Pacific Symposium on Biocomputing 12:209-220(2007)

EXTRACTING SEMANTIC PREDICATIONS FROM MEDLINE CITATIONS FOR PHARMACOGENOMICS

CAROLINE B. AHLERS, [1] MARCELO FISZMAN, [2] DINA DEMNER-FUSHMAN, [1]
FRANÇOIS-MICHEL LANG, [1] THOMAS C. RINDFLESCH[1]

[1]*Lister Hill National Center for Biomedical Communications,*
National Library of Medicine
Bethesda, Maryland 20894, USA
[2]*The University of Tennessee, Graduate School of Medicine*
Knoxville, Tennessee 37920, USA

We describe a natural language processing system (Enhanced SemRep) to identify core assertions on pharmacogenomics in Medline citations. Extracted information is represented as semantic predications covering a range of relations relevant to this domain. The specific relations addressed by the system provide greater precision than that achievable with methods that rely on entity co-occurrence. The development of Enhanced SemRep is based on the adaptation of an existing system and crucially depends on domain knowledge in the Unified Medical Language System. We provide a preliminary evaluation (55% recall and 73% precision) and discuss the potential of this system in assisting both clinical practice and scientific investigation.

1. Introduction

We discuss the development of a natural language processing (NLP) system to identify and extract a range of semantic predications (or relations) from Medline citations on pharmacogenomics. Core research in this field investigates the interaction of genes and their products with therapeutic substances. Discoveries hold considerable promise for treatment of disease [1], as clinical successes, notably in oncology, demonstrate. For example, Gleevec is a first-line therapy for chronic myelogenous leukemia, as it attacks the mutant BCR-ABL fusion tyrosine kinase in cancer cells, leaving healthy cells largely unharmed [2].

Automatic methods, including NLP, are increasingly used as important aspects of the research process in biomedicine [3,4,5,6]. Current NLP for pharmacogenomics concentrates on co-occurrence information without specifying exact relations [7]. We are developing a system (called Enhanced SemRep in this paper) which complements that approach by representing assertions in text as semantic predications. For example, the predications in (2) are extracted from the sentence in (1).

1)These findings therefore demonstrate that dexamethasone is a potent inducer of multidrug resistance-associated protein expression in rat

hepatocytes through a mechanism that seems not to involve the classical glucocorticoid receptor pathway.

2)Dexamethasone STIMULATES Multidrug Resistance-Associated Proteins
Dexamethasone NEG_INTERACTS_WITH Glucocorticoid receptor
Multidrug Resistance-Associated Proteins PART_OF Rats
Hepatocytes PART_OF Rats

Enhanced SemRep is based on two existing systems: SemRep [8,9] and SemGen [10,11]. SemRep extracts semantic predications from clinical text, and SemGen was developed from SemRep to identify etiologic relations between genetic phenomena and diseases. Several aspects of these programs were combined and modified to identify a range of relations referring to genes, drugs, diseases, and population groups. The enhanced system extracts pharmacogenomic information down to the gene level, without identifying more specific genetic phenomena, such as mutations (e.g., CYP2C9*3), single nucleotide polymorphisms (e.g., C2850T), and haplotype information. In this paper we describe the major issues involved in developing Enhanced SemRep for pharmacogenomics.

2. Background

2.1. *Natural Language Processing for Biomedicine*

Several NLP systems identify relations in biomedical text. Due to the complexity of natural language, they often target particular semantic relations. In order to achieve high recall, some methods rely mainly on co-occurrence of entities in text (e.g. Yen et al. [12] for gene-disease relations). Some approaches use machine learning techniques to identify relations, for example Chun et al. [13] for gene-disease relations. Syntactic templates and shallow parsing are also used, by Blaschke et al. [14] for protein interactions, Rindflesch et al. [15] for binding, and Leroy et al. [16] for a variety of relations. Friedman et al. [17] use extensive linguistic processing for relations on molecular pathways, while Lussier et al. [18] use a similar approach to identify phenotypic context for genetic phenomena.

In pharmacogenomics, methods for extracting drug-gene relations have been developed, based on co-occurrence of drug and gene names in a sentence [19, 7]. The system described in [19] is limited to cancer research, while Chang et al. [7] use machine learning to assign drug-gene co-occurrences to one of several broad relations, such as genotype, clinical outcome, or pharmacokinetics. The system we present here (Enhanced SemRep) addresses a

wide range of syntactic structures and specific semantic relations pertinent to pharmacogenomics, such as STIMULATES, DISRUPTS, and CAUSES. We first describe the structure of the domain knowledge in the Unified Medical Language System (UMLS) [20], upon which the system crucially depends.

2.2. *The Unified Medical Language System*

The Metathesaurus and the Semantic Network are components of the UMLS representing structured biomedical domain knowledge. In the current (2006AB) release, the Metathesaurus contains more than a million concepts. Editors combine terms from constituent sources having similar meaning into a concept, which is also assigned a semantic type, as in (3).

3) **Concept**: fever; **Synonyms**: pyrexia, febrile, and hyperthermia;
Semantic Type: 'Finding'

The Semantic Network is an upper level ontology of medicine. Its core structure consists of two hierarchies (entities and events) of 135 semantic types, which represent the organization of phenomena in the medical domain.

4) Entity
Physical Object
Anatomical Structure
Fully Formed Anatomical Structure
Gene or Genome

Semantic types serve as arguments of "ontological" predications that represent allowable relationships between classes of concepts in the medical domain. The predicates in these predications are drawn from 54 semantic relations. Some examples are given in (5).

5) 'Gene or Genome' PART_OF 'Cell'
'Pharmacologic Substance' INTERACTS_WITH 'Enzyme'
'Disease or Syndrome' CO-OCCURS_WITH 'Neoplastic Process'

Semantic interpretation depends on matching asserted semantic predications to ontological semantic predications, and the current version of SemRep depends on the unedited version of the UMLS Semantic Network for this matching. One of the major efforts in the development of Enhanced SemRep was to edit the Semantic Network for application in pharmacogenomics.

2.3. *SemRep and SemGen*

SemRep: SemRep [8,9] is a rule-based symbolic natural language processing system developed to extract semantic predications from Medline citations on clinical medicine. As the first step in semantic interpretation, SemRep produces

an underspecified (or shallow) syntactic analysis based on the SPECIALIST Lexicon [21] and the MedPost part-of-speech tagger [22]. The most important aspect of this processing is the identification of simple noun phrases. In the next step, these are mapped to concepts in the Metathesaurus using MetaMap [23]. The structure in (7) illustrates syntactic analysis with Metathesaurus concepts and semantic types (abbreviated) for the sentence in (6).

6) Phenytoin induced gingival hyperplasia

7) [[head(noun(phenytoin)), metaconc('Phenytoin':[orch,phsu]))],
[verb(induced)], [head(noun(['gingival hyperplasia')),
metaconc('Gingival Hyperplasia':[dsyn]))]]

The structure in (7) serves as the basis for the final phase in constructing a semantic predication. During this phase, SemRep relies on "indicator" rules which map syntactic elements (such as verbs and nominalizations) to predicates in the Semantic Network, such as TREATS, CAUSES, and LOCATION_OF. Argument identification rules (which take into account coordination, relativization, and negation) then find syntactically allowable noun phrases to serve as arguments for indicators. If an indicator and the noun phrases serving as its syntactic arguments can be interpreted as a semantic predication, the following condition must be met: 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, in (7) the indicator *induced* maps to the Semantic Network relation in (8).

8) 'Pharmacological Substance' CAUSES 'Disease or Syndrome'

The concepts corresponding to the noun phrases *phenytoin* and *gingival hyperplasia* can serve as arguments because their semantic types ('Pharmacological Substance' (phsu) and 'Disease or Syndrome' (dsyn)) match those in the Semantic Network relation. In the final interpretation (9), The Metathesaurus concepts from the noun phrases are substituted for the semantic types in the Semantic Network relation.

9) Phenytoin CAUSES Gingival Hyperplasia

SemGen: SemGen [10,11] was adapted from SemRep in order to identify semantic predications on the genetic etiology of disease. The main consideration in creating SemGen was the identification of gene and protein names as well as related genomic phenomena. For this SemGen relies on ABGene [24], in addition to MetaMap and the Metathesaurus.

Since the UMLS Semantic Network does not cover molecular genetics, ontological semantic relations for this domain were created for SemGen. The allowable relations were defined in two classes: gene-disease interactions (ASSOCIATED_WITH, PREDISPOSE, and CAUSE) and gene-gene interactions (INHIBIT, STIMULATE, and INTERACTS_WITH).

3. Methods

The development of Enhanced SemRep for pharmacogenomics began with scrutiny of the pharmacogenomics literature to identify relevant predications not identified by either SemRep or SemGen. Approximately 1000 Medline citations were retrieved with queries containing drug and gene names. From these, 400 sentences were selected as containing assertions most crucial to pharmacogenomics, including genetic (gene-disease), genomic (gene-gene), and pharmacogenomic (drug-gene, drug-genome) relations; in addition relations between genes and population groups; relations between disease and population groups; and pharmacological relations (drug-disease, drug-pharmacological effect, drug-drug) were scrutinized. Examples of relevant assertions include:

10) N-acetyltransferase 2 plays an important role in Alzheimer's Disease. (gene-disease)
Ticlopidine is a potent inhibitor for CYP2C19. (drug-gene)
Gefitinib and erlotinib for tumors with epidermal growth factor receptor (EGFR) mutations or increased EGFR gene copy numbers. (drug-gene)
The CHF patients with the VDR FF genotype have higher rates of bone loss. (gene-disease and gene-process)

After processing these 400 sentences with SemRep, errors were analyzed and categorized for etiology. It was determined that the majority of errors were missed predications that could be accounted for under three broad categories: a) the Semantic Network, b) errors in argument identification due to "empty" heads, and c) Gene name identification. For Enhanced SemRep, gene name identification was addressed by adding ABGene [24] to the machinery provided by MetaMap and the Metathesaurus. The other classes of errors required more extensive modifications.

3.1. *Modification of Semantic Network for Enhanced SemRep*

The UMLS Semantic Network was substantially modified in enhanced SemRep. New ontological semantic predications were added and the definitions of others were modified. In order to accommodate semantic relations crucial to pharmacogenomics, semantic types stipulated as arguments of ontological semantic predications were reorganized into groups reflecting major categories in this field.

Semantic Types: Semantic groups have been defined to organize the finer grained UMLS semantic types into broader semantic categories relevant to the clinical domain [25]. For Enhanced SemRep, five semantic groups (Substance, Anatomy, Living Being, Process, and Pathology) were defined to permit

systematic and comprehensive treatment of arguments in predications relevant to pharmacogenomics. These semantic groups are used to stipulate allowable arguments of the ontological semantic predications defined for each domain. Each group for pharmacogenomics is defined as:

11) Substance: 'Amino Acid, Peptide, or Protein', 'Antibiotic', 'Biologically Active Substance', 'Carbohydrate', 'Chemical', 'Eicosanoid', 'Element, Ion, or Isotope', 'Enzyme', 'Gene or Genome', 'Hazardous or Poisonous Substance', 'Hormone', 'Immunologic Factor', 'Inorganic Chemical', 'Lipid', 'Neuroreactive Substance or Biogenic Amine', 'Nucleotide Sequence', 'Organic Chemical', 'Organophosphorous Compound', 'Pharmacologic Substance', 'Receptor', 'Steroid', 'Vitamin'
12) Anatomy: 'Anatomical Structure', 'Body Part, Organ, or Organ Component', 'Cell', 'Cell Component', 'Embryonic Structure', 'Fully Formed Anatomical Structure', 'Gene or Genome', 'Neoplastic Process', 'Tissue'
13) Living Being: 'Animal', Archaeon', 'Bacterium', 'Fungus', 'Human', 'Invertebrate', 'Mammal', 'Organism', 'Vertebrate', 'Virus'
14) Process: 'Acquired Abnormality', 'Anatomical Abnormality', 'Cell Function', 'Cell or Molecular Dysfunction', 'Congenital Abnormality', 'Disease or Syndrome', 'Finding', 'Injury or Poisoning', 'Laboratory Test Result', 'Organism Function', 'Pathologic Function', 'Physiologic Function', 'Sign or Symptom'
15) Pathology: 'Acquired Abnormality', 'Anatomical Abnormality', 'Cell or Molecular Dysfunction', 'Congenital Abnormality', 'Disease or Syndrome', 'Injury or Poisoning', Mental or Behavioral Disorder', 'Pathologic Function', 'Sign or Symptom'

In addition to grouping semantic types, semantic types assigned to two classes of Metathesaurus concepts were manipulated to handle the following generalizations.

16) Proteins are also genes. Concepts assigned the semantic type 'Amino Acid, Peptide, or Protein' are also assigned the semantic type 'Gene or Genome' ("Cytochrome P-450 CYP2E1" now has 'Gene or Genome' in addition to 'Amino Acid, Peptide, or Protein')
17) Group members are human. Concepts assigned the semantic type 'Group' (or its descendants) are also assigned the semantic type 'Human'. ("Child" now has 'Human' in addition to 'Age Group').

Predications: Predications for the pharmacogenomics domain were defined in the following categories (18-23). Ontological predications are defined by specifying allowable arguments, that is semantic types in the stipulated semantic

groups. The predications in (18-23) constitute a type of schema [26] for representing pharmacogenomic information.

18) *Genetic Etiology*:
{Substance} ASSOCIATED_WITH OR PREDISPOSES OR CAUSES {Pathology}
19) *Substance Relations* :
{Substance} INTERACTS_WITH OR INHIBITS OR STIMULATES {Substance}
20) *Pharmacological Effects*:
{Substance} AFFECTS OR DISRUPTS OR AUGMENTS {Anatomy OR Process}
21) *Clinical Actions*:
{Substance} ADMINISTERED_TO {Living Being}
{Process} MANIFESTATION_OF {Process}
{Substance} TREATS {Living Being OR Pathology }
22) *Organism Characteristics*:
{Anatomy OR Living Being} LOCATION_OF, {Substance}
{Anatomy} PART_OF {Anatomy OR Living Being}
{Process} PROCESS_OF {Living Being}
23) *Co-existence*:
{Substance} CO-EXISTS_WITH {Substance}
{Process} CO-EXISTS_WITH {Process}

3.2. *Empty Heads*

"Empty" heads [27,28] are a pervasive phenomenon in pharmacogenomics text. An example is *variants* in (24).

24) We saw differential activation of CYP2C9 variants by dapsone.

Nearly 80% of the 400 sentences in the training set contain at least one empty head. These structures impede the process of semantic interpretation. In SemRep the semantic type of the Metathesaurus concept corresponding to the head of a noun phrase qualifies that noun phrase for use as an argument. For example, from (24) we want to use the noun phrase *CYP2C9 variant* as an argument of STIMULATES, which requires that the semantic type of its object be a member of the Substance group. However, the semantic type of the head concept "Variant" is 'Qualitative Concept'.

As has been noted (e.g. [28]), such words are not really empty (in the sense of having no semantic content). A complete interpretation would take the meaning of empty heads into account. However, that is beyond the present capabilities of the Enhanced SemRep system. It is possible to get a partial interpretation of structures containing this phenomenon by ignoring the empty head [27].

We enumerated several categories of terms which we identified as semantically empty heads. These include general terms for genetic and genomic phenomena (*allele, mutation, polymorphism*, and *variant*), measurements (*concentration, levels*), and processes (*synthesis, expression, metabolism*). During processing in Enhanced SemRep, words from these lists that have been labeled as heads are hidden and the word to their left is relabeled as head. After this processing, *CYP2C9* becomes the head (with semantic type 'Gene or Genome', a member of the Substance group) in *CYP2C9 variants* above, thus qualifying as an argument of STIMULATES.

3.3. *Evaluation*

Enhanced SemRep was tested for recall and precision using a gold standard of 300 sentences randomly generated from the set of 36,577 sentences containing drug and gene co-occurrences found on the Web site [29] referenced by Chang and Altman [7]. These sentences were annotated by three physicians (CBA, DD-F, MF) for the predications discussed in the methods section. That is, we did not mark up all assertions in the sentences, only those representing a predication defined in Enhanced SemRep. A total of 850 predications were assigned to these 300 sentences by the annotators.

4. Results

Enhanced SemRep generated 623 predications from the 300 sentences in the test collection. Of these, 455 were true positives, 168 were false positives, and 375 were false negatives, reflecting recall of 55% (95% confidence interval 49% to 61%) and precision of 73% (95% confidence interval 65% to 81%).

We also calculated results for the groups of predications defined in the categories (18-22) above. Recall and precision for the predications in the five categories are: Genetic Etiology (ASSOCIATED_WITH, CAUSES, PREDISPOSES): 74% 74%; Substance Relations (INTERACTS_WITH, INHIBITS, STIMULATES): 50% 73%; Pharmacological Effects (AFFECTS, DISRUPTS, AUGMENTS): 41% 68%; Clinical Actions (ADMINISTERED_TO, MANIFESTATION_OF, TREATS): 54% 84%; Organism Characteristics (LOCATION_OF, PART_OF, PROCESS_OF): 63% 71%.

5. Discussion

5.1. *Error Analysis*

We assessed the etiology of errors separately for recall and precision. In considering both false negatives and false positives for Enhanced SemRep, the etiology of error was almost exclusively due to characteristics in SemRep before

enhancement, not to changes introduced for Enhanced SemRep. Word sense ambiguity was responsible for almost a third (28%) of all errors. For example, in interpreting (25), *inhibition* was wrongly mapped to the Metathesaurus concept "Psychological Inhibition," thus allowing the system to generate the false positive "CYP2C19 AFFECTS Psychological Inhibition."

25) Ticlopidine inhibition of phenytoin metabolism mediated by potent inhibition of CYP2C19.

Difficulty in processing coordinate structures caused more than a third (35%) of the false negatives seen in our evaluation. For example, in processing (26), although Enhanced SemRep identified the predication "Fluorouracil INTERACTS_WITH DPYD gene," it missed "mercaptopurine INTERACTS_WITH thiopurine methyltransferase."

26) The cytotoxic activities of mercaptopurine and fluorouracil are regulated by thiopurine methyltransferase (TPMT) and dihydropyrimidine dehydrogenase (DPD), respectively.

5.2. *Processing Medline citations on CYP2D6*

We processed 2849 Medline citations containing variant forms of CYP2D6 with Enhanced SemRep, which produced 36,804 predications, 22,199 of which were unique. 5219 total and 2310 unique predications contained CYP2D6 as an argument, with the remaining predications representing assertions about other genes, drugs, and diseases. The 5219 total predications containing CYP2D6 were analyzed according to two predication categories (Genetic Etiology and Substance Relations), and the results were compared with relations listed for this gene on the PharmGKB Web site [30].

Genetic Etiology: 267 total predications represented CYP2D6 as an etiologic agent (CAUSES, PREDISPOSES, or ASSOCIATED_WITH) for a disease. The most frequent of these are the following: Parkinson's disease (35 occurrences), carcinoma of the lung (21), tardive dyskinesia (15), Alzheimer's disease (9), bladder carcinoma (8). All of the above relations were judged to be true positives. Only carcinoma of the lung occurs in PharmGKB. Of the 4 PharmGKB CYP2D6-disease relations not obtained by SemRep (hepatitis C, ovarian carcinoma, pain, and bradycardia), two were found not to contain the disease name in the referenced citation (ovarian carcinoma and pain).

Substance Relations: Enhanced SemRep retrieved 1128 total predications involving CYP2D6 and a drug. Sixty-nine drugs occurred 3 or more times in those predications. Forty-one of the 69 were in PharmGKB and 28 were not. Sixty-eight were true positives. For example, The following drugs (all true positives) were interpreted by Enhanced SemRep as inhibiting CYP2D6:

quinidine (45 occurrences in 1128 predications with CYP2D6), paroxetine (34), fluoxetine (27), fluvoxamine (8), sertraline (8). Quinidine and sertraline are not in PharmGKB. SemRep also retrieved predications that the following drugs (all true positives) interact with CYP2D6: bufuralol (27), antipsychotic agents (25) dextromethorphan (21 occurrences), venlafaxine (19), debrisoquin (18). Bufuralol is not in PharmGKB. The PharmGKB relations SemRep failed to capture were CYP2D6 interactions with cocaine, levomepromazine, maprotiline, trazodone, and yohimbine. Two of these entries (levomepromazine and maprotiline) were found not to be based on the content of Medline citations.

6. Conclusion

We discuss the adaptation of an existing NLP system to apply in the pharmacogenomics domain. The major changes for developing Enhanced SemRep from SemRep involved modifying the semantic space stipulated by the UMLS Semantic Network. The output of Enhanced SemRep is in the form of semantic predications that represent assertions from Medline citations expressing a range of specific relations in pharmacogenomics. The information provided by Enhanced SemRep has the potential to contribute to systems that go beyond traditional information retrieval to support advanced information management applications for pharmacogenomics research and clinical care. In the future we intend to adapt the summarization and visualization techniques developed for clinical text [31] to the pharmacogenomic predications generated by Enhanced SemRep.

Acknowledgments

This study was supported in part by the Intramural Research Programs of the National Institutes of Health, National Library of Medicine. The first author was supported by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

1. Halapi E, Hakonarson H. Advances in the development of genetic markers for the diagnosis of disease and drug response. *Expert Rev Mol Diagn*. 2002 Sep;2(5):411-21.

2. Druker BJ, Talpaz M, Resta DJ, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N Engl J Med*. 2001 Apr 5;344(14):1031-7.
3. Yandell MD, Majoros WH. Genomics and natural language processing. *Nature Reviews Genetics* 2002;3(8):601-10.
4. K. Bretonnel Cohen and Lawrence Hunter Natural language processing and systems biology. In Dubitzky and Pereira, *Artificial intelligence methods and tools for systems biology*. Springer Verlag, 2004.
5. Hirschman L, Par JC, Tsujii J, Wong L, Wu CH. Accomplishments and challenges in literature data mining for biology. *Bioinformatics* 2002;18(12):1553-61.
6. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics* 2006;7:119-29.
7. Chang JT, Altman RB. Extracting and characterizing gene-drug relationships from the literature. *Pharmacogenetics*. 2004 Sep;14(9):577-86.
8. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J of Biomed Inform*. 2003 Dec;36(6):462-477.
9. Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In Chen, Fuller, Hersh, and Friedman, *Medical informatics: Knowledge management and data mining in biomedicine*. Springer, 2005, pp. 399-422.
10. 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.
11. 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(1):291.
12. Yen YT, Chen B, Chiu HW, Lee YC, Li YC, Hsu CY. Developing an NLP and IR-based algorithm for analyzing gene-disease relationships. *Methods Inf Med.* 2006;45(3):321-9.
13. Chun HW, Tsuruoka Y, Kim J-D, Shiba R, Nagata N, Hishiki T, Tsujii J. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. *Pac. Symp. Biocomput.* 2006:4-15.
14. Blaschke C, Andrade MA, Ouzounis C, Valencia A: Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology:* Edited by Lenauer T, Schneider R, Bork P, Brutlag DL, Glasgow JI, Mewes H-W, Zimmer R: San Francisco, CA: Morgan Kaufman Publishers, Inc; 1999:60-67.

15. Rindflesch TC, Rajan JV, Hunter L. Extracting molecular binding relationships from biomedical text. *Proceedings of the ANLP-NAACL* 2000:188-95. Association for Computational Linguistics.
16. Leroy G, Chen H, Martinez JD: A shallow parser based on closed-class words to capture relations in biomedical text. *J Biomed Inform*. 2003, 36(3):145-158.
17. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A: GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 2001, 17 Suppl 1:S74-S82.
18. Lussier YA, Borlawsky T, Rappaport D, Liu Y, Friedman C. PhenoGO: assigning phenotypic context to Gene Ontology annotations with natural language processing. *Pac Symp Bio*. 2006:64-75.
19. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* 2000, 517–528.
20. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical language System: An informatics research collaboration. *J Am Med Inform Assoc* 1998 Jan-Feb;5(1):1-11.
21. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care*. 1994;235-9.
22. Smith L, Rindflesch T, Wilbur WJ. MedPost: a part-of-speech tagger for biomedical text. *Bioinformatics*. 2004;20(14):2320-1.
23. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proc AMIA Symp*. 2001;17-21.
24. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text . *Bioinformatics*. 2002;18(8):1124-32.
25. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Medinfo* 2001;10(Pt 1):216-20.
26. Friedman C, Borlawsky T, Shagina L, Xing HR, Lussier YA. Bio-ontology and text: bridging the modeling gap. *Bioinformatics*. 2006 Jul 26.
27. Chodorow, Martin S., Roy I. Byrd, and George E. Heidom (1985). Extracting Semantic Hierarchies from a Large On-Line Dictionary. *Proceedings of the 23rd Annual Meeting of the ACL*, pp. 299-304.
28. Guthrie L, Slater BM, Wilks Y, Bruce R. Is there content in empty heads? *Proceedings of the 13th conference on Computational linguistics.* 1990; v3: 138 – 143.
29. http://bionlp.stanford.edu/genedrug/
30. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res*. 2002 Jan 1;30(1):163-5.
31. M Fiszman, TC Rindflesch, H Kilicoglu. Abstraction Summarization for Managing the Biomedical Research Literature. *Proc HLTNAACL Workshop on Computational Lexical Semantics*, 2004.