CausalTriad: Toward Pseudo Causal Relation Discovery and Hypotheses Generation from Medical Text Data

Sendong Zhao
School of Computer Science and Technology
Harbin Institute of Technology
sdzhao@ir.hit.edu.cn

Meng Jiang
Department of Computer Science and Engineering
University of Notre Dame
mjiang2@nd.edu

Ming Liu
School of Computer Science and Technology
Harbin Institute of Technology
mliu@ir.hit.edu.cn

Bing Qin
School of Computer Science and Technology
Harbin Institute of Technology
bqin@ir.hit.edu.cn

Ting Liu
School of Computer Science and Technology
Harbin Institute of Technology
tliu@ir.hit.edu.cn

ABSTRACT

Deriving pseudo causal relations from medical text data lies at the heart of medical literature mining. Existing studies have utilized extraction models to find pseudo causal relation from single sentences, while the knowledge created by causation transitivity – often spanning multiple sentences – has not been considered. Furthermore, we observe that many pseudo causal relations follow the rule of causation transitivity, which makes it possible to discover unseen casual relations and generate new causal relation hypotheses. In this paper, we address these two issues by proposing a factor graph model to incorporate three clues to discover causation expressions in the text data. We propose four types of triad structures to represent the rules of causation transitivity among causal relations. Our proposed model, called CausalTriad, uses textual and structural knowledge to infer pseudo causal relations from the triad structures. Experimental results on two datasets demonstrate that (a) CausalTriad is effective for pseudo causal relation discovery within and across sentences; (b) CausalTriad is highly capable at recognizing implicit pseudo causal relations; (c) CausalTriad can infer missing/new pseudo causal relations from text data.

KEYWORDS

Medical Literature Mining; Causal Discovery; Causation Transitivity Rules; Hypothesis Generation

ACM Reference Format:
Sendong Zhao, Meng Jiang, Ming Liu, Bing Qin, and Ting Liu. 2018. CausalTriad: Toward Pseudo Causal Relation Discovery and Hypotheses Generation from Medical Text Data. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3233547.3233555

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233555

1 INTRODUCTION

Causal knowledge from medical data lies at the heart of medical research. Big text data in the medical domain is being published every day, such as medical literature, clinical text, and posts in online healthcare forums. Mining psuedo medical causal knowledge from these text data can be useful for new medical knowledge discovery, adverse drug side effects detection and even for clinical decision support after confirmation. For example, it is extremely difficult and costly to determine unknown drug side effects and adverse drug reactions via clinical trials. However, one can obtain drug side effects and adverse drug reactions via causal relation discovery from online medical forums [11, 15]. Psuedo medical causal relation in this paper represent semantic-level causal relations in medical text. They are potentially verified true causal knowledge, or those that either have not been identified previously or do not have evidence to support them. For convenience, we use "causal relation" or "causal knowledge" to represent the psuedo medical causal relation.

Numerous efforts have been made to extract causal relations from text data [8, 9, 24]. However, these studies have several limitations. First, these traditional methods of causal extraction can hardly extract causal relations across sentences. A notable exception on cross-sentence relation extraction is Peng et al. [20], which applied graph LSTMs to general cross-sentence relation extraction, but heavily relied on syntactic and dependency structures between sentences. Second, medical causal relations are frequently expressed implicitly, which requires causal inference. Even when causation is explicit, there is a wide variation in how it is expressed. But rare studies encode causal inference to enhance implicit and explicit causal relation extraction. Third, these studies focus on extracting existing causal relations in text rather than inferring unseen causal relation candidates.

These three mentioned drawbacks are also key challenges of discovering causal relations from text data.

Due to the above limitations of previous studies and challenges, a very natural question would be *can we extract medical causal relation within/across sentences and infer unseen causations from the same text?* To address these issues, in this paper, we propose three clues for medical causal relation discovery from text data as illustrated below. These three clues consider both textual and structural information.

First, inspired by Hume's comments on causal relation [5] [1], it is natural to assume that two medical entities co-occurring frequently are more likely to be a causal relation candidate. Some recent studies have utilized this assumption for causal extraction [6]. As shown in Figure 1, medical entities that co-occur with a high frequency are likely to be a causal relation.

Second, transitivity rules are powerful for recognizing causal relations, especially for those with few textual supports. Paul and Hall [12] suggest that "preserving transitivity is a basic desideratum for an adequate analysis of causation". Take the cause-effect candidate (*food allergies* $\longrightarrow$ *post nasal drip*) in Figure 1 as an example. This pair co-occurs in only five text examples. However, given the cause-effect pairs "(*food allergies* $\overset{\text{CAUSE}}{\longrightarrow}$ *acid reflux*)" and "(*acid reflux* $\overset{\text{CAUSE}}{\longrightarrow}$ *post nasal drip*)", it is easy to infer "(*food allergies* $\overset{\text{CAUSE}}{\longrightarrow}$ *post nasal drip*)". Therefore, it is clear that causation transitivity rules are extremely useful in determining those causal relations with few textual supports, making it interesting to study how we can leverage transitivity rules to discover causal relations. By considering transitivity rules, it becomes possible to leverage observable causal relations for discovering unseen casual relations and generating new causal relation hypotheses.

Third, the causal relation tends to have special contexts including domain-related words, causal triggers and connectives, such as "help to improve," "increase the risk of," "is caused by," etc. However, we aim at causal relation discovery by taking supports from all potential contexts of every appearance of the causal relation rather than causal relation extraction from a specific context only. To use all potential contexts for causal relation discovery, we consider two methods to model contextual information of causal relations: 1) synthetic contexts; and 2) contexts based on word embedding. Both are to combine context from the whole dataset but in different ways.

Combining evidence from both textual supports and structural inferences, the above three clues are better equipped to discover causal relations. Furthermore, they are complementary in several ways. First, these three clues include both textual and structural knowledge. Cause-effect co-occurrence and contextual information are the textual knowledge while the causation transitivity rule is the structural knowledge. Second, while causal relation recognition based on high frequency co-occurrence obviously gives preference to frequently co-occurring causal pairs, causation

[1] If two objects are causal, then they should be always co-occurring.

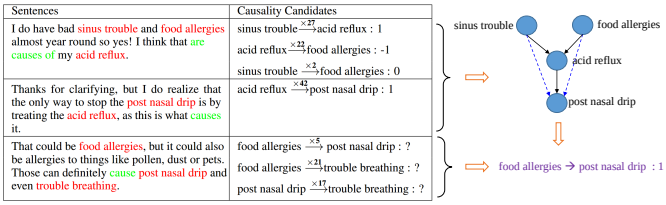

Figure 1: Examples of utilizing causation transitivity rule for inferring unseen pseudo medical causal relation (*food allergies* $\overset{\text{Cause}}{\longrightarrow}$ *post nasal drip*). The solid directed lines denote observed causal relations and dotted directed lines denote the unseen causal relations

transitivity rules are designed to identify causal relations with few textual supports except for those that follow the transitivity rule. Third, incorporating contextual causal information from the text can eliminate those frequently co-occurring medical entities which are not causal.

To incorporate these clues into a joint framework, we propose CAUSALTRIAD which unifies them into a factor graph model. Specifically, for modeling causation transitivity rules we segment the large network composed of cause-effect candidates into triad structures. Each triad structure consists of two causal relations which share a common object. On these triad structures we define different types of rules to encode causation transitivity in the cause-effect candidates network. CAUSALTRIAD infers links in triads with both textual and structural knowledge for causal relation discovery. Experimental results on two different datasets show that (1) CAUSALTRIAD is effective for causal extraction within and across sentences; (2) CAUSALTRIAD is highly capable at extracting implicit causation; (3) CAUSALTRIAD can infer unseen casual relations and generate new causal relation hypotheses from text data.

2 RELATED WORK

Discovering causal relations from data lies at the heart of medical research. Conducting controlled experiments, the standard and most popular method for causal discovery, is too costly in medicine. Causal knowledge can also be derived from purely observational data [10]. Many studies have been dedicated to extracting causal relations from structured data [1, 17]. Unfortunately, these studies utilize small data or synthetic data, which greatly limits the scalability of their findings for use in real and massive data. Alternatively, we can also discover causal relations in text data like medical literature, patients' records, posts in health forums, etc.

In computational linguistics, many studies deal with the extraction of causal relations from text [8, 19, 23, 24, 30]. While causal relations naturally have an attached network structure, making it possible to leverage structural inference for causal relation discovery, none of these studies leverage this important resource. Consequently, they can only extract existing causal relations within single sentences, but cannot

discover causal relations across sentences and go further to infer unseen causal relations in the data.

In general, cross-sentence relation extraction has received little attention, even in the supervised-learning setting. Among the limited amount of prior work, The study of Peng et al. [20] is a notable one that utilizes graph LSTMs to extract relations across sentences, by constructing document graph through syntactic features and a dependency link between the root nodes of parse trees. Its method of relation extraction involves intricate syntactic analysis of sentences, leading to unsatisfactory performance due to its failure to account for the diversity of expressions, especially for medical forum posts.

3 PROBLEM DEFINITION

We define the medical causal candidate network derived from causal pair candidates of medical knowledge.

Definition 3.1 (Medical Cause-Effect Candidates Network). The network, denoted as $G = (V, E)$, is a directed network that captures the causal relation between medical entities e_i and e_j. V is the set of medical entities. $E \subset V \times V$ is the set of causal candidates extracted according to the co-occurrence of medical entities.

To segment the network G of large scale for causal relation discovery, we generate triad structures for those causal pair candidates which share common nodes. These triad structures make it convenient to exploit transitivity rules to determine causal relations.

Definition 3.2 (Triad Structure). Each triad structure is denoted by $T = (\{e_i, e_j, e_k\}, \{e_{ij}, e_{jk}, e_{ki}\})$, where $e_i \in V$ is the ith medical entity and e_{ij} is the relation between e_i and e_j. The corresponding $c_{ij} = 1$ on e_{ij} denotes a **cause-to-effect** relation, $c_{ij} = -1$ for **effect-to-cause** relation between e_i and e_j, $c_{ij} = 0$ for **null** relation. Among these three edges $\{e_{ij}, e_{jk}, e_{ki}\}$, the triad structure requires there are at least two edges which are observed in the network G. At each round of our system, we generate different set of triad structures.

This network G should be dense because there would be an edge if two entities e_i and e_j co-exist in the same text.

Given these definitions, we can formally formulate the problem studied in this work.

Problem 1 (Causal Relation Discovery from Triad Structures). *Given the medical cause-effect candidate network G with small fraction of edges labeled by human, we segment this network into many corresponding triad structures T. On these different triad structures we define types of triads to represent the rules of causation transitivity. Our task is to determine the causal relation and its direction of the left e_{ij} without human annotation in triad structures via incorporating both triads and textual supports.*

This is a fundamental problem in medical literature mining. The obtained medical causal knowledge can be exploited to power downstream applications, such as new medical knowledge discovery, adverse drug side effects detection and even for clinical decision support after confirmation with medical expertise. Three clues are utilized to determine causal relation in each triad structure.

4 OUR METHOD

In this section, we present a factor graph model on the triad structures to integrate both textual and structural information as factors in a unified framework for discovering causal relations from text data.

4.1 Medical Cause-Effect Candidates Matching

Mining causal relation from medical text data first requires the recognition of instances of the medical entities involved in causation. There already have been many studies on medical entity extraction from Electronic Health Records (EHR) [7] and biomedical literature [2, 18, 28]. Studying medical entity extraction is not within the scope of this paper. Therefore, we just collect dictionaries directly from existing databases to match instances of medical entities in medical literature and forum text. For English medical entities, we use the medical dictionary from the Dryad data package [7] which is extracted from millions of clinical narratives. For TCM entities, we use two TCM databases, i.e., TCMonline and TCMID [27].

In each window of sentences, i.e., consecutive n sentences, we exploit these medical dictionaries to match medical entities and pair each of them into several causal pair candidates. Every two pairs with a common object generate a triad structure. For example, there are two candidates (A,B) and (B,C), we take (A,B,C) as a triad structure. In the following steps, we introduce the approach to model clues for determining the causal relation in each triad structure.

4.2 Causal Association

In this section, we model the clue that frequently co-occurring medical entities are more likely to be a causation. To measure and predict the existence of causation between medical entities, we define the association as follows.

$$CA(e_{ij}) = I(e_i, e_j) \times D(e_i, e_j) \times Max(u_i, u_j) \qquad (1)$$

In which we measure the co-occurrence of two medical entities, the distance of two medical entities in text and the frequency of that co-occurrence.

In Suppes' probabilistic theory of causation [25], he highlighted that entity e_i is a possible cause of entity e_j, if e_j happens more frequently with e_i than by itself, i.e. $P(e_j|e_i) > P(e_j)$. This can be easily rewritten as $\frac{P(e_j,e_i)}{P(e_j)P(e_i)} > 1$, similar to the definition of pointwise mutual information:

$$I(e_i, e_j) = \log \frac{P(e_i, e_j)}{P(e_i)P(e_j)}$$

which is positive when $\frac{P(e_j,e_i)}{P(e_j)P(e_i)} > 1$.

We award pairs that co-exist closer, while penalizing those are further apart in texts, by incorporating the distance

measure of [13]:

$$D(e_i, e_j) = -log\frac{|sent(e_i) - sent(e_j)| + 1}{2 \times WS}$$

where $sent(e)$ gives the sentence number (index) in which e occurs and WS indicates the window-size (of sentences) used. If e_i and e_j are drawn from the same sentence, the numerator of the above fraction will return 1. If e_i and e_j appear in multiple sentences, we choose a co-occurrence with the smallest $|sent(e_i) - sent(e_j)|$ and assign the value to the above formula. In our work, we set WS to 3 and thus, if e_i occurs in sentence n, the furthest sentence that e_j will be drawn from, is sentence $n + 2$.

The term $Max(u_i, u_j)$ aims to model the frequency of co-occurence of two medical entities e_i and e_j. u_i and u_j are defined as follows.

$$u_i = \frac{P(e_i, e_j)}{\max_k(P(e_i, e_k)) - P(e_i, e_j) + \varepsilon}$$

$$u_j = \frac{P(e_i, e_j)}{\max_k(P(e_k, e_j)) - P(e_i, e_j) + \varepsilon}$$

in which, we set $\varepsilon = 0.01$ to avoid 0 in the denominator. u_i will be maximized if there is no e_k which makes (e_i, e_k) has higher frequency than (e_i, e_j). u_j has the same treatment.

4.3 Modeling Contextual Information

We observe that causal relations in the text tend to share special contexts including domain-related words, causal triggers, and connectives, such as "help to improve," "increase the risk of," "is caused by," etc. This observation is the main inspiration of this section. There are broadly two different approaches to encode context. In this section, we introduce these two approaches to model contextual information for causal relation candidates i.e., synthetic symbolic context and implicit context based on word embedding.

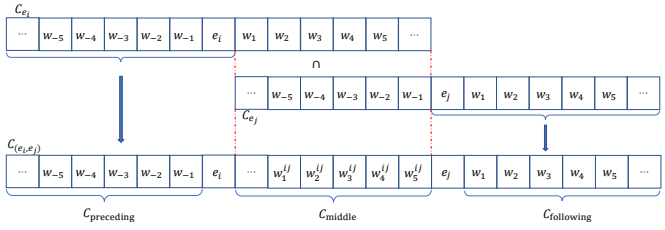

Figure 2: Synthesizing synthetic context of a causal pair candidate (e_i, e_j).

4.3.1 Synthetic Context. We use explicit symbolic approach to generate synthetic context of causal pair candidates as shown in Figure 2. Note that C_{e_i} is the synthetic context of entity e_i, $\{w_{-k}\}$ is the preceding contextual bag-of-words (consist of unigrams and bigrams) in which every unigram/bigram co-exists at least once in front of e_i in text. k is the index of unigram/bigram in vocabulary, "-" represents the unigram/bigram is a preceding context of e_i. $\{w_k\}$ is the following contextual bag-of-words (consist of unigrams and bigrams) in which every unigram/bigram co-exists at least once behind e_i in text. C_{e_j} follows the same rule to generate the synthetic context of e_j. Most importantly, we synthesize the synthetic context of causal pair candidate i.e., $C_{(e_i,e_j)}$, by carrying out concatenating and merging on C_{e_i} and C_{e_j} as shown in Figure 2.

In particular, 1) the generated synthetic context $C_{(e_i,e_j)}$ takes bag-of-words $\{w_{-k}\}$ with attached $TFIDF(w_{-k}, e_i)$ in C_{e_i} as its preceding context, i.e., $\mathbf{C}_{\text{preceding}}$. 2) It takes bag-of-words $\{w_k\}$ with attached $TFIDF(w_k, e_j)$ in C_{e_j} as its following context, i.e., $\mathbf{C}_{\text{following}}$. 3) By matching the $\{w_k\}$ in C_{e_i} and the $\{w_{-k}\}$ in C_{e_j}, $C_{(e_i,e_j)}$ gets common bag-of-words $\{w_k^{ij}\}$ of those two as its middle context between entities, i.e., $\mathbf{C}_{\text{middle}}$. $\{w_k^{ij}\}$ takes $TFIDF(w_k^{ij}, (e_i, e_j))$ in $\mathbf{C}_{\text{middle}}$ as its corresponding value. We define $TFIDF(w, e)$ as follows:

$$TFIDF(w, e) = TF(w, e) \times IDF(w, e),$$

$$IDF(w) = \log\frac{N + 1}{N(w, e) + 1} + 1,$$

where $TF(w, e)$ is the frequency of the co-occurrence of w and e. N is the number of medical entities in all causal relation candidates. $N(w, e)$ is the number of medical entities which co-occurs with w.

To get those contextual words, we remove stop words excluding prepositions from those sentences with medical entity e_i and then count the most frequent preceding contextual bag-of-words $\{w_{-k}\}$ and following contextual bag-of-words $\{w_k\}$ after stemming. Here, we define contextual bag-of-words to be composed of unigrams and bigrams.

The context vector $\mathbf{e}_{ij}$ of the causal relation candidate (e_i, e_j) is the vector generated by conducting $TFIDF$ on concatenating $\mathbf{C}_{\text{preceding}}$, $\mathbf{C}_{\text{middel}}$ and $\mathbf{C}_{\text{following}}$, i.e.,

$$\begin{aligned}\mathbf{e}_{ij} =& TFIDF(\mathbf{C}_{\text{preceding}}, e_i) \\ &\circ TFIDF(\mathbf{C}_{\text{middle}}, (e_i, e_j)) \\ &\circ TFIDF(\mathbf{C}_{\text{following}}, e_j)\end{aligned} \tag{2}$$

where $\circ$ is the operation for concatenating vectors.

4.3.2 Implicit Context Based on Word Embedding. We borrow the idea from word embedding [16, 21] to generate implicit context of causal relation candidates. We try ways to embed the context of causal relation candidates as (1) the difference of two vectors; (2) the sum of two vectors; (3) the concatenation of two vectors. Therefore, we take

$$\begin{aligned}\mathbf{e}_{ij} &= v(e_j) - v(e_i), \\ \mathbf{e}_{ij} &= v(e_j) + v(e_i), \\ \mathbf{e}_{ij} &= v(e_i) \circ v(e_j)\end{aligned} \tag{3}$$

as the representation of the causal relation e_{ij}, where $v(e_j)$ and $v(e_i)$ are achieved by carrying out word2vec [16] on our medical text data.

By the above two ways, it is flexible to leverage contextual information for causal relation discovery. Furthermore, it becomes easy to leverage observed/obtained causal relation for the non-existing causal relations discovery because they tend to share some contextual information, such as causal trigger, connectives and domain-related words.

4.4 Triads of Causation Transitivity Rules

Modeling transitivity rules are important for causal relation discovery. Given two causal pairs ($e_k \overset{\text{CAUSE}}{\longrightarrow} e_i$) and ($e_i \overset{\text{CAUSE}}{\longrightarrow} e_j$), it is easy to get the causal relation ($e_k \overset{\text{CAUSE}}{\longrightarrow} e_j$) via simple inference. However, modeling this transitivity for causal extraction from the text is far from reach in existing studies. In this section, we use two functions to model causation transitivity rules. The function $A(\cdot)$ is used for modeling causation transitivity and directions and the function $T(\cdot)$ is the supplement to $A(\cdot)$.

We define four types of triads to represent causation transitivity rules in the function $A(\cdot)$, which are shown in Figure 3. In triad (a), e_i is a cause in a causal relation e_{ij} with the label c_{ij} as well as an effect in another causal relation e_{ki} with the label c_{ki}, thus the direction between e_k and e_j is **cause-to-effect**. In triad (b), e_i is an effect in a causal relation e_{ij} with the label c_{ij} as well as a cause in another causal relation e_{ki} with the label c_{ki}, thus the direction between e_k and e_j is **effect-to-cause**. In triads (c) and (d), the relation between e_j and e_k is uncertain. Here, we model the uncertain cases (c) and (d) because there might be stronger relation between e_j and e_k than **null** relation in these two cases. Hence, it is meaningful to preserve the possibility of the causal relation between e_j and e_k rather than making arbitrary decisions. We call these four triads with specific rules attached **angle rules** of causation transitivity. We model them in the following function.

$$A(a_{kij}) = \begin{cases} 2, & |c_{ki}| + |c_{ij}| = 2,\ \textsc{Cause}(e_i, c_{ij}) \\ & \textsc{Effect}(e_i, c_{ki}) \\ 1, & |c_{ki}| + |c_{ij}| = 2,\ \textsc{Cause}(e_i, c_{ki}) \\ & \textsc{Effect}(e_i, c_{ij}) \\ -1, & |c_{ki}| + |c_{ij}| = 2,\ \textsc{Effect}(e_i, c_{ij}) \\ & \textsc{Effect}(e_i, c_{ki}) \\ -2, & |c_{ki}| + |c_{ij}| = 2,\ \textsc{Cause}(e_i, c_{ij}) \\ & \textsc{Cause}(e_i, c_{ki}) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where e_i is a vertex representing a medical entity, c_{ij} and c_{ki} are labels of two edges linked to e_i shown in Figure 3. (c_{ki}, e_i, c_{ij}) composes the causal angle a_{kij}. $\textsc{Cause}(e_i, c_{ki})$ represents that e_i is the cause of a relation e_{ki} with corresponding label c_{ki}. $\textsc{Effect}(e_i, c_{ki})$ represents e_i is the effect of a relation e_{ki} with corresponding label c_{ki}. The first four cases in this equation refer to four types of triads with angle rules attached in Figure 3. Note that, at the beginning, the value of c_{ki} and c_{ij} comes from the training data and our system is a bootstrapping implementation.

However, the angle rules should only exist in those three medical entities which have at least two edges among them. We call this rule **triadic rule**. Modeling this rule is the supplement to angle rules and also benefits causal relation recognition. The function $T(t_{kij})$ is to model the triadic rule.

$$T(t_{kij}) = \begin{cases} 1, & |c_{ki}| + |c_{ij}| + |c_{jk}| \geq 2 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

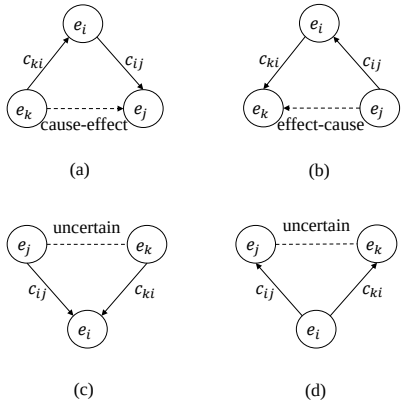

Figure 3: Angle rules: 4 types of triads to represent the rules of causation transitivity for causal relation discovery. The transitivity rules in first two triads, (a) and (b), is obvious. The differences between (a) and (b) is the direction of causation, i.e., cause-to-effect and effect-to-cause. The causal relation in last two triads is uncertain.

where $t_{kij} = \{c_{ki}, c_{ij}, c_{jk}\}$ is the label set on each causal triad structure which is constructed by connecting e_i, e_j and e_k.

4.5 Factor Graph Model for Triads

The medical cause-effect candidate network is composed of triad structures. Each triad structure T consists of three vertexes $\{e_i, e_j, e_k\}$ and three edges $\{e_{ki}, e_{ij}, e_{jk}\}$ with corresponding labels $\{c_{ki}, c_{ij}, c_{jk}\}$. On each triad structure of the medical cause-effect candidates network, we can use angle rules and the triadic rule to model causation transitivity. Therefore, the causal relation discovery problem in cause-effect candidate network can be cast as a problem of inferring the values of each $C = \cup_{(ij)} c_{ij}$ of corresponding triad structure. Specifically, we use the joint posteriori probability $Pr(C|\mathbf{X})$ to model the distribution of C. $\mathbf{X}$ corresponds to all observed and hidden variables, respectively.

A factor graph provides an explicit way to factorize the global joint probability by introducing local variable and factor nodes. To encode three clues for causal relation discovery, we present CAUSALTRIAD to take causal association, contextual information and causation transitivity rules as different local factors, which is shown in Figure 4. Therefore, the joint probability can be factorized as association factors $F_{\text{association}}(e_{ij})$ which incorporates Eqn. (**??**), contextual information factors $F_{\text{context}}(e_{ij})$ which incorporates Eqn. (2) or Eqn. (3), causation triad factors including angle rule factors $F_{\text{angle}}(a_{kij})$ which incorporates Eqn. (4) and triadic rule factors $F_{\text{triad}}(t_{ijk})$ which incorporates Eqn. (5).

The association factors $F_{\text{association}}(e_{ij})$ are used to capture the dependencies between c_{ij} and the causal association $CA(e_{ij})$. The contextual information factors $F_{\text{context}}(e_{ij})$ are used to capture the dependencies between c_{ij} and the contextual information of the causal candidate e_{ij}. We

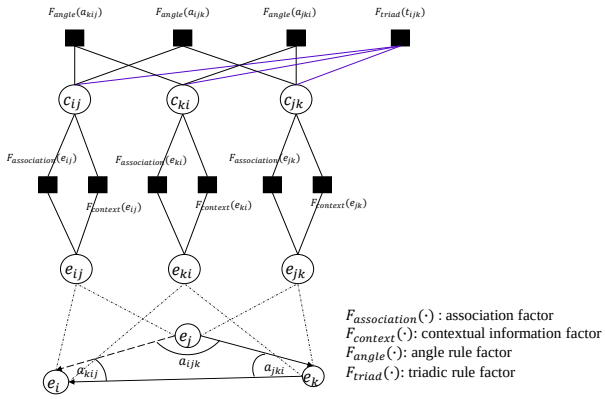

Figure 4: *Factor graph for each triad structure.* It models three clues as factors for causal relation discovery. The dotted directed line denotes the unseen causal relation to be inferred.

use angle rule factors $F_{\text{angle}}(a_{kij})$ to incorporate the rules illustrated in Figure 3. The triadic rule factors $F_{\text{triad}}(t_{ijk})$ are utilized to capture the compatibility among the labels of edges in each triad structure T.

$$Pr(C|\mathbf{X}) = \prod_{e_{ij}} Pr(c_{ij}|F_{\text{association}}(e_{ij}))Pr(c_{ij}|F_{\text{context}}(e_{ij})) \prod_{e_{ij}} Pr(c_{ij}|F_{\text{angle}}(a_{kij}))Pr(c_{ij}|F_{\text{triad}}(t_{ijk}))$$

where C is the set of all relation labels (i.e., c_{ki}, c_{ij} and c_{kj}) in triad structures T. $\mathbf{X}$ corresponds to all observed and hidden variables, respectively.

All these four factors can be instantiated by exponential-linear functions:

$$Pr(c_{ij}|F_{\text{association}}(e_{ij})) = \frac{1}{Z_{\text{association}}} \exp\{\alpha CA(e_{ij})\},$$

where $CA(e_{ij})$ refers to Eqn. (1) and

$$Z_{\text{association}} = \sum_{c' \in \{-1,0,1\}} Pr(c'|F_{\text{association}}(e_{ij}))$$

is local normalization term. α is the weight of $CA(e_{ij})$.

$$Pr(c_{ij}|F_{\text{context}}(e_{ij})) = \frac{1}{Z_{\text{context}}} \exp\{\beta \mathbf{w} \cdot \mathbf{e}_{ij}\},$$

where $\mathbf{e}_{ij}$ refers to Eqn. (2) or Eqn. (3) and

$$Z_{\text{context}} = \sum_{c' \in \{-1,0,1\}} Pr(c'|F_{\text{context}}(e_{ij})$$

is the local normalization term. β is the weight of $\mathbf{w} \cdot \mathbf{e}_{ij}$ and $\mathbf{w}$ is the weight vector of the context feature $\mathbf{e}_{ij}$.

$$Pr(c_{ij}|F_{\text{angle}}(a_{ikj})) = \frac{1}{Z_{\text{angle}}} \exp\{\delta A(a_{ikj})\},$$

where $A(a_{ikj})$ corresponds to 5 cases in Eqn. (4) which is to model 4 types of angle rules in Figure 3. δ is the weight of $A(a_{ikj})$.

$$Z_{\text{angle}} = \sum_{c' \in \{-1,0,1\}} Pr(c'|F_{\text{angle}}(a_{ikj}))$$

is the local normalization term.

$$Pr(c_{ij}|F_{\text{triad}}(t_{ijk})) = \frac{1}{Z_{\text{triad}}} \exp\{\gamma T(t_{ijk})\},$$

where $T(t_{ijk})$ corresponds to 2 cases in Eqn. (5), γ is the weight of $T(t_{ijk})$ and

$$Z_{\text{triad}} = \sum_{c' \in \{-1,0,1\}} Pr(c'|F_{\text{triad}}(t_{ijk}))$$

is the local normalization term.

We propose to use a semi-supervised learning algorithm to estimate parameters of the model, i.e., $\{\alpha, \beta, \delta, \gamma\}$ and $\mathbf{w}$, which enables us to determine whether each candidate is a true causal relation or not, i.e., to predict the labels of each edge of cause-effect candidate network based on estimated parameters. In order to learn parameters automatically, we make use of DeepDive [29] which is an openly available tool for using factor graphs to perform learning and inference.

5 EXPERIMENTAL EVALUATION

5.1 Data Preparation

We construct two datasets HealthBoards and TCM to evaluate CausalTriad. The basic statistics of these two datasets are shown in Table 1.

Table 1: Basic Statistics of two datasets TCM and HealthBoards

Dataset	# of abstract/post	# of pairs	# of labeled pairs
TCM	106,151	9,572	3,828
HealthBoards	37,432	12,093	4,836

HealthBoards: HealthBoards data consists of post messages on health and medical issues such as diseases, symptoms, medicines, and side-effects, etc. We download 37,432 posts from 7 selected topics in HealthBoards[2]. From these posts, we recognize pairs of medical entities with co-occurrence in a window with determined length as causal relation candidates.

TCM: Traditional Chinese Medicine (TCM) data consists of the abstracts of 106,151 papers published in major Chinese TCM journals, which almost cover all aspects of TCM research. All the abstracts are extracted from the China National Knowledge Infrastructure (CNKI). We perform word segmentation on these TCM abstracts via LTP [4] and leverage the mentioned TCM dictionaries to match medical entities.

5.2 Compared Methods

We compare CausalTriad with other models that can be used for causal extraction. The compared methods are as follows. i.e., the approach based on **Causal Patterns** [23], **Causal Association**, **Causal Context**, the **Minimally Supervised Method** [6], **CNN based Method** [19] and **Graph LSTMs** [20]. The causal patterns are regular expressions containing selected connectors. We consider the

[2]https://www.healthboards.com/

causal connectors with highest frequencies used in [26]. Four connectors with highest frequencies are selected for causal relation extraction, i.e., "cause", "is (the) cause(s) of", "lead to" and "increase the risk of". As a factor in CausalTriad, the details of causal association and causal context is described in section 4.2 and section 4.3 separately. We feed causal association and causal context into the factor graph model and simplify our CausalTriad with the single factors only.

5.3 Experimental Setting and Evaluation Metrics

For baselines, we tune hyper-parameters for best results. For our methods, we tried different values on WS and ε and the best configuration is $WS = 3$ and $\varepsilon = 0.01$ in the component of causal association. To construct document graph for implementing Graph LSTMs of Peng et al. [20], we utilize LTP [4] to parse TCM abstracts and follow Peng et al. [20] to use SPLAT [22], Stanford CoreNLP [14] on HealthBoards posts. The intra-sentential edges of document graph are based on the dependencies derived from syntactic parses. The inter-sentential edges of document graph are created by adding an edge between the dependency roots of adjacent sentences.

We randomly labeled a fraction (i.e., 40%) of candidates of the causal pair in each dataset. For each of these candidates, we ask one graduate computer science student and two students who specialize in medicine and healthcare to annotate each candidate in the two datasets. For each candidate, the three annotators read the corresponding TCM abstracts or HealthBoards posts to distinguish whether it appears as a semantic causal relation at least once. We measure the inter-annotator agreement using the kappa coefficient [3]. The kappa value on TCM abstracts is 0.87 and on HealthBoards posts is 0.84, which indicates a good strength of agreement.

We split human labeled causal relation candidates of each dataset into training/validation/test set respectively, with the ratio of 3:1:1. To examine the power of CausalTriad for inferring unseen casual relations and generating new causal relation hypotheses, 60% unlabeled candidates is also used for both training and testing. We sample 200 cases from those obtained causal relations in each dataset through CausalTriad and analyze their specific types of causal relation, such as Drug–against–Disease. In the test set, we separate causal pairs across sentences from those within single sentences to evaluate how well CausalTriad can discover the causal relation across sentences. In addition, we distinguish explicit causal relations from implicit causal relations to check the performance of CausalTriad for implicit casual relation extraction.

We compute true precision, true recall and true F-measure for comparisons with baselines. We exploit true F-measure to evaluate causal relation extraction within single sentences, causal relation extraction across sentences and implicit causal relation extraction.

Table 2: Examples of unseen medical causal relations obtained from HealthBoards posts and TCM literature abstracts.

Cause	Effect
acid reflux	post nasal drip
low-carb diet	acid reflux
Ranitidine	headache
PPI (proton pump inhibitors)	stomach pain
Prilosec	diarrhea
小柴胡汤 (xiaocaihu docection)	疟疾 (malaria)
六味地黄丸 (six ingredient rehmannia pill)	糖尿病 (diabetes)
糖尿病 (diabetes)	多尿 (increased urine output)
何首乌 (polygoni multiflori)	血虚证 (blood deficiency sign)
枸杞子 (Goji)	肝肾亏虚证 (liver and kidney deficiency)

5.4 Bootstrapping Implementation

Our method for determining causal relation from the triad structure itself is naturally suitable for a bootstrapping implementation. We use obtained causal relations in each triad structure to infer the unseen/new relations and generate new triad structures. On the one hand, once a causal relation is determined, we can further use obtained causal relations to identify the unseen/new relations. On the other hand, the new obtained causal relations can potentially generate new triad structures. In each round of determining causal relations in all triad structures, our classifier recognizes new causal relation and removes old causal relation candidates which the classifier do not confirm. Therefore, we have different sets of triad structures in each round due to adding new ones and removing old ones. We use early stopping to identify the point when adding additional new causal relation is not worthwhile. In every round, if a causal relation candidate is involved in multiple triad structures, we choose the highest probability of $c_{ij} = 1$, 0 or -1 as its corresponding value. In the end, we average the probability of each causal candidate in multiple rounds to determine the causal relation.

5.5 Results and Analysis

We design experiments to evaluate the capability of CausalTriad for medical causal relation discovery. Each of them presents different advanced aspects of our method.

We first examine the capability of our CausalTriad for discovering new causal relations and inferring unseen casual relations and generating new causal relation hypotheses. We sample some obtained causal relations through CausalTriad which are not labeled as causation by annotators or do not co-occur in text. In other words, these causal relations can be considered as unseen/new causal relations because they are not observed in the training data. We choose some obtained examples and present them in Table 2. The first five examples are from posts of the HealthBoards, and

Table 3: Comparison of causal relation discovery with baselines on two datasets (%).

Method	TCM			HealthBoards		
	True P	True R	True F-measure	True P	True R	True F-measure
Causal Patterns [23]	-	-	-	89.4	24.5	38.5
Causal Context	49.7	65.4	56.5	44.5	63.2	52.2
Causal Association	52.6	85.3	65.1	45.1	87.1	59.4
MiniSupervise [6]	54.3	88.6	67.3	46.4	89.3	61.1
CNN based [19]	76.4	85.3	80.6	73.7	76.6	74.6
Graph LSTMs [20]	78.8	90.8	**84.4**	72.4	74.1	73.2
-Association	66.3	67.4	66.8	57.7	73.2	64.5
-Context	62.3	87.6	72.8	59.0	81.1	68.3
-Transitivity	68.3	85.1	75.7	64.6	79.2	71.2
CausalTriad (Embedding,+)	61.5	86.4	71.8	59.2	80.7	68.3
CausalTriad (Embedding,○)	67.4	89.2	76.8	59.5	82.6	69.2
CausalTriad (Embedding,−)	72.3	90.7	80.5	67.4	83.9	74.8
CausalTriad (Synthetic)	74.8	91.8	82.4	69.2	85.4	**76.4**

the following five examples are from abstracts of TCM literature. In addition, we sample 200 obtained cases from the obtained causal relations for each dataset and analyze their types. These causal pairs from TCM have roughly **6 different types**, 1) "Disease–cause–Symptom"; 2) "Formula–against–Disease"; 3) "Herb–against–Disease"; 4) "Formula–relieve–Symptom"; 5) "Herb–relieve–Symptom" and 6) "the other". These causal pairs from HealthBoards have roughly **4 different types**, 1) "Disease–bring–Disease"; 2) "Drug–against–Disease"; 3) "Disease–cause–Symptom" and 4) "the other". The type distribution of these causal relations on two datasets is shown in Figure 5. From these different types of causal relations, it is easy to instantiate reasonable specific patterns of causation transitivity rules as follows.

- Formula–against–Disease–cause–Symptom⇒ Formula–relieve–Symptom
- Herb–against–Disease–cause–Symptom⇒ Herb–relieve–Symptom
- Drug–against–Disease–bring–Disease⇒ Drug–against–Disease
- Disease–bring–Disease–cause–Symptom⇒Disease–cause–Symptom

Note that the first two patterns are from TCM and the last two patterns are from HealthBoards.

Furthermore, we find some causal relations of medical entities which do not co-exist in text but medical experts have confirmed that they truly are causal relations. It is obvious that these causal relations are derived through inference by the transitivity rule. This demonstrates the capability of causation transitivity rules for inferring unseen causal relations and generating new causal relation hypotheses. This capability of our model potentially gives a high-efficiency and low-cost way to update medical causal knowledge. Consider you can extract drug side-effects and adverse drug reactions on medical forums from patients' reports, generate new potential effects of particular treatments and find unknown causes of particular disease through the inference of our CausalTriad. At least CausalTriad can provide high-quality medical causal relation candidates through causal discovery from text, therefore, significantly lower the cost of experimental verification for medical causal relations.

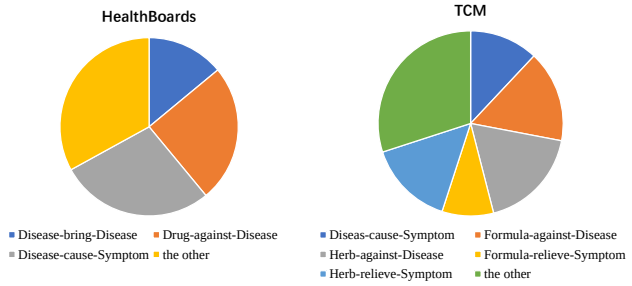

Figure 5: Distributions of types of medical causal relations obtained from two datasets.

We second explore how CausalTriad outperforms baseline models for causal relation discovery and the contribution of each component in our model. Table 3 shows the comparisons of causal relation discovery with baselines on two datasets. Since CausalTriad is a model based on bootstrapping, to get the best performance, we iteratively update the label of each causal relation candidate until the labels no longer change. The iterative update of label of each candidate simultaneously uses its self-attributes and inferred neighboring labels. The results indicate that our method outperforms baselines on HealthBoards posts and is comparable with the recent advance of cross-sentence relation extraction, i.e., Graph LSTMs, on TCM abstracts. The fail of Graph LSTMs on HealthBoards posts might due to the fact that it heavily relies on the syntactic dependency parser, while the posts in HealthBoards is narrative and informally expressed. Compared with the most recent advance of deep learning models, i.e., CNN based and Graph LSTMs, on relation extraction, our CausalTriad is more flexible with less dependence on non-trivial NLP technique.

We third examine how well our CausalTriad can discover the causal relation across sentences. In the labeled causal pairs, we first manually separate causal pairs across sentences from those within single sentences. Then, we evaluate the performance of causal discovery within and across sentences respectively. Table 4 shows the results of our CausalTriad on discovering causal relations within single sentences and across sentences. In terms of cross-sentences causal relation extraction, CausalTriad is comparable with the state-of-the-art cross-sentence relation extraction method [20] but

do not rely on non-trivial NLP techniques like syntactic dependency parsing. Unlike existing causal extraction models which extract causal relation from single sentences, our CausalTriad can discover causal relation across sentences effectively. Most importantly, CausalTriad gives a new view and a flexible way to discover causal relations across sentences.

Table 4: Comparison of causal relation discovery within single sentence and across sentences on two datasets (F-measure %).

Method	TCM		HealthBoards	
	Single	Cross	Single	Cross
MiniSupervise [6]	68.7	66.1	63.4	57.6
CNN based [19]	84.3	73.4	75.8	73.3
Graph LSTMs [20]	**84.2**	**84.7**	75.5	71.7
CausalTriad	83.1	82.3	**76.6**	**76.5**

We fourth check the ability of our CausalTriad for extracting implicit causal relations. Extracting implicit causal relation is difficult because 1) there are no explicit marks to indicate the causal relation and 2) implicit causal relation extraction usually requires inference. Figure 6 shows the performance of extracting implicit causal relations, which indicates that our method gets promising performance. This advantage might rely on incorporating the evidence of co-occurrence and structural inference in CausalTriad.

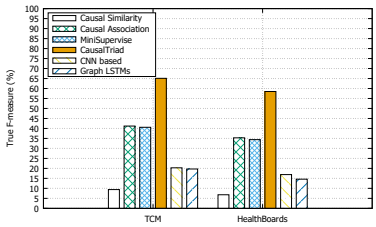

Figure 6: Comparison of semantic implicit causal relation extraction on two datasets

5.6 Influence Factors

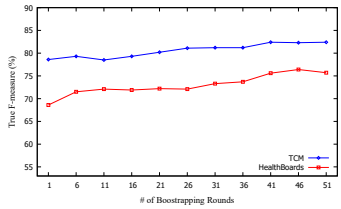

Figure 7: Effect of number of bootstrapping rounds in CausalTriad

Does the number of bootstrapping rounds matter? We want to investigate how the bootstrapping round affects the performance of CausalTriad for medical causal relation discovery. To make CausalTriad better use of the causation transitivity rules in triad structures, we run an iterative classification algorithm for causal relation discovery tasks. Iterative classification gives a simple and approximate way to implement the concept of collective inference. It exploits known causal pairs to assist the unseen causal relation identification in each triad structure and verify the obtained causal relation as well. Specifically, we design an iterative classifier that takes the results from the original factor graph model as initial values of all the labels. Then the iterative classifier iteratively updates the label of each causal relation candidate in the following rounds, until labels do not change anymore. Figure 7 presents how the bootstrapping rounds affects the performance of CausalTriad for medical causal relation discovery. Bootstrapping yields improvement over the original classification method with an absolute gain of 3.8 points in TCM data and 7.8 in HealthBoards data.

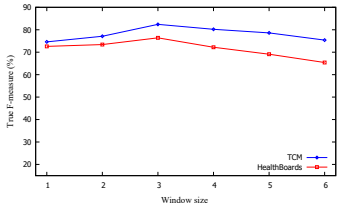

Figure 8: Effect of the window size WS in CausalTriad

Does the window size WS matter? We investigate how the window size WS affects the performance of CausalTriad for medical causal relation discovery from text data. We vary WS on two datasets from 1 to 6, increasing by 1. Results of the effect of window size WS on true F-measure are given in Figure 8. We can see that the best WS configuration is 3 for both "TCM" and "HealthBoards". Greater or less than the best configurations will cause a worse performance.

How dose the size of labeled training data affect performance? As shown in Figure 9, we check the performance of methods on the textual causal relation discovery by incrementally augmenting labeled training data. The results on both TCM and HealthBoards are really convincing that CausalTriad is much less dependent on the size of labeled training data compared to CNN based and Graph LSTMs, even though it gets better performance for more training data. Especially on the condition that small training data, CausalTriad significantly outperforms the other two baseline models.

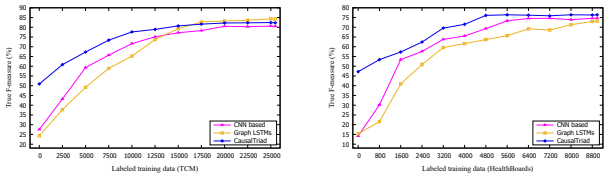

Figure 9: Comparison of methods by incrementally augmenting labeled training data.

5.7 Error Analysis

Finally, we present error analysis of CausalTriad on two datasets. As we check all error cases in obtained medical causal relations, we find that they originate from three reasons. **First**, some errors happen in medical causal relation recognition due to noises in medical entity dictionaries especially the medical dictionary from the Dryad data package [7]. **Second**, the self-contradiction happens in some causal relation candidates. For example, it is clear that "obesity" cause "cardiopathy" from some textual supports and you can see "cardiopathy" cause "obesity" alternatively from other supports, which makes it difficult to determine which one is true. **Third**, the setting window-size slightly affects the result on both datasets. If we set the window-size as 1, we would miss many true causal pairs spanning multiple sentences, leading to a low recall. If we set the window-size larger than 1, more noisy candidates would be introduced, leading to a low precision.

6 CONCLUSION

In this paper, we aimed at discovering causal relation from medical text data. However, existing studies utilized extraction models to extract causal relation from single sentences, while the knowledge of causation transitivity that often lies across sentences was not considered. To address this issue, we propose CausalTriad to incorporate clues for causal relation discovery from texts. Specifically, we segment the large network composed of cause-effect candidates into triad structures. On these triad structures, we define different types of rules to model causation transitivity to address the inference of unseen causal relations and extraction of medical causation including causation transitivity. Experimental results on two datasets demonstrate that CausalTriad is effective for textual medical causal relation discovery.

7 ACKNOWLEDGMENTS

We thank Lillian Foote, Charles Argon, Zhengpeng Xiang and the anonymous reviewers for their insightful comments and suggestions. The corresponding author of this paper is Ting Liu. This work was supported by the National Natural Science Foundation of China (NSFC) via grant 61632011, 61772156 and 61472107.

REFERENCES

[1] David Arbour, Dan Garant, and David Jensen. 2016. Inferring Network Effects from Observational Data. In *KDD'16*. 715–724.
[2] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics* 13, 1 (2012), 1.
[3] S. Siegel Castellan. 1988. *Non-Parametric Statistics for The Behavioral Sciences*. McGraw-Hill,. 3 pages.
[4] Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Coling'10*. 13–16.
[5] Angela M Coventry. 2006. *Hume's Theory of Causation*. A&C Black.
[6] Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally Supervised Event Causality Identification. In *EMNLP'11*. 294–303.
[7] Samuel G Finlayson, Paea LePendu, and Nigam H Shah. 2014. Building the graph of medicine from millions of clinical narratives. *Scientific data* 1 (2014).
[8] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong Hoon Oh. 2015. Generating event causality hypotheses through semantic relations. In *AAAI'15*. 2396–2403.
[9] Christopher Hidey and Kathleen McKeown. 2016. Identifying Causal Relations Using Parallel Wikipedia Articles. In *ACL'16*. 1424–1433.
[10] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. 2009. Nonlinear causal discovery with additive noise models. In *NIPS'09*. 689–696.
[11] Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015. Text and data mining techniques in adverse drug reaction detection. *Comput. Surveys* 47, 4 (2015), 56.
[12] Paul L. A. and Hall Ned. 2013. *Causation: A User's Guide*. Oxford University Press.
[13] Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database* 49, 2 (1998), 265–283.
[14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David Mcclosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL'14*.
[15] Alejandro Metke-Jimenez and Sarvnaz Karimi. 2015. Concept Extraction to Identify Adverse Drug Reactions in Medical Forums: A Comparison of Algorithms. *CoRR* (2015).
[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. [n. d.]. Distributed representations of words and phrases and their compositionality. In *NIPS'13*.
[17] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *JMLR* 17, 32 (2016), 1–102.
[18] Aurélie Névéol, Sonya E Shooshan, Susanne M Humphrey, James G Mork, and Alan R Aronson. 2009. A recent advance in the automatic indexing of the biomedical literature. *Journal of biomedical informatics* 42, 5 (2009), 814–823.
[19] Thien Huu Nguyen and Ralph Grishman. 2015. Relation Extraction: Perspective from Convolutional Neural Networks. In *NAACL'15*. 39–48.
[20] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics* 5 (2017), 101–115.
[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP'14*. 1532–1543.
[22] Chris Quirk, Pallavi Choudhury, Jianfeng Gao, Hisami Suzuki, Kristina Toutanova, Michael Gamon, Wen Tau Yih, Lucy Vanderwende, and Colin Cherry. 2012. MSR SPLAT, a language analysis toolkit. In *NAACL'12*. 21–24.
[23] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning Causality for News Events Prediction. In *WWW*. 909–918.
[24] Mehwish Riaz and Roxana Girju. 2014. Recognizing causality in verb-noun pairs via noun and verb semantics. In *EACL'14*.
[25] Patrick Suppes. 1970. *A probabilistic theory of causality*. North-Holland Publishing Company Amsterdam.
[26] Phillip Wolff and Grace Song. 2003. Models of causation and the semantics of causal verbs. *Cognitive Psychology* 47, 3 (2003), 276–332.
[27] R. Xue, Z. Fang, M. Zhang, Z. Yi, C. Wen, and T. Shi. 2013. TCMID: Traditional Chinese Medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Research* 41 (2013).
[28] Hong Yu and Eugene Agichtein. 2003. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics* 19 (2003), 1340–1349.
[29] Ce Zhang. 2015. *DeepDive: a data management system for automatic knowledge base construction*. Ph.D. Dissertation. Citeseer.
[30] Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Bin Wang, and ChengXiang Zhai. 2017. Constructing and Embedding Abstract Event Causality Networks from Text Snippets. In *WSDM'17*. 335–344.