

Integrating a Hypernymic Proposition Interpreter into a Semantic Processor for Biomedical Texts

Marcelo Fiszman MD PhD, Thomas C. Rindflesch PhD, and Halil Kilicoglu MS
National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland 20894

ABSTRACT

Semantic processing provides the potential for producing high quality results in natural language processing (NLP) applications in the biomedical domain. In this paper, we address a specific semantic phenomenon, the hypernymic proposition, and concentrate on integrating the interpretation of such predications into a more general semantic processor in order to improve overall accuracy. A preliminary evaluation assesses the contribution of hypernymic propositions in providing more specific semantic predications and thus improving effectiveness in retrieving treatment propositions in MEDLINE abstracts. Finally, we discuss the generalization of this methodology to additional semantic propositions as well as other types of biomedical texts.

INTRODUCTION

Accurate identification of semantic propositions (concepts and relations between concepts) in biomedical text would support enhanced effectiveness in biomedical applications such as focused information retrieval, computerized decision support systems, quality improvement, and medical research. Several approaches to semantic interpretation are being pursued in the medical informatics community.^{1,2,3,4,5} In our knowledge-based framework (called SemRep), we use underspecified syntactic analysis and structured domain knowledge from the Unified Medical Language System® (UMLS®)⁶ to identify semantic propositions in biomedical text.^{7,8}

In this paper, we address a specific semantic phenomenon, the hypernymic proposition, in which two concepts, one more specific (hyponym) and the other more general (hypernym), are in a taxonomic relationship. This is illustrated by the relationship between “**Tacrolimus**” and “**Immunomodulator**” in *Tacrolimus, a macrolide immunomodulator, is believed to control atopic dermatitis*. The hypernymic proposition in this sentence is represented as: “Tacrolimus-ISA-Immunomodulator.” We use ISA to cover any of several meanings of the hypernymic proposition, including subset / superset, generalization / specialization, kind-of, and role value restrictions.^{9,10}

Currently, in the sentence above, SemRep identifies the associative relationship “Immunomodulator-

TREATS-Atopic Dermatitis.” Although, this is correct, it is not precise. It would be more useful to identify “Tacrolimus” as the more specific semantic subject of TREATS in this sentence. There are two requirements for providing that information: a) a method for accurately identifying hypernymic propositions and b) the integration of that method into SemRep.

Although research has addressed the automatic enhancement of taxonomies based on syntactic patterns learned from corpora,¹¹ the semantic interpretation of these structures has not been addressed. We are developing a top down approach to interpret hypernymic propositions in biomedical text that uses syntactic patterns but also relies on medical domain knowledge from the UMLS to validate the propositions.¹² The focus of the present study is to explore the integration of this hypernymic proposition interpreter as a module in SemRep. Currently, we constrain this integration to treatment propositions in MEDLINE citations, but we discuss the extension of this methodology to additional semantic propositions as well as other types of biomedical texts.

BACKGROUND

Linguistic structure of the hypernymic proposition

Three syntactic phenomena commonly encode the hypernymic proposition: verbs, appositive structures, and nominal modification. Among verbs the most frequent is *be* when used as a main verb. In such instances, the hyponym is usually the subject and the hypernym is represented by the complement, as in (1). Other verbs like *remain* may also encode this relationship.

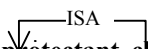
(1) **Lamivudin** is a **nucleoside analogue** with potent antiviral properties.

In appositive structures, two noun phrases must be contiguous. Three kinds of appositive cues can then mark the second noun phrase: commas, parentheses, or lexical items (*including, such as, particularly, and especially*). As an example, consider (2) where “Haloperidol” is the hyponym and “Antipsychotic drugs” is the hypernym.

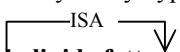
(2) The treatment of schizophrenia with old typical **antipsychotic drugs** such as **haloperidol** can be problematic.

In nominal modification, both concepts of the hypernymic proposition are represented in a single noun phrase. In these cases, the head may represent either the hypernym or the hyponym, while the modifier represents the other. Sentences (3) and (4) are examples of both situations.

(3) Clinical trial of the **neuroprotectant clome-thiazole** in coronary artery bypass surgery.



(4) Plasma **phospholipids fatty acids** were measured.



UMLS Resources

All three UMLS knowledge sources, the Metathesaurus,[®] the Semantic Network, and the SPECIALIST Lexicon are used in SemRep. Our interpreter for hypernymic propositions relies heavily on semantic groups from the Semantic Network and hierarchical relationships from the Metathesaurus.

Recent research by McCray et al.¹³ and Perl et al.¹⁴ to reduce the conceptual complexity of medical knowledge represented in the Semantic Network has resulted in the development of semantic groups. We use the semantic groups devised by McCray et al. to constrain the identification of hypernymic propositions. Their groups organize the 134 semantic types in the Semantic Network into 15 coarse grained aggregates. Currently we use the group Chemicals & Drugs, which contains such semantic types as ‘Pharmacologic Substance’, ‘Antibiotic’, and ‘Biologically Active Substance’. We also depend on (direct or indirect) hierarchical relationships in the Metathesaurus to identify hypernymic propositions.

SemRep

SemRep identifies associative semantic propositions in biomedical text. During processing, an underspecified syntactic parser¹⁵ depends on lexical look-up in the SPECIALIST lexicon and the Xerox Part-of-Speech Tagger.¹⁶ MetaMap¹⁷ matches noun phrases to concepts in the Metathesaurus and determines the semantic type for each concept. Argument identification is based on dependency grammar rules that enforce syntactic constraints. Indicator rules map syntactic phenomena to predicates in the Semantic Network, which imposes semantic validation for the associative relationships constructed; for example:

(5) **Alfuzosin** is effective in the *treatment of benign prostatic hyperplasia*

A semantic indicator rule links the nominalization *treatment* with the Semantic Network predicate “Pharmacologic Substance-TREATS-Disease or Syndrome.” Since the semantic types of the syntactic arguments identified for *treatment* in this sentence match the corresponding semantic types in the predi-

cation from the Semantic Network, the predication in (6) is constructed.

(6) Alfuzosin-TREATS-Prostatic Hypertrophy, Benign

METHODS

Interpreting Hypernymic Propositions

We are developing a system called **SemSpec** (Semantic Specification) to interpret hypernymic propositions as a module in SemRep. SemSpec¹² first identifies syntactic structures that potentially encode hypernymic propositions. After syntactic arguments have been identified, MetaMap matches them to concepts in the Metathesaurus. Such concepts are then subjected to semantic validation. In the current version, the semantic types must occur within the semantic group Chemicals & Drugs and the concepts themselves must be in a hierarchical relationship in the Metathesaurus.

As an example, consider the instance of nominal modification highlighted in (7).

(7) A number of reports show that [**the biguanide metformin**] improves ovarian function.

On the basis of the underspecified parse, in which the head and the modifier are identified for the noun phrase in bold, MetaMap matches the Metathesaurus concepts “Biguanides” and “Metformin” and their respective semantic types (‘Organic Chemical’ and ‘Pharmacologic Substance’). Since the semantic types belong to the semantic group Chemicals & Drugs, the Metathesaurus hierarchical file is consulted, and it is determined that “Biguanides” is an ancestor of “Metformin,” thus allowing the construction of (8).

(8) Metformin-ISA-Biguanides.

If a verb and its arguments encode a hypernymic proposition, the verb is most commonly a from of *be*. In our methodology, based on underspecified syntactic analysis, the identification of verbal arguments is subject to two syntactic constraints (in addition to semantic validation): A verb must occur between its potential arguments and there can be no more than four phrases intervening between the arguments, including the phrase containing the verb. As an example, (9) has three intervening phrases between the arguments of *be*. The hypernymic proposition identified by SemSpec is given in (10).

(9) **Oral ketorolac** [can] [be] [recommended] [as an **analgesic**] [for postoperative pain]

(10) Oral ketorolac-ISA-Analgesics

Incorporating SemSpec into SemRep

We can exploit SemSpec to identify more accurate predications in SemRep, based on meta-rules that

determine the more specific subject of a proposition generated by SemRep. We have devised such rules only for the TREATS predicate, but they can be generalized to include other predicates (see discussion).

- I. <Hypernym Y>-TREATS-<Object of TREATS predication>
- II. <Hyponym X>-ISA-<Hypernym Y>
- III. IF (I and II) THEN IV
- IV. <Hyponym X>-TREATS(SPEC)-<Object of TREATS predication>

Figure 1 - Meta-rule to retrieve a more specific treatment predication

The meta-rule for TREATS is given in Figure 1. If the hypernym concept of the hypernymic proposition (II) matches the subject of TREATS (I), we can create a new predication (IV), which substitutes the hyponym of the hypernymic proposition for the hypernym in the subject of the original predication. (SPEC) indicates that the semantic subject of TREATS is now the more specific concept.

As an example, we apply the rule in Figure 1 to the sentence in (11)

- (11) Market authorization has been granted in France for **pilocrapine**, an old **parasymphathomimetic agent**, in the *treatment of xerostomia*

SemRep retrieves (12) for (11) and SemSpec retrieves (13).

- (12) Parasympathomimetic Agents-TREATS-Xerostomia
 (13) Pilocrapine-ISA-Parasympathomimetic Agents
 From Figure 1, (I) and (II) are true, therefore (14) is generated.

- (14) Pilocrapine-TREATS(SPEC)-Xerostomia

If either the hyponym or the object of the TREATS predicate is coordinated and if the coordination algorithm from SemRep accurately identifies them, the meta-rule of Figure 1 applies to the coordinated concepts. (15) is an example with coordination of the object of TREATS. The predications generated after the meta-rule has applied to both objects of TREATS are given in (16).

- (15) **Vancomycin** is the **antibiotic** of choice for resistant **staphylococcal infections** and **bacterial endocarditis**
 (16) Vancomycin-TREATS(SPEC)-Staphylococcal infections; Vancomycin-TREATS(SPEC)-Bacterial endocarditis

Evaluation

We conducted a preliminary evaluation to test the performance of SemRep in extracting more specific

treatment propositions with and without the integration of the SemSpec module. We used a previously tagged (by MF) set of 340 sentences from MEDLINE citations that were retrieved using the Haynes methodological filter¹⁸ for the treatment purpose category without contents terms. These citations had also previously been subjected to a filter to enrich the prevalence of hypernymic propositions. We measured recall and precision (with 95% CI) for the SemRep program on the extraction of treatment propositions from this tagged sample, with and without the incorporation of the SemSpec module.

RESULTS

The total number of treatment propositions marked in the pre-tagged sample was 339. The performance of SemRep in extracting more specific treatment propositions increased by 7%, from 39% (34-44%) to 46% (41-51%) recall by adding the SemSpec module. The gain in recall by SemRep was not at the cost of reducing precision: Precision, at 78% (74-82%), was one point higher with SemSpec than without 77% (73-81%).

72 of the 339 treatment propositions marked in the sample were of the (SPEC) type; the meta-rule in Figure 1 is required to identify these propositions. Out of those 72, SemRep, after incorporating SemSpec, was able to identify 24 and still missed 48. Two false positives were produced.

DISCUSSION

We have presented a methodology to interpret hypernymic propositions based on underspecified syntactic analysis and domain knowledge provided by the UMLS. We have further focused on incorporating such interpretation into a generic semantic processor, SemRep. We were able to increase recall for SemRep in extracting more accurate predications by adding a meta-rule for treatment predications.

Since precision was higher than recall, we concentrate on analyzing false-negatives. When SemRep is running without SemSpec, recall errors are due to a variety of phenomena, including word sense ambiguity, missing semantic indicator rules, and missing Semantic Network relationships, as well as errors in processing coordination and other syntactic structures. In SemSpec, false negatives are caused by missing hierarchical relationships and concepts in the Metathesaurus, missing synonyms, and coordination problems.

The 48 missed treatment propositions that are more specific (SPEC) were due in part to SemRep (19 cases) and in part to SemSpec (29 cases). As an example of a SemSpec problem, consider the sentence in (17), in which a hierarchical relationship is not

represented in the Metathesaurus.

- (17) **Leukotriene receptor antagonists** are a new class of **anti-inflammatory drugs** that have clinical efficacy in the *management of asthma, allergic rhinitis, and inflammatory bowel disease*.

SemRep is able to recognize the TREATS propositions between the hypernym (“Anti-Inflammatory Agents”) and the diseases. However, SemSpec does not identify the hypernymic proposition between “Leukotriene Antagonists” and “Anti-Inflammatory Agents,” since these concepts do not appear in a hierarchical relationship in the Metathesaurus. Therefore, the meta-rule from Figure 1 cannot be applied.

As an example of a SemRep problem, consider the sentence in (18).

- (18) **Amisulpiride** is a selective **antipsychotic drug** with potent efficacy in exacerbations of **schizophrenia**.

In this sentence, SemSpec correctly interprets the hypernymic relationship between “Amisulpiride” and “Antipsychotic Agents.” However, SemRep does not correctly identify the semantic heads of macro-noun phrases like *exacerbations of schizophrenia*, and thus fails to recognize that *schizophrenia* is the object of the TREATS predication in this sentence.

A more interesting example of a false negative can be exemplified by (19).

- (19) **Topiramate** is being evaluated *for other neurological conditions such as migraine, neuropathic pain and essential tremor*.

In order for SemSpec to determine that this sentence asserts that “Migraine”, “Neuropathic Pain”, and “Essential Tremor” are hyponyms of “Neurological Disorders,” it must be expanded to include the semantic group Disorders. In addition, the meta-rule in Figure 1 must be generalized to refer to the hyponyms of the objects of TREATS as well as subjects.

We have recently expanded SemSpec to include the semantic groups Disorders as well as Procedures; however we have not yet evaluated its effectiveness. After processing a set of general MEDLINE abstracts, we noted that the interpretation of other predications could benefit from incorporating hypernymic propositions. In the interpretation (21) of (20), for example, the meta-rules in Figure 1 could be generalized to include OCCURS_IN, thus generating the more specific predication (22).

- (20) **Capillary hemangiomas** are rare **benign vascular tumors** that tend to *occur in children*
- (21) Hemangioma, Capillary-ISA-Neoplasms, Vascular Tissue; Neoplasms, Vascular Tissue-

OCCURS_IN-Child

- (22) Hemangioma, Capillary-OCCURS_IN(SPEC)-Child

We have also informally explored biomedical text other than MEDLINE citations. The National Library of Medicine’s MEDLINE^{plus}® contains links to a medical encyclopedia, which has definitions for thousands of concepts, including diseases, procedures, medications, and medical tests. An example from the encyclopedia is given in (23). The (SPEC) predications in (24) are based on meta-rules for LOCATION_OF and CAUSES.

- (23) **Mycoplasma Pneumonia** is an **infection** of the lung *caused by Mycoplasma pneumoniae*
- (24) Mycoplasma Pneumonia-ISA-Infection;
Lung-LOCATION_OF-Infection;
Lung-LOCATION_OF(SPEC)-Mycoplasma Pneumonia;
Mycoplasma pneumoniae-CAUSES-Infection;
Mycoplasma pneumoniae-CAUSES(SPEC)-Mycoplasma Pneumonia

Another potential application is in molecular biology. Currently, a system called SemGen is being developed to identify associative relationships between genes and diseases.¹⁹ Given information regarding hierarchical relationships in this domain, SemSpec could be incorporated into SemGen. Currently, SemGen interprets the predication (26) from (25).

- (25) Evidence supports that **PIK3CA** is an **oncogene** in **cervical cancer**
- (26) Oncogene-ASSOCIATED_WITH-Cervical Cancer

If SemSpec were able to identify (27), (28) could be generated (based on extension of the meta-rules to include ASSOCIATED_WITH).

- (27) PIK3CA-ISA-Oncogene
- (28) PIK3CA-ASSOCIATED_WITH(SPEC)-Cervical Cancer

The claim that the incorporation of SemSpec would improve the effectiveness of SemRep has been supported by the results of the evaluation, which indicate and increase recall of 7%; however, there are limitations to this study.

SemSpec must be tested in the context of each associative predicate SemRep identifies. In this preliminary study, we have applied SemSpec only to TREATS predicates. Further, the sample was enriched to include sentences more likely to have a hypernymic proposition. The true prevalence of these in the general MEDLINE literature is unknown. Therefore, the overall contribution of SemSpec to interpreting more accurate treatment propositions is an overestimation. In addition, only one expert, a

overestimation. In addition, only one expert, a developer of the system (MF), marked the treatment predications in the sample. Reliability of a gold standard is important when assessing any performance of an NLP system.²⁰ Finally, SemSpec has only been evaluated for the semantic group Chemicals & Drugs. It remains to be seen how the system will generalize when other groups are included, and how this will affect its incorporation into SemRep.

CONCLUSION

We have shown how the interpretation of a specific kind of semantic proposition, the hypernymic proposition, can be integrated into a general semantic processor to identify more specific treatment propositions in MEDLINE abstracts. This might be useful for information retrieval applications. We also discussed the generalization of this methodology to additional semantic propositions as well as other types of biomedical texts.

Acknowledgements

We acknowledge Alan Aronson, Olivier Bodenreider, and Bisharah Libbus, for their contributions to this project. The first author is 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

- Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp 2000;:270-4.
- Hahn U, Romacker M, Schulz S. MEDSYNDIKATE--design considerations for an ontology-based medical text understanding system. Proc AMIA Symp 2000;:330-4.
- Johnson SB, Aguirre A, Peng P, et al. Interpreting natural language queries using the UMLS. Proc Annu Symp Comput Appl Med Care 1993;:294-8.
- Christensen L; Haug PJ, Fiszman M. MPLUS: A probabilistic medical language understanding system. Proc Workshop on NLP in the Biomedical Domain, Assoc Comp Linguistics 2002;:29-36.
- Rassinoux AM, Wagner JC, Lovis C, et al. Analysis of medical texts based on a sound medical model. Proc Annu Symp Comput Appl Med Care 1995;:27-31.
- Humphreys BL, Lindberg DA, Schoolman HM, et al. The Unified Medical Language System: An informatics research collaboration. J Am Med Inform Assoc 1998 Jan-Feb;5(1):1-11.
- Rindflesch TC, Bean CA, Sneiderman CA. Argument identification for arterial branching predications asserted in cardiac catheterization reports. Proc AMIA Symp 2000;:704-8.
- Srinivasan P, Rindflesch T. Exploring Text Mining from MEDLINE. Proc AMIA Symp. 2002;:722-6.
- Brachman RJ. What IS-A is and isn't: an analysis of taxonomic links in semantic networks. Computer 1983;16(10):30-6.
- Bodenreider O, Burgun A, Rindflesch TC. Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. Proc Conf Terminology and Artificial Intelligence 2001;:11-21.
- Hearst MA. Automatic acquisition of hyponyms from large text corpora. Proc Int Conf on Comput Linguistics (COLING) 1992;:539-45.
- Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. Submitted, 2003.
- McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001;10(Pt 1):216-20.
- Perl Y, Chen Z, Halper M, et al. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network. J Biomed Inform. 2003 Jun;35(3):194-212.
- Rindflesch TC. Integrating natural language processing and biomedical domain knowledge for increased information retrieval effectiveness. Proc Ann Dual-use Technologies and Applications Conference 1995;:260-5.
- Cutting D, Kupiec J, Pedersen J, et al. A practical part-of-speech tagger. Proc Conf Applied NLP. 1992;:133-40.
- Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp 2001;:17-21.
- Haynes RB, Wilczynski N, McKibbon KA, et al. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc. 1994 Nov-Dec;1(6):447-58.
- Rindflesch TC, Libbus B, Hristovski D, et al. Semantic Relations Asserting the Etiology of Genetic Diseases. Proc AMIA Symp 2003, Submitted.
- Hripesak G, Kuperman GJ, Friedman C, et al. A reliability study for evaluating information extraction from radiology reports. J Am Med Inform Assoc. 1999;6(2):143-50.