Journal of Biomedical Informatics 82 (2018) 189–199

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations

Gokhan Bakal[b], Preetham Talari[c], Elijah V. Kakani[c], Ramakanth Kavuluru[a,b,*]

[a] Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, United States
[b] Department of Computer Science, University of Kentucky, United States
[c] Division of Hospital Medicine, Department of Internal Medicine, University of Kentucky, United States

ARTICLE INFO

Keywords:
Information extraction
Relation prediction
Semantic graph patterns

Background: Identifying new potential treatment options for medical conditions that cause human disease burden is a central task of biomedical research. Since all candidate drugs cannot be tested with animal and clinical trials, in vitro approaches are first attempted to identify promising candidates. Likewise, identifying different causal relations between biomedical entities is also critical to understand biomedical processes. Generally, natural language processing (NLP) and machine learning are used to predict specific relations between any given pair of entities using the distant supervision approach.

Objective: To build high accuracy supervised predictive models to predict previously unknown treatment and causative relations between biomedical entities based only on semantic graph pattern features extracted from biomedical knowledge graphs.

Methods: We used 7000 *treats* and 2918 *causes* hand-curated relations from the UMLS Metathesaurus to train and test our models. Our graph pattern features are extracted from simple paths connecting biomedical entities in the SemMedDB graph (based on the well-known SemMedDB database made available by the U.S. National Library of Medicine). Using these graph patterns connecting biomedical entities as features of logistic regression and decision tree models, we computed mean performance measures (precision, recall, F-score) over 100 distinct 80–20% train-test splits of the datasets. For all experiments, we used a positive:negative class imbalance of 1:10 in the test set to model relatively more realistic scenarios.

Results: Our models predict *treats* and *causes* relations with high F-scores of 99% and 90% respectively. Logistic regression model coefficients also help us identify highly discriminative patterns that have an intuitive interpretation. We are also able to predict some new plausible relations based on false positives that our models scored highly based on our collaborations with two physician co-authors. Finally, our decision tree models are able to retrieve over 50% of treatment relations from a recently created external dataset.

Conclusions: We employed semantic graph patterns connecting pairs of candidate biomedical entities in a knowledge graph as features to predict treatment/causative relations between them. We provide what we believe is the first evidence in direct prediction of biomedical relations based on graph features. Our work complements lexical pattern based approaches in that the graph patterns can be used as additional features for weakly supervised relation prediction.

1. Introduction

Biomedical processes are inherently composed of interactions between various types of entities. Typically, these interactions are captured, for computational convenience, as binary relations connecting a subject entity to an object entity through a predicate (or relation type). For example, the relation ("Tamoxifen", *treats*, "Breast Cancer") indicates that the subject entity Tamoxifen is related to the object entity breast cancer via the relation type or predicate *treats*. Besides *treats*, relations with other types of associative predicates such as *causes*, *prevents*, and *inhibits* are also interesting for biomedical research. Often different relations are put together to derive new relations, also termed knowledge discovery. Given that we have established that relations are central to biomedical research, a natural question that arises is how to obtain these relations. Relations that are already discovered and considered common knowledge in the clinical and biomedical communities are typically manually recorded and distributed in public knowledge bases like the Unified Medical Language System (UMLS) Metathesaurus.

* Corresponding author at: Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, United States.
E-mail addresses: mgokhanbakal@uky.edu (G. Bakal), preetham.talari@uky.edu (P. Talari), elijah.kakani@uky.edu (E.V. Kakani), ramakanth.kavuluru@uky.edu (R. Kavuluru).

https://doi.org/10.1016/j.jbi.2018.05.003
Received 9 August 2017; Received in revised form 31 January 2018; Accepted 9 May 2018
Available online 12 May 2018
1532-0464/

However, relations that are not well-known and accepted by the scientific community but are being discovered by particular individuals are often reported in research articles that are subject to peer review.

Given the exponential growth [20] of scientific literature, it is unrealistic to manually review all articles published on a given topic for discovering/extracting any previously unknown relationships between biomedical entities such as treatment/causative relations or drug-drug interactions. Therefore, natural language processing (NLP) and machine learning [12,21] are being increasingly used to *extract* biomedical relations from free text documents. For instance, the treatment relation example discussed earlier in this section may be extracted from the sentence – *"We conclude that Tamoxifen therapy is more effective for early stage breast cancer patients."* However, NLP extractions are essentially based on evidence present in particular sentences and are prone to two types of errors. First, the NLP techniques themselves might not be foolproof given they are computational models and second, and more important, the evidence found in a particular sentence might be circumstantial and not something that is universally accepted. Nevertheless, extraction of the same relation from multiple sentences is indicative of the strength of the relation if it is being reported by multiple research projects.

In this work, we take a completely different approach to predict potential relations between arbitrary pairs of biomedical entities. We refrain from NLP approaches that look at individual sentences to extract a potential relation. Instead, we build a large graph of relations (given a relation translates to a labeled edge) extracted using NLP approaches from scientific literature and exploit semantic path patterns over this graph to build models for specific predicates. That is, instead of looking at what a particular sentence conveys, we model the prediction problem at a global level and build models that output probability estimates of whether a pair of entities participates in a particular relation. Our models are trained with graph pattern features over the well-known knowledge graph SemMedDB [15] extracted from biomedical literature by researchers at the National Library of Medicine (NLM) using the rule based SemRep NLP tool [26].

In our approach, a different binary classification model is trained for each predicate. The ability to identify potentially viable drugs, procedures, and other therapeutic agents for treating different conditions that cause disease burden among humans is at the heart of biomedical research. Similarly causative relations of different biomedical phenomena are also critical in understanding the complex processes that underlie diseases and treatments. Hence, we demonstrate our approach specifically using the *treats* and *causes* predicates. Our method also generalizes to other predicates and can also complement other lexical and syntactic pattern based distant supervision [11,22] approaches for relation extraction. The following are specific contributions of this paper.

1. We propose a novel and intuitive graph pattern feature based approach to predict treatment and causative relations between any given pair of biomedical entities using logistic regression (LR) and decision tree models.
2. We discuss and present details about the potential of graph patterns in terms of coverage and utility of top patterns identified through coefficients of our best LR model.
3. Based on inputs from practicing physicians, we analyze false positives with high probability estimates output by our model to assess their expert based ground truth labels. We also assess the abilities of our best models to recall treatment relations from an external drug repositioning dataset.

We note that this paper is a major extension of a conference paper we published earlier [3] just for the *treats* predicate where a balanced dataset was used during training. In this paper, in addition, we experiment with imbalanced datasets, extend our prior results to the *causes* predicate, conduct qualitative analyses of top graph patterns identified by our models, and assess the reliability of our approach through physician inputs and experiments on a newer drug repositioning dataset.

2. Background: knowledge acquisition and approaches

High level knowledge in biomedicine typically involves interactions between named entities (e.g., genes, drugs, diseases). For example, there is typically a treatment relation between a drug and a disease. Here, these interactions are generally called relations that connect a subject entity (*drug*) with an object entity (*disease*) through a predicate (*treats*). Beyond just named entities, a set of meaningful relations extracted from a dataset can also be construed as a more specific kind of knowledge. However, indirect or implicit relations might exist and can be obtained by putting together several known relations as a sequence where the entities at either end of the sequence are now seen as participating in a new relation. This can only happen when the nature of entities and predicates along the sequence is meaningful to derive this new connection. For example, consider this sequence of two relations (obtained from SemMedDB) with *stimulates* and *treats* predicates in that order:

$$\text{Mercaptopurine} \xrightarrow{stimulates} \text{Cytarabine triphosphate} \xrightarrow{treats} \text{Leukemia}$$

From the two constituent relations taken in this order, we can now see a potential new relation: $\text{Mercaptopurine} \xrightarrow{treats} \text{Leukemia}$. In fact, this is known to be a valid relation between Mercaptopurine and Leukemia. This information is called implicit knowledge if it is not explicitly expressed in textual narratives or structured data sources but can be inferred from existing disparate pieces of evidence. Many simple paths or sequences of relations clearly do not lead to new relations. Even when they do, they may not sometimes be interpretable in a biomedical sense due to missing additional context. Hence there are cases where a compact subgraph connecting a pair of entities is essential in inferring a new implicit relation [7].

Binary relations are typically encoded as (subject, predicate, object) triples and can be obtained from (1). well known expert curated knowledge bases, (2). applying NLP techniques to free text from literature, or (3). employing global lexico-syntactic pattern based methods. Due to excessive time consumption involved in manual curation, knowledge bases are generally not complete/exhaustive [41]. NLP approaches can be used to extract relations from particular sentences using the linguistic structure of a sentence (syntactic/dependency parse) especially involving the spans of named entities that occur in it [1,12,14,16,32]. Even though such systems are popular for relation extraction, they are error prone and might result in extraction of coincidental outcomes that cannot be considered general knowledge. Furthermore, implicit relations that are not necessarily asserted in a sentence cannot be obtained through such approaches. However, NLP extractions can be used as a basis to develop more advanced techniques that aim toward a global relation prediction modeling paradigm. This process of distilling the literature and gleaning actionable information that drastically reduces researcher efforts in dealing with information explosion is termed as literature-based discovery (LBD).

An alternative approach for the LBD is distant supervision [22] (also called "weak supervision") when there are many predicates and manual labeling of sentences with relations is impractical. In this approach, pairs of entities, which come from a high quality knowledge base, are known to participate in specific relations and are used to search literature to identify sentences that contain both of them. Such sentences are used as training instances for the corresponding predicates to learn lexico-syntactic patterns that could be used as features in supervised models or in ranking new relations using unsupervised approaches [40]. Although distant supervision offers a convenient approach to overcome the labeled data scarcity issue, a disadvantage is that existence of a pair of entities in a sentence does not directly mean that the

sentence is expressing the particular relation existing in the knowledge base. Another important disadvantage is that the knowledge base could be incomplete and hence negative example pairs (those that do not participate in a relation in the knowledge base) may not be true negatives. Even though these disadvantages were comparatively mentioned and obviated by some other approaches [29,31,33,41], few researchers have addressed these issues especially in biomedicine. In our current effort, we propose an approach that is very different from these existing popular methods by relying on graph path patterns extracted from a large graph of extracted relations using NLP approaches. As such, our approach is not "extracting" a relation form a particular sentence, but is rather "predicting" a relation based on graph patterns connecting the corresponding entities.

A well-known method for LBD is to utilize the *pathway schemas* or *discovery patterns* [13] along with relations extracted by an NLP system such as NLM's SemRep program [30]. These discovery patterns are essentially designed by domain experts and typically connect a drug and a target disease. The task is then to identify a list of pairs connected by the thusly designed pathways [42]. However, this assumes the availability of experts' time and can be tedious with scalability issues specifically if it were to be undertaken for multiple diseases and other predicates besides *treats*. Recent approaches leverage distributional semantics to automatically infer discovery patterns. Specifically, the predication-based semantic indexing [9,10] approach identifies specific patterns by encoding entities and relations linking them in a vector space using the random indexing approach [8]. This approach typically outputs a ranked list of patterns the top few of which are used to retrieve potential new treatments for target diseases. Our approach deviates from this retrieval style framework involving a few top patterns by taking a traditional probabilistic route to output a probability estimate of a treatment relation holding between an input drug-disease pair by considering all patterns that connect them. As such, our approach starts with a closed discovery assumption [38] but can later be adapted to an open discovery scenario by identifying concept pairs connected by top patterns discovered under the closed assumption (details in Section 6). Besides treatment relations, identifying the cause of a condition or symptom is also of significance from a prevention perspective. Hence causative relations are also included in well known knowledge bases [15,25] and also used in discovery applications [36,42]. Thus in this effort our focus will be on predicting both treatment and causative relations.

3. Semantic patterns over knowledge graphs

In this section, we describe the graph pattern features and their extraction for predicting relations. Our basic intuition is simple: different entity pairs participating in a particular relation type (that is, linked via a specific predicate) are potentially connected in "similar" ways to each other where the connections are paths between them in knowledge graphs extracted from scientific literature. This is analogous to the NLP variant where a particular type of relation manifests with certain lexico-syntactic patterns surrounding the entity pair mentions in free text, the central idea exploited in distant supervision. In our approach, we need two essential components:

1. a broad scoped and large knowledge graph over which paths connecting candidate entity pairs can be obtained and
2. an approach to identify similar paths connecting entities, so we can abstract or "lift" specific paths to high level semantic graph patterns to be subsequently used as features in a supervised classifier.

3.1. SemMedDB knowledge graph

For this effort, we build a large knowledge graph of relations obtained from SemMedDB [15,27], a large database of (subject, predicate, object) relationships extracted from biomedical citations (titles & abstracts). SemMedDB is a public resource made available by the NLM, which uses their NLP tool SemRep [26,30] to extract "semantic predications" from biomedical text. SemMedDB is produced by running SemRep on all biomedical citations made available thorough the PubMed search system. The relations recorded in this database are called semantic predications given SemRep normalizes textual mentions of entities to unique UMLS Metathesaurus concepts (that is, performs named entity recognition) and the predicates are also based upon those available in the UMLS semantic network [23]. Each of the UMLS concepts also has at least one semantic type [24], which is essentially a classification system to categorize different biomedical entities. As such, the relations in SemMedDB represent a semantic summary of biomedical citations currently indexed by the PubMed search system. Our knowledge graph is essentially a directed graph with labeled edges formed from the relations in SemMedDB. The scope of this graph is very broad in a thematic sense given its edges are not limited to a particular biomedical topic. It is also large in that it has 14.3 million unique edges[1] connecting over 3 million nodes. It has already been used for literature based discovery and analysis of clinical documents [6,7,19,42,43].

3.2. Specific paths & semantic patterns

To abstract specific paths between entities over the SemMedDB graph to semantic patterns, we exploit an intuitive heuristic – simply replace the concepts along the path with their corresponding semantic type sets (given a concept can have more than one type) and retain the directions of the edges and edge labels as they are. For example, consider a sample graph showing a couple of paths between the drug `Lexapro (L)` and the condition `major depressive disorder (MDD)` in Fig. 1. We only employ simple paths (that is, without cycles) and ignore directionality when computing paths (but retaining it after paths are identified). Thus, we have the following two paths between `L` and `MDD`: (L, ingredient_of^{-1}, E, is_a, SUI, treats, MDD) and (L, ingredient_of^{-1}, E, treats, ND, treats^{-1}, SUI, treats, MDD), where the intermediate nodes are `Escitalopram` (E), `Serotonin Uptake Inhibitors` (SUI), and `Nonulcer Dyspepsia` (ND).

For notational convenience we encode the reverse direction with a superscript of −1 on the predicate. To obtain the patterns, we replace the specific entities with their semantic type sets. Thus, the corresponding two patterns are

$$(ingredient\,of^{-1},\{oc,ps\},isa,\{ps\},treats) \quad (1)$$

and

$$(ingredient\,of^{-1},\{oc,ps\},treats,\{f\},treats^{-1},\{ps\},treats), \quad (2)$$

where *oc, ps,* and *f* are abbreviations of the semantic types *organic chemical*, *pharmacologic substance*, and *finding* respectively. A pattern of length l (i.e., based on a path of length l) has l predicates and $l-1$ semantic types in the representation we use for this effort as shown in these examples (Eqs. (1) and (2)). Note that patterns do not include the entities being connected, but only the semantic types of the intermediate notes and the predicates along the path. By replacing specific intermediate entities with their semantic types we aim to capture high level patterns that connect candidate entity pairs. Although we just showed two paths, there are usually many others with a variety of edge types (over 50 different predicates) connecting related entities. We reiterate that our main hypothesis is that these patterns will act as highly discriminative features in identifying entity pairs that participate in a particular type of relationship. Here we clarify that although we refer to the SemMedDB graph as a knowledge graph (for general

[1] Although SemMedDB (Ver. 22) has over 63 million relations, there are many duplicates given a relation can be extracted from multiple sentences due to the semantic mapping to UMLS concepts and semantic network predicates.

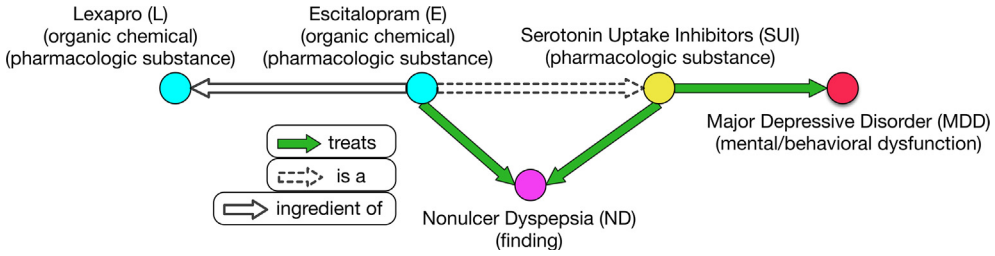

Fig. 1. A sample graph of biomedical relations.

understanding), the precision of NLM's SemRep tool used to build SemMedDB is known to be around 75% [15]. However, the advantage of our approach is that our prediction is not directly dependent on the correctness of each and every relation in the knowledge graph, rather on the general patterns found within it. Hence, any knowledge graph with reasonable quality will suffice although high quality graphs should yield better results. This was also observed to be the case by Cohen et al. [10] in their distributional semantics framework.

For extraction of the paths from the knowledge graph, from our literature review, there are no efficient implementations for computing *all* simple paths of an arbitrary length between two given nodes in large graphs, although many well known algorithms (e.g., modified breadth first search) exist for identifying shortest paths. In general, finding all simple paths becomes extremely expensive with lengths greater than three simply because the number of such paths could increase drastically in dense graphs. Our implementation for lengths ≤ 3 is based on straightforward heuristics that maintain precomputed lists of neighbors for each node in the knowledge graph. Specifically, to determine length two paths between nodes $e1$ and $e2$, we simply look at nodes in $\mathcal{N}(e1) \cap \mathcal{N}(e2)$ where $\mathcal{N}(e)$ denotes neighbors of node e. To identify length three paths, we look for edge membership for pairs in $\mathcal{N}(e1) \times \mathcal{N}(e2)$ in our knowledge graph.

In this effort, we are exclusively interested in predicting treatment/causative relationships and hence we chose this particular example for *treats* predicate from Fig. 1. The two example patterns we show here have a nice high level meaning. In the first pattern, we see that a *pharmacologic substance* (SUI) is a hypernym for another (E) (whose main ingredient is the source (L)) and is known to treat a *dysfunction* (MDD). The second pattern is similar except that it has two pharmacologic substances (SUI and E) both treating a common second condition (ND) while one of them (SUI) treats the target condition (MDD) and the source (L) is the ingredient of the other. However, in general, the patterns themselves do not need to have interesting or meaningful interpretations, but when considered together they should be reasonably predictive of the particular predicate that is of interest to us. In this specific example, it turns out that the treatment relationship also holds for our candidate pair (L, MDD). Essentially, we expect to leverage machine learned models to automatically weight different patterns based on their predictive power rather than human experts having to manually identify interesting patterns, a highly impractical task with the explosion of biomedical knowledge.

3.3. Primitive semantic type patterns

Henceforth we call the patterns discussed in Section 3.2 *compound* type patterns given an entity is replaced with the set of all semantic types assigned to it. However, there is a different way to look at semantic patterns where we split these compound patterns into potentially multiple primitive patterns to generate simpler and more generic patterns. In order to generate *primitive* patterns, we replace each set of types for the nodes in the compound pattern with just one of the constituent semantic types. Thus, we derive primitive patterns from the compound patterns simply by considering all possible combinations of constituent semantic types for each entity in the compound patterns. If

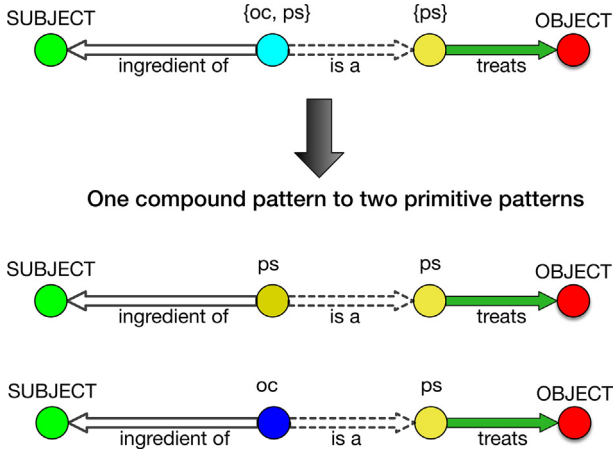

Fig. 2. An example of primitive patterns generated from a compound pattern (from Eq. (1)).

we consider the first pattern in Eq. (1) as an example, the derived two primitive patterns will be as in Fig. 2. So for a compound pattern of length l, the number of corresponding primitive patterns is $\prod_{i=1}^{l-1} |\mathcal{S}(e_i)|$, where e_i are intermediate nodes along the path and $\mathcal{S}(e)$ denotes the set of semantic types for entity e in the UMLS. Entities joined by the original compound pattern are now considered to be connected by all the primitive patterns generated from it. The primitive patterns form a more generic feature space when compared with their compound counterparts.

4. Datasets for *treats* and *causes* predicates

In this section we outline how we chose positive and negative examples to build the two datasets for experiments with graph pattern features introduced in Section 3.

4.1. Positive examples from the Metathesaurus

We derive our positive examples dataset from the UMLS Metathesaurus's MRREL table [25] that has over 26 million manually curated relations that are sourced from different biomedical terminologies. Among these we also have several *treats* and *causes* relations which are used for our experiments. We needed an external human vetted resource like the relations in UMLS given our knowledge graph is derived from a computationally curated relation database. We curated a set of around 7000 unique treatment relations (entity pairs connected through the *treats* predicate) and 2918 unique causative relations (entity pairs connected through the *causes* predicate) connecting UMLS concepts from the MRREL table. For each predicate, we divided positive example datasets into 80% (5600 for *treats* and 2334 for *causes*) forming the training set and 20% (1400 for *treats* and 584 for *causes*) constituting the test set split. Although there were more positive examples in MRREL, these counts are based on pairs that had at least one path connecting them in the SemMedDB

graph. This is necessary given we cannot make any prediction (given there is no information) if the entities are not connected in the graph from which we plan to extract patterns.

4.2. Selection of negative examples

Considering concerns identified in Section 2 to select negative examples for distant supervision we carefully choose negative examples in our dataset using the following two steps.

1. Every predicate in the UMLS semantic network, including *treats* and *causes*, has a set of domain/range semantic type constraints. That is, based on expert consultation NLM prescribes which types of entities can take the role of the subject and object for each relation. All such possible and allowable subject-object semantic entity type combinations for each predicate are available in three tables with the SRSTR prefix [25] in the UMLS. We first randomly select a pair of entities (from over 3 million unique UMLS concepts) that satisfies these domain/range constraints for the predicate for which we want to build the pattern based model.
2. For each pair selected in step 1, we check to see if the pair is connected via the predicate of interest to us either in the UMLS MRREL table or in the SemMedDB relation database. If it does not already occur in our knowledge bases, we include it as a negative example in our dataset.

This two-step process selects fairly hard-to-classify negative examples since they satisfy the domain/range constraints but do not participate in a relationship represented by the predicate for which we want to built the model. Checking for membership in both the UMLS and SemMedDB resources minimizes concerns surrounding incomplete knowledge bases. Since we want to predict treatment (causative) relations based on graph patterns, if the knowledge graph already has a *treats* (*causes*) edge between our candidate pairs, the prediction could become trivial and the whole process would be self-deceiving. Therefore, we deleted any existing *treats* (*causes*) edges between entities in all training/test positive pairs from the knowledge graph (note that negative example selection already ensures this) to guarantee a fair analysis of the predictive ability of graph patterns.

5. Experiments and results

Elaborate experimentation is essential to identify performance trends across different aspects of our relation prediction problem including dataset constitution and model features and parameters. In this section we first outline some experimental configuration basics before moving on to specific models built for this effort. We start with the LR model and build upon its findings to experiment with decision tree models.

5.1. LR model configurations and findings

We use the well known LR algorithm to predict whether an input pair of entities participates in treatment or causative relationship by building two separate binary classification models. The features for the LR models are frequencies of patterns connecting the input entity pair as discussed in Section 3. The specific implementation used is the LR classifier based on the LIBLINEAR formulation made available through the Python scikit-learn [28] machine learning library. Parameter tuning for the regularization coefficient in the LR model did not yield any noticeable gains and hence we chose to leave it at $C = 1$, the default value in scikit-learn. Performances assessment in this effort are based on standard measures of precision, recall, and F-score. All experiments were repeated using **hundred distinct 80–20% train-test splits** of the full dataset so as to account for chance and to derive average scores and confidence intervals.

In our earlier paper [3], we experimented with a balanced training dataset (equal number of positive and negative instances) considering imbalanced scenarios for our test dataset. In the universe of all pairs of entities that satisfy domain/range constraints for a predicate, most are going to be false. For *treats*, an arbitrary drug-disease pair would not have a treatment relationship. So we increased the numbers of negative examples in the test to double that of the positive examples. We extended this imbalance with positive:negative ratios of 1:5 and 1:10. With a balanced training dataset, the performance gradually decreased as the test set imbalance increased. We kept the training dataset balanced to ensure that there is enough signal for the model to learn patterns for positive instances. This style of oversampling of the positive class is not uncommon in these cases where the class we care about is a rare one. In our preliminary results the performance also increased with the length of the patterns. That is, considering all patterns of length ≤ 3 resulted in better F-score when compared with considering patterns of length ≤ 2 or one. All of our prior experiments outlined in our conference paper were done with compound patterns.

In this extended version, we always keep the 1:10 imbalance in the test set given large imbalance is inherent to the true distribution for *treats/causes*. We then experiment with various imbalance scenarios in the training dataset. This is to see whether increasing the number of negative instances in the training dataset would result in performance gains on the imbalanced test dataset. The negative examples are chosen as discussed in Section 4.2. The full dataset size depends on the training dataset imbalance selected. For example, for the balanced training dataset and 1:10 imbalanced test set scenario, for the *treats* predicate, we have 7000 positive examples (5600 for training, 1400 for test) and 19,600 negative examples (5600 for training, 14,000 for test). When the imbalance is 1:10 in both training and test datasets, the corresponding counts are 7000 positive examples (5600 for training, 1400 for test) and 70,000 negative examples (56,000 for training, 14,000 for test). Note that these count configurations are limited by the number of positive examples available (Section 4.1).

Another parameter to select is the number of patterns to be included in the feature space. When classifying text with word *n*-gram features, researchers typically ignore all *n*-grams whose frequency is less than a small threshold (mostly set to five). That is, all *n*-grams that occurred in fewer than five documents (regardless of class membership) are ignored in populating feature vectors. We have a similar situation here with an overwhelming number of patterns of length ≤ 3 connecting entities that have a *treats* or *causes* relation. We had over 50 million unique compound patterns for *treats* and nearly 25 million such patterns for the *causes* dataset. To reduce noise and address computational efficacy concerns, we chose those patterns that occurred as connectives for at least 500 entity pairs for *treats* and 100 pairs for *causes* in the corresponding datasets. This rendered the feature spaces to manageable sizes of around 600,000 unique patterns for *treats* and 200,000 for *causes*.

The overall architecture of our method is shown in Fig. 3. Although we are currently discussing LR models, any supervised learning algorithm can be used with graph pattern based features.

5.1.1. Balanced training dataset scenario

As we mentioned earlier in this section, the balanced models have equal number of positive and negative examples in training dataset; the test set always has ten times as many negative examples as the positive ones to model realistic scenarios. In Table 1, we show the average precision, recall, and F-scores computed over hundred distinct splits of the full dataset for *treats*. The performance gains between the 1000 and 500 pattern frequency thresholds are not substantial. We see a precision gain of around 4% and a recall loss of 0.3% for each threshold when using primitive patterns over compound patterns.

Performances when using primitive patterns are also superior for *causes* as shown in Table 2 except for the higher pattern frequency threshold of 1000. The actual F-scores are lower for causative relations when compared with treatment relations. The 95% confidence intervals

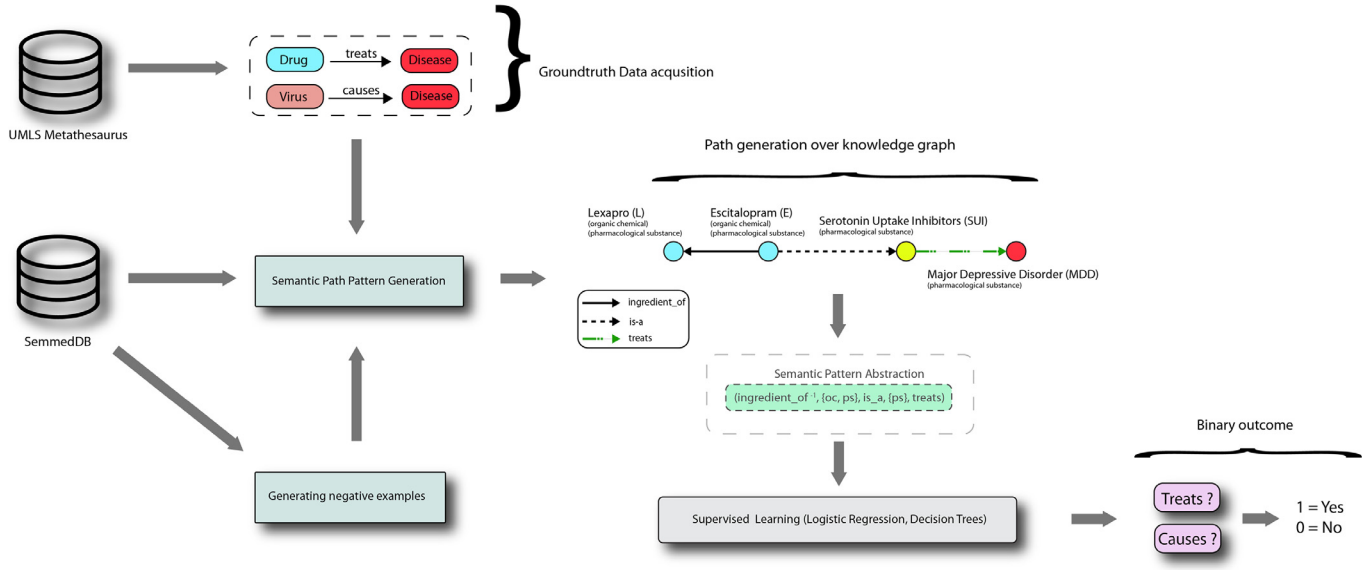

Fig. 3. Schematic of graph pattern based relation prediction.

we computed for F-scores have widths ≈ 0.01 when using primitive and compound patterns; thus they do not overlap for both predicates. Thus overall, primitive patterns are more effective for the balanced training dataset scenarios.

5.1.2. Imbalanced training dataset scenarios

While keeping the 1:10 positive to negative class test imbalance, we wanted to see the effect of increasing the imbalance in the training dataset in contrast with the scenario in Section 5.1.1. From Tables 3 and 4, we notice that the imbalance setting where $|N| = 10 \cdot |P|$ in training datasets gives the best overall F-score when compared with situations with less imbalance (including in comparison with top scores in Tables 1 and 2). Furthermore, the 95% confidence interval widths for the top F-scores in Tables 3 and 4 are very small – 0.0011 (for *treats*) and 0.0036 (for *causes*). The improvements are not as substantial for *treats* but are prominent for *causes* when training set imbalance is increased; for the latter predicate, however, the recall goes down with increase in training set imbalance which is compensated by an increase in precision leading to an overall better F-score. Lowering the minimum pattern frequency yields marginal improvements for *treats* compared with corresponding gains for *causes*.

Note that our improvements in Tables 3 and 4 with imbalanced training datasets are using compound patterns. Contrary to our observations in balanced training dataset scenarios (Section 5.1.1), we noticed that compound patterns provided major gains over primitive patterns for the imbalanced scenarios. Furthermore, the number of patterns is substantially higher for primitive patterns (at least twice as many) leading to additional efficiency concerns. We show our observations for *causes* in Table 5 when using primitive patterns. When comparing these scores with those in Table 4, it is clear that compound patterns are better overall in imbalanced training dataset cases, which offer the best case for improving test score performances. The 95% confidence intervals we computed for F-measures in the last row and last column in Tables 4 and 5 have widths <0.01 and hence do not overlap. Thus the improvements with compound patterns are statistically significant. Although we do not show the results here, we observed a similar trend with compound patterns outperforming primitive patterns when considering patterns of length ≤ 2 for both *treats* and *causes*. We believe this reversal in performance trend for primitive patterns is due to the fact that imbalanced training datasets lead to an explosion of unique patterns from the negative examples. When this happens, the generic and simpler primitive patterns may lose their discriminative power in comparison with the more specific compound patterns.

Table 1
Balanced training data: test set scores with patterns of length ≤ 3 for treatment relations.

Pattern Type	Min. Frequency: 1000			Min. Frequency: 500		
	Precision	Recall	F-score	Precision	Recall	F-score
Compound	0.675	0.926	0.781	0.683	0.928	0.786
Primitive	0.717	0.924	0.807	0.721	0.925	**0.810**

The bold scores indicate the highest values of F-score (indicating best performances of the models).

5.2. Experiments with decision trees

In Section 5.1, we exclusively studied application of logistic regression models to predict relations. In this section, we discuss additional experiments we conducted with decision trees [4] to explore nonlinear models that are also interpretable. We use the same approach as in Section 5.1 to come up with average scores over hundred distinct runs with 80% used for training and 20% for testing. Our results are shown in Table 6 for the imbalanced training dataset scenario with 1:10 imbalance in the test set (given this turned out to be the best configuration based on results from Tables 3 and 4). Given deeper trees can model more complex relationships by trading off interpretability, we experimented with scenarios where the maximum depth is restricted to five and when it is left unconstrained. As can be noticed from the table, the recall is much better when depth is not constrained. The scores are also slightly better than the best results obtained through LR models (from Tables 3 and 4).

5.3. Recall analysis using an external dataset: repoDB

Our experiments reported thus far involved positive relations that are well known and recorded in the UMLS. Although our performance scores reported are on held-out datasets, it is possible that patterns connecting well known relations may not be present in newly discovered relations. repoDB [5] is a database that has over 6000 FDA approved drugs and corresponding indications collected from the regularly updated DrugCentral platform [35]. As such repoDB is expected

Table 2
Balanced training data: test set scores with patterns of length ≤ 3 for causative relations.

Pattern Type	Min. Frequency: 1000			Min. Frequency: 500			Min. Frequency: 100		
	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
Compound	0.446	0.744	0.554	0.472	0.776	0.583	0.478	0.811	0.598
Primitive	0.400	0.736	0.518	0.510	0.756	0.609	0.546	0.791	**0.645**

The bold scores indicate the highest values of F-score (indicating best performances of the models).

Table 3
Imbalanced training data: test set scores with *length* ≤ 3 compound patterns for treatment relations.

Imbalance in training set	Min. Frequency: 1000			Min. Frequency: 500		
	Prec.	Rec.	F-score	Prec.	Rec.	F-score
$\|N\| = 2{\cdot}\|P\|$	0.979	0.962	0.970	0.981	0.964	0.973
$\|N\| = 4{\cdot}\|P\|$	0.988	0.964	0.976	0.988	0.966	0.977
$\|N\| = 10{\cdot}\|P\|$	0.992	0.966	0.979	0.992	0.968	**0.980**

The bold scores indicate the highest values of F-score (indicating best performances of the models).

to contain the latest FDA approved drugs. To test our best (both LR and decision tree) models,

1. We removed UMLS treatment relations from the list of all FDA approved drug-disease indications from repoDB.
2. We also removed relations whose entities are already connected through a *treats* edge in SemMedDB.
3. Out of the remaining approved drug-disease relations, we removed those that do not have at least one path connecting the involved entities in the SemMedDB graph from which we derived our patterns.

After these filtering processes, we were left with 2739 new treatment relations. We built 100 different models for the *treats* predicate based on positive examples used in Sections 5.1 and 5.2 with a fresh set of negative examples chosen for each of the models. Next, we computed the average recall by running them on these 2739 instances we obtained from repoDB. Our results shown in Table 7 indicate that we are able to recall over 50% of the approved drugs that are at least connected with one path in SemMedDB. Decision trees (without max depth constraints) proved to be much better than LR models. Primitive patterns seems to help LR models while both types of patterns resulted in similar performances when using decision trees. Using imbalanced training data that gave us over 95% F-scores for held-out UMLS treatment relations turned out to be ineffective to retrieve repoDB relations. Under-sampling the majority class when the minority class is of high relevance is a tested method [37] and appears to work well for our situation too.

6. Qualitative analyses of LR models: informative patterns and new hypotheses

In Section 5, we focused exclusively on quantitative evaluation of our methods and showed that best results are obtained by using compound patterns and imbalanced training datasets for both predicates. But it is also important to analyze the patterns qualitatively in terms of their discriminative power and their suitability in discovering previously unknown relations.

6.1. Exploring highly discriminative patterns

In order to assess the predictive contribution of different graph pattern features, we conducted an additional experiment to identify patterns that correlate well with positive examples. During the process of building hundred different models based on hundred distinct 80–20% train-test splits of the full datasets, we stored model coefficients for all features. Subsequently, we ranked all patterns based on the average coefficient value across the hundred models. If β is the model coefficient of a pattern in the LR model, we know e^{β} is the odds ratio of that pattern with respect to the positive class [17]. Hence, ranking patterns in the descending order of the corresponding average model coefficient values is equivalent to ranking them based on their importance toward the positive prediction for the corresponding predicate. Thus we ranked all patterns accordingly and made them available as supplementary materials along with this manuscript. The patterns can also be searched and visualized using an online interface: http://patterns.mgokhanbakal.net/. In order to assess the sensitivity of the top patterns, we considered the top hundred patterns selected as per this ranking. For *treats*, the top 100 patterns cover 43% of the 7000 instances. For *causes*, the top 100 connected 25% of the full positive instance dataset. This indicates that our method is able to identify high quality patterns that can be used to query knowledge graphs for generating potential new hypotheses.

Another objective is to manually examine these patterns and see if they are meaningful or informative in some sense. We show some interesting patterns in Fig. 4 from our full pattern list for *treats* predicate. Patterns P1–P4 are those obtained from the top 100 patterns among nearly 600,000 unique patterns ranked. P1 indicates the situation where two drugs treat a common condition (node 2) and given one of them (node 3) treats our target condition, it is also plausible for our source substance to treat the target. P2 has a similar structure except we

Table 4
Imbalanced training data: test set scores with *length* ≤ 3 compound patterns for causative relations.

Imbalance in training set	Min. Frequency: 1000			Min. Frequency: 500			Min. Frequency: 100		
	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
$\|N\| = 2{\cdot}\|P\|$	0.744	0.724	0.732	0.868	0.775	0.819	0.865	0.846	0.855
$\|N\| = 4{\cdot}\|P\|$	0.851	0.698	0.766	0.922	0.760	0.833	0.924	0.837	0.878
$\|N\| = 10{\cdot}\|P\|$	0.950	0.646	0.769	0.967	0.745	0.842	0.967	0.816	**0.885**

The bold scores indicate the highest values of F-score (indicating best performances of the models).

Table 5
Imbalanced training data: test set scores with *length* ≤ 3 primitive patterns for causative relations.

Imbalance in training set	Min. Frequency: 1000			Min. Frequency: 500			Min. Frequency: 100		
	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
$\|N\| = 2{\cdot}\|P\|$	0.531	0.685	0.597	0.626	0.710	0.665	0.667	0.728	0.696
$\|N\| = 4{\cdot}\|P\|$	0.619	0.667	0.642	0.721	0.684	0.702	0.761	0.689	**0.723**
$\|N\| = 10{\cdot}\|P\|$	0.762	0.597	0.669	0.799	0.623	0.700	0.817	0.621	0.705

The bold scores indicate the highest values of F-score (indicating best performances of the models).

Table 6
Imbalanced training data: average test set scores using decision trees with compound patterns.

Predicate type	Max depth = 5			No depth constraint		
	Precision	Recall	F-score	Precision	Recall	F-score
Treats	0.998	0.815	0.897	0.994	0.987	**0.990**
Causes	0.994	0.506	0.669	0.922	0.887	**0.904**

The bold scores indicate the highest values of F-score (indicating best performances of the models).

Table 7
Balanced training data: average recall on repoDB.

Model	Compound	Primitive
LR	18.6%	43.1%
Decision tree	**52.9%**	50.7%

have a common therapeutic procedure that uses the two medications (a source and another intermediate antibiotic). P3 involves the patient group semantic type (e.g., cancer patients) that is connected to a condition via the *process-of* predicate. It also uses a class membership relation as the first edge to form a meaningful pattern connecting the instance of a class of drugs to a target condition affecting a patient group. P4 involves two conditions (an intermediate one and the target condition) that share an immunologic factor and the pattern connects the source to the target via a treatment relation involving the intermediate condition. Thus we see that patterns identified through our approach appear to have an intuitive semantic interpretation.

Patterns P5–P8 are also high scoring patterns that appeared in the top 1% of the full ranked list. We show them in the figure given a recent effort by Cohen et al. [10] also identified them as top scoring reasoning pathways for cancer therapies. In fact, all pathways identified by them show up in the top 1% of our ranked pattern list. We do not show the semantic types of intermediate nodes given Cohen et al. do not consider types as part of their reasoning pathways. So each of their pathways can match multiple patterns in our list; hence, we show counts of our unique patterns (or unique type combinations) that match the corresponding pathway in parentheses next to the ID for P5–P8 in Fig. 4. However, as we pointed out in Section 2, Cohen et al.'s work takes a retrieval approach to identify a few top patterns, while we focus on building a high accuracy predictive probabilistic model that is also interpretable through its feature coefficients.

One interesting observation here is that most of the patterns for treatment relations shown in Fig. 4 have a *treats* edge in them. In fact, among the top 1000 treatment (causative) relation patterns 646 (241) contain a *treats* (*causes*) edge. This is not surprising for treatment relations given certain drugs and procedures treat clusters of diseases that share certain characteristics. Thus, even though the predicted relation may not be there in the SemMedDB graph, other treatment relations involving the subject medication might be indicative of its therapeutic potential for the target condition. Intuitively, this also conveys the general motivation behind computational drug repurposing efforts that are popular these days [2,18,39]. Another aspect of note is that most top patterns are of length three. Of the top 10,000 patterns for each predicate, the count of length two patterns is only 24 for *treats* and 47 for *causes*. This might simply be because of the fact that, in general, paths of length three are much more common than length two associations in SemMedDB. Hence length three patterns offer a much larger feature space to exploit for our predictive models.

6.2. Discovering new relations

Our evaluations thus far focused on hand curated relations already recorded in the UMLS or repoDB. However, we thought it would be more interesting to see if our approach can discover new plausible

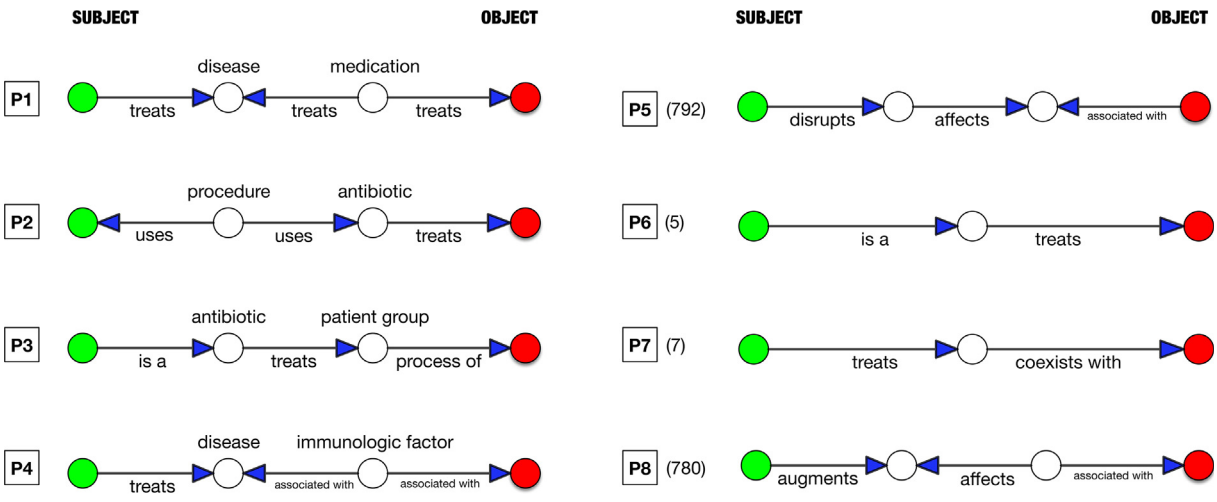

Fig. 4. Example discriminative *treats* patterns obtained through our methods.

relations that are not already known. From Sections 5.1.2 and 5.2, it is clear that our approach achieves very high precision for both treatment and causative relations. However, we still have some false positives (FPs). The intuition is that high scoring FPs could actually be plausible new relations that are not already known to the medical community. To this end, we chose 10,000 new negative examples for *treats* and *causes* that are not part of the negative examples chosen to be in our training datasets used in Section 5 and the additional examples used in this section. We built 100 different models for each predicate changing only the negative examples as was done in Section 5.3. We applied these hundred models to each of the 10,000 negative examples chosen for this experiment.

Next, we needed a way to carefully choose high confidence positive predictions for expert review. For this, we finalized a list of all negative examples that were predicted as positive by at least 90 models (that is with output probability ≥ 0.5) and have an average probability estimate of at least 0.9 (overall hundred models). This process resulted in a total of 181 instances for *treats* and 138 instances for *causes*. These potentially new relations were independently reviewed by two practicing inpatient physicians (co-authors Drs. Talari and Kakani) at the University of Kentucky hospital for biomedical plausibility. After independent annotations, both physicians resolved their disagreements. 33% of *treats* FPs and 28% of the *causes* FP instances examined were deemed plausible by the physicians. Thus we are able to identify relations that are not in our knowledge bases but are still plausible. However, we needed experts to also assess novelty based on their knowledge. Although these FPs are plausible positive instances, they could just be common knowledge to experts and are simply not available in UMLS and SemMedDB. Among the *treats* instances that were deemed plausible, 68% were also identified as potentially novel findings by the physicians. This proportion is only 5.5% for *causes*; so most plausible FPs were already known to the experts despite their absence in our sources.

Among the manually reviewed FP examples, the experts chose a few plausible and novel examples for each relation (*treats* and *causes*) and came up with the corresponding plausibility explanations as follows. This was done to simulate the discovery process using our approach and offers additional evidence of practical relevance of our methods.

Plausibility of new treatment relations identified.

1. **Gentamicin Sulfate** $\xrightarrow{TREATS}$ **Anthrax disease**: Gentamycin Sulfate is an antibiotic that is used on the outside of the body (topical). It belongs to aminoglycoside class of antibiotics. It acts by disrupting the normal cycle of ribosomes, which are the structures present inside a cell. This disrupts initiation of protein synthesis inside the cells. These antibiotics are directed primarily against aerobic gram-negative bacilli class of bacteria but have limited activity against gram-positive class of bacteria. Bacillus Anthracis bacteria causing Anthrax is classified as gram-positive rod. Cutaneous Anthrax disease can be potentially treated by topical gentamicin sulfate with the above rationale.
2. **Orbifloxacin** $\xrightarrow{TREATS}$ **Dysentery — Infectious Diarrhea**: Orbifloxacin is an antibiotic mainly used in animals. It belongs to fluoroquinolone class of antibiotics. Fluoroquinolone class of antibiotics are used in human beings for the treatment of Dysentery and Infectious diarrhea. Orbifloxacin is a fluoroquinolone antibiotic, so there is a biological plausibility for it be used in human beings for the treatment of Dysentery and Infectious diarrhea, as the mechanism of action of these group of drugs is the same.
3. **Zorubicin** $\xrightarrow{TREATS}$ **Acute Myelomonocytic Leukemia**: Zorubicin is a medication that belongs to Anthracyclin class of drugs. Anthracyclin class of drugs are used in the treatment of cancers including leukemia. Therefore, it is biologically plausible for it to be used in the treatment of acute myelomonocytic leukemia.
4. **Ziconotide** $\xrightarrow{TREATS}$ **Nonspecific Urethritis**: Ziconotide is a synthetic peptide related to the marine snail toxin ω-conotoxin, which selectively blocks N-type calcium channels at the cellular level. It is used in patients with chronic pain by injecting this substance into the spinal canal. With this rationale, this drug can be used to treat pain from Nonspecific urethritis as well through the same mechanism of action.
5. **Miocamycin** $\xrightarrow{TREATS}$ **Staphylococcus Aureus Pneumonia**: Miocamycin is an antibiotic that belongs to Macrolide class antibiotics. Macrolide antibiotics have activity against many classes of bacteria including gram-positive cocci class of bacteria. Staphylococcus Aureus is a gram-positive cocci class of bacteria. With this rationale, miocamycin can be used to treat Pneumonia caused by Staphylococcus Aureus bacteria.

Plausibility of new causative relations identified.

1. **Human Metapneumovirus** $\xrightarrow{CAUSES}$ **Systemic Lupus Erythematosus (SLE)**: The etiology of SLE is unknown and is probably multifactorial. Interplay of genetic predisposing factors, environmental factors, and hormonal factors is thought to play a role. Among environmental factors, various viruses are thought to stimulate the body's immune network. For example, people with SLE are known to have high levels of autoantibodies to Epstein Barr virus and certain retroviruses. Thus the role of immune response to Human metapneumovirus infection in the etio-pathogenesis of SLE is a topic that warrants additional exploration.
2. **Maternal Fetal Infection Transmission** $\xrightarrow{CAUSES}$ **Autoimmune Diseases**: The etiology of many autoimmune diseases is unknown. During immune development in the fetus, maturing lymphocytes in thymus and bone marrow are exposed to several antigens and those immune cells reacting to self-antigens are selectively inactivated via apoptosis (programmed cell death) or by induction of anergy. Thus the involvement of fetal infection during gestation with the process of self-antigen recognition is worth further analysis.
3. **Human Herpes Virus 6** $\xrightarrow{CAUSES}$ **IgG Gammopathy**: Human herpes virus 6 (HHV-6) was first isolated in patients with lymphoproliferative disorders. HHV-6 infection has been associated with a prolonged mononucleosis like syndrome with prolonged lymphadenopathy and encephalitis. Associations between HHV-6 and diseases such as multiple sclerosis and neoplasia have been proposed but remain unproven. HHV-6 antigens and DNA have reportedly been detected in some malignant tissues such as lymphomas. Hence HHV-6 may play a role in IgG gammopathies (increased immunoglobulins belonging to Ig-G class due to abnormal proliferation of some bone marrow cells) such as Monoclonal gammopathy of uncertain significance (MGUS) deserving additional attention.

7. Limitations

Despite the evidence of effectiveness of our models from Sections 5 and 6, our methodology has a few important limitations.

1. A main caveat of our approach pertains to the "true" recall of our models. Specifically, in Sections 4.1 and 5.3 we make an important assumption that the positive examples we chose for our experiments must be connected in the SemMedDB graph with at least one path. Otherwise, our feature vector will be a zero vector and will have no information for prediction. Similar to the assumption that the relation must be expressed in the sentence for NLP approaches and the assumption that candidates must co-occur in at least one sentence for pattern based distant supervision approaches, our assumption is also reasonable. In fact, our constraint is less restrictive than those of other approaches given we do not require subject-object co-occurrence but that they simply be participants in relations with other entities and have enough shared context to have a path of length

≤ 3. Nevertheless, our recall values should be qualified with the constraint of being relevant only for recoverable instances.

2. Another issue with our core method is the need to extract paths connecting entities of length $\leq k$. Although we handled $k = 3$ using straightforward heuristics, we are not aware of a simple way to do the same for $k > 3$. One could argue that longer paths may not be essential and may adversely affect the prediction accuracy. However, even for $k = 3$, as the knowledge graph becomes denser with new findings connecting more and more entities in it, the simple task of enumerating all paths can be very expensive. In this case, one might need to resort to effective heuristics to prune which edge types to include in the analyses, a common practice in graph based knowledge discovery [7,10].
3. In our approach, we need to train a binary model for each predicate separately. This may be general practice when we are interested in specific relation types but ideally an approach that jointly learns a single model capable of predicting one or more viable predicates among several possible (say, the 54 predicates from the UMLS semantic network) may be preferable. Recent advances in neural networks especially with the multi-class or multi-label cross entropy loss [34] may enable such a unified model.

8. Conclusion

Treatment and causative relations are central to knowledge discovery in biomedicine. In this paper, we employed semantic graph patterns connecting pairs of candidate entities as the sole set of features to predict treatment and causative relations between them. We exploited a well-known biomedical relation database, SemMedDB, to build a knowledge graph with over 14 million edges extracted from scientific literature. We then used this graph to derive features and also select suitable negative training instances for predictive modeling experiments.

Evidenced by the results presented in Section 5, we have successfully verified our hypothesis that semantic patterns over knowledge graphs can be powerful predictors of treatment and causative relations. Specifically in Section 5.3 we demonstrated that supervised *treats* models trained with graph pattern features can also recall newly approved drugs along with the corresponding indications from an external dataset. In Section 6, we analyzed the top patterns informed by model coefficients and demonstrated their interpretability for gaining insights into the prediction process. Additionally, we sought human expert assessments to demonstrate the utility of the proposed approach in identifying potentially novel and previously unknown relations. Even though our central idea is straightforward and intuitive, it can be naturally generalized to other predicates such as *disrupts* and *prevents*, and also to other domains of interest where knowledge graphs of reasonable quality are available. In addition to potential extensions indicated in Section 7, an important research direction is to adapt our work to *n*-ary relations or events such as protein-protein interactions where besides the participating entities and the interaction type, additional attributes such as cell/tissue type where the interaction took place are also to be predicted. For now, our results in this effort demonstrate that semantic patterns over knowledge graphs hold great promise for global relation prediction in biomedicine.

Conflict of interest

The authors declared that there is no conflict of interest.

Acknowledgement

We thank reviewers for their constructive criticism that helped improved the quality of this manuscript. We are grateful for the support of the U.S. National Library of Medicine through NIH grant R21LM012274 and also thankful for partial support offered by the U.S. National Center for Advancing Translational Sciences via grant UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We also acknowledge the Ministry of National Education, Republic of Turkey, for providing financial support to Gokhan Bakal with full scholarship for his doctoral studies.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2018.05.003.

References

[1] Asma Ben Abacha, Pierre Zweigenbaum, Automatic extraction of semantic relations between medical entities: a rule based approach, J. Biomed. Semant. 2 (5) (2011) S4.
[2] Christos Andronis, Anuj Sharma, Vassilis Virvilis, Spyros Deftereos, Aris Persidis, Literature mining, ontologies and information visualization for drug repurposing, Brief. Bioinform. 12 (4) (2011) 357–368.
[3] Gokhan Bakal, Ramakanth Kavuluru, Predicting treatment relations with semantic patterns over biomedical knowledge graphs, International Conference on Mining Intelligence and Knowledge Exploration, Springer, 2015, pp. 586–596.
[4] Leo Breiman, Jerome Friedman, Charles J. Stone, Richard A. Olshen, Classification and Regression Trees, CRC Press, 1984.
[5] Adam S. Brown, Chirag J. Patel, A standard database for drug repositioning, Sci. Data 4 (2017) 170029.
[6] Delroy Cameron, Olivier Bodenreider, Hima Yalamanchili, Tu Danh, Sreeram Vallabhaneni, Krishnaprasad Thirunarayan, Amit P. Sheth, Thomas C. Rindflesch, A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications, J. Biomed. Inform. 46 (2) (2013) 238–251.
[7] Delroy Cameron, Ramakanth Kavuluru, Thomas C. Rindflesch, Amit P. Sheth, Krishnaprasad Thirunarayan, Olivier Bodenreider, Context-driven automatic subgraph creation for literature-based discovery, J. Biomed. Inform. 54 (2015) 141–157.
[8] Trevor Cohen, Roger Schvaneveldt, Dominic Widdows, Reflective random indexing and indirect inference: a scalable method for discovery of implicit connections, J. Biomed. Inform. 43 (2) (2010) 240–256.
[9] Trevor Cohen, Dominic Widdows, Roger W. Schvaneveldt, Peter Davies, Thomas C. Rindflesch, Discovering discovery patterns with predication-based semantic indexing, J. Biomed. Inform. 45 (6) (2012) 1049–1065.
[10] Trevor Cohen, Dominic Widdows, Clifford Stephan, Ralph Zinner, Jeri Kim, Thomas Rindflesch, Peter Davies, Predicting high-throughput screening results with scalable literature-based discovery methods, CPT: Pharmacometr. Syst. Pharmacol. 3 (10) (2014) e140.
[11] Mark Craven, Johan Kumlien, et al., Constructing biological knowledge bases by extracting information from text sources, in: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB), 1999, pp. 78–86.
[12] Katrin Fundel, Robert Küffner, Ralf Zimmer, RelEx – relation extraction using dependency parse trees, Bioinformatics 23 (3) (2007) 365–371.
[13] Dimitar Hristovski, Carol Friedman, Thomas C. Rindflesch, Borut Peterlin, Exploiting semantic relations for literature-based discovery, in: AMIA Annual Symp., 2006.
[14] Ramakanth Kavuluru, Christopher Thomas, Amit P. Sheth, Victor Chan, Wenbo Wang, Alan Smith, Armando Soto, Amy Walters, An up-to-date knowledge-based literature search and exploration framework for focused bioscience domains, Proc. of the 2nd ACM SIGHIT Health Informatics Symposium, ACM, 2012, pp. 275–284.
[15] Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, Thomas C. Rindflesch, Semmeddb: a pubmed-scale repository of biomedical semantic predications, Bioinformatics 28 (23) (2012) 3158–3160.
[16] Sun Kim, Haibin Liu, Lana Yeganova, W. John Wilbur, Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach, J. Biomed. Inform. 55 (2015) 23–30.
[17] Daniel G. Kleinbaum, Mitchell Klein, Logistic Regression: A Self-Learning Text, Statistics for Biology and Health, Springer New York, 2010.
[18] Jiao Li, Si Zheng, Bin Chen, Atul J. Butte, S. Joshua Swamidass, Zhiyong Lu, A survey of current trends in computational drug repositioning, Brief. Bioinform. 17 (1) (2016) 2–12.
[19] Ying Liu, Robert Bill, Marcelo Fiszman, Thomas Rindflesch, Ted Pedersen, Genevieve B. Melton, Serguei V. Pakhomov, Using semrep to label semantic relations extracted from clinical text, in: AMIA Annual Symposium Proceedings, vol. 2012, American Medical Informatics Association, 2012, p. 587.
[20] Zhiyong Lu, PubMed and beyond: a survey of web tools for searching biomedical literature, Database: J. Biol. Databases Curat. (2011).
[21] Yuan Luo, Özlem Uzuner, Peter Szolovits, Bridging semantics and syntax with graph algorithm – state-of-the-art of extracting biomedical relations, Brief. Bioinform. 18 (1) (2017) 160–178.
[22] Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky, Distant supervision for relation extraction without labeled data, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL, 2009, pp. 1003–1011.

[23] National Library of Medicine, Current Hierarchy of UMLS Predicates, 2003. < http://www.nlm.nih.gov/research/umls/META3_current_relations.html > .
[24] National Library of Medicine, Current Hierarchy of UMLS Semantic Types, 2003. < http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html > .
[25] National Library of Medicine, Unified Medical Language System Reference Manual, 2009. < http://www.ncbi.nlm.nih.gov/books/NBK9676/ > .
[26] National Library of Medicine, SemRep – NLM's Semantic Predication Extraction Program, 2013. < http://semrep.nlm.nih.gov > .
[27] National Library of Medicine, Semantic MEDLINE Database, 2016. < http://skr3.nlm.nih.gov/SemMedDB/ > .
[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (Oct) (2011) 2825–2830.
[29] Sebastian Riedel, Limin Yao, Andrew McCallum, Modeling relations and their mentions without labeled text, Machine Learning and Knowledge Discovery in Databases, Springer, 2010, pp. 148–163.
[30] Thomas C. Rindflesch, Marcelo Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, J. Biomed. Inform. 36 (6) (2003) 462–477.
[31] Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al., Modeling missing data in distant supervision for information extraction, Trans. Assoc. Comput. Linguist. 1 (2013) 367–378.
[32] Isabel Segura-Bedmar, Paloma Martinez, Cesar de Pablo-Sánchez, Using a shallow linguistic kernel for drug–drug interaction extraction, J. Biomed. Inform. 44 (5) (2011) 789–804.
[33] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, Christopher D Manning, Multi-instance multi-label learning for relation extraction, in: Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2012, pp. 455–465.
[34] Tung Tran, Ramakanth Kavuluru, Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks, J. Biomed. Inform. (2017) S138–S148.
[35] Oleg Ursu, Jayme Holmes, Jeffrey Knockel, Cristian G. Bologa, Jeremy J. Yang, Stephen L. Mathias, Stuart J. Nelson, Tudor I. Oprea, Drugcentral: online drug compendium, Nucl. Acids Res. 45 (D1) (2017) D932–D939.
[36] Özlem Uzuner, Brett R. South, Shuying Shen, Scott L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (5) (2011) 552–556.
[37] Byron C. Wallace, Kevin Small, Carla E. Brodley, Thomas A. Trikalinos, Class imbalance, redux, In Data Mining (ICDM), in: 2011 IEEE 11th International Conference on, IEEE, 2011, pp. 754–763.
[38] Marc Weeber, Henny Klein, Lolkje de Jong-van den Berg, Rein Vos, et al., Using concepts in literature-based discovery: simulating Swanson's raynaud–fish oil and migraine–magnesium discoveries, J. Assoc. Inform. Sci. Technol. 52 (7) (2001) 548–557.
[39] Hua Xu, Melinda C. Aldrich, Qingxia Chen, Hongfang Liu, Neeraja B. Peterson, Qi Dai, Mia Levy, Anushi Shah, Xue Han, Xiaoyang Ruan, et al., Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality, J. Am. Med. Inform. Assoc. (2014) amiajnl–2014.
[40] Rong Xu, Alex Morgan, Amar K. Das, Alan Garber, Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon, in: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Association for Computational Linguistics, 2009, pp. 63–70.
[41] Wei Xu, Raphael Hoffmann, Le Zhao, Ralph Grishman, Filling knowledge base gaps for distant supervision of relation extraction, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL, 2013, pp. 665–670.
[42] Rui Zhang, Michael J. Cairelli, Marcelo Fiszman, Halil Kilicoglu, Thomas C. Rindflesch, Serguei V. Pakhomov, Genevieve B. Melton, Exploiting literature-derived knowledge and semantics to identify potential prostate cancer drugs, Cancer Inform. 13 (2014) 103–111.
[43] Rui Zhang, Michael J. Cairelli, Marcelo Fiszman, Graciela Rosemblat, Halil Kilicoglu, Thomas C. Rindflesch, Serguei V. Pakhomov, Genevieve B. Melton, Using semantic predications to uncover drug–drug interactions in clinical data, J. Biomed. Inform. 49 (2014) 134–147.