DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–15
doi: 10.1093/database/baw025
Original article

Original article

Discovering biomedical semantic relations in PubMed queries for information retrieval and database curation

Chung-Chi Huang and Zhiyong Lu*

National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA

*Corresponding author: Tel: (301) 594-7089; Fax: (301) 480-2290; Email: Zhiyong.lu@nih.gov

Citation details: Huang,C.-C. and Lu,Z. Discovering biomedical semantic relations in PubMed queries for information retrieval and database curation. *Database* (2016) Vol. 2016: article ID baw025; doi:10.1093/database/baw025

Received 6 November 2015; Revised 5 February 2016; Accepted 14 February 2016

Abstract

Identifying relevant papers from the literature is a common task in biocuration. Most current biomedical literature search systems primarily rely on matching user keywords. Semantic search, on the other hand, seeks to improve search accuracy by understanding the entities and contextual relations in user keywords. However, past research has mostly focused on semantically identifying biological entities (e.g. chemicals, diseases and genes) with little effort on discovering semantic relations. In this work, we aim to discover biomedical semantic relations in PubMed queries in an automated and unsupervised fashion. Specifically, we focus on extracting and understanding the contextual information (or context patterns) that is used by PubMed users to represent semantic relations between entities such as '*CHEMICAL-1 compared to CHEMICAL-2*.' With the advances in automatic named entity recognition, we first tag entities in PubMed queries and then use tagged entities as knowledge to recognize pattern semantics. More specifically, we transform PubMed queries into context patterns involving participating entities, which are subsequently projected to latent topics via latent semantic analysis (LSA) to avoid the data sparseness and specificity issues. Finally, we mine semantically similar contextual patterns or semantic relations based on LSA topic distributions. Our two separate evaluation experiments of chemical-chemical (*CC*) and chemical–disease (*CD*) relations show that the proposed approach significantly outperforms a baseline method, which simply measures pattern semantics by similarity in participating entities. The highest performance achieved by our approach is nearly 0.9 and 0.85 respectively for the *CC* and *CD task* when compared against the ground truth in terms of normalized discounted cumulative gain (nDCG), a standard measure of ranking quality. These results suggest that our approach can effectively identify and return related semantic patterns in a ranked order covering diverse bio-entity relations. To assess the potential utility of our automated top-ranked patterns of a given relation in semantic search, we performed a pilot study on frequently sought semantic relations in PubMed and observed improved literature retrieval effectiveness based on post-hoc human relevance evaluation. Further investigation in larger tests and in real-world scenarios is warranted.

Introduction

Many natural language queries are submitted to search engines on the Web every day, and an increasing number of online search engines target domain-specific search services. For example, Yelp (www.yelp.com) facilitates restaurant searching while PubMed (www.ncbi.nlm.nih.gov/pubmed) retrieves scholarly publications in biomedicine.

Today's search engines typically treat natural language queries as lists of terms and retrieve documents containing those terms. However, documents with different words but similar semantics may be overlooked. Take the search engine in biomedical domain, PubMed (1), for example. Semantically similar as the queries *chlorthalidone vs hydrochlorothiazide* and *chlorthalidone versus hydrochlorothiazide* are, PubMed returns 2.5 times more relevant articles when users compare these two drugs using *versus* than using *vs*. Such performance difference in retrieval effectiveness may be reduced and/or the levels of user satisfaction may be maintained if queries of similar semantic meaning were presented at search time. In this regard, this paper learns to discover semantic relations between bio-concepts (such as chemicals and diseases) on the Web for possible help of biocuration and retrieval effectiveness. Specifically, this paper aims to identify semantically similar context words (like the *vs* and *versus* example), referred to as context patterns thereafter, in PubMed queries that assert specific relations between two entities. We focus on semantically understanding PubMed queries with exactly two bio-entities as bio-NLP research in entity relations has long focused on relations between dual entities: chemical–disease relations (2), protein-protein interaction (3), gene events (4), drug-drug interaction (5) and disease co-morbidities (6).

We present a novel unsupervised framework, SIP (semantically similar pattern finder), that discovers two-argument context patterns that are semantically similar but lexically different. Table 1 shows example SIP discovery of synonymous context patterns associated with semantic bio-relations involving chemicals/drugs (denoted as *#C*) and diseases (denoted as *#D*). SIP leverages the semantic information of biological entities in Web queries to differentiate pattern semantics, based on observations that semantically similar patterns such as *#C induced #D* and *#D due to #C* share significantly more chemical and disease pairs among Web queries than patterns like *#C induced #D* and *treatment of #D with #C* which are not semantically similar. Intuitively, SIP estimates patterns' semantic similarity by their distributional similarity, whether their distributional contexts are participating entities or semantic topics. In specific, the SIP framework discovers patterns of similar semantics in three main steps. First, it determines patterns' participating entities which constitute entity space. Next, SIP transforms entity space into latent topic space for pattern semantics analysis/understanding. It learns the transformation by analyzing PubMed queries using latent semantic analysis (LSA). Finally, SIP yields pattern pairs with high distributional similarity in LSA topics and proposes them as semantically similar patterns.

Our SIP framework is unique as it targets biomedical queries, gaining importance in Web searches and biomedical research (1, 7). Second, SIP leverages search crowds' wisdom (i.e. user entities in Web queries) to discern context patterns' semantics and estimate patterns' semantic similarity. This makes SIP unsurpervised requiring no training/seed data for related pattern discovery. Third, SIP serves as one of the pioneering work to analyze pattern semantics based on real-world user queries in either NLP or bio-NLP community. Last but not least, SIP exploits LSA to project entities in queries into lower-dimension latent topics, avoiding specificity in entity mentions, and SIP transforms the problem of finding semantically similar patterns into one of finding patterns with distributional similarity in LSA topics.

The results of our work can benefit biocuration and semantic information retrieval. For example, the automated semantically similar patterns can be used by biocurators for

Table 1. Example SIP synonymous context patterns

Semantic relation	Context patterns in user queries
Drug comparison	*#C vs #C*, *#C compare #C*, *#C compare to #C*, *#C comparison #C*, *#C versus #C*, *comparison between #C and #C*, ...
Drug combination	*combine #C and #C*, *#C and #C combination*, *#C in combination with #C*, *#C with #C*, ...
Chemical cause disease	*#C cause #D*, *#D after #C*, *#D due to #C*, *#D from #C*, *#C induce #D*, *#D induce by #C*, *#D associate with #C*, ...
Chemical treat disease	*#D treatment #C*, *treatment of #D with #C*, *#C treatment for #D*, *#D treat with #C*, *#D therapy #C*, ...

Pattern words are stemmed.

assisting bio-relation curation and article triaging (e.g. (8)), or can be passed on to search engines to expand search results for better recall of relevant documents (e.g. (9)). This paper focuses on discovering semantically similar patterns and its evaluation, together with its real-life applications in two use cases (see Application Section for more details).

Related work

Curating relationships between biological entities and concepts is an active task carried out by many groups such as CTD (gene–disease–chemical) (10), BioGrid (protein-protein interaction) (11) and PharmaGKB (drug-gene) (12). The proposed work could potentially contribute to improved curation quality and productivity in two main ways: a) our discovered patterns could be directly used by curators to locate relevant papers more effectively (i.e. with better coverage and precision) in their routine literature search; and b) our patterns could be integrated into automated text-mining systems for assisting relation curation.

Semantic search, or searching with semantics, has been an area of active research for improving keyword-based retrieval systems by taking semantics into account. Semantics of the documents to be searched or semantics of the search terms may be leveraged in the process. In biomedicine, understanding the semantics of user queries has received much attention since (13, 14). For instance, (15) analyzes query length, query specificity and query clarity of TREC and CLEF shared tasks. Another interesting work (16) imposes position constraint on search terms in retrieved documents. Such in-proximity constraint aims to preserve semantic relations of search terms in multi-word queries. Moreover, past research has studied the effectiveness of semantically expanding queries on biological entities, concepts, or controlled vocabulary for improved retrieval performance (17, 18). Following this line of trend and term disambiguation (19), here we aim to understand the semantics of biomedical queries on a deeper level than individual concepts, but in the form of context patterns and entity relations.

In contrast to the previous work, we are the first to examine the applicability of LSA in query/pattern semantics and to discover semantically similar context patterns in user queries, inspired by the success of using LSA for lexical similarity estimation (20). Furthermore, compared to (21)'s single drug side-effect pattern recognition, we automatically discover bio-relational patterns related to diverse semantics of #*C compared with* #*C*, #*C in combination with* #*C*, #*C* #*C interaction*, #*C induced* #*D*, *treatment of* #*D with* #*C*, #*D* #*C deficiency*, *dietary* #*C and* #*D*, etc. simply by using bio-entities in PubMed queries as knowledge. The unsupervised nature of our framework makes it highly scalable: needing no seeds, it can easily be extended to cover various entity types (e.g. genes) and to understand the semantics of corresponding relations (e.g. #*G responsible for* #*D* where #*G* denotes genes).

The unsupervised SIP framework

Problem statement

We now formally state the problem that we are addressing: We are given a collection of PubMed queries QL and a context pattern p that specifies a biological relationship between two entities. Our goal is to automatically discover a reasonable-sized set of patterns in QL that are semantically similar to p in biomedical search context. For this, we represent queries in QL as context patterns in entity space and project such representations into latent topic space using LSA, such that patterns' semantic similarity can be estimated by their distributional similarity among LSA latent topics and those patterns having high LSA topic similarity with p can be proposed as its paraphrases. Figure 1 summarizes the workflow of our method while Figure 2 elaborates on semantically similar patterns identification at run-time. Detailed process is discussed in the following sections.

Transforming spaces

We propose to address the problem of finding semantically similar context patterns in an unsupervised manner by finding patterns with high distributional similarity in LSA-learned latent topics. Figure 1 outlines the procedure to transform PubMed queries into patterns in entity space and LSA space for this purpose. Algorithm 1 shows the corresponding steps. Note that we consider SIP unsupervised in that SIP does not require any training/seed data for pattern semantics understanding.

In the first step, we perform stemming and named entity recognition on PubMed queries QL. We use (22) to stem query words (e.g. reduce third-person singular verb 'induces' to the base form 'induce' and plural noun 'differences' to singular 'difference') for pattern analysis. We then use tmChem (23), DNorm (24) and GNormPlus (25) to recognize chemical/drug, disease/disorder and gene/protein in queries, respectively. These are state-of-the-art entity recognition tools that are publicly available (http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/). Although different text genres may lead to different performance, they in general can achieve 0.8–0.9 in F-measure based on previous benchmarking evaluations (2). Sample stemmed and semantically tagged queries are shown in Table 2 where <*X*> denotes the start of an entity while </*X*> the end, and in our paper *X* can be *C*, *D* and *G* which respectively correspond to a chemical, disease and gene entity. Note that our bio-entities are identified in a greedy fashion with priority given to longer text spans.

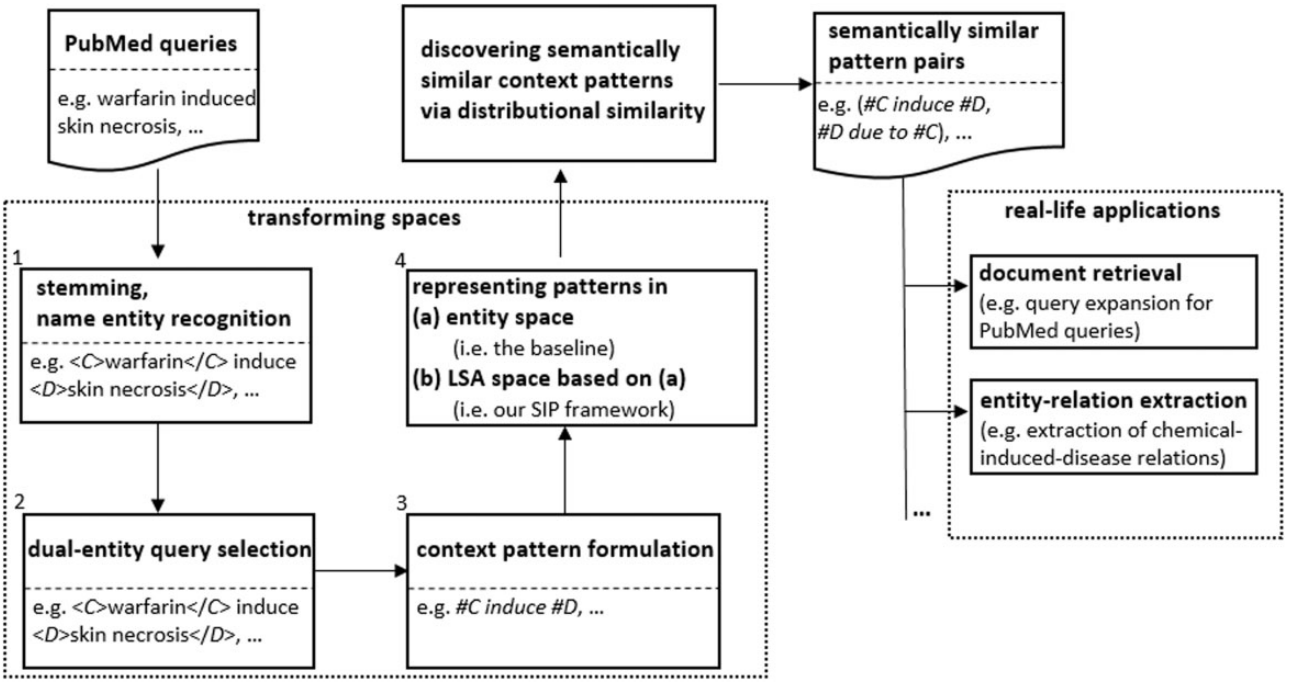

Figure 1. Workflow and application of SIP.

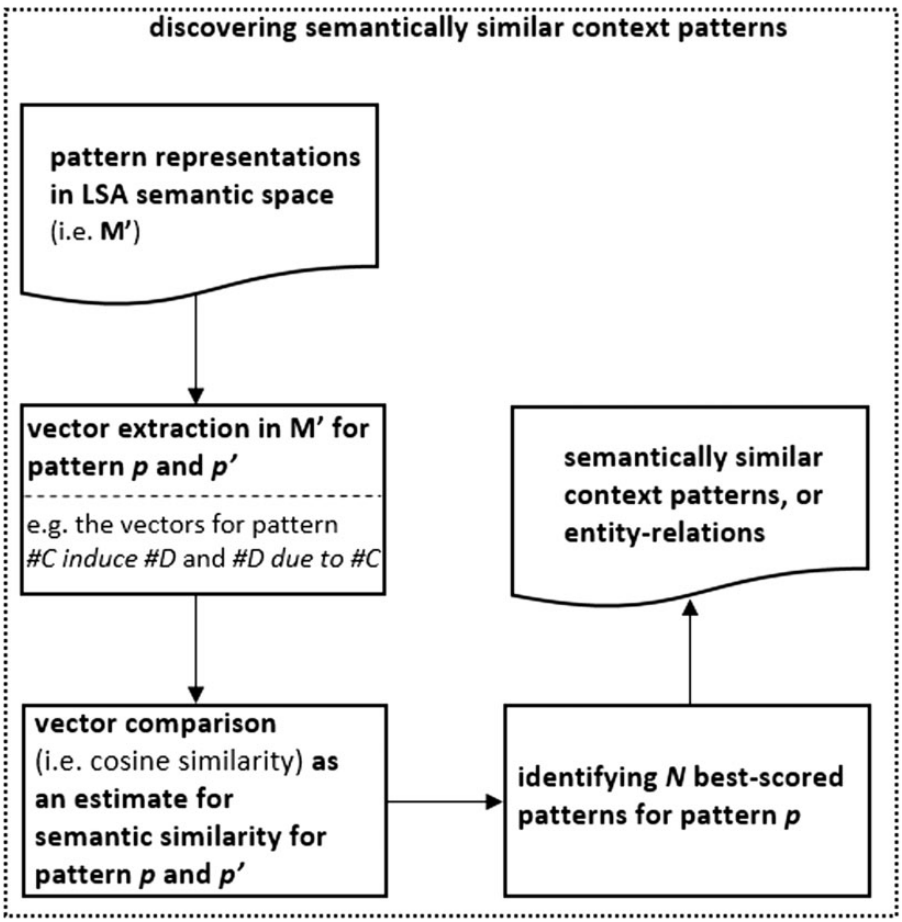

Figure 2. Semantically similar pattern finding.

Algorithm 1. Space transformation.
(1) Stem queries in *QL* and locate entities in queries
(2) Identify and collect queries with dual entities into *QL′*
(3) Formulate dual-entity queries in *QL′* into context patterns
(4a) Transform pattern strings into patterns in the space of their participating entities and represent patterns in entity space as matrix **M**
(4b) Transform patterns in entity space, **M**, into patterns in LSA space by dimension reduction from the number of entity pairs to that of LSA latent topics

Step 2 of Figure 1 collects queries with exactly two entities into QL'. In contrast to using entity seeds for pattern recognition (21), unsupervised SIP leverages participating entity pairs in user queries to semantically constrain the 'contexts' of the queries' non-entity words (i.e. Step 4a and 4b), thus understanding the semantic relations between entities. The wisdom of search crowds and searchers' perception, encoded in search queries, are also valued in (26, 27), and our experiments in Experiments Section suggest user entity pairs in Web queries serve as good knowledge to capture query/pattern semantics.

In the third step, dual-entity queries in QL' are formulated and mapped to distinct context patterns. This is done based on recognized named entity types. For instance, semantically tagged query *chlorthalidone*</C> *vs* <C> *hydrochlorothiazide*</C> becomes pattern #C *vs* #C (see Table 2). Note that this paper focuses on patterns (a) involving two chemicals (e.g. #C *vs* #C) and (b) between a chemical and a disease (e.g. #D *due to* #C). Hereafter, we denote the former *CC task*, discovering semantically similar chemical–chemical patterns, and the latter *CD task*, discovering those of chemical–disease.

Inspired by distributional similarity (28–30), Step 4a learns a pattern's semantics by its contextual/participating entities in PubMed queries, i.e. entity space. For example, the pattern #D *associate with* #C is distributionally and semantically associated with a set of disease–chemical entity pairs in the query log: <*skin necrosis*, *warfarin*>, <*myocardial infarction*, *isoproterenol*>, <*intraoperative floppy iris syndrome*, *tamsulosin*>, etc. We use matrix $\mathbf{M}_{i\times j}$ to represent our context patterns in entity space where i denotes the number of unique patterns and j the number of unique co-occurring entity pairs. Matrix element $\mathbf{M}[x,y]$ verifies the reference of the entity pair y in the pattern x in QL': value 1 indicates the reference exists, 0 otherwise. Our *CC/CD task* has its own **M**, ensuring subsequent LSA transformation and semantically similar pattern finding are confined to a specific entity type pair. Table 3 shows sample **M** for *CC task* while Table 4 shows the **M** for *CD task*. As we can see in these two sample **M**'s, the contextual entity pairs (reflected by zeros and ones) coarsely categorize the patterns into upper-left and bottom-right groups. This is genuinely how SIP learns to discern pattern semantics.

Learning pattern semantics by patterns' *specific* participating entities, however, come with issues of data sparseness and specificity: a certain entity pair could only be mentioned in a handful of patterns, and entities may be topically-related (e.g. *carcinoma* and *tumor* are related to *cancer*, *malignant melanoma* to *skin cancer*, *simvastatin* to *statin* and *simvastatin* to *lovastatin*). Therefore, we further transform entity space into latent topic space to avoid these issues (Step 4b). Specifically, we leverage LSA (31) to learn entity pairs' semantic topics and to reduce dimensionality from the number of distinct entity pairs (j) to the number of distinct LSA semantic topics (t) where $t \ll j$. This equates to transforming pattern representations in entity space, **M**, into pattern representations in LSA topic space, $\mathbf{M}'$.

LSA constructs the t-topic semantic space by a number of steps, namely, performing rank-reduced singular value decomposition on the matrix in entity space, retaining t largest (significant) singular values and approximating the matrix in

Table 2. Example stemmed, semantically tagged queries and corresponding context patterns

User query	Stemmed and semantically tagged query	Context pattern
Chlorthalidone vs hydrochlorothiazide	<C>chlorthalidone</C> vs < C>hydrochlorothiazide</C>	*#C vs #C*
Switching from clopidogrel to prasugrel	switch from <C>clopidogrel</C> to < C>prasugrel</C>	*switch from #C to #C*
Sodium hypochlorite and chlorhexidine gluconate interaction	<C>sodium hypochlorite</C> and <C>chlorhexidine gluconate</C> interaction	*#C and #C interaction*
Megestrol acetate for treatment of anorexia-cachexia syndrome	<C>megestrol acetate</C> for treatment of < D>anorexia-cachexia syndrome</D>	*#C for treatment of #D*
Isoproterenol induced myocardial infarction	<C>isoproterenol</C> induce <D>myocardial infarction</D>	*#C induce #D*
Tamoxifen side effects breast cancer	<C>tamoxifen</C> side effect <D>breast cancer</D>	*#C side effect #D*

Table 3. Reference matrix **M** for the *CC task*

Entity Pair\Pattern	*#C vs #C*	*#C versus #C*	*#C compare to #C*		*#C and #C interaction*	*#C interaction with #C*	*#C and #C drug interaction*
#C: chlorthalidone #C: hydrochlorothiazid	1	1	1		0	0	0
#C: albuterol #C: levalbuterol	1	1	1		0	0	0
#C: omeprazole #C: ranitidine	1	1	0	...	0	0	0
#C: pazopanib #C: sunitinib	1	1	1		0	0	0
⋮				⋱			
#C: sodium hypochlorite #C: chlorhexidine gluconate	0	0	0		1	1	0
#C: warfarin #C: amoxicillin	0	0	0		1	1	1
#C: amlodipine #C: simvastatin	0	0	0		1	1	1
#C: voriconazole #C: tacrolimus	0	0	0		1	0	1

Table 4. Reference matrix **M** for the *CD task*

Entity Pair\Pattern	*#D due to #C*	*#D associate with #C*	*#C induce #D*		*#C in #D treatment*	*Treatment of #D with #C*	*#D and #C therapy*
#C: warfarin #D: skin necrosis	1	1	1		0	0	0
#C: isoproterenol #D: myocardial infarction	0	1	1		0	0	0
#C: tamsulosin #D: intraoperative floppy iris syndrome	1	1	1	...	0	0	0
#C: omeprazole #D: acute pancreatitis	1	1	1		0	0	0
⋮				⋱			
#C: sodium bicarbonate #D: cancer	0	0	0		1	1	1
#C: methotrexate #D: rheumatoid arthritis	0	0	0		1	1	1
#C: glucosamine #D: osteoarthritis	0	0	0		1	1	1
#C: clonidine #D: diabetic diarrhea	0	0	0		1	1	0

the least-squares sense. Finally, a lower-dimension i-by-t matrix approximation ($\mathbf{M}'$) to the original i-by-j matrix ($\mathbf{M}$) is learned in an attempt to model pattern semantics in terms of t LSA topics. Note that although similar method such as probabilistic LSA (pLSA) (32) could also be used for rank reduction, pLSA does not outperform LSA in both our tasks.

In this paper, we refer to SIP as an unsupervised framework because it requires no *specific* manually annotated seeds or training data for pattern semantics analysis. Although the open-source entity recognition tools (i.e. (23–25)) used in Step 1 need entity annotations, such annotations and these tools are not designed and re-trained for the purpose of discovering context patterns with similar meaning, and entity recognition can always be achieved by less-satisfying dictionary methods.

Discovering semantically similar patterns

Once context patterns are semantically recognized in LSA space as $\mathbf{M}'_{i\times t}$, instead of their lexical forms, SIP estimates patterns' semantic similarity by their distributional similarity in LSA latent topics. SIP proposes semantically similar candidate patterns using the procedure in Figure 2. Algorithm 2 shows the detailed steps.

Algorithm 2. Semantically similar pattern discovery.
(1) Initialize matrices **Sim** and **List**
(2) Extract LSA topical vectors for pattern p, p' from $\mathbf{M}'$ where $p \neq p'$
(3) Calculate distributional similarity for LSA vectors of p and p' as an estimate of semantic similarity of p and p'
(4) Record similarity scores in **Sim** and store the highest-scored N patterns for each p in **List**

First, matrix $\mathbf{Sim}_{i\times i}$ is initialized to record (semantic) similarity scores between patterns and $\mathbf{List}_{i\times N}$ to store each pattern's top-scored N patterns in similarity. Similar to space transformations, finding candidates of semantically similar patterns is done independently from one entity type pair to another. As a result, the similarity calculation of chemical–chemical patterns does not concern that of chemical–disease patterns, and i refers to the number of the unique patterns in our *CC task* or that in our *CD task*.

Next, for pattern p and p' ($p \neq p'$), SIP first extracts their LSA topic vectors from $\mathbf{M}'_{i\times t}$. These vectors represent the patterns in LSA space and describe pattern semantics in t LSA topics. Then, SIP estimates the semantic similarity of patterns p and p' by the cosine similarity of their LSA t-topic distributions as

$$\mathrm{cosSim}(\mathbf{V}_p, \mathbf{V}_{p'}) = \frac{\mathbf{V}_p \cdot \mathbf{V}_{p'}}{|\mathbf{V}_p|\ |\mathbf{V}_{p'}|} = \sum_{t'=1}^{t} (\mathbf{V}_p[t'] \times \mathbf{V}_{p'}[t']) \Big/ \left(\sqrt{\sum_{t'=1}^{t} \mathbf{V}_p^2[t']} \times \sqrt{\sum_{t'=1}^{t} \mathbf{V}_{p'}^2[t']} \right)$$

where $\mathbf{V}_x$ denotes the LSA vector for pattern x and $\mathbf{V}_x[t']$ denotes the scalar component of $\mathbf{V}_x$ along the axis of LSA topic t' ($1 \leq t' \leq t$).

For each pattern p, SIP yields a set of patterns whose similarity scores are among its top N as its semantically similar candidates. At last, sets of paraphrasable pattern pairs are obtained. Table 1 shows example discovery of semantically similar context patterns on our working prototype.

Experiments

SIP is designed to learn the semantics of context patterns by entities involved. Although both scholarly publications and Web queries provide such information (i.e. the entities that patterns keep), we prefer Web queries because user queries tend to bond entities in proximity. As such, SIP is trained and evaluated over Web queries. In this section, we first present our PubMed query data, for discovering semantically similar entity relations or context patterns and the process to construct our test set. Then, we describe the parameter settings for SIP and outline the evaluation process. Finally, experimental results are reported and discussed.

Knowledge source and test set

Knowledge source: PubMed queries

A total of six-month's worth of 35 968 309 PubMed queries (24.3 million unique queries) was collected for our experiment of pattern semantics understanding. Queries with exactly two entities were stemmed, entity-tagged and re-formulated into context patterns following the procedure in Figure 1 for semantically similar pattern finding in Figure 2. Table 5 shows some frequent dual-entity context patterns or entity relations in PubMed queries. Frequent chemical–chemical patterns cover relations of drug/chemical comparison (e.g. *#C versus #C*), interaction (e.g. *#C and #C interaction*) and so on, whereas frequent chemical–disease patterns cover semantics of chemical-induced side effects (e.g. *#C induce #D*), drugs' therapeutic effects (e.g. *treatment of #D with #C*), etc.

Table 5. Example frequent context patterns in PubMed queries

Chemical–chemical context pattern	Chemical–disease context pattern
#C *and* #C	#D *and* #C
#C *versus* #C	#C *induce* #D
#C *and* #C *interaction*	#D *treatment* #C
#C *and* #C *combination*	*treatment of* #D *with* #C
#C *plus* #C	#C #D *review*
#C *with* #C	#D *with* #C
comparison of #C *and* #C	#D *child* #C
interaction between #C *and* #C	#D *induce by* #C
#C *oxidase* #C	#D *due to* #C
#C *dehydrogenase* #C	#D *treatment with* #C
combine #C *and* #C	*role of* #C *in* #D
#C *transporter* #C	#C *metabolism and* #D

Patterns are shown in the order of descending frequency and words are stemmed.

Table 6. Example test patterns in our *CC* and *CD* tasks

CC task's test pattern	*CD task*'s test pattern
#C *vs* #C	#C *induce* #D
#C *and* #C *interaction*	#D *treatment* #C
#C *and* #C *combination*	#D *with* #C
#C *plus* #C	#D *child* #C
#C *with* #C	#D #C *supplement*
Comparison of #C *and* #C	*dietary* #C *and* #D
Switch #C *to* #C	#C *and the risk of* #D
#C *after* #C	#C *intake and* #D
#C *and* #C *resistance*	#D #C *therapy*
#C *and* #C *abuse*	#D *due to* #C
#C *and* #C *side effect*	*refractory* #D #C

Words are stemmed.

Test set construction

We constructed our test set semi-automatically in two steps. We first ordered PubMed context patterns according to their frequency and the diversity of their participating entity pairs in our query log. We then manually examined the top-ranked patterns and considered a pattern suitable for testing if it is a common, general biomedical pattern (in contrast to specific ones such as #C *oxidase* #C and #C *transporter* #C) and it should not be ambiguous about entity relations. Our final test set consisted of 68 chemical–chemical and 120 chemical–disease testing patterns (see Table 6 for examples). For each of these patterns, we performed the evaluation on the list of top-ranked similar patterns returned by SIP.

System settings and evaluation process

System settings for SIP

We evaluated SIP framework on different numbers of LSA topics: 10, 20, 40, 60, 80, 100, 150, 200 and 300. We started with a small topic number of 10 and increased the number faster to 300 because of the fact that 300 ± 100 topics have been used to analyze lexical semantics of general documents (33) and that, compared to full-text general documents, we had a much smaller and constrained vocabulary. On the other hand, to avoid possible noise in Web queries, we restricted SIP to the most frequent 500, 1000, 1500, 2000, 2500 and 3000 chemical–chemical/chemical–disease entity pairs in PubMed queries when constructing *CC*/*CD task*'s entity space in Figure 1.

Evaluation process

All 54 system settings for SIP (9 different numbers of LSA topics $\times$ 6 different numbers of frequent entity pairs) were evaluated in our *CC* and *CD* tasks. In evaluation, candidate semantically similar pattern pairs were pooled from the 54 SIP alternatives and our baseline, and were manually judged for semantic similarity. As the authors concurred on each other's semantic judgement most of the time (85%) in prior-experiment analysis, only one of the authors examined the pooled results blindfolded. In total, 1687 unique pattern pairs in *CC task* and 3609 unique pairs in *CD task* were manually evaluated and annotated as:

Strict match. A pattern pair is considered to be strict-match if, in biomedical context, its patterns are semantically the same (e.g. #C *induce* #D and #D *due to* #C) or highly similar (e.g. #D *child* #C and *pediatric* #D #C).

Relaxed match. A pattern pair is considered to be relaxed-match if, in biomedical context, its patterns are semantically related and one of its patterns entails or contextually subsumes the other. For example, #C *reduce* #C and #C *effect on* #C are relaxed-match semantically similar patterns since #C *reduce* #C entails #C *effect on* #C, whereas #C *induce* #D and #C *induce* #D *in rat* are relaxed-match since #C *induce* #D subsumes the contexts of #C *induce* #D *in rat* (the same applies to #C *induce* #D and #C *induce* #D *treatment*).

No match. A pattern pair is considered to be no-match if it is neither one of the above.

Based on the annotations, standard information retrieval measures—mean reciprocal rank (MRR) and normalized discounted cumulative gain (nDCG) (34)—were used to evaluate system ability to return relevant, semantically similar, patterns at top N positions. While MRR measures the effort to locate the first true semantically similar pattern pair in the candidate list (the closer it is to 1 means less effort), nDCG measures system performance in ranking true semantically similar pairs earlier in the list (the closer it is to 1 means better performance).

In our experiments, systems were expected to discover **strict-match** pattern pairs. However, finding **relaxed-match** ones could also be beneficial to biocuration and information retrieval. For instance, #*C reduce* #*C* depicts a specific context of its relaxed-match counterpart #*C effect on* #*C* and narrows down information need in search, and #*C induce* #*D treatment* provides the also-want-to-know for its relaxed-match #*C induce* #*D* which indicates an opportunity of automatic query suggestion/completion (35). As a result, we also evaluated systems on finding relaxed-match pattern pairs. Specifically, system performance on discovering **strict-match/relaxed-match** semantically similar patterns was measured in terms of MRR@*N* and averaged nDCG@*N* where $N = 1$, 3, 5 or 10. And since similar trends were observed across different values of *N*, we only present the results with $N = 3$ in the next subsection for simplicity.

Evaluation results

Results of chemical–chemical (CC) semantic relations

The performance of SIP on finding **strict-match** chemical–chemical patterns (i.e. the **strict-match** *CC task*) is summarized in Figure 3. In this figure, histograms represent the (a) MRR and (b) nDCG performance of different SIP settings concerning the LSA topics and the most frequent entity pairs, and colors are used to differentiate LSA topic numbers. For instance, green bars, labelled as T60, denote the SIP performance when set with 60 LSA topics. And 60-topic SIP (i.e. green bars) performed differently when accompanied with different numbers of frequent entity pairs (i.e. 500, 1000, 1500, 2000, 2500, 3000): 60-topic SIP achieved around 0.4 MRR using 500 frequent entity pairs but achieved around 0.8 MRR using 3000. The results of T200 and T300 are omitted as system performance degraded drastically after T100 (i.e. T150, T200 and T300).

Figure 3 also plots SIP's best performance (i.e. the solid lines) with respect to each number of entity pairs used. For example, when using 2000 frequent entity pairs, SIP achieved the best 0.81 MRR with 20 LSA topics, thus T20 0.81 labelled. For comparison, the dotted lines represent the performance of our baseline, which simply estimated patterns' semantic similarity by the cosine similarity of their *specific* participating entity pairs in the queries without using LSA topic information. In other words, our baseline is basically SIP framework excluding the component of latent semantic analysis (i.e. Step 4b in Figure 1).

As shown in Figure 3, the performance of smaller topic numbers ($t \leq 80$) tends to improve with increasing entity pairs and their performance becomes steady at 2500–3000 entity pairs: increasing the number of frequent entity pairs from 500 to 1000 gave MRR and nDCG the largest margin of improvement whereas increasing from 1000 to 1500 yielded the second largest. Nonetheless, with larger topic numbers ($t \geq 150$), SIP did not always benefit from the entity pair increase and did not perform well.

Encouragingly, SIP with small topic numbers significantly outperformed the baseline which tends not to benefit from using more entity pairs either. SIP achieved the highest MRR score of 0.86 and the highest nDCG score of 0.87 when as few as 20 LSA topics were used with 3000 entity pairs. And a MRR and nDCG above 0.85 indicate that the first-ranked candidate pairs were almost always correct.

Results of chemical–disease (CD) semantic relations

Using the same **strict-match** criterion and figure configuration in Figure 3, Figure 4 summarizes the results on

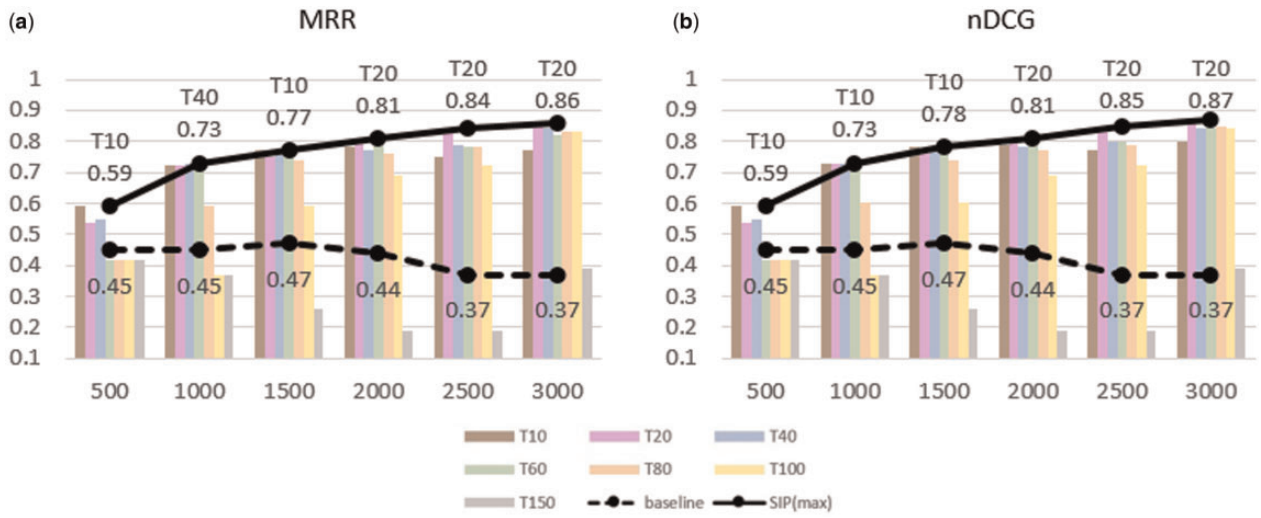

Figure 3. System performance on the *CC task* with different LSA topic numbers (10–150) and different numbers of the most frequent entity pairs (500–3000). **Strict match** is required. The solid line represents best-performing SIP while the dotted line represents the baseline.

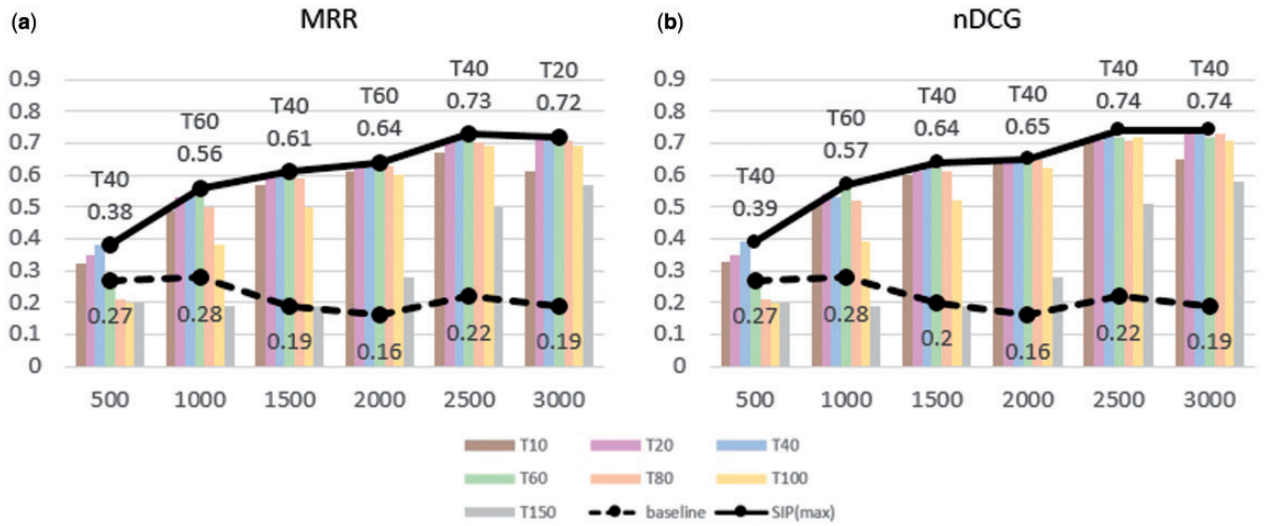

Figure 4. System performance on the *CD task* with different LSA topic numbers (10–150) and different numbers of the most frequent entity pairs (500–3000). **Strict match** is required. The solid line represents best-performing SIP while the dotted line represents the baseline.

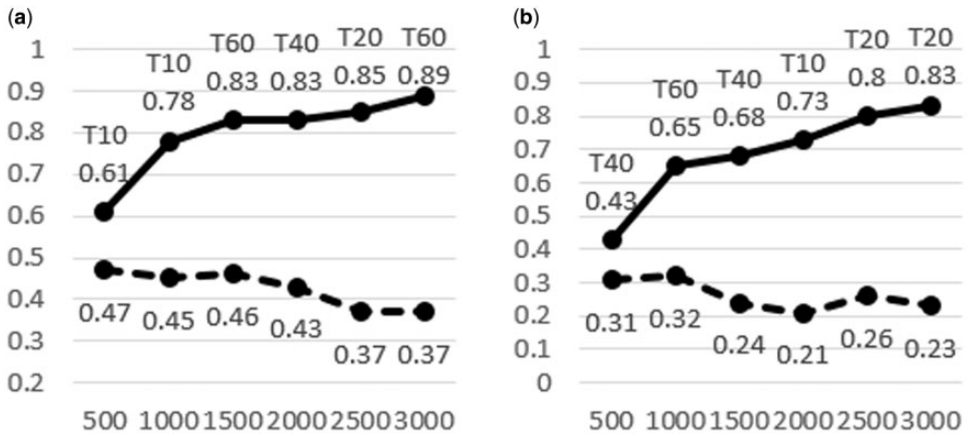

Figure 5. nDCG results on our (a) *CC task* and (b) *CD task* when both **strict-match** and **relaxed-match** are allowed. The solid line represents best-performing SIP while the dotted line represents the baseline.

discovering semantically similar chemical–disease semantic relations. Similar to the *CC task*, SIP generally benefited from more entity pairs in the *CD task* and the 500–1000 entity pair increase led to SIP's largest margin of improvement. Again SIP significantly outperformed the baseline by a large margin. One thing worth mentioning is that, compared to the *CC task*, both SIP and the baseline yielded lower performance: while SIP dropped from a MRR of 0.86 to 0.73 and a nDCG of 0.87 to 0.74, the baseline drastically dropped to a MRR of 0.28 and a nDCG of 0.28. This is mainly because our *CD task* contained a broader spectrum of semantic contexts/relations (i.e. the chemical–disease relations in PubMed queries were more diverse).

Since discovering **relaxed-match** patterns can also be beneficial, we further examined system performance with both **strict-** and **relaxed-match** patterns allowed. Figure 5 reports corresponding nDCG results on our *CC* and *CD* tasks. As expected, SIP gained from relaxing the matching criterion and achieved an improved performance of nDCG closer to 0.9 and 0.85 in semantically understanding the chemical–chemical and chemical–disease patterns, respectively.

Overall, entities in user queries serve as good knowledge to differentiate query semantics. Projecting user entities into LSA latent topics further helps discover semantically similar entity-relations, or context patterns, on the Web. Also, compared to word sense induction in general documents, smaller LSA topic numbers in the range of 30 ± 10 can yield the best results for biomedical **strict-match** *CC* and *CD* tasks. And using numbers as small as the top frequent 2500–3000 entity pairs in the PubMed query log can achieve satisfying performance across diverse semantic relations.

Discussion

Tables 7 and 8 show SIP-proposed **strict-match** patterns in the *CC* and *CD* task respectively. They are consolidated across different SIP settings @ $N = 10$. SIP effectively discovered synonymous patterns for a spectrum of entity relations (e.g. *#C vs #C* and *comparison between #C and #C*, *combine #C and #C* and *#C in combination with #C* and *#C induce #D* and *#D due to #C*). One thing worth mentioning is that SIP discovered multiple-sense pattern *#C with #C* and *#D with #C* and semantically associated them with senses *#C combine with #C* and *#C interaction with #C* for *CC* (Table 7), and senses *#D associate with #C* and *#D treatment with #C* for *CD* (Table 8) respectively. Although SIP focused on context patterns instead of words to avoid ambiguity and yielded satisfying results, it would be interesting to investigate the impact of such ambiguous context patterns (e.g. short context patterns containing only prepositions) on SIP in the future.

In the experiments, we compared SIP with our baseline to highlight the importance of LSA space transformation. We did not directly compare our method with other pattern recognition methods such as (21) that are based on entity co-occurrence at either sentence or abstract level because our goal was to discover information needs (relations) that are frequently sought by PubMed users. Additionally, SIP was designed to work in query space where semantics (in queries) tend to be clear and specific, and entities and relations (in queries) tend to bond in proximity, which may not hold for some entity relations in literature space. Finally, SIP was designed to discover semantically similar patterns without supervision. That is, no training/seed data are required for the purpose of pattern semantics understanding. Therefore, it is not straightforward to compare SIP with traditional methods which require and start with entity seeds describing the pattern (e.g. (21) leverages chemical–disease entity seeds having chemical–induced–disease relation to recognize the said relation). Nonetheless, when examining the output of (21), we observed complementary results for the chemical–cause–disease relation. For instance, SIP had query-specific *#D from #C* but missed *#D during #C*.

Table 7. Example synonymous chemical–chemical patterns identified by SIP with different settings

Semantic relation	Test pattern	SIP semantically similar patterns
Drug comparison	*#C vs #C*	*#C compare #C*, *#C compare to #C*, *#C comparison #C*, *#C versus #C*, *comparison between #C and #C*, ...
Drug combination	*combine #C and #C*	*#C and #C combination*, *#C in combination with #C*, *#C with #C*, ...
Drug combination/ interaction	*#C with #C*	*#C combine with #C*, *#C interaction with #C*, ...
Drug change	*switch #C to #C*	*switch from #C to #C*, *#C to #C*, ...

Pattern words are stemmed.

Table 8. Example synonymous chemical–disease patterns identified by SIP with different settings

Semantic relation	Test pattern	SIP semantically similar patterns
Chemical cause disease	*#C induce #D*	*#C cause #D*, *#D after #C*, *#D due to #C*, *#D from #C*, *#D induce by #C*, ...
Chemical treat disease	*#D treatment #C*	*#C treatment for #D*, *#D therapy #C*, ...
Chemical cause/treat disease	*#D with #C*	*#D associate with #C*, *#D treatment with #C*, ...
Role/use of chemical in disease	*#C in #D*	*role of #C in #D*, *use of #C in #D*, ...

Pattern words are stemmed.

Applications of SIP-derived semantically similar patterns

In this section, we apply SIP **strict-match** pattern pairs to two specific biomedical tasks: biomedical document retrieval and bio-entity relation extraction (See Figure 1). We show that SIP output can benefit the process of biocuration and semantic information retrieval (IR).

Biomedical document retrieval

In the field of biology and life sciences where entities have abundant alias, retrieving documents containing the exact user search words may not be sufficient. As a result, PubMed (1) uses Medical Subject Headings (i.e. MeSH terms) expansion by default (9) and searches for query

Table 9. PubMed responses to query submission (a) *albuterol vs levalbuterol* and (b) *albuterol vs levalbuterol OR albuterol versus levalbuterol* where (a) is the original user query while (b) is (a)'s new query expanded using SIP pattern knowledge

(a) Search results for the original query *albuterol vs levalbuterol*	(b) Search results for the new query *albuterol vs levalbuterol OR albuterol versus levalbuterol*
1. **Albuterol** and **levalbuterol** use...	1.–16. the same as the search results on the left
2. **Levalbuterol** compared to racemic **albuterol**...	17. ...cost comparison of **levalbuterol versus albuterol**...
3. Comparison of **levalbuterol** and racemic **albuterol**...	18. **Levalbuterol versus albuterol** for...
4. Comparison of racemic **albuterol** and **levalbuterol**...	19. Efficacy of racemic **albuterol versus levalbuterol** used...
5. ... comparison of nebulized **levalbuterol** and **albuterol**...	20. Hospital readmissions following initiation...
6. The effects of racemic **albuterol** versus **levalbuterol**...	21. **Levalbuterol versus albuterol.**
7. Low-dose **levalbuterol** ... in comparison with... **albuterol.**	22. Repeated β2-adrenergic receptor agonist therapy...
8. Comparison of ... **levalbuterol** ... S-**albuterol** ...	23. ...comparison of **levalbuterol versus** racemic **albuterol**...
9. ...comparison of **levalbuterol** and **albuterol**...	24. ...evaluation of **levalbuterol versus** racemic **albuterol**...
10. Evaluation of **levalbuterol** metered dose...	25. ...comparing **levalbuterol** with racemic **albuterol**...
11. A comparison of **levalbuterol** with racemic **albuterol**...	26. ...study of **levalbuterol** and racemic **albuterol** ...
12. **Levalbuterol vs** racemic **albuterol**...	27. ... **levalbuterol** compared with racemic **albuterol**...
13. Long-term safety study of **levalbuterol**...	28. ...efficacy of racemic **albuterol versus levalbuterol**...
14. **Albuterol vs. levalbuterol**...	29. ...effectiveness of **levalbuterol versus** racemic **albuterol.**
15. High-dose continuous nebulized **levalbuterol** for...	30. **Levalbuterol versus** racemic **albuterol** in...
16. 16. Evaluation of the efficacy and safety of **levalbuterol**...	31. An evaluation of **levalbuterol** HFA in...
	32. **Levalbuterol** aerosol ... a nonelectrostatic **versus** a...
	33. Risk **versus** benefit considerations for...
	34. In vitro estimations of in vivo jet nebulizer efficiency...
	35. Evaluation of the utilization patterns of leukotriene...

Note that we bold-face PubMed highlighting of query words and show only PubMed titles for simplicity.

words not only in documents but also associated MeSH headings. By doing so, PubMed alleviates the issue of biomedical term mismatch between document words and query words. Take the query *albuterol* for example. PubMed will return documents containing *albuterol* and documents without *albuterol* but (annotated) with the same MeSH heading as *albuterol*. Thus, documents not containing *albuterol* but containing the synonyms of *albuterol* such as proventil, salbutamol and ventolin, will also be returned. Nonetheless, since general terms/phrases are out of the scope of MeSH headings, PubMed can still suffer from general-purpose vocabulary mismatch during search.

Consider a real user query *albuterol vs levalbuterol*. PubMed's retrieval effectiveness could be improved if PubMed semantically understands the query by exploiting SIP synonymous pattern pairs, (#*C vs* #*C*, #*C versus* #*C*) in this case, and returns the accumulated search results from both the original query *albuterol vs levalbuterol* and SIP-motivated counterpart *albuterol versus levalbuterol* (see Table 9). As shown, PubMed retrieves relatively 118% more documents for the new query (35 documents vs 16 documents). In addition, examining the retrieved PubMed titles shows that with SIP's query expansion, *albuterol versus levalbuterol* for *albuterol vs levalbuterol*, one can obtain relatively 100% more relevant documents (22 vs 11) in this case. Retrieving more relevant documents is essential to biocuration, semantic IR, and article triaging of many biomedical shared challenges (36).

The benefit of SIP in semantic IR, alleviating vocabulary mismatch that is not covered by MeSH, can also be observed in another real user query *methotrexate combined with tofacitinib* where SIP proposes pattern #*C combine with* #*C* and #*C in combination with* #*C* are synonymous (see Table 10). Based on the first-page PubMed responses shown in Table 10, PubMed clearly achieves better retrieval performance with the expanded query *methotrexate combined with tofacitinib OR methotrexate in combination with tofacitinib*. In the near future, the applicability of SIP patterns in PubMed literature search as query expansion will be examined more extensively and quantitatively.

Biomedical relation extraction

Similar to many bio-NLP challenge tasks such as chemical–disease relation extraction (2), protein-protein interaction extraction (3), drug-drug interaction extraction (5) and identification of gene events (4) and disease co-morbidities (6), SIP focuses on two-argument, dual-entity, relations. In this subsection, we examine SIP applicability in a real-life relation extraction problem and compare SIP

Table 10. PubMed responses to query submission (a) *methotrexate combined with tofacitinib* and (b) *methotrexate combined with tofacitinib OR methotrexate in combination with tofacitinib* where (a) is the original user query while (b) is (a)'s new query expanded using SIP pattern knowledge

(a) Search results for the original query *methotrexate combined with tofacitinib*	(b) Search results for the new query *methotrexate combined with tofacitinib OR methotrexate in combination with tofacitinib*
1. **Tofacitinib** in combination with nonbiologic disease-	1. **Tofacitinib** in **combination** with nonbiologic disease-
2. modifying antirheumatic drugs...	modifying antirheumatic drugs...
3. ...**tofacitinib** (CP-690,550) **combined** with **methotrexate**...	2. ...**tofacitinib** (CP-690,550) **combined** with **methotrexate**...
4. ...of Novel DMARDs as Monotherapy and in Combination	3. ...of Novel DMARDs as Monotherapy and
5. with **Methotrexate**...	in **Combination** with **Methotrexate**...
6. Recent progress and perspective in JAK inhibitors...	4. **Tofacitinib** with **methotrexate**...
7. In vitro and in vivo analysis of a JAK inhibitor in...	5. ...safety of **tofacitinib**, with or without **methotrexate**,...
8. Serum 14-3-3η level is associated with severity...	6. **Tofacitinib** (CP-690,550) in **combination** with
9. Pharmacotherapy options in rheumatoid arthritis.	**methotrexate**...
	7. The JAK inhibitor **tofacitinib** suppresses...
	8. Systematic review of **tofacitinib**: a new drug for...
	9. ...studies of **tofacitinib** in patients with rheumatoid...
	10. ...**tofacitinib** (CP-690,550) versus placebo in
	combination with background **methotrexate**...
	11. Efficacy of conventional synthetic disease-modifying
	antirheumatic drugs, glucocorticoids and **tofacitinib**:...
	12. ...the treatment of rheumatoid arthritis: position on the
	use of **tofacitinib**.
	13. **Tofacitinib**: The First Janus Kinase (JAK)...
	14. **Tofacitinib** or adalimumab versus placebo...
	15. **Tofacitinib**: a review of its use...
	16. In vitro and in vivo analysis of a JAK inhibitor in...
	17. Efficacy and safety of **tofacitinib** for treatment...
	18. Summaries for patients: **tofacitinib** for the treatment...
	19. Current and future oral systemic therapies for psoriasis.
	20. **Tofacitinib** for the treatment of moderate to...

Note that we bold-face PubMed highlighting of query words and show only the PubMed titles of the retrieval results on the first pages for simplicity.

effectiveness in helping biocuration with simple co-occurrence method.

Specifically, we exploit SIP **strict-match** patterns to address the problem of 2016 BioCreative chemical–disease relation extraction subtask (2): extraction of chemical–induced–disease (CID) relations. We experiment on the 2016 official development set, consisting of 500 PubMed abstracts and extract chemical–induced–disease (CID) relations in a number of steps. First, starting with a representative CID pattern *#C induce #D* (words are stemmed), we consolidate its SIP **strict-match** patterns from different settings. This process examines newly-discovered patterns and adds their synonymous patterns iteratively. In total, we collect 24 SIP context patterns associated with the CID relation including *#C cause #D*, *#D due to #C*, *#D associate with #C*, *#D cause by #C* and *#C and the risk of #D*. Second, for any chemical–disease pair in a PubMed abstract, we extract its context (or contextual words) in the abstract and manage to best match the contextual words to a SIP pattern out of the 24, if any. Table 11 shows example PubMed contextual words surrounding chemical–disease pairs and contextual words' best-matched SIP CID patterns. Note that in this step we require the chemical–disease pair to appear in the same sentence. Finally, we consider the chemical–disease pairs whose PubMed contextual words have matched our SIP patterns to be candidates having CID relation.

As Table 12 shows, the above pattern-matching approach assisted by SIP output outperforms co-occurrence baselines relatively by 47 and 10% where co-occurring chemical–disease pairs in abstracts and sentences are proposed as CID candidates. We believe that, without the computational overhead of stemming and machine learning/training, such approach can be the first step to help accelerate biocuration and that its performance in relation extraction can be further improved if incorporated more CID patterns and/or co-developed with machine learning techniques.

Table 11. Example PubMed contextual words of chemical–disease pairs in the same sentence and their best-matched SIP patterns related to chemical–induced–disease relation

PubMed ID	PubMed contextual words of chemical–disease pair	Best-matched SIP pattern
1928887	#D due to #C overdose	*#D due to #C*
2564649	#D is a well-known side effect that is associated with high-dose #C	*#D associate with #C*
3015327	#D caused by a single dose of #C in rats	*#D cause by #C*
6504332	#C-induced #D in a neurologically-impaired child	*#C induce #D*
11573852	#C intake as well as withdrawal can also cause #D.	*#C cause #D*
15266215	#D risk was consistently higher for users of #C	*#D risk #C*

Pattern words are stemmed.

Table 12. System performance on extraction of chemical–induced–disease (CID) relation

	Precision (%)	Recall (%)	*f*-Measure (%)
Abstract-level co-occurrence	19.2	99.7	32.3
Sentence-level co-occurrence	30.6	74.7	43.4
24 SIP-derived CID patterns	55.8	41.8	47.8

Summary and future work

We have introduced an unsupervised method for discovering semantically similar patterns/relations on the Web. The method involves representing Web queries as context patterns in both entity space and LSA topic space, and estimating patterns' semantic similarity by their distributional similarity in LSA topics. In this work, we pioneer in examining the applicability of LSA in query-based pattern semantics analysis and the applicability of biological entities in PubMed queries in automatic discovery of synonymous biomedical patterns/relations without seed entities that pattern recognition methods generally require. Two separate task-oriented evaluations (*CC* and *CD* tasks) show that entities in user queries and LSA entity-to-topic space transformation can contribute to biomedical semantic relation discovery. In addition, we explore the applications of our SIP-derived synonymous patterns in biomedical document retrieval and relation extraction and discuss the potentials in helping the process of biocuration and semantic information retrieval. In future work, we plan to extensively and quantitatively examine the benefits of our discovered patterns in biocuration, query suggestion and PubMed search engine serving million users.

Funding

This work was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

Conflict of interest. None declared.

Acknowledgements

This work was supported [in part] by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. The authors would like to thank MD. Michael Simmons for his help of proofreading this article.

References

1. Islamaj Dogan,R., Murray,G.C., Neveol,A. *et al*. (2009) Understanding PubMed user search behavior through log analysis. *Database (Oxford)*, 2009, bap018
2. Wei,C.H., Peng Y.Leaman R. *et al*. (2015) Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In: *Proceedings of The fifth BioCreative challenge evaluation workshop*, pp. 154–166.
3. Krallinger,M., Leitner,F., Rodriguez-Penagos,C. *et al*. (2008) Overview of the protein–protein interaction annotation extraction task of BioCreative II. *Genome Biol*., 9, S4.
4. Kim,J.D., Ohta T.Pyysalo S. *et al*. (2009) Overview of BioNLP'09 shared task on event extraction. In: *Proceedings of the Workshop on BioNLP: Shared Task*, p. 1–9.
5. Segura-Bedmar,I. Martínez,P. and Sánchez-Cisneros,D. 2011 The 1st DDIExtraction-2011 challenge task: extraction of drug–drug interactions from biomedical texts. In: *Proceedings of the 1st challenge task on drug–drug interaction extraction*, p. 1–9.
6. Uzuner,O. (2009) Recognizing obesity and comorbidities in sparse data. *J. Am. Med. Inform. Assoc*., 16, 561–570.
7. Paul,M.J, White,R.W. and Horvitz,E. (2015) Diagnoses, decisions, and outcomes: web search as decision support for cancer. In: *Proceedings of WWW*, p. 831–841.
8. Kim,S., Kim,W., Wei,C.H. *et al*. (2012) Prioritizing PubMed articles for the Comparative Toxicogenomic Database utilizing semantic information. *Database (Oxford)*, 2012, bas042.
9. Lu,Z., Kim,W., and Wilbur,W.J. (2009) Evaluation of query expansion using MeSH in PubMed. *Inf. Retr. Boston*, 12, 69–80.
10. Mattingly,C.J., Colby,G.T., Forrest,J.N. *et al*. (2003) The comparative toxicogenomics database (CTD). *Environ. Health Perspect*., 111, 793–795.
11. Breitkreutz,B.J., Stark,C., and Tyers,M. (2002) The GRID: the general repository for interaction datasets. *Genome Biol*., 3, PREPRINT0013.
12. Klein,T.E. and Altman,R.B. (2004) PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. *Pharmacogenomics J*., 4, 1.

13. Neveol,A., Islamaj Dogan,R., and Lu,Z. (2011) Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *J. Biomed. Inform.*, 44, 310–318.
14. White,R. and Horvitz,E. (2009) Cyberchondria: studies of the escalation of medical concerns in Web search. *ACM Trans. Inf. Syst.*, 27(4), Article No. 23.
15. Tamine,L., Chouquet,C. and Palmer,T. (2015) Analysis of biomedical and health queries: lessons learned from TREC and CLEF evaluation benchmarks. *J. Assoc. Inf. Sci. Technol.*, 66(12), 2626–2642.
16. Kim,J.J., Pezik,P., and Rebholz-Schuhmann,D. (2008) MedEvi: retrieving textual evidence of relations between biomedical concepts from medline. *Bioinformatics*, 24, 1410–1412.
17. Lu,Y., Fang,H., and Zhai,C.X. (2009) An empirical study of gene synonym query expansion in biomedical information retrieval. *Inf. Retr.*, 12, 51–68.
18. Pasche,E., Gobeill,J., Vishnyakova,D. *et al.* (2013) Use of controlled vocabularies to improve biomedical information retrieval tasks. *Stud. Health Technol. Inf.*, 192, 1068.
19. Liu,H., Johnson,S.B., and Friedman,C. (2002) Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J. Am. Med. Inf. Assoc.*, 9, 621–636.
20. Gliozzo,A., Giuliano,C. and Strapparava,C. (2005) Domain kernels for word sense disambiguation. In: *Proceedings of ACL*, p. 403–410.
21. Xu,R. and Wang,Q. (2014) Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. *J. Biomed. Inf.*, 51, 191–199.
22. Tsuruoka,Y. and Tsujii,J. (2005) Bidirectional inference with the easiest-first strategy for tagging sequence data. In: *Proceedings of EMNLP*, p. 467–474.
23. Leaman,R., Wei,C.H., and Lu,Z. (2015) tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminf.*, 7, S3.
24. Leaman,R., Islamaj Dogan,R., and Lu,Z. (2013) DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29, 2909–2917.
25. Wei,C.H., Kao,H.Y., and Lu,Z. (2015) GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *Biomed. Res. Int.*, 2015 (2015), Article ID 918710.
26. Paşca,M. (2007) Organizing and searching the World Wide Web of facts – step two: harnessing the wisdom of the crowds. In: *Proceedings of WWW*, p. 101–110.
27. Jain,A. and Pennacchiotti,M.(2010) Open entity extraction from Web search query logs. In: *Proceedings of COLING*, p. 510–518.
28. Lin,D. (1998) Automatic retrieval and clustering of similar words. In: *Proceedings of ACL*, p. 768-774.
29. Turney,P.D. (2001) Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: *Proceedings of EMCL*, p. 491–502.
30. Chen,H, Lin,M. and Wei,Y. (2006) Novel association measures using web search with double checking. In: *Proceedings of COLING/ACL*, p. 1009–1016.
31. Řehůřek,R. and Sojka,P. (2010) Software framework for topic modelling with large corpora. In: *Proceedings of LREC Workshop: New Challenges for NLP Frameworks*, p. 46–50.
32. Hofmann,T. (1999) Probabilistic latent semantic analysis. In: *Proceedings of Uncertainty in Artificial Intelligence*, p. 289–296.
33. Foltz,P.W., Kintsch,W. and Landauer,T.K. (1998) The measurement of textual coherence with latent semantic analysis. *Discourse Process*, 25, 285–307.
34. Jarvelin,K. and Kekalainen,J. (2002) Cumulated gain-based evaluation of IR technologies. *ACM Trans. Inf. Syst.*, 20, 422–446.
35. Lu,Z., Wilbur,W.J., McEntyre,J.R. *et al.* (2009) Finding query suggestions for PubMed. *AMIA Annu. Symp. Proc.*, 2009, 396–400.
36. Huang,C.C. and Lu,Z. (2015) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief Bioinf.* DOI: 10.1093/bib/bbv024.