THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

An Intramural Research Division of the U.S. National Library of Medicine

A Report to the Board of Scientific Counselors September 2012 TR-2012-002

Semantic Knowledge Representation Project: Advanced Information Management for Biomedical Research

Thomas C. Rindflesch, PhD, Principal Investigator
Marcelo Fiszman, MD, PhD
Halil Kilicoglu, PhD
Graciela Rosemblat, PhD
Dongwook Shin, PhD
Michael J. Cairelli, DO
Guocai Chen, PhD
Christopher M. Miller, MD
T. Elizabeth Workman, PhD

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

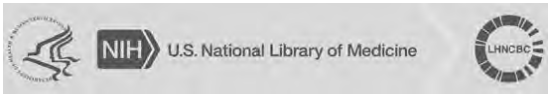

Table of Contents

1. Background

It is increasingly challenging for researchers and health professionals to exploit the extensive textual resources provided by NLM. High throughput natural language processing applications can supplement Library information retrieval services such as PubMed. The Semantic Knowledge Representation (SKR) project conducts basic research in symbolic natural language processing based on the UMLS knowledge sources. In addition, project staff develop methodologies and applications for advanced biomedical information management.

2. Project Objectives

A core resource is the SemRep program, which extracts semantic propositions from biomedical text. A facility is under development for extending SemRep to process proposition-modifying information, including speculations, opinions, evidence, and attitudes. Processing is knowledge intensive and relies on the UMLS Metathesaurus and Semantic Network as an ontology underpinning the identification of propositions in biomedical text. SemRep was originally developed for biomedical research. A general methodology is being devised for extending its domain, currently to influenza epidemic preparedness, health promotion, and medical informatics language processing.

SKR efforts support innovative information management approaches in biomedicine, as well as basic research. The Semantic MEDLINE Web application integrates information retrieval, advanced natural language processing, automatic summarization, and visualization into a single Web portal. The project team is using semantic predications to find publications that support critical questions used during the creation of clinical practice guidelines (with support from NHLBI). Other work exploits predications and graph theory for automatic summarization of biomedical text. Significant research is being devoted to developing and applying the literature-based discovery paradigm using semantic predications.

3. Project Significance

In consonance with the National Library of Medicine (NLM) long-term commitment to basic research, the Semantic Knowledge Representation project contributes to the NLM leadership in medical informatics by providing enhanced access to the biomedical research literature.

Project research using literature-based discovery to investigate declining sleep quality in aging men was recently published in *Sleep*, the premier venue for sleep research [1].

A repository of SemRep predications extracted from the entire set of PubMed citations (SemMedDB) is made available for research as a knowledge resource that can assist in hypothesis generation and literature-based discovery in biomedicine [2]. The current version of the repository consists of approximately 57.6 million semantic predications extracted from 21 million citations (dated 6/30/2012 or earlier) stored as a MySQL database and is available for non-commercial use at http://skr3.nlm.nih.gov/SemMedDB/. A UMLS Metathesaurus license is required.

Project staff are collaborating with several academic researchers in exploiting the semantic predications in SemMedDB:

- Trevor Cohen (U. Texas-Houston): Mathematical models for inferencing and discovery [3]
- J. Caleb Goodwin (U. Texas-Houston): Exploiting the ACT-R model of human memory for literature-based discovery modeling [4]
- Dimitar Hristovski (U. Ljubljana, Slovenia): Implemented system (BITOLA) for literature-based discovery [5]
- Guilherme Del Fiol (U. Utah): Automatic generation of patient-specific knowledge summaries to support clinical decision making [6]
- Serguei Pakhomov (U. Minnesota): Automatic semantic labeling of related terms in clinical text [7]

Dr. Fiszman has presented Semantic MEDLINE at the Biomedical Informatics course sponsored by NLM at the Marine Biology Laboratory in Woods Hole, MA. (http://hermes.mbl.edu/education/courses/special_topics/med.html)

Susan Roy (Library Associate) is investigating ways librarians can use tools such as Semantic MEDLINE (SM) to create connections with scientists to foster collaborations by promoting scientific innovation. [8]

4. Methods and Procedures

4.1. SemRep

SemRep [9,10] provides partial semantic interpretation of the biomedical research literature (MEDLINE citations). The system is rule based (guided by underspecified linguistic analysis) and symbolic (dependent on structured domain knowledge in the UMLS). Textual content is represented as semantic predications consisting of Metathesaurus concepts as arguments and Semantic Network relations as predicates. We exploit a modified version of the UMLS for SemRep processing.

SemRep first produces a partial syntactic analysis, relying on the SPECIALIST Lexicon [11] and the MedPost part-of-speech tagger [12]. Noun phrases in this structure are assigned Metathesaurus concepts (and semantic types) using MetaMap [13]. For example, parser output (2) for the text in (1) is the basis for further interpretation.

(1) … fish oils can protect against coronary heart disease …

(2)
[NP[mod(fish),head(oils),metaconc(Fish Oils),semtype(Pharmacologic Substance)],
[aux(can)],
[verb(protect)],
NP[prep(against),mod(coronary),mod(heart),head(disease),metaconc(Coronary heart disease),semtype(Disease or Syndrome)]]

Several mechanisms are involved in interpreting (2) as a semantic predication. First, "indicator" rules map syntactic elements such as verbs and nominalizations to predicates in the Semantic Network (e.g. TREATS, PREVENTS, or LOCATION_OF). The indicator rule needed for the current example is (3).

(3) *protect against* (verb) → PREVENTS

Following the application of indicator rules, argument identification rules establish syntactic relations between indicators and the heads of simple noun phrases serving as arguments. In order for an indicator and its syntactic arguments to be interpreted as a semantic predication, the semantic types of the Metathesaurus concepts for the noun phrases must match the semantic types serving as arguments of the indicated predicate in the Semantic Network. For example, PREVENTS allows the semantic type arguments in the Semantic Network ontological predication shown in (4).

(4) Pharmacologic Substance PREVENTS Disease or Syndrome

Since the semantic types in (4) match the semantic types of the syntactic subject and object of the indicator *protect* in (2), the semantic predication (5) is created, which substitutes the Metathesaurus concepts from the relevant arguments for the semantic types in the Semantic Network relation.

(5) Fish Oils PREVENTS Coronary heart disease

There is a principled distinction in SemRep between those aspects that apply to English in general and those specific to a particular domain. The underlying syntactic processing as well as argument identification are general, while indicator rules are domain specific (to a large extent). Metathesaurus concepts and relationships in the Semantic Network needed for interpretation are also domain specific. Due to this SemRep characteristic, in migrating to new domains, only domain-specific aspects need to be enhanced.

The core ontological space of SemRep is clinical medicine. This includes concepts and relationships about characteristics of diseases (such as etiology, body location, symptoms and comorbidities), descriptions of organisms, (including physiologic attributes and epidemiologic characteristics), as well as methods for diagnosing and treating patients and diseases. Some aspects of molecular biology are also addressed, including genetic etiology of disease and gene and protein interactions [14,15]. Finally, pharmacogenomics, which includes much of the foregoing in addition to the pharmacologic effect of substances on both anatomy and physiologic function, is also covered by SemRep [16].

The UMLS has been modified to accommodate the way this ontological space is expressed in the research literature. Our major modification to the Metathesaurus has been to block synonyms that are not generally valid, but only true in a limited domain covered by one of the constituent terminologies. For example, in addition to referring to the perception, the natural phenomenon and the disorder, "Cold" is also a synonym of

"Chronic Obstructive Airway Disease," "Cold Therapy," "Cold brand of chlorpheniramine-phenylpropanolamine," and "Colds homeopathic medication." Such "infelicitous" (spurious) synonyms often adhere to a pattern, in which one of them is a substring of the other. For example "Influenza" is a synonym of "Influenza Vaccines" in the Metathesaurus (in addition to being a synonym of "Flu"). We use the term "dysonym" for synonyms that are in a substring relationship and are only synonymous contextually, not paradigmatically. SemRep includes processing to eliminate dysonyms from the Metathesaurus. Spurious synonymy generates spurious ambiguity.

The UMLS Semantic Network forms the core of the SemRep ontology, but we have added and deleted some predications and changed the meaning of others by manipulating allowable semantic type arguments. Of the 54 relations in the UMLS Semantic Network, the following changes have been made. Twelve relations are used with some changes to semantic type arguments (part_of, location_of, precedes, affects, treats, complicates, prevents, produces, causes, uses, diagnoses, method_of). Seven relations have been redefined (associated_with, co-occurs_with, disrupts, interacts_with, occurs_in, process_of, manifestation_of). Seven relations not originally in the UMLS Semantic Network have been added (coexists_with, administered_to, stimulates, inhibits, converts_to, augments, predisposes). Other relations in UMLS are not currently interpreted by SemRep.

Recent work on SemRep has addressed additional syntactic structures, namely comparatives and arguments of nominalizations. A major enhancement is underway to extend SemRep beyond propositional meaning by interpreting embedding predications. Finally, we are developing a method for migrating SemRep to additional domains beyond clinical medicine and basic biomedical research.

4.1.1. Comparative structures

The range of comparative expressions in English is extensive and complex [17,18]. SemRep recognizes a subset of such constructions [19], those in which two drugs are compared with respect to a shared characteristic (e.g. how well they treat some disease). The compared terms are expressed as noun phrases (primary and secondary), which can be considered to be conjoined. Their relative merit is indicated by position on a scale, and the shared characteristic is expressed as a predicate outside the comparative structure. An adjective or noun is used to denote the scale, and words such as *than, as, with,* and *to* serve as cues to identify the compared terms, the scale, and the relative position of the terms on the scale.

For example, SemRep comparative processing generates the predications in (7) from (6).

(6) Amoxicillin-clavulanate was not as effective as ciprofloxacin for treating uncomplicated bladder infection in women.

(7) SCALE(effectiveness)
Amoxicillin-Potassium Clavulanate Combination COMPARED_WITH Ciprofloxacin
Amoxicillin-Potassium Clavulanate Combination LOWER_THAN Ciprofloxacin

Ciprofloxacin TREATS Infective cystitis
Amoxicillin-Potassium Clavulanate Combination TREATS Infective cystitis
Infective cystitis PROCESS_OF Woman

To evaluate the effectiveness of the developed methods we created a test set of 300 sentences containing comparative structures. The overall effectiveness of comparative processing was .70 recall and .96 precision (.81 F-score).

4.1.2. Arguments of nominalizations

SemRep was recently enhanced for effective interpretation of a wide range of patterns used to express arguments of nominalization in clinically oriented biomedical text [20]. Nominalizations are pervasive in the scientific literature, yet few text mining systems adequately address them, thus missing a wealth of information. We limited this enhancement to nominalizations with two overt arguments. These fall into one of several patterns noted by [21], including those in which both arguments are to the right of the nominalization, cued by prepositions (*treatment of fracture with surgery*), the nominalization separates the arguments (*fracture treatment with surgery, surgical treatment for fracture*), and both arguments precede the nominalizations, as modifiers of it (*surgical fracture treatment* and *fracture surgical treatment*).

We found additional patterns in the clinical domain, including those in which the subject appears to the right marked by a verb (*the treatment of fracture is surgery*) or as an appositive (*the treatment of fracture, surgery*), and those in which the subject appears to the left and the nominalization is either in a prepositional phrase (*surgery in the treatment of fracture*, *surgery in fracture treatment*) or is preceded by a verb or is parenthetical (*surgery is (the best) treatment for fracture*; *surgery is (the best) fracture treatment; surgery, the best fracture treatment*). One pattern, in which both arguments are on the right and the subject precedes the object, is seen most commonly when the nominalization has a lexically specified cue (e.g. *the contribution of stem cells to kidney repair*).

Based on these generalizations, we enhanced SemRep and evaluated the system by assessing the algorithm independently and by determining its contribution to SemRep generally. The first evaluation demonstrated the strength of the method through an F-score of 0.646 (P=0.743, R=0.569), which is more than 20 points higher than the baseline. The second evaluation showed that overall SemRep results were increased to F-score 0.689 (P=0.745, R=0.640), approximately 25 points better than without nominalization processing.

4.1.3. Embedding predications

Recent research in the SKR project underpins a major expansion of SemRep expressiveness [22], which is currently limited to semantic propositions. Written communication is rarely a sequence of simple propositions. More often, in addition to simple assertions, authors express subjectivity, such as beliefs, speculations, opinions, intentions, and desires. Furthermore, they link statements of various kinds to form a coherent discourse that reflects their pragmatic intent.

Kilicoglu [22] contributes to the understanding of extra-propositional meaning in natural language understanding, by providing a comprehensive account of the semantic phenomena that occur beyond simple assertions and examining how a coherent discourse is formed from lower level semantic elements. Our approach is linguistically based, and we propose a general, unified treatment of the semantic phenomena involved, within a computationally viable framework. We identify semantic embedding as the core notion involved in expressing extra-propositional meaning. The embedding framework is based on the structural distinction between embedding and atomic predications, the former corresponding to extra-propositional aspects of meaning. It incorporates the notions of predication source, modality scale, and scope. We develop an embedding categorization scheme and a dictionary based on it, which provide the necessary means to interpret extra-propositional meaning with a compositional semantic interpretation methodology.

We distinguish four basic classes of embedding predicates: modal, relational, valence shifter and propositional, each class further divided into subcategories. In a nutshell, modal and valence shifter predicates are concerned with lower level extra-factual phenomena, introducing modal scales or providing meaning shifts with respect to these modal scales as well as with respect to polarity, respectively. On the other hand, relational predicates largely operate at the higher discourse coherence level, whereas the propositional predicates function at the basic propositional level.

In the embedding framework, predication construction is a bottom-up, compositional process that builds mainly on the following components:

- Syntactic dependency parse of each sentence in the document
- Word information, including lemma, part-of-speech, and positional information
- The embedding predicate dictionary
- (Optionally) additional semantic information associated with the document, in the form of semantic terms and atomic predications. Additional semantic information allows the framework to integrate with a relation extraction system, such as SemRep.

Using the components listed above, first, a semantic embedding graph representing the content of the document is constructed and semantic dependencies are made explicit, guided by transformation rules. Next, predications are composed by traversing the embedding graph in a bottom-up manner, guided by several compositional operations, such as argument identification and source propagation. Limited coreference resolution is also performed in predication composition.

SemRep and the embedding framework will be integrated to the fullest extent. The most immediate consequence of such integration could be determining the factual status of the predications in the Semantic MEDLINE database. Are they speculations, facts or counter-facts? What is the level of confidence associated with a predication?

The success of the framework in this task can be evaluated on the small corpus of SemRep relations recently annotated [23] or in a post-hoc analysis of a small set of

randomly selected relations from the database. This integration can also serve literature-based discovery and hypothesis generation tasks, for which SemRep relations have been exploited previously [1,3,24]. The embedding predication framework can enhance the value of semantic predications that contribute to these tasks by determining whether they are supported by strong, compelling evidence, based on their factual status and explicitness of evidence.

Since the MEDLINE database covers more or less all biomedical and life sciences research from mid-20th century forward, this level of information also gives us the ability to track how the scientific knowledge changes diachronically. One can, for example, assume that when a particular piece of biomedical information first appears in the literature (captured as a predication by SemRep), its factual status is more tentative, and in later periods, the same information is supported by more evidence or perhaps refuted by counter-evidence, which can be captured via embedding framework and aggregation over the entire predication database. This is essentially similar to the idea of capturing paradigm shifts, proposed by [25] albeit at a much larger scale.

4.1.4. Domain migration

SemRep was originally devised for the clinical domain and was subsequently extended to genetic etiology [14,15], and pharmacogenomics [16]. We have recently devised a domain-independent methodology that allows us to leverage existing UMLS knowledge by adapting well-known ontology engineering phases partially based on [26], and integrating them with the knowledge sources afforded by the UMLS, extending coverage within a newly defined semantic space. The ontological and terminological extensions implemented in the system to apply to other domains have been successfully deployed in medical informatics knowledge processing, disaster information management [27], and public health promotion [28].

To extend SemRep coverage to a new domain we draw on the 4-stage ontology engineering approach in [26]:

1. *Specification and conceptualization* define the ontology purpose and scope and provide the concepts, vocabulary, and relationships for ontology design.

2. *Formalization* establishes the ontological hierarchies and relationships (IS_A /PART_OF) to develop a domain ontology and sub-ontologies for use in the implementation phase.

3. *Implementation* of new domain knowledge focuses on compatibility with SemRep formatting and knowledge integration with existing UMLS concepts and relations.

4. *Evaluation and maintenance* focus on user evaluation of different components and technical/formative evaluation as the ontology is built.

The case study that follows illustrates the application of our methodology to extend SemRep to the field of medical informatics, and allows, for example, the predication (9) to be extracted from sentence (8).

(8) *The authors used information retrieval technology to search automatically for sentences in MEDLINE abstracts that support these 851 DIP interactions*. (PMID 18628915)
(9) MEDLINE (Information Construct) LOCATION_OF Sentences (Linguistic Artifact)

The UMLS semantic type was changed for some existing concepts. For example, the semantic type for Metathesaurus concept "Discharge summary" was changed from 'Intellectual Product' to 'Information Construct'. New concepts for which a domain-appropriate UMLS semantic type was available were added to the supplemental file, such as the expression "Semantic Processing," which was given UMLS semantic type 'Machine Activity'. Domain-relevant non-mapping noun phrases were analyzed to determine synonymy based on semantic analysis. For example "qa" was added as a synonym (variant) of UMLS "Question answering" ('Machine Activity'). Other domain-relevant non-UMLS-mapping noun phrases that bore no semantic similarity to existing Metathesaurus concepts were added as potential new concepts.

Twenty four new domain-specific semantic relations were defined to connect some of the new or redefined concepts. Examples include ANALYZES, CATEGORIZES, and COLLECTS. Discovering the links between text expressions and the relations they correspond to involves a manual process. Rules to map from the text to the new predicates were also stipulated. For example, text expressions that link to the predicate ANALYZES are: *analyze, analysis, examine, explore, insight, investigate, review, study, treat.*

To evaluate the precision of the predications generated by the enhanced version of SemRep, an evaluation was carried out by an expert in the field who did not participate in the ontology development process. The 500-citation evaluation set contained 3,775 sentences and 2,092 predications, of which only the first 304 predications were evaluated for correctness, which corresponded to the first 663 sentences in the set. The evaluation yielded precision of 0.90.

Influenza epidemic information management
Explosion of disaster health information results in information overload among response professionals. A recent project extended SemRep to influenza epidemic information management. We characterized concepts and relationships commonly used in disaster health-related documents on influenza pandemics, as the basis for adapting the UMLS to the domain [27]. Three major themes emerged from analyzing the training documents on influenza epidemic; the first pertains to conducting surveillance, along with preventing and controlling disease. The second refers to actions of supervision, cooperation, and sponsorship among the organizations and entities involved in epidemics management. The third relates to dissemination of information during epidemics. We evaluated extended SemRep for influenza epidemic management on predications extracted from a test set of 371 sentences. The total number of predications retrieved was 604. Of these, 404 were correct and 200 were false positives, yielding precision of 0.67.

Public health promotion

Public health professionals require good information about successful health promotion policies and programs that might be considered for application within their own communities. We have extended the SemRep ontology to this domain [28]. To identify the main domain themes in public health promotion, we extracted 775 sentences representative of this domain and clustered them into subcategories per topic similarity. We identified three major themes: *Assessment* has to do with identifying and monitoring populations to identify health problems. *Policy Development* deals with the design and implementation of programs and policies to address health problems. *Assurance* deals with appropriate implementation of effective programs. In an evaluation set of 2,163 sentences from 218 MEDLINE citations, SemRep identified 773 predications. Of these, 658 were correct (true positive) and 115 were false positives, yielding precision of 0.85.

4.2. Exploiting semantic predications

4.2.1. Automatic summarization

We have developed a semantic abstraction approach to automatic summarization in the biomedical domain [29], which takes SemRep predications as input. There are four phases to our summarization processing.

Phase 1 (*relevance*), a condensation process, identifies predications on a given topic and is controlled by a semantic schema. Predications must conform to this schema in order to be included; such predications are called core predications. We currently rely on four schemas representing points of view in biomedicine: treatment [29]; diagnosis [30]; substance interactions [31]; and molecular biology [16].

Phase 2 (*connectivity*) is a generalization process that retrieves all predications sharing an argument with one of the core predications.

Phase 3 (*novelty*) provides further condensation by eliminating predications that have a generic argument, as determined by hierarchical depth in the Metathesaurus.

Phase 4 (*saliency*) is the final transformation phase. Frequency of occurrence for arguments, predicates, and predications are calculated, and those occurring more frequently than the average are kept in the condensate; others are eliminated.

We exploit graph theory for focusing summaries based on more than 500 citations [32]. The method is based on degree centrality, which measures connectedness in a graph. Four categories of clinical concepts related to treatment of disease were identified and presented as a summary of input text. A baseline was created using term frequency of occurrence. The system was evaluated on summaries for treatment of five diseases compared to a reference standard produced manually by two physicians. The results showed that recall for system results was 0.72, precision was 0.73, and F-score was 0.72. The system F-score was considerably higher than that for the baseline (0.47).

We have developed a clique-clustering method to automatically summarize graphs of semantic predication produced from large numbers of PubMed citations (titles and

abstracts) [33]. Cliques are identified from frequently occurring predications with highly connected arguments filtered by degree centrality. Themes contained in the summary were identified with a hierarchical clustering algorithm based on common arguments shared among cliques. The validity of the clusters in the summaries produced was compared to the Silhouette-generated baseline for cohesion, separation, and overall validity. The theme labels were also compared to a reference standard produced with major MeSH headings. For 11 topics in the testing data set, the overall validity of clusters from the system summary was 10% better than the baseline (43% versus 33%). While compared to the reference standard from MeSH headings, the results for recall, precision, and F-score were 0.64, 0.65, and 0.65 respectively.

Figure 1 illustrates the final graphical summary for 16,799 citations on schizophrenia. Four themes are highlighted in color: Etiology (yellow), Procedure treatment (green), Drug treatment (violet), and Disease comorbidities (gray). Notably in this summary, delusions and hallucinations are seen as comorbidities of schizophrenia, while dopamine, glutamate and neurotransmitters are associated with its etiology. Drug treatment constitutes the largest cluster; in addition to representing major drugs for schizophrenia (linked by blue TREATS arcs), it shows comparison between two drugs (purple arcs, COMPARED_WITH), and some adverse effects resulting from the drugs, such as weight gain and tardive dyskinesia (red arcs, CAUSES).

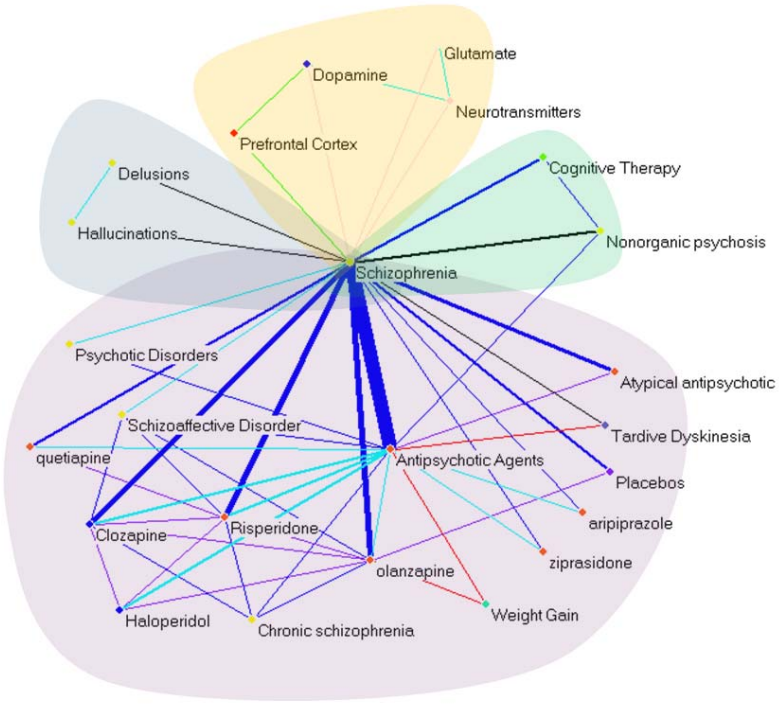

Figure 1. Summary and theme partitions for schizophrenia

4.2.2. Semantic MEDLINE

To support more effective biomedical information management, we have developed Semantic MEDLINE [34,35], which integrates document retrieval, advanced natural language processing, automatic summarization, and visualization into a single Web

portal. The application is intended to help manage the results of PubMed searches by condensing core semantic content in the citations retrieved. Output is presented as a connected graph of semantic relations, with links to the original MEDLINE citations. The ability to connect salient information across documents helps users keep up with the research literature and discover connections which might otherwise go unnoticed. Semantic MEDLINE can make an impact on biomedicine by supporting scientific discovery and the timely translation of insights from basic research into advances in clinical practice and patient care.

Semantic MEDLINE is implemented as a three-tier, Java EE-based Web application (Figure 2), which allows separation of user interface, application logic, and data storage. We leverage mature open-source technologies to the extent possible. The application runs in a Tomcat servlet container on an Apache http server and has been developed using the Apache Struts Web application framework (http://struts.apache.org/). This encourages the use of the MVC (Model-View-Controller) paradigm to provide a clean separation of application model, navigational code, and page design code through the use of Java Servlet API.

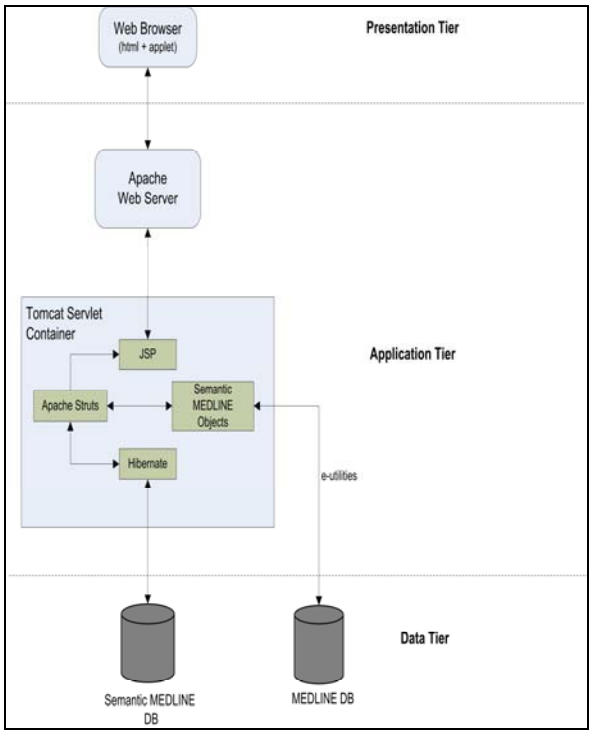

Figure 2. Semantic MEDLINE architecture

A MySQL database is used to store Semantic MEDLINE data, which includes semantic predications extracted from MEDLINE citations in addition to UMLS Metathesaurus and Entrez Gene data. The database tables are pre-populated from plain text files that contain SemRep output and Metathesaurus/Entrez Gene data using Perl scripts. The Hibernate object/relational mapping (ORM) tool (http://www.hibernate.org/) provides enhanced database access through database connection pooling and query caching. Semantic MEDLINE supports PubMed searching through NCBI's Entrez Programming Utilities

API (http://eutils.ncbi.nlm.nih.gov/) to provide real-time access to PubMed records, retrieved and manipulated in XML format.

To visualize the semantic condensates as graphs in Semantic MEDLINE, we developed a Flash application using the Adobe Flex framework (http://www.adobe.com/products/flex) and the Flare visualization toolkit (http://flare.prefuse.org/), the ActionScript extension of the Prefuse toolkit written in Java. Nodes in a graph represent arguments in SemRep predications, and the arcs predicates. We enhanced the visualization capabilities provided by Flare by linking the semantic predications in the graph to external structured biomedical resources.

Arcs are linked to the MEDLINE citations from which the corresponding predications were extracted, while nodes are linked to three resources in addition to Entrez Gene: the UMLS Semantic Navigator [36], Online Mendelian Inheritance in Man (OMIM) [37], and Genetics Home Reference [38]. Linking to the UMLS Semantic Navigator uses Metathesaurus concept identifiers (CUI) and allows the user to view the context of a predication argument in the UMLS hierarchy.

SemRep is not fast enough to accommodate Semantic MEDLINE in real time. We therefore run SemRep on the MEDLINE database in an off-line process and store the extracted predications in the MySQL database as they become available.

We describe a scenario exploiting the components of Semantic MEDLINE to elucidate relaxin, a peptide hormone originally connected with parturition and more recently found to have a wider range of physiological implications. The user issues the PubMed query "relaxin" to Semantic MEDLINE, retrieving 349 citations, which generate 2899 predications. These are summarized and displayed as a graph (Figure 3) which provides an informative overview of the characteristics of relaxin as extracted from the retrieved citations. Hierarchical structure in the Metathesaurus, accessible from graph nodes, provides general information about the entities that relaxin is involved with. For example, two of these are shown to be peptides:

- Angiotensin II → Angiotensins → peptide hormone
- Adenylate Cyclase → Intracellular Signalling Peptides and Proteins → Peptides

Perusal of predicate types in the graph elucidates the major characteristics of relaxin in a principled way. For example, it can be seen that "Relaxin" ISA Hormones, CAUSES: Premature Birth, AFFECTS Renal fibrosis, and INTERACTS_WITH RXFP2. The user can follow links to retrieve more detailed information on selected aspects of the graph. Figure 3 illustrates the citation from which the predication "Relaxin INTERACTS_WITH RXFP2" was extracted, for example.

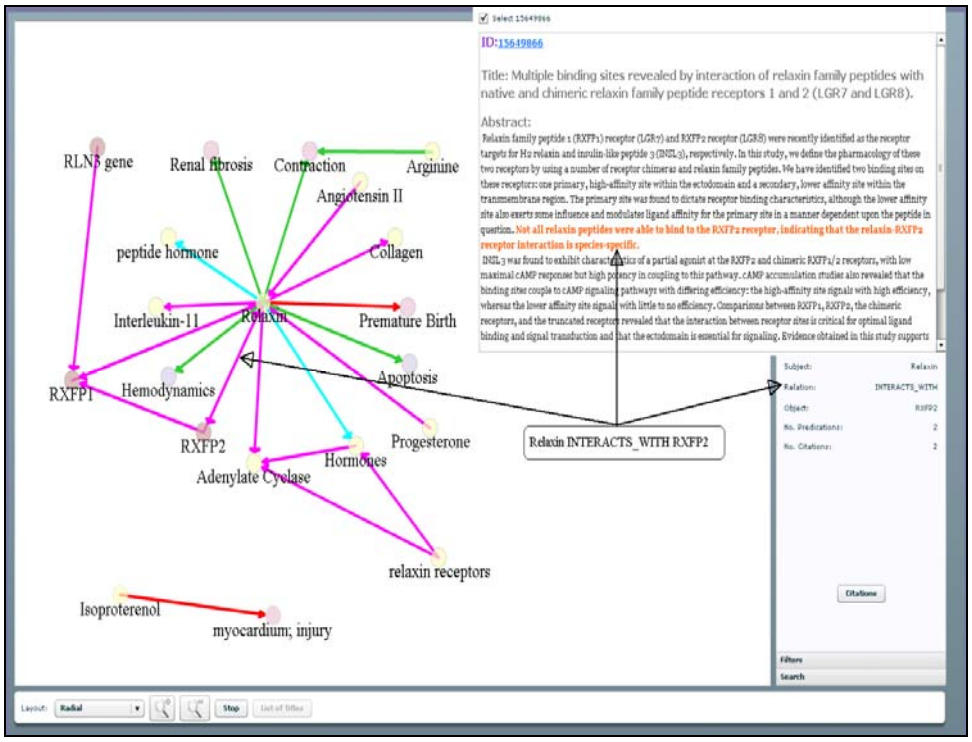

Figure 3. Visualizing summarization results for Relaxin search, with Relaxin INTERACTS_WITH RXFP2 relation highlighted.

4.2.3. Information management

With support from the National Heart Lung Blood Institute (NHLBI), NIH, we have adapted Semantic MEDLINE for a production environment for clinical practice guidelines, one of the main resources for communicating evidence-based practice to health professionals [39,40]. During guideline development, questions that express a knowledge gap are answered by finding relevant citations in MEDLINE and other biomedical databases. Determining citation relevance involves extensive manual review. We propose an automated method for finding relevant citations based on guideline question classification, semantic processing, and rules that match question classes with semantic predications. In this initial study, we focused on a pediatric cardiovascular risk factor guideline. The overall performance of the system was 0.40 recall, 0.88 precision (F0.5-score 0.71), and 0.98 specificity. We show that relevant and nonrelevant citations have clinically different semantic characteristics and suggest that this method has the potential to improve the efficiency of the literature review process in guideline development.

Successful guidelines depend on literature that is both relevant to the questions posed and based on high quality research in accordance with evidence-based medicine. Meeting these standards requires extensive manual review. We describe a system that combines symbolic semantic processing with a statistical method for selecting both relevant and high quality studies. We focused on a cardiovascular risk factor guideline, and the overall performance of the system was 0.56 recall, 0.91 precision (F0.5-score 0.81). If quality of the evidence is not taken into account, performance drops to 0.62 recall and 79%

precision (F0.5-score 0.75). We suggest that this system can potentially improve the efficiency of the literature review process in guideline development [40].

We are currently adapting the Semantic MEDLINE technology for analyzing NIH grant applications. With support and collaboration from the Office of Portfolio Analysis, Division of Program Coordination and Strategic Initiatives (DPCPSI), Office of the Director, NIH, we are developing the SPA (Semantic Portfolio Analyst) application. The goal is to have a tool to facilitate scientific portfolio analysis by exploiting semantic processing. SPA will help NIH portfolio analysts explore grant application content for emerging biomedical discoveries and innovative research opportunities.

In modifying Semantic MEDLINE for portfolio analysis we left the basic design unchanged, but added a sophisticated search engine that produces ranked output [41] and accommodated grant application format. As part of this effort, we have enhanced SemRep to extract predications in the domain of genomic engineering. Based on the domain migration method discussed above, SemRep now identifies the predications in (11) from the sentence (10), for example.

(10) We will apply mass spectrometry to simultaneously quantify the phosphorylation kinetics of hundreds of specific tyrosines in the signaling pathways downstream of EGFR.

(11) Spectrum Analysis, Mass MEASURES Phosphorylation kinetics
Phosphorylation kinetics COEXISTS_WITH Signal Pathways

We have extracted 2,175,737 predications from 309,515 NIH grant applications (2008 through 2011).

4.2.4. Literature-based discovery

Drug therapies are often used effectively without their underlying mechanism being completely understood. In earlier work [42], we exploit the literature-based discovery paradigm to investigate these mechanisms and propose a discovery pattern that draws on semantic predications extracted from MEDLINE citations. The use of semantic predications and the discovery pattern provides a way to uncover previously unnoticed associations between pharmacologic and bioactive substances on the one hand and bioactive substances and disorders on the other. In this paper, we concentrate on research investigating the use of antipsychotic agents used for treatment of cancer. Our method resulted in five biomolecules that may provide a link between the antipsychotic agents and cancer: brain-derived neurotrophic factor, CYP2D6, glucocorticoid receptor, PRL, and TNF.

We recently presented an extension to literature-based discovery that goes beyond making discoveries to a principled way of navigating through selected aspects of a specified biomedical domain [43]. The method is a type of “discovery browsing” that guides the user through the research literature on a specified phenomenon. Poorly understood relationships may be explored through novel points of view, and potentially

interesting relationships need not be known ahead of time. In a process of "cooperative reciprocity" the user iteratively focuses system output, thus controlling the large number of relationships often generated in literature-based discovery systems. The underlying technology exploits SemRep semantic predications represented as a graph of interconnected nodes (predication arguments) and edges (predicates) (Figure 4). The system suggests paths in this graph, which represent chains of relationships. The methodology is illustrated with depressive disorder and focuses on the interaction of inflammation, circadian phenomena, and the neurotransmitter norepinephrine. Insight provided may contribute to enhanced understanding of the pathophysiology, treatment, and prevention of this disorder.

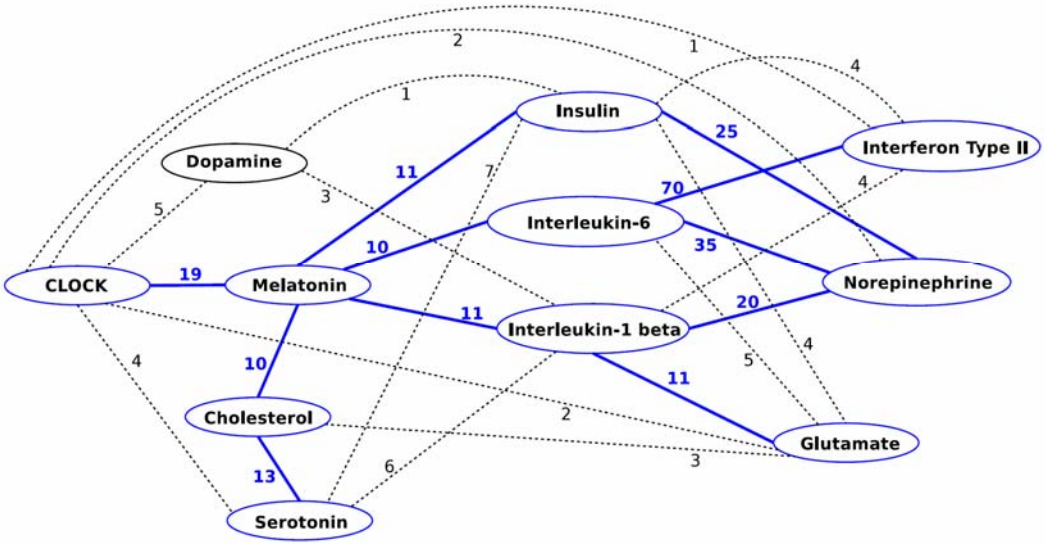

Figure 4. Graph representing the top twenty paths. Numbers on the edges show the number of predication instances.

We have also exploited semantic predications and the Semantic MEDLINE application to pursue an LBD method that involves user interaction to develop a hypothesis incrementally and iteratively [1]. In closed discovery research we propose cortisol as a mechanistic explanation for the observation that aging men suffer both from reduced levels of testosterone and degraded sleep quality. There are two parts to the methodology: (a) using Semantic MEDLINE to gain insight into the relationships between testosterone and sleep and, guided by the first part, (b) getting predications on testosterone and cortisol specifically. In the first part of the research, queries were issued to PubMed. For each query, predications extracted from the citations retrieved were presented by Semantic MEDLINE as a graph with links to the relevant text (e.g. Figure 5). Each graph was then manually inspected for predications involving testosterone and sleep. The full-text articles of the citations from which such predications had been extracted were then read for their contribution to the developing hypothesis. The follow-up query retrieved 16 predications linking testosterone and cortisol. One of these [44], in research not involving sleep, reports that testosterone significantly suppresses cortisol (which enhances wakefulness), which increases with age.

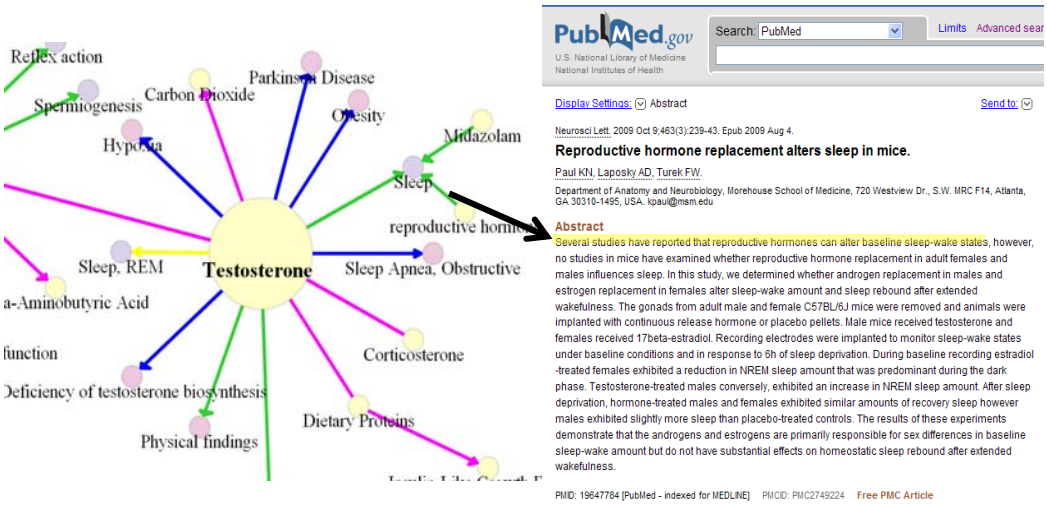

Figure 5. Semantic MEDLINE graph showing predications extracted from citations on testosterone.

We are currently exploring the etiology of restless legs syndrome (RLS) using the LBD paradigm [45]. The Semantic MEDLINE natural language processing application was used to implement this work. Based on the prominent relationship between dopamine and iron in the pathophysiology of RLS, a query consisting of these substances was issued to the application. Core content in citations retrieved was then visualized as a graph of interconnected semantic relationships. This graph was systematically examined for known facts that might underpin novel hypotheses about the etiology of RLS and revealed that melanin is involved in the etiology of Parkinson's disease (PD), through both dopamine and iron. Since there are similarities between RLS and PD involving these substances, we hypothesize that melanin is also significant in the etiology of RLS.

In work in progress, we are using LBD guided by systems biology/system medicine to enhance knowledge of the pathophysiology of sleep apnea in order to propose more effective drug therapies than those currently in use for this disorder. We used semantic predications and graph theory analysis to implement a discovery method. Based on the assessment that sleep apnea is necessarily a neurologic phenomenon [46], we concentrate on the neurology (neurotransmitters in particular) of sleep apnea. The results are initially relevant to central sleep apnea and neurologic aspects of obstructive sleep apnea. We hypothesize that in OSA acetylcholine and glutamate are abnormally decreased, while GABA is abnormally increased. We therefore propose rational drug design to redress this imbalance

5. Evaluation

We have conducted numerous evaluations assessing SemRep accuracy and the effectiveness of applications we are developing based on semantic predications. Several evaluations assessed SemRep accuracy focusing on particular predicates. In [9] we evaluated 830 instances of ISA, with precision of 0.83. In [14] we judged 1,124 sentences containing predications interpreting assertions on the genetic etiology of disease (ASSOCIATED_WITH, PREDISPOSES, NEG_ASSOCIATED_WITH, CAUSES, NEG_PREDISPOSES, NEG_CAUSES). Precision was 0.76. We constructed a reference

standard for predications on pharmacogenomics, annotating 623 relevant predicates (ASSOCIATED_WITH, CAUSES, PREDISPOSES, INTERACTS_WITH, INHIBITS, STIMULATES, AFFECTS, DISRUPTS, AUGMENTS, ADMINISTERED_TO, MANIFESTATION_OF, TREATS, LOCATION_OF, PART_OF, PROCESS_OF) [16]. Results of SemRep evaluation were: recall 0.55 (95% confidence interval 0.49 to 0.61) and precision 0.73 (95% confidence interval 0.65 to 0.81).

In the context of research on automatic summarization, we have also evaluated SemRep effectiveness by concentrating on the content of the text processed. In [29] we performed a linguistic evaluation on the condensates generated for four diseases: migraine, angina pectoris, Crohn's disease, and pneumonia. The input for each summary was 300 MEDLINE citations. 306 predications (ISA, CAUSES, TREATS, LOCATIONS_OF, OCCURS_IN, CO-OCCURS_WITH) were found to have 0. 66 precision. In [47] we conducted a similar evaluation on four additional diseases (gout, hyperthyroidism, migraine, and chest pain) described in an online medical encyclopedia. Overall precision was 0.87. In [31] we performed a linguistic evaluation on the quality of INTERACTS_WITH and the AFFECTS predications after the Saliency phase of our automatic abstraction summarization. A sample of ten drugs was categorized as follows: Central nervous system: citalopram, paroxetine, phenytoin, and selegiline; Antiviral: efavirenz; Heart: enalapril; Gastrointestinal: lansoprazole and ranitidine; Vascular: sumatriptan; Skin: voriconazole. 203 predications were evaluated, with 68% precision

Two evaluations have been conducted on sentences with a specified linguistic structure addressed by SemRep. In [19] we created a reference standard of 300 sentences containing comparative structures. Recall and precision overall for all comparative structures were 0.70 and 0.96 respectively. In a second evaluation, three-hundred sentences from 239 MEDLINE titles and abstracts were selected for annotating a test set for arguments of nominalizations [20]. We then performed two evaluations. The first evaluated nominalizations in isolation, while the second assessed the effect of the enhancements on overall semantic interpretation in SemRep. The results for the first evaluation were 0.57 recall and 0.74 precision. For the second evaluation, recall was 0.64 and precision was 0.75, while the baseline (with no nominalization processing) was 0.33 recall and 0.64 precision.

In [48] we evaluated the ability of our automatic semantic abstraction summarization system to identify useful drug interventions for fifty-three diseases in MEDLINE citations. The evaluation methodology used existing sources of evidence-based medicine as surrogates for a physician-annotated reference standard. Mean average precision (MAP) and a clinical usefulness score developed for this study were computed as performance metrics. The automatic summarization system significantly outperformed the baseline in both metrics. The MAP gain was 0.17 ($p < 0.01$) and the increase in the overall score of clinical usefulness was 0.39 ($p < 0.05$).

In [23], we present a multi-phase gold standard annotation study, in which we annotated 500 sentences randomly selected from MEDLINE abstracts on a wide range of biomedical topics with 1371 semantic predications. We measured interannotator

agreement and analyzed the annotations closely to identify some of the challenges in annotating biomedical text with relations based on an ontology or a terminology. While the resulting gold standard is mainly intended to serve as a test collection for SemRep development, we believe that the lessons learned are applicable generally.

We are in the process of exploiting the Lister Hill Center Usability Lab to conduct user-centered evaluations for Semantic MEDLINE development.

6. Project Status and Future Plans

After considerable concentration on underpinning development, SKR research has attained a level of maturity poised for meaningful impact on biomedical research. SemRep has been applied to all of MEDLINE, with the 57 million extracted semantic predications made available to the community. Evaluation suggests that accuracy is sufficient for practical application, and several universities are exploiting this resource for both clinical use and basic biomedical research, particularly in the literature-based discovery paradigm. Additionally, NIH staff are looking to SKR applications for developing clinical practice guidelines and scientific portfolio analysis. SKR staff, including postdoctoral fellows, are using SemRep predications and the Semantic MEDLINE application with notable success in developing and exploiting literature-based discovery methodology. We plan to intensify this effort, supported by continuing enhancement of SemRep capabilities (e.g. domain migration and embedding predications) as well as improvements to Semantic MEDLINE, particularly focused on visualization and better integration with existing resources, such as the UMLS.

Project staff also plan to cooperate with broader initiatives aimed at exploiting Semantic MEDLINE on the Semantic Web. As an example, Dr. George Strawn, Director, National Coordination Office for Networking and Information Technology Research and Development, White House Office of Science and Technology Policy, has supported the porting of Semantic MEDLINE to a YarcData uRiKA graph appliance, a purpose-built computer for real-time relationship analysis on big data graphs. Dr. Rindflesch has been invited to participate in the 14th SOA e-Government Conference, to be held in October, to discuss this project.

7. References

1. Miller CM, Rindflesch TC, Fiszman M, Hristovski D, Shin D, Rosemblat G, Zhang H, Strohl KP. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. Sleep. 2012 Feb 1;35(2):279-85.

2. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. [submitted to Bioinformatics].

3. Cohen T, Widdows D, Schvaneveldt RW, Davies P, Rindflesch TC. Discovering discovery patterns with predication-based Semantic Indexing. J Biomed Inform. 2012 Jul 26. [Epub ahead of print]

4. Goodwin JC. Cohen T, Rindflesch TC. Discovery by scent: Closed literature-based discovery system based on the Information Foraging Theory. [submitted to The First International Workshop on the role of Semantic Web in Literature-Based Discovery].

5. Hristovski D, Rindflesch T, Peterlin B. Using Literature-based Discovery to Identify Novel Therapeutic Approaches. Cardiovasc Hematol Agents Med Chem. 2012 Jul 27. [Epub ahead of print]

6. Jonnalagadda S, Del Fiol G, Medlin Jr. R, Weir C, Mostafa J, Liu H. Fiszman M. Automatically extracting sentences from MEDLINE citations to support clinicians' information needs [submitted to the 2nd IEEE conference on Healthcare Informatics, Imaging, and Systems Biology (HISB 2012)].

7. Liu Y, Bill R, Fiszman M, Rindflesch TC, Pedersen T, Melton GB, Pakhomov SV. Using SemRep to label semantic relations extracted from clinical text. AMIA Annu Symp Proc. 2012.

8. Roy S, Rindflesch TC. Creating connections with scientists utilizing Semantic MEDLINE [submitted to 2012 MLA Quad Chapter Meeting].

9. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462-77.

10. Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In Chen H, Fuller S, Friedman C, Hersh W (eds.) Medical Informatics: Knowledge Management and Data Mining in Biomedicine. New York: Springer, 2005:399-422.

11. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994:235-9.

12. Smith L, Rindflesch T, Wilbur WJ. MedPost: a part-of-speech tagger for biomedical text. Bioinformatics 2004;20(14):2320-1.

13. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May-Jun;17(3):229-36.

14. Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. Semantic relations asserting the etiology of genetic diseases. AMIA Annu Symp Proc. 2003:554-8.

15. Masseroli M, Kilicoglu H, Lang FM, Rindflesch TC. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. BMC Bioinformatics. 2006 Jun 8;7:291.

16. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput. 2007:209-20.

17. Ryan K. Corepresentational grammar and parsing English comparatives. Proc 19th Annual Meeting Assoc Comp Linguistics 1981:13-18.

18. Huddleston R, Pullum GK. 2002. The Cambridge Grammar of the English Language. Cambridge University Press, 2002.

19. Fiszman M, Demner-Fushman D, Lang FM, Goetz P, Rindflesch TC. Interpreting comparative constructions in biomedical text. Proceedings of the BioNLP Workshop, Association for Computational Linguistics 2007:137-44.

20. Kilicoglu H, Fiszman M, Rosemblat G, Marimpietri S, Rindflesch TC.. Arguments of nominals in semantic interpretation of biomedical text. Proceedings of the BioNLP Workshop, Association for Computational Linguistics 2010:46-54.

21. Cohen KB, Palmer M, Hunter L. 2008. Nominalization and alternations in biomedical language. PLoS ONE 2008:3(9): e3158.

22. Kilicoglu H. Embedding predications. Concordia University Ph.D. dissertation, 2012.

23. Kilicoglu H, Rosemblat G, Fiszman M, Rindflesch TC. Constructing a semantic predication gold standard from the biomedical literature. BMC Bioinformatics. 2011 Dec 20;12:486.

24. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc. 2006:349-53.

25. Lisacek F, Chichester C, Kaplan A. Discovering Paradigm Shift Patterns in Biomedical Abstracts: Application to Neurodegenerative Diseases. Proceedings of the

First International Symposium on Semantic Mining in Biomedicine (SMBM) 2005:41-50.

26. Kuziemsky CE, Lau F. A four stage approach for ontology-based health information system design. Artificial Intelligence in Medicine 2010;50:133-48

27. Keselman A, Rosemblat R, Kilicoglu H, Fiszman M, Jin H, Shin D, Rindflesch TC. Adapting semantic natural language processing technology to address information overload in influenza epidemic management. Journal of the American Society for Information Science and Technology 2010;61(12):2531-43.

28. Rosemblat G, Resnick MP, Auston I, Shin D, Sneiderman CA, Rindflesch TC. Extending SemRep to the public health domain [submitted to Journal of the American Medical Informatics Association].

29. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. HLT/NAACL Workshop on Computational Lexical Semantics 2004:76-83.

30. Sneiderman C, Demner-Fushman D, Fiszman M, Rosemblat G, Lang FM, Norwood D, Rindflesch TC. Semantic processing to enhance retrieval of diagnosis citations from Medline. AMIA Annu Symp Proc. 2006:1104.

31. Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in Medline citations. AMIA Annu Symp Proc. 2006:254-8.

32. Zhang H, Fiszman M, Shin D, Miller CM, Rosemblat G, Rindflesch TC. Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform. 2011 Oct;44(5):830-8.

33. Zhang H, Fiszman M, Shin D, Wilkowski B, Rindflesch TC. Clustering cliques for graph-based summarization of the biomedical research literature [submitted to BMC Bioinformatics].

34. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple AM, Rindflesch TC. Semantic MEDLINE: A Web application to manage the results of PubMed searches. Proceedings of the Third International Symposium for Semantic Mining in Biomedicine 2008:69-76.

35. Rindflesch TC, Kilicoglu H, Fiszman M, Rosemblat G, Shin D. Semantic MEDLINE: An advanced information management application for biomedicine. Information Systems and Use 2011;31(1,2):15-21.

36. Bodenreider, O. A semantic navigation tool for the UMLS. AMIA Annu Symp Proc. 2000.

37. Hamosh A, Scott AF, Amberger J., Bocchini C, Valle D, McKusick VA Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 2001;30(1):52-5.

38. Mitchell, J. A., Fun, J., and McCray, A.T. (2004) Design of Genetics Home Reference: A new NLM consumer health resource. Journal of the American Medical Informatics Association. 11(6):439-47.

39. Fiszman M, Ortiz E, Bray BE, Rindflesch TC. Semantic processing to support clinical guideline development. AMIA Annu Symp Proc. 2008 Nov 6:187-91.

40. Fiszman M, Bray BE, Shin D, Kilicoglu H, Bennett GC, Bodenreider O, Rindflesch TC. Combining relevance assignment with quality of the evidence to support guideline development. Stud Health Technol Inform. 2010;160(Pt 1):709-13.

41. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc. 2007 May-Jun;14(3):253-63.

42. Ahlers CB, Hristovski D, Kilicoglu H, Rindflesch TC. Using the literature-based discovery paradigm to investigate drug mechanisms. AMIA Annu Symp Proc. 2007 Oct 11:6-10.

43. Wilkowski B, Fiszman M, Miller CM, Hristovski D, Arabandi S, Rosemblat G, Rindflesch TC. Graph-based methods for discovery browsing with semantic predications. AMIA Annu Symp Proc. 2011:1514-23.

44. Rubinow DR, Roca CA, Schmidt PJ, et al. Testosterone suppression of CRH-stimulated cortisol in men. Neuropsychopharmacology 2005;30:1906-12.

45. Miller CM, Koo B, Rindflesch TC, Strohl KP. Literature-based discovery suggests neuromelanin in the etiology of restless legs syndrome. Poster presented at SLEEP 2012, 27th Annual Meeting of the Associated Professional Sleep Societies.

46. Strohl KP. Rebuttal from Dr. Strohl. Am J Respir Crit Care Med 2003;168:273.

47. Fiszman M, Rindflesch TC, Kilicoglu H. Summarization of an online medical encyclopedia. MEDINFO 2004:506-10.

48. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13.