Boosting Causal Embeddings via Potential Verb-Mediated Causal Patterns

Zhipeng Xie[*] and **Feiteng Mu**
Shanghai Key Laboratory of Data Science, Fudan University
School of Computer Science, Fudan University
{xiezp, 17210240011}@fudan.edu.cn

Abstract

Existing approaches to causal embeddings rely heavily on hand-crafted high-precision causal patterns, leading to limited coverage. To solve this problem, this paper proposes a method to boost causal embeddings by exploring potential verb-mediated causal patterns. It first constructs a seed set of causal word pairs, then uses them as supervision to characterize the causal strengths of extracted verb-mediated patterns, and finally exploits the weighted extractions by those verb-mediated patterns in the construction of boosted causal embeddings. Experimental results have shown that the boosted causal embeddings outperform several state-of-the-arts significantly on both English and Chinese. As by-products, the top-ranked patterns coincide with human intuition about causality.

1 Introduction

Causation or causality is a special semantic relationship between one cause process (or state) and another process (or state), where the cause is partially responsible for the effect, and the effect is partially dependent on the cause. Causation plays an important role in human thinking and reasoning. Whenever something serious happens, human usually tries to determine its causes and its effects.

Human is able to figure out causalities in text for at least two reasons. On the one hand, human can recognize lexical units that explicitly indicate causalities, such as `because`, `cause`, `result in` and `as a consequence`, which makes it possible to detect causalities even when human has no knowledge about the meaning of the cause or the effect. On the other hand, human can understand the meaning of language and has both commonsense and domain-specific knowledge, and thus can detect causalities even when there are no explicit causal lexical units. These two cognitive characteristics get mixed together and assign human the ability of causality understanding and causal reasoning. The results of causality understanding and causal reasoning can enrich human's knowledge, the other way round.

It is challenging for computer to perform such a task. For many years, NLP community has made a lot of efforts to causality-related tasks, resulting in a lot of proposed methods. Ealiest approaches were based on hand-coded domain-specific knowledge bases to extract explicit causal knowledge from text [Kaplan and Berry-Rogghe, 1991] and to rank the extracted possible causalities [Girju and Moldovan, 2002]. These works achieved satisfactory precision but low recall. Newer approaches employed machine learning algorithms to extract explicit and implicit causalities from text [Girju, 2003; Chang and Choi, 2006; Hashimoto *et al.*, 2015a], which relies heavily on feature engineering. Automatic detection of causalities has played a fundamental role in various downstream applications such as question answering [Oh *et al.*, 2013], event prediction [Radinsky *et al.*, 2012] and future scenario generation [Hashimoto *et al.*, 2014].

In recent years there has been an increasingly interest in using word embeddings [Collobert *et al.*, 2011; Mikolov *et al.*, 2013] as an alternative to traditional hand-crafted features. More recent studies have focused on how to build and use task-specific word embeddings [Tang *et al.*, 2014; Boros *et al.*, 2014; Nguyen and Grishman, 2014; Hashimoto *et al.*, 2015b; Li *et al.*, 2017]. As for the cause-effect relationship, several methods have been proposed in [Sharp *et al.*, 2016; Xie and Mu, 2019] to construct causality-specific word embeddings (or causal embeddings in short) from a set of causal phrase pairs extracted via a handful of high-precision patterns.

The construction of causal embeddings is a valuable causality-related task in several aspects. Firstly, causal word pairs are important lexical knowledge resource which is helpful to causality extraction [Chang and Choi, 2006]. Causal embeddings can provide a simple way to measure the causal affinity of word pairs, or equivalently, provide a large scale collection of causal word pairs. Secondly, with causal embeddings, the causal affinity between two phrases can be measured as the maximal causal affinity of all the word pairs between them, which can facilitate detecting implicit causality detection and measuring causal strengths of potential patterns. Last but not the least, causal embeddings are complementary to vanilla word embeddings in downstream applications such as why-question answering [Sharp *et al.*, 2016; Xie and Mu, 2019].

As for the causal embedding methods, existing works

[*] Contact Author

[Sharp *et al.*, 2016; Xie and Mu, 2019] usually have to first extract a large number of causal phrase pairs by a handcrated set of strong causal patterns, and then construct causal embeddings from them. The resulting causal embeddings have shown high precision, but relatively low coverage, because a large number of textual causalities are connected only by weak patterns. To solve this problem, we propose a novel method (called *BoostedMaxMatching*, or *BMM* in short) to boost the causal embeddings by exploring potential weak causal patterns. This paper makes the following contributions:

- Firstly, it proposes an extraction method for verb-mediated patterns, defines an inversion operation of patterns to handle the causal directionality problem, and designs an algorithm to learn pattern weights that indicate their causality-indicating strengths.
- Secondly, it proposes a method to boost causal embeddings by fusing the existing strong causal phrase pairs and the extactions from (possibly weak) verb-mediated patterns. Experimental results have shown that the resulting causal embeddings have achieved high precision and recall on both Chinese and English test data.
- Finally, our method also ouputs verb-mediated patterns together with their causal strengths, as byproducts. Examination of them has shown that the top-ranked patterns coincide with human intuition about causality.

2 Related Work

To the best of our knowledge, there are two existing works on causality-specific embeddings [Sharp *et al.*, 2016; Xie and Mu, 2019] which are closely related to this paper.

Sharp et al. [2016] did the first work on causal embeddings and proposed a series of *cEmbed*-family methods. They generated a causality-specific embeddings and demonstrated that these dedicated embeddings are helpful in a downstream causal QA task. Their *cEmbed*-family methods for embedding construction are based on the Skip-Gram algorithm [Mikolov *et al.*, 2013], by treating an effect phrase as the context of its cause and a cause phrase as the context of its effect. However, such treatment is based on the assumption that each word pair between a cause-effect phrase pair is causally related. Due to the fact that the assumption is far from reality and thus introduces too much noise, the performance of cEmbed-family models is undistinguished.

Xie and Mu [2019] proposed three methods (Pairwise Matching, Max Matching, and Attentive Matching) for building causal embeddings from a corpus of cause-effect phrase pairs. In order to transfer causal relationship from phrase-pair level to word-pair level, *Max-Matching* method assumes that there is at least one word pair carrying the causality information of the positive phrase pair, which can be thought of as a special case of multi-instance learning. *Attentive-Matching* takes the assumption that for a causal phrase pair, there must exist at least one close interaction between a cause word and the effect phrase and another between an effect word and the cause phrase. Both *Max-Matching* and *Attentive-Matching* have achieved satisfactory performance on both English and Chinese corpora.

The research work on vanilla and task-specific word embeddings is also related to our work. In recent years, there is an upsurge of deep learning in natural language processing [Collobert *et al.*, 2011], where distributed representation of words serves as the basis. The neural methods that learn vanilla distributed representation of words (called *word embeddings*) usually capture only co-occurrence relationships between words [Mikolov *et al.*, 2013]. Although such a general-purpose word embeddings are helpful for various NLP tasks, the acquisition of generality is often at the cost of losing specificity to a certain degree. Tang et al. [2014] proposed learning sentiment-specific word embedding for sentiment analysis, where sentiment information is encoded into the continuous representation of words such that it can separate *good* and *bad* to opposite ends of the spectrum. Hashimoto et al. [2015b] proposed a novel method to train word embeddings for semantic relation classification, by predicting words between noun pairs using lexical relation-specific features. Li et al. [2017] developed a tailored neural network to learn contradiction-specific word embedding.

3 Method

Existing works on causal embedding (inclusive of Sharp et al. [2016], Xie and Mu [2019]) extract a set of causal phrase pairs by a handful of hand-crafted causal patterns, and then build causal embeddings from the phrase pairs, which may suffer from limited coverage. To remedy this deficiency, we propose a novel method to take weak-causality patterns into consideration, in order to boost the coverage of the learned causal embeddings. The architecture of our method is illustrated in Figure 1, which consists of four main components:

- **Triple extraction via verb-mediated patterns:** This paper focuses on the verb-mediated patterns and treats their extractions;
- **Causal word pair seeding:** A high-quality seed set of causal word pairs is generated, which will provide supervision information to the phase of pattern weighting;
- **Pattern weighting and scoring:** The candidate patterns are weighted and scored according to their ability to extract causal pairs;
- **Boosted causal embeddings via weak-causality patterns:** The phrase pairs extracted by the verb-mediated patterns are weighted and filtered according to the pattern scores and then get fused with the previously extracted high-precision causal phrase pairs, to construct the boosted causal embeddings.

3.1 Triple Extraction via Verb-Mediated Patterns

Verb-mediated patterns that indicate causation are a most frequent kind of explicit intra-sentential causal pattern [Girju, 2003]. In this paper, we focus on three simplest easy-to-extract forms, which work on English corpus as follows, after parsing the corpus with the SpaCy[1] dependency parser:

- for each transitive verb with active voice, locate its `nsubj` or `csubj` as the subject phrase, its `dobj` or

[1] http://spacy.io

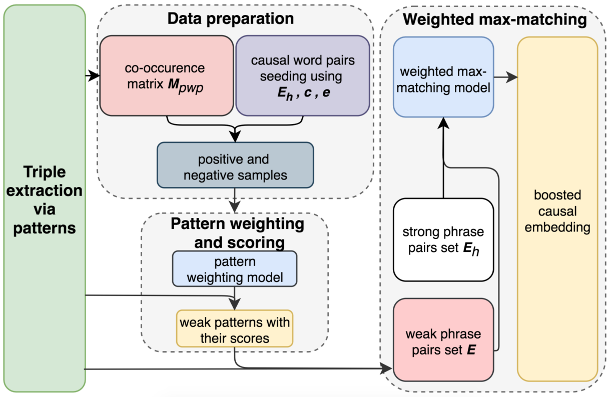
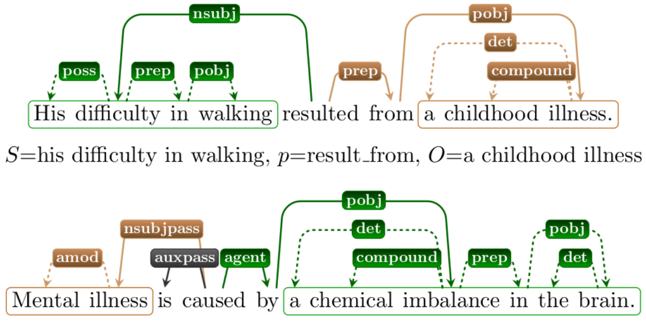

Triple extraction via patterns
Data preparation
co-occurence matrix M_{pwp}
causal word pairs seeding using E_h , c , e
positive and negative samples
Pattern weighting and scoring
pattern weighting model
weak patterns with their scores
Weighted max-matching
weighted max-matching model
strong phrase pairs set E_h
weak phrase pairs set E
boosted causal embedding

Figure 1: The architecture of the boosted causal embeddings.



S=his difficulty in walking, p=result_from, O=a childhood illness

S=a chemical imbalance in the brain, p=cause, O=mental illness

Figure 2: Examples of triple extractions

`ccomp` extended by its `oprd`, `xcomp`, and `acomp` as the object phrase, and use the verb's lemma as the pattern;

- for each transitive verb with passive voice, locate its `agent` as the subject phrase, its `nsubjpass` or `csubjpass` as the object phrase, and use the verb's lemma as the pattern;
- for each intransitive verb (with active voice), locate its `nsubj` or `csubj` as the subject phrase, the `pobj` or `pcomp` of its `prep` as the object phrase, and use the verb's lemma plus the preposition as the pattern.

A phrase is valid if it contains at least a content word (i.e., noun, verb, adjective or adverb). Only the extractions whose subject and object phrases are both valid will be kept. Figure 2 illustrates several extraction examples. On Chinese corpus, the dependency parser in LTP toolkit[2] is used instead and some minor modifications are made to the rules accordingly.

Applying the above rules on a raw text corpus, we can obtain a dataset $\mathbb{E}$ of triples $\langle S, p, O\rangle$, where S is the subject phrase, O is the object phrase, and p is the pattern that is either simply a transitive verb or an intransitive verb plus a following preposition.

For these causality-indicating patterns, the causal directionality deserves our attention. Some of them such as `prompt` have their subject phrases as cause, for example in the first sentence "*The powerful typhoon prompts a huge flood*", while the others such as `result_from` have their object phrases as cause, for example in the second sentence "*A huge flood resulted from the powerful typhoon*", where `(typhoon, flood)` is known to be a causal word pair. Therefore, it is desirable to create a new pattern by inverting an existing one. Specifically, for a given pattern p, we can make its inverted version p^T by inverting the subject phrase and the object phrase of p. That is, for each triple $\langle S, p, O\rangle \in \mathbb{E}$, we will add its inverted triple $\langle O, p^T, S\rangle$ into $\mathbb{E}$. For example, the triple ⟨`A huge flood, result_from, the powerful typhoon`⟩ extracted from the second sentence will be inverted to ⟨`the powerful typhoon, result_from_inv, A huge flood`⟩, where `result_from_inv` is the inverted pattern of `result_from`. Clearly, inverting the patterns will double the size of dataset $\mathbb{E}$. Let $P = \{p_1, p_2, \ldots, p_K\}$ denote the set of patterns in $\mathbb{E}$, and $\gamma(p)$ denote the set of subject-object phrase pairs extracted with a pattern $p \in P$.

In addition, these verb patterns vary significantly in the strength of indicating causality, due to the ambiguity of verbs. Some verb patterns such as `cause` and `result_from` always convey strong causal signal, some patterns such as `contain` and `consist_of` have weak or even no signal of causality.

3.2 Seed Set of Causal Word Pairs

At the starting point of our method, it is assumed that there are

- a collection $\mathbb{E}_h$ of high-precision causal phrase pairs,
- a cause embedding function $\mathbf{c}$ which maps a word w to its cause embedding $\mathbf{c}(w)$ in a d-dimensional Euclidean space $\mathcal{S}^c$ (called the cause space), i.e. $\mathbf{c}(w) \in \mathbb{R}^d$, and
- an effect embedding function $\mathbf{e}$ which maps a word w to its effect embedding in another d-dimensional Euclidean space $\mathcal{S}^e$ (called the effect space), i.e. $\mathbf{c}(w) \in \mathbb{R}^d$,

where the embeddings $\mathbf{c}$ and $\mathbf{e}$ are built up from $\mathbb{E}_h$ by one of the cEmbed-family methods [Sharp *et al.*, 2016] or the Max-Matching and Attentive-Matching algorithms [Xie and Mu, 2019]. To measure the causality of word pair (w_1, w_2), the causal interaction score $cs(w_1, w_2)$ is defined as the inner product between the cause embedding of w_1 and the effect embedding of w_2:

$$cs(w_1, w_2) = \mathbf{c}(w_1)^\top \cdot \mathbf{e}(w_2) \tag{1}$$

Identifying Candidate Causal Word Pairs

Given a high-precision causal phrase pair $(C, E) \in \mathbb{E}_h$, the word pair (c, e) with the highest causal interaction score betweeen C and E is thought of as the dominant word pair, that is:

$$(c, e) = \underset{(w_1, w_2): w_1 \in C, w_2 \in E}{\arg\max} \; cs(w_1, w_2) \tag{2}$$

A word pair (w_1, w_2) is a candidate causal word pair if it serves as the dominant word pair of at least λ_1 phrase pairs in $\mathbb{E}_h$ and its causal interaction score is not less than λ_2 (λ_1 and λ_2 are set to 30 and 0.55 by default).

[2] http://github.com/HIT-SCIR/pyltp

Generating Seed Causal Word Pairs by Pruning

The candidate causal word pairs are then pruned according to the irreflexive and antisymmetric properties of causal relationship:

- We remove all the word pairs (w, w) whose cause and effect words are identical, because of the irreflexity of causal relation;
- For two different words, $w_1 \neq w_2$, if (w_1, w_2) and its reverse (w_2, w_1) belong to the seed set at the same time, we remove both of them, because the causal relation is antisymmetric.

The resulting set of seed causal word pairs is denoted as $\mathbb{S}$, which will provide supervision information for learning pattern weights.

3.3 Pattern Weighting and Scoring

Since different patterns usually have different strengths of causality, we would like to assign a weight to each pattern. The problem is how to learn these weights and where the supervision information comes from. In Section 3.1, we have built up the connections between patterns and phrase pairs in $\mathbb{E}$, which is equivalent to a co-occurrence matrix of patterns and phrase pairs (or PPP matrix in short), denoted as $\mathbf{M}_{ppp}$. The dataset $\mathbb{E}_h$ can serve as the gold standard, but it is not helpful in practice to tell us which phrase pairs in $\mathbf{M}_{ppp}$ are positive (or causally-related), because of the sparsity of phrase pairs (i.e., a phrase pair seldom co-occurs more than one time in the text). Therefore, we cannot infer the pattern weights directly on the level of phrase pairs.

Pattern Weighting

To solve this problem, let us move to the level of word pairs and seek for possible solution. We first build connections between patterns and word pairs in a so-called co-occurrence matrix of patterns and word pairs (PWP matrix in short, denoted as $\mathbf{M}_{pwp}$), where the element $f_k(w_1, w_2)$ at the row of word pair (w_1, w_2) and the column of pattern p_k indicates how many times w_1 and w_2 are linked by p_k:

$$f_k(w_1, w_2) = |\{(S, O) \in \gamma(p_k) : w_1 \in S, w_2 \in O\}|. \quad (3)$$

For the i-th word pair in $\mathbf{M}_{pwp}$, it is labeled as positive ($y_i = 1$) if it belongs to the seed set $\mathbb{S}$; otherwise, it is negative ($y_i = 0$). Since the positives and negatives are ready and the examples (or word pairs) are described by a feature set $\{f_1, \ldots, f_K\}$, we propose a pattern-weighting method as following, similar to logistic regression.

The causal affinity ca of a word pair (w_1, w_2) is measured by the average weight of all its features:

$$ca(w_1, w_2) = \sigma\left(\frac{\sum_{k=1}^{K} a_k \times f_k(w_1, w_2)}{\sum_{k=1}^{K} f_k(w_1, w_2)}\right) \quad (4)$$

where $\sigma(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$ is the sigmoid function that makes normalization.

The weight vector $\mathbf{a} = (a_1, \ldots, a_K)$ is trainable to minimize the cross-entropy loss with l_2 regularization:

$$\begin{aligned} L = &-\frac{1}{N}\sum_{i=1}^{N} y_i \log ca(wp_i) + (1 - y_i)\log(1 - ca(wp_i)) \\ &+ \lambda\|\mathbf{a}\| \end{aligned} \quad (5)$$

where wp_i is the i-th word pair in $\mathbf{M}_{pwp}$, $N = |\mathbf{M}_{pwp}|$ is the total number of word pairs in $\mathbf{M}_{pwp}$, and λ is the regularization coefficient that is set to 10^{-5} by default. It is expected that causality-indicating patterns get positive weights and word pairs connected by many causality-indicating patterns will receive high causal affinities.

Pattern Scoring

Given a pattern p_i, it is of interest to know the average causal affinity of the phrase pairs it connects, measured by the score $s(p_i)$:

$$s(p_i) = \frac{1}{\|\gamma(p_i)\|} \sum_{(C,E)\in\gamma(p_i)} CA(C, E), \quad (6)$$

where the causality affinity CA of phrase pair (C, E) is calculated as the maximum causal affinity among the word pairs between C and E:

$$CA(C, E) = \max_{c\in C, e\in E} ca(c, e) \quad (7)$$

Patterns whose scores are less than a threshold (0.55 by default), together with the phrase pairs they connect, are removed from $\mathbb{E}$. For each phrase pair $PP \in \mathbb{E}$, the score of its pattern p is assigned as its weight ($v(PP) = s(p)$); while for each phrase pair $PP \in \mathbb{E}_h$, the weight $v(PP)$ is set to 1.0. As a result, we can merge $\mathbb{E}$ and $\mathbb{E}_h$ into a set of weighted phrase pairs, denoted as $\mathbb{E}_w$. Section 3.4 will build the boosted causal embeddings on it.

3.4 Weighted Max-Matching for Boosted Causal Embeddings

Now that we have obtained the weighted phrase pair set: $\mathbb{E}_w = \{(S_i, O_i)|1 \leq i \leq |\mathbb{E}_w|\}$, the task is to construct boosted causal embeddings, $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{e}}$ from it. This task looks similar to the one solved by [Sharp *et al.*, 2016] and [Xie and Mu, 2019], except that all the phrase pairs have their associated weights. Therefore, we propose a variant of *Max-Matching* algorithm, called *Weighted Max-Matching*, to take the weight information into consideration.

For a given word pair (w_1, w_2), the boosted causal score $\tilde{cs}(w_1, w_2)$ is defined as the inner product between the boosted cause embedding of w_1 and the boosted effect embedding of w_2:

$$\tilde{cs}(w_1, w_2) = \tilde{\mathbf{c}}(w_1)^\top \cdot \tilde{\mathbf{e}}(w_2) \quad (8)$$

It is different from the causal interaction score defined by Equation (1) in that it uses different embeddings $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{e}}$.

Similar to the Max-Matching model, its weighted variant also treats the task of learning causal embeddings as a multi-instance learning problem [Maron and Lozano-Pérez, 1998]. A phrase pair (C, E) can be mapped to a bag of word pairs $\{(c, e)|c \in C, e \in E\}$. It assumes that the bag contains at least one positive word pair if the phrase pair is positive, and all the word pairs are negative if the phrase pair is negative. That is, for a given positive phrase pair (C, E), it is expected that we can find a word pair (c, e) between C and E such that the cause word $c \in C$ has a large boosted causal score with the effect word $e \in E$.

The boosted causal score of a phrase pair (C, E) is defined as the maximum causal interaction score among all the word pairs between C and E:

$$\tilde{cs}(C, E) = \max_{c \in C, e \in E} \tilde{cs}(c, e) \tag{9}$$

In other words, only the word pair with the highest causal interaction score is selected as the positive word pair, which serves as the representative of the positive phrase pair.

The boosted variant has adopted the same negative sampling strategy as *Max-Matching* algorithm. Let $\mathbb{N}$ denote the set of all the sampled negative phrase pairs. Then, the loss J to minimize is calculated by summing all the cross-entropy losses of the training examples:

$$\begin{aligned} J = & \sum_{(C,E) \in \mathbb{E}_w} v((C, E)) \cdot \log(\sigma(\tilde{cs}(C, E)) \\ & + \sum_{(C,E) \in \mathbb{N}} \sum_{c \in C, e \in E} \log(1 - \sigma(\tilde{cs}(c, e)) \end{aligned} \tag{10}$$

where $v((C, E))$ is the weight of the phrase pair (C, E).

4 Evaluation on English Corpus

To make an evaluation on English, we make use of a high-precision corpus $\mathbb{E}_h$ of 815,233 cause-effect phrase pairs which was extracted with a set of 13 hand-crafted rules from Gigaword and Simple English Wikipedia. Both the rules and the corpus are taken from [Sharp *et al.*, 2016][3]. The Max-Matching algoirthm [Xie and Mu, 2019] is used to build the initial causal embeddings[4] $\mathbf{c}$ and $\mathbf{e}$ because of its superior performance.

We compare our causal embedding method against several state-of-the-arts:

- *vEmbed*: The vanilla word embeddings trained on raw text corpus by the skip-gram algorithm [Mikolov *et al.*, 2013] with a sliding window of 5;
- *cEmbed*: The cEmbed method [Sharp *et al.*, 2016] treats the effect phrase as the context of the cause, and uses a variant of Skip-Gram to train the causal embeddings;
- *cEmbedBi*: The bidrectional model [Sharp *et al.*, 2016] trains a second model treating causes as context, and ranks word pairs by averaging it and *cEmbed*;
- *cEmbedBiNoise*: The noise-aware bidirectional model [Sharp *et al.*, 2016] improves *cEmbedBi* by weighting word pairs with their likelihoods of being truly causal;
- *Attentive-Mathing*: It assumes that there are close interactions between a cause word and the effect phrase, and also between an effect word and the cause phrase [Xie and Mu, 2019];
- *Max-Matching*: It is built upon the assumption that there is at least one word pair carrying the causality information of positive phrase pair [Xie and Mu, 2019];

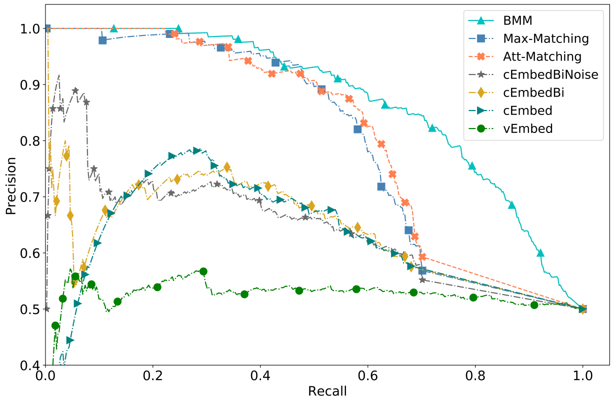


Figure 3: Precision-recall curves of the compared models to rank causal word pairs above non-causal pairs

- *BoostedMaxMatching* (*BMM* in short): The proposed method in this paper, running on the corpus of English Wikipedia[5].

4.1 Causal Word Pair Ranking

Our method together with several state-of-the-arts are evaluated on an external set of word pairs drawn from the SemEval 2010 Task 8[6] [Hendrickx *et al.*, 2010], which is originally a multi-way classification of semantic relations between nominals. This external set contains 1730 nominal pairs in total, 865 of which are from the Cause-Effect relation and an equal number of which are from the other eight relations.

Figure 3 shows the precision-recall (PR) curves for these embedding methods. Two facts can be observed from it:

- First, our *BMM* method has achieved much higher coverage for it has covered almost all the test word pairs. However, the curves of all the other causal embedding models become straight at tail, because about 30% of test word pairs have at least one word missing from their training examples, which means that they have only 70% coverage of the test word pairs;
- Second, our *BMM* method is much more precise than *cEmbed*-family methods (inclusive of *cEmbed*, *cEmbedBi* and *cEmbedBiNoise*). As for *Attentive-Matching* and *Max-Matching* methods, their precisions both have a sharp drop after the recalls exceed 0.5, while the curve of *BMM* runs more smoothly, even on the range of high recall values.

As the result of the combination of the two observations, the PR curve of *BMM* runs above all the others, at almost all the positions, resulting in the highest Area-Under-PRCurve (AUC) value as shown in Table 1.

4.2 Top-Ranked English Patterns

Besides the boosted causal embeddings, our method outputs potential patterns with the calculated causal strengths

[3] http://clulab.cs.arizona.edu/data/emnlp2016-causal/

[4] http://www.ke.fudan.edu.cn/data/causal/en_causal_embeds.zip

[5] http://dumps.wikimedia.org/enwiki/20181120/enwiki-20181120-pages-articles-multistream.xml.bz2

[6] http://www.kozareva.com/downloads.html

vE	cE	cEB	cEBN	AM	MM	BMM
0.52	0.62	0.64	0.64	0.81	0.80	**0.87**

Table 1: The AUC values. Method names are abbreviated: *vEmbed* (vE), *cEmbed* (cE), *cEmbedBi* (cEB), *cEmbedBiNoise* (cEBN), *Attentive-Matching* (AM) and *Max-Matching* (MM).

Rank 1-7	Rank 8-14	Rank 15-21
`spark`	`aggravate`	`stir`
`precipitate`	`ignite`	`pave`
`exacerbate`	`force`	`arouse`
`worsen`	`result_from_inv`	`reduce`
`provoke`	`inflict`	`fuel`
`stem`	`prompt`	`alleviate`
`trigger`	`increase_with_inv`	`contribute_to`

Table 2: The top-21 English verb-mediated patterns in causal strengths. The patterns ending with `_inv` are the inverted patterns of their prefixes.

by Equation (6), which can help detect and extract explicit (potential) causalities from text. The top-21 verb-mediated patterns are listed in Table 2. It can be observed that:

- Based on human examination, all the listed patterns can indicate causality in some degree, and coincide with human intuition about causality;
- Among the top-21 verbs, 18 of them are excitatory, and only 3 are inhibitory. We also find that the excitatory and inhibitory verbs ususally have high ranks, which is related to the work of [Hashimoto *et al.*, 2012];
- Interestingly, an inverted pattern `result_from_inv` is top-ranked, indicating that `result_from` is a effect-to-cause pattern. Such information is conducive to improving the accuracy of causality extraction.

5 Evaluation on Chinese Corpus

To make evaluation on Chinese, we use the `Sogou`[7] set of 517,746 high-precision causal phrase pairs which were extracted by Xie and Mu [2019] from the SogouCS[8] news corpus [Wang *et al.*, 2008]. The causal embeddings[9] constructed by the Max-Matching algorithm [Xie and Mu, 2019] are used to generate the seed set of causal word pairs.

5.1 Causal Word Pair Identification

To evaluate the boosted causal embeddings on the task of identifying causal word pairs from causal phrase pairs, we use an external set of 355 phrase pairs[10] extracted from a Chinese encyclopedia corpus, where each phrase pair was annotated with the causal word pair that indicates the causality between the phrases[Xie and Mu, 2019].

Given a cause-effect phrase pair $pp = (C, E)$ where $C = c_1c_2 \dots c_m$ and $E = e_1e_2 \dots e_n$, all the word pairs between C and E get ranked according to their interaction scores $cs(\cdot, \cdot)$ in Equation 1. For a word pair $wp = (c, e)$ where $c \in C$ and $e \in E$, let $r(wp, pp)$ denote the rank of word pair wp with respect to the phrase pair pp. If the 1st ranked word pair is the same as the annotated causal word pair, then we say that the causal embedding has correctly identified the causal word pair. The accuracy of a causal embedding model on the test dataset is defined as the percentage of the correctly identified causal word pairs. The mean reciprocal rank (MRR) is calculated as:

$$MRR = \frac{1}{|D|} \sum_{pp \in D} \frac{1}{r(ann(pp), pp)} \tag{11}$$

where $ann(pp)$ denotes the annotated word pair for the phrase pair $pp \in D$.

Model	Accuracy	MRR
vEmbed	8.4%	0.272
cEmbed	18.9%	0.345
cEmbedBi	19.2%	0.344
cEmbedBiNoise	18.3%	0.334
Max-Matching	42.9%	0.586
Attentive-Matching	42.1%	0.572
BMM	**44.3%**	**0.601**

Table 3: Quantitative performance on Baidu test data of the causal embeddings trained from Sogou corpus

Rank 1-7	Rank 8-14	Rank 15-21
致使(cause)	激发(arouse)	受到(get)
加剧(aggravate)	削弱(alleviate)	凸显(highlight)
迫使(force)	备受(-)	有助于(contribute_to)
遭到(suffer_from)	打击(hit)	避免(avoid)
致(incur)	带来(bring_about)	促成(facilitate)
令(make)	源于_inv(stem_from_inv)	赔偿(compensate_for)
掩盖(hide)	伴随(follow_with)	产生(produce)

Table 4: The top-21 Chinese verb-mediated patterns in causal strengths. The patterns ending with `_inv` are the inverted patterns of the prefixes.

The accuracies and MRRs are reported in Table 3. It can be seen that our Boosted Max-Matching model has achieved the highest performance on both accuracy and MRR, among all the competitors. In addition, the top-21 patterns ranked by their causal strengths are listed in Table 4, where observations can be made similar to Section 4.2.

6 Conclusion

In this paper, we propose a novel method to boost causal embeddings by exploring potential verb-mediated causal patterns. It first learns the pattern weights that indicate their causal strengths, and then boosts causal embeddings by fusing the existing strong causal phrase pairs and the extactions from (possibly weak) verb-mediated patterns. Experimental results have shown that the resulting causal embeddings have achieved high precision and recall on both Chinese and English. In addition, the top-ranked patterns, as byproducts, coincide with human intuition about causality.

Acknowledgements

This work is supported by National Key Research and Development Program of China (No.2018YFB1005100). We are grateful to anonymous reviewers for their valuable comments.

[7] http://www.ke.fudan.edu.cn/data/causal/sg_hp_extractions.txt

[8] http://www.sogou.com/labs/resource/cs.php

[9] http://www.ke.fudan.edu.cn/data/causal/sg_causal_embeds.zip

[10] http://www.ke.fudan.edu.cn/data/causal/bk_eval.txt

References

[Boros *et al.*, 2014] Emanuela Boros, Romaric Besançon, Olivier Ferret, and Brigitte Grau. Event role extraction using domain-relevant word representations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1852–1857, 2014.

[Chang and Choi, 2006] Du-Seong Chang and Key-Sun Choi. Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. *Information processing & management*, 42(3):662–678, 2006.

[Collobert *et al.*, 2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[Girju and Moldovan, 2002] Roxana Girju and Dan Moldovan. Mining answers for causation questions. In *AAAI symposium on mining answers from texts and knowledge bases*, 2002.

[Girju, 2003] Roxana Girju. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83. Association for Computational Linguistics, 2003.

[Hashimoto *et al.*, 2012] Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 619–630. Association for Computational Linguistics, 2012.

[Hashimoto *et al.*, 2014] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *ACL (1)*, pages 987–997, 2014.

[Hashimoto *et al.*, 2015a] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. Generating event causality hypotheses through semantic relations. In *AAAI*, pages 2396–2403, 2015.

[Hashimoto *et al.*, 2015b] Kazuma Hashimoto, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka. Task-oriented learning of word embeddings for semantic relation classification. *arXiv preprint arXiv:1503.00095*, 2015.

[Hendrickx *et al.*, 2010] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[Kaplan and Berry-Rogghe, 1991] Randy M Kaplan and Genevieve Berry-Rogghe. Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition*, 3(3):317–337, 1991.

[Li *et al.*, 2017] Luyang Li, Bing Qin, and Ting Liu. Contradiction detection with contradiction-specific word embedding. *Algorithms*, 10(2):59, 2017.

[Maron and Lozano-Pérez, 1998] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Nguyen and Grishman, 2014] Thien Huu Nguyen and Ralph Grishman. Employing word representations and regularization for domain adaptation of relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 68–74, 2014.

[Oh *et al.*, 2013] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. Why-question answering using intra-and intersentential causal relations. In *ACL (1)*, pages 1733–1743, 2013.

[Radinsky *et al.*, 2012] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918. ACM, 2012.

[Sharp *et al.*, 2016] Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 138–148. Association for Computational Linguistics, 2016.

[Tang *et al.*, 2014] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1555–1565, 2014.

[Wang *et al.*, 2008] Canhui Wang, Min Zhang, Shaoping Ma, and Liyun Ru. Automatic online news issue construction in web environment. In *Proceedings of the 17th international conference on World Wide Web*, pages 457–466. ACM, 2008.

[Xie and Mu, 2019] Zhipeng Xie and Feiteng Mu. Distributed representation of words in cause and effect spaces. In *AAAI*, 2019.