Knowledge-Based Systems 258 (2022) 109935

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Joint event causality extraction using dual-channel enhanced neural network

Jianqi Gao [a], Hang Yu [b,*], Shuang Zhang [c]

[a] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[b] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[c] ACRE Coking and Refractory Engineering Consulting Corporation (Dalian), MCC, Dalian, China

ARTICLE INFO

Article history:
Received 8 May 2022
Received in revised form 8 August 2022
Accepted 20 September 2022
Available online 27 September 2022

Keywords:
Event causality extraction
Natural language processing
Dual-channel enhanced neural network
Graph convolutional network

ABSTRACT

Event Causality Extraction (ECE) plays an essential role in many Natural Language Processing (NLP), such as event prediction and dialogue generation. Recent research in NLP treats ECE as a sequence labeling problem. However, these methods tend to extract the events and their relevant causality using a single collapsed model, which usually focuses on the textual contents while ignoring the intra-element transitions inside events and inter-event causality transition association across events. In general, ECE should condense the complex relationship of intra-event and the causality transition association among events. Therefore, we propose a novel dual-channel enhanced neural network to address this limitation by taking both global event mentions and causality transition association into account. To extract complete event mentions, a Textual Enhancement Channel(TEC) is constructed to learn important intra-event features from the training data with a wider perception field. Then the Knowledge Enhancement Channel(KEC) incorporates external causality transition knowledge using a Graph Convolutional Network (GCN) to provide complementary information on event causality. Finally, we design a dynamic fusion attention mechanism to measure the importance of the two channels. Thus, our proposed model can incorporate both semantic-level and knowledge-level representations of events to extract the relevant event causality. Experimental results on three public datasets show that our model outperforms the state-of-the-art methods.

1. Introduction

Event causality extraction (ECE) is a challenging task in Information Extraction (IE) that automatically extract event descriptions and identify causal relations between events. For example, "*Tax revenue decline is caused by fiscal decentralization reforms*". There exists a cause–effect relation, i.e., causality association, between "*fiscal decentralization reforms*" and "*tax revenue decline*". This pattern reflects the logical relationship between events, which is of great value for many NLP tasks such as recommendation system [1], event prediction [2,3], and question answering [4]. However, due to the ambiguity of event description, limited dataset, and long-distance dependence of event causality, designing an effective ECE method is still a topic worthy of long-time research.

Recent research on ECE can be divided into three types, including rule-based methods [5,6], statistical methods [7,8], and deep learning methods [9,10]. Rule-based methods that rely on pattern matching consider massive hand-crafted linguistic features, including lexical, syntactic, and semantic patterns, to extract causal events. Generally, these methods have the following problems: (1) Regardless of extensive manual efforts, it is impossible to enumerate all causal language expressions; (2) The diversity of semantics in natural language (e.g., lexicon ambiguities) leads to extraction errors when performing template matching. Compared with rule-based methods, statistical methods learn causality from annotated corpus through much effort of feature engineering, which requires careful design of features and can only be applied to limited domains.

In recent years, deep learning has achieved success in many NLP tasks. Among these neural networks, the convolutional neural network (CNN) [11] has the advantage of extracting n-gram features that have been widely used in relation extraction. In addition, the language model [12] (e.g., Bidirectional Encoder Representations from Transformers, BERT) is served as the token-wise encoder of neural networks and have achieved state-of-the-art performance in many NLP tasks. However, there exist shortcomings in either pre-trained model BERT and its variants or CNNs: (1) BERT contains a lot of non-domain knowledge, and its improvement on domain-specific tasks with less data is limited;

* Corresponding author.
E-mail addresses: 193139@sjtu.edu.cn (J. Gao), yuhang@shu.edu.cn (H. Yu), zhangshuang@acre.com.cn (S. Zhang).

https://doi.org/10.1016/j.knosys.2022.109935
0950-7051/

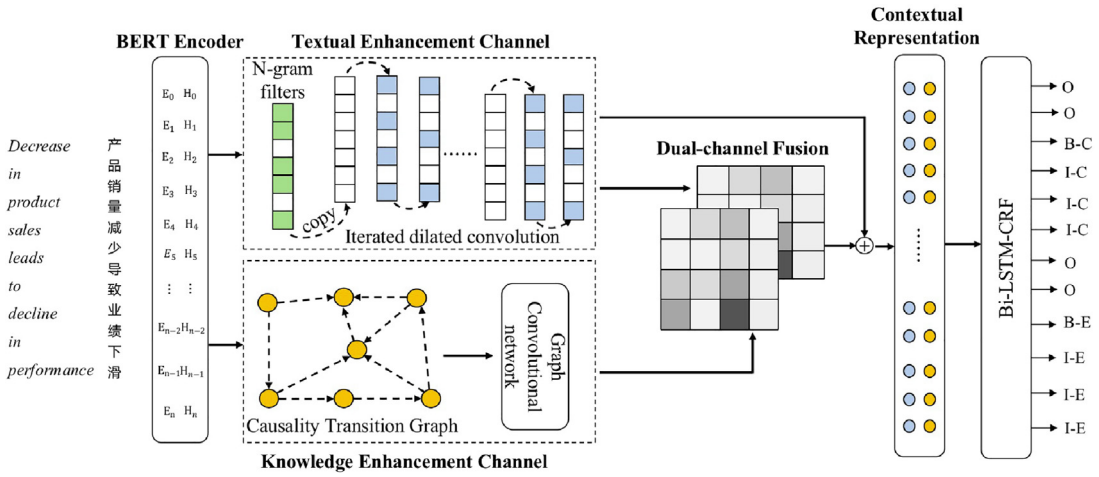

Fig. 1. Overview of dual-channel enhanced neural network.

(2) CNN can only extract local n-gram features of text, its ability to perceive context is limited; (3) Deep models are prone to overfit on insufficient datasets; (4) Due to the ambiguity of event mentions and the long-distance dependence of cause and effect, it is difficult for the model to extract complete event causality just relying on the model itself and limited annotated dataset.

To solve the above four problems, we propose a novel dual-channel enhanced neural network for ECE. As shown in Fig. 1, our proposal combines event semantics learned from the text and causality transition among events obtained from a Causality Transition Graph (CTG). To utilize global textual context and CTG, we design a dual-channel enhancement architecture to leverage multiple sources, such as textual description and graph structure, to generate rich-attribute embeddings for events, simultaneously encapsulating causality relations. First, the pre-trained model BERT is applied to generate the token-wise representations for the input sentence. Then, above the BERT encoder, the Textual Enhancement Channel (TEC) passes the hidden representations into an iterated dilated convolution neural network to encode the global textual information. However, differing from previous works [13], we learn the probability distribution of n-grams from the training data using Naive Bayes and initialize the convolutional kernel with centroid vectors of the cause/effect n-gram clusters before the fine-tuning procedure. In parallel with TEC, we design another Knowledge Enhancement Channel (KEC) by constructing a CTG from a causal corpus using causal indicators to integrate event causality transition associations. Finally, to enhance the model's ability to identify the complex causality transition of inter-event in the sentence, we use the attention mechanism to link the representation of TEC and KEC together. The contributions of the paper are four folds:

- We propose a dual-channel enhanced model that integrates knowledge obtained from labeled data and domain unstructured text into the model, which can fully consider intra-element transitions inside events and inter-event causality transition association across events.
- We propose a TEC to enhance event extraction on cause and effect by integrating the important n-gram filters learned from labeled data into iterated dilated convolutions, capturing semantic features inside events with global contextual information.
- We propose an effective method to construct KEC incorporating causality transition associations obtained from CTG, which can improve the model's ability to identify complex causal relationships between events.
- Experiments conducted on three datasets demonstrate that the dual-channel enhancement strategies are interrelated and effective, and achieve state-of-the-art performance on both in-domain and out-domain datasets.

2. Related work

ECE can be divided into template matching, statistical learning, and neural networks. In this section, we will briefly introduce the three parts.

2.1. Template matching on ECE

ECE based on template matching uses causal indicators to construct adaptive semantic templates and extract causal events from text. Khoo et al. [5] apply adaptive templates and linguistic clues to ECE. Girju et al. [6] obtain lexical patterns that can express causality from knowledge bases such as WordNet, and sort the obtained patterns through the coarse-grained semantic constraints. To reduce manual participation, Ittoo et al. [14] extract complex causal relations from domain texts through a minimally supervised algorithm without relying on manual rules. However, template matching requires a lot of manpower to carefully design the template, and the accuracy and generalization of the template matching still need to be improved.

2.2. Statistical learning on ECE

Methods of statistical learning turn ECE into a classification task. Inui et al. [7] propose a computational model for ECE, which can extract four types of event causality, including cause, effect, precondition, and means. Blanco et al. [8] further present a supervised method to extract causality from open domain texts. However, the above model can only deal with explicit causality, and the accuracy of implicit causality extraction is low. To solve this problem, Yang et al. [15] develop an ECE system that is able to extract more complex causal relations between two noun phrases represented by fixed verbs or prepositions. However, extensive feature engineering and possible noise reduce the accuracy and applicability of the model.

2.3. Neural networks on ECE

In recent years, deep learning has become the mainstream method of natural language processing, among which the most widely used are Recursive Neural Networks (RNN) [16,17] and

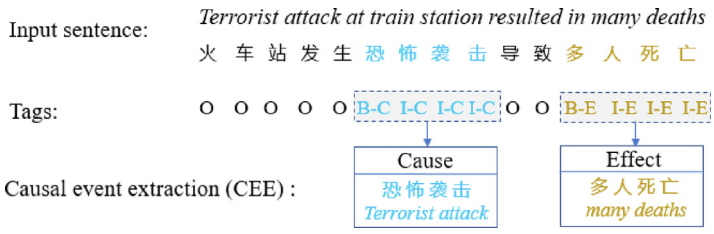

Fig. 2. Example of event causality extraction, the description of cause and effect are extracted at once.

Convolutional Neural Networks (CNN) [18,19]. Socher et al. [20] present a method for relation classification using a recursive neural network. To represent relations more compactly, Ebrahimi et al. [21] use RNN for relation classification based on the shortest path of two entities in the dependency graph. However, RNN-based methods require dependency analysis which may introduce error propagation.

CNN can automatically extract n-gram semantic features from the text and has been proven suitable for ECE due to its excellent local feature extraction and discriminative representation capabilities [11,22]. Furthermore, Santos et al. [23] propose a ranking CNN algorithm for relation classification, and reduce the influence of artificial classes on experimental results. However, the last layer of neurons in the convolution of conventional CNN models can only obtain a small piece of information in the original input data. Since the contextual information in the text may affect the label of the input text, the perceptual field of CNN needs to be further improved. To solve this problem, Yu et al. [24] propose dilated convolutions to increase the perception field of the filter. Since increasing the depth of the dilated convolution on limited data can easily lead to overfitting, Strubell et al. [25] propose an improved method called Iterated Dilated Convolutional Neural Network (IDCNN), which greatly increases the perception range and stability of the model by using a recursive approach. For ECE, Liang et al. [26] propose a novel multi-level causality detection network to detect text with event causality by combining the advantages of feature engineering in providing prior knowledge and neural networks in capturing contextual information. Furthermore, Jin et al. [9] propose a cascaded multi-structure neural network to improve the accuracy of inter-sentence and implicit causality extraction. However, due to complex of ECE, just relying on a small amount of labeled data and the model itself, the improvement of ECE is limited. In response, Wang et al. [10] propose a join extraction framework to incorporate the prior knowledge like frequent event causality mentioned into the convolution kernel. Li et al. [27] incorporate multiple knowledge into the embedding representation to generate hybrid embedding representations. Besides, Li et al. [13] obtain causal knowledge from labeled data and external knowledge bases such as WordNet and FrameNet, respectively, and integrate it into the convolution initialization, which improves the overall performance of the model on the ECE task.

2.4. Neural network on graph

Due to the challenge of heterogeneous graph data, people have done extensive and in-depth research on how to apply the deep learning method to the graph [28,29]. Previous research treat neural network on the graph as a form of recurrent neural network. However, their method requires repeated application of the contraction maps as the activation function until the representation of the node reaches a stable state. This restriction was alleviated by adding gated recurrent units and improving the back-propagation optimization strategy [30]. As for Graph Convolutional Network (GCN), Bruna et al. [31] extend the convolutional neural network to a graph. Then Defferrard et al. [32] use Chebyshev polynomials to obtain graph convolution to remove expensive Laplacian eigen-decomposition. Based on previous work [31,32], Kipf et al. [33] further simplify spectral graph convolutions via a localized first-order approximation. Recently, people have done a lot of research on the application of graph neural networks on NLP, such as text classification [34], event argument extraction [35], document-level graphs for relation extraction [36], and document-level graphs for event causality identification [37].

3. The proposed approach

In this section, we will introduce our model architecture as shown in Fig. 1. For both the textual enhancement channel and knowledge enhancement channel, we use BERT as a token encoder.

3.1. Notations

Event causality extraction (ECE) is a subtask of information extraction. It aims to extract event phrases containing causal relationships from texts, as shown in Fig. 2.

In ECE, let $l = (x_1, x_2, x_3, \ldots, x_n)$ denote a sentence consisting of several tokens x_i. $y = (B{-}C, I{-}C, B{-}E, I{-}E, O)$ is the label set. Each element in y represents x_i is the beginning of the cause, a continuation of cause, the beginning of the effect, a continuation of effect, non-target word. The ECE model needs to output the probability that x_i belongs to each label in y.

3.2. Textual enhancement channel

In the field of NLP, the convolution operation can be seen as the semantic feature extraction of a sentence. In our model of ECE, the iterated dilated convolution layer aims to extract feature map of intra-event mentions and capture contextual information in the sentence. Inspired by [10,13,38], the construction of the textual enhancement channel is shown in Fig. 3. Firstly, we extract important cause/effect n-gram semantic features from the labeled data. Secondly, the semantic features of n-gram with similar semantics are divided together using the clustering method to generate high-level n-gram semantic representation. Finally, the high-level n-gram representation is fed into the initialization process of iterated dilated convolution. In this operation, we use high-level n-gram features for part of the filters, and the remaining positions are randomly initialized, allowing the model to learn more useful features by itself. Therefore, the textual enhancement channel consists of two steps: n-gram selection and filter initialization.

N-gram Selection. Events can be represented by meaningful phrases composed of several ordered words, and convolution can extract n-gram features from the text. Therefore, n-gram semantic features (e.g., tri-gram: decline in performance) can be used as prior knowledge for convolution initialization to enable the

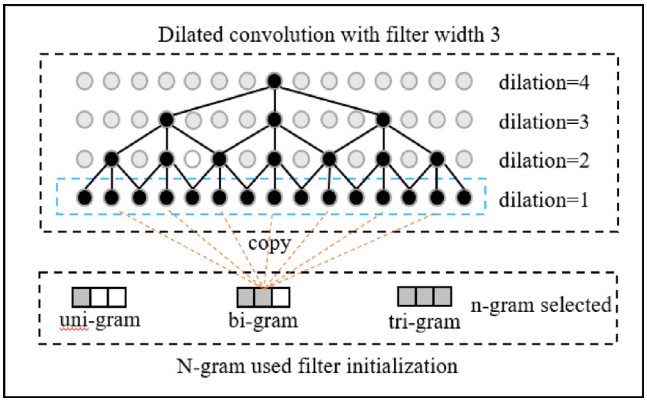

Fig. 3. Textual enhancement channel, the important n-gram semantic features selected for filters are copied to the first layer of a bottom-up dilated CNN block.

model to extract complete event mentions. To select effective n-gram semantic features from labeled data, Naive Bayes is applied to select effective n-gram as follows:

$$score = \frac{\left(n_c^i + b\right) / \| n_c \|_1}{\left(n_e^j + b\right) / \| n_e \|_1} \tag{1}$$

where c and e are the cause and effect, respectively. n_c^i is the number of sentences that contain n-gram i in cause c. $\| n_c \|_1$ is the number of n-gram in cause c, b is a smoothing parameter. Intuitively, when the length of the n-gram is close to the length of the event description in the sentence, it is easier to capture the semantic information and boundary of event mentions. Therefore, we choose the event length that accounts for the largest proportion in the labeled data as the parameter of n-gram, and select the top 20% of n-gram as the semantic feature for convolution initialization.

Filter Initialization. In NLP, the convolution filter is usually a $m \times n$ dimensional matrix, where m represents the width of the convolution filter, and n is the embedding dimension of each token. The convolutional operator for each token x_t can be calculated as follows:

$$c_t = W_c \bigoplus_{k=0}^{r} x_{t \pm k} \tag{2}$$

where $\oplus$ is vector concatenation. W_c is the filter width of r tokens.

To enlarge the perception field of convolution, dilated convolution [24] performs a wider effective input width by skipping over δ inputs at a time. The dilated convolution can be represented as follows:

$$c_t = W_c \bigoplus_{k=0}^{r} x_{t \pm k\delta} \tag{3}$$

where δ is the dilation width, when dilation width $\delta > 1$, dilated convolution can provide a wider perceptual field than simple convolution without adding additional parameters.

We can obtain top k cause/effect n-grams through Formula (1), and encode each n-gram with BERT to generate embedding representations. Since the limited filters in the CNN are not enough to use all the n-gram semantic features, we use K-means to divide the similar n-gram features together, and thereby the cluster centroid vector can be used as an abstract representation of an n-gram class. In order to assign the centroid vector to the convolution kernel, we set the number of clusters equal to the number of filters. Considering the impact of contextual information on cause/effect events, we input the centroid vector to the center position of the filter, and the remaining positions of the filter are randomly initialized. This operation can make the model learn more useful features itself. For dilated convolution, simply increasing the depth of the stacked dilated convolution can easily cause an overfitting problem. Therefore, we feed the final generated filters into the bottom-up first layer of iterated dilated convolutions [25], which enables the model to extract complete event mentions with a wider perceptual field and desirable generalization capabilities.

3.3. Knowledge enhancement channel

TEC can extract complete event mentions with full consideration of contextual information. Naturally, Knowledge Enhancement Channel (KEC) can be designed to enhance the model's ability to capture the causal relationship between events. Considering that there are a large number of explicit causality in the network, which can be used to improve the accuracy of the model in event causality extraction. The KEC is constructed through the following steps, including Causality Transition Graph (CTG) construction and GCN encoding based on CTG.

3.3.1. Causality transition graph construction

Fig. 4 shows the main procedure for constructing a Causality Transition Graph (CTG). We crawl a large number of news texts from the Internet, and split them into sentences. CTG can be constructed through three steps: causal sentence recognition, event nuggets detection, and causality transition calculation.

Causal Sentence Recognition. Causal sentence recognition aims to identify sentences containing causal relationships from unlabeled news texts using Causal Indicator Words (CIW), including CIW construction, CIW expansion, CIW disambiguation. Each CIW lexicon and its corresponding example sentences are shown in Table 1.

Firstly, we construct a CIW lexicon. According to the composition and part of speech of CIW, CIW can be divided into unary CIW conjunctions, unary CIW verbs, and dual CIW conjunctions. In addition to the above three cases, CIW also has some irregular phrase descriptions (e.g., irregular CIW: have the role in). Secondly, we expand the obtained CIW lexicon. The recall of the initially constructed CIW lexicon is low, we use HowNet,[1] WordNet[2] and word2vec [39] to perform synonym expansion for CIW. In order to ensure the accuracy and objectivity of CIW, we use the voting strategy to correct each CIW. Finally, we disambiguate the expanded CIW lexicon. The causal sentences can be identified by template matching. However, due to the multiple meanings of some CIW, the recognition of causal sentences by template matching may lead to some errors. For example, for the word "so", as a conjunction it can be equivalent to "lead to" and "so that", but as an adverb of degree, it is equivalent to "very" and "quite". Through comparative analysis, it can be found that the wrong recognition of causal sentences generally has the following two characteristics: (1) the part of speech of CIW in the sentence has changed; (2) CIW becomes part of a phrase. Based on these two properties. We use Language Technology Platform (LTP) [40] and Natural Language Toolkit (NLTK)[3] to identify the parts of speech in Chinese and English, respectively. Partially wrong causal sentences are identified by judging whether the CIW becomes part of a phrase or whether the part-of-speech of CIW has changed.

Event Nuggets Detection. Event nuggets are a meaningful semantic unit that can describe an event, which can be a single

[1] http://www.yuzhinlp.com.
[2] https://wordnet.princeton.edu.
[3] http://www.nltk.org/.

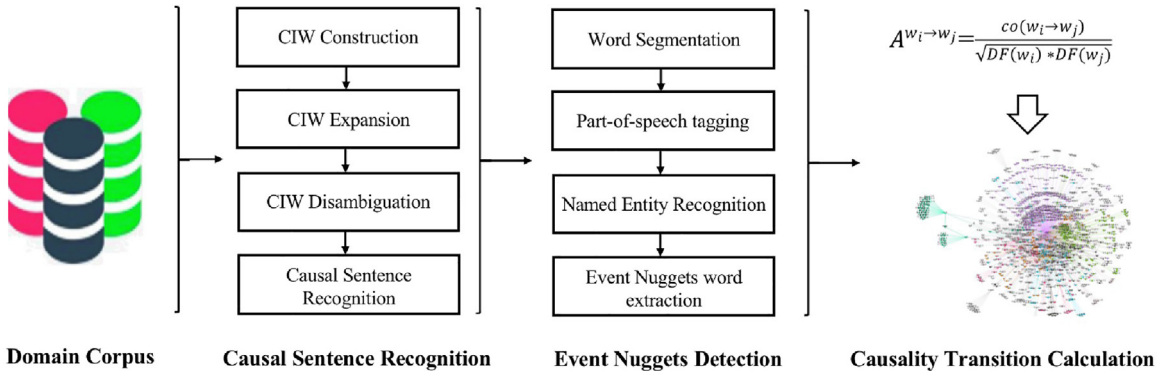

Fig. 4. An overview of causality transition graph construction.

Table 1
Examples of different types of causal indicator words and their corresponding causal sentences.

Name of CIW	CIW	Example of causal sentence
Unary CIW conjunctions	于是，所以，致使... then, so, result in ...	“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，导致，引发 ... indicate, lead to, bring about ...	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)... (the reason, because), (the reason, due to) ...	**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) ... (is the reason for),(is the result of) ...	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.

word or a phrase [41,42]. A single word is generally a verb, noun, and adverb, which refers to an event type. For a phrase, it is a complete semantic unit composed of multiple words (continuous or discontinuous). The following are two examples of event nuggets, the word in bold face is event nuggets.

- Many people were **killed** in **car accident**.
- **Calluses** are caused by a **skin abnormality**.

It can be seen that events contain the following two properties in terms of composition and part of speech. (1) An event consists of a single word or phrases of multiple words; (2) Events generally consist of verbs or common nouns, followed by adjectives and adverbs. Therefore, we extract the main part of the events in the text through the following steps. Firstly, LTP is applied to do word segmentation for Chinese datasets. We perform part-of-speech tagging and named entity recognition on the corpus, LTP and NLTK are applied to process the Chinese corpus and English corpus, respectively. Then we sequentially remove stop words and specific entities (e.g., people, organizations, and places), and pick out verbs, common nouns, adjectives, and adverbs from the sentence. The selected content in the sentence can be considered as event nuggets $l_m (l_m \subseteq l)$.

Causality Transition Calculation. Through the construction of CIW, we can get a large number of CIW, and further obtain a large number of explicit causal sentences through template matching. Besides, the obtained CIW usually have a clear direction. For example, lead to, owing to, bring about, etc. are indicative of causal direction. Therefore, CIW can be divided into order matching from cause to effect, middle matching from cause to effect, and order matching from effect to cause according to its position in the sentence. We deal with the above three cases as follows: (1) For order matching from cause to effect, the second word of CIW is used as separation. The left part is the cause, and the right part is the effect. (2) For the order matching from effect to cause, the second word of CIW is used as separation. The left part is the effect, and the right part is the cause. (3) For the middle matching from cause to effect, the current word is used as separation. The left part is the cause, and the right part is the effect.

Association Link Network (ALN) is a kind of semantic link network that can be used to effectively associate and organize various resources on the Internet [43,44]. Inspired by ALN, we design an ALN-based Causality Transition Graph (CTG) to model causality transition in massive texts. CTG can be represented by graph $g = (w, e)$, and each node w_i is the keyword of the event description, each edge $(w_i \to w_j) \in e$ is the causality transition weight from the word w_i to word w_j. Given a set of pre-processed causal sentences through event nuggets detection, we can construct CTG as follows:

$$D^{w_i \to w_j} = \frac{co(w_i \to w_j)}{\sqrt{DF(w_i) * DF(w_j)}} \tag{4}$$

where $co(w_i \to w_j)$ is the co-occurrence frequency from cause word w_i to effect word w_j, $DF(w_i)$ is the number of sentences containing the word w_i.

3.3.2. GCN encoding based on CTG

In this section, we use Graph Convolutional Networks (GCN) [33] to encode the causality transition information between nodes in CTG. Given a graph $G = (V, E)$, which contains node $v_i \in V$ and edge $e_{ij} (e_{ij} = (v_i, v_j) \in E)$ with weights w_{ij}. The input of GCN consists of two parts: node representations denoted by $X = \{x_i\}_{i=1}^{N}$, where x_i is the feature vector of node v_i, and the weighted adjacency matrix of the graph $A \in R^{NXN}$ where $A_{ij} = w_{ij}$.

For a sentence l (assuming it has N tokens), the word representations X can be obtained by the BERT encoder. We extract event nuggets $l_m \subseteq l$ through event nuggets detection, and the matched weight $w_{ij} \in A$ of causality transition between words in l_m can be obtained from CTG. However, it should be noted that BERT has different encoding granularity for English and Chinese, which use word encoding and character encoding, respectively. For Chinese data, the nodes in CTG are represented by words, and BERT

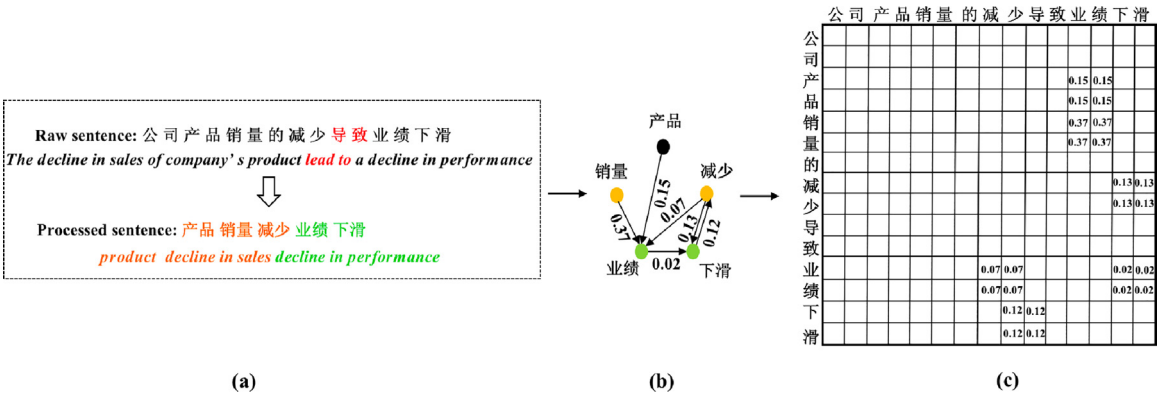

Fig. 5. Example of constructing matrix of causality transition for Chinese. (a) Causal event nuggets obtained After causal sentence recognition and event nuggets detection, color text in processed sentence denotes causal event mentions. (b) the weight of causality transition between words obtained from CTG. (c) the adjacency matrix of causality transition between words.

encodes each token by character, making GCN unable to encode the subgraph of CTG due to the different dimensions of the input matrix. To solve this problem, we use the method shown in Fig. 5 to generate the final weighted adjacency matrix. If the weight of causality transition from Chinese word w_i to Chinese word w_j is w_{ij}, the weight of character $c_i \in w_i$ to character $c_j \in w_j$ is w_{ij}. This operation retains the association information between words inside the event and reflects the causality transition association between words across the event.

After we get the embedding matrix X and the weighted adjacency matrix of causality transition A, the layer-wise propagation rule of multi-layer GCN can be summarized as follows:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{5}$$

where $\tilde{A} = A + I_N$, $\tilde{D}$ is a degree matrix of $\tilde{A}$, and I_N is the identity matrix. $H^{(0)} = X$ is fed into the input layer of GCN, which contains origin word features and sequence information. $H^{(l)} \in R^{N\times M_l}$ is the activation matrix that contains hidden information of the vertices in the lth layer. $W^{(l)}$ is a trainable weight matrix. $\sigma(\cdot)$ denotes the activation function. Such propagation rules can be considered as a differentiable generalization of the Weisfeiler–Lehman algorithm [33].

3.4. Dual-channel fusion

The textual enhancement channel can extract the features of event mentions with wider perception fields. The knowledge enhancement channel captures the causality transition between events. Intuitively, considering that the knowledge enhancement channel can provide supervised guidance information for the learning process of determining the causal relationship between feature maps of the textual enhancement channel, and the information in the two parts may have different priorities, we use the multi-head attention mechanism [45] to link the two parts together to represent the global preference of event causality. The formulation of multi-head attention is as follows:

$$H_{HEAD} = concat\,(head_1, head_2, \ldots, head_h)\,W \tag{6}$$

where $head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$, $Q \in \mathbb{R}^{t\times d}$, $K \in \mathbb{R}^{t\times d}$, and $V \in \mathbb{R}^{t\times d}$ are the input of attention, representing the query matrix, key matrix, and value matrix, respectively. The parameter matrices of ith linear projection is $W_i^Q \in \mathbb{R}^{n\times(\frac{d}{h})}$, $W_i^K \in \mathbb{R}^{n\times(\frac{d}{h})}$, $W_i^V \in \mathbb{R}^{n\times(\frac{d}{h})}$, and the attention values of h heads are concatenated together. Considering that the TEC retains the semantic information and sequence information of the text, and the output of dual-channel fusion reflects the contribution of each feature in TEC under the guidance of KEC. The outputs of TEC and attention structure are cascaded to output token-wise contextual representations.

3.5. Object function

In order to make full use of contextual information, the token-wise contextual representations is fed to Bi-directional Long Short-Term Memory (BiLSTM) [46], which can learn the semantic features of sentences from both directions. Assuming the output sequences of BiLSTM is $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$, the two hidden vectors can be concatenated into $[\overrightarrow{h_t}, \overleftarrow{h_t}]$ to generate the final representation.

Adjacent labels usually have strong dependencies in the obtained label sequence. Therefore, we use Conditional Random Field (CRF) [47] to decode the label sequence output by BiLSTM, which can fully consider the sequence and correlation between labels. Given sentence l and its predicted label sequence $y = (y_1, y_2, \ldots, y_n)$, the score of CRF can be expressed as follows:

$$score(l, y) = \sum_{i=1}^{n+1} A_{y_{i-1},y_i} + \sum_{i=1}^{n} P_{i,y_i} \tag{7}$$

where A is the transition matrix in the CRF layer, A_{y_{i-1},y_i} represents the transition score from label y_{i-1} to label y_i. P is the score matrix output by BiLSTM, P_{i,y_i} is the confidence score of word i belongs to label y_i. The convergence conditions of the model by minimizing the loss function are as follows:

$$E = \log \sum_{y\in Y} \exp^{s(y)} - \text{score(l, y)} \tag{8}$$

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

4. Experiment and evaluation

4.1. Datasets

We conduct experiments on three benchmark datasets to verify the effectiveness of our method. The dataset contains two Chinese domain datasets and one English non-domain dataset, including the financial dataset, CEC (Chinese emergency corpus) dataset[4] and SemEval-2010 task 8 dataset [48]. The financial dataset is a Chinese dataset, and we crawl a large number of news reports from financial websites such as Jinrongjie[5] and Hexun.[6]

4 https://github.com/shijiebei2009/CEC-Corpus.
5 http://www.jrj.com.cn/.
6 http://www.hexun.com/.

Table 2
Statistics result of the three datasets.

Statistics	Financial	CEC	SemEval2010
Data size	2270	1026	1003
Average sentence length	57.94	31.14	18.54
Distance between cause and effect	13.49	10.24	5.33
Length of cause with the largest proportion	4(41%)	2(31%)	1(85%)
Length of effect with the largest proportion	4(23%)	4(27%)	1(91%)

These texts contain a certain number of causal events. We split these news reports into sentences and annotated causal events on the corpus, finally getting 2270 causality instances. For the CEC dataset, it is a public event ontology corpus that contains six event types, including fires, earthquakes, outbreaks, terrorist attacks, traffic accidents, and food poisonings. We annotate causal events on the corpus, and obtain 1026 causality instances. For the SemEval-2010 task 8 dataset, it contains 10 717 annotated samples, and each sample is a sentence annotated with entity pairs (e_1, e_2) and their relationship. We re-annotate sentences containing causal relations and obtained 1003 causal sentence instances. The details of the three datasets are shown in Table 2.

4.2. Experimental setting

For CTG construction, we constructed domain-specific Chinese CTG (including Chinese financial CTG and Chinese emergency CTG) and non-domain English CTG for different datasets. The large-scale news corpus used to construct CTG is crawled from the related website, and the domain corpus is obtained by matching the pre-defined domain lexicon.

For the above three datasets, We use BERT to encode text to generate embedded representations. The hyper-parameters of the model during training are as follows: (1) the n-gram filter used for convolution initialization on the three datasets of financial, CEC, and SemEval-2010 task 8 are 4, 2, and 1, respectively; (2) the number of epochs for model training is 70; (3) the batch size for the training set is 8; (4) the learning rate of the optimizer adam used for optimization is 1×10^{-5}.

In order to ensure the reliability of the experimental results, all the data is shuffled with different random seeds before training, and is divided into a training set, validation set, and test set with a proportion of 80%, 10% and 10%, respectively. We use the average F1 value of ten-fold cross-validation as the evaluation metric, and the final result is the average value of the ten macro-averaged F1 scores.

4.3. Method for comparison

We apply some classical methods to the above three datasets to verify the effectiveness of the proposed method.

- **BiLSTM+CRF**: This is a basic model for information extraction, which use BiLSTM to capture the sequence and contextual information of texts, and CRF is applied to decode the obtained tag sequence.
- **CNN+BiLSTM+CRF** [49]: The model uses CNN to extract multiple n-gram semantic features, and the BiLSTM+CRF layer is applied to capture the dependencies between features.
- **CSNN** [9]: The author uses CNN to extract text features, and an association between semantic features is established using the self-attention mechanism.
- **BERT+CISAN** [10]: The authors integrate important event causality mentions into the convolution initialization, and BERT is served as an embedded encoder for CISAN to replace GloVe word embedding.

Table 3
Macro-averaged F1 scores of various methods on three datasets, results are shown in percentages.

Model	Financial	CEC	SemEval2010
BiLSTM+CRF	74.75	68.74	73.20
CNN+BiLSTM+CRF	74.31	71.68	74.20
CSNN	74.59	70.61	73.71
BERT+CISAN	77.09	75.93	77.65
BERT+SCITE	78.20	74.13	77.41
Our model	**79.89**	**82.27**	**80.02**

- **BERT+SCITE** [27]: The method incorporates multiple prior knowledge into the embedding representation to generate hybrid embedding representations, and a multi-head attention mechanism is applied to learn the dependencies between causal words. We encode sentences using BERT instead of hybrid embedding representations.

4.4. Results and analysis

Table 3 shows the results of our method compared with other baselines. Besides, BERT+CISAN and BERT+SCITE are two strong benchmarks to verify the effectiveness of our dual-channel enhanced method.

It can be seen that the performance of the CNN-based method (e.g., CNN+BiLSTM+CRF and CSNN) is better than BiLSTM + CRF. In the conventional information extraction tasks (e.g., named entity recognition), BiLSTM can capture the sequence information and dependencies between tokens, and CRF can adjust the label sequence according to the distribution and interrelation of the predicted label sequence to generate the best label sets. Therefore, BiLSTM+CRF is more suitable for conventional information extraction tasks. However, ECE is a joint extraction task of phrase-level event extraction and causality recognition. CNN can extract the semantic n-gram features of several consecutive tokens, which is crucial for phrase-level information extraction tasks. Experimental results also demonstrate the importance of introducing CNN for ECE.

In addition, BERT-based methods significantly outperform non-BERT methods. This is mainly because the pre-trained model BERT is trained from a large-scale corpus, which contains a lot of prior knowledge and has been proven to be suitable for many NLP tasks. Compared with BERT-based methods, our model is 2.8%, 6.34%, and 2.37% higher than BERT+CISAN and 1.69%, 8.14%, and 2.61% higher than BERT+SCITE on financial, CEC, and SemEval2010, respectively. Event causality extraction is a joint extraction task of event extraction and causal relationship identification. The two BERT-based methods only consider prior knowledge of an event, but it ignores the influence of a large amount of causal knowledge that exists on the Internet on the model. Our method integrates prior knowledge and data features into the model through the knowledge enhancement channel and textual enhancement channel to improve the model's ability to extract causal events. Among them, the textual enhancement channel feeds the semantic features of important event mentions to the iterated dilated convolutions, capturing features of different sizes inside events with global contextual information. The

Table 4
Ablation analysis of our proposed method on three datasets, "–" means to remove a component from the model in order.

Model	CEC	Financial	SemEval2010
Our model	**82.27**	**79.89**	**80.02**
-KEC	77.52	77.77	78.21
-Attention	76.10	76.16	75.93
-TEC	75.15	74.95	73.96
-BERT	68.26	71.81	68.55

Table 5
Results of n-gram filters with different lengths on three datasets.

Model	Ngram	Financial	CEC	SemEval2010
Our model	Unigram	79.61	81.52	**80.02**
	Bigram	79.11	**82.27**	78.67
	Trigram	78.91	81.32	79.07
	Quagram	**79.89**	81.79	79.26

knowledge enhancement channel improves the model's ability of causality identification between events by using GCN to learn the causality transition information between nodes in the CTG. Moreover, financial and CEC are two domain datasets, our method can achieve better performance on the three datasets, and the performance on the domain datasets is better than the non-domain dataset (SemEval 2010). Namely, our method also has certain applicability to non-domain datasets.

4.5. Ablation experiments

Table 4 is an ablation analysis to show whether each part of our method has a positive role in the ECE task. We perform -kec, -attention, -tec, and -BERT in order. First, we perform the -KEC operation. After removing KEC, the output c_t of TEC is concatenated with self-attention operation of c_t (this operation is proved to be effective in BERT+CISAN), and is fed into the BiLSTM+CRF layer. Then, we perform the -attention operation to directly feed the output c_t of TEC into the BiLSTM+CRF layer. Next, we perform the -TEC operation to initialize the convolution randomly. Finally, we perform the -BERT operation to initialize the model's embedding layer randomly, which can be seen as the comparison baseline model IDCNN+BiLSTM+CRF.

It can be seen that after removing the relevant parts of the model in order, the performance of the model gradually decreases, indicating that all parts of the model are useful for ECE tasks. The contribution of BERT has the greatest impact on the model because BERT is a deep transformer neural network trained from a large-scale corpus. As a result, it can learn good representation for each token and adjust its parameters through fine-tuning to make the model obtain state-of-the-art performance.

In addition, TEC is designed to help the model extract a complete event description based on global contextual information, and KEC is used to assist the model in identifying the causal relationship between features of different granularity. Then the attention mechanism links the two parts together, and the features of event description and causality transition can be matched in a targeted manner. Finally, the model can pick out the optimal feature combination. Therefore, all these layers simultaneously contribute to obtaining SOTA performance.

4.6. Comparison w.r.t semantic convolutional filters

To find the appropriate n-gram length for different datasets, we designed the experiment as shown in Table 5 to explore the influence of different n-gram lengths on the experimental results.

Table 6
F1-score of various methods under different sentence length intervals on the financial dataset.

Model	(0,16)	(16,32)	(32,48)	(48,64)	(64,80)	(>80)
BiLSTM+CRF	78.26	57.50	71.43	57.58	61.11	56.68
CNN+BiLSTM+CRF	83.33	71.01	72.26	63.49	63.46	75.95
CSNN	83.33	68.29	75.00	67.74	63.16	76.34
BERT+CISAN	100	70.93	73.17	69.72	76.19	84.40
BERT+SCITE	100	78.32	75.82	70.43	**82.27**	80.49
Our model	**100**	**82.30**	**77.46**	**70.74**	69.84	**86.71**

The result shows that the best length of n-grams for financial, CEC, and SemEval2010 is 4, 2, and 1, respectively. By analyzing the differences among different datasets, we find that the event lengths with the largest proportions in the financial dataset, CEC dataset, and SemEval2010 dataset are also 4, 2, and 1, respectively. This phenomenon proves that this paper's textual enhancement channel can better capture common event mentions. For example, in the English dataset, "*the fire cause a slight injury on the ventral side of the neck and at the base of horns*". In this sentence, the cause is "*fire*", and the effect is "*injury*". We think the convolution is more sensitive to a such event when the frequent event mentions are expressed in a single word.

4.7. Comparison w.r.t. epoch

Fig. 6 shows the learning curve of our methods compared with other baselines. It can be seen the iteration result of the BERT model (e.g., our model, BERT+CISAN and BERT+SCITE) is significantly better than the NON-BERT model. Besides, the convergence speed of the BERT+CISAN and BERT+SCITE outperforms the non-BERT model. However, we find an interesting phenomenon, i.e., our method does not have a very high starting point at the beginning of the iteration, but it converged quickly after three iterations. This may be caused by two reasons: (1) The search for causality transition in the knowledge enhancement channel is a dynamic process, so the model needs several rounds of iteration to find the appropriate parameters; (2) Due to the complexity of our model and insufficient training data, the model needs several rounds of iterations to find suitable weights until converge. Even so, the time taken for our model to converge is almost the same as BERT+CISAN.

4.8. Comparison w.r.t. sentence length

Table 6 shows the performance of various methods for sentences of different lengths on the financial dataset. We divide the test set into six intervals according to the length of sentences. It can be seen that our results are better than other methods in most cases, which means that our method has better results in both long-range dependency and short-range dependency causality. This is mainly because both KEC and TEC of our model have the ability to perceive global contextual information. TEC can perceive more contextual information by iterated dilated convolution to extract phrase-level event semantic features, and the CTG used in KEC contains global causality transition information. Therefore, the combination of KEC and TEC can enable the model to capture more contextual information from the text and improve the model's performance on the ECE task.

5. Conclusion and future work

We present a novel dual-channel enhanced neural network for Event Causality Extraction (ECE). The proposed method improves the model's ability of ECE through the Textual Enhancement Channel (TEC) and Knowledge Enhancement Channel (KEC). The

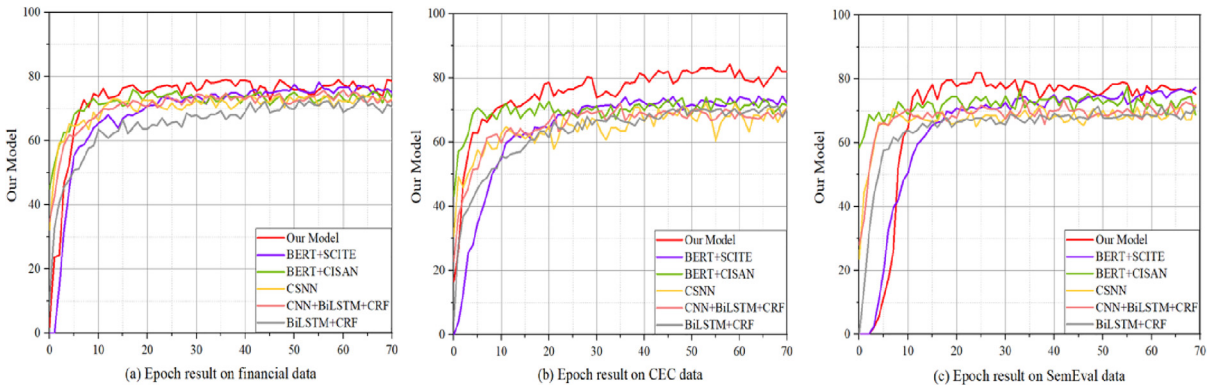

Fig. 6. The learning curve of different methods on the three test datasets.

TEC uses Naive Bayes and unsupervised clustering to generate important n-gram semantic features on cause and effect from labeled data, and then the features are applied to the initialization of iterated dilated convolutions, allowing the model to extract more complete event mentions while considering rich contextual information. Meanwhile, the KEC models the external causal knowledge as Causality Transition Graph (CTG), which can be constructed from the related corpus. Then we use Graph Convolutional Networks (GCN) to capture the complex information transition of inter-event causality learned from CTG. As a result, our method can capture intra-element transitions inside events and inter-causality association across events. The proposed method is fully automatic without sophisticated feature engineering, and the performance of our approach has been experimentally verified on three datasets.

In future work, we will try a more effective method to build a causality transition graph without relying on too much external corpus, and explore potential applications of our model in other non-domain-specific tasks.

CRediT authorship contribution statement

Jianqi Gao: Conceptualization, Methodology, Software, Writing – original draft, Data curation. **Hang Yu:** Writing – review & editing, Visualization, Supervision. **Shuang Zhang:** Writing & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The work presented in this paper was supported by the Shanghai Yangfan Program, China (22YF1413600).

References

[1] P. Sulikowski, T. Zdziebko, K. Coussement, K. Dyczkowski, K. Kluza, K. Sachpazidu-Wójcicka, Gaze and event tracking for evaluation of recommendation-driven purchase, Sensors 21 (4) (2021) 1381.
[2] Z. Li, X. Ding, T. Liu, Constructing narrative event evolutionary graph for script event prediction, 2018, arXiv preprint arXiv:1805.05081.
[3] F. Nawaz, N.K. Janjua, O.K. Hussain, PERCEPTUS: Predictive complex event processing and reasoning for IoT-enabled supply chain, Knowl.-Based Syst. 180 (2019) 133–146.
[4] Y. Zhou, Y. Chen, J. Zhao, Y. Wu, J. Xu, J. Li, What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 14638–14646.
[5] C.S. Khoo, J. Kornfilt, R.N. Oddy, S.H. Myaeng, Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing, Literary and Linguistic Computing 13 (4) (1998) 177–186.
[6] R. Girju, D.I. Moldovan, et al., Text mining for causal relations, in: FLAIRS Conference, 2002, pp. 360–364.
[7] T. Inui, K. Inui, Y. Matsumoto, Acquiring causal knowledge from text using the connective marker tame, ACM Trans. Asian Lang. Inf. Process. (TALIP) 4 (4) (2005) 435–474.
[8] E. Blanco, N. Castell, D. Moldovan, Causal relation extraction, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008.
[9] X. Jin, X. Wang, X. Luo, S. Huang, S. Gu, Inter-sentence and implicit causality extraction from chinese corpus, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2020, pp. 739–751.
[10] Z. Wang, H. Wang, X. Luo, J. Gao, Back to prior knowledge: Joint event causality extraction via convolutional semantic infusion, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2021, pp. 346–357.
[11] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, 2014, arXiv preprint arXiv:1404.2188.
[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[13] P. Li, K. Mao, Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts, Expert Syst. Appl. 115 (2019) 512–523.
[14] A. Ittoo, G. Bouma, Extracting explicit and implicit causal relations from sparse, domain-specific texts, in: International Conference on Application of Natural Language to Information Systems, Springer, 2011, pp. 52–63.
[15] X. Yang, K. Mao, Multi level causal relation identification using extended features, Expert Syst. Appl. 41 (16) (2014) 7171–7181.
[16] M.-T. Luong, R. Socher, C.D. Manning, Better word representations with recursive neural networks for morphology, in: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, 2013, pp. 104–113.
[17] R. Socher, C.C.-Y. Lin, A.Y. Ng, C.D. Manning, Parsing natural scenes and natural language with recursive neural networks, in: ICML, 2011.
[18] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377.
[19] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Adv. Neural Inf. Process. Syst. 28 (2015).
[20] R. Socher, B. Huval, C.D. Manning, A.Y. Ng, Semantic compositionality through recursive matrix-vector spaces, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 1201–1211.
[21] J. Ebrahimi, D. Dou, Chain based RNN for relation classification, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1244–1249.
[22] T.H. Nguyen, R. Grishman, Relation extraction: Perspective from convolutional neural networks, in: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015, pp. 39–48.
[23] C.N.d. Santos, B. Xiang, B. Zhou, Classifying relations by ranking with convolutional neural networks, 2015, arXiv preprint arXiv:1504.06580.
[24] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, 2015, arXiv preprint arXiv:1511.07122.
[25] E. Strubell, P. Verga, D. Belanger, A. McCallum, Fast and accurate entity recognition with iterated dilated convolutions, 2017, arXiv preprint arXiv:1702.02098.

[26] S. Liang, W. Zuo, Z. Shi, S. Wang, J. Wang, X. Zuo, A multi-level neural network for implicit causality detection in web texts, Neurocomputing (2022).
[27] Z. Li, Q. Li, X. Zou, J. Ren, Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings, Neurocomputing 423 (2021) 207–219.
[28] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, Vol. 2, 2005, pp. 729–734.
[29] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Netw. 20 (1) (2008) 61–80.
[30] Y. Li, D. Tarlow, M. Brockschmidt, R. Zemel, Gated graph sequence neural networks, 2015, arXiv preprint arXiv:1511.05493.
[31] J. Bruna, W. Zaremba, A. Szlam, Y. LeCun, Spectral networks and locally connected networks on graphs, 2013, arXiv preprint arXiv:1312.6203.
[32] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, Adv. Neural Inf. Process. Syst. 29 (2016).
[33] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.
[34] L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 7370–7377.
[35] A.P.B. Veyseh, T.N. Nguyen, T.H. Nguyen, Graph transformer networks with syntactic and semantic structures for event argument extraction, 2020, arXiv preprint arXiv:2010.13391.
[36] F. Christopoulou, M. Miwa, S. Ananiadou, Connecting the dots: Document-level neural relation extraction with edge-oriented graphs, 2019, arXiv preprint arXiv:1909.00228.
[37] M.T. Phu, T.H. Nguyen, Graph convolutional networks for event causality identification with rich document-level structures, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 3480–3490.
[38] S. Li, Z. Zhao, T. Liu, R. Hu, X. Du, Initializing convolutional filters with semantic features for text classification, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1884–1889.
[39] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. 26 (2013).
[40] W. Che, Z. Li, T. Liu, Ltp: A chinese language technology platform, in: Coling 2010: Demonstrations, 2010, pp. 13–16.
[41] J. Araki, T. Mitamura, Open-domain event detection using distant supervision, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 878–891.
[42] T. Mitamura, Y. Yamakawa, S. Holm, Z. Song, A. Bies, S. Kulick, S. Strassel, Event nugget annotation: Processes and issues, in: Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, 2015, pp. 66–76.
[43] X. Luo, Z. Xu, J. Yu, X. Chen, Building association link network for semantic link on web resources, IEEE Trans. Autom. Sci. Eng. 8 (3) (2011) 482–494.
[44] Y. Liu, X. Luo, J. Xuan, Online hot event discovery based on association link network, Concurr. Comput.: Pract. Exper. 27 (15) (2015) 4001–4014.
[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[46] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[47] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 2001.
[48] I. Hendrickx, S.N. Kim, Z. Kozareva, P. Nakov, D.O. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz, Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals, 2019, arXiv preprint arXiv:1911.10422.
[49] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, 2016, arXiv preprint arXiv:1603.01354.