Exploiting Semantic Relations for Literature-Based Discovery

Dimitar Hristovski,[1] PhD, Carol Friedman,[2] PhD, Thomas C Rindflesch,[3] PhD, Borut Peterlin,[4] MD PhD

[1]Institute of Biomedical Informatics, Medical Faculty, University of Ljubljana
[2]Department of Biomedical Informatics, Columbia University, New York
[3]National Library of Medicine, Bethesda, Maryland
[4]Division of medical genetics, UMC, Slajmerjeva 3, Ljubljana, Slovenia
e-mail: dimitar.hristovski@mf.uni-lj.si

We propose using semantic predications to enhance literature-based discovery (LBD) systems, which currently depend exclusively on co-occurrence of words or concepts in target documents. In this paper, the predications, which are produced by the combined application of two natural language processing systems, BioMedLEE and SemRep, are coupled with an LBD system BITOLA. Initial experiments suggest this approach can uncover new associations that were not possible using previous methods.

INTRODUCTION

Literature-based discovery (LBD) is a method for automatically generating hypotheses for scientific research by finding overlooked implicit connections in the research literature. Discoveries have the form of relations between two primary concepts, for example a drug as a treatment for a disease or a gene as the cause of a disease. Swanson [1] introduced a paradigm in which such relations are discovered in bibliographic databases by uncovering a third concept (such as a physiologic function) that is related to both the drug and the disease. The discovery of the third concept allows a relation between the primary concepts, which was latent in the literature, to become explicit, thus constituting a potential discovery.

Current literature-based discovery systems (for example [1-8]) use simple concept co-occurrence as their primary mechanism. No semantic information about the nature of the relation between concepts is provided. The use of co-occurrence has several drawbacks, since not all co occurrences underlie “interesting” relations: (a) Users must read large numbers of Medline citations when reviewing candidate relations; (b) systems tend to produce large numbers of spurious relations; and, finally, (c) there is no explicit explanation of the discovered relation.

In this paper we address these deficiencies by enhancing the literature-based paradigm with the use of semantic relations to augment co-occurrence processing. We combine the output of two natural language processing systems to provide these predications: SemRep [9] and BioMedLee [10]. On the basis of explicit semantic predications, the user can ignore relations which are either uninteresting (thus reducing the amount of reading required) or wrong (eliminating false positives). Analysis using predications can support an explanation of potential discoveries.

BACKGROUND

Literature-based discovery

The methodology in literature-based discovery relies on the notion of concepts relevant to three literature domains: X, Y, and Z. In a typical scenario, X concepts are those associated with some disease and Z concepts relate to a drug that treats the disease. Y concepts might then be physiological or pathological functions, symptoms, or body measurements. Concepts in X and Y are often discussed together, as are those in Y and Z. However, concepts from X and Z may not appear together in the same research paper. Discovery is facilitated by using particular Y concepts to draw attention to a connection between X and Z that had not been previously noticed.

In implementation, usually all the Y concepts in a bibliographic database related to the starting concept X are first computed. Then the Z concepts related to Y are found. Those Y concepts that appear in both X and Z provide the link from X to Z. The user then checks whether X and Z appear together in the research literature; if they do not, a potentially useful relation has been discovered. This relation needs to be confirmed or rejected using human

Figure 1. Discovery pattern *Maybe_Treats*. *Maybe_Treats1* proposes Z1 (drug or substance) as a n treatment for disease X because Z1 causes opposite change to Y1 (function or substance) and the change Y1 is a characteristic of disease X. *Maybe_Treats2* proposes Z2 as a new treatment of X because there i similar disease, X2, and drug Z2 is known to treat X2.

judgment, laboratory methods, or clinical investigations.

In a discovery reported by Swanson [1], the X domain was Raynaud's disease. Of the many Y terms co-occurring with this disorder, blood viscosity and platelet aggregation were found to co-occur also with a Z term, fish oil (rich in eicosapentaenoic acid). Fish oil (Z) reduces blood viscosity and platelet aggregation (Y), which are increased in Raynaud's disease (X), and thus fish oil was proposed as a new treatment for Raynaud's disease. However, in the original attempt done by Swanson and all the subsequent replications of this discovery, what is "increased" in relation to a disease and what can be used "to decrease" it, has been left to be extracted by the user by reading relevant Medline citations. This is exactly where we want to improve the state-of-the-art in LBD.

Swanson (together with Smalheiser) has published several other medical discoveries using this methodology, and have developed a LBD system called Arrowsmith [2]. There are a few other LBD systems such as ours (BITOLA) [3,4], and those developed by others ([5,6,7,8] for example).

Natural language processing

BioMedLEE captures a large variety of genotypic and phenotypic information and relations from the literature. BioMedLEE is a recent adaptation of MedLEE [11,12], which was developed to structure and encode clinical information in the patient record. BioMedLEE is based on a grammar formalism that combines syntax and semantics and uses MedLEE's lexicon derived from clinical documents, the UMLS, and other online biomedical knowledge sources, such as some of the ontologies in the Open Biomedical Ontologies (OBO) (http://obo.sourceforge.net/) consortium . However, this work focuses on use of the concepts in the UMLS Metathesaurus only.

SemRep [9] is a symbolic natural language processing system for identifying semantic relations in biomedical text. The program currently focuses on Medline citations emphasizing treatment of disease. Linguistic processing is based on an underspecified (shallow) parse tree supported by the SPECIALIST Lexicon. Medical domain knowledge is provided by the UMLS Metathesaurus, accessed using MetaMap [13]. Identification of semantic relations is guided by the UMLS Semantic Network. SemRep identifies a variety of semantic predications. For this project, the most relevant relation (predication) is *Treats*.

METHODS

In order to exploit semantic predications in literature-based discovery, we introduce the notion of a discovery pattern, which contains a set of conditions to be satisfied for the discovery of new relations between concepts. The conditions are combinations of relations between concepts extracted from Medline citations. In this paper we deal with the *Maybe_Treats* pattern, which has two forms: *Maybe_Treats1* and *Maybe_Treats2* (Fig. 1). In both forms the goal is to propose potentially new treatments, and the two can work in concert: proposing either two different new treatments (complementary) or the same treatments by using different discovery reasoning (reinforcement). The following reasoning is used as a novelty check for the proposed new treatments (stated informally in terms of the X, Y, Z paradigm): It is a discovery that drug Z maybe treats disease X if there is currently no evidence in the medical literature that drug Z is already used to treat disease X.

The two forms are different in the way they generate new candidate treatments Z. The first form *Maybe_Treats1* is satisfied when there is a change in