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, Slovenia*
[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

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

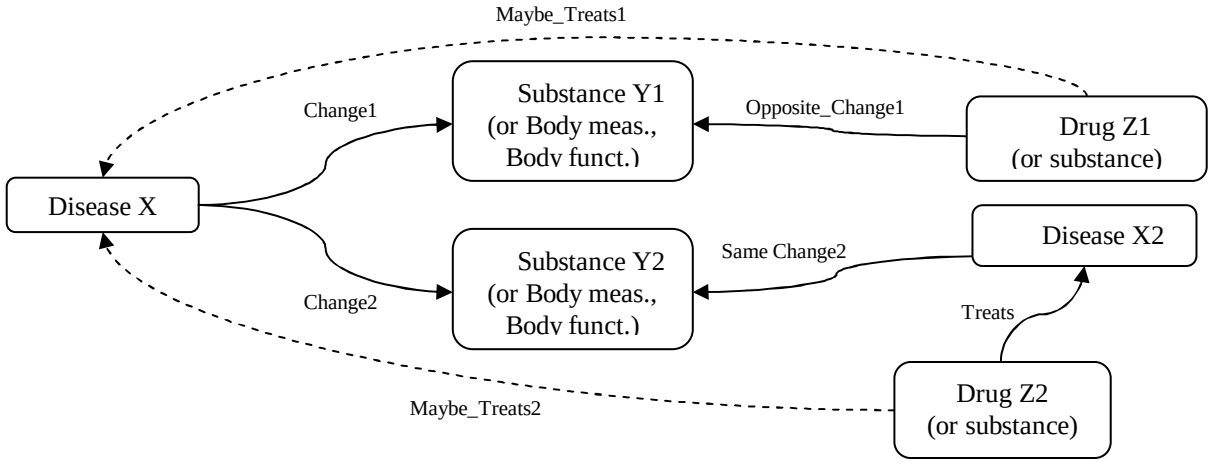

Figure 1. Discovery pattern *Maybe_Treats*. *Maybe_Treats1* proposes Z1 (drug or substance) as a new treatment for disease X because Z1 causes opposite change to Y1 (function or substance) and the change of Y1 is a characteristic of disease X. *Maybe_Treats2* proposes Z2 as a new treatment of X because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

Maybe_Treats1 is satisfied when there is a change in a substance, body function, or body measurement (concept Y) associated with the starting disease X, and there is also an opposite change in concept Y associated with the concept Z. An example of this form is Swanson's proposal of fish oil (Z) as a new treatment for Raynaud's disease (X). Fish oil (Z) reduces blood viscosity (Y) which was reported in the literature to be increased in patients with Raynaud's.

In *Maybe_Treats2*, in order to find a potentially new treatment for a starting disease X we first search for another disease X2 which has similar characteristics (Y2 substances or functions changed in the same direction (both increased or decreased). Then we propose as a potentially new treatment for disease X the drugs (Z2) already used to treat the disease X2, if there is no evidence in the literature that Z2 is already used to treat X. An example of this might be what we have observed while performing this research. In patients with Huntington disease (HD) the level of insulin is often decreased. The level of insulin is also decreased in Diabetes Mellitus. Therefore, treatments for diabetes might also be used for HD. The relations *Associated_with_change* and *Treats* are used to extract known facts from the biomedical literature. The relations *Maybe_Treats1* and *Maybe_Treats2* predict potentially new treatments based on the known facts extracted by *Associated_with_change* and *Treats*.

Associated_with_change is used to extract a relation where one concept is associated with a change in another concept (e.g. a disease associated with an increase in the level of a substance). For the extraction of *Associated_with_change* we use BioMedLee. The relation *Treats* is used to extract drugs known to treat a disease according to the literature. The major purpose of this relation in our approach is to eliminate the drugs already known to be used for treatment from the list of drugs or chemicals that have not been used yet, but seem promising. Additionally, in the *Maybe_Treats2* form, the *Treats* relation is used to find existing treatments to similar diseases. *Treats* relations are identified by SemRep.

It is possible to use the *Maybe_Treats* pattern (both forms) for several discovery tasks depending on what input is provided. If a drug Z is provided as input then the pattern will try to generate diseases X that might be treated. If a disease X is provided as input then the pattern will try to generate drugs Z that might be used to treat the disease X. If both a disease X and a drug Z are provided as input, then the pattern will test whether the drug might be used to treat the disease. If it can, the pattern can generate an explanation through the intermediate concepts Y. For example, the drug Z might be used to treat X because Y is increased in disease X and Z has been reported to decrease the level of Y.

Num.	System	Extracted Relations	Sentence (or fragment)
1.	BL	Associated_with(oxidative stress, iron, increase)	reducing the oxidative stress associated with increased iron levels
2.	SR	Treats(coenzyme Q10,Huntington Disease)	Oral administration of CoQ10 significantly decreased elevated lactate levels in patients with Huntington's disease.
3.	BL	Associated_with(Raynaud's, blood viscosity, increase)	Local increase of blood viscosity during cold-induced Raynaud's phenomenon.
4.	BL	Associated_with(Raynaud's, viscosity, increase)	Increased viscosity might be a causal factor in secondary forms of Raynaud's disease, …
5.	BL	Associated_with(eicosapentaenoic acid, blood viscosity, decrease)	We recently reported that eicosapentaenoic acid (EPA) also reduces whole blood viscosity.
6.	BL	Associated_with(eicosapentaenoic acid, blood viscosity, decrease)	A statistically significant reduction in whole blood viscosity was observed at seven weeks in those patients receiving the eicosapentaenoic acid rich oil.
7.	BL	Associated_with(Huntington's disease, insulin, decrease)	Huntington's disease transgenic mice develop an age-dependent reduction of insulin mRNA expression and diminished expression of key regulators of insulin gene transcription, …

Table 1. Examples of extracted relations by BioMedLee (BL) or SemRep (SR). The relation *Associated_with* shown in column 3, represented a shortened form of *Associated_with_change*.

RESULTS

To illustrate the *Maybe_Treats1* discovery pattern, we show how Swanson's Raynaud's discovery [1] could be replicated. This example also illustrates integration of semantic relation extraction with an existing (co-occurrence based) LBD system.

We used the BITOLA [3,4] LBD system (available at http://www.mf.uni-lj.si/bitola/) and searched for Raynaud's as the starting concept X. Then among the related concepts Y limited to the semantic group *Physiology*, we found *Blood Viscosity* in the eighth place and *Platelet Aggregation* in the seventeenth place out of 230 concepts from the *Physiology* group that co-occur with Raynaud's.

We then submitted the citations in which Raynaud's co-occurs with either *Blood Viscosity* or *Platelet Aggregation* to BioMedLee, which produced five relations in which Raynaud's was associated with an increase in blood viscosity (examples 3. and 4. in Table 1.) and one in which Raynaud's was associated with platelet aggregation.

In the next step we used BITOLA to search for concepts co-occurring with blood viscosity or platelet aggregation. Among others we found *Eicosapentaenoic acid,* which can be found in large quantities in fish oil. After processing the relevant Medline citations with BioMedLee, we obtained several relations in which eicosapentainoic acid was associated with a reduction in blood viscosity (examples 5 and 6 in Table 1). By combining examples 3 and 4 with 5 and 6 we can conclude that eicosapentainoic acid (Z) (and consequently food rich it this acid such as fish oil) might be used to treat Raynaud's (X) because blood viscosity (Y) is increased in Raynaud's and eicosapentainoic acid reduces blood viscosity.

To illustrate the *Maybe_Treats2* form of the *Maybe_Treats* discovery pattern, we selected *Huntington disease* as a test case. Huntington disease (HD) is an autosomal-dominant inherited neurodegenerative disorder that is characterized by the insidious progressive development of mood disturbances, behavioural changes, involuntary choreiform movements and cognitive impairments. Onset is most commonly in adulthood, with a typical duration of 15-20 years before premature death. No successful treatment is currently available. We constructed a set of all 5511 Medline citations (in January, 2006) in which Huntington Disease occurs as a MeSH heading. We first submitted this set to SemRep, which extracted 30,103 relations, out of which 2139 were *Treats* relations. Of these, 740 *Treats* relations contained Huntington disease as an argument. These represent current treatments for Huntington (example 2 in Table 1).

Our strategy then was to find relations between HD and changes in substances or body functions which could be potential therapeutic targets for HD. For this we submitted the Huntington citations to BioMedLee, which extracted 18360 relations, of which 1912 contained a change, 310 of which were associated with Huntington disease.

From the 310 relations, a clinician who is an expert in HD, selected 35 interesting concepts representing neurotransmitters, their receptors or other biologic substances changed in HD. The next step was to find diseases in which these concepts were changed in the same way as in HD.

We discovered an interesting potential new treatment for HD – insulin, which was one of the substances found to be **decreased** in HD (example 7 in Table 1). It is known that HD patients develop diabetes mellitus about seven times more often than matched healthy control individuals [14]. The reason for this is unclear, although inappropriate insulin secretion is a potential reason. The transgenic HD mouse model also develops an age-dependent reduction of insulin mRNA expression and diminished expression of key regulators of insulin gene transcription [15].

Strong evidence from studies in humans and animal models suggests the involvement of energy metabolism defects, which may contribute to excitotoxic processes, oxidative damage, and altered gene regulation in the pathogenetic mechanism of HD. Reduced glucose metabolism in affected brain areas of HD patients is a well documented fact used for diagnostic purposes.

We then searched for diseases other than HD with reduced levels of insulin. Expectedly the system identified diabetes mellitus. We thus concluded that insulin treatment, used for diabetes mellitus, might be an interesting drug for HD. Insulin might improve glucose metabolism in the brains of HD patients and thus slow down the pathogenetic process.

DISCUSSION

Although there are clear advantages in using semantic relation extraction for LBD, there are also some issues that have to be addressed. One is scalability. Ideally all of Medline needs to be processed to support the system we propose. The other issue is accuracy in semantic relation extraction. In further work we plan to evaluate the

performance of semantic relation extraction and the performance of LDB based on that extraction. Because of these issues, we believe that for the near future, the best approach would be the integration of semantic relation extraction with co-occurrence based LBD. In further work we plan to integrate the BITOLA LBD system with SemRep and BioMedLee.

Another research contribution is the use of two natural language processing systems, namely SemRep and BioMedLee, to extract the kind of relations they are best at capturing. This entailed developing a common format for each system's output. To our knowledge this is the first time two different natural language processing systems have been utilized together to capture different types of semantic relations.

CONCLUSIONS

We presented a new method aimed at improving literature-based discovery. It is based on semantic predications, which are extracted from text using the combined results of two natural language processing systems. The proposed system has the potential to produce a smaller number of false positive discoveries while, at the same time, facilitating user evaluation and review of potentially new relations. Finally, it can support explanation of the discovery produced.

ACKNOWLEDGEMENTS

The part of this research done at Columbia University was supported by grants LM007659 and LM008635 from the National Institutes of Health. This study was supported in part by the Intramural Research Programs of the National Institutes of Health, National Library of Medicine.

References

1. Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986 Autumn;30(1):7-18.
2. Swanson, D.R., Smalheiser, N.R.: An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif. Intell. 91 (1997) 183-203.
3. Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. MEDINFO 2001.
4. Hristovski, D., Peterlin, B., Mitchell, J. A. and Humphrey, S. M. (2005), 'Using literature-based discovery to identify disease candidate genes', Int. J. Med. Inform., Vol. 74(2–4), pp. 289–298.
5. Weeber M, Klein H, Aronson AR, Mork JG, Jong-Van Den Berg L, Vos R. Text-based discovery in biomedicine: the architecture of the DAD-system. Proc AMIA Symp. 2000;(20 Suppl):903-7.
6. Srinivasan, P. and Libbus, B. (2004), 'Mining MEDLINE for implicit links between dietary substances and diseases', Bioinformatics, Vol. 20, Suppl 1, pp. I290–I296.
7. Gordon MD, Lindsay RK. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. J Am Soc Inf Sci 1996; 47(2):116-128.
8. Fuller, S. S., Revere, D., Bugni, P. F. and Martin, G. M. (2004), 'A knowledgebase system to enhance scientific discovery: Telemakus', Biomed. Digit Libr., Vol. 1(1), p. 2.
9. 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 2003;36:462-477.
10. Lussier YA, Borlawski T, Rappaport D, Liu Y, Friedman C. PhenoGO: assigning phenotypic context to Gene Ontology annotations with natural language processing. Pac Symp Bio. 2006:64-75.
11. Friedman C, Alderson P, Austin J, Cimino JJ, and Johnson SB. A general natural language text processor for clinical radiology. Journal of American Medical Informatics Association, March 1994, 1(2):161--174.
12. Friedman C, Shagina L, Lussier YA, Hripcsak G. Automated encoding of clinical documents based on natural language processing.J Am Med Inform Assoc. 2004 Sep-Oct;11(5):392-402. Epub 2004 Jun 7
13. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp 2001:17-21.
14. Ristow M. Neurodegenerative disorders associated with diabetes mellitus. J Mol Med. 2004; 82:510-2.
15. Andreassen OA, Dedeoglu A, Stanojevic V, et al. Huntington's disease of the endocrine pancreas: insulin deficiency and diabetes mellitus due to impaired insulin gene expression. Neurobiol Dis. 2002;11:410-24.