Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 36 (2003) 462–477

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text

Thomas C. Rindflesch* and Marcelo Fiszman

Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received 4 November 2003

Abstract

Interpretation of semantic propositions in free-text documents such as MEDLINE citations would provide valuable support for biomedical applications, and several approaches to semantic interpretation are being pursued in the biomedical informatics community. In this paper, we describe a methodology for interpreting linguistic structures that encode hypernymic propositions, in which a more specific concept is in a taxonomic relationship with a more general concept. In order to effectively process these constructions, we exploit underspecified syntactic analysis and structured domain knowledge from the Unified Medical Language System (UMLS). After introducing the syntactic processing on which our system depends, we focus on the UMLS knowledge that supports interpretation of hypernymic propositions. We first use semantic groups from the Semantic Network to ensure that the two concepts involved are compatible; hierarchical information in the Metathesaurus then determines which concept is more general and which more specific. A preliminary evaluation of a sample based on the semantic group Chemicals and Drugs provides 83% precision. An error analysis was conducted and potential solutions to the problems encountered are presented. The research discussed here serves as a paradigm for investigating the interaction between domain knowledge and linguistic structure in natural language processing, and could also make a contribution to research on automatic processing of discourse structure. Additional implications of the system we present include its integration in advanced semantic interpretation processors for biomedical text and its use for information extraction in specific domains. The approach has the potential to support a range of applications, including information retrieval and ontology engineering.
Published by Elsevier Inc.

Keywords: Natural language processing; Semantic processing; Knowledge representation; Information extraction

1. Introduction

Biomedical information management applications involving text either retrieve documents or extract information. Enabling technologies include word-based statistical methods and semantic processing to identify concepts and relations. Although statistical methods provide considerable success, improvement is needed.

Semantic predications identified in MEDLINE[1] citations on top of methodological search filters [1] increase the precision of retrieved citations [2]. For extracting information from text (concepts or relationships), statistical and pattern matching approaches have been attempted, but symbolic natural language processing has generally been more successful [3–10].

In response to these considerations, a number of researchers in biomedical informatics are examining the use of natural language processing for a range of applications, including medical knowledge acquisition, medical literature indexing and searching, automatic coding of clinical text, and processing molecular biology information (see [11,12]). Providing high quality results (including semantic propositions) with accuracy in the general case remains a matter for investigation, however.

In this paper, we propose a detailed analysis of the hypernymic proposition in English, a structure which

* Corresponding author. Fax: 1-301-496-0673.
E-mail address: tcr@nlm.nih.gov (T.C. Rindflesch).

[1] MEDLINE, Unified Medical Language System (UMLS), and Metathesaurus are registered trademarks of the National Library of Medicine.

1532-0464/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.jbi.2003.11.003

has not previously been the focus of extensive semantic interpretation methodologies. We investigate the processing of this structure as a way of improving a particular approach to semantic interpretation that supports enhanced access to biomedical documents and information. The hypernymic proposition involves two concepts that are in a taxonomic ('ISA') relationship, one semantically more specific, the hyponym, and the other more general, the hypernym. This is illustrated by the relationship between *modafinil* (hyponym) and *stimulant* (hypernym) in the sentence *modafinil is a novel stimulant that is effective in the treatment of narcolepsy*. This semantic structure appears frequently in scientific text and appears to function as a discourse phenomenon for accommodating the flow of new and old information.

Our proposal for identifying and interpreting hypernymic propositions is devised as an addition to a general knowledge-based semantic interpreter (called SemRep) that uses underspecified syntactic analysis and structured domain knowledge from the UMLS to identify semantic predications in biomedical text. SemRep has been used for processing coronary catheterization reports [13], extracting molecular biology information from the research literature [14,15], and identifying drug therapies in MEDLINE citations [16].

Currently, when interpreting the sentence above, SemRep identifies the semantic proposition "Stimulants TREATS Narcolepsy." Although this is correct, it would be more useful to identify "Modafinil" (hyponym of "Stimulant") as the semantic subject of TREATS in this sentence. The program discussed in this paper (called SemSpec) determines that "Modafinil" is a hyponym of "Stimulant" in this sentence and thus supplies the information that SemRep needs to provide the more precise semantic interpretation. These semantic propositions can potentially serve as MEDLINE indexing terms to support high-precision results from search queries such as "List all immunomodulators used to treat Crohn's Disease."

The research discussed here serves as a paradigm for investigating details of the interaction between domain knowledge and linguistic structure in natural language processing, and could also make a contribution to research on automatic processing of discourse structure and ontology engineering as well as information extraction in specific domains, such as pharmacology.

In the remainder of the paper, we first review research on natural language processing in the biomedical domain. Then, after introducing the permissible syntactic configurations encoding hypernymic propositions, we focus on the UMLS knowledge that supports the interpretation of these propositions and provide a brief overview of the SemRep approach to semantic interpretation. The Methods section concentrates on the SemSpec program. Finally, we describe a preliminary evaluation of the effectiveness of this processing and discuss improvements needed and directions for future work.

2. Background

2.1. NLP in the biomedical domain

Several research groups are developing and applying natural language processing methodologies in biomedical informatics, and systems vary along several dimensions. The complexity of natural language dictates that semantic interpretation be focused in scope, typically by domain of discourse; many applications are designed to interpret clinical text of a certain type, for example discharge summaries or imaging reports, such as chest X-rays or mammograms. The majority of this work is knowledge based, and the specific domain guides the type and amount of knowledge used [17]. Often this is drawn from existing resources, such as the Unified Medical Language System (UMLS) [18] or the GALEN ontology [19], but several systems rely solely on locally developed knowledge bases. Further, system restrictions may be imposed on the basis of syntactic structure; some process only noun phrases or just those phrases covered by a semantic grammar. Finally, various linguistic formalisms are used, including semantic grammars, definite clause and dependency grammars, as well as bottom-up chart parsers. Below, we discuss some of the NLP systems developed in the biomedical domain. (For more comprehensive reviews, see [11,12].)

MedLEE [20,21] builds on semantic models derived from the linguistic string project (LSP) [22] and is guided by a semantic grammar that consists of patterns of semantic classes, such as degree + change + finding, which would match *mild increase in congestion*. These classes are defined in a semantic lexicon, and Friedman et al. [23] discuss use of the UMLS in constructing such a lexicon. MedLEE has been evaluated for several clinical applications [5,24,25].

The AQUA system [26] was developed to interpret natural language queries issued by users to an information retrieval system. The parser uses standard definite clause grammars enhanced by an operator grammar, with the support of a semantic lexicon compiled from the UMLS Metathesaurus and Semantic Network. The final semantic representation is in the form of conceptual graphs. Although AQUA was developed for clinical queries, it has recently been applied to clinical data and MEDLINE citations, which are ranked based on a conceptual graph-matching algorithm [2].

The RECIT system [27] concentrates on processing noun phrases and is composed of a proximity processor, a typology of concepts, a dictionary with syntactic and semantic information, a set of conceptual relationships, and a set of canonical concepts. The semantic information relies on the model developed by the GALEN project [28].

Rosario et al. [10] describe an approach to the semantic interpretation of noun phrases and nominal compounds based on the semantic information contained in a large lexical hierarchy, the National Library of Medicine's Medical Subject Headings (MeSH). Part of the challenge addressed by their research is to determine the possible semantic relations that can obtain among the components of a nominal construction.

SymText [29] uses probabilistic Bayesian networks to represent semantic types and relations. Syntactic knowledge comes from augmented transition networks, and the system depends on a set of reports to train the network for a specific medical domain. SymText has been evaluated for clinical applications [6,30,31]. In a recent upgrade to SymText (called MPLUS) Bayesian networks are represented in an object-oriented format and a bottom-up chart parser provides syntactic analysis. In addition, MPLUS uses an abstract semantic language to link Bayesian network types to each other in a predication format [32].

The MENELAS system [33] is a multilingual text understanding system (French, Dutch, and English) built to extract information from patient discharge summaries. Domain knowledge resides in a locally developed ontology, and linguistic relations are projected to the reference model using morphosyntactic analysis. Output is in the form of an annotated parse tree that is subject to a semantic analyzer that heuristically selects the best representation using a semantic lexicon and semantic rules. MENELAS was evaluated for coding a subset of discharge summaries [34].

Hahn et al. [35] have developed a natural language processor called MEDSYNDIKATE to automatically acquire knowledge from medical reports. Grammatical knowledge comes from a lexicon and a fully specified dependency grammar. Conceptual knowledge comes from a locally developed ontology that consists of a set of axioms for concept roles with corresponding type restrictions for role fillers. In addition to sentence level analysis, MEDSYNDIKATE uses a centering algorithm to resolve anaphoric expressions at the discourse level [36]. The system has recently been evaluated for semantic propositions in sample medical texts [37].

Our approach to natural language processing differs from those described here in two major ways: the linguistic formalism used and the source of the domain knowledge. As will be seen below, syntactic structures are represented by two mechanisms, a shallow categorial parser and an underspecified dependency grammar. It should be noted that although these are both incomplete, they apply to English syntax in general and are not crafted for the biomedical domain. The domain knowledge for our system is taken directly from the UMLS rather than being compiled manually. Although the UMLS knowledge sources were not intended as ontologies and will not ultimately support extensive inferencing without enhancement, the breadth of coverage they provide supports the application of our system to a variety of medical subdomains with a minimum of effort.

2.2. Linguistic structure of hypernymic propositions

Before describing the way in which we automatically interpret the hypernymic proposition in our system, we provide some general discussion of this phenomenon based on examples seen in a study of MEDLINE citations pertaining to treatment (mostly drug therapy). Although the structure types encountered (and addressed in this study) are not exhaustive, they constitute a useful illustrative sample.

A hypernymic proposition is a semantic structure in which two concepts, a hyponym and a hypernym, are in a taxonomic relation. In English, there are three major syntactic strategies for encoding such a proposition: with verbs, appositives, or nominal modification. We provide a few examples of these structures culled from our sample.

In configurations involving verbs, the specific concept is most often represented by a noun phrase that is the subject of *be* and the general is represented by its complement.

(1) **Nimodipine** is an isopropyl **calcium channel blocker** which readily crosses the blood–brain barrier.

Verbs other than *be*, such as *remains*, are occasionally seen in this structure.

(2) **Amoxicillin** remains a reliable first-choice **antibiotic** in the treatment of lower respiratory infections.

The appositive structure consists of two noun phrases occurring next to each other. There are variations on how the second noun phrase is marked; it can be set off by commas (the second does not always appear), parentheses, or lexical items such as *including, such as*, and *particularly*.

(3) **Arginine**, a **semiessential amino acid**, has been shown to increase wound collagen accumulation in rodents and humans.
(4) From the time of extubation, patients had access to an **opioid** (**oxycodone**) via a patient-controlled analgesia device.
(5) **Non-steroidal anti-inflammatory drugs** such as **indomethacin** attenuate inflammatory reactions.

Hearst [38] reports on other appositive patterns that encode hypernymic propositions (although she does not address the general interpretation of these propositions). Examples include *works by such authors as Herrick,* *Goldsmith, and Shakespeare*, in which the hypernym precedes and is marked by *such*, while the hyponym follows marked by *as*. She also reports on coordinate structures in which the initial members of the construction are hyponymic to the final member, which is marked by *other*: *temples, treasuries, and other important civic buildings*. We have so far not addressed these patterns, since we did not encounter them in the sample used to develop our system. They could be accommodated without major effort.

In nominal modification, both concepts in a hypernymic proposition may be represented in the same simple noun phrase. In such instances, either the general or the specific may be represented by the head of that noun phrase, while the modifier represents the other argument.

(6) The **anticonvulsant gabapentin** has proven effective for neuropathic pain.
(7) An increase in blood pressure was also seen in patients who were taking adjunctive **antihypertensive medications** prior to withdrawal of omapatrilat.

Based on a sample of approximately 1000 sentences containing hypernymic propositions, the relative frequencies of the syntactic structures we encountered are as follows. About 20% are encoded as arguments of verbs (most frequently *be*); somewhat under 40% appear as appositives (of all types); finally, somewhat over 40% are found as modifier and head in the simple noun phrase. For a more detailed analysis of the distribution of the structures we encountered, see Section 4.

In this study, we have not articulated the semantics of the relationship between the two arguments of what we call the hypernymic proposition. We assume that this relationship is taxonomic, but have not systematically investigated its semantic value regarding either the intent of the author's assertion in the text encountered or the relationships between concepts found in the Metathesaurus. We shall simply refer to the predicate of the hypernymic proposition as ISA, with the assumption that this is a cover term for what may in fact be several semantic values. Brachman [39] offers a number of alternatives for the meaning of ISA, including "subset/superset," "generalization/specialization," and "kind-of." Burgun and Bodenreider [40] and Bodenreider et al. [41] investigate in further detail the semantics of hierarchical relations, with particular emphasis on the UMLS.

Although the focus in this study is on the interaction of syntax and domain knowledge in expressing hypernymic propositions, we make brief note of the discourse function of this phenomenon. Understanding and analyzing the structure of discourse plays an important role in advanced natural language processing [42].

Chafe [43] describes discourse structure as the way in which a speaker (writer) uses syntactic structures to impart information to a listener (reader). An important aspect of this strategy is the distinction between given (or old) information and information that the speaker assumes is being introduced to the listener as new. Hypernymic propositions provide a means of facilitating the flow of information by accommodating this distinction and can be thought of as definitions embedded in a discourse.

Definitions impart new information (the definiens) in terms of old, or already accessible, information (the definiendum). Bodenreider and Burgun [44] describe one type of definition that follows what they call the Aristotelian pattern of genus and differentia, in which the definiendum is in a taxonomic relation with the first part of the expression serving as the definiens. That is, the definition is a hypernymic proposition. The definitional nature of the hypernymic proposition provides a mechanism for serving the same function in a discourse, where the specific concept is the new information and the general is the old.

In MEDLINE citations discussing a specific drug therapy for a particular problem, it is common for a hypernymic proposition to appear early in the abstract, functioning as a definition that provides a context of old information for the new information being introduced, namely the characteristics of the drug in question. For example:

(8) **Mizolastine** provides effective symptom relief in patients suffering from perennial allergic rhinitis: ... [Title of abstract]
(9) **Mizolastine** is a nonsedating **H1 histamine receptor antagonist** with additional antiallergic properties. [First sentence of abstract]

Before discussing the program we have devised for identifying hypernymic propositions in MEDLINE citations, we describe the UMLS knowledge sources that provide the domain knowledge on which our processing depends. We first use semantic types from the Semantic Network to ensure that the two concepts involved are compatible. We then appeal to hierarchical information in the Metathesaurus to determine which concept is more general and which more specific.

2.3. Unified Medical Language System

The Unified Medical Language System (UMLS) project [18] is a long-term National Library of Medicine research and development effort designed to facilitate the retrieval and integration of information from multiple machine-readable biomedical information sources. The UMLS has three components: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon. In addition to supporting information management applications, structured domain knowledge contained in these knowledge sources can be exploited for research in

NLP, such as the effort described here to identify hypernymic predications in MEDLINE citations.

The SPECIALIST Lexicon and associated lexical access tools [45] provide syntactic information about terms in general and medical English. Both simple and multiword lexical entries are included, and each entry has been assigned one or more part-of-speech labels. Spelling variants, inflectional forms, and complement information for verbs further support NLP applications.

The Metathesaurus is a large repository of concepts (nearly 777,000 in the 2002 version) drawn from more than 60 vocabularies, classifications, and coding systems. During compilation, the structure of source terminologies is preserved; however, terms that have equivalent meanings are organized into unique concepts, which form the organizational core of the Metathesaurus. Associative and hierarchical relationships between concepts either come from the source terminologies or are added by editors. In this study, we make extensive use of these relationships in order to identify hypernymic propositions; the two arguments of such a predication must be in a (direct or indirect) hierarchical relationship, loosely defined to include Parent, Child, as well as Broader and Narrower.

It is important to note that due to varying semantics in source vocabularies, many of the relationships we use to support interpretation of hypernymic propositions are not strictly accurate for this purpose. For example, "Tylenol" is related to "Acetaminophen" by the Narrower relation in the Metathesaurus, although something like BRAND_OF would be more correct. In other instances, however, the relationship can be profitably construed as hierarchical. "Aspirin," for example, is in a Broader relationship with "Analgesics," "Salicylates," and "Cyclooxygenase Inhibitors." These limitations notwithstanding, it is our experience (supported by the evaluation of this project) that domain knowledge from the Metathesaurus can provide effective support for natural language processing directed at the interpretation of hypernymic propositions.

Each Metathesaurus concept is also assigned one or more semantic types such as 'Disease or Syndrome' or 'Pharmacologic Substance' that categorize concepts in the biomedical domain. There are 134 semantic types in the 2002 release of the UMLS, and the Semantic Network [46] organizes these into two single-inheritance hierarchies, one for entities and one for events. In addition, associative relations are assigned between semantic types; these semantic propositions represent knowledge that is accepted as being valid in the biomedical domain, such as

(10) 'Body Part, Organ, or Organ Component' HAS_PART 'Cell.'
'Body Location or Region' LOCATION_OF 'Anatomical Abnormality.'
'Pharmacologic Substance' TREATS 'Disease or Syndrome.'

Recent research by McCray et al. [47] aimed at reducing the conceptual complexity of the medical knowledge represented in the Semantic Network has resulted in the development of semantic groups. Subject to principles of semantic validity, parsimony, completeness, exclusivity, naturalness, and utility, such groups organize the 134 semantic types in the Semantic Network into 15 coarse-grained aggregates such as Anatomy, Activities and Behaviors, Living Beings, and Chemicals and Drugs. Based on the distribution of relationships in the Semantic Network, Perl et al. [48–50] have proposed alternative groups to those devised by McCray et al. In this work, we rely on the groups of McCray et al; our methodology can accommodate other configurations, although results will differ.

We use semantic groups to constrain the identification of hypernymic propositions; the Metathesaurus concepts that serve as arguments of such propositions must have semantic types that belong to the same semantic group. (In addition, as noted above, the concepts must be in a hierarchical relationship.) In the version of the program discussed here, we used only the group Chemicals and Drugs. This group consists of 26 semantic types, a few examples of which are 'Pharmacologic Substance,' 'Antibiotic,' 'Biologically Active Substance,' 'Hormone,' 'Enzyme,' 'Vitamin,' 'Steroid,' and 'Immunologic Factor.'

In the next section, we briefly describe how UMLS domain knowledge is used in SemRep, which forms the basis of SemSpec. In the subsequent section describing SemSpec, we discuss and illustrate the specific way that we exploit semantic groups and Metathesaurus hierarchical relationships to support effective semantic interpretation of hypernymic propositions.

2.4. The SemRep system: general semantic interpretation

SemRep is a natural language processing system designed to recover semantic propositions from biomedical text using underspecified syntactic analysis and structured domain knowledge from the UMLS [13–15]. Also see [7] for a related approach (although one that does not use the UMLS). After input and tokenization, text is submitted to an underspecified parser that relies on the syntactic information in the SPECIALIST Lexicon. Part-of-speech ambiguities are resolved with the Xerox Part-of-Speech Tagger [51]. For example, (11) is given the underspecified syntactic analysis in (12).

(11) New fluoroquinolones such as ofloxacin are beneficial in the treatment of chronic obstructive airways disease exacerbation requiring mechanical ventilation.

(12) [[mod(adj(new)),head(noun(fluoroquinolones), metaconc('Fluoroquinolones': [orch, phsu]))],
[prep('such as'),head(noun(ofloxacin),metaconc ('Ofloxacin': [orch,phsu]))],
[aux(are)],
[head(adj(beneficial))],
[prep(in),det(the),head(noun(treatment))],
[prep(of),mod(adj(chronic)),mod(adj(obstructive)), mod(noun(airways)), mod(noun(disease), head(noun(exacerbation),metaconc('Chronic obstructive airways disease exacerbated':[dsyn]))],
[verb(requiring)],
[head(noun(['mechanical ventilation')),punc('.')]].

In this analysis, simple noun phrases are identified and given a partial internal analysis. The head is identified and modifiers occurring to the left of the head other than determiners are marked as modifiers regardless of their part-of-speech label. Prepositional phrases are treated as simple noun phrases whose first element is a preposition. Other syntactic categories, including verbs, auxiliaries, and conjunctions are simply given their part-of-speech label and put into a separate phrase.

Referring expressions such as *fluoroquinolones* in (12) are augmented with Metathesaurus concepts and semantic types. (The semantic types are abbreviated: 'Disease or Syndrome' (dsyn); 'Organic Chemical' (orch); 'Pharmacologic Substance' (phsu).) This domain knowledge is acquired through MetaMap [52,53], a flexible, knowledge-based application that uses the SPECIALIST Lexicon along with rules for morphological variants to determine the best mapping between the text of a noun phrase and a concept in the Metathesaurus.

The interpretation of semantic propositions depends on this underspecified analysis enriched with domain knowledge and is driven by syntactic phenomena that "indicate" semantic predicates, including verbs, prepositions, nominalizations, and the head-modifier relation in simple noun phrases. Rules are used to map syntactic indicators to predicates in the Semantic Network. For example, there is a rule that links the nominalization *treatment* with the predicate TREATS.

Domain restrictions are enforced by a meta-rule stipulating that all semantic propositions identified by SemRep must be sanctioned by a predication in the Semantic Network. This rule ensures that syntactic arguments associated with *treatment* in the analysis of (12) must have been mapped to Metathesaurus concepts with semantic types that match one of the permissible argument configurations for TREATS, such as 'Pharmacologic Substance' and 'Disease or Syndrome'.

Further syntactic constraints on argument identification are controlled by statements expressed in a dependency grammar. For example, the rules for nominalizations state that one possible argument configuration is for the object to be marked by the preposition *of* occurring to the right of the nominalization and that one possible location for the subject is anywhere to the left of the noun phrase containing the nominalization.

During semantic interpretation of the predication on *treatment* in (12), choosing the noun phrase *ofloxacin* (which maps to a concept with semantic type 'Pharmacologic Substance') as the subject and *chronic obstructive airways disease exacerbation* (mapped to a concept with semantic type 'Disease or Syndrome') as the object allows all constraints to be satisfied. The final interpretation is the semantic proposition in (13), where the Metathesaurus concepts are arguments of the predicate from the Semantic Network.

(13) Ofloxacin TREATS Chronic obstructive airways disease exacerbated.

SemRep also addresses noun phrase coordination [54] by taking advantage of semantic types. This processing begins before the interpretation of semantic propositions. On the basis of the underspecified syntax enhanced with domain knowledge, an attempt is made to determine whether each coordinator is conjoining noun phrases or something other than noun phrases. For a coordinator determined to be conjoining noun phrases, the semantic type of the noun phrase immediately to the right of that coordinator is examined. The noun phrase immediately to the left of the coordinator and noun phrases occurring to the left of that noun phrase (and separated from it either by another coordinator or by a comma) are examined to see whether they are semantically consonant. In the current formulation of the coordination algorithm, semantic consonance means that the semantic types are identical.

For example in (14), *inflammatory bowel disease* has been mapped to a concept with semantic type 'Disease or Syndrome'; *allergic rhinitis* and *asthma* also have been mapped to concepts with this semantic type and thus these three noun phrases are considered to be coordinate.

(14) ... a new class of anti-inflammatory drugs that have clinical efficacy in the management of asthma, allergic rhinitis, and inflammatory bowel disease.

During the process of semantic interpretation, if a coordinate noun phrase is found to be an argument of a semantic predicate, then all noun phrases coordinate with that noun phrase must also be arguments of a predication with that predicate. During the semantic processing of (14), for example, once the first predication in (15) has been constructed, the other two are automatically generated by virtue of the coordinate status of *asthma*.

(15) Anti-Inflammatory Agents TREATS Asthma.
Anti-Inflammatory Agents TREATS Allergic rhinitis, NOS.
Anti-Inflammatory Agents TREATS Inflammatory Bowel Diseases.

In order to identify and interpret hypernymic propositions, we have developed SemSpec as a module within SemRep. SemSpec processing depends on underspecified syntactic analysis enhanced with concepts and semantic types and follows the general SemRep framework, including the use of indicator rules to map between syntactic phenomena and semantic predicates, dependency grammar constraint on argument identification, and the notion of domain restrictions on allowable arguments.

3. Methods

3.1. SemSpec: the interpretation of hypernymic propositions

Fig. 1 provides an overview of our approach to the extraction of semantic predications from text and indicates where SemSpec fits within this system. SemSpec takes advantage of the linguistic processing in SemRep by first identifying the syntactic structures that potentially indicate hypernymic propositions (semantic indicator rules), including verbs, appositives, and the modifier head relationship in the simple noun phrase. After potential syntactic arguments have been identified, regardless of the structure in which they were found, they are subjected to uniform semantic constraints based on the UMLS. However, due to the semantic characteristics of the hypernymic proposition being retrieved, this knowledge is exploited differently than it is in SemRep. Rather than using the overt stipulations of the associative predications in the Semantic Network for semantic constraints on argument identification, SemSpec calls on semantic groups from the Semantic Network and hierarchical relationships from the Metathesaurus to constrain the arguments of the hypernymic proposition.

We first discuss the syntactic processing that allows SemSpec to identify the potential arguments in a hypernymic proposition in the three syntactic structures we address. As an example of how SemSpec identifies hypernymic propositions encoded in the simple noun phrase, consider the sentence (16), for which SemRep processing and MetaMap identify the noun phrase in (17).

(16) Caffeine increases cortical arousal by serving as an antagonist to the **[inhibitory neurotransmitter adenosine]**.
(17) [det(the), mod(adj(inhibitory),metaconc('inhibitors': chvf)), mod(noun(neurotransmitter),metaconc('Neurotransmitters':nsba)), head(noun(adenosine),metaconc('Adenosine':bacs))].

SemSpec examines each simple noun phrase for a modifier immediately to the left of the head. If the semantic types assigned to the Metathesaurus concepts for both the modifier and the head belong to the same semantic group, the Metathesaurus is consulted to determine whether the corresponding concepts are in a hierarchical relationship. In this example, the concept of the modifier has semantic type 'Neuroreactive Substance or Biogenic Amine' (nsba), and the head concept has 'Biologically Active Substance' (bacs); both are members of the semantic group Chemicals and Drugs. Further, it is determined that the concepts "Neurotransmitters" and "Adenosine" are in a hierarchical relation in the Metathesaurus and that the former is an ancestor of the latter. Based on these syntactic and semantic constraints, SemSpec interprets the noun phrase (17) as the proposition (18).

(18) Adenosine ISA Neurotransmitters.

Appositive structures comprise two contiguous noun phrases, the second of which may be set off simply by commas or may be marked by overt cues such as parentheses or lexical items such as *including* and *such as*.

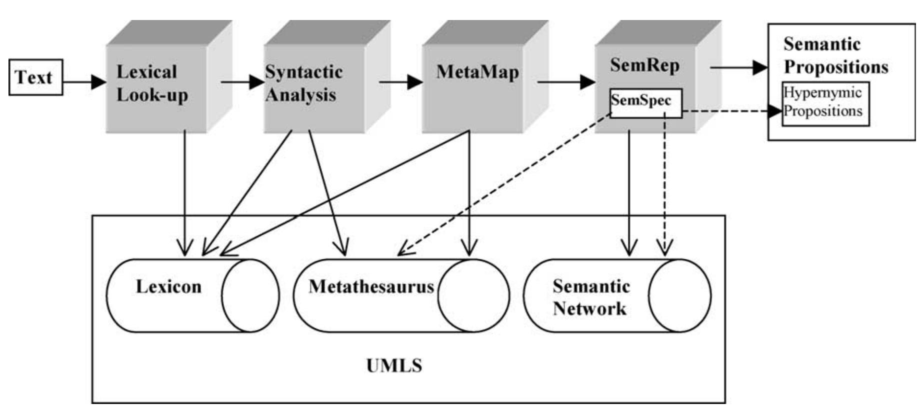

Fig. 1. General overview of semantic processing. SemSpec, a module within SemRep, interprets hypernymic propositions only.

(19) New **fluoroquinolones** such as **ofloxacin** are beneficial in the treatment of COPD.

In processing (19), in which the second phrase is unambiguously introduced by *such as*, the relevant syntactic analysis is

(20) [mod(new), head(noun(fluoroquinolones),metaconc('Fluoroquinolones':phsu))]
[prep('such as'), head(noun(ofloxacin),metaconc('Ofloxacin':phsu))].

After affirming that the semantic types in these two noun phrases are in the same semantic group, it is determined from the Metathesaurus that "Fluoroquinolones" is an ancestor of the "Ofloxacin" and the following predication is generated.

(21) Ofloxacin ISA Fluoroquinolones.

Out of context, appositives marked only by commas are ambiguous with items in a series coordination structure, as for example in

(22) ... tricyclic antidepressants, monoamine oxidase inhibitors, and antiepileptic agents...

In (22), the two noun phrases *tricyclic antidepressants* and *monoamine oxidase inhibitors* occurring together separated by a comma could be analyzed as an appositival structure asserting a hierarchical relation (if the entire structure of the sentence is not considered). In fact, the concept "Monoamine Oxidase Inhibitors" is (wrongly) in a hierarchical relationship with "Antidepressive Agents, Tricyclic" in the Metathesaurus. Yet, the intent of the author in (22) is that these two concepts be considered as coordinate and not in apposition. SemSpec uses SemRep's coordination facility to check whether two noun phrases separated by a comma have already been determined to be coordinate. If so, they cannot be analyzed as being in an appositive relation, even when the relevant concepts are in a hierarchical relationship, as in (22). In order for SemSpec to interpret a hypernymic proposition, all syntactic and semantic conditions must be met. In cases such as these, the syntactic requirements are not met.

The sentence in (23) contains an instance of an appositive structure marked by commas that does not involve coordination. The noun phrases *clonidine* and *an a-2 adrenergic agonist* were determined by SemRep not to be coordinated, and thus SemSpec processes them as a hypernymic proposition and retrieves the proposition in (24).

(23) Clinical observations suggest that **clonidine**, an **a-2 adrenergic agonist**, may improve diabetic gastropathy symptoms.
(24) Clonidine ISA Adrenergic Agonists.

SemSpec faces a particular challenge when interpreting hypernymic propositions based on arguments of verbs. In addition to semantic generalizations, the dependency grammar rules inherited from SemRep are subject to general constraints that prevent crossing dependency lines and disallow the reuse of arguments without license (coordination, for example). Further, the subject of a verb must occur to its left and the object (or complement) may appear on the right. Although these rules are generally effective, the underspecified categorial analysis to which they apply does not provide detailed structural cues in support of argument identification [55]. For identifying hypernymic predications based on verbs, we augment these rules with an intervention constraint.

In order for a verb to encode a hypernymic proposition, it must occur between its potential arguments, and the number of phrases (as determined by the underspecified analysis) intervening between the arguments can be no more than four, including the phrase containing the verb. This distance measure was chosen on the basis of experimentation with a training set (described below in Section 5).

In our study, if a hypernymic proposition is encoded by a verb, it is a form of *be* in the vast majority of cases, and we thus limit our discussion to this verb. (The analysis does not distinguish between *be* as an independent verb and as an auxiliary.) For example, the sentence fragment (25) is given the underspecified syntactic analysis shown schematically in (26).

(25) **Amisulpride** is to date the only atypical **antipsychotic** ...
(26) [Amisulpride] [is] [to date] [the only atypical antipsychotic].

The noun phrases *amisulpride* and *the only atypical antipsychotic* are separated by two intervening phrases (*is* and *to date*), and thus are correctly considered by SemSpec to be potential arguments of *is* in this sentence. Further semantic processing permits the following hypernymic proposition to be constructed.

(27) AMISULPRIDE ISA Antipsychotic Agents.

The following example illustrates the effective application of this constraint to disallow a relationship that is not asserted in the text.

(28) [The use] [of **desmopressin**] [in patients] [with primary nocturnal enuresis] [**is**] [based] [on the hypothesis] [of a nocturnal lack] [of endogenous **arginine vasopressin**].

Although *is* occurs between the noun phrases *of desmopressin* and *of endogenous arginine vasopressin* in (28), the number of intervening phrases between these potential arguments is greater than four; SemSpec thus does not interpret the highlighted phrases as being arguments of *is* in this sentence. It is important to note that "Desmopressin" appears in the Metathesaurus as a

I. Verbs INDICATES 'ISA'
II. Appositive Structure INDICATES 'ISA'
III. [Modifier, Head]$_{NP}$ INDICATES 'ISA'

Appositive Structure → Parenthesis | Comma | Lexical
Parenthesis → NP (NP)
Coma → NP , NP ,
Lexical → NP Cues NP

Verbs: *be, remain*, etc.
Cues: *such as, including, particularly*, etc.

Fig. 2. Informal representation of semantic indicator rules for hypernymic propositions. (*Note.* the parentheses and commas in these rules are leaf nodes and not meta-symbols.)

descendant of "Arginine Vasopressin." Without the imposition of the intervention constraint, SemSpec would retrieve a hypernymic proposition that has face-value validity, but which is not asserted in this sentence. Fig. 2 summarizes the semantic indicator rules we have so far used to identify hypernymic propositions in SemSpec.

Above, we noted how SemSpec exploits SemRep coordination processing to eliminate incorrect interpretations of hypernymic propositions involving appositives. The ability of SemRep to identify coordinate noun phrases is also used by SemSpec to identify coordinate arguments of hypernymic propositions, as in

(29) **Captopril, enalapril**, and **lisinopril** are **angiotensin-converting enzyme inhibitors** widely prescribed for hypertension.

Prior to SemSpec processing, SemRep identifies *captopril, enalapril*, and *lisinopril* as being coordinate in this sentence. SemSpec then determines that the concept "Lisinopril" is in a hierarchical relation with "Angiotensin-Converting Enzyme Inhibitors" and applies the SemRep rule that stipulates that when a noun phrase is analyzed as an argument of a predication, all noun phrases coordinate with that noun phrase must be arguments of similar predications. The application of this rule during the semantic interpretation of (29) produces the following predications.

(30) Captopril ISA Angiotensin-Converting Enzyme Inhibitors.
Enalapril ISA Angiotensin-Converting Enzyme Inhibitors.
Lisinopril ISA Angiotensin-Converting Enzyme Inhibitors.

3.2. Training and evaluation

We used two samples of MEDLINE citations for training and testing SemSpec. The training set was based on 6000 MEDLINE citations (titles and abstracts) from the year 2001 retrieved using the Haynes methodological filter [1] for treatment, without content terms. Sentences in these citations were subjected to a second filter ensuring that at least two concepts having a semantic type from the semantic group Chemicals and Drugs were present in each sentence. Three hundred and forty sentences meeting the criteria of both filters were selected for developing the system, and 175 hypernymic propositions were identified by hand (by MF) in these sentences.

A second sample serving as a test set was compiled from MEDLINE citations disjoint from those used for training. Approximately 3000 citations were retrieved using the same Haynes methodological filter and limited by date from January through August, 2002. In processing these citations, SemSpec identified 830 hypernymic propositions, which were then assessed by a professional indexer and a clinician (415 for each judge), neither of whom had worked on the project.

In the test set, the judges were asked to evaluate the propositions identified by SemSpec, but not to identify propositions asserted in the text that were missed by the system. Therefore, we were able to identify true and false positives, but not false negatives in this sample. When comparing the hypernymic propositions produced by SemSpec against those marked in the training set, we identified false negatives in addition to true and false positives. An exact match of the entire predication was required for a SemSpec predication to be considered correct.

4. Results

The distribution of syntactic structures encoding the correct predications in both the training and test sets is given in Table 1. We list appositive structures in separate entries according to the marking of the second noun phrase of the construction: parentheses, comma, and other appositive cues (*such as, including*, etc.) in this table. In the text we have encountered, *remains* is the only verb other than *be* that encodes these propositions. Both samples are too small to be representative of the

Table 1
Distribution of syntactic patterns for correct hypernymic propositions in the evaluated samples

Syntactic pattern	Training set		Test set	
	Count	%	Count	%
Modifier head	34	19.4	277	40.1
Verb *be*	69	39.4	148	21.4
Parentheses	45	25.7	158	22.9
Comma	12	6.9	82	11.9
Other appositive cues	13	7.4	23	3.3
Other verbs	2	1.1	2	0.3
Total	175	100	690	100

true distribution of the syntactic patterns encoding hypernymic propositions in biomedical scientific text. Differences between the frequencies in the two samples probably reflect this fact. Regarding evaluation, as noted above, the judges assessed 830 hypernymic propositions generated by SemSpec from the test set sample. Six hundred and ninety of these were considered correct, while 140 were marked as false positives, resulting in precision of 83%. Since we were not able to determine false negatives in the test set, we provide an estimation of recall from the training sample. Out of 175 hypernymic propositions marked in this sample, SemSpec correctly identified 121 and missed 54, giving a recall figure of 69%. Although this result is likely to be an overestimation of the performance of the system, error analysis of the false negatives provides valuable guidance for future enhancements.

5. Discussion

In the following discussion, we note the quantitative results of error analysis on false positives from the test set and false negatives form the training sample, providing illustrative examples as well as strategies for addressing both types of errors.

5.1. False positives

Of the 140 false positives encountered in the test set, almost all could be placed in four categories: Mistakes due to misidentification of arguments of *be* (40), coordination (41), word sense ambiguity (48), and Metathesaurus relations (10).

5.1.1. Misidentification of arguments of be

As noted above, the underspecified syntactic analysis is not adequate by itself to support the identification of arguments of verbs. We also noted that the analysis proceeds on the assumption that the semantic constraints based on UMLS domain knowledge would provide support for argument identification at an acceptable level of accuracy. The results of our evaluation bear out that supposition; however, a number of errors remain.

One reason for misidentifying arguments of *be* is that two concepts separated by a form of *be* in a sentence may not be syntactic arguments of that verb, yet may be related hierarchically in the Metathesaurus, as in

(31) … several **[cephalosporins]** [were] [monitored] [in a 52-year-old man] [after a selective systemic anaphylaxis attributable] [to **cefuroxime]**, …

Since there is a form of *be* occurring between the concepts “Cephalosporins” and “Cefuroxime” in this sentence, and because the number of phrases (including *were*) intervening between these concepts is four, SemSpec retrieves the predication (32). Although this predication is not incorrect from the point of view of the domain, it is not asserted in this sentence, and hence is an error. Errors of this sort are not necessarily eliminated by domain knowledge.

(32) Cefuroxime ISA Cephalosporins.

One possible way to improve the accuracy of argument identification based on underspecified syntax might be to reduce the number of phrases that are allowed to intervene between arguments of a verb. However, noun phrases occurring in close proximity to a verb are often not in fact its arguments, as in (33), where the noun phrase whose head is *anticonvulsants* is not an argument of *is*, but rather of the verb form *is combined*.

(33) Adverse effects are infrequent when the drug is used alone, but become more frequent when **lamotrigine** [is] [combined] [with other **anticonvulsants**].

Although allowing four intervening phrases does not always provide correct results, it appears to be optimal. Fig. 3 illustrates how the performance measures for the identification of arguments of *be* varied in the training set by allowing the distance between arguments of *be* to range from one to six intervening phrases.

There are other constraints that we could impose in identifying arguments of *be*, given the resources of the underspecified syntactic analysis. As noted earlier, the underspecified syntactic analysis does not identify auxiliaries. We could approximate such identification by considering the item immediately to the right of a form of *be*. If it is a participle (either present or past) we could analyze that form of *be* as an auxiliary and prevent it from encoding a hypernymic proposition. For example, the presence of *combined* immediately to the right of *is* in (33) disallows it from encoding a hypernymic proposition in that sentence.

It would also be possible to exploit the order of the two arguments in a hypernymic proposition. Currently we do not stipulate the order of the hypernym and the

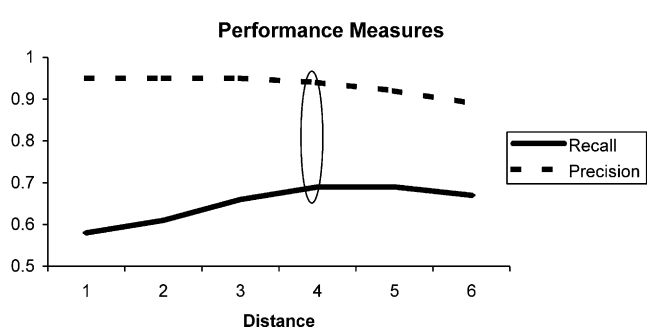

Fig. 3. Performance measures as a function of the distance between arguments of *be*. The circle across the lines represents the best level of performance in the training set.

hyponym in the syntax. In appositives, the syntax does not specify which precedes, and so the hierarchical structure in the Metathesaurus is relied on to specify the order of the arguments in the semantic proposition. However, the hyponym normally comes first in hypernymic predications encoded by *be*. A constraint stipulating this order would prevent the generation of the false positive in (31) above, since the noun phrase encoding the hypernym (*cephalosporins*) precedes the noun phrase encoding the hyponym (*cefuroxime*).

5.1.2. Coordination

The coordination processing used by SemSpec led to two classes of false positive errors. As introduced earlier, SemSpec relies on a constraint stating that if two noun phrases are coordinate, they cannot be interpreted as arguments in a hypernymic proposition (or any predication). This constraint is as effective as the algorithm for identifying coordinate noun phrases, which has deficiencies. Comparative structures are similar to coordinate noun phrases, and comparatives are not yet handled adequately by the SemRep coordination algorithm on which SemSpec depends. For example, in (34), *amisulpride* and *typical antipsychotics* are in a comparative relationship.

(34) Regarding positive symptoms, **amisulpride** was as effective as **typical antipsychotics**, ...

If that relationship had been detected by the program, these noun phrases would not have been allowed to be interpreted as arguments of the intervening *was*, and the false positive predication "AMISULPRIDE ISA Antipsychotic Agents" would not have been generated. Often comparative noun phrases are cued by formulas such as "*more* ADJ *than* " or "*as* ADJ *as*" and can be recognized on the basis of the underspecified syntactic analysis.

The way in which SemRep (and hence SemSpec) handles the consequences of coordinate noun phrases sometimes led to a second class of false positive. We stated above that when two noun phrases have been determined to be coordinate, if one of them is analyzed as an argument in a hypernymic proposition, then the other one must also participate in a hypernymic proposition having an identical predicate and second argument.

Although this rule has felicitous consequences (without a check in the Metathesaurus) when the hypernymic proposition is syntactically encoded by the verb *be*, it can lead to error when the predication is based on an appositive, as in (35). The predications (36) and (37) are going to be generated.

(35) The combination of **valsartan** and **hydrochlorothiazide** (a **thiazide diuretic**), administered once daily, has been evaluated in the treatment of patients with hypertension.

(36) Hydrochlorothiazide ISA Diuretics, Thiazide.

(37) Valsartan ISA Diuretics, Thiazide.

Although the noun phrases *valsartan* and *hydrochlorothiazide* are coordinate in this sentence, the author only asserts a hierarchical relationship between "Hydrochlorothiazide" and "Diuretics, Thiazide" and not between "Valsartan" and "Diuretics, Thiazide." In fact, valsartan is an angiotensin-converting enzyme inhibitor and not a diuretic. This problem can be resolved by ensuring that the arguments of all hypernymic proposition are checked in the Metathesaurus before the predication is constructed, even if coordinate noun phrases are involved.

5.1.3. Word sense ambiguity

The Metathesaurus represents many senses of ambiguous English words, and word sense ambiguity underlies nearly a third of the false positives generated. Although such ambiguity is a problem in any NLP application, in this project, branded drug names being ambiguous with non-drug names pose a particular challenge. For example, "Relief" is a Metathesaurus synonym for "Relief brand of phenylephrine." This causes SemSpec to generate a false positive hypernymic proposition when the noun phrase *of relief medication* is encountered in (38), for example.

(38) Accelerated return to normal activities, and reduced interference with sleep, consumption of **relief medication** and incidence of complications leading to antibacterial use were also observed with zanamivir.

When MetaMap encounters this noun phrase it retrieves two concepts from the Metathesaurus for *relief:* "Feeling relief" and "Relief brand of phenylephrine." The head of this noun phrase, *medication*, maps to "Pharmaceutical Preparations" (with semantic type 'Pharmacologic Substance'). Since this noun phrase is analyzed as a modifier followed by a head, and since one of the concepts referred to by the modifier has semantic type 'Pharmacologic Substance', SemSpec incorrectly generates the hypernymic proposition asserting that "Relief brand of phenylephrine" is a hyponym of "Pharmaceutical Preparations," which is true, but was not the intent of the author of (38).

A second example of a false positive due to word sense ambiguity illustrates the interaction of this phenomenon with inflectional variation. During normal MetaMap processing, inflectional variation is normalized. For example, *test, tests, tested, and testing* are all treated as the base form *test*. This permits robust matching between text tokens and Metathesaurus forms, without interference from noun plurals and verb tense marking. However, in the face of word sense ambiguity, this can lead to errors, as in

(39) The **tested drug** was allowed to retain for 1 min.

In this sentence, the modifier *tested* in the noun phrase *the tested drug* is normalized by MetaMap to *test*. This token maps to the Metathesaurus concept "TEST," which is a synonym for a particular form of Ethanesulfonic acid; and *drug* maps to "Pharmaceutical Preparations." These concepts then allow SemSpec to interpret this noun phrase as "TEST ISA Pharmaceutical Preparations." We are exploring several approaches to resolving word sense ambiguity in order to address this class of errors.

5.1.4. Metathesaurus relationships

It is rarely the case that false positive errors are due exclusively to Metathesaurus relationships; usually incorrect mapping between text and concepts as well as syntactic processing is also involved. For example, consider the following example:

(40) A total of 1471 children with non-severe pneumonia were randomly assigned to 25 mg/kg amoxicillin or 4 mg/kg **trimethoprim** plus 20 mg/kg **sulphamethoxazole** (**co-trimoxazole**).

In (40), due to the inclusion of dosage information, the syntactic analysis does not support mapping *4 mg/kg trimethoprim plus 20 mg/kg sulphamethoxazole* to the correct concept, "Trimethoprim-Sulphamethoxazole Combination." If this had been done, SemSpec would have established a relationship between this concept and "Co-Trimoxazole." Instead, the text was mapped to two concepts, "Trimethoprim" and "Sulphamethoxazole." Appositive processing then led to a check in the Metathesaurus for a relationship between "Co-Trimoxazole" and "Sulphamethoxazole," which was found. This relationship, however, is Broader and thus not strictly hierarchical. The false positive error generated while processing (40) illustrates inherent limitations in using thesaurus relationships as taxonomic relationships.

5.2. False negatives

We used the training set sample to analyze false negatives. The 54 errors of this type fall into four categories: mistakes in interpreting the modifier head relation in simple noun phrases (17), errors due to missing Metathesaurus hierarchical relations (14) and Metathesaurus coverage (9), and other syntactic problems (14), half of which are due to coordination processing.

5.2.1. Modifier head relation in simple noun phrases

The etiology of a number of false negatives is illustrated by an analysis of the fragment (41) for which SemSpec retrieves the predications in (42).

(41) **Fluoxetine** is the only **antidepressant medication** that ...

(42) Fluoxetine ISA Pharmaceutical Preparations.
Antidepressive Agents ISA Pharmaceutical Preparations.

The first predication is derived from the text *Fluoxetine is . . .medication* and the second is the interpretation of the noun phrase *antidepressant medication*. Both are correct, but we would further like to identify the predication (43) from (41).

(43) Fluoxetine ISA Antidepressive Agents.

In order to do this we would need to introduce a meta-rule that could derive this predication from the two predications in (42) under the syntactic conditions that obtain in (41).

This problem is to a large extent resolved by representation in the UMLS. Classes of pharmacologic substances, for example antidepressant medications, antiviral agents, or anti-schizophrenic drugs, are often represented directly as Metathesaurus terms. Although the term "Antidepressant Medication" does not appear, "Antidepressive Agents," "Antidepressant Drugs," and "Antidepressants" occur as synonyms. When text such as that in (44) is encountered, SemSpec is able to retrieve the predication (45), based on the Metathesaurus synonyms "Antidepressants" and "Antidepressive Agents."

(44) **Fluoxetine** is the only **antidepressant** that...

(45) Fluoxetine ISA Antidepressive Agents.

A related problem is encountered in processing (46), for which no predication is retrieved. However, in this case, *analog* and *vitamin D* do not appear in a hierarchical relationship, nor do *Cacipotriol* and *analog*.

(46) **Calcipotriol** is a **vitamin D analog**...

An acronym in the middle of a noun phrase impedes SemSpec processing. For example, the (*ACE*) in *angiotensin-converting enzyme* (*ACE*) *inhibitors* interferes with MetaMap's ability to map this phrase to the Metathesaurus concept "Angiotensin-Converting Enzyme Inhibitors." We note that acronyms appearing at the end of a complete concept do not interfere with MetaMap, however. The text *platelet-derived growth factor (PDGF)* is correctly mapped to the concept "Platelet-Derived Growth Factor." Several recent works address acronyms in medical text [56–58], and MetaMap is also being enhanced to deal with this phenomenon.

5.2.2. Coordination

A number of false negative errors are related to the coordination processing used by SemSpec. Some of these are due to the fact that the criterion for semantic consonance that must obtain among the conjuncts of a coordinate structure is too stringent. For example, from the text in (47), SemRep does not analyze *hormone* and *antioxidant* as being coordinate due to the fact that the former has the semantic type 'Hormone' and the latter has 'Pharmacologic Substance.'

(47) **Melatonin** is a **hormone** and **antioxidant** produced by the pineal gland ...

Since the SemRep coordination processing did not coordinate these noun phrases, SemSpec missed the predication "Melatonin ISA Antioxidants." The SemRep coordination algorithm was devised before the availability of the semantic groups in the UMLS Semantic Network and needs to be revised to take advantage of that facility.

Another problem involving coordination is seen in the following sentence.

(48) All tests were performed before and after administration of one of five different **antihistamines** (**cetirizine**, **loratadine**, **ebastine**, **fexofenadine**, **mizolastine**).

The coordination algorithm requires a conjunction to appear before the last element of a coordinated series of noun phrase. Although the elements enclosed in parentheses in (48) are intended to be coordinate, a conjunction does not appear in the list, and thus SemSpec only retrieves the predication "Cetirizine ISA antihistamines." The appearance of a series of elements that are intended to be coordinate, but without the appearance of a conjunction as in (48), is not common in scientific text. Dispensing with the requirement for a conjunction in the coordination algorithm would no doubt lead to more problems than it would solve.

A final problem involving coordination is illustrated by the terms in bold in the following sentence.

(49) The "atypical" profile of the new **antipsychotics**, **clozapine**, **olanzapine**, **quetiapine**, **and risperidone** has been linked to combined antagonism of serotonin 2 and dopamine 2 receptors.

The coordination algorithm incorrectly analyzed all the elements in bold in (49) as being coordinate, since the term to the right of the conjunction and all the contiguous terms to the left have consonant semantic types, and all the terms to the left are separated only by a comma. The correct analysis of this series is that *clozapine* is the first member of the coordinate structure of which *risperidone* is the last member. The term *antipsychotics* is not a member of this structure, but, rather, is in an appositive relation with the coordinate terms.

The coordination algorithm was formulated without regard to hierarchical relations. It might be profitable to revise the algorithm to disallow the left-most element of a coordinate series from being in a hierarchical relationship with the next member of the coordination to its right. Such a provision would not permit *antipsychotics* to be analyzed as a member of the series coordination in (49), which would allow it to be in apposition to all the coordinate terms. This in turn would form the basis for retrieving missed hierarchical relations in this sentence.

5.2.3. Metathesaurus coverage

The UMLS has broad coverage of the biomedical domain, and thus only a few false negative errors were due to concepts in the text not being found in the Metathesaurus. Examples can be seen in (50) and (51), where the hypernymic terms, *noradrenaline reuptake inhibitor* and *anti-fibrotic agent* do not appear in the Metathesaurus.

(50) The clinical profile of **reboxetine**, a selective **noradrenaline reuptake inhibitor**, was compared with...

(51) **Colchicine** is an **anti-fibrotic agent**.

Other, more prevalent, false negatives were due to relations not present in the Metathesaurus. In some instances, concepts share a common ancestor, but are not in a direct descent relationship. In (52) through (55), we provide some examples of concepts that were asserted in text as being in a hierarchical relationship but did not appear in such a relationship in the Metathesaurus.

(52) There has been much interest in **lidocaine**, **a sodium channel blocker**, used clinically ...

(53) Data from experimental studies indicated that **antioxidants**, e.g., **acetylcysteine**, may prevent radiocontrast-induced nephropathy.

(54) Dexketoprofen is strongly bound to **plasma proteins**, such as **albumin**.

(55) This study examined whether **kava**, the herbal **anxiolytic**, produces improvement in anxiety disorder.

The concepts "Lidocaine" and "Sodium Channel Blockers" occur in the Metathesaurus, but are not in a relationship other than both being descendants of "Cardiovascular Agents." "Antioxidants" and "Acetylcysteine" have a common parent, "Chemical Actions." "Plasma Proteins," and "Albumin" have the common ancestor "Proteins" but "Albumin" is not a child of "Plasma Proteins." "Kava Preparation" and "Anti-Anxiety Agents" do not appear in any kind of relationship.

5.3. Limitations

Our preliminary evaluation of SemSpec has several limitations. First, we only evaluated the system on one semantic group and we further restricted the sample by applying a filter that was more likely to retrieve citations containing concepts from the semantic group Chemicals and Drugs. We have expanded the system to other semantic groups but have not yet performed a formal evaluation on this data.

Our reliance on training data for calculating recall diminishes the reliability of this metric as an indicator of effectiveness. However, error analysis of the false negatives identified in the training sample serves as a valuable guide for improving SemSpec. The test set is independent from the training set; unfortunately, we were unable to measure recall in this sample. This limitation in our evaluation method was determined by the difficulty in finding experts not involved with the project to mark up text with semantic predications. A third limitation of this study is that we used only two expert raters to assess the test sample. It has been noted that inter-rater variation [59] has an effect on evaluation reliability. In future evaluations, we would like to use more judges and measure inter-rater variation.

5.4. Future work

We intend to use Semspec to improve SemRep's performance in semantic interpretation generally. The underspecified approach sometimes produces results that are not wrong, but are not as precise as could be achieved with a more complete analysis. SemRep's limitations can be seen particularly in relativizing structures. For example from (56), SemRep is able to extract (57), involving the more general term in a hypernymic proposition.

(56) This study demonstrates that **netilmicin** is a safe and effective **antibiotic** that can be used as a first choice treatment of acute **bacterial conjunctivitis**.
(57) Antibiotics TREATS Conjunctivitis, Bacterial.

However, it would be more accurate to construct a proposition asserting that netilmicin treats acute bacterial conjunctivitis. Toward this goal, SemSpec is able to produce (58), connecting the general term with its more specific partner.

(58) Netilmicin ISA Antibiotics.

We could exploit SemSpec output by devising special rules to determine the more specific subject of TREATS in sentences exhibiting the structure seen in (56). If we are able to match the hypernym concept of the hypernymic proposition with the subject of the TREATS predication, we can then create a third predication following the schema given informally in (59). Based on this, the predication in (60) can be generated in order to more accurately represent the semantic interpretation of (57).

(59) <Hyponym> TREATS(SPEC) <Object of TREATS predication>
(60) Netilmicin TREATS(SPEC) Conjunctivitis, Bacterial.

In addition, we intend to address the problems discussed in the failure analysis to improve performance. When the system has been enhanced, we will investigate its use for extracting hypernymic propositions outside the MEDLINE database.

The National Library of Medicine's MEDLINE*plus* facility contains links to a medical encyclopedia that has definitions for thousands of concepts, including diseases, procedures, medications, and medical diagnostic tests. These are presented in definitional sections and are in free-text format. One interesting application would be to parse the definitions and extract hypernyms and hyponyms (see [60] for a related approach). These might be useful for enhancing retrieval and categorization of Web pages in the encyclopedia section of MEDLINE*plus*.

As an example consider the following definition from the medical encyclopedia.

(61) **Cholangiocarcinoma** is a **malignant (cancerous) growth** in one of the ducts that carries bile from the liver to the small intestine.

The hypernymic predication in (62) was retrieved from (61) after a slight modification to SemSpec to include the semantic group Disorders.

(62) Cholangiocarcinoma ISA Malignant Neoplasms.

Although our approach so far has been to use the Metathesaurus to support the interpretation of hypernymic propositions, we could take the opposite direction and use patterns found in the research literature to audit hierarchical relationships in the Metathesaurus. This could be used to validate relationships or add relationships not currently represented. One third of the false negatives encountered while evaluating SemSpec are due to potential hierarchical relationships not represented in the Metathesaurus.

6. Conclusion

We have presented a methodology for investigating the interaction of domain knowledge and linguistic structure, concentrating on the interpretation of hypernymic propositions in MEDLINE citations. After discussing the linguistic structure of this phenomenon, we described the underspecified syntactic processing and UMLS domain knowledge we exploit in our system. Crucial information is provided by semantic groups from the Semantic Network and hierarchical relationships from the Metathesaurus. The results of a preliminary evaluation are encouraging and error analysis provides a guide for improvements. The methodology described can make a contribution to improvements in high quality natural language processing in the biomedical domain, and has the potential to support a range of applications, including information retrieval and extraction as well as ontology engineering.

Acknowledgments

We gratefully acknowledge Alan Aronson, Olivier Bodenreider, Susanne Humphrey, Halil Kilicoglu, Alexa

McCray, Charles Sneiderman, and Songmao Zhang for their contributions to this project. The second 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 US Department of Energy and the National Library of Medicine.

References

[1] Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994;1(6):447–58.

[2] Mendonca EA, Johnson SB, Seol YH, Cimino JJ. Analyzing the semantics of patient data to rank records of literature retrieval. In: Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain of the Association of Computational Linguistics; 2002. p. 69–75.

[3] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. Proc Int Conf Intell Syst Mol Biol 1999:77–86.

[4] Rosario B, Hearst M. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In: Proceedings of Conference on Empirical Methods in Natural Language Processing; June 2001. p. 82–90.

[5] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995;122(9):681–8.

[6] Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7(6):593–604.

[7] Grishman R, Huttunen S, Yangarber R. Information extraction for enhanced access to disease outbreak reports. J Biomed Inform 2002;35(4):236–46.

[8] Hahn U, Romacker M, Schulz S. MEDSYNDIKATE—a natural language system for the extraction of medical information from findings reports. Int J Med Inf 2002;67(1–3):63–74.

[9] 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):S74–82.

[10] Rosario B, Hearst M, Fillmore C. The descent of hierarchy, and selection in relational semantics. In: Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Association for Computational Linguistics; 2002. p. 247–54.

[11] Spyns P. Natural language processing in medicine: an overview. Methods Inf Med 1996;35(4–5):285–301.

[12] Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999;74(8):890–5.

[13] Rindflesch TC, Bean CA, Sneiderman CA. Argument identification for arterial branching predications asserted in cardiac catheterization reports. Proc AMIA Symp 2000:704–8.

[14] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 2000:517–28.

[15] Rindflesch TC, Rajan J, Hunter L. Extracting molecular binding relationships from biomedical text. In: Proceedings of the Sixth Applied Natural Language Processing Conference, Association for Computational Linguistics; 2000. p. 188–95.

[16] Srinivasan P, Rindflesch T. Exploring text mining from MEDLINE. Proc AMIA Symp 2002:722–6.

[17] Baud RH, Lovis C, Rassinoux AM, Scherrer JR. Alternative ways for knowledge collection, indexing and robust language retrieval. Methods Inf Med 1998;37(4–5):315–26.

[18] Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The unified medical language system: an informatics research collaboration. J Am Med Inform Assoc 1998;5(1):1–11.

[19] Amaral MB, Roberts A, Rector AL. NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logistic constructs. Proc AMIA Symp 2000:76–80.

[20] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–74.

[21] Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp 2000:270–4.

[22] Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994;1(2):142–60.

[23] Friedman C, Liu H, Shagina L, Johnson S, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Proc AMIA Symp 2001:189–93.

[24] Knirsch CA, Jain NL, Pablos-Mendez A, Friedman C, Hripcsak G. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. Infect Control Hosp Epidemiol 1998;19(2):94–100.

[25] Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res 2000;33(1):1–10.

[26] Johnson SB, Aguirre A, Peng P, Cimino J. Interpreting natural language queries using the UMLS. Proc Annu Symp Comput Appl Med Care 1993:294–8.

[27] Rassinoux AM, Wagner JC, Lovis C, Baud RH, Rector A, Scherrer JR. Analysis of medical texts based on a sound medical model. Proc Annu Symp Comput Appl Med Care 1995:27–31.

[28] Rector AL, Nowlan WA. The GALEN project. Comput Methods Programs Biomed 1994;45(1–2):75–8.

[29] Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff S. A natural language understanding system combining syntactic and semantic techniques. Proc Annu Symp Comput Appl Med Care 1994:247–51.

[30] Fiszman M, Haug PJ. Using medical language processing to support real-time evaluation of pneumonia guidelines. Proc AMIA Symp 2000:235–9.

[31] Gundersen ML, Haug PJ, Pryor TA, van Bree R, Koehler S, Bauer K, Clemons B. Development and evaluation of a computerized admission diagnoses encoding system. Comput Biomed Res 1996;29(5):351–72.

[32] Christensen L, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. In: Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Association for Computational Linguistics; 2002. p. 29–36.

[33] Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, Boisvieux JF. A multi-lingual architecture for building a normalised conceptual representation from medical language. Proc Annu Symp Comput Appl Med Care 1995:357–61.

[34] Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, Boisvieux JF. Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. Proc AMIA Annu Fall Symp 1997:590–4.

[35] Hahn U, Romacker M, Schulz S. How knowledge drives understanding—matching medical ontologies with the needs of medical language processing. Artif Intell Med 1999;15(1):25–51.

[36] Hahn U, Romacker M, Schulz S. MEDSYNDIKATE—design considerations for an ontology-based medical text understanding system. Proc AMIA Symp 2000:330–4.

[37] Romacker M, Schulz S, Hahn U. Streamlining semantic interpretation for medical narratives. Proc AMIA Symp 1999:925–9.

[38] Hearst, MA. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING); 1992. p. 539–45.

[39] Brachman RJ. 1983. What IS-A is and isn't: an analysis of taxonomic links in semantic networks. Computer 1983;16(10):30–6.

[40] Burgun A, Bodenreider O. Aspects of the taxonomic relation in the biomedical domain. In: Collected papers from the Second International Conference "Formal Ontology in Information System"; 2001 p. 222–33.

[41] Bodenreider O, Burgun A, Rindflesch TC. Lexically suggested hyponymic relations among medical terms and their representation in the UMLS. In: Proceedings of the Conference on Terminology and Artificial Intelligence; 2001. p. 11–21.

[42] Hahn U, Romacker M, Schulz S. Discourse structures in medical reports—watch out! The generation of referentially coherent and valid text knowledge bases in the MEDSYNDIKATE system. Int J Med Inf 1999;53(1):1–28.

[43] Chafe WL. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In: Li CN, editor. Subject and topic. New York: Academic Press; 1975. p. 25–56.

[44] Bodenreider O, Burgun A. Characterizing the definitions of anatomical concepts in WordNet and specialized sources. In: Proceedings of the First Global WordNet Conference; 2002. p. 223–30.

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

[46] McCray AT. Representing biomedical knowledge in the UMLS Semantic Network. High-performance medical libraries: advances in information management for the virtual era. Meckler Publishing; 1993. p. 45–55.

[47] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001;10(Pt 1):216–20.

[48] Chen Z, Perl Y, Halper M, Geller J, Gu H. Partitioning the UMLS semantic network. IEEE Trans Inf Technol Biomed 2002;6(2):102–8.

[49] Perl Y, Chen Z, Halper M, Geller J, Zhang L, Peng Y. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network. J Biomed Inform 2002;35(3):194–212.

[50] Zhang L, Perl Y, Halper MH, Geller J, Cimino JJ. Enriching the Structure of the UMLS Semantic Network. Proc AMIA Symp 2002:939–43.

[51] Cutting D, Kupiec J, Pedersen J, Sibun P. A practical part-of-speech tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing; 1992. p. 133–40.

[52] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp 2001:17–21.

[53] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM indexing initiative. Proc AMIA Symp 2000:17–21.

[54] Rindflesch TC. Integrating natural language processing and biomedical domain knowledge for increased information retrieval effectiveness. in: Proceedings of the Fifth Annual Dual-use Technologies and Applications Conference; 1995. p. 260–5.

[55] Gildea D, Palmer M. The necessity of parsing for predicate argument recognition. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic; 2002. p. 146–239.

[56] Liu H, Aronson AR, Friedman C. A study of abbreviations in MEDLINE abstracts. Proc AMIA Symp 2002:464–9.

[57] Wren JD, Garner HR. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. Methods Inf Med 2002;41(5):426–34.

[58] Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in electronic articles. J Am Med Inform Assoc 2002;9(3):262–72.

[59] Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. J Am Med Inform Assoc 1999;6(2):143–50.

[60] Klavans, J.L., Brian W. Extracting taxonomic relationships from online definitional sources using LEXING. In: Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), Roanoke, Virginia; 2001. p. 257–8.