Literature-Based Knowledge Discovery using Natural Language Processing

D. Hristovski, C. Friedman, T.C. Rindflesch, and B. Peterlin

Abstract Literature-based discovery (LBD) is an emerging methodology for uncovering nonovert relationships in the online research literature. Making such relationships explicit supports hypothesis generation and discovery. Currently LBD systems depend exclusively on co-occurrence of words or concepts in target documents, regardless of whether relations actually exist between the words or concepts. We describe a method to enhance LBD through capture of semantic relations from the literature via use of natural language processing (NLP). This paper reports on an application of LBD that combines two NLP systems: BioMedLEE and SemRep, which are coupled with an LBD system called BITOLA. The two NLP systems complement each other to increase the types of information utilized by BITOLA. We also discuss issues associated with combining heterogeneous systems. Initial experiments suggest this approach can uncover new associations that were not possible using previous methods.

1 Introduction

Literature-based discovery (LBD) is a method for automatically generating hypotheses for scientific research by finding overlooked implicit connections in the research

D. Hristovski
Institute of Biomedical Informatics, Medical Faculty, University of Ljubljana, Vrazov trg 2/2, 1104 Ljubljana, Slovenia
dimitar.hristovski@mf.uni-lj.si

C. Friedman
Department of Biomedical Informatics, Columbia University, 622 West 168 St, New York, NY 10032, USA

T.C. Rindflesch
National Library of Medicine, Bethesda, Maryland, USA

B. Peterlin
Division of medical genetics, UMC, Slajmerjeva 3, Ljubljana, Slovenia

P. Bruza and M. Weeber (eds.), *Literature-based Discovery,*
Springer Series in Information Science and Knowledge Management 15.

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 [2–12] use 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 chapter 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 [13] and BioMedLee [14]. 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.

2 Background

2.1 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, all the Y concepts in a bibliographic database related to the starting concept X are usually computed first. Then the Z concepts related to Y are found. Those Y concepts that appear with 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 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. Swanson has published several other medical discoveries using this methodology. However, in his original work (and in all subsequent replications of this discovery), what is increased in relation to the disease and what can be used to decrease it, must be determined by reading relevant Medline citations. This is exactly where we want to improve the state-of-the-art in LBD.

Several methods are being pursued in current LBD systems (for a more detailed review see [15]). Some systems extract concepts from the titles and abstracts of Medline citations (often using MetaMap [16]), while others use the assigned MeSH descriptors to represent concepts in citations. All systems use co-occurrence to determine which concepts are in a relationship, although some augment co-occurrence with other derived relation measures. Usually the semantic types of the concepts are used to filter out unneeded relations and concepts.

Swanson and Smalheiser have developed a system called Arrowsmith [2], which uses co-occurrence of words or phrases from the title of Medline citations. The BITOLA system (Hristovski et al. [3,4]) uses association rules as a relation measure between concepts. In general, association rule mining [17] finds interesting associations and/or correlation relationships among large set of data items. In BITOLA a data item corresponds to a Medline citation and is represented as a set of concepts. For each citation, the concepts are the assigned MeSH headings and additionally gene symbols extracted from the titles and abstracts of Medline citations. For example, the association rule *Multiple Sclerosis* → *Optic Neuritis* tells us that there is probably some association between *Multiple Sclerosis* and *Optic Neuritis*, but does not tell us the semantic nature of this association.

Weeber et al. [5] use MetaMap to identify UMLS concepts in titles and abstracts and use concept co-occurrence as a relation measure. For filtering, they use UMLS semantic types. For example, the semantic type of one of the co-occurring concepts might be set to *Disease or Syndrome* and the other to *Pharmacologic Substance*, thus only co-occurrences between a disease and a drug are found. Lindsay and Gordon [6] use an approach similar to Arrowsmith but add various information retrieval techniques to assign weight to the terms being manipulated. Gordon and Dumais [7] employ a statistical method called latent semantic indexing to assist in LBD. Wren [8] uses mutual information measures for ranking target terms based on their shared associations. Srinivasan [12] developed a system, called Manjal, which uses MeSH terms as concepts and term weights instead of simple term frequencies. For ranking, the system uses an information retrieval measure based on term co-occurrence. Pratt [9] uses MetaMap to extract UMLS concepts from the titles of Medline citations and then uses association rules as a relationship measure between concepts.

The Telemakus system [10] is different from the rest of the systems mentioned in so far as it uses manually extracted relationships to represent the research findings. Each relationship is a pair of concepts from the article's figure and title legends.

The semantic relation between the concepts is not extracted. The manual relation extraction method has two consequences: the positive one is that the method has high precision and the negative one is that it is time consuming and thus currently used in only two relatively narrow domains.

Recently Hu [11] presented a system called Bio-SbKDS where MeSH terms are used as concepts. This system uses the relations between semantic types from the UMLS Semantic Network for two purposes: to filter out uninteresting concepts, and to guess the semantic relation between concepts. In other words, if two concepts co-occur in a Medline citation, the relation between the corresponding semantic types of these two concepts is used as the semantic relation between the concepts. This is only an approximation because there is no guarantee that if the concepts co-occur they are semantically related and also there is ambiguity in the UMLS Semantic Network because often more then one semantic relation is present between two semantic types. However, this approach seems to work quite well in replicating Swanson's Raynaud's – fish oil discovery.

Our method differs from all the above methods because we use natural language processing (NLP) techniques to augment co-occurrences with specific types of relations, which are obtained as a result of using two different NLP systems.

2.2 Natural Language Processing

2.2.1 BioMedLEE Natural Language Processing System

BioMedLEE captures genotypic and phenotypic information and relations from the literature, and is a recent adaptation of MedLEE [18, 19], which was developed to structure and encode telegraphic clinical information in the patient record. BioMedLEE is based on a symbolic grammar formalism that combines syntax and semantics, using a lexicon to specify semantic and syntactic classes for words and phrases in the domain. The lexicon consists of a modified and augmented version of MedLEE's lexicon, which was derived from clinical documents, the UMLS (Unified Medical Language System) [20], and other online biomedical knowledge sources, but this work focuses on use of the concepts that correspond to UMLS Metathesaurus concepts only. BioMedLEE consists of a number of different text processing modules, each of which aims to regularize specific aspects of text processing while minimizing loss of information. The following is a brief summary of the primary modules and the resources they use:

a. *Abbreviation and Parenthesis Component*: This module identifies abbreviations explicitly defined in the article, and tags them so that the subsequent modules will be able to substitute the full form in place of the abbreviation. For example, HD, in *Huntington Disease (HD)* will be assumed to be *Huntington Disease* throughout the article. Other parenthetical expressions may be tagged so that they will be ignored during parsing.

b. *Biomolecular Named Entity Recognition and Normalization*: This module uses part of speech tagging to recognize the boundaries of noun phrases, and then identifies ones that appear to be biomolecular entities, such as the names of genes, gene products, and other substances. The terms that are biomolecular entities are then matched against a database of biomolecular entities using regular expressions that allow for certain variations (e.g. *il-2, il 2, il2*). When a match is found, the term is tagged so that the tag includes the semantic category (e.g. gene/gene product, substance), and the target output form. For example, after tagging is performed (we assume here that the tagging module used a database of UMLS genes and proteins to normalize biomolecular entities), the tagged output for the sentence "*Axonal transport of N-terminal huntingtin suggests pathology of corticostriatal projections associated with HD"* will be "*Axonal transport of N-terminal <phr sem="gp"t="UMLS: C1415504_hd gene"> huntingtin</phr> suggests pathology of corticostriatal projections associated with <phr sem="disease" t="Huntington's disease">HD</phr>"*. The tag around *huntingtin* has an attribute, which is a semantic category **sem** with value **gp** representing the category **gene/gene product** and a target form attribute **t**, which, in this case, is the UMLS code previously generated by the tagger. In addition, there is a tag around, *HD*, with a semantic category **disease** and target form **Huntington's disease**, which is the full form that occurred previously in the article along with the abbreviation *HD*.

c. *Preprocessing Component*: This module determines section and sentence boundaries, and performs lexical lookup for the remaining parts of the sentence that were not tagged in b. above. This would include phenotypic entities, such as anatomical locations, diseases, and processes, as well as functional English words. For example, "corticostriatal" would be identified as an anatomical concept, and "suggest" would be identified as a relation that could connect two biomedical entities. The relations are semantic relations that have been categorized based on linguistic characteristics and are not necessarily UMLS relations.

d. *Parser*: This module extracts, structures, and encodes phenotypic and genotypic entities and relations for tagged text from the previous module using a grammar and a lexicon to parse and structure the output, and a coding table to map the normalized output to ontological codes. The output is in an XML form based on a representational schema of the domain, called PGschema [21], which represents genotypic and phenotypic entities, their ontological codes, modifiers, and relations between the entities. Figure 1 shows an example of a simplified output form generated by BioMedLEE for the above tagged sentence, where some of the nested tags have been manually indented to facilitate readability of the output structure. This output differs from output generated by systems that use co-occurrence of terms because BioMedLEE found actual relations "suggest" and "associated with" in the text. The relation "suggest" connects "axonal transport of hd gene" with a second nested relation "associated with", whose first argument is "pathology" with an anatomical modifier "corticostriatal" and whose second argument is "Huntington's disease".

```
<relation v = "suggest">
  <bodyfunc v = "transport"><arg v = "1"></arg>
    <bodyloc v="axon"></bodyloc><cellcomp v ="N-terminal"></cellcomp>
    <gene_gproduct v = "UMLS: C1415504_hd gene" idref="p126">
    </gene_gproduct>
    <code v ="UMLS:C0004462_axonal transport"</code>
  </bodyfunc>
  <relation v = "associated with"><arg v = "2"></arg>
    <problem v = "pathology"><arg v = "1"></arg>
    <bodyloc v = "corticostriatal"></bodyloc>
    </problem>
    <problem v = "Huntington's disease"><arg v = "2"></arg>
      <code v = "UMLS:C0020179_huntington disease"></code>
    </problem>
  </relation>
</relation>
```

Fig. 1 Simplified XML output generated by BioMedLee for a sample sentence

2.2.2 SemRep Natural Language Processing System

SemRep [13] is a symbolic natural language processing system for identifying semantic predications in biomedical text. The current focus is on Medline citations. Linguistic processing is based on an underspecified (shallow) parse structure supported by the SPECIALIST Lexicon [22] and the MedPost part-of-speech tagger [23]. Medical domain knowledge is provided by the UMLS. Predications produced by SemRep consist of Metathesaurus concepts as arguments of a Semantic Network relation.

For this project, the most important relation is TREATS; however, SemRep identifies additional semantic predications representing various aspects of biomedicine. The core relations addressed refer to clinical actions (e.g. TREATS, PREVENTS, ADMINISTERED_TO, MANIFESTATION_OF) and organism characteristics (LOCATION_OF, PART_OF, PROCESS_OF). SemRep has recently been enhanced to address pharmacogenomics text [24]. Relations in this semantic area refer to substance interactions and pharmacologic effects (AFFECTS, CO-EXISTS_WITH, DISRUPTS, AUGMENTS, INTERACTS_WITH, INHIBITS, STIMULATES), as well as genetic etiology (ASSOCIATED_WITH, PREDISPOSES, CAUSES). The majority of SemRep's relations are drawn from the Semantic Network; however, several have been defined to extend the coverage of that ontology, including ADMINISTERED_TO, CO-EXISTS_WITH, and PREDISPOSES.

Each semantic relation serves as the predicate of an ontological predication that controls SemRep processing. The arguments in these predications are UMLS semantic types, such as 'Human' or 'Anatomical Structure', which can, for example, appear in the predication "Anatomical Structure PART_OF Human." All predications extracted from text by SemRep must conform to an ontological predication.

Semantic interpretation is based on the underspecified parse structure, in which simple noun phrases are enhanced with corresponding Metathesaurus concepts by

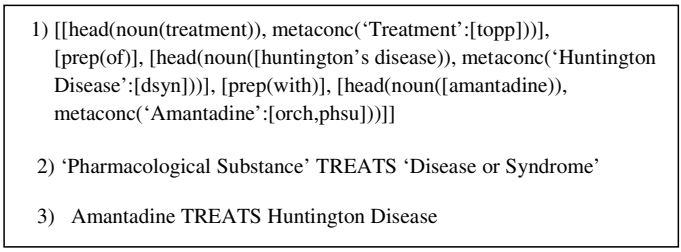

Fig. 2 SemRep processing of treatment of Huntington's disease with amantadine

MetaMap [16]. For example, processing of the phrase *treatment of Huntington's disease with amantadine* produces the structure seen in (1) in Fig. 2. The noun phrase *Huntington's disease* has been mapped to the concept "Huntington's disease," with semantic type 'Disease or Syndrome' (dsyn).

The parse structure enhanced with Metathesaurus concepts serves as the basis for the final phase in constructing a semantic predication. During this phase, SemRep applies "indicator" rules which map syntactic elements (such as verbs and nominalizations) to the predicate of an ontological predication. Argument identification rules (which take into account coordination, relativization, and negation) then find syntactically allowable noun phrases to serve as arguments for indicators. If an indicator and the noun phrases serving as its syntactic arguments can be interpreted as a semantic predication, the following condition must be met: The semantic types of the Metathesaurus concepts for the noun phrases must match the semantic types serving as arguments of the indicated ontological semantic predication. For example, in Fig. 2 *treatment* is an indicator for TREATS, with the corresponding ontological predication seen in (2) in Fig. 2. The concepts corresponding to the noun phrases *amantadine* and *Huntington's disease* can serve as arguments of TREATS because their semantic types ('Pharmacological Substance' (phsu) and 'Disease or Syndrome' (dsyn)) match those in the ontological predication. In the final interpretation, (3) in Fig. 2, the Metathesaurus concepts from the noun phrases are substituted for the semantic types in the ontological predication.

3 Methods

3.1 Discovery Patterns

3.1.1 The Relations Maybe_Treats1 and Maybe_Treats2

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. 4). In both forms the goal is to propose potential new treatments, and the two can work in concert: proposing either two different new treatments (complementarity) or the same treatment 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 discovery patterns are different in the way they generate new candidate treatments Z. The first form *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 concept Z. In other words, we first try to find the characteristics of a disease X with regard to a change in the level of substance or measurement Y in patients with this disease. Then we look for a drug or chemical Z that can cause an opposite change in the same substance or measurement Y. That is, if the Y concept decreases in association with the X disease, we expect it to increase in association with the Z drug, or vice versa. An example of the first form is the reasoning used by Swanson to propose fish oil (Z) as a new treatment for Raynaud's disease (X). Fish oil (Z) was proposed because it reduces blood viscosity (Y) which was reported in the literature to be increased in patients with Raynaud's.

In using *Maybe_Treats2* to find a potential new treatment for a starting disease X we first search for another disease X2 that has characteristics similar to X (Y2 substance or function is either increased or decreased in both X and X2). Then we propose as a new treatment for disease X the drug (Z2) already used to treat 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 (type 1). Therefore, treatments for diabetes might also be used for HD.

We can formally define the two forms of the *Maybe_Treats* discovery pattern using the predications in Figs. 3 and 4.

3.1.2 The Relations Associated_with_change and Treats

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 in which 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

```
Maybe_Treats(Drug_Z, Disease_X)  IF
        Maybe_Treats1(Drug_Z, Disease_X)  OR
        Maybe_Treats2(Drug_Z, Disease_X).

Maybe_Treats1(Drug_Z, Disease_X) IF
        Associated_with_change(Disease_X,Subst_Y,Change_Y11) AND
        Associated_with_change (Drug_Z, Subst_Y,Change_Y12)    AND
        Opposite_Change(Change_Y11, Change_Y12)      AND
        NOT Treats(Drug_Z, Disease_X).
Opposite_Change("Increase", "Decrease").
Opposite_Change("Decrease", "Increase").

Maybe_Treats2(Drug_Z2, Disease_X) IF
        Associated_with_change (Disease_X,Subst_Y2,Change_Y21)  AND
        Associated_with_change (Disease_X2,Subst_Y2,Change_Y22) AND
        Same_Change(Change_Y21,Change_Y22)  AND
        Treats(Drug_Z2, Disease_X2) AND
        NOT Treats(Drug_Z2, Disease_X).
Same_Change("Increase", "Increase").
Same_Change("Decrease", "Decrease").
```

Fig. 3 Formal definition of the discovery pattern *Maybe_Treats*

from the list of drugs or chemicals that have not been used, 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.

The relation *Associated_with_change* is a higher level relation and is based on basic BioMedLee relations. In this research, we used three methods to derive *Associated_with_change* where the first two are the most credible. The first is based on the binary *Increase* or *Decrease* relations. For example, for the sentence "*Speech production increases cerebral blood flow* in HD patients", BioMedLEE extracts *Increase(Speech production, cerebral blood flow)*. In this example, although the binary relation associated with "increase" was extracted, the relation "in HD patients" was lost because BioMedLEE did not recognize that the abbreviation HD referred to Huntington's disease.

The second method is to use binary relations in which one of the arguments has a change such as *Increase* or *Decrease* associated directly with the argument. The relation can be any of those that indicate some kind of an association between its arguments, such as *associated_with, exhibited, due to, suggest, results from*. For example, from the sentence "*Huntington's disease* brains all *exhibited* a marked *decrease* in *substance P* fiber density in the substantia nigra and globus pallidus" BioMedLee extracts *Exibit(Huntington disease, Substance P/decrease)*.

The third way to derive *Associated_with_change* relations is to exploit phrase or sentence level co-occurrence of concepts with which a change is associated with one of the concepts. In other words, we extract all the concepts from a phrase or sentence and if there is at least one concept with a change directly associated with it, we then assume that that concept is related to the other concepts in the same phrase or sentence. Obviously, this is the least credible way of deriving *Associated_with_change* relations; however, it significantly improves recall. For example,

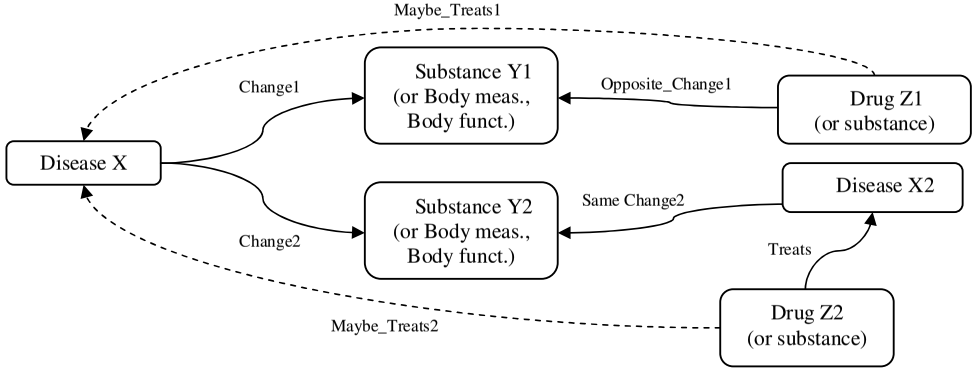

Fig. 4 Discovery pattern *Maybe_Treats*. *Maybe_Treats1* proposes Z1 (drug or substance) as a new treatment for disease X because Z1 causes opposite change to Y1 (function or substance) and the change of Y1 is a characteristic of disease X. *Maybe_Treats2* proposes Z2 as a new treatment of X because there is a similar disease, X2, and drug Z2 is known to treat X2

from the sentence "In *Huntington's disease*, there is a *decrease* of the *neuropeptides*, substance P, enkephalins, and cholecystokinin in the striatonigral system, whereas in *Parkinson's disease* an *increase* of substance P is found in the substantia nigra", BioMedLee extracts *co_occurs(Huntington's disease, Neuropeptides/decrease)* which is correct, but from the same sentence the system also extracts *co_occurs (Parkinson's disease/increase, Huntington's disease)*, which is not correct.

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, the pattern will try to generate diseases X that might be treated. If a disease X is provided as input, the pattern will try to generate drug Z that might be used to treat the disease X. If both a disease X and a drug Z are provided as input, 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.

3.2 Integrated BioMedLEE and SemRep Output Format

The output formats normally provided by BioMedLee and SemRep are different from each other, and therefore it was not straightforward to combine the use of both systems. To enable the integration of the output of the two systems for the purpose of this research, we developed a *common output format*, the specification of which is still evolving. Currently, the common format contains three types of lines: *text*, *entity* and *relation*. Each type of line is a delimited list of fields. The input to both systems is a set of Medline citations. Each Medline citation is broken into a sequence of sentences and each sentence is processed separately. For each sentence, a line of type *text* is first generated to present the actual text of the current sentence. Then a line of type *entity* is generated for each biomedical entity (concept) extracted from the current sentence regardless of whether the entity is part of a relation or not. Finally, all the relations between the entities from the current sentence are generated as lines of type *relation*.

Table 1 shows the fields used in the common format. All three types of lines start with fields 1–6. The first field is the system identification to indicate which system generated the line, the second is the PubMed identification number, followed by the subsection abbreviation, which indicates whether the sentence comes from the methods, conclusions, results or some other subsection of a structured abstract. The fourth field specifies whether the sentence is from the title or the abstract. The fifth field specifies the sentence identification, which is slightly different for each system because different methods are used to recognize sentence boundaries. The sixth field identifies the row type, which is one of *text*, *entity* or *relation*. This field determines the format of the rest of the line.

For a line of type *text*, the next field is the actual text of the sentence, which for BioMedLee is in a tagged text format where the tags are linked to the entities,

Table 1 The common format used to represent BioMedLee and SemRep output. There are three types of lines: text, entity and relation. The first six fields are used by all three types of lines. The next fields are specific for each line type

Field number	Description (*example value*)
1	BL (*BioMedLEE*) or SE (*SemRep*)
2	PubMed ID
3	Subsection of abstract (*objective*, *results*)
4	Section of abstract ti(*title*) or ab (*abstract*)
5	Sentence id
6	Line type, one of: *'text'*, *'entity'*, *'relation'*
7. Text	Sentence text
If line type is 'entity' then next fields	
7. Entity	Entity type (*T047* or *disease*)
8. Entity	CUI
9. Entity	Preferred name
10. Entity	Change term (*increase*)
11. Entity	Degree term (*low*)
12. Entity	Negation (*not*)
13. Entity	MetaMap score
14. Entity	Begin character or phrase position
15. Entity	End character position of matched phrase
If line type is 'relation' then next fields	
7–15. Relation	Argument1 related fields
16. Relation	Name of relation (*treat*, *increase*)
17. Relation	Negation of explicit relation or empty
18. Relation	Begin character or phrase position of relation indicator
19. Relation	End character position of relation indicator
20–28. Relation	Argument2 related fields

relations, and modifiers, and for SemRep is plain text. For an *entity* type of line, there are fields specifying the type of entity, UMLS CUI (Concept Unique Identifier), preferred entity name, change and degree of associated change, location of the entity, MetaMap score, and location of the entity in the actual text (start and end position).

For a line of type *relation*, fields 7–15 describe the first argument of the relation in the same format as *entity* line; subsequent fields describe the semantic relation, including the name of the relation, whether it is negated or not, and the start and end positions of the relation in the text. Finally, the second argument of the relation is described in the same way as the first argument in fields 20–28. The specification of the arguments of the relations is currently redundant for ease of experimentation. At a subsequent stage the entities and relations will be associated with identifiers and then arguments of the relations will just be identifiers.

Some of the fields in the common format are specific for only one system, in which case the other system leaves these fields empty. Sometimes the two systems fill a particular field in a different way or format. For example, SemRep uses UMLS semantic types as entity type and BioMedLee uses its own types. BioMedLee

identifies the part of the actual text as a phrase identifier within a tagged text format while SemRep uses start and end character positions within a plain text string.

Some of the results presented here were obtained by directly processing the common output format by Unix shell scripts and Perl scripts. Some of the results were produced using SQL statements after the common format output generated by BioMedLee and SemRep was postprocessed with Perl scripts and loaded into a relational database management system.

4 Results

In this section we first replicate Swanson's Raynaud's discovery using the *Maybe_Treats1* discovery pattern. Then we present two hypothetically new therapeutic approaches: one for Huntington disease, based on the *Maybe_Treats1* discovery pattern and one for Parkinson's disease, based on *Maybe_Treats2*. Although we have not done a formal evaluation of our approach, at the end of this section we show evaluation results for the two important components of our methodology, BioMedLee and SemRep.

4.1 Rediscovering Fish Oil for Raynaud's Disease

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 2) 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 2). 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.

Table 2 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*

Number	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, ...

4.2 Insulin for Huntington Disease

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, behavioral changes, involuntary choreiform movements and cognitive impairments. Onset is most common in adulthood, with a typical duration of 15–20 years before premature death. No successful treatment is currently available. We constructed the set of all 5,511 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 2,139 were *Treats* relations. Of these, 740 Treats relations contained Huntington disease as an argument. These represent current treatments for Huntington (example 2 in Table 2).

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 18,360 relations, of which 1,912 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 than assumed that drugs and treatments which are successfully used to treat diseases associated with the same changes in substances and body functions as in HD would be potential new treatments for HD.

By using this approach 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 2). Although insulin has been attempted for immediate relief of one of the symptoms (chorea) of HD [25], we have not found research on insulin as a general treatment for this disease.

It is known that HD patients develop diabetes mellitus about seven times more often than matched healthy control individuals [26]. 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 [27].

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.

4.3 Gabapentin for Parkinson's Disease

This example illustrating the *Maybe_Treats1* pattern for Parkinson's disease uses the same set of articles used for *Maybe_Treats2* above. We selected Parkinson's disease as a starting concept in a modified version of Bitola which integrates co-occurrence based association rules with semantic relations extracted by BioMedLee and SemRep. This version of Bitola is in early development phase and is not yet publicly available.

In order to find potential therapies for the disease, our discovery strategy was first to identify Y concepts (Neuroreactive Substance or Biogenic Amine or Biologically Active Substance), characterized by a "decrease" of some substance in Parkinson's disease and in the second step to find all Z concepts (pharmacological substances) with the opposite change. We limited Y concepts by "change" and got five different concepts. Two of them, *levodopa* and *dopamine* are the mainstream of therapy for decades. The next two of the concepts, *Hommovanilic acid* and *Substance P*, were not selected due to inappropriate context of the relations. A relevant relation

was identified in the following sentence: "Postmortem brain studies indicate that patients with *Parkinson's disease* have *decreased* basal ganglia *gamma-aminobutyric acid* function in addition to profound striatal dopamine deficiencies." for *gamma-aminobutyric acid (GABA)*.

In the second step we searched for all Z concepts (pharmacological substance) characterized by an "opposite change". Six substances, all antiepileptics, were identified which were related to GABA in an appropriate way: *gabapentin*, *Vigabatrin*, *Tiagabine* and *Topiramate*, *methamphetamine* and *milacemide* through the following sentences: "*Gabapentin*, probably through the activation of glutamic acid decarboxylase, *leads to* the *increase* in synaptic *GABA*", "*GVG (Vigabatrin) caused* a significant *increase* in *GABA* release, even at concentrations as low as 25 μM", "*Tiagabine* is an antiepileptic drug, which *increases GABA* via selective blockade of GABA reuptake", "*Topiramate increased* brain *GABA*, homocarnosine, and pyrrolidinone to levels that could contribute to its potent antiepileptic action in patients with complex partial seizures." "These results support the hypothesis that long-term administration of *methamphetamine increases* the activity of the striatonigral *GABA* system and thereby reduces the sensitivity of postsynaptic GABA receptors in the SNR." And "The results show that *milacemide increases* the *GABA* content in the GABA pool which is associated with the striatonigral neurons."

GABA is ubiquitous in the nervous system and regarded widely as the principal inhibitory neurotransmitter of the brain. It is also considered as one of the principal vehicles for inhibition in Parkinson's disease. Furthermore, production of inhibitory transmitter GABA in the subthalamic nucleus (STN), suppressing the hyperactive STN, is considered as one of the strategies for gene therapy in the treatment of Parkinson's disease.

In this way we identified selected antiepileptics as a possible therapy for Parkinson's disease. Indeed, some potential benefit of Gabapentin and Toprimate in treatment of Parkinson's disease has been already mentioned in the literature [32, 33].

4.4 Evaluation of BioMedLee and SemRep

Although BioMedLEE has not yet been evaluation for use in LBD, it has been evaluated for two different applications. In Lussier [14], BioMedLEE was combined with a phenotypic ontological organizing system, PhenOS, to create a new system called PhenoGO. PhenoGO associates contextual information with GOA annotations [28] by adding phenotypic information to the protein and GO pairs specified in GOA. The overall PhenoGO system was evaluated for extracting and coding anatomical and cellular information associated with the pairs and for assigning the code to the correct pairs. The results of the evaluation demonstrated that PhenoGO has a precision of 91% and a recall of 92%. Although the results have been computed for the entire PhenoGO system and not for BioMedLEE separately, the high performance of PhenoGO is an indicator of the performance of BioMedLEE because the

relations among the genes, GO terms, and phenotypes were determined based on BioMedLEE.

In Borlawsky [29], BioMedLEE was used for a clinical application geared to facilitating clinical practice using Evidence-Based Medicine (EBM). This involved extracting and coding disease, therapy, and drug concepts and their relations from textual sections of Cochrane Reviews, the best standard for obtaining evidence-based medicine. Although BioMedLEE was designed for capturing phenotypic and genotypic relations and not designed for clinical applications or processing of Cochrane Reviews, the study showed that the pertinent information could be extracted and correlated with an overall recall of 80.3% and precision of 75.2%. The most frequent cause of error was due to differences in the semantic classification assigned by BioMedLEE and by the expert, who manually coded the information. For example, the expert manually parsed 'hearing loss' as a problem, but the NLP engine alternatively parsed the phrase as a compositional phrase consisting of a process *hearing* with a change modifier *loss*, which is also correct. Thus, it is likely that performance can be increased by expanding the guidelines to permit certain variations in semantic categorization between the expert and system and by refining the system specifically for the clinical domain, which is not as broad as the complete biomedical domain.

The effectiveness of SemRep in extracting semantic predications from biomedical text has been evaluated in several contexts [24, 30, 31]. In two of these [30, 31], accuracy was assessed after the predications had been subjected to an automatic summarization algorithm. In [30], 306 predications (for predicates ISA, CAUSES, CO-OCCURS_WITH, LOCATION_OF, OCCURS_IN, TREATS) extracted from 1,200 Medline citations were evaluated. Of these, 203 predications were determined to be correct (66% precision). In [eval2], for predicates AFFECTS, CAUSES, COMPLICATES, DISRUPTS, INTERACTS_WITH, ISA, PREVENTS, and TREATS, 148 of 189 predications extracted from 130 Medline citations were judged as correct (78% precision). SemRep was tested for both recall and precision in [24], using a gold standard of 300 sentences randomly generated from 36,577 sentences drawn from a set of Medline citations containing drug and gene co-occurrences. In addition to the predicates addressed in the first two evaluations, predications having such predicates as INHIBITS, STIMULATES, and DISRUPTS were also assessed. SemRep extracted 623 predications from the 300 sentences in the test collection. Of these, 455 were true positives, 168 were false positives, and 375 were false negatives, reflecting recall of 55% (95% confidence interval 49–61%) and precision of 73% (95% confidence interval 65–81%).

5 Discussion and Further Work

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. We presented some general performance evaluation of semantic relation extraction, but in further work we plan to evaluate specifically the extraction of *Associated_with_change* and *Treats*, which are the most important relations in our method. We also plan to evaluate the performance of the overall LDB method. 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 better integrate the BITOLA LBD system with SemRep and BioMedLee. Currently, the user has to run the three systems separately and the output is combined with various scripts in a way which is not very user-friendly.

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. We plan to combine BioMedLee's change detection with SemRep's relations in order to obtain a larger number of binary relations with a change. Namely, SemRep may find a binary relation whereas BioMedLEE may not, but BioMedLEE may have found a change in one of the arguments of the relation that SemRep found. Currently we have a large number of unary change relations which are not associated directly with another concept. Another way to improve the extraction of change relations is by analyzing the cases in which the change was not captured and creating better extraction rules.

Yet another research contribution is the notion of a *discovery pattern* which is based on semantic relations and allows more precise hypothesis generation. Here we have presented one such pattern, *Maybe_Treats*, but we plan to develop other discovery patterns as well.

We plan to develop a user-friendly web-based interface which will allow public access to our methodology. It should allow among other things ranking of potentially new discoveries based on a heuristic ranking procedure not yet developed.

6 Conclusions

Literature-based discovery (LBD) is a method for automatically generating hypotheses from the research literature. Currently LBD systems depend exclusively on co-occurrence based methods for finding relations between concepts. We presented a new method aimed at improving LBD. It is based on semantic predications, which are extracted from text using the combined results of two natural language processing systems. Additionally, the change associated with the arguments of the predications, is also extracted. We also introduced the notion of a *discovery pattern*. 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.

Using our methodology we successfully replicated Swanson's Raynaud's – fish oil discovery. Furthermore, we generated some interesting potentially new therapeutic approaches for Huntington disease and for Parkinson's disease.

We believe that the future of literature-based discovery lies in developing specific discovery patterns for particular discovery tasks based on semantic relations further integrated with co-occurrence-based approaches.

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 **30** (1986) 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., Peterlin, B., Mitchell, J.A., Humphrey, S.M.: Using literature-based discovery to identify disease candidate genes. Int J Med Inform **74** (2005) 289–298
4. Hristovski, D., Stare, J., Peterlin, B., Dzeroski, S.: Supporting discovery in medicine by association rule mining in Medline and UMLS. Medinfo **10** (2001) 1344–1348
5. Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., de Jong-van den Berg, L.T., Vos, R.: Text-based discovery in biomedicine: the architecture of the DAD-system. Proc AMIA Symp (2000) 903–907
6. Gordon, M.D., Lindsay, R.K.: 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 **47** (1996) 116–128
7. Gordon, M.D., Dumais, S.: Using latent semantic indexing for literature based discovery. J Am Soc Inf Sci **49** (1998) 674–685
8. Wren, J.D.: Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics **5** (2004) 145
9. Pratt, W., Yetisgen-Yildiz, M.: LitLinker: capturing connections across the biomedical literature. In Proceedings of the 2nd International Conference on Knowledge Capture. ACM Press, Sanibel Island, FL, USA (2003)
10. Fuller, S.S., Revere, D., Bugni, P.F., Martin, G.M.: A knowledgebase system to enhance scientific discovery: Telemakus. Biomed Digit Libr **1** (2004) 2
11. Hu, X.: Mining novel connections from large online digital library using biomedical ontologies. Libr Manage **26** (2005) 261–270
12. Srinivasan, P., Libbus, B.: Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics **20 Suppl 1** (2004) I290–I296
13. Rindflesch, T.C., Fiszman, M.: The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform **36** (2003) 462–477
14. Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y., Friedman, C.: PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. Pac Symp Biocomput (2006). pp. 64–75
15. Weeber, M., Kors, J.A., Mons, B.: Online tools to support literature-based discovery in the life sciences. Brief Bioinform **6** (2005) 277–286

16. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp (2001) 17–21
17. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of association rules. In: Fayyad, U. (ed.): Advances in Knowledge Discovery and Data mining. MIT Press, Cambridge, MA (1996), pp. 307–328
18. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., Johnson, S.B.: A general natural-language text processor for clinical radiology. J Am Med Inform Assoc **1** (1994) 161–174
19. Friedman, C., Shagina, L., Lussier, Y., Hripcsak, G.: Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc **11** (2004) 392–402
20. Humphreys, B.L., Lindberg, D.A., Schoolman, H.M., Barnett, G.O.: The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc **5** (1998) 1–11
21. Friedman, C., Borlawsky, T., Shagina, L., Xing, H.R., Lussier, Y.A.: Bio-ontology and text: bridging the modeling gap. Bioinformatics **22** (2006) 2421–2429
22. McCray, A.T., Srinivasan, S., Browne, A.C.: Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care (1994) 235–239
23. Smith, L., Rindflesch, T., Wilbur, W.J.: MedPost: a part-of-speech tagger for biomedical text. Bioinformatics **20** (2004) 2320–2321
24. Ahlers, C., Fiszman, M., Demner-Fushman, D., Lang, F.-M., Thomas, C.R.: Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput (2007) 209–220
25. Quinn, N.P., Lang, A.E., Marsden, C.D.: Insulin-induced hypoglycaemia does not abolish chorea. J Neurol Neurosurg Psychiatry **45** (1982) 1169–1170
26. Ristow, M.: Neurodegenerative disorders associated with diabetes mellitus. J Mol Med **82** (2004) 510–529
27. Andreassen, O.A., Dedeoglu, A., Stanojevic, V., Hughes, D.B., Browne, S.E., Leech, C.A., Ferrante, R.J., Habener, J.F., Beal, M.F., Thomas, M.K.: Huntington's disease of the endocrine pancreas: insulin deficiency and diabetes mellitus due to impaired insulin gene expression. Neurobiol Dis **11** (2002) 410–424
28. Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R.: The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res **32** (2004) D262–D266
29. Borlawsky, T., Friedman, C., Lussier, Y.: Generating executable knowledge for evidence-based medicine using natural language and semantic processing. AMIA Annu Symp Proc (2006)
30. Fiszman, M., Rindflesch, T.C., Kilicoglu, H.: Abstraction summarization for managing the biomedical research literature. Proc HLTNAACL Workshop on Computational Lexical Semantics (2004) 76–83
31. Fiszman, M., Rindflesch, T., Kilicoglu, H.: Summarizing drug information in Medline citations. Proc AMIA Annu Symp (2006)
32. Van Blercom, N., Lasa, A., Verger, K., Masramón, X., Sastre, V.M., Linazasoro, G: Effects of gabapentin on the motor response to levodopa: a double-blind, placebo-controlled, crossover study in patients with complicated Parkinson disease. Clin Neuropharmacol **27** (2004) 124–128
33. Silverdale, M.A., Nicholson, S.L., Crossman, A.R., Brotchie, J.M.: Topiramate reduces levodopa-induced dyskinesia in the MPTP-lesioned marmoset model of Parkinson's disease. Mov Disord **20** (2005) 403–409