent
Anterior descending branch of left coronary artery LOCATION_OF Acquired stenosis

SemRep has recently been enhanced to address hypernymic propositions (Fiszman et al., 2003), in which a more specific concept is asserted to be in a taxonomic relation with a more general concept. For example, SemRep is able to extract the predication (22) as a representation of the relationship between *posaconazole* and *antifungal agent* in (21).

(21) Posaconazole is a potent broad-spectrum azole antifungal agent in clinical development for the treatment of invasive fungal infections.

(22) posaconazole ISA Antifungal Agents.

The interpretation of hypernymic predications depends on the arguments involved being in a hierarchical relationship in the Metathesaurus.

4.4.1 Evaluation of SemRep

Preliminary evaluation of SemRep has been conducted on a collection of 2,000 sentences from MEDLINE citations, concentrating on drug treatments for disease. Initial focus has been on a core set of semantic predicates, such as TREATS, LOCATION_OF, CO-OCCURS_WITH, and OCCURS_IN. Precision and recall on this test collection are 78% and 49%

respectively. The majority of the false positive errors (contributing to diminished precision) are due to word sense ambiguity. For example, in (23), *concentration* maps to the corresponding Metathesaurus concept with semantic type 'Mental Process'.

(23) . . .the mean fluorescein concentration in the cornea of the lyophilisate group was two times higher than at baseline.

This mapping allows the incorrect predication (24) to be constructed, in which the cornea is interpreted as the location of a mental process.

(24) Cornea <1> LOCATION_OF Concentration

A significant percentage of false negative errors are due to current deficiencies in processing comparative structures. For example, SemRep retrieves the predication (26) while interpreting (25), but fails to identify that co-trimoxazole treats pneumonia, which is also asserted in the sentence.

(25) The purpose of this study was to compare the clinical effectiveness of co-trimoxazole with amoxicillin for treatment of childhood pneumonia

(26) Amoxicillin TREATS Pneumonia

4.5 Comparison of AQUA, PROTEUS-BIO, and SemRep

The three systems discussed are intended to provide useful results, without attempting a full semantic analysis. AQUA uses operator grammar to manipulate traditional syntactic constituent structure. The flexibility of this formalism allows the system to focus on grammatical structures relevant to the interpretation of users' queries to MEDLINE. Domain knowledge used by AQUA is based on the UMLS, and a wide range of semantic topics are accommodated. PROTEUS-BIO is intended to retrieve timely information from Web documents in a specific content area, namely infectious disease outbreaks. It uses partial constituent structure for noun phrases and verb groups, along with robust pattern matching in cooperation with specially constructed knowledge sources to achieve practical results in a limited domain. SemRep also relies on partial constituent structure, and in addition uses an underspecified dependency grammar for argument identification. It exploits the UMLS knowledge sources without modification. Although limited in the semantic relations it addresses, SemRep applies to a wide range of syntactic structures asserting the treatment of disease in the biomedical research literature.

5. APPLICATION OF SEMREP

Above, we briefly mentioned applications for semantic interpretation in the discussion of AQUA and PROTEUS-BIO. We now consider recent applications of SemRep. This program serves as the basis for several ongoing research initiatives in biomedical information management, including efforts directed at automatic summarization of the results of PubMed searches and extracting molecular biology information from text.

5.1 Automatic Summarization

Automatic summarization is an important emerging application in the biomedical domain. With the growing emphasis on evidence-based medicine it is important for physicians to keep abreast of the research literature. This is challenging due to the large size of the MEDLINE database. For example, a PubMed query on the treatment of diabetes, limited to articles published in 2003 and having an abstract in English, finds 3,621 items; further limitation to articles describing clinical trials still returns 390 citations.

One goal of automatic summarization in biomedicine is to provide practitioners with current, focused information on the treatment of specific diseases, including summaries with pointers to the most relevant citations. SemRep is being used as the basis for an automatic summarization application in the abstraction paradigm (Fiszman et al., 2004), in which the semantic interpretation of text is manipulated, rather than the text itself (extraction summarization).

The system we are developing takes as input a list of semantic predications produced by SemRep from a set of documents on a specified disorder topic. The output is a conceptual condensate (in graphical format) containing just those predications that represent key information in the input documents. There are links to the original text that generated the propositions.

The core of the method is a transformation process that condenses and generalizes the input predications, guided by four principles (27) that use semantic information from the UMLS and frequency of occurrence of concepts and relations in the input predications.

(27) Relevance: Include predications on the topic of the summary
Novelty: Do not include predications that the user already knows
Connectivity: Also include “useful” additional predications
Saliency: Only include the most frequently occurring predications

Relevance processing condenses the list of predications by ensuring that they conform to a schema describing disorders (Jacquelinet et al.,

2003) that contains general statements such as "{Treatment} treats {Disorders}." "Domains" such as {Treatment} and {Disorder} define sets of UMLS semantic types derived from the semantic groups. Predications conforming to the schema are called "core predications." Novelty provides further condensation by eliminating predications having generic arguments, as determined by hierarchical depth in the Metathesaurus. For example, predications containing arguments such as "Patients" and "Pharmaceutical Preparations" are eliminated by the Novelty principle.

Connectivity is a generalization process that identifies predications occurring in neighboring semantic space of the core, namely non-core predications that share an argument with a core predication. For example, from "Naproxen TREATS Osteoarthritis," non-core predications such as "Naproxen ISA NSAID" are included in the condensate. Finally, the Saliency principle calculates frequency of occurrence of arguments, predicates, and predications; those occurring less frequently than the average are eliminated from the final condensate (Hahn and Reimer, 1999).

Figure 14-1 is a conceptual condensate summarizing the 300 most recent citations retrieved by a PubMed search using the query "Diabetes Mellitus, Type II" (a MeSH term). SemRep generated 3,092 semantic predications from the input documents, and the transformation process reduced these to 73 predications (only the unique types are given in Figure 14-1).

The summary of type 2 diabetes given in Figure 14-1 provides an overview of the latest research on interventions for this disorder. Insulin is becoming increasingly important in this regard and is included in the summary. Traditionally, oral pharmacotherapy has been the treatment of choice, as shown by the appearance of metformin in the condensate. New drugs such as pioglitazone (thiazolinediones) and acarbose (both are included in Figure 14-1) are showing promise in either treating or preventing type 2 diabetes.

The conceptual condensate can be viewed from the perspective of citations rather than predications and doing so may have implications for improving information retrieval effectiveness. Of the 300 citations summarized, 52 contributed at least one predication to the final condensate. The three citations that contributed at least four predications are all highly relevant to the treatment of type 2 diabetes. For example, one of these has the title "Effect of antidiabetic medications on microalbuminuria in patients with type 2 diabetes." Of the citations that contributed a single predication to the conceptual condensate, only one directly discusses the treatment of type 2 diabetes; others are about related issues, for example: “Persistent remodeling of resistant arteries in type 2 diabetic patients on anti-hypertensive treatment.”

5.2 Information Extraction in Molecular Genetics

A second application of SemRep currently being pursued investigates the use of NLP for studying the etiology of genetic diseases. The focus of this work is to identify semantic predications in the research literature that assert a relationship either between a gene and a disease or between two genes implicated in a disease. The underlying technology is a program called SemGen (Rindflesch et al., 2003), which is a modification of SemRep. SemGen has the same core structure as SemRep, and processing other than mapping noun phrases to semantic concepts is identical in the two programs.

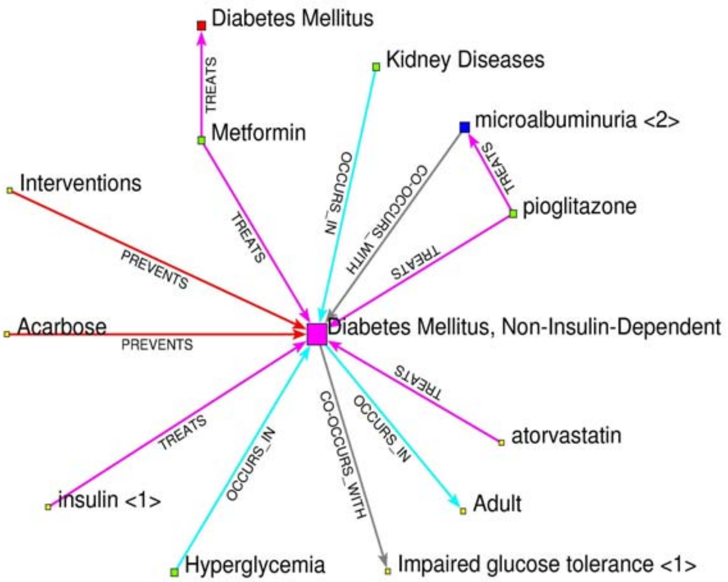

Figure 14-1. Conceptual condensate summarizing 300 citations on type 2 diabetes

While enhancing SemRep to construct SemGen, a program called ABGene (Tanabe and Wilbur, 2002) was added in order to augment MetaMap processing for genetic terminology. ABGene is based on part-of-speech tagging technology and uses several statistical and empirical methods to identify gene names that do not occur in the UMLS Metathesaurus.

The domain knowledge underpinning semantic interpretation specific to the etiology of genetic diseases that SemGen relies on was constructed by hand. This knowledge substitutes for the UMLS Semantic Network in SemRep. The allowable arguments of the semantic predications addressed by SemGen are characterized by two semantic classes: disorders and genetic phenomena. Disorders are defined as concepts having the UMLS semantic types in the Disorder semantic group. For genetic phenomena, concepts with semantic types from the semantic group Gene (including semantic type 'Gene or Genome', for example) are augmented with output from ABGene.

The relevant predicates for gene-disease relationships are ASSOCIATED_WITH, PREDISPOSE, and CAUSE. The subject of these predicates is a genetic phenomenon and the object is a disorder. Predicates defined for gene interactions are INTERACT_WITH, STIMULATE, and INHIBIT. Both arguments of these predicates are genetic phenomena. For example, SemGen extracts the gene-disease interaction predication (29) from (28) and the gene-gene predication (31) from (30).

(28) An elevated frequency of the CYP2D6*4 allele has been found in Parkinson's disease.

(29) cyp2d6*4 allele ASSOCIATED_WITH Parkinson Disease

(30) PDX-1 interacts with multiple transcription factors and coregulators, including the coactivator p300, to activate the transcription of the insulin gene and other target genes within pancreatic beta cells.

(31) pdx-1 STIMULATE insulin

We are pursuing research on several fronts that exploits SemGen output in bioinformatics applications. One project compares a curated database to the current literature. OMIM (Online Mendelian Inheritance in Man) is an information resource on genetic diseases that has nearly 15,000 hand-curated entries describing clinical phenotypes and associated genes. We have used SemGen output as the basis for comparing OMIM entries on a particular disorder to MEDLINE citations (Libbus et al., 2004). The goal was to explore the possibility of automatically suggesting recent research to supplement OMIM information.

For example, we ran SemGen on OMIM text for Alzheimer's disease and also on the output of a PubMed query on that disorder, limited to citations that postdate the most recent OMIM entry. We then automatically compared the SemGen predications from OMIM to those from MEDLINE. We were most interested in discovering predications that occurred in MEDLINE, but not in OMIM, and the following are examples of such predications.

(32) TGFB1 ASSOCIATED_WITH Amyloid deposition
MAPT INTERACT_WITH HSPA8
CD14 STIMULATE amyloid peptide

On the basis of this kind of output, SemGen can potentially serve as an important tool for researchers in scanning a large number of citations and providing information that could promote hypothesis generation and scientific discovery.

Finally, visualization techniques can be used to construct gene-gene interaction networks automatically from predications extracted from text by SemGen. Such networks provide an easily accessible overview of the molecular mechanisms implicated in genetic disease. As an example, Figure 14-2 is a partial network for some of the predications describing the genes that interact with the leptin gene (LEP). The relationships illustrated were extracted from documents discussing diabetes and genes and may provide insight into the genetic underpinnings of that disorder.

Figure 14-2, for example, indicates that LEP inhibits insulin (INS), while INS stimulates LEP. This feedback relationship is involved in appetite suppression and is perturbed during diabetes or obesity. Further, LEP, which is elevated in obesity, stimulates the gene AKT, which ultimately leads to the formation of new vessels underlying diabetic retinopathy.

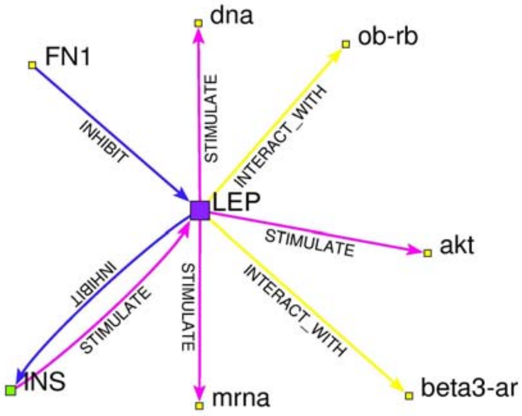

Figure 14-2. Some LEP gene interactions extracted by SemGen from text on diabetes

6. CONCLUSION

The development of NLP systems for semantic interpretation in the biomedical research literature is motivated by the need to support emerging applications that focus on the manipulation of information rather than documents. Implemented systems address a range of information management tasks, including automatic summarization, connecting patient records with the research literature, question answering, literature-based scientific discovery, and the extraction of information to support molecular biology research, as well as enriched query processing and document manipulation.

A variety of linguistic formalisms are used for semantic processing. Due to the complexity of natural language, practical systems focus on biomedical subdomains as well as specific syntactic structures and semantic relations. The identification of semantic concepts and predications in the research literature relies on structured domain knowledge, such as the UMLS. This large resource includes lexical information to support NLP, and the content it contains is organized hierarchically and as an upper-level ontology of biomedicine.

Two examples of the application of semantic interpretation to the biomedical research literature include automatic summarization for the treatment of disease and extraction of molecular biology information on the etiology of genetic disorders. Visualization techniques can profitably be used to give users an overview of extracted information. Continued development of semantic processing systems in biomedicine promises to provide professionals with more powerful tools for effectively exploiting online textual resources.

REFERENCES

Aronson, A. R. (2001). “Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program,” in *Proceedings of the AMIA Symposium*, 17-21.

Cimino, J. J. (1996). “Linking Patient Information Systems to Bibliographic Resources,” *Methods of Information in Medicine,* **35**, 122-6.

Cimino, J. J. and Barnett, G. O. (1993). “Automatic Knowledge Acquisition from MEDLINE,” *Methods of Information in Medicine,* **32**, 120-30.

Christensen, L., Haug, P. J., and Fiszman, M. (2002). “MPLUS: A Probabilistic Medical Language Understanding System,” *ACL Workshop on Natural Language Processing in the Biomedical Domain,* 29-36.

Fiszman, M., Rindflesch, T. C., and Kilicoglu, H. (2003). “Integrating a Hypernymic Proposition Interpreter into a Semantic Processor for Biomedical Text,” in *Proceedings of the AMIA Symposium*, 239-43.

Fiszman, M., Rindflesch, T. C., and Kilicoglu, H. (2004). "Abstraction Summarization for Managing the Biomedical Research Literature," in *Proceedings of the Workshop on Computational Lexical Semantics*, 76-83. HLT-NAACL.

Fuller, S., Revere, D., Bugni P., and Martin, G.M. "Telemakus: A Schema-based Information System to Promote Scientific Discovery," *Journal of the American Society for Information Science and Technology,* in press.

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., and Johnson, S. B. (1994). "A General Natural-language Text Processor for Clinical Radiology," *Journal of the American Medical Informatics Association*, **1**(2), 161-74.

Grishman, R., Huttunen, S., and Yangarger, R. (2002). "Information Extraction for Enhanced Access to Disease Outbreak Reports," *Journal of Biomedical Informatics,* **35**(4), 236-46.

Hahn, U. and Reimer, U. (1999). "Knowledge-based Text Summarization: Salience and Generalization Operators for Knowledge Base Abstraction," in I. Mani (Ed.), *Advances in Automatic Summarization*, Cambridge, MA: MIT Press. 215-32.

Hahn, U., Romacker, M., and Schulz, S. (2002). "MEDSYNDIKATE—Design Considerations for an Ontology-based Medical Text Understanding System," in *Proceedings of the AMIA Symposium*, 330-4.

Humphreys, B. L., Lindberg, D.A., Schoolman, H.M., and Barnett, G.O. (1998). "The Unified Medical Language System: An Informatics Research Collaboration," *Journal of the American Medical Informatics Association,* **5**(1), 1-11.

Jacquelinet, C., Burgun, A., Delamarre, D., Strang, N., Djabbour, S., Boutin, B., and Le Beux, P. (2003). "Developing the Ontological Foundations of a Terminological System for End-stage Diseases, Organ Failure, Dialysis and Transplantation," *International Journal of Medical Informatics,* **70(**2-3), 317-28.

Jacquemart, P. and Zweigenbaum P. (2003). "Towards a Medical Question-answering System: A Feasibility Study," *Stud Health Technol Inform,* **95,** 463-8.

Johnson, S. B., Aguirre, A., Peng, P., and Cimino, J. J. (1993). "Interpreting Natural Language Queries Using the UMLS," in *Proceedings of the AMIA Symposium,* 294-8.

Johnson, S. B. and Gottfried, M. (1989). "Sublanguage Analysis as a Basis for Controlled Medical Vocabulary," *SCAMC*, 519-23.

Leroy, G., Chen, H., and Martinez, J.D. (2003). "A Shallow Parser Based on Closed-class Words to Capture Relations in Biomedical Text," *Journal of Biomedical Informatics,* **36**(3), 145:58.

Libbus, B., Kilicoglu, H., Rindflesch, T. C., Mork, J. G., and Aronson, A. R. (2004). "Using Natural Language Processing, Locus Link, and the Gene Ontology to Compare OMIM to MEDLINE," in *Proceedings of the Workshop on Linking the Biological Literature, Ontologies and Databases: Tools for Users*, 69-76. HLT-NAACL.

McCray, A. T. (2003). "An Upper-level Ontology for the Biomedical Domain," *Comp Funct Genom,* **4,** 80-4.

McCray, A. T., Burgun, A., and Bodenreider, O. (2001). "Aggregating UMLS Semantic Types for Reducing Conceptual Complexity," in *Medinfo,* 10(Pt 1), 216-20.

McCray, A. T., Srinivasan, S., and Browne, A. C. (1994). "Lexical Methods for Managing Variation in Biomedical Terminologies," *SCAMC*, 235-9.

Mendonça, E. A. and Cimino, J. J. (2000). "Automated Knowledge Extraction from MEDLINE Citations," in *Proceedings of the AMIA Symposium*, 575-9.

Mendonça, E. A., Johnson, S. B., Seol, Y., and Cimino, J. J. (2002). "Analyzing the Semantics of Patient Data to Rank Records of Literature Retrieval," *ACL Workshop on Natural Language Processing in the Biomedical Domain*, 69-76.

Pakhomov, S. V., Ruggieri, A., and Chute, C. G. (2002). "Maximum Entropy Modeling for Mining Patient Medication Status from Free Text," in *Proceedings of the AMIA Symposium*, 587-91.

Rindflesch, T. C., and Fiszman, M. (2003). "The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text," *Journal of Biomedical Informatics,* **36**(6),462-77.

Rindflesch, T. C., Libbus, B., Hristovski, D., Aronson, A. R., and Kilicoglu, H. (2003). "Semantic Relations Asserting the Etiology of Genetic Diseases," in *Proceedings of the AMIA Symposium,* 554-8.

Smith, L., Rindflesch, T., and Wilbur, W. J. (2004). "MedPost: A Part of Speech Tagger for Biomedical Text," *Bioinformatics,* in press.

Sowa, J. F. (2000). *Knowledge Representation*, Pacific Grove: Brooks/Cole.

Srinivasan P. and Libbus, B. (2004). "Mining MEDLINE for Implicit Links Between Dietary Substances and Diseases," *Bioinformatics,* **20,** i290-i296.

Taira, R. K., and Soderland, S. G. (1999). "A Statistical Natural Language Processor for Medical Reports," in *Proceedings of the AMIA Symposium*, 970-4.

Tanabe, L. and Wilbur, W. J. (2002). "Tagging Gene and Protein Names in Biomedical Text," *Bioinformatics,* **18**(8), 1124-32.

SUGGESTED READINGS

Friedman, C. and Hripcsak, G. (1999). "Natural Language Processing and its Future in Medicine," *Acad Med,* **74**(8), 890-5.

Reviews several current NLP methodologies in biomedicine with discussion of potential applications.

Jurafsky, D. and Martin J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed., Upper Saddle River: Prentice Hall.

Provides an overview of techniques for natural language processing, both rule-based and statistical approaches.

Rindflesch, T. C. and Aronson, A.R. (2002). "Semantic Processing for Enhanced Access to Biomedical Knowledge," in *Real World Semantic Web Applications*, V. Kashyap and L. Shklar (Eds.), IOS Press, 157-72.

Gives an overview of SemRep and MetaMap with examples of their application.

ONLINE RESOURCES

UMLS documentation:
http://www.nlm.nih.gov/research/umls/documentation.html

Semantic Knowledge Representation Project at NLM:
http://skr.nlm.nih.gov

SPECIALIST Lexicon and lexical tools:
http://specialist.nlm.nih.gov

MetaMap Transfer (MMTx):

http://mmtx.nlm.nih.gov

QUESTIONS FOR DISCUSSION

1. Why is it important to pursue research on semantic processing of the biomedical literature? Discuss biomedical applications (other than those noted in this chapter) that could benefit from semantic representation.
2. What are the levels of knowledge required for semantic processing? List the steps required for semantic interpretation of "low dose aspirin for the prevention of myocardial infarction," if SemRep is used as the semantic processor.
3. Discuss strengths and limitation as well as similarities and differences of the systems designed to provide semantic interpretation of the biomedical literature (AQUA, PROTEUS-BIO, and SemRep).
4. What is automatic summarization and why is it important in the biomedical domain? What is the importance of semantic processing as the basis for automatic summarization?
5. Discuss differences between task-oriented evaluation of semantic processing and evaluation of the accuracy of semantic predications identified in text. Which one do you think is harder and why?