Constructing and Embedding Abstract Event Causality Networks from Text Snippets

Sendong Zhao*, Quan Wang†, Sean Massung‡,
Bing Qin*, Ting Liu* , Bin Wang†, ChengXiang Zhai‡
*School of Computer Science and Technology , Harbin Institute of Technology
†Institute of Information Engineering, Chinese Academy of Sciences
‡Computer Science Department, University of Illinois at Urbana-Champaign
{sdzhao, bqin, tliu}@ir.hit.edu.cn, {wangquan, wangbin}@iie.ac.cn,
{massung1,czhai}@illinois.edu

ABSTRACT

In this paper, we formally define the problem of representing and leveraging abstract event causality to power downstream applications. We propose a novel solution to this problem, which build an abstract causality network and embed the causality network into a continuous vector space. The abstract causality network is generalized from a specific one, with abstract event nodes represented by frequently co-occurring word pairs. To perform the embedding task, we design a dual cause-effect transition model. Therefore, the proposed method can obtain general, frequent, and simple causality patterns, meanwhile, simplify event matching. Given the causality network and the learned embeddings, our model can be applied to a wide range of applications such as event prediction, event clustering and stock market movement prediction. Experimental results demonstrate that 1) the abstract causality network is effective for discovering high-level causality rules behind specific causal events; 2) the embedding models perform better than state-of-the-art link prediction techniques in predicting events; and 3) the event causality embedding is an easy-to-use and sophisticated feature for downstream applications such as stock market movement prediction.

Keywords

Causality, Event causality network, Event prediction, Embedding methods, Stock price movement prediction

1. INTRODUCTION

Causality is the relation between one event (the cause) and a second event (the effect), where the second event is understood as a consequence of the first [5]. Furthermore, causality is not only a kind of knowledge but also the basis behind reasoning and making sense of the unknown. As a kind of knowledge, causality is an essential resource for question answering and decision making. To be able to answer the question *What causes tumors to shrink?*, one would require a large causality repository [18].

However, how to leverage the power of event causality for downstream applications is very challenging and has never been seriously studied before. Some researchers tried methods to use event causality for predicting events [25] and generating future scenarios [15, 14]. Nevertheless, there are three major drawbacks to such studies. First, they focus solely on causality between specific events and fail to discover general causality patterns. For example, they can only detect "*A massive 8.9-magnitude earthquake hit northeast Japan on Friday → a large amount of houses collapsed*", but cannot create the concise and abstract causality pattern "*earthquake → house collapse*". Second, event matching is the key to event prediction and future scenario generation. However, the underlying symbolic nature of tuple matching in [25] or phrase matching in [15] greatly limits the flexibility of event matching, and might degrade the accuracy of event prediction or future scenario generation. Third, although causality itself is an important resource for reasoning and prediction, the symbolic form of causality derived from [25, 15, 14] is hard to generalize to other applications.

Due to these limitations, we propose to 1) build an abstract news event causality network from which we can obtain general, frequent, and simple causality patterns; and 2) embed the causality network into a continuous vector space to simplify event matching and make it easy to use for other applications, resulting in two benefits: 1) general causality patterns help people to better understand the high-level causality rules behind specific causal events; 2) compared to matching tuples or noun phrases, matching in continuous vector spaces is easy to manipulate and much more flexible, which might result in better performance in event prediction or future scenario generation. Besides, event causality embedding may improve application tasks such as stock market prediction for two reasons: 1) low-dimensional dense vectors can effectively alleviate feature sparsity issues and causality gives a reasonable way to embed events into low-dimensional dense vectors; and 2) the causality also benefits from finding the causal factors of

*Corresponding author

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/3018661.3018707

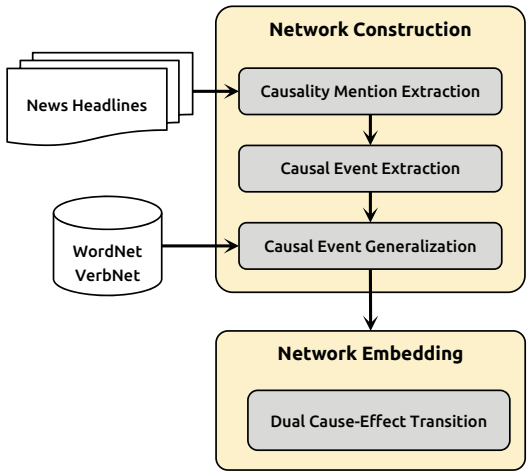

Figure 1: Our proposed framework for abstract event causality mining.

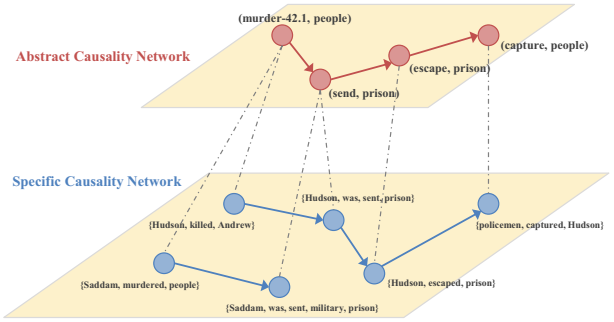

Figure 2: A novel hierarchical causality network. The blue nodes and blue solid directed edges between them on the bottom layer constitute the specific causality network. The red nodes and red dashed directed edges between them on the top layer constitute the abstract causality network. The dashed undirected edges linking the top and bottom layers represent specifications from abstract events to specific ones.

the stock market. Figure 1 sketches our novel framework, which consists of two key parts: a novel causality network construction and a novel causality network embedding.

To construct the causality network, given a collection of text snippets (in our case, news headlines), we propose a novel algorithm that works as follows. We first use textual causality connectors (such as "X because Y" and "X leads to Y") to identify pairs of causality mentions. Then, we extract specific causal events from the identified mentions to obtain high-quality and readable causal pairs, where each specific event is represented by a set of verbs and nouns occurring in the corresponding mention with their original order. The result is a specific causality network where each node is associated with a specific event. In order to discover general patterns which can reveal high-level causality rules and reduce the sparsity of the network, for each specific event represented by nouns and verbs, we generalize the nouns to their hypernyms in WordNet [22] (e.g. the hypernym of the noun "chips" is "dishes"), and the verbs to their classes in VerbNet [30] (e.g. "kill" belongs to the "murder-42.1" class). We propose a novel hierarchical causality generation method to build an abstract causality network on top of the specific one. The nodes of the abstract causality network are frequently co-occurring word pairs, which are general enough and have been demonstrated to be effective in representing certain types of events [29]. In this way, we discover general, frequent, and simple causality patterns which can reveal high-level causality rules behind specific causal events. Figure 2 gives a specific example of the novel hierarchical causality network.

After constructing the abstract causality network, we embed it into a continuous vector space to simplify event manipulation while preserving the inherent cause-effect structure of the original network. Specifically, we design a new dual cause-effect transition model to perform the embedding task. This model abstracts events as vectors in the embedding space, and models the cause-effect relation in different ways. Differing from the existing relation embedding models [8, 7], the proposed dual cause-effect transition model is the first to encode many-to-many relations and enhances the ability to model asymmetry so as to better model the event causality. Specifically, it represents the cause-to-effect and effect-to-cause relations as different transition vectors $\mathbf{t}$ and τ. With this model, we encode several key attributes of event causality, i.e., asymmetry, many-to-many, and transitivity of event causality. Given the learned embeddings, predicting events or generating future scenarios is then formulated as link prediction on the network. By learning and operating on latent continuous representations, causality network embedding can not only simplify event matching, but also capture some unobservable but intrinsic characteristics of causality. All these aspects enhance the link prediction task. Furthermore, this kind of continuous vector form of events which encode the causal context (causes and effects of events) makes it an easy-to-use and sophisticated feature for downstream applications, such as predicting stock market movement.

The contributions of this work are threefold. **First**, we propose a new concept called "abstract causality network" and devise a novel hierarchical causality generation method to generalize specific causality networks. The resulting abstract causality networks contain plenty of general, frequent, and simple causality patterns, which help people to better understand high-level causality rules behind specific causal events. They are much less sparse than specific networks, making it possible to perform the following embedding step. **Second**, for a better use of event causality, we propose to embed abstract causality networks into continuous vector spaces by designing a novel dual cause-effect transition model which encodes key attributes of event causality. To our knowledge, this is the first work on constructing and embedding abstract causality networks. **Third**, our experiment indicates that the embedding of event causality is a valuable resource for stock market prediction.

The rest of the paper is organized as follows. Section 2 introduces the process of constructing event causality networks from text. Section 3 details the proposed dual cause-effect transition model that embed causality networks into continuous vector spaces. Section 4 reports experiments

and results. Section 5 discusses related work, followed by the conclusion and future work in Section 6.

2. CAUSALITY NETWORK CONSTRUCTION

The first stage of our work is to construct a hierarchical causality network from text. In this paper, we use the New York Times headlines archived from 1987 to 2007. In comparison with content of news articles, headlines have some advantages on event causality network construction. First of all, a news headline usually contains at least one event, but a sentence in an article body might not. Hence, there are more events that can be extracted per sentence in news headlines than in news content. Besides, news headlines usually have simpler structures than sentences in news content, and thus are easier to process during causality mention extraction (section 2.1) and causal event extraction (section 2.2). These advantages guided us in making the decision to use headlines to construct the event causality network. The construction process consists of three steps: 1) causality mention extraction, 2) causal event extraction, and 3) causal event generalization.

2.1 Causality Mention Extraction

This step identifies potential cause-effect pairs from unstructured natural language text. As the amount of data may be extremely large (in our experiments, spanning millions of headlines), getting human-annotated pairs is impossible. We therefore provide a procedure similar to the one used in [25], which can automatically identify mentions of causal events from natural language text.

We first consider the textual causality connectors described in [38], and ask human annotators to judge how frequently these connectors express true causality in our data. We sample 100 headlines for each connector and ask human annotators to judge whether these headlines convey causality or not. Each headline is judged by two annotators, and is deemed to express true causality if both annotators think so. Then, we calculate the frequency of true causality for each connector. The four connectors with the highest frequencies are finally selected for causality mention extraction. They are "because", "because of", "lead to" and "after". In fact, more than 80% of event causality in our data is evoked by these four connectors[1].

We then construct a set of rules to extract mentions of causal events. Each rule follows the template of <**Pattern, Constraint, Priority**>, where Pattern is a regular expression containing a selected connector, Constraint is a syntactic constraint on sentences to which the pattern can be applied, and Priority is the priority of the rule if several rules are matched. For instance, given the connector "after", we use the pattern "after [sentence1], [sentence2]" to extract causality mentions with a constraint that [sentence1] cannot start with a number. As a consequence, this pattern can match the sentence "after Iraq war, the oil price falls", but will not match the sentence "After 2 Years, US Capitol Restoration Project Is Complete" It is obvious that (Iraq war, the oil price falls) is a causality but (2 years, US Capitol Restoration Project Is Complete) is not a causality. After applying the rules, we obtain pairs of causality mentions, with one tagged as a cause and the other tagged as the effect. Alternatively, we can use a more sophisticated method to extract causal relation, such as [40]. In this study, however, we put emphasis on the precision of extracted causality rather than the recall. A more sophisticated method also makes this work hard to reproduce.

[1]We sample a subset of the whole dataset and use all connectors mentioned in [38] to extract causality mentions and count the proportion of true causality evoked by these four connectors.

2.2 Causal Event Extraction

The second step is to extract causal events from causality mentions identified in the previous step. Causal events are represented as tuples extended from (Subject, Predicate, Object) triples [25] or as noun phrases [15] in previous work. However, in our case news headlines are usually brief, with a large fraction of constituents omitted, which makes it difficult to extract tuples or noun phrases. We claim that representing each specific event simply by a set of verbs and nouns with their original order is a better choice. Formally, $E = \{W_i | W_i \in Verbs \cup Nouns\}$ in which $Verbs$ and $Nouns$ is the set of verb and nouns in headlines respectively. In particular, the order of W_i is determined by a partial parser [1] and the original order in headlines, which is for the sake of readability of events. This will be illustrated in details using the following examples and three reasons:

Example 1: *Williams retired* [because of] *overuse*, agent says.

Example 2: [After] *the school shootings in Colorado*[,] *the government adopted effective measures to maintain school safety.*

Example 3: [After] *G.O.P. triumph*[,] *Bush sets security department and tax cut as priorities.*

First, events mentioned in news headlines are rarely represented in complete subject-predicate-object structures due to brevity, making it infeasible to extract tuples for most cases. Consider the effect mention "*Williams retired*" in Example 1. It cannot be extracted as a tuple or a noun phrase. However, if we use verbs and nouns, it can be successfully represented by (Williams, retired).

Second, representing events by tuples or noun phrases is likely to miss important information about the events. Consider Example 2 with the effect mention "*the government adopted effective measures to maintain school safety*". Neither the subject-predicate-object triple (the government, adopted, effective measures) nor the noun phrase "*school safety*" contains complete information about the event. However, if we use the verb-and-noun representation (government, adopted, measures, maintain, school, safety), no important information will be lost.

Third, in some cases, multiple tuples or noun phrases can be extracted, and it is difficult to decide which one is the best. For example, given the effect mention "*Bush sets security department and tax cut as priorities*" in Example 3, a regular triple extraction tool will probably extract two triples: (Bush, sets, security department) and (tax, cut as, priorities).[2] It is a non-trivial task to decide which one to be selected as the event representation. However, representing events by verbs and nouns requires no such choices. Furthermore, the performance of tuple extraction

[2]The triple (tax, cut as, priorities) is extracted by mistakenly taking "cut" as a verb.

or noun phrase extraction relies heavily on several more sophisticated NLP tools such as Reverb [12].

From these examples, we can see that representing events simply by ordered sets of verbs and nouns is a better choice. This method can be applied to almost any text data (not just news headlines), covering more comprehensive information about events and placing less reliance on other NLP tools. After extracting causal events, we can construct a specific causality network. Each node of the network corresponds to a specific event, represented by an ordered set of verbs and nouns. Each edge is associated with a cause-effect pair, directed from the cause to the effect.

2.3 Causal Event Generalization

The causality network constructed in the previous step focuses solely on causality between *specific* events, but fails to discover *general* causality patterns between related events. For example, it can only detect "*a massive 8.9-magnitude earthquake hit northeast Japan on Friday → a large amount of houses collapsed*", but cannot summarize to find the general causality pattern of "*earthquake hit → house collapse*". Such general causality patterns are extremely useful in discovering high-level causality rules behind specific causal events. We therefore propose to generalize the specific causality network to get general, frequent, and simple causality patterns. The key idea is to build a hierarchical causality network as illustrated in Figure 2. The bottom layer of the hierarchy consists of specific causal events, and the top layer consists of abstract causal events, generalized from the specific ones.

We first use WordNet [22] and VerbNet [30] to generalize words occurring in specific causal events. Given a specific event represented by a set of nouns and verbs, we generalize each noun to its hypernym in WordNet (e.g. the noun "chips" is generalized to "dishes"), and each verb to its class in VerbNet (e.g. "kill" is generalized to "murder-42.1[3]"). The word generalization step eliminates the negative effect of word diversity, and can help us discover frequent patterns from massive specific causal events.

Next, we devise a hierarchical causality generation method to build an abstract causality network on top of the specific one. We use **f**requently **co**-occurring word **pa**irs (FCOPA) such as (murder-42.1, people), (capture, people), (escape, prison), and (send, prison) to represent abstract events, i.e., nodes of the abstract causality network. A FCOPA refers to two words that co-occur in more than a minimum support of five specific events. Such word pairs are general enough and have been demonstrated to be effective in representing certain types of events [29]. Actually, it is not very difficult to expand these kind of word pair to the set of three or four words. In this study we use word pairs as abstract nodes because of the specific features of news headlines besides the practice of [29]. Given an abstract event, any specific event containing the two words is taken as a specification of the abstract one. For example, in Figure 2, both (Saddam, murdered, people) and (Hudson, killed, Andrew) are specifications of (murder-42.1, people). Edges of the abstract causality network are generated according to those of the specific one. An edge is created from abstract event A to abstract event B if there exists an edge from specific event A_i to specific event B_j, where A_i and B_j are specifications of A and B respectively. For example, in Figure 2, there exists an edge from (Hudson, killed, Andrew) to (Hudson, was, sent, prison) in the specific causality network. So we create an edge between the corresponding abstract events in the abstract causality network, from (murder-42.1, people) to (send prison). Figure 2 provides an illustration of the hierarchical causality network. The blue nodes and the blue solid directed edges between them constitute the specific causality network. The red nodes and the red solid directed edges between them constitute the abstract causality network. The dashed undirected edges linking red nodes and blue nodes represent specifications from abstract events to specific ones.

The advantages of building an abstract causality network are three-fold. First, the abstract causality network contains plenty of general, frequent, and simple causality patterns, which help people to better understand high-level causality rules behind specific causal events, such as "(murder-42.1, people) → (send, prison)→(escape, prison) → (capture, people)". Second, the abstract causality network generalizes much better than the specific one. Given a totally new event extracted from natural language text and then represented by a set of ordered verbs and nouns, matching it exactly with existing specific events is almost impossible. It is much more likely that the new event can be generalized to an abstract event that already exists, and then can be analyzed in the abstract causality network. Third, the abstract causality network is much more dense than the specific one, making it possible to perform the following embedding stage.

3. CAUSALITY NETWORK EMBEDDING MODEL

If we were to use the abstract network directly to do event prediction and stock market prediction, we would run into some problems: 1) event matching is the key to event prediction. However, the underlying symbolic nature of tuple matching in [25] or phrase matching in [15] greatly limits the flexibility of event matching; 2) the symbolic form of causality is hard to generalize to applications such as stock market prediction. To address these problems, we further embed event causality network into a continuous vector space, simplifying event manipulation while preserving the inherent cause-effect structure of the original network. We employ a new model to perform the embedding task: the dual cause-effect transition model. In this paper, we denote scalars by lowercase letters, such as x; vectors by boldface lowercase letters, such as $\mathbf{x}$; and matrices by boldface upper case letters, such as $\mathbf{X}$.

How to model cause-effect relation between events detimines the informativeness of event representations and the ability of event causality for downstream application such as event prediction and stock movement prediction, thus is the core part of our embedding model. In order to simulate event causality as much as possible, our model must embrace the essential characteristics of event causality, i.e., asymmetry, many-to-many and transitivity. The asymmetry and transitivity of event causality are easy to understand. Without asymmetry, our model can not distinguish cause event and effect event. Without transitivity, our model can not capture the long-term event causality. Besides these two characteristics, many-to-many is extremly important especially for event prediction. Without many-to-many, given an event, our model theoretically is just able to find

[3]the form of verb class in VerbNet.

one cause or one effect, which greatly deviates the reality. For example, a case may like that a cause c_1 has two effects e_1 and e_2, i.e., $f_t(c_1) = \{e_1, e_2\}$, the effect e_1 has two causes c_1 and c_2, i.e., $f_\tau(e_1) = \{c_1, c_2\}$. In this many-to-many causality case, f_t and f_τ is neither the same nor linear related. Therefore, we can take the path **cause-to-effect** and the path **effect-to-cause** as different additions $+\mathbf{t}$ and $+\tau$ so as to encode many-to-many, asymmetry and transitivity of event causality. In other words, by modeling these different transitions we encode three key attributes of event causality, i.e., the asymmetry of event causality, the many-to-many relation of event causality and the transitivity of event causality. Specifically, we propose a novel dual cause-effect (Dual-CET) model by designing a new energy function $f(c, e)$ defined on each pair,

$$f(c, e) = \|\mathbf{c} + \mathbf{t} - \mathbf{e}\|_1 + \|\mathbf{e} + \tau - \mathbf{c}\|_1 \quad (1)$$

where true cause-effect pairs are assumed to have low energies.

To learn the event embeddings $\{\mathbf{x}\}$ and the transition vectors $\{\mathbf{t}, \tau\}$, we consider a ranking criterion. Intuitively, given a true cause-effect pair (c, e), if the cause c or the effect e is missing, we would like the model to be able to predict the correct event. The objective of training is to learn the energy function f so that it can successfully rank the true pair (c, e) below all other possible pairs, i.e.,

$$f(c, e) < f(c', e),\ \forall c' \in \mathcal{E} - \{c\},$$
$$f(c, e) < f(c, e'),\ \forall e' \in \mathcal{E} - \{e\},$$

where $\mathcal{E}$ is the set containing all events in the abstract causality network. Towards this goal, the following margin-based ranking loss is minimized:

$$\min_{\{\mathbf{x}\},\mathbf{t},\tau} \sum_{(c,e)\in\mathcal{P}^+} \sum_{(c',e')\in\mathcal{P}^-} \left[\gamma + f(c, e) - f(c', e')\right]_+ + \frac{\alpha}{\|\mathbf{t} + \tau\|_2} \quad (2)$$

where $\mathcal{P}^+$ is the set of true cause-effect event pairs found in the abstract causality network, $\mathcal{P}^-$ contains corrupted pairs constructed by replacing the cause or the effect in (c, e), $\gamma > 0$ is a margin separating true cause-effect pairs and corrupted pairs, and $[x]_+ = \max(0, x)$ denotes the positive part of x, and $\alpha > 0$ is a parameter. The optimization in Eq. (2) favors lower energies for true cause-effect pairs than for corrupted pairs, and is thus a natural implementation of the intended criterion. The optimization is carried out by stochastic gradient descent in mini-batch mode. We enforce the constraints that the embedding of each event $\|X\| = 1$ to avoid overfitting.

4. EXPERIMENTS

In this study, we use two different datasets. For evaluating the performance model on causal event prediction, we use the New York Times corpus (catalog number LDC2008T19) which contains over 1.8 million articles written and published by the New York Times between January 1, 1987 and June 19, 2007, with article metadata provided by the New York Times Newsroom. For simplicity, in the following sections, we refer to our data set as NYT. To predict stock price movement using event causality, we use financial news from Reuters and Bloomberg over the period from October 2006 to November 2013. This time span witnesses a severe economic downturn in 2007-2010, followed by a modest recovery in 2011-2013. There are 106,521 documents in total from Reuters News and 447,145 from Bloomberg News. News titles and contents are extracted from HTML. The timestamps of the news are also extracted, for alignment with stock price information.

Table 1: 10 frequent causality patterns discovered in the NYT data.

Cause Event	Effect Event
(Dow, drop)	(stocks, fall)
(fear, appear-48.1.1)	(price, fall)
(ban, flag)	(public, disorder)
(dollar, lower)	(await, trader)
(neglect-75-1-1, cleanup)	(wildlife, die)
(murder-42.1, people)	(send, prison)
(near, election)	(scandal, appear-48.1.1)
(vote, draw)	(accusation, appeal)
(shock, therapy)	(brain, damage)
(make, dish.n.02)	(cooking-45.3, root_vegetable.n.01)

4.1 Causality Pattern Case Study

The first experiment is to test the effectiveness of the event causality network in discovering general, frequent, and simple causality patterns. We construct the causality network using the steps described in Section 2. Specifically, we use four textual causality connectors "because", "because of", "lead to", and "after" as well as the associated rules to identify causality mentions. We further use the Stanford POS Tagger [31] to recognize verbs and nouns in the identified causality mentions, determine the order of verbs and nouns via partial parser [1] and the original word order in headlines, and represent each specific event by the ordered set of verbs and nouns. Then, we employ the hierarchical causality generation method to build an abstract causality network on top of the specific one. Each node of the abstract causality network corresponds to a FCOPA, standing for an abstract event. We finally extract 1729 abstract events and 3134 cause-effect relationships between them from our NYT data.

Table 1 shows 10 frequent causality patterns between abstract events in the NYT data. Here, the frequency of a causality pattern is calculated by using the specific causality network. Given a causality pattern $A \rightarrow B$ between abstract events A and B, we count the number of edges from specific event A_i to specific event B_j, where A_i and B_j are specifications of A and B respectively. The edge number is taken as the frequency of that causality pattern. We randomly select 10 from the top 50 most frequent causality patterns and list them in Table 1. From the results, we can see that our method is indeed effective in discovering general, frequent, and simple causality patterns which can reveal high-level causality rules behind specific causal events, such as "(Dow, drop) $\rightarrow$ (stocks, fall)", "(murder, people) $\rightarrow$ (send, prison)", and "(near, election) $\rightarrow$ (scandal, appear-48.1.1)", etc.

Given these causality patterns, we can conduct some meaningful analyses, e.g., analyzing possible causes and effects of a given event and discovering causality chains. A subgraph of the abstract causality network is shown in

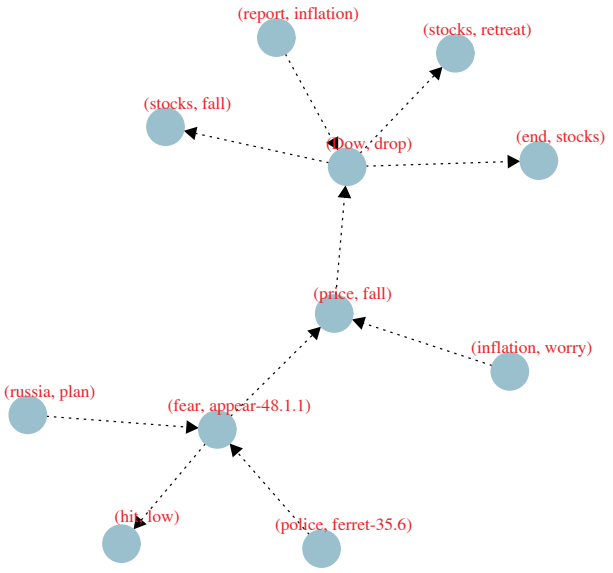

Figure 3: A subgraph of the abstract causality network.

Figure 3. On the graph, given the event (Dow, drop),[4] we can retrieve all its causes such as (price, fall) and (report, inflation), and its effects such as (stocks, fall), (stocks, retreat), and (stocks, end). Furthermore, we can discover interesting causality chains such as "(Russia, plan)→(fear, appear-48.1.1)→ (price, fall)→ (Dow, drop)→ (stocks, retreat)". By inspecting the original related news, we found that the New York Times had large amount of data concerning Russia's unsuccessful military and economic plans which had huge effects on the stock market in the 1990s.

4.2 Causal Event Prediction

The second experiment is to test the effectiveness of the Dual-CET model proposed in Section 3 in predicting causal events. The prediction task is formulated as link prediction on the event causality network, i.e., predicting a cause-effect pair (c, e) with the cause c or the effect e missing.

4.2.1 Evaluation Criterion

To evaluate, we split all the 3134 cause-effect event pairs into a training/validation/test set, with the ratio of 8:1:1. The first is used for model training, the second for hyper-parameter tuning, and the third for evaluation.

For each test cause-effect pair, the cause event is removed and replaced by each of the events in the dictionary in turn. Energies (or dissimilarities) of these corrupted cause-effect pairs are first computed by a specific link prediction method (which are detailed in the following section), and then sorted in ascending order. The rank of the original correct event is finally stored. Similarly, we can get another rank by corrupting the effect event. Aggregated over all test pairs, we report Hits@10, i.e., the proportion of cause (effect) events which we can find at least one correct event ranked in the top 10 corresponding effect (cause) events, which follows the previous studies in multi-relational embedding [8, 7].

[4]The Dow Jones, also called the Dow for short, is a stock market index.

4.2.2 Baseline Methods

First, we compare the Dual-CET model with several state-of-the-art link prediction models summarized in [20] and two embedding models [8, 7] proposed by Bordes et al. Then, the event represented by tuple is taken as the comparison with the event represented by FCOPA for event prediction.

The first four link prediction baselines are **undirected** methods and the following two link prediction baselines are **directed** methods. The last two baselines are embedding-based models which were originally designed for mining multi-relational knowledge graphs of named entities. Lastly, **Jaccard's coefficient**, **Common neighbors** and **Adamic/Adar** are neighborhood-based methods. SimRank$_\zeta$ and Katz$_\beta$ are path-based methods.

Random Given a cause event, we choose an event as the effect randomly and vice versa.

Common neighbors [23] defines the similarity between two events as the number of neighbors they have in common.

Jaccard's coefficient [28] is a statistic used for comparing the similarity and diversity of sample sets.

Adamic/Adar [2] comes out as the winner in the review paper of link prediction methods [20]. This method refines the simple counting of common neighbors by weighting rarer edges more heavily.

Katz$_\beta$ [17] defines the similarity between two events by summing over all paths linking them, exponentially damped by length to count short paths more heavily.

SimRank$_\zeta$ [16] is a randomized version of SimRank by adding a random value ζ^l to the measurement function for encoding the directions of the causality network, which make it an asymmetric link prediction method.

Causal-SE is the causal specialization of SE [8]. Unlike the SE model which was originally designed for multi-relational name entities, the Causal-SE model focuses solely on a single relation, i.e., the cause-effect relation.

Causal-TransE is the causal specialization of TransE [7]. Unlike the TransE model which was originally designed for multi-relational named entities, the Causal-TransE model focuses solely on a single relation, i.e., the cause-effect relation.

In all the baseline methods, we use $Dissim(x, y) = 1 - Sim(x, y)/MAX$ to obtain the dissimilarity between two events x and y, where $Sim(x, y)$ in the measurement function of the above non-embedding baselines and MAX denotes the maximum similarity value.

It is natural to represent events as tuples, and [25] use tuples to represent causal events for event prediction. For comparison, we extract a causal events network in which each node is represented as a tuple. In the comparison, we use the same specific event causality network but different representations of events (the one use ordered bag of verbs and nouns, and the other uses tuples). The first is generalized to an abstract event causality network with nodes represented as FCOPAs. In the second network, each verb and each noun in the tuple are generalized to its classes in VerbNet and its hypernym in WordNet.

4.2.3 Implementation Details

For Random, which does not need model training, we just use the test set to evaluate. For Common neighbors and

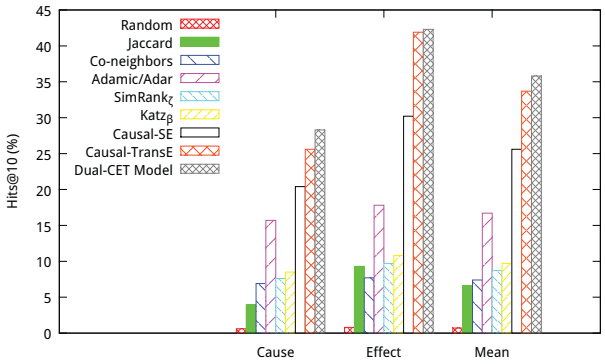

Figure 4: Results on event prediction of all the methods on Hits@10 of the test set by representing event with FCOPA.

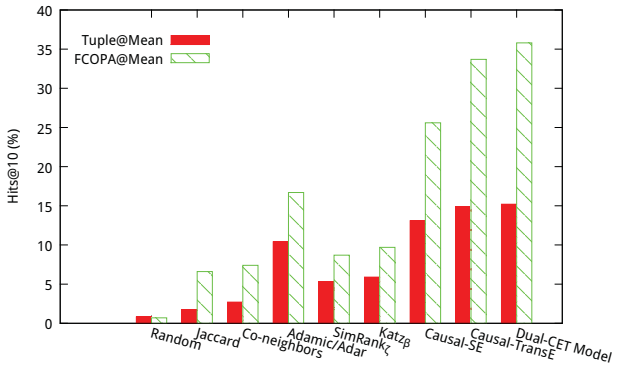

Figure 5: The Mean Hits@10 of all the methods to represent event with FCOPA and Tuple separately, in which FCOPA refers to frequently co-occurring word (verbs and nouns) pairs.

Jaccard's coefficient (which do not have hyperparameters), we use the training set for model training and the test set for evaluation. For the rest of the models including SimRank$_\zeta$, Katz$_\beta$, the Causal-SE model, the Causal-TransE model, and the Dual-CET model (which *do* contain hyperparameters), we use the training set for model training, the validation set for hyperparameter tuning, and the test set for evaluation.

For SimRank$_\zeta$ and Katz$_\beta$ we tune the hyper-parameters ζ and β in $(0, 1)$ with steps of 0.05. The best configurations are $\zeta = 0.8$ for SimRank$_\zeta$ and $\beta = 0.85$ for Katz$_\beta$. Training for SimRank$_\zeta$ is limited to 50 epochs over the training set.

We select the learning rate λ for stochastic gradient descent among $\{0.001, 0.01, 0.1\}$, the margin γ among $\{1, 2, 10\}$, and the latent dimension k among $\{20, 50\}$ on the validation set. Optimal configurations are $k = 20$, $\lambda = 0.01$, $\gamma = 2$ for the Causal-SE model, $k = 50$, $\lambda = 0.01$, $\gamma = 1$ for the Causal-TransE model and $k = 40$, $\lambda = 0.01$, $\gamma = 1$, $\alpha = 0.1$ for the Dual-CET model . Training is limited to at most 1,000 epochs over the training set. The best models are selected by early stopping on the validation sets (by monitoring Hits@10).

4.2.4 *Results and Discussions*

Figure 4 shows the results of all the models on the test set, where "Cause" means predicting causes, "Effect" is predicting effects, and "Mean" is the Hits@10 values averaged over the two cases. From the results, we can see that the embedding-based Causal-SE, Causal-TransE, and Dual-CET models perform significantly and consistently better than the non-embedding-based ones. By learning and operating on latent representations, the embedding-based models are able to capture some unobservable but intrinsic characteristics of causality, and hence perform better in predicting causal events. Between the three embedding based models, our proposed Dual-CET model consistently outperforms Causal-SE and Causal-TransE. The reason might be that the Dual-CET model has fewer parameters to learn than Causal-SE, which greatly reduces the risk of over-fitting. On the other hand, the Dual-CET model captures key specific attributes of event causality by modeling cause-to-effect and effect-to-cause as different transitions, which benefits event prediction. All the models perform better in predicting effects than in predicting causes. This is because in our data an event gets more causes than effects on average (2.88 causes vs. 2.45 effects), making predicting causes harder than predicting effects.

Table 2: Examples of the Dual-CET model in predicting effect events.

Input Cause	Top-5 Predicted Effects
(Dow, drop)	(stocks, retreat) (stocks, end) (report, inflation) (fall, bond) (price, fall)
(Iraq, war)	(iraq, attack) (murder-42.1-1, bomb) (talks, lag) (fall, oil) (price, fall)
(near, election)	(other_cos-45.4, turmoil) (rise, tension) (new, premier) (election, year) (tv, ad)

In order to verify the effectiveness of embedding methods (i.e. Causal-SE, Causal-TransE and Dual-CET) in learning asymmetric cause and effect relationships, we conduct the following sanity check: (1) train the model; (2) record the loss on the training set; (3) swap cause and effect in all training pairs; (4) evaluate the model on this modified set; (5) record the difference on loss with step 2. After swapping causes and effects in training pairs, the loss increases by 33.9% for the Causal-SE, 10.5% for Causal-TransE and 47.4% for the Dual-CET model, which also demonstrate the ability of the Dual-CET model to capture the asymmetry of event causality.

Figure 5 compares performance of predicting events by using FCOPA and Tuple as event representations, where "@Mean" indicates the Hits@10 values averaged over causes and effects Hits@10. The comparison shows that it performs better to predict events if we use FCOPA to represent events. In Section 2.2, we explained several advantages of FCOPA. In addition, it is easier to get a high-level representation of events and obtain a more dense event causality network, which are crucial for link prediction in the event causality network.

Table 3: Overall stock price movement prediction results (%).

	Accuracy
ET+SVM (Ding [10])	53.72
EoC$_{\textbf{Causal-TransE}}$+SVM	55.41
EoC$_{\textbf{Dual-CET}}$+SVM	**56.76**
ET+DNN (Ding [10])	56.08
EoC$_{\textbf{Causal-TransE}}$+DNN	57.77
EoC$_{\textbf{Dual-CET}}$+DNN	**59.80**

Table 2 further gives several examples of the Dual-CET model in predicting effect events. For each given cause, the top five most plausible effects are listed. The results demonstrate the capability of the Dual-CET model in causal event prediction.

4.3 Stock Prediction Using Event Causality

Predicting stock price movements is of clear interest to investors, public companies and governments. Recent studies [35, 10] have applied text mining techniques to help analyze the effect of text on the Web on stock market prediction, finding that events reported in news are important evidence. Particularly, [10] proposed event-based stock price movement prediction, which demonstrated that news is a very effective resource to predict stock price movement (i.e., increase or decrease of future stock prices).

We use textual causality connectors that are used in [40] and extract 12,482 events with 20,392 causality pairs on Ding's data [10]. The stock price movement data from 02/10/2006 to 21/02/2013 is our training set and the stock price movement data from 22/02/2013 to 21/11/2013 is our testing set. Our experiments are carried out on long-term (one month) stock price movements using 12,482 events during this period. We mainly focus on predicting the increase or decrease of Standard & Poor's 500 stock (S&P) index. The binary classification problem is to determine whether the index price will increase or decrease the next month. In this experiment, we take the event causality embedding as the input to Ding's prediction framework. In order to check the effectiveness of the embedding of event causality for stock price movement prediction in comparison to event tuples, we evaluate two models (SVM and deep neural networks) taking an input event tuple (ET) and embedding of event causality (EoC) generated from Causal-TransE and Dual-CET to predict stock price movement on the same dataset.

We follow the same evaluation metrics as Ding. In terms of SVM, we just use SVM with a linear kernel. For the parameter C, we perform a grid search to get the best value. For the DNN, we use a neural network with two hidden layers and the number of nodes in each hidden layer is $2n+1$, where n is the number of input nodes.

Table 3 shows the results. The comparison of ET, EoC generated from Causal-TransE and EoC generated from Dual-CET on two classic models show that event causality embedding improves the performance of stock price movement prediction, demonstrating the embedding of event causality is a valuable resource. The reason why the EoC outperforms the ET might rely on two aspects. First, low-dimensional dense vectors can effectively alleviate the problem of feature sparsity which is usually a disaster for machine learning prediction models. The causality between events gives a reasonable way to embed events into continuous vectors. Second, the causality also enables finding the causal factors of stock price movement. As shown in Figure 3, event "(Russia, plan)" has an effect on stocks: "(stocks, retreat)". Therefore, this kind of high-level events causality embedding is quite meaningful for stock price movement prediction.

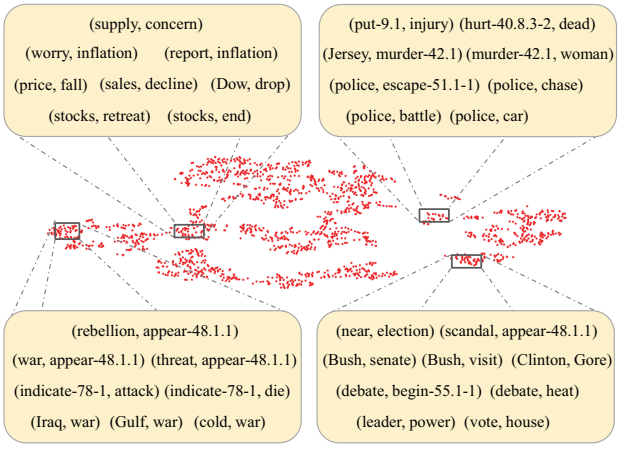

Figure 6: The distribution of events in the NYT data. The results indicate that similar events (or semantically related events) indeed lie close to each other in the embedding space. The events in the top left corner cluster are about "economy and stocks", the top right corner about "police and criminals", the bottom left corner "wars", and the bottom right corner "election and voting".

4.4 Event Embedding Analysis

The third experiment is to test the effectiveness of the learned embeddings in clustering similar events. To do so, we use the event embeddings learned by the Dual-CET model with the optimal parameter settings determined in the event prediction task. We create a 2D plot using t-SNE [32][5]. Figure 6 shows the distribution of events in the NYT data. The results indicate that similar events (or semantically related events) indeed lie close to each other in the embedding space. The events in the top left corner cluster are about "economy and stocks", the top right corner about "police and criminals", the bottom left corner "wars", and the bottom right corner "election and voting". All these observations demonstrate the effectiveness of the embedding models in similar event clustering.

5. RELATED WORK

To our knowledge, there is no previous study which considers the task we confront: 1) building a hierarchical causality network to discover high-level abstract causality rules, and 2) embedding the causality network into a continuous vector space for better event matching and prediction. This section reviews three lines of related

[5]http://lvdmaaten.github.io/tsne/

work: causality extraction, event prediction, and embedding learning.

Event Causality Extraction. In computational linguistics, many studies deal with extraction of causality relations from text. Event causality extraction is a fundamental task because causality between events can be used in many applications. Causality is an effective basis to predict future [25]. In order to predict future events, [25] extracted causal relations between events from a large-scale news corpus. Similarly, [15] proposed a supervised method of extracting event causality from the web to generate future scenarios.

There are some approaches proposed concerning the automatic extraction and detection of causality. For event causality extraction and detection, clues used by previous methods can be roughly categorized as causality patterns [25, 40], words in context [24], associations among words [11] and the semantics of predicates and nouns [13, 27]. Among the four types of clues, causality patterns perform the best in causality extraction. This work employs pattern-based techniques to extract causality between events. In terms of event extraction, we compared our method with Reverb on the event prediction task.

Event Prediction. Event prediction is to predict future events based on those that already happened. Association rules [26, 19, 3] and causality rules [25, 15, 39] are typical clues for prediction. Association rules perform well if there are plenty of events associated with each other. But they fail to predict entirely new events which never happened before or rare events which happened only a few times. Causality is a strong principle to predict the future with good interpretability. In causality based methods, prediction is conducted via event matching, making such methods capable of handling new/rare events. However, the event representations of tuples in [25] and noun phrases in [15] greatly limits the flexibility of event matching. Our work differs in that we construct an abstract event causality network which shows more essential casual rules and further embed the event causality network into a continuous vector space so as to simplify the matching and the prediction.

Embedding Learning in Text. In text mining, the embedding idea has been widely used based on language model [4]. Usually, it is used to learn the embedding of each word. Several NLP tasks can be improved by this kind of word embedding, which are confirmed by Collobert and Weston [9]. Similarly, Weston et al. [37] have successfully applied the embedding idea for matching queries and images. This work mapped queries and images into a common semantic space following meaningful data representations and state-of-the-art results. Bordes et al. [6] adapted this model to a (very) small custom knowledge base for language understanding. In order to solve the problem of embedding entities and relationships of multi-relational data in low-dimensional vector spaces, Bordes et al. proposed several models including SE [8], TransE [7]. Based on these studies of Bordes, improved versions of relation embedding models are proposed [36, 34, 21, 33]. All these studies demonstrate that encoding data in distributed embeddings induce gains in performance.

We proposed a novel Dual-CET model to learn embeddings of causal events. In comparison with SE and TransE, the Dual-CET model considers the specific attributes of event causality by modeling cause-to-effect and effect-to-cause as different transitions.

6. CONCLUSION

In this paper, we studied the problem of building an event causality network from text and using the causality network for future event prediction. We proposed a new concept "abstract causality network" and a hierarchical causality generation method to build the abstract causality network on top of a specific one. From this network, we can obtain plenty of general, frequent, and simple causality patterns which reveal high-level causality rules behind specific causal events. Furthermore, we designed a novel Dual-CET model which encoded key attributes of causality to embed the causality network into a continuous vector space. Future event prediction was formulated as a link prediction task on the causality network using event embeddings. By learning and operating on latent continuous representations, the embedding methods could greatly improve the flexibility of event matching and hence improve the accuracy of event prediction.

We experimented with a New York Times corpus containing over 1.8 million news headlines. Experimental results demonstrated 1) the effectiveness of the abstract causality network in discovering general, frequent, and simple causality patterns, 2) the superiority of the embedding methods over state-of-the-art link prediction techniques in future event prediction, and 3) the use and value of the event causality embedding in stock market movement prediction.

In the future, we would like to apply the discovered causality rules in more applications, such as medical cause-and-effect discovery. Predicting specific causal events instead of abstract ones is another focus.

7. ACKNOWLEDGMENTS

We are grateful to Prof. Wangxiang Che, Jing Liu, Jiang Guo and the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Key Basic Research Program of China (973 Program) via grant 2014CB340503 and the National Natural Science Foundation of China (NSFC) via grants 61133012, 61472107 and 61402465. This work was done while the author was visiting Institute of Information Engineering Chinese Academy of Sciences and University of Illinois at Urbana-Champaign.

8. REFERENCES

[1] S. Abney. Part-of-speech tagging and partial parsing. In *Corpus-Based Methods in Language and Speech Processing*, pages 118–136. Springer Netherlands, 1997.

[2] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[3] G. Amodeo, R. Blanco, and U. Brefeld. Hybrid models for future event prediction. In *CIKM'11*, pages 1981–1984, 2011.

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *JMLR*, 3:1137–1155, 2003.

[5] B. Berofsky. The cement of the universe: A study of causation, 1977.

[6] A. Bordes, N. Usunier, R. Collobert, and J. Weston. Towards understanding situated natural language. In *AISTATS'10*, pages 65–72, 2010.

[7] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS'13*, pages 2787–2795, 2013.

[8] A. Bordes, J. Weston, R. Collobert, and Y. Bengio. Learning structured embeddings of knowledge bases. In *AAAI'11*, 2011.

[9] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML'08*, pages 160–167, 2008.

[10] X. Ding, Y. Zhang, T. Liu, and J. Duan. Using structured events to predict stock price movement: An empirical investigation. In *EMNLP'14*, pages 1415–1426, 2014.

[11] Q. X. Do, Y. S. Chan, and D. Roth. Minimally supervised event causality identification. In *EMNLP'11*, pages 294–303, 2011.

[12] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP'11*, pages 1535–1545, 2011.

[13] C. Hashimoto, K. Torisawa, S. De Saeger, J.-H. Oh, and J. Kazama. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *EMNLP'12*, pages 619–630, 2012.

[14] C. Hashimoto, K. Torisawa, J. Kloetzer, and J.-H. Oh. Generating event causality hypotheses through semantic relations. In *AAAI'15*, 2015.

[15] C. Hashimoto, K. Torisawa, J. Kloetzer, M. Sano, I. Varga, J.-H. Oh, and Y. Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *ACL'14*, pages 987–997, 2014.

[16] G. Jeh and J. Widom. Simrank: A measure of structural-context similarity. In *KDD'02*, pages 538–543, 2002.

[17] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[18] Z. Kozareva. Cause-effect relation learning. In *TextGraphs-7 Workshop on Graph-based Methods for Natural Language Processing*, pages 39–43, 2012.

[19] B. Letham, C. Rudin, and D. Madigan. Sequential event prediction. *Machine learning*, 93(2-3):357–380, 2013.

[20] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031, 2007.

[21] Y. Lin, Z. Liu, H. Luan, M. Sun, S. Rao, and S. Liu. Modeling relation paths for representation learning of knowledge bases. In *AAAI'15*.

[22] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[23] M. E. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64, 2001.

[24] J.-H. Oh, K. Torisawa, C. Hashimoto, M. Sano, S. De Saeger, and K. Ohtake. Why-question answering using intra- and inter-sentential causal relations. In *ACL'13*, pages 1733–1743, 2013.

[25] K. Radinsky, S. Davidovich, and S. Markovitch. Learning causality for news events prediction. In *WWW '12*, pages 909–918, 2012.

[26] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *WSDM'13*, pages 255–264, 2013.

[27] M. Riaz and R. Girju. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *SIGDIAL'14*, page 161, 2014.

[28] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[29] E. Schubert, M. Weiler, and H.-P. Kriegel. Signitrend: Scalable detection of emerging topics in textual streams by hashed significance thresholds. In *KDD'14*, pages 871–880, 2014.

[30] K. K. Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2005.

[31] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL'03*, pages 173–180, 2003.

[32] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9(2579-2605):85, 2008.

[33] Q. Wang, J. Liu, Y. Luo, B. Wang, and C. Lin. Knowledge base completion via coupled path ranking. In *ACL'16*, pages 1308–1318, 2016.

[34] Q. Wang, B. Wang, and L. Guo. Knowledge base completion using embeddings and rules. In *AAAI'15*, pages 1859–1865, 2015.

[35] W. Y. Wang and Z. Hua. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *ACL'14*, 2014.

[36] Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119, 2014.

[37] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81:21–35, 2010.

[38] P. Wolff and G. Song. Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47(3):276–332, 2003.

[39] S. Zhao and T. Liu. Causality and its applications in social media: A survey. *Ruan Jian Xue Bao/ Journal of Software*, 25(12):2733–2752, 2014.

[40] S. Zhao, T. Liu, S. Zhao, Y. Chen, and J.-Y. Nie. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950, 2016.