The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

Distributed Representation of Words in Cause and Effect Spaces

Zhipeng Xie, Feiteng Mu
School of Computer Science
Shanghai Key Laboratory of Data Science
Fudan University, Shanghai, China
xiezp@fudan.edu.cn

Abstract

This paper focuses on building up distributed representation of words in cause and effect spaces, a task-specific word embedding technique for causality. The causal embedding model is trained on a large set of cause-effect phrase pairs extracted from raw text corpus via a set of high-precision causal patterns. Three strategies are proposed to transfer the positive or negative labels from the level of phrase pairs to the level of word pairs, leading to three causal embedding models (Pairwise-Matching, Max-Matching, and Attentive-Matching) correspondingly. Experimental results have shown that Max-Matching and Attentive-Matching models significantly outperform several state-of-the-art competitors by a large margin on both English and Chinese corpora.

Introduction

Causality, also referred to as causation, is a fundamental concept in human thinking and reasoning (Stukker, Sanders, and Verhagen 2008). It indicates a special semantic relation between one process (the cause) with another process or state (the effect), where the cause is partially responsible for the effect, and the effect is partially dependent on the cause. Causality is commonly expressed, explicitly or implicitly, in text of most natural languages, which is of great value and has been exploited for various applications such as why-question answering (Oh et al. 2013), event prediction (Radinsky, Davidovich, and Markovitch 2012), and future scenario generation (Hashimoto et al. 2014).

One key to the success of these downstream applications lies in the automatic construction of a large causality base. Many methods have been proposed to mine causalities from text corpora, which can be categorized into classes. The oldest approaches used 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. New 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.

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 (Levy and Goldberg 2014b). Although such a general-purpose word embeddings is helpful for various NLP tasks, the acquisition of generality is often at the cost of losing specificity to a certain degree. Some work has been devoted to enhance word embeddings for specific tasks. Task-specific word embeddings capture task-specific word similarity. For example, if the task is about POS tagging, two nouns "cat" and "man" might be considered similar by the model (Faruqui et al. 2016). 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.

With vanilla word embedding, the underlying rationale of its wide applicability is the inference rule based on similarity, which states that if A is similar to B, and A has some property or has a semantic relation R with C, then it can be hypothesized that B also has the property or has the relation R with C. Such an inference rule may work for some properties and semantic relations, but definitely not applicable to all of them, such as causal relation. We conjecture that it explains why there is little research work of embedding-driven deep learning in causality-related tasks. The reason that the similarity-based inference rule is not applicable to causality exists possibly in the fact that causality is very sensitive to the small variation of semantics. For example, "*salmonella*" and "*bacillus acidi lactici*" have high similarity in their vanilla word embeddings because they are both bacteria. However, "*salmonella*" is the cause of some diseases, while "*bacillus acidi lactici*" is non-pathogenic. Such a high sensitivity results in the inapplicability of vanilla word embedding to causality-related tasks, and at the same time, necessitates the development of causality-specific word embedding

techniques.

In this paper, we continue the line of task-specific word embedding research, and devote to causal word embedding. We develop three models to learn causal embeddings from the corpus of cause-effect phrase pairs extracted from a text corpus , and present quantitative evaluation results on both English corpus and Chinese corpus. In particular, we make the following contributions in this paper:

- We propose three methods (Pairwise Matching, Max Matching, and Attentive Matching) for building causal embeddings from a corpus of cause-effect phrase pairs, by transferring causal relationship from phrase-pair level to word-pair level. Performance evaluation has shown that Max Matching and Attentive Matching models perform much better than several state-of-the-art competitors, by a large margin, on both English and Chinese corpora. Here, Max Matching model 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, and Attentive Matching considers the causal relations between words and phrases.
- For Chinese corpus, we hand-craft a set of high-precision causal patterns and extract nearly 1.5 million of cause-effect phrase pairs from raw corpus. The quality of those extractions is satisfactory, with precision approaching 90%.
- We define a new task of identifying causal word pairs from cause-effect phrase pairs, and manually annotate a dataset of cause-effect phrase pairs accordingly. This annotated dataset can be used for evaluating the performance of causal embeddings, to promote the research on causal embeddings.

Related Work

To the best of our knowledge, there are two research papers (Sharp et al. 2016; Zhao et al. 2017) that are closely related to this paper.

Sharp et al. (2016) generated a causality-specific embeddings and demonstrated that these dedicated embeddings is 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 an 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. This assumption is far from reality and introduces too much noise. The cEmbed-family models perform undistinguished in our performance evaluation.

Zhao et al. (2017) proposed to construct an abstract causality network from the specific one, and then to learn network embedding by a dual cause-effect transition model, in order to achieve generalization ability. However, their approach has two shortcomings. Firstly, the construction of abstract causality network depends on the availability of knowledge resources, which limits its applicability. Secondly, the rationality of generalizing a noun to its hypernym need further clarification, because it is doubtful whether such a generalization will preserve the causal-effect relationship.

Causal Embedding Models

Given a corpus of N cause-effect phrase pairs $\mathbb{D} = \{(C_i, E_i)|1 \leq i \leq N\}$, we use $V^c = \{c|\exists C_i \text{ such that } c \in C_i\}$ to denote the vocabulary of cause words, and $V^e = \{e|\exists E_i \text{ such that } e \in E_i\}$ to denote the vocabulary of effect words. For a cause word $c \in V^c$, its cause embedding is a vector $\mathbf{c}(c)$, and the effect embedding of effect word $e \in V^e$ is $\mathbf{e}(e)$, both of size d (200 by default).

In this paper, we solve the causal embedding problem in a classification framework. Each phrase pair $(C, E) \in \mathbb{D}$ is thought of as a positive example, because C and E are causally related. Negative example (C, E) can be generated by choosing randomly one phrase C from all the cause phrases in $\mathbb{D}$ and one phrase E from all the effect phrases. As a result, we obtain a dataset of positive examples and negative examples for classification, where examples are phrase pairs.

Let $t(C, E)$ denote the class label of phrase pair (C, E). The class label is assigned clearly at the phrase-pair level. In order to learn causal embeddings for words, we have to transfer the class information from phrase-pair level to word-pair level. This problem can be solved easily for negative examples: for a given negative phrase pair (C, E), we assume that all word pairs (c, e), where $c \in C$ and $e \in E$, are negative (i.e., not causally related).

However, for positive examples, the situation is more complex, because we do not know which word pair dominates the causality between the phrase pair. Next, we propose three strategies to deal with positive phrase pairs, leading to three different causal embedding models.

Model 1: Pairwise-Matching

The first strategy assumes that for each positive phrase pair, all the word pairs between the cause phrase and the effect phrase are treated as positive (i.e., being causally related). With this strategy, for a phrase pair (C, E), no matter positive or negative, all the word pairs between C and E will inherit the class label from the phrase pair. The corresponding method works in a straightforward way as follows.

As illustrated in Figure 1, for a given word pair (c, e), Pairwise-Matching model first calculates a causal interaction score $cs(c, e)$ as the inner product between the cause embedding of c and the effect embedding of e:

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

and then uses sigmoid function to transform the score into a probability as the prediction of the class label:

$$p(c, e) = \sigma(cs(c, e)) = \frac{\exp(cs(c, e))}{1 + \exp(cs(c, e))} \tag{2}$$

Given a training set $\mathbb{D}$ of phrase pairs, we choose the cross-entropy loss function as the objective J_{PM} to minimize:

$$J_{PM} = \sum_{(C,E)\in\mathbb{D}} \sum_{c\in C, e\in E} -t(C, E) \log p(c, e) - (1 - t(C, E)) \log(1 - p(c, e)) \tag{3}$$

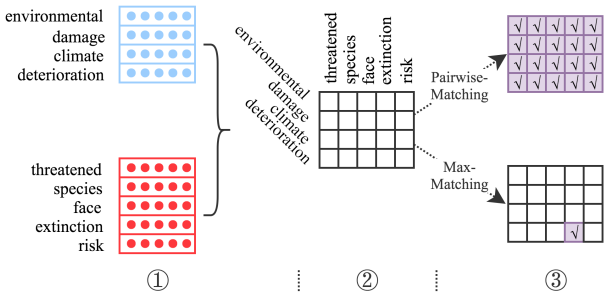

Figure 1: The architecture of Pairwise-Matching and Max-Matching models. Here, ① is "the cause and effect embeddings", ② is "the causal interaction scores between words", the upper branch of ③ corresponds to Pairwise-Matching model, and the lower branch of ③ corresponds to Max-Matching one.

The focal loss (used in the following two models) was also tried, but we found that it is very unstable for pairwise-matching model. We conjecture that the reason is that focal loss is sensitive to noise, and the assumption of pairwise-matching introduces too much noise.

This model is similar to cEmbed-family models (Sharp et al. 2016). Quantitative evaluation shows that they have similar performance . The difference exists in the negative sampling part: Pairwise-Matching samples negative examples at the phrase level, while cEmbed uses downsampling technique to sample negative examples at the word level.

Model 2: Max-Matching

The second strategy treats the task of learning causal embedding as a multi-instance learning problem (Maron and Lozano-Pérez 1998), where each phrase pair (C, E) can be mapped to a bag of word pairs $\{(c, e)|c \in C, e \in E\}$. We assume that the bag does not contain positive word pairs if the phrase pair is negative, and the bag contains at least one positive word pair if the phrase pair is positive. In other words, it requires that there is at least one word pair carrying the causality information of the positive example. That is, for a given cause-effect 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 causal interaction with the effect word $e \in E$.

The 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:

$$cs_{MM}(C, E) = \max_{c \in C, e \in E} cs(c, s) \tag{4}$$

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 (as illustrated in Figure 1.

The Max-Matching method also uses a sigmoid function to transform the causal score $cs_{MM}(C, E)$ into the predicted probability that the phrase pair (C, E) is positive or causally related, as follows:

$$p_{MM}(C, E) = \frac{\exp(cs_{MM}(C, E))}{1 + \exp(cs_{MM}(C, E))} \tag{5}$$

The Max-Matching method uses focal loss (Lin et al. 2017) as the loss function, which focuses training on a sparse set of hard examples and prevents the vast number of easy negatives from overwhelming the model during training. We adopt focal loss function here, because during learning causal embedding, there are a vast number of easy negatives, and focal loss is suitable for this situation, especially for the max-matching strategy. For a given predicted probability p and a target class label t, the focal loss $fl(p, t)$ is defined as:

$$fl(p, t) = \begin{cases} -\alpha(1-p)^\gamma \log(p) & \text{if } t = 1 \\ -(1-\alpha)p^\gamma \log(1-p) & \text{if } t = 0 \end{cases} \tag{6}$$

where α and γ are two hyperparameters, which are set to 0.8 and 2.0 by default.

The total focal loss J_{MM} for the training dataset $\mathbb{D}$ is calculated by summing all the focal losses of positive phrase pairs and the focal losses of all the negative word pairs:

$$\begin{aligned} J_{MM} = & \sum_{\substack{(C,E) \in \mathbb{D} \\ t(C,E)=1}} fl(p_{MM}(C, E), 1) \\ & + \sum_{\substack{(C,E) \in \mathbb{D} \\ t(C,E)=0}} \sum_{\substack{c \in C \\ e \in E}} fl(p(c, e), 0) \end{aligned} \tag{7}$$

where $p(c, e)$ is defined in Equation 2.

Model 3: Attentive-Matching

The third method considers the interaction between a cause word with the effect phrase and the interaction between an effect word with the cause phrase. It is expected that there is at least one cause word that has a close causal interaction with the effect phrase, and also at least one effect word that causally interacts with the cause phrase.

The architecture of Attentive-Matching model is illustrated in Figure 2 to implement this idea. We firstly define the attentive representation $\mathbf{C}^{att}$ of the cause phrase C with respect to the effect phrase E and the attentive representation $\mathbf{E}^{att}$ of the effect phrase E with respect to the cause phrase C, as follows:

$$\mathbf{C}^{att} = \sum_{i=1}^{m} a_i^C \mathbf{c}(c_i) \tag{8}$$

$$\mathbf{E}^{att} = \sum_{j=1}^{n} a_j^E \mathbf{e}(e_j) \tag{9}$$

where m and n are the lengths of C and E respectively, a_i^C is the attention weight of the cause word c_i on condition of the effect phrase E, and a_j^E is the attention weight of the effect word e_j on condition of the cause phrase C:

$$a_i^C = \frac{\sum_{j'=1}^{n} \exp(cs(c_i, e_{j'}))}{\sum_{i'=1}^{m} \sum_{j'=1}^{n} \exp\left(cs(c_{i'}, e_{j'})\right)} \tag{10}$$

$$a_j^E = \frac{\sum_{i'=1}^{m} \exp(cs(c_{i'}, e_j))}{\sum_{i'=1}^{m} \sum_{j'=1}^{n} \exp\left(cs(c_{i'}, e_{j'})\right)} \tag{11}$$

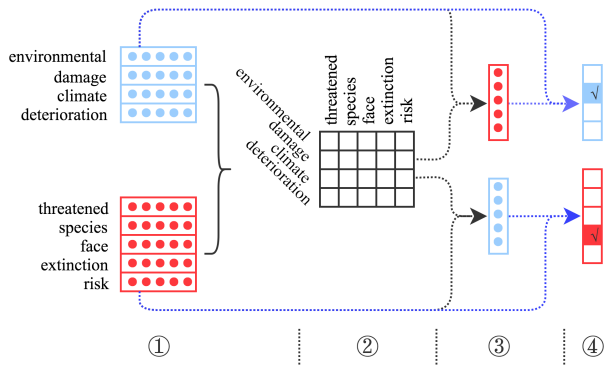

Figure 2: The architecture of Attentive-Matching model. Here, ① is "the cause and effect embeddings", ② is "the causal interaction scores between words", ③ is "the attentive representations of phrases", and ④ stands for "the causal interaction scores between words and phrases".

The causal interaction score between a cause word c_i and the effect phrase E is defined as the inner product of the cause embedding of c_i and the attentive representation of E with respect to C:

$$cs_{AM}(c_i, E) = \mathbf{c}(c_i)^\top \cdot \mathbf{E}^{att} \tag{12}$$

Similarly, the causal interaction score between the cause phrase C and an effect word e_j is calculated as:

$$cs_{AM}(C, e_j) = \mathbf{e}(e_j)^\top \cdot \mathbf{C}^{att} \tag{13}$$

Given a phrase pair (C, E), its causal score from cause to effect is defined as:

$$\overrightarrow{cs}_{AM}(C, E) = \max_{c_i \in C} cs_{AM}(c_i, E) \tag{14}$$

and its causal score from effect to cause is defined as:

$$\overleftarrow{cs}_{AM}(C, E) = \max_{e_j \in E} cs_{AM}(C, e_j) \tag{15}$$

Finally, both the scores are transformed into two probabilities $\overrightarrow{p}_{AM}(C, E)$ and $\overleftarrow{p}_{AM}(C, E)$ by sigmoid function, respectively. Therefore, for each positive phrase pair (C, E), there are two focal losses: one is from cause to effect $fl(\overrightarrow{p}_{AM}(C, E), 1)$, the other is from effect to cause $fl(\overleftarrow{p}_{AM}(C, E), 1)$.

The total focal loss J_{AM} is the summation of all the positive losses and all the negative losses:

$$J_{AM} = \sum_{\substack{(C,E)\in\mathbb{D} \\ t(C,E)=1}} fl(\overrightarrow{p}_{AM}(C, E), 1) + fl(\overleftarrow{p}_{AM}(C, E), 1) + \sum_{\substack{(C,E)\in\mathbb{D} \\ t(C,E)=0}} \sum_{\substack{c\in C \\ e\in E}} fl(p(c, e), 0) \tag{16}$$

Model Training

We use simple gradient descent algorithm to train our models, with learning rate of 0.005. Other related hyperparameters are listed as follows. The number of training epochs are set to 30, and the batch size is 256. The words whose frequencies are less than 8 are pruned. The cause embeddings and the effect embeddings have the same dimensionality of 200. The negative sampling rate is 10, which means that we samples 10 negative phrase pairs for each positive phrase pair.

Evaluation on English Corpus

To make an evaluation on English, we build our models on a corpus of 815,233 cause-effect phrase pairs which was extracted with a set of 13 rules from Gigaword and Simple English Wikipedia. Both the rules and the corpus are taken from (Sharp et al. 2016)[1].

Direct Evaluation: Ranking Word Pairs

The models are evaluated on an external set of word pairs drawn from the SemEval 2010 Task 8 (Hendrickx et al. 2010)[2], 865 of which were from the Cause-Effect relation and an equal number of which were randomly selected from the other eight relations. These pairs are then ranked and the goal is to rank the causal pairs above the others.

We compare our causal embedding models against several baseline and state-of-the-art models:

- *Look-up*: A word pair is ranked by the number of times that it appears in the extracted cause-effect phrase pairs.
- *vEmbed*: The vanilla word embeddings trained on raw text corpus with the skip-gram algorithm and 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 the variant of Skip-Gram implemented by Levy and Goldberg (2014a) to train the causal embeddings.
- *cEmbedBi*: The bidirectional embedding model (Sharp et al. 2016) trains a second embedding model by reversing the input, which treats the cause phrase as the context of the effect, and ranks word pairs by the average of the scores returned by the two unidirectional causal embedding models.
- *cEmbedBiNoise*: The noise-aware bidirectional model makes an improvement on *cEmbedBi* by weigh a word pair by the likelihood that they are truly causal, which is approximated by the pointwise mutual information (PMI).

Figure 3 shows the precision-recall (PR) curve for these models and baselines. The *Max-Matching* and *Attentive-Matching* models perform consistently much better than other existing models and baselines including cEmbedBiNoise, cEmbedBi and cEmbed, which is attributed to their more-reasonable assumptions. The *Pairwise-Matching* model performs similarly to the *cEmbed*-family models, because they all share the same assumption that all word pairs between a cause-effect phrase pair are causally related.

It is also noted that the curves of all causal embedding models become straight at tail, which is caused by the fact

[1] http://clulab.cs.arizona.edu/data/emnlp2016-causal/
[2] http://www.kozareva.com/downloads.html

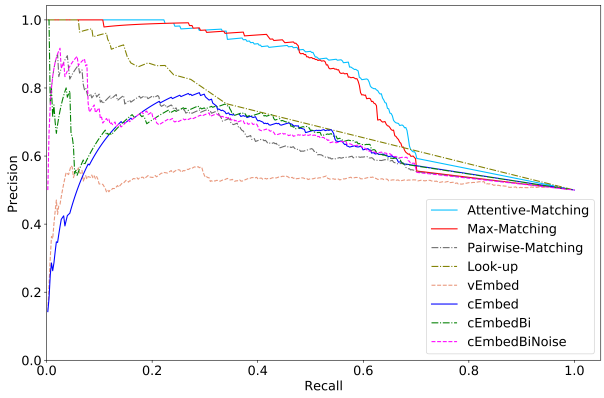

Figure 3: Precision-recall curve of the compared models to rank causal word pairs above non-causal pairs

Model	P@1
Random	16.4
CR	24.3
CR+vEmbed	34.6
CR+vEmbed+allCausalEmbed	37.9

Table 1: Performance on the Causal QA Task

that about 30% of test word pairs have at least one word missing from the training corpus.

Indirect Evaluation: Causal Question-Answering

Here, we would like to investigate whether the learned causal embeddings can help the downstream causal question-answering task. To do so, we conduct five-fold cross validation on a dataset of 3031 causal questions extracted from the Yahoo! Answers corpus (Sharp et al. 2016), where each question has at least four alternatives (candidate answers). All the alternatives are generated by the community, and one of them is voted as the top answer. Our task is to identify the top answer from the alternatives. With random guess, the P@1 is 16.4%, which means that each causal question has more than 5 candidate answers on average.

Here, we use a standard answer reranking architecture (Jansen, Surdeanu, and Clark 2014), where the candidate answers are initially ranked by a shallow information retrieval method called candidate retrieval (CR). For each embedding model, no matter vanilla or causal, four features are computed for each question-answer pair: the maximum, the minimum, the average cosine similarity between word pairs, and the overall similarity between the composite question and answer vectors, where the composition vector is calculated as the mean vector of all its words. Then, the SVM ranker (a SVM classifier adapted for ranking) is applied to re-rank the candidate answers, with the initial CR score and the embedding-related features.

The following facts can be observed from Table 1 which lists the precision-at-one scores (P@1):

- The shallow candidate retrieval method (CR) achieves the P@1 of only 24.31%, which indicates the difficulty of this causal QA task.
- With CR+vEmbed(word2vec), the P@1 is 34.6%, which illustrates the vanilla word embeddings are helpful.
- By augmenting the above configuration (CR+word2vec) with the derived features from all the causal embedding models, the P@1 gets increased to 37.9%, which demonstrates that causal embeddings are complementary to the vanilla word embeddings and make their own contribution in the downstream QA task.

We shall explore how to use CNN and RNN to build up the representations of questions and answers on the basis of vanilla embeddings and causal embeddings, as future work.

Evaluation on Chinese Corpus

Chinese Corpus construction

To build up causal embeddings for Chinese words, we have to collect a sufficiently large number of Chinese cause-effect phrase pairs from raw text corpus. It is too costly to be done by human annotation, so we make use of a few high-precision hand-crafted (causal) patterns instead.

Causality is pervasive in natural language. Part of them are marked by a variety of explicit causative devices (or called causal patterns) that can be categorized into two types: the causal verbs that mark causal relations within one clause, and the causal connectives that mark causal relations between clauses. However, these causal patterns usually signal different causal strengths. Since our emphasis here is put on the high precision of the extractions, we choose only a limited number of hand-crafted strong patterns that have sufficiently high precision, some of which is listed in Table 2. Additionally, in Chinese, some strong causal connectives can only indicate the existence of causalities, but can not determine the exact position of the cause and effect text spans. For example, the pattern "$\cdots$，因为 (because)$\cdots$" signaling the fact that there exists causality and the cause phrase follows the connective "因为 (because)" closely, but the position of effect phrase is ambiguous, which may appear before the connective or after the cause phrase. Therefore, we choose the patterns that not only signal strongly the existence of causality, but also be unanimouos in the location of cause phrase and effect phrase. The cause and effect phrases are extracted according to the dependency parses of causality sentences with a Chinese dependency parser PyLTP[3]. The extracted causal phrase pairs are further filtered by requiring that both the cause phrase and the effect phrase should have at least one content word (i.e., noun, verb or adjective).

We apply the above causal patterns on two raw Chinese corpora, the Baike corpus and the SogouCS corpus, where Baike is a 10GB data crawled from a Chinese encyclopedia website and SogouCS[4] (Wang et al. 2008) is the news data on the web, yielding to about 1.5 million of cause-effect phrase pairs in total. The statistics of the extractions are listed in Table 3.

To assess the quality of the extracted causality pairs, we randomly sample 500 cause-effect phrase pairs from each

[3]https://github.com/HIT-SCIR/pyltp

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

Pattern Types	Patterns
Causal Connectives	“因为<C>, 所以<E>” (“because <C>, thus <E>”); …… “之所以<E>, 是因为<C>” (“The reason <E> is because <C>”); …… “<C>, 所以<E>” (“<C>, so <E>”); ……
Causal verbs	导致 (reasult in), 造成 (bring about), 引起 (give rise to), 使得 (make), 使 (make), 引来 (lead to), 促使 (cause), 引发 (trigger), 招致 (incur)

Table 2: The causal patterns used to automatically extract cause-effect pairs. For connective patterns, <C> denotes the position of cause phrase, and <E> denotes the phrase position. For verb patterns, the cause phrase is alway on the left side of the verb, and the effect phrase on the right.

Corpus	Pattern Type	#Extracted CE Pairs
Baike	causal connectives	356,654
	causal verbs	619,526
Sogou	causal connectives	233,126
	causal verbs	284,620
Total.		**1,493,926**

Table 3: Statistics of the extracted CE Pairs

subset and let annotators judge whether these pairs express causality. The results are shown in Table 4. It can be seen that the extracted cause-effect pairs have satisfactory precision of nearly 90%, and the extractions by causal verbs have slightly higher precision than those by causal connectives.

Corpus	Pattern Type	# Samples	Precision
Baike	causal connectives	500	88.4%
	causal verbs	500	89.6%
Sogou	causal connectives	500	89.4%
	causal verbs	500	90.4%

Table 4: Quality of the extracted CE Pairs

Quantitative Evaluation

To assess our causal embedding models, we use a simple causality-related task: identifying the causal word pair that carries the most significant causal information from a cause-effect phrase pair. For example, the causal word pair “(缺陷 (flaw), 爆炸 (explode))” from the cause-effect phrase pair in Table 5. Identifying causal word pairs helpful in interpreting a cause-effect phrase pair, and can be used to extend the potential causality patterns and then to help identify implicit causality in text.

Cause Phrase	一枚 密封圈 的 缺陷 the flaw in a seal
Effect Phrase	飞机 在 起飞 后 爆炸 the airplane exploded after taking off

Table 5: A simple cause-effect phrase pair

To evaluate the performance of causal embedding models, we let two annotators to mark the causal word pairs from cause-effect phrase pairs, obtaining 715 pairs for which the two annotators got the same annotating results: 360 pairs for Baike corpus and 355 pairs for Sogou, where the overall inter-annotator agreement is near to 80%.

Model	**Accuracy**	**MRR**
Look-up	6.8%	0.184
vEmbed	12.7%	0.305
cEmbed	19.1%	0.338
cEmbedBi	23.5%	0.404
cEmbedBiNoise	24.1%	0.400
Pairwise-Matching	24.9%	0.424
Max-Matching	**58.6%**	**0.720**
Attentive-Matching	53.0%	0.674

Table 6: Quantitative performance on Sogou test data of the causal embeddings trained from Baike corpus

Model	**Accuracy**	**MRR**
Look-up	8.4%	0.190
vEmbed	8.4%	0.272
cEmbed	18.9%	0.345
cEmbedBi	19.2%	0.344
cEmbedBiNoise	18.3%	0.334
Pairwise-Matching	19.5%	0.348
Max-Matching	**42.9%**	**0.586**
Attentive-Matching	42.1%	0.572

Table 7: Quantitative performance on Baidu test data of the causal embeddings trained from Sogou corpus

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:

$$MRR = \frac{1}{|D|} \sum_{pp \in D} \frac{1}{r(ann(pp), pp)} \tag{17}$$

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

Cause Word	cEmbed	Pairwise-Matching	Attentive-Matching	Max-Matching
爆炸 (explosion)	爆炸 (explosion) 死亡 (death) 遇难 (killed) 财产 (asset) 伤亡 (casualties)	爆炸 (explosion) 里氏 (the Richter scale) 相撞 (collision) 坠毁 (crash) 交火 (Fire)	大火 (big fire) 烧伤 (burns) 水污染 (water pollution) 罹难 (die) 重伤 (serious injury)	大火 (big fire) 死伤 (casualties) 伤 (wound) 受伤 (injured) 重伤 (serious injury)
战争 (warfare)	战争 (warfare) 遭到 (suffer) 军事 (military) 战斗 (battle) 和平 (peace)	出兵 (dispatch troops) 战争 (warfare) 阿富汗 (afghanistan) 歼灭 (annihilate) 法西斯 (fascist)	战乱 (War) 动乱 (disturbance) 毁灭 (destroy) 流离失所 (homeless) 灭亡 (perish)	动乱 (disturbance) 苦难 (suffering) 流离失所 (homeless) 灾祸 (disaster) 反抗 (resist)
暴雨 (rainstorm)	倒塌 (collapse) 水位 (water level) 山体 (massif) 泥石流 (mud-rock flow) 滑坡 (landslide)	暴雨 (rainstorm) 降雨 (rainfall) 大暴雨 (heavy rain) 洪涝 (waterlogging) 洪灾 (flooding)	洪灾 (floods) 淹 (inundate) 内涝 (waterlogging) 泥石流 (mud-rock flow) 洪水 (waterflood)	淹 (inundate) 洪水 (waterflood) 洪灾 (floods) 内涝 (waterlogging) 受灾 (victim)

Table 8: Examples of top-5 effect words retrieved for the given cause words. For the cause word "爆炸 (explosion)", it can be seen that the top-5 effect words retrieved by the Attentive-Matching model and the Max-Matching model are all correct. However, the words retrieved by cEmbed or Pairwise-Matching model are not the case. For example, for the cause word "爆炸 (explosion)", "财产 (asset)" is retrieved by vEmbed and "里氏 (the Richter scale)" is retrieved by Pairwise-Matching model.

Based on the two Chinese corpora and their respective annotated causal word pairs, we train the causal embeddings on one corpus, and then evaluate their performance on the annotated word pairs of the other corpus. The accuracies and MRRs are reported in Table 6 and Table 7. It can be seen that *cEmbed*-family models and *Pairwise-Matching* have achieved similar accuracy and MRR values, which are far below those of Attentive-Matching model and Max-Matching model. *Max-Matching* is the best model, no matter in Accuracy or in MRR value. The reason is because both the *Pairwise-Matching* and *cEmbed*-family models assume all word pair between a cause-effect phrase pair are causally related, yielding too much noise of false positives which will definitely do harm to the process of learning causal embeddings. With *Max-Matching*, only one word pair with the highest interaction score is used as positive, all the other word pairs do not take part in the training, which has effectively controlled the noise. As to *Attentive-Matching*, it can be thought of as a soft version of Max-Matching. The false negatives are controlled in some degree, because most words (no matter cause words or effect words) will receive small attentive weights.

The causal embeddings trained from Baike corpus have better performance than those trained from Sogou corpus, partly because the Baike corpus of nearly 1 million pairs is larger than Sogou corpus of only half million pairs (see Table 3) and the Baike corpus can provide a higher coverage.

Qualitative Observation

To give more insight into the quality of different causal embedding models, we train causal embedding models on the merged Baidu and Sogou corpus, and show the retrieved top-5 effect words to a given set of cause words in Table 8. Due to the space limitation, the results from effect to cause are not shown here.

It can be observed that all the top-5 words retrieved by *Max-Matching* and *Attentive-Matching* are always causally related to the queries. The quality of the words retrieved by *Pairwise-Mathing* and *cEmbed* is much lower.

As a conclusion, the qualitative observation, in accordance with the quantitative evaluation, manifests the superiority of *Max-Matching* and *Attentive-Matching* in the task of learning causal embeddings.

Conclusion and Future Work

This paper proposes three models for learning causal embeddings, with the following contributions:

- We propose three strategies to transfer the class label from the level of phrase pairs to the level of word pairs, leading to three causal embedding models.
- We construct automatically a large Chinese corpus of cause-effect phrase pairs by a few hand-crafted high-precision causal patterns, and annotate manually the causal word pairs that carry causal information between cause phrases and their effect phrases. Such an annotated dataset is helpful for evaluating causal embedding models.
- Quantitative evaluation demonstrates that *Attentive-Matching* and *Max-Matching* models outperform the existing models substantially, no matter on English or Chinese corpus.

In the future, we would like to improve the coverage of learned causal embeddings by making use of numerous weak causal patterns and explore the potential applicability of the learned causal embeddings in the downstream application such as event prediction and scenario generation.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No.61732004), National Key Research and

Development Program of China (No.2018YFC0830901) and National High-Tech R&D Program of China (863 Program) (No. 2015AA015404). We are grateful to the anonymous reviewers for their valuable comments.

References

Chang, D.-S., and Choi, K.-S. 2006. Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. *Information processing & management* 42(3):662–678.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Faruqui, M.; Tsvetkov, Y.; Rastogi, P.; and Dyer, C. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, 30–35.

Girju, R., and Moldovan, D. 2002. Mining answers for causation questions. In *AAAI symposium on mining answers from texts and knowledge bases*.

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

Hashimoto, C.; Torisawa, K.; Kloetzer, J.; Sano, M.; Varga, I.; Oh, J.-H.; and Kidawara, Y. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *ACL (1)*, 987–997.

Hashimoto, C.; Torisawa, K.; Kloetzer, J.; and Oh, J.-H. 2015a. Generating event causality hypotheses through semantic relations. In *AAAI*, 2396–2403.

Hashimoto, K.; Stenetorp, P.; Miwa, M.; and Tsuruoka, Y. 2015b. Task-oriented learning of word embeddings for semantic relation classification. *arXiv preprint arXiv:1503.00095*.

Hendrickx, I.; Kim, S. N.; Kozareva, Z.; Nakov, P.; Ó Séaghdha, D.; Padó, S.; Pennacchiotti, M.; Romano, L.; and Szpakowicz, S. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 33–38. Uppsala, Sweden: Association for Computational Linguistics.

Jansen, P.; Surdeanu, M.; and Clark, P. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 977–986.

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

Levy, O., and Goldberg, Y. 2014a. Dependency-based word embeddings. In *ACL (2)*, 302–308.

Levy, O., and Goldberg, Y. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, 2177–2185.

Li, L.; Qin, B.; and Liu, T. 2017. Contradiction detection with contradiction-specific word embedding. *Algorithms* 10(2):59.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2999–3007.

Maron, O., and Lozano-Pérez, T. 1998. A framework for multiple-instance learning. In *Advances in neural information processing systems*, 570–576.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Oh, J.-H.; Torisawa, K.; Hashimoto, C.; Sano, M.; De Saeger, S.; and Ohtake, K. 2013. Why-question answering using intra-and inter-sentential causal relations. In *ACL (1)*, 1733–1743.

Radinsky, K.; Davidovich, S.; and Markovitch, S. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, 909–918. ACM.

Sharp, R.; Surdeanu, M.; Jansen, P.; Clark, P.; and Hammond, M. 2016. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 138–148. Association for Computational Linguistics.

Stukker, N.; Sanders, T.; and Verhagen, A. 2008. Causality in verbs and in discourse connectives: Converging evidence of cross-level parallels in dutch linguistic categorization. *Journal of Pragmatics* 40(7):1296–1322.

Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; and Qin, B. 2014. 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, 1555–1565.

Wang, C.; Zhang, M.; Ma, S.; and Ru, L. 2008. Automatic online news issue construction in web environment. In *Proceedings of the 17th international conference on World Wide Web*, 457–466. ACM.

Zhao, S.; Wang, Q.; Massung, S.; Qin, B.; Liu, T.; Wang, B.; and Zhai, C. 2017. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 335–344. ACM.