SemMedDB: A PubMed-Scale Repository of Biomedical Semantic Predications

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C. Rindflesch

Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, 20894, USA

ABSTRACT
Summary: Effective access to the vast biomedical knowledge present in the scientific literature is challenging. Semantic relations are increasingly used in knowledge management applications supporting biomedical research to help address this challenge. We describe SemMedDB, a repository of semantic predications (subject-predicate-object triples) extracted from the entire set of PubMed citations. We propose the repository as a knowledge resource that can assist in hypothesis generation and literature-based discovery in biomedicine as well as in clinical decision-making support.


Contact: kilicogluh@mail.nih.gov

1 INTRODUCTION
Scientific discoveries depend on synthesis of knowledge from the literature and generation of novel hypotheses for testing. In biomedicine, the overwhelming size of the literature makes it difficult for researchers to identify promising new avenues, which may involve insights from seemingly unrelated subfields of the domain. Large-scale information extraction in the form of semantic relations is increasingly proposed for advanced knowledge management and discovery systems (Hristovski et al., 2006; Björne, et al., 2010; Cohen et al., 2010; Björne, et al., 2012).

In this note, we describe SemMedDB, a repository of semantic predications extracted from the titles and abstracts of all PubMed citations by SemRep (Rindflesch and Fiszman, 2003), a rule-based semantic interpreter. Elements of semantic predications are drawn from the UMLS knowledge sources (Bodenreider, 2004); the subject and object pairs correspond to Metathesaurus concepts, and the predicate to a relation type in an extended version of the Semantic Network. For example, SemRep extracts the predication Infection-CAUSES-Guillain-Barre Syndrome from the sentence Infections can trigger GBS. By normalizing the free text describing a relation to UMLS domain knowledge, SemRep provides the ability to combine and link knowledge from various sources as well as aggregate knowledge at PubMed scale. SemRep extracts 30 predicate types, largely relating to clinical medicine (e.g. TREATS, DIAGNOSES, ADMINISTERED_TO, PROCESS_OF), substance interactions (e.g. INTERACTS_WITH, INHIBITS, STIMULATES), genetic etiology of disease (e.g., ASSOCIATED_WITH, CAUSES, PREDISPOSES), and pharmacogenomics (e.g., AFFECTS, AUGMENTS, DISRUPTS). For a full list and descriptions of these predicate types, we refer the reader to Kilicoglu et al. (2011). Several evaluations have focused on different domains and linguistic structures. Results are summarized in Table 1. Numbers in the first column indicate the number of predications evaluated.

Table 1. Evaluation of SemRep predications

<table>
<thead>
<tr>
<th>Evaluation Type</th>
<th>Reference</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene-disease relations (1124)</td>
<td>Rindflesch et al., 2003</td>
<td>76%</td>
<td></td>
</tr>
<tr>
<td>Pharmacogenomics (623)</td>
<td>Ahlers et al., 2007</td>
<td>73%</td>
<td>55%</td>
</tr>
<tr>
<td>Hypernymic relations (830)</td>
<td>Rindflesch and Fiszman, 2003</td>
<td>83%</td>
<td></td>
</tr>
<tr>
<td>Comparative structures (300)</td>
<td>Fiszman et al., 2007</td>
<td>96%</td>
<td>70%</td>
</tr>
<tr>
<td>Nominalizations (300)</td>
<td>Kilicoglu et al., 2010</td>
<td>75%</td>
<td>57%</td>
</tr>
</tbody>
</table>

The SemMedDB repository consists of information regarding semantic predications extracted from PubMed citations by preprocessing and storing for efficient access. The repository underpins Semantic MEDLINE (http://skr3.nlm.nih.gov/SemMed), a Web application that incorporates PubMed-based information retrieval with semantic predications, automatic summarization, and visualization (Kilicoglu et al., 2008; Rindflesch et al., 2011). In this note, we propose the repository as a stand-alone, large-scale knowledge resource that can be exploited independently of Semantic MEDLINE. With the ability to access the full extent of the repository, the users can circumvent the ‘search-summarize-visualize’ model of Semantic MEDLINE and apply advanced data mining algorithms directly to support biomedical research as well as clinical practice, for hypothesis generation and literature-based discovery.

2 OVERVIEW OF SEMMEDDB
The SemMedDB repository is implemented primarily as a MySQL relational database that consists of tables holding information regarding PubMed citations, relevant UMLS knowledge, and the semantic predications. A brief description of the content of each table and the approximate number of records in it are provided in Table 2. The UMLS-related tables (CONCEPT*) contain information from a modified version of the UMLS 2006AA release, adapted for SemRep. The descriptions of data fields in each table are explained in detail in the online documentation. The entity-relationship diagram of the database is also provided.

The preprocessing of PubMed for SemMedDB takes about two months. Citations are first retrieved from PubMed using the NCBI E-utils API and then processed by SemRep in a distributed computing environment. The latest version of the repository has approximately 57.6M predications extracted from 21M citations (dated 6/30/2012 or earlier). We process
newly added citations at regular intervals and update the repository with the extracted predications.

3 USE CASES

SemMedDB is being used to support a range of biomedical applications, especially for literature-based discovery and hypothesis generation. Miller et al. (2012) proposed a mechanistic link between cortisol, testosterone and age-related sleep quality decline, while Wilkowski et al. (2011) applied graph-theoretical notions to elucidate the relation between sleep and depression. More recently, Goodwin et al. (2012) replicated these results by exploiting Information Foraging Theory. Cohen et al. (2012) used analogical reasoning in a high-dimensional vector space to suggest drug therapies. Similarly, Hristovski et al. (2012) proposed predication space for drug target discovery and drug repurposing, while Hristovski et al. (2010) combined predications with DNA microarray data to generate novel hypotheses on Parkinson disease. We are currently combining predications with microarray data to automatically generate gene regulatory networks.

SemMedDB has formed the basis of several clinical applications. Jonnalagadda et al. (2012) generate therapy-oriented summaries to support clinical decision-making, while Liu et al. (2012) elucidate the association between medical concepts co-occurring in clinical reports. The repository also underpinned a recent study to identify interactions between drugs mentioned in clinical reports.

4 CONCLUSION

We described the SemMedDB repository, which makes the semantic content of all PubMed citations as extracted by SemRep available to the research community. The utility of the repository for hypothesis generation, literature-based discovery and clinical decision-making has so far been demonstrated from within the Semantic MEDLINE paradigm and independently. Given its scale and size, this repository can further serve as the basis of advanced data mining techniques and assist in uncovering novel relationships in biomedicine.

Funding: This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

REFERENCES


Table 2. Brief descriptions of SemMedDB tables

<table>
<thead>
<tr>
<th>Name</th>
<th># of records</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>CITATION</td>
<td>21M</td>
<td>Metadata relevant for each PubMed citation</td>
</tr>
<tr>
<td>SENTENCE</td>
<td>119.1M</td>
<td>Sentences from each PubMed citation</td>
</tr>
<tr>
<td>CONCEPT</td>
<td>1.3M</td>
<td>Relevant information about UMLS Metathesaurus concepts</td>
</tr>
<tr>
<td>CONCEPT_SEMTYPE</td>
<td>1.5M</td>
<td>One-to-many relationships between concepts and their semantic types from UMLS Semantic Network</td>
</tr>
<tr>
<td>PREDICATION</td>
<td>12.9M</td>
<td>Unique predications</td>
</tr>
<tr>
<td>PREDICATION_ARGUMENT</td>
<td>27.5M</td>
<td>Links between each predication and its subject and object contained in CONCEPT table</td>
</tr>
<tr>
<td>SENTENCE_PREDICATION</td>
<td>57.6M</td>
<td>Links between a sentence and a predication extracted from it</td>
</tr>
<tr>
<td>PREDICATION_AGGREGATE</td>
<td>57.6M</td>
<td>Convenience table that aggregates information from all of the tables above for more efficient access</td>
</tr>
</tbody>
</table>


Klicicglo, H. et al. (2011) Constructing a semantic predication gold standard from the biomedical literature. BMC Bioinformatics, 12, 486.


