BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 23 2012, pages 3158–3160
doi:10.1093/bioinformatics/bts591

Data and text mining Advance Access publication October 8, 2012

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

Associate Editor: Jonathan Wren

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.

Availability and implementation: The SemMedDB repository is available as a MySQL database for non-commercial use at http://skr3.nlm.nih.gov/SemMedDB. An UMLS Metathesaurus license is required.

Contact: kilicogluh@mail.nih.gov

Received on August 1, 2012; revised on September 8, 2012; accepted on September 25, 2012

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 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 (Björne *et al.,* 2010, 2012; Cohen *et al*., 2010; Hristovski *et al.,* 2006).

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 Unified Medical Language System (UMLS) knowledge sources (Bodenreider, 2004); the subject and object pair corresponds 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.

The SemMedDB repository consists of information regarding semantic predications extracted from PubMed citations by preprocessing and stored 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

Table 1. Evaluation of SemRep predications

Evaluation type	Reference	Precision (%)	Recall (%)
Gene-disease relations (1124)	Rindflesch *et al*., 2003	76	—
Pharmacogenomics (623)	Ahlers *et al*., 2007	73	55
Hypernymic relations (830)	Rindflesch and Fiszman, 2003	83	—
Comparative structures (300)	Fiszman *et al*., 2007	96	70
Nominalizations (300)	Kilicoglu *et al*., 2010	75	57

*To whom correspondence should be addressed.

Table 2. Brief descriptions of SemMedDB tables

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

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 2 months. Citations are first retrieved from PubMed using the NCBI E-utilities API and then processed by SemRep in a distributed computing environment. The latest version of the repository has ~57.6 M predications extracted from 21 M citations (dated June 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, whereas 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.

Conflict of Interest: none declared.

REFERENCES

Ahlers,C.B. *et al.* (2007) Extracting semantic predications from Medline citations for pharmacogenomics. In *Pacific Symposium on Biocomputing*. World Scientific, Maui, HI, USA, pp. 209–220.

Björne,J. *et al.* (2010) Scaling up biomedical event extraction to the entire PubMed. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP'10)*. Association of Computational Linguistics, Uppsala, Sweden, pp. 28–36.

Björne,J. *et al.* (2012) PubMed-scale event extraction for post-translational modifications, epigenetics and protein structural relations. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP'12)*. Association of Computational Linguistics, Montreal, Canada, pp. 82–90.

Bodenreider,O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

Cohen,T. *et al.* (2010) EpiphaNet: an interactive tool to support biomedical discoveries. *J. Biomed. Discov. Collab.*, **5**, 21–49.

Cohen,T. *et al.* (2012) Many paths lead to discovery: analogical retrieval of cancer therapies. In *Proceedings of the Sixth International Conference on Quantum Interaction (QI'12)*. Springer, Paris, France (in press).

Fiszman,M. *et al.* (2007) Interpreting comparative constructions in biomedical text. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP'07)*. Association of Computational Linguistics, Prague, Czech Republic, pp. 137–144.

Goodwin,J.C. *et al.* (2012) Discovery by scent: closed literature-based discovery system based on the information foraging theory. In *IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*. IEEE, Philadelphia, PA, USA, pp. 232–239.

Hristovski,D. *et al.* (2006) Exploiting semantic relations for literature-based discovery. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, Washington, DC, USA, pp. 349–353.

Hristovski,D. *et al*. (2010) Combining semantic relations and DNA microarray data for novel hypothesis generation. In: Blaschke,C. and Shatkay,H. (eds.) *ISMB/ECCB2009, Lecture Notes in Bioinformatics*. Springer, Heidelberg, pp. 53–61.

Hristovski,D. *et al*. (2012) Using literature-based discovery to identify novel therapeutic approaches. *Cardiovasc. Hematol. Agents. Med. Chem*., (in press).

Jonnalagadda,S. *et al*. (2012) Automatically extracting sentences from Medline citations to support clinicians' information needs. *J. Am. Med. Inform. Assn*, (in press).

Kilicoglu,H. *et al*. (2008) Semantic MEDLINE: a web application to manage the results of PubMed searches. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*. Turku Centre for Computer Science (TUCS), Turku, Finland, pp. 69–76.

Kilicoglu,H. *et al*. (2010) Arguments of nominals in semantic interpretation of biomedical text. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP'10)*. Association of Computational Linguistics, Uppsala, Sweden, pp. 46–54.

Kilicoglu,H. *et al*. (2011) Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinformatics*, American Medical Informatics Association, Chicago, IL, USA, **12**, 486.

Liu,Y. *et al*. (2012) Using SemRep to label semantic relations extracted from clinical text. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, Chicago, IL, USA.

Miller,C.M. *et al*. (2012) A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. *Sleep*, **35**, 279–285.

Rindflesch,T.C. and Fiszman,M. (2003) The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J. Biomed. Inform*., **36**, 462–477.

Rindflesch,T.C. *et al*. (2003) Semantic relations asserting the etiology of genetic diseases. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, Washington, DC, USA, pp. 554–558.

Rindflesch,T.C. *et al*. (2011) Semantic MEDLINE: an advanced information management application for biomedicine. *Inform. Services Use*, **31**, 15–21.

Wilkowski,B. *et al*. (2011) Graph-based methods for discovery browsing with semantic predications. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, Washington, DC, USA, pp. 1514–1523.