Received: 20 April 2021 | Revised: 17 June 2021 | Accepted: 22 July 2021

DOI: 10.1002/cpe.6572

RESEARCH ARTICLE

WILEY

Chinese causal event extraction using causality-associated graph neural network

Jianqi Gao[1] | **Xiangfeng Luo**[1,2] | **Hao Wang**[1,2]

[1]School of Computer Engineering and Science, Shanghai University, Shanghai, China

[2]Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China

Correspondence
Xiangfeng Luo and Hao Wang, School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China.
Emails: luoxf@shu.edu.cn; wang-hao@shu.edu.cn

Funding information
Ministry of Industry and Information Technology project of the Intelligent Ship Situation Awareness System, Grant/Award Number: MC-201920-X01; National Natural Science Foundation of China, Grant/Award Numbers: 61991415, 91746203; Shanghai Science and Technology Young Talents Sailing Program in 2021, Grant/Award Number: 21YF1413900

Abstract

Causal event extraction (CEE) aims to identify and extract cause-effect event pairs from texts, which is a fundamental task in natural language processing. Recent research treat CEE as a sequence labeling problem. However, the linguistic complexity and ambiguity of textual description results in the low accuracy of extractors. To address the above issues, considering the prior knowledge like the causal network constructed based on the causal indicators, which can represent information transition between cause and effect, may helpful for CEE. In this article, we propose causality-associated graph neural network to incorporate in-domain knowledge by taking important causal words into account. External causal knowledge is modeled as causal associated graph (CAG). Then we use graph neural networks (GNN) to capture the complex relationship of intraevent mentions and interevent causality in a sentence based on the relationship obtained from CAG. Finally, sentence sequence and prior causal knowledge of GNN embedding are fed into multiscaled convolution and bidirectional long short-term memory networks. Experimental results on two datasets show that our method outperforms the state-of-the-art baseline.

KEYWORDS

causal event extraction, causality-associated graph neural network, cause-effect event pairs, sequence labeling

1 | INTRODUCTION

Causal event extraction (CEE) is a joint extraction task of events and causality. It is a fundamental task for event logic graph construction that automatically extracts cause/effect events from plain text, and determines whether there exists a causal relationship between the cause event and the effect event. CEE is of great value for various intelligent applications, such as event prediction, recommendation system and question answering.

As shown in Table 1, there are mainly two types of causal relations in CEE, including explicit relation and implicit relation. Explicit relation usually contains causal indicators such as lead to, due to, because of, and so forth, while implicit causality does not. For example, "*The price of raw material rises sharply, leading to a sharp increase in feed cost*." In this sentence, the cause is "*the price of raw material rises*," the effect is "*a sharp increase in feed cost*." Understanding the event causality in a sentence is essential to many artificial intelligence applications. However, there are still exist two problems in previous CEE methods: (1) The accuracy of phrase-level event extraction is relatively low due to the ambiguous description of events; (2) The long-distance dependence make it difficult to determine implicit causality. These difficulties prevent conventional methods to accurately extract causal events just rely on limited training data. The main reason is the lack of necessary background knowledge. For example, "*John killed someone, and was sent to prison*" and "*As Hudson murdered Andrew, he was sent to prison.*" The news text contains a large number of similar causal patterns, which can assist the model to improve the accuracy of CEE.

TABLE 1 Example of causal event extraction

Forms	Sentence	Causal-effect
Explicit causal sentences	The price of raw materials rise sharply, **leading to** a sharp rise in feed cost.	"the price of raw materials rise" → "sharp rise in feed cost"
Implicit causal sentence	The Country adopts macro-control, and the price of pork begin to fall.	"macro-control" → "the price of pork begin to fall"

Existing methods can be roughly divided into rule-based methods, statistical methods and deep learning methods. Rule-based methods[1-4] extract causal events from text using template matching approach, requiring extensive manual efforts to construct patterns such as lexical patterns, syntactic patterns, and semantic patterns. However, it is impossible to enumerate all templates due to the complexity of natural language. Moreover, pattern matching has relatively poor ability of generalization, resulting in low recall of CEE. Although machine learning methods[5-8] partially solve above problems based on rich features, it relies on sophisticated feature engineering and consumes external time and manual efforts.

In recent years, deep learning methods relying on neural network models have achieved great success in natural language processing. These methods extract high-dimensional features from data in a popular end-to-end manner. Fu et al.[9] transform CEE into sequence labeling. Convolutional neural network (CNN)[10,11] models that emphasized n-gram features have proved to be suitable for relation extraction. Besides, Dasgupta and Dunietz et al.[12,13] apply long short-term memory (LSTM) to CEE to capture the long-term dependence of sentences. Furthermore, Ma and Jin et al.[14,15] combine CNN and LSTM to take advantage of both architectures. Though neural networks have achieved state-of-the-art performance, CEE is still remains a hard problem due to its complexity and ambiguity. Besides, insufficient training data make the deep model easy to overfit and drastic fluctuations.

To overcome the above limitations, considering that we can find a large number of sentences containing causal relations based on causal indicators, which can be used to construct causal graph, may benefit both event and causality extraction. We propose a novel neural method for CEE with the associated graph, the model incorporate causality-associated graph (CAG) as prior knowledge into the graph neural network (GNN) model to assist CEE. GNN[16,17] are designed to represent the graph. They have been widely used in natural language processing.[18,19] For CEE, we first construct an undirected graph of CAG from large-scale domain text. Compared with conventional sequential methods, our model is more sensitive to event phrases and their causal relationships, which is difficult to be captured by previous methods.

Figure 1 illustrates the workflow of the proposed causality-associated graph neural network (CA-GNN) method. First, we construct CAG using external knowledge, such as causal trigger tables. Second, we use GNN to model both the intraevent mentions and interevent causality in a sentence. Finally, we fed the encoded input sentence sequence and the prior knowledge of CAG into a multiple convolution layer and a single BiLSTM+CRF layer sequently. The main contributions of this work are summarized as follows:

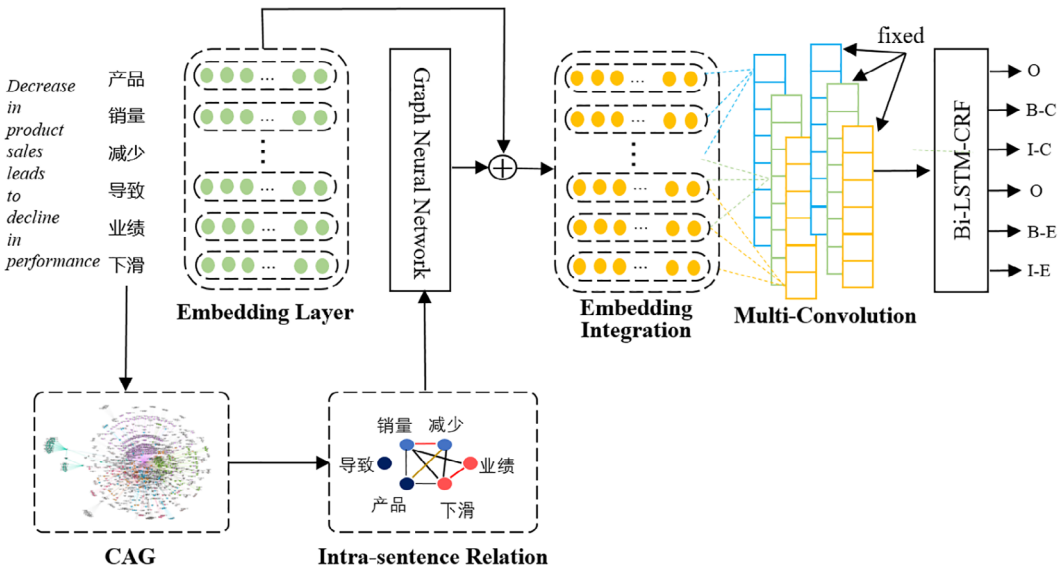

FIGURE 1 Overview of the causality-associated graph neural network framework

- This work enhances event causality by integrating the prior knowledge of event causality using a CAG, capturing intraelement transitions inside events and intercausality across events.
- We use GNN to learn the semantic information of the nodes and their rich connections in CAG, and integrate the learned graph embedding into CA-GNN to generate global semantic representations.
- Experiments conducted on the real-world Chinese datasets show that CA-GNN significantly outperforms state-of-the-art methods.

2 | RELATED WORK

In this section, we give a brief review of some related works on CEE in different perspectives, including template matching, statistical learning, neural networks and graph model.

2.1 | Template matching

Kontos and Sidiropoulou[3] use hand-crafted causal patterns to discover causal relationships. Garcia and Daniela et al.[4] develop a system, called COATIS that use the contextual information and heuristic rules to extract event causality. Furthermore, Girju et al.[1] suggest to divide the CEE into two steps: The first step is to mine lexicon-syntactic patterns that can express causality such as the most frequent pattern like *NP*1 *causal* – *verb NP*2, where *NP*1 and *NP*2 can be found in the lexical knowledge base like WordNet. The second step is to validate and rank the obtained patterns according to some semantic constraints on *NP*1, *NP*2, and *causal* – *verb* through a WordNet-based course-grained process. To reduce human participation, Ittoo and Bouma[2] present a minimally supervised algorithm that extract both explicit and implicit causal relations from domain-specific sparse texts without relying on hand-coded knowledge. However, template matching requires much manual efforts, and the generalization of template matching is far from satisfied.

2.2 | Statistical learning

Unlike template matching, statistical learning turns CEE into a classification problem. Girju[5] prove that statistical learning is a very effective method to discover causality by using C4.5 decision tree. Blanco et al.[6] improve it by constructing seven types of features. However, this model can only handles explicit causal patterns, for example, < *VerbPhrase Relator Cause* >. Rink and Harabagiu[7] use support vector machines as the classifier based on lexical, syntactic and semantic features. Although this model achieves promising results, it cannot extract more complex implicit causality. To solve this challenge, Yang and Mao[8] propose a multilevel relation extraction algorithm to recognize all potential causal relations on the basis of dependency/constituency grammar trees. However, huge efforts in feature engineering still limit the use of those models.

2.3 | Neural networks

In recent years, deep learning methods have been widely used in the field of natural language processing. Fu et al.[9] convert CEE into a sequence labeling problems. Nguyen and Grishman[10] prove that convolutional neural networks can significantly improve the performance of relation extraction, which can automatically learn features from sentences with multiple window sizes for filters and pretrained word embeddings. To reduce the impact of artificial classes, Santos et al.[11] propose the ranking CNN algorithm to minimize novel pairwise ranking loss function. Unlike convolutional neural networks, LSTM can obtain sequence information and long-term dependencies of text. Dasgupta et al.[12] propose linguistically informed recursive neural network, using word-level embeddings and other linguistic features to detect causality. Dunietz et al.[13] developed the DeepCx system, which effectively incorporates causal expression patterns by using LSTM to enhance causal relationship extraction. Jin et al.[15] combine the advantage of CNN and LSTM, and propose cascaded multistructure neural network to extract intersentence or implicit causal relations. However, it is difficult for conventional neural networks to capture the more complex causality in a sentence, and the lack of training data makes the deep model easy to overfit. To tackle this problem, some researchers try to incorporate prior knowledge into neural networks. Li et al.[20] propose use transfer contextual string embeddings to solve the problem of insufficient training data. Moreover, Li and Mao[21] propose a knowledge-oriented convolutional neural network for causal relation extraction, which consists a knowledge-oriented channel that incorporates human prior knowledge and a data-oriented channel that learns important features of causality from the data.

2.4 | Neural network on graph

GNN has achieved great success in natural language processing, and is designed to represent the graph, for example, social network and knowledge graph, which can improve the performance of deep model. Gori et al.[22] and Scarselli et al.[16] propose a GNN to process a variety of graph data structures such as acyclic graph, cyclic graph, directed graph, and undirected graph. Li et al.[17] add gated recurrent units and modern back-propagation optimization techniques to compute gradients. Recently, GNN is widely used in natural language processing. Li et al.[18] construct an event graph to utilize the event network information, and apply scaled graph neural network to script event prediction. Zhang et al.[19] propose a novel text classification TextING by using GNN with inductive word representations based on their local structures, which can also effectively produce embeddings for unseen words in the new document, the performance of TextING is better than mainstream text classification tools such as TextCNN,[23] TextRNN,[24] FastText,[25] and graph-based methods for text classification TextGCN.[26] Furthermore, Xu et al.[27] proposes a sequence labeling method to extract causal relations incorporating graph attention network based on syntactic dependency graph, and achieve desirable results. Therefore, with its powerful representation ability, GNN can learn rich semantic information of various graph data, which is very suitable for various tasks of NLP.

3 | OUR MODEL

In this section, we will introduce our proposed CA-GNN, we formulate the problem of CEE at first, then show how to apply domain prior knowledge and GNN into CEE, and give the details how to construct CA-GNN.

3.1 | Notations

CEE is a joint extraction task of events and causality, it aims to extract causal event pairs from texts as shown in Figure 2. Here, we give a formulation of this problem.

In CEE, let $I = (w_1, w_2, w_3, \ldots, w_n)$ denote a sentence consisting of word sequence, where $w_i \in I$ is a word. Label set y can be represented as $y = (O, B-C, I-C, B-E, I-E)$, where each element in y means w_i is nontarget word, the beginning of cause, continuation of cause, the beginning of effect, continuation of effect. Under a CEE model, for each word w_i, we output the probabilities for all labels in label set y.

3.2 | Constructing CAG

Figure 3 shows a high-level overview of the approach. We crawled a large number of domain news reports from the Internet, and split them into sentences, then CAG is constructed through three steps: domain causal sentence recognition, event nuggets detection, and association calculation between words.

3.2.1 | Domain causal sentence recognition

The goal of domain causal sentence recognition is to generate high-quality domain causal corpus automatically from unannotated text. We construct causal indicator lexicon at first. To improve the accuracy and recall of causal sentence recognition, we expand the causal indicator words (CIW), and disambiguate the CIW in the sentence.

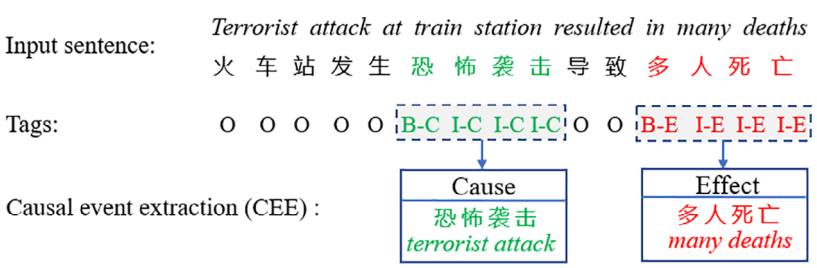

FIGURE 2 An example of CEE from raw emergency corpus, the cause and effect event pairs are extracted at once

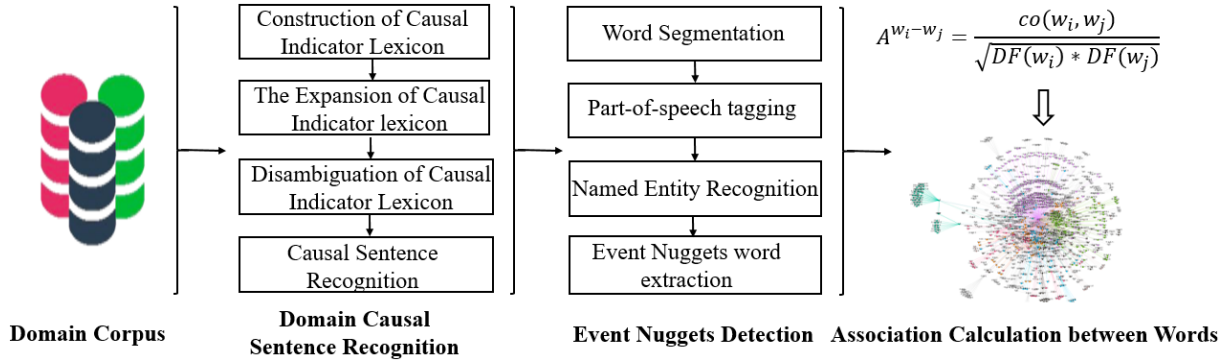

FIGURE 3 An overview of causal associated graph construction

Name of CIW	CIW	Example of causal sentence
Unary CIW Conjunctions	于是，所以，致使，以至于，因而，因为，由于，来源于，依赖 … then, so, result in, so that, therefore, because, due to, come from, depend on …	"问题地图"的出现，**源于**个别商家国家版图意识薄弱。 The emergence of the "problem map" caused by the weak awareness of individual merchants' national territory
Unary CIW- Verbs	标志着，意味着，推进，导致，引发，造成，促使 … indicate, mean, push on, lead to, bring about, cause, contribute to…	城市的扩大**使得**商品的品种增加。 The expansion of cities increase the variety of commodities.
Dual CIW Conjunctions	(之所以,因为), (之所以,由于), (之所以,缘于), (因为,从而), (因为,为此) … (The reason, because), (the reason, due to), (the reason, because of), (because, thus), (because, so)…	**只要**股价上涨，减持乃至清仓减持**就**鱼贯而来 As long as the stock price rises, shareholding reductions and even liquidation reductions will continue
Irregular CIW	(是,的原因), (是,的结果), (起,作用) … (is the reason for), (is the result of), (have the role in) …	短期美元升值，对油价**起**到一种抑制**作用**。 Short-term dollar appreciation has a negative effect on oil prices.

FIGURE 4 Examples of causal indicator lexicon and their corresponding causal sentences

Construction of CIW According to the composition of CIW, it can be divided into unary CIW and Dual CIW. According to the part of speech, CIW can be divided into conjunctions, adverbs, and verbs. Besides, the indicator word also contains some irregular causal indicator phrases, which are called irregular CIW. In order to ensure the objectivity of screening CIW, the voting method is adopted to review each CIW. The lexicon of CIW is organized as shown in Figure 4.

The expansion of causal indicator lexicon Based on the basic CIW, the recall of the identified causal sentence is relatively low. Through the analysis of unrecognized sentences, we find that the CIW in some causal sentences are mostly synonyms of unary causal indicator verbs. Therefore, HowNet* and word2vec[28] are applied to synonym expansion.

Disambiguation of causal indicator lexicon As Chinese words often has multiple meanings, recognizing causal sentences through template matching will cause errors. We analyze the identified causal sentences and find that the cause of the recognition error usually has the following two characteristics: (1) The CIW becomes part of a certain word; (2) The part of speech of the CIW has changed. Therefore, we use language technology platform (LTP)[29] to segment sentences, and identify the part of speech for each word. Some wrong causal sentences can be identified by matching words and parts of speech.

3.2.2 | Event Nuggets detection

Event Nuggets is defined as a semantically meaningful unit that expresses an event, it can be either a single word (verb, noun, or adjective) or a phrase (multiword).[30,31] For single-word event nuggets, it meet the definitions of event types/subtypes. Below are some examples of single-word event nuggets. The words in bold face are event nuggets.

*http://www.yuzhinlp.com

- The **attack** by insurgents occurred on Saturday.
- Wenchuan was severely affected by the **earthquake**.
- There was a **fire** in this place last year.

Multiword event nuggets represent single events of a complete semantic unit. Below are some examples of multiword event nuggets. The words in bold face are event nuggets.

- The company was **punished** by the China Securities Regulatory Commission for **inflating profits**.
- The news describes the **shipping accident**.
- His **death sentence** was **carried out**.

We can see that the event mainly has the following two properties: (1) An event nugget can be either a single word or a continuous or discontinuous multiword phrase. (2) Most verbs can be seen as event, which represent physical actions, followed by nouns, adjectives, and adverbs. Multiword event nuggets take various forms such as verb+noun, verb+particle/adverb, noun+noun, and so forth.

Based on the above two characteristics, we extract the main part of event in a sentence through the following two steps. First, LTP[29] is used to do word segmentation, part-of-speech (POS)-tagging and named entity recognition including person, location, and organization. Second, we pick out verbs, nouns, adjectives, and adverbs, and removed stop words and specific named entity (e.g., name of person, organization, and place). The rest of the sentence can be considered as the main part of the event nuggets.

3.2.3 | Association calculation between words

To discover the causal relationship between words, we construct CAG by using association link network (ALN).[32,33] ALN is a kind of semantic link network used to establish associations between different resources.[33] Liu et al.[32] propose an ALN-based event detection algorithm, which is used to timely discover newly occurring hot events.

The relationship between words in the text can be modeled as ALN $g = (w, e)$, each node w_i represents a word, each edge (w_i, w_j) means there is an association between word w_i and word w_j. Given a set of preprocessed sentences, we construct CAG as follows:

$$A^{w_i - w_j} = \frac{co(w_i, w_j)}{\sqrt{DF(w_i) * DF(w_j)}}, \tag{1}$$

where $co(w_i, w_j)$ is the cooccurrence frequency of word w_i and w_j, $DF(w_i)$ means the number of sentences containing the word w_i.

3.3 | Embedding integration

Then we present how to obtain latent vectors of nodes using GNN. Scarselli et al.[16] propose GNN model, and apply it to supervised classification which can directly process most of the practically useful types of graphs. Li et al.[17] further introduce gated recurrent units to GNN. GNN are well-suited for discovering relationship between words, as it can extract features both nodes and their rich connections.

For each sentence l, we can get event nuggets l_m. For $(w_i, w_j) \in l_m$, the association A_w of (w_i, w_j) in l_m can be obtained from CAG, and the recursive update procedure of GNN is as follows:

$$a^t_{w,i} = A_{w,i:}[w^{t-1}_1, w^{t-1}_2, \ldots, w^{t-1}_n]^{\mathrm{T}} + b, \tag{2}$$

$$z^t_{w,i} = \sigma(W_z a^t_{w,i} + U_z w^{t-1}_i), \tag{3}$$

$$r^t_{w,i} = \sigma(W_r a^t_{w,i} + U_r w^{t-1}_i), \tag{4}$$

$$\widetilde{w^t_i} = \tanh(W a^t_{w,i} + U(r^t_{w,i} \odot w^{t-1}_i)), \tag{5}$$

$$w^t_i = (1 - z^t_{w,i}) \odot w^{t-1}_i + z^t_{w,i} \odot \widetilde{w^t_i}, \tag{6}$$

where matrix $A \in R^{n \times n}$ is defined as association weight between words in CAG. The construction of matrix A_w is shown in Figure 5. $[w^{t-1}_1, w^{t-1}_2, \ldots, w^{t-1}_n]$ is the list of word embedding in sentence l, Equation (2) is the step that passes information between different words and their

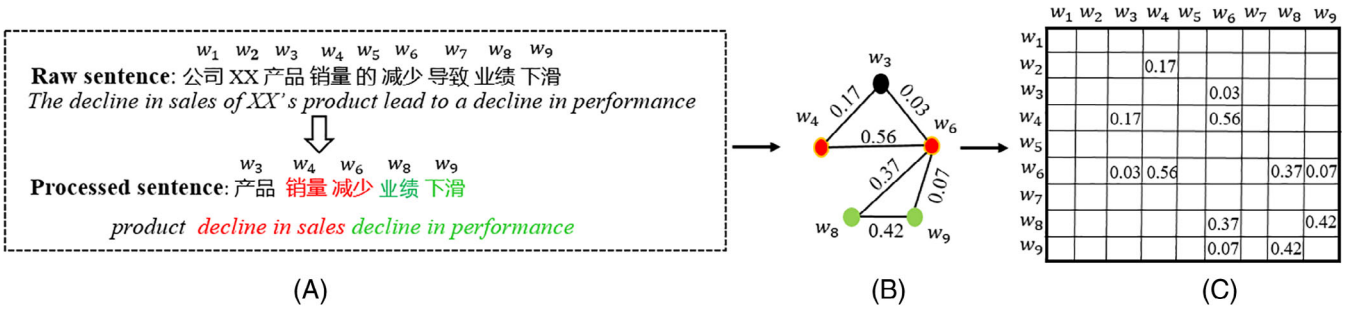

FIGURE 5 Example of constructing Association matrix between words in a sentence. (A) Examples of processed sentences, color denotes causal event mentions. (B) Association weight of processed sentences obtained from causal associated graph. (C) Association matrix between words in a sentence

association weight edges. The remaining is GRU-like updates that use the information of other nodes and the previous moment to update the hidden state of each node in the next moment. $z_{w,i}$ and $r_{w,i}$ are the rest and update gates, $\sigma = \frac{1}{1+e^{-x}}$ is the sigmoid function and $\odot$ is the elementwise multiplication operator. After updating all parameters until convergence, we can get the final word vector of GNN encoding which integrate the association information between words obtained from CAG.

After obtaining word vector of GNN encoding, we combine it with the original word-level word embedding, which keeps the original word sequence in the sentence.

To sum up, our embedding integration including two parts: (1) GNN encoding that combine intraphrase event mentions and interphrase causal relationship on CAG; (2) Original pretrained word embedding.

3.4 | Multiconvolution and BiLSTM+CRF layer

CEE can be seen as two steps. First, we need to extract a complete event mentions. Then it need to determine which is the cause or effect. Due to the ambiguous event mentions, we extract important features from embedding integration by using parallel convolution kernels with varying windows.[34] Through the following process, the causal event semantic characteristics are encoded into the filters.

$$c_i^n = f(W \cdot h_{i:i+n-1} + b) \tag{7}$$

$$c_i = [c_i^{n_0} \oplus c_i^{n_1} \oplus c_i^{n_2}], \tag{8}$$

where W is the weights of initialized filter, b is a bias term, and f is a nonlinear function such as sigmoid, ReLU,[35] and so forth. We use three convolution layers with different convolution windows $n = 3, 4, 5$ for the convolution operation. Finally, the overall feature maps of a window can be included in a single vector, and we concatenate all the vectors generated by different convolution operations.

LSTM is a variant of recurrent neural network, which is used to solve the problem of gradient vanishing.[36] The LSTM used in bidirectional long short-term memory (BiLSTM)[37] mainly consists of three parts, including an input gate i_t, an output gate o_t and a cell activation vectors v_t. BiLSTM uses two LSTM layers to learn the valid characteristics of each token in the sequence based on the past and future context information. One LSTM layer processes left-to-right information, and the other from right to left. Given an input sequence f_t generated by multiple convolution operations, the context features can be captured as follows:

$$\overrightarrow{h_t} = \overrightarrow{\text{LSTM}}(\overrightarrow{f_t}, \overrightarrow{h_{t-1}}) \tag{9}$$

$$\overleftarrow{h_t} = \overleftarrow{\text{LSTM}}(\overleftarrow{f_t}, \overleftarrow{h_{t+1}}) \tag{10}$$

$$H = [\overrightarrow{h_t}, \overleftarrow{h_t}]. \tag{11}$$

Then we can get a probability matrix P with dimensions $m \times n$, where n is the number of words and m is the number of tags.

3.5 | Object function

Conditional random field[38] can obtain the label of a given sequence in the global optimal chain, and take the interaction between adjacent labels into consideration. Given sentence l and its prediction sequence $y = y_1, y_2, \ldots, y_n$, CRF score can be obtained by using the following formula:

$$\text{score}(l, y) = \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} + \sum_{i=1}^{n} P_{i, y_i}, \tag{12}$$

where P_{i,y_i} is the prediction probability of the ith word in the sentence with label y_i, and the transition probability from label y_{i-1} to y_i is A_{y_{i-1},y_i}. The final convergence condition is to minimize the loss function which can be expressed by the following formula:

$$E = \log \sum_{y \in Y} \exp^{s(y)} - \text{score}(s, y), \tag{13}$$

where Y is the set of all possible label sequences in a sentence.

4 | EXPERIMENT AND EVALUATION

In this section, we describe the datasets across different domains and the baseline methods applied for comparison.

4.1 | Datasets

In order to verify our model, we conduct our experiments on two benchmark datasets. The first dataset is Chinese emergency corpus (CEC), which is an event ontology corpus publicly available[†]. It contains six categories: outbreaks, earthquakes, fires, traffic accidents, terrorist attacks, and food poisonings. The second is financial dataset, we crawled a large scale of Chinese financial news reports from the internet, such as Jinrongjie[‡]and Hexun[§]. This dataset contains a large number of financial articles involving causal relations.

The articles are splited into sentences by full point, semicolon and exclamation point. We invite three annotators to annotate the data. Each annotator needs to determine whether it is a causal event description. If it does, they need to annotate which part are the cause or effect. The annotated data such as *< Cause >Decreased sales< Cause > of company X products led to a < Effect >decline in performance< Effect >*. Then BIO is used to mark the sentences ("B-X" is the beginning of the cause or effect, "I-X" represents the remaining part of the cause or effect, "O" means a part that neither a cause nor an effect). Finally, we annotate 1026 causal sentences from CEC corpus, and 2270 causal sentences from financial corpus.

4.2 | Experimental setting

For the fairness of the experiment, all data preprocessing is done by LTP. Based on two large domain data (financial data and disaster data), we use the word2vec[28] to generate 100-dimensional word embedding vectors, which are used in parameter initialization of the model rather than random initialization. For all the two datasets, we set the number of training epochs to 100, the learning number of words in a sentence to 25, the size of the batch to 16, the learning rate for adam to 5×10^{-5}. To prevent overfitting, the dropout rate of training process is 0.5.

We use 8/10 of the data as the training set, 1/10 as the validation set, and 1/10 as the test set. Both CEC and financial data is shuffled with different random seeds before the cross-validation, and the evaluation metric is macro-averaged F1 score calculated from 10-fold cross-validation. We take the average value of the 10 macro-averaged F1 scores as the final result.

4.3 | Baseline methods

To prove the effective of our method, we compare it with the following baselines.

- **BiLSTM+CRF**[39] proposes a basic model for sequence tagging, which uses BiLSTM to mine past and future input features and capture sentence level tag information with CRF.
- **CNN+BiLSTM+CRF**[14] is used for POS tagging. They first use CNN to encode a word into character-level representation, and feed them into BiLSTM to capture context information of each word. Finally, a sequential CRF is used to jointly decode labels for the whole sentence.

[†]https://github.com/shijiebei2009/CEC-Corpus
[‡]http://www.jrj.com.cn/
[§]http://www.hexun.com/

- **CSNN**[15] uses CNN and self-attention to capture features relationship, and the higher-level phrase representations are feed into BiLSTM and CRF layer for CEE.
- **BERT+CSNN** is a very effective baseline language model for CEE, which uses pretrained BERT[40] trained from large-scale unlabeled corpus as input.

4.4 | Comparison with baseline methods

To demonstrate the overall performance of our method, we compare it with other state-of-the-art baselines including BiLSTM+CRF, CNN+BiLSTM+CRF, CSNN, and BERT+CSNN. The overall performance in terms of precision *P*, recall *R*, and F1 scores F_1 is shown in Table 2.

CA-GNN integrates CAG as prior knowledge into CEE. In this model, we jointly consider the complex intraphrase event mentions as well as interphrase causal relationship. According to the experiment, the best performance on two datasets verify the effectiveness of the proposed method.

The performance of traditional sequence labeling algorithm BiLSTM+CRF is relatively poor. Although LSTM-based methods can capture both past and future contextual information, and are very effective in sequence labeling such as named entity recognition. However, unlike named entity recognition, phrase-level event mentions often contain several consecutive words and their expressions are relatively ambiguous, resulting in poor performance of BiLSTM+CRF. Even so, CNN+BiLSTM+CRF and CSNN achieve better performance than BiLSTM+CRF, demonstrating the importance of introducing CNN for CEE. These two methods use CNN to extract phrase-level features of several adjacent words, which is crucial for phrase-level event extraction, then BiLSTM is used to capture the sequence and long-term dependency information of convolution features.

CEE is more complicated than conventional event extraction, and the accuracy of CEE by combining convolution and LSTM still needs to be improved. We can see that the results of BERT+CSNN have been greatly improved compared with CSNN. BERT is a pretrained transformer network with multihead attention over 12 (base-model) or 24 layers (large-model) that can be set for various downstream NLP tasks, and achieves promising results, including question answering, sentence classification. It learns a good representation for each word by using self-supervised learning method on a large amount of corpus, which means pretrained transformer network contains a large amount of nondomain prior knowledge such as causal events, named entities, and so forth. Compared with the state-of-the-art methods like BERT+CSNN, CA-GNN further considers complex phrase-level causal transitions between words in a sentence, which can capture more long-term dependence and implicit connections between words. Therefore, though BERT+CSNN uses a pretrained model with a deeper neural network, we can see the performance of our method is still higher than BERT+CSNN. CA-GNN integrates domain causal knowledge and phrase-level event mentions into undirected graph CAG, and uses GNN to encode the semantic information of nodes and edges in CAG. Besides, CA-GNN adopts multiple convolution to get features of graph encoding and original word vector, and apply BiLSTM to further discover the sequence and causal relationship among features. On the contrary, BERT is trained on large-scale nondomain data, which may not be sufficient for domain CEE. Other sequence labeling models, such as CNN+BiLSTM+CRF and CSNN, they usually learn more obvious causal events, when the description of the event is ambiguous, or the causal relationship is more obscure, conventional models are ineffective to cope with this situation.

4.5 | Ablation experiments

To study the contribution of each component in CA-GNN, we conducted ablation experiments on the two dataset and display the results on Table 3, the comparison methods are -MultCNN and -GNN. -MultCNN is the CA-GNN model without Multiple convolution, -GNN is the CA-GNN model

TABLE 2 Comparison experiment of causal event extraction with other baselines over two datasets

	Financial			CEC		
Method	**P**	**R**	**F1**	**P**	**R**	**F1**
BiLSTM+CRF	68.26	70.83	69.51	73.26	**73.52**	73.39
CNN+BiLSTM+CRF	75.25	73.73	74.47	74.44	73.29	73.86
CSNN	75.87	73.55	74.67	74.32	73.52	73.91
BERT+CSNN	74.95	77.58	76.23	74.22	75.00	74.61
CA-GNN	**78.86**	**77.60**	**78.23**	**76.30**	73.05	**74.64**

Note: Bold values are the best results in the comparison method.
Abbreviation: CA-GNN, causality-associated graph neural network.

TABLE 3 Ablation experiments of causal event extraction on two datasets

	Financial			CEC		
Method	**P**	**R**	**F1**	**P**	**R**	**F1**
CA-GNN	**78.86**	**77.60**	**78.23**	**76.30**	**73.05**	**74.64**
-MultCNN	73.25	72.37	72.80	74.51	72.52	73.49
-GNN	78.83	75.54	77.15	74.73	72.70	73.70

Note: -MultCNN is the CA-GNN model without multiple convolution, -GNN is the CA-GNN model without graph neural network. Bold values are the best results in the comparison method.
Abbreviations: CA-GNN, causality-associated graph neural network; GNN, graph neural network.

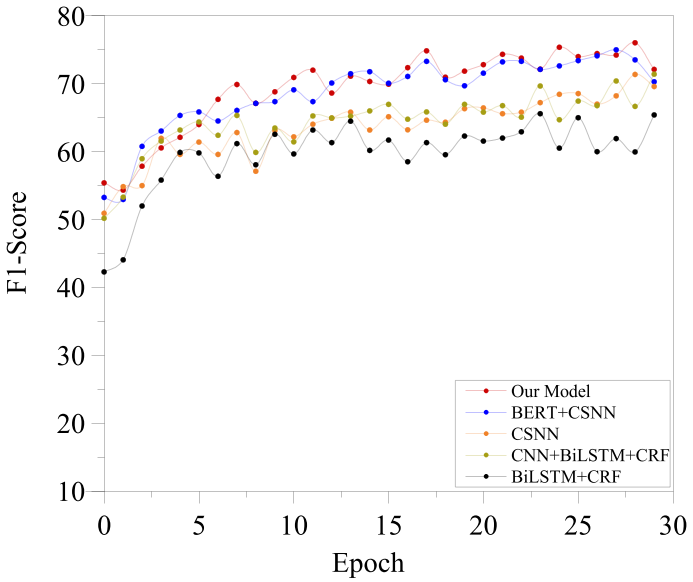

FIGURE 6 F1-Score of different methods on the financial test dataset under multiple training epoch

without GNN. The results show that both the multilayer convolution layer and the GNN coding layer play a positive role in CEE compared with the conventional method BiLSTM+CRF.

It can be seen that the F1 score of -MultCNN (only GNN and BiLSTM +CRF) increased by 1.7% on average on the two datasets compared with BiLSTM+CRF. The F1 score of -GNN (only MultCNN and BiLSTM+CRF) increased by 3.98% on average on the two datasets. By contrast, CA-GNN achieves the state-of-the-art performance by combining GNN and MultCNN, and the average F1 score on the two datasets increased by 4.99%. This is probably that the CEE is a pipeline process, multiconvolution can extract more complete event representation, GNN can give more guidance information to determine the cause and effect of the event.

4.6 | Comparison w.r.t. epoch

We further compare the convergence speed and performance of our model and other baseline models on finical dataset. As shown in Figure 6, our model begin to converge after 10 iterations. However, the NON-BERT model starts with a low F1 score at the beginning of the iteration, whose performance dramatically decreases when *epoch* < 10.

Due to the lack of prior knowledge and the limitation of dataset size, the performance of NON-BERT model is inferior to BERT+CSNN. We can see that although our model lacks BERT, the performance of CA-GNN iteration is still slightly better than BERT+CSNN, demonstrating the effectiveness of combining domain prior knowledge and GNN to extract causal event.

4.7 | Comparison w.r.t. sentence length

Figure 7 shows the performance of several baseline models with different sentence length on the finical dataset. We split the test dataset into six parts according to the sentence length, the proportions of each sentence length are 3%, 21.5%, 20.5%, 16.5%, 13.5%, and 25%. The BERT+CSNN is

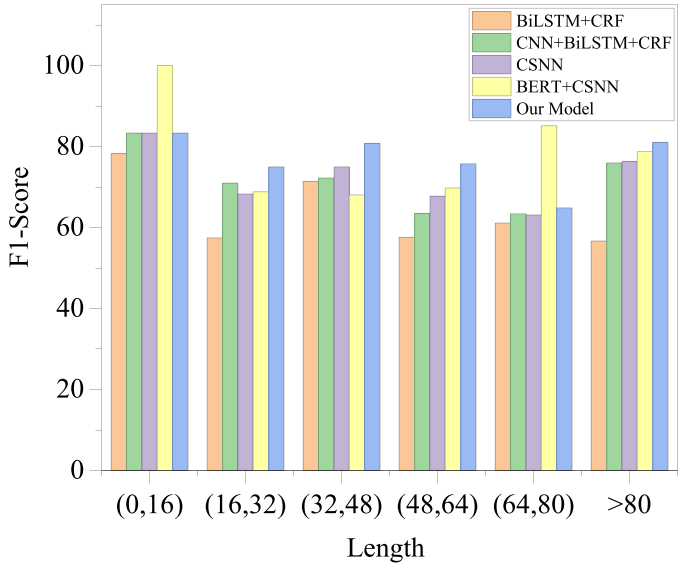

FIGURE 7 F1-score of our method compared with other methods on the financial dataset under different sentence length intervals

a strong baseline that outperforms BiLSTM+CRF, CNN+BiLSTM+CRF, and CSNN when the sentence length is more than 48 characters. However, BERT+CSNN has no obvious advantage when the sentence length is less than 48 characters. By contrast, CA-GNN not only gives higher results over short sentences, but also shows its effectiveness and robustness when the sentence length is more than 48 characters. It gives a higher F1 score in most cases compared with the NON-BERT baselines, which indicates that global sentence semantics and long-range dependency can be captured under the combination of domain prior knowledge and the graph structure.

4.8 | Case study

Figure 8 illustrates an example that demonstrates the effectiveness of CA-GNN. We can easily find that BiLSTM+CRF method is inferior to CNN+BiLSTM+CRF and CSNN, and it even misses the effect event in the implicit causality extraction. CEE is a phrase-level event relationship extraction task, CNN can extract phrase-level features of several adjacent words, and BiLSTM is used to capture the sequence and long-term dependency

Method	Implicit Causality	Explicit Causality
Sentence	公司连续两年 <cause>**亏损**<cause>，该公司可能被<effect>**深交所暂停债券上市交易**<effect>. The company's **losses** for two consecutive years，and may be suspended from trading in its bonds by the Shenzhen Stock Exchange.	较大的<cause>**汇兑损失**<cause>,使归属于母公司股东的<effect>**净利润同比下降**<effect>. Larger **exchange losses** caused a **proportional decrease in net profit** belonging to shareholders of the parent company
CALGNN	**Cause:** 亏损 **Effect:** 深交所暂停债券上市交易 **Cause:** Losses **Effect:** suspension of bond listing and trading by Shenzhen Stock Exchange	**Cause:** 汇兑损失 **Effect:** 净利润同比下降 **Cause:** Exchange losses **Effect:** Proportional decrease in net profit
BERT+CSNN	**Cause:** 亏损 **Effect:** 深交所暂停债券上市交易 **Cause:** Losses **Effect:** suspension of bond listing by Shenzhen Stock Exchange	**Cause:** 汇兑损失 **Effect:** 净利润同比下降 **Cause:** Exchange losses **Effect:** Proportional decrease in net profit
CSNN	**Cause:** 亏损 **Effect:** 深交所暂停债券上市 **Cause:** Losses **Effect:** suspension of bond listing and trading by Shenzhen Stock Exchange	**Cause:** 汇兑损失 **Effect:** 同比下降 **Cause:** Exchange losses **Effect:** Proportional decrease
CNN+BiLSTM+CRF	**Cause:** 亏损 **Effect:** 公司可能被深交所暂停债券上市交易 **Cause:** Losses **Effect:** may be suspended from trading in its bonds by the Shenzhen Stock Exchange	**Cause:** 汇兑损失 **Effect:** 同比下降 **Cause:** Exchange losses **Effect:** Proportional decrease
BiLSTM+CRF	**Cause:** 亏损 **Effect:** None **Cause:** Losses **Effect:** None	**Cause:** 较大的汇兑损失 **Effect:** 净利润同比下降 **Cause:** Larger exchange losses **Effect:** Proportional decrease in net profit

FIGURE 8 Examples of different baseline methods for implicit causality and explicit causality extraction

information of convolution features. However, due to the ambiguity of the event mentions, it is difficult to fully learn these complex information by the model itself. Knowledge-driven CA-GNN and pretrained BERT+CSNN strengthen the ability of model to extract causal events by introducing external knowledge, yielding better results for CEE.

5 | CONCLUSION

In this article, we present a novel approach CA-GNN for CEE. Our method can effectively integrate domain knowledge into the model and improve the accuracy of CEE. The whole model can be divided into three parts. First, we use CAG to represent domain causal knowledge. Then we apply GNN to learn accurate embedding from CAG, the generated word vector can capture the complex relationship of intraphrase event mentions and interphrase causality in a sentence. Besides, we apply multiconvolution to extract phrase-level text features, and further use BiLSTM to learn long-term causal sequence information. The performance of our approach has been experimentally verified on two datasets for CEE.

Although CA-GNN can improve the performance of the model, there are still some shortcomings. CEE can be seen as a pipeline task. First, it needs to extract the complete event mentions, then determine which is the cause or effect. It may be more appropriate to model intraelement events and intercausality across events separately. In future work, we will try to build a pipeline model to further improve the accuracy of CEE based on existing research.

ACKNOWLEDGMENTS

The research reported in this article was supported in part by the Natural Science Foundation of China under the grant No. 91746203, Ministry of Industry and Information Technology project of the Intelligent Ship Situation Awareness System under the grant No. MC-201920-X01, and National Natural Science Foundation of China under the grant No. 61991415.

DATA AVAILABILITY STATEMENT

This article mainly uses two datasets, including CEC dataset and financial dataset. For CEC dataset, the data that support the findings of this study are openly available in https://github.com/shijiebei2009/CEC-Corpus. For financial dataset, the raw data is obtained from http://www.jrj.com.cn/ and http://www.hexun.com/. We reannotate the two datasets, and the data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Girju R, Moldovan DI. Text mining for causal relations. Proceedings of the FLAIRS Conference; 2002:360-364.
2. Ittoo A, Bouma G. Extracting explicit and implicit causal relations from sparse, domain-specific texts. Proceedings of the International Conference on Application of Natural Language to Information Systems; 2011:52-63.
3. Kontos J, Sidiropoulou M. On the acquisition of causal knowledge from scientific texts with attribute grammars. *Int J Appl Expert Syst*. 1991;4(1):31-48.
4. Garcia D COATIS, an NLP system to locate expressions of actions connected by causality links. Proceedings of the International Conference on Knowledge Engineering and Knowledge Management; 1997:347-352.
5. Girju R. Automatic detection of causal relations for question answering. Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering; 2003:76-83.
6. Blanco E, Castell N, Moldovan DI. Causal relation extraction. *Lrec*. European Language Resources Association; 2008;66:74.
7. Rink B, Harabagiu S. Utd: classifying semantic relations by combining lexical and semantic resources. Proceedings of the 5th International Workshop on Semantic Evaluation; 2010:256-259.
8. Yang X, Mao K. Multi level causal relation identification using extended features. *Expert Syst Appl*. 2014;41(16):7171-7181.
9. Fu J, Liu Z, Liu W, Zhou W. Event causal relation extraction based on cascaded conditional random fields. *Pattern Recogn Artif Intell*. 2011;24(4):567-573.
10. Nguyen TH, Grishman R. Relation extraction: perspective from convolutional neural networks. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing; 2015:39-48.
11. Santos CND, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks; 2015. arXiv preprint arXiv:1504.06580.
12. Dasgupta T, Saha R, Dey L, Naskar A. Automatic extraction of causal relations from text using linguistically informed deep neural networks. Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue; 2018:306-316.
13. Dunietz J, Carbonell JG, Levin L. DeepCx: a transition-based approach for shallow semantic parsing with complex constructional triggers. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018:1691-1701.
14. Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf; 2016. arXiv preprint arXiv:1603.01354.
15. Jin X, Wang X, Luo X, Huang S, Gu S. Inter-sentence and implicit causality extraction from Chinese corpus. Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining; 2020:739-751.
16. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. *IEEE Trans Neural Netw*. 2008;20(1):61-80.
17. Li Y, Tarlow D, Brockschmidt M, Zemel R. Gated graph sequence neural networks; 2015. arXiv preprint arXiv:1511.05493.
18. Li Z, Ding X, Liu T. Constructing narrative event evolutionary graph for script event prediction; 2018. arXiv preprint arXiv:1805.05081.
19. Zhang Y, Yu X, Cui Z, Wu S, Wen Z, Wang L. Every document owns its structure: inductive text classification via graph neural networks; 2020. arXiv preprint arXiv:2004.13826.
20. Li Z, Li Q, Zou X, Ren J. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing*. 2021;423:207-219.

21. Li P, Mao K. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Syst Appl*. 2019;115:512-523.
22. Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. Proceedings. 2005 IEEE International Joint Conference on Neural Networks; Vol. 2, 2005:729-734.
23. Kim Y. Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP; 2014:1746-1751.
24. Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning; 2016. arXiv preprint arXiv:1605.05101.
25. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification; 2016. arXiv preprint arXiv:1607.01759.
26. Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. Proceedings of the AAAI Conference on Artificial Intelligence; Vol. 33, 2019:7370-7377.
27. Jinghang X, Wanli Z, Shining L, Ying W. Causal relation extraction based on graph attention networks. *J Comput Res Develop*. 2020;57(1):159.
28. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality; 2013. arXiv preprint arXiv:1310.4546.
29. Che W, Li Z, Liu T. Ltp: a Chinese language technology platform. *Coling 2010*; Demonstrations Volume; 2010:13-16.
30. Araki J, Mitamura T. Open-domain event detection using distant supervision. Proceedings of the 27th International Conference on Computational Linguistics; 2018:878-891.
31. Mitamura T, Yamakawa Y, Holm S, et al. Event nugget annotation: processes and issues. Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation; 2015:66-76.
32. Liu Y, Luo X, Xuan J. Online hot event discovery based on association link network. *Concurr Comput Pract Exper*. 2015;27(15):4001-4014.
33. Luo X, Xu Z, Yu J, Chen X. Building association link network for semantic link on web resources. *IEEE Trans Autom Sci Eng*. 2011;8(3):482-494.
34. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. *J Mach Learn Res*. 2011;12(ARTICLE):2493-2537.
35. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. *Icml*; Omnipress; 2010.
36. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780.
37. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition; 2016. arXiv preprint arXiv:1603.01360.
38. Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data; 2001.
39. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging; 2015. arXiv preprint arXiv:1508.01991.
40. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding; 2018. arXiv preprint arXiv:1810.04805.

How to cite this article: Gao J, Luo X, Wang H. Chinese causal event extraction using causality-associated graph neural network. *Concurrency Computat Pract Exper*. 2022;34(3):e6572. https://doi.org/10.1002/cpe.6572