

## Research Article

# Joint Event Relation Identification Based on Multiscale Convolutional Neural Network and Sharing Strategy

Yi Zhang <sup>1,2</sup>, Wanhua Cao <sup>1,2</sup>, Yuanbin Wang <sup>2</sup>, Juntao Liu <sup>2</sup>, Yongqi Chen <sup>3</sup>,  
and Xiao Wei <sup>3</sup>

<sup>1</sup>College of Computer Science and Technology, Harbin Engineering University, Harbin, China

<sup>2</sup>Wuhan Digital Engineering Research Institute, Wuhan, China

<sup>3</sup>School of Computer Engineering and Science, Shanghai University, Shanghai, China

Correspondence should be addressed to Yi Zhang; [y Zhang85@hrbeu.edu.cn](mailto:y Zhang85@hrbeu.edu.cn)

Received 25 June 2022; Accepted 25 August 2022; Published 9 September 2022

Academic Editor: Kapil Sharma

Copyright © 2022 Yi Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

At present, most of the event relation identification work mainly focuses on the sequential temporal and explicit causal relation between events. These methods usually ignore the role of synchronous temporal and implicit causal relation in sentences, which makes the semantic understanding of the model deviate from the text. In this paper, we propose a joint event relation identification model. The model uses bidirectional GRU and multiscale convolutional neural network to obtain the context semantic features and multiscale local semantic features of text, respectively. Then, these two kinds of features are fused to fully obtain the semantics of the text itself. In addition, we build encoders and decoders of event temporal and causal relation, respectively, to obtain the event temporal and causal semantic features from text. In this process, considering the correlation between event timing and causality, we use three different parameter sharing strategies to realize the interaction between event temporal and causal semantic features. The experimental results on the legal field dataset we constructed show that our model has made significant improvements compared with the baseline model. Through experimental analysis, our method can effectively improve the identification performance of synchronous temporal and implicit causality relation.

## 1. Introduction

Event relation identification [1] in text is an important research topic in the field of information extraction and natural language processing, especially in some specific fields, such as the judicial field. Mastering the causal and temporal relation between events can provide support for the analysis of the cause and development of the case. Among the possible relation types between events, this paper focuses on the joint identification of temporal and causal relation.

For the causal relation of events, it mainly depends on the explicit causal indicators in the text, such as “cause,” “so,” “therefore,” and “because.” By constructing the causal rule base [2] and using the model to learn the causal relation features in the text, the causal relation between events can be identified. In example 1 shown in Figure 1, due to the causal indicator

“cause,” the model can easily identify the “cause-effect” relation between event A and event B. However, in the face of some implicit causal relation, that is, when there are no explicit causal indicators in the text, the semantic features of causality are obscure, and it is difficult for the model to learn the causal features. In example 2 shown in Figure 1, there are no indicators in the text that can directly express cause and effect, so it is difficult to obtain the “cause-effect” relation between event A and event B.

For the temporal relation of events, the event chain is constructed according to the shortest dependent path [3] between events or the sequence of events in the text, and then, the event chain [4] is optimized to distinguish the temporal relation. In example 3 shown in Figure 2, by analyzing the sequence of events in the text, it is clear that there is a “before” relation between event A and event B and a “before” relation between

Example 1:	The burst has been caused by water hammer pressure.	
Event A: Burst	Event B: Pressure	Relation: Cause-effect (B, A)
Example 2:	The radiation from the atomic bomb explosion is a typical acute radiation	
Event A: Radiation	Event B: Bomb explosion	Relation: Cause-effect (B, A)

FIGURE 1: Examples of causal indicators.

Example 3:	People have predicted his demise so many times, and the US has tried to hasten it on several occasions. Time and again, he endures.	
Event A: Predicted	Event B: Hasten	Event C: Endures
Time series: A, B, C		
Example 4:	Mr.Erdogan accepted the Israeli apology, the prime minister's office said. Mr.Erdogan has long sought an apology for the raid in May 2010 on the Mavi Marmara, which was part of a flotilla that sought to break Israel's blockade of Gaza.	
Event A: Accepted	Event B: Said	Event C: Sought
Time series: C, A, B		

FIGURE 2: Example of causal event chain.

event B and event C. However, there are some differences between the sequence of events in the text and the sequence of their occurrence. When the difference is large, it is difficult to optimize the event chain, resulting in the difficulty of identifying the sequence of events. In example 4 shown in Figure 2, the sequence of events in the text is “A, B, C,” but temporal relation between them is “C, before, A,” “A, before, B.” Due to the great difference between the initial event chain and the real results, the mainstream methods are difficult to completely optimize the initial event chain.

In this paper, we propose a joint event relation identification model of timing and causality. Our contributions include the following:

- (i) The fusion of context and multiscale local semantic features is used to fully mine the semantic information contained in the text, so as to provide semantic support for causal and temporal relation feature mining
- (ii) Temporal and causal relation encoders and decoders are constructed, respectively, to amplify the causal and temporal features of events in the text semantic information, and the correlation learning of timing and causality is realized by using the shared parameter strategy
- (iii) We experiment on the constructed legal domain dataset, and the experimental results show that the performance of our proposed model is better than that of the baseline model

## 2. Related Work

At present, the research of event relation identification mainly focuses on event causal and temporal relation identification. Among them, event causality identification mainly excavates the causality between ordered event pairs, and event timing identification mainly distinguishes the sequential timing and synchronous timing between events.

*2.1. Event Causality Identification.* For the study of event causality, the method based on template matching was mainly used in the early stage. Kaplan and Berry-Rogghe [5] used manual weaving rules to establish domain knowledge base and used knowledge reasoning technology to identify the causal relation between events. By combining with cue phrase and pattern matching, Khoo and Kornfilt [6] extracted causal language pattern rules for English corpus in the field of medicine and achieved good results in event causal recognition. Bethard et al. [7, 8] annotated the event timing and causality at the same time and used the manually annotated timing relation to assist the causality classifier to extract the causality between events. Mostafazade et al. [9] proposed an event semantic annotation model CaTeRS, which provides an annotation tool for the joint identification of event timing relation and causality. Mirza and Tonelli [10] combined the prediction results of event causality to make auxiliary judgment on the time sequence relation, so as to realize the correlation between event timing and causality. Using the constraints and linguistic rules between time series and causality, Ning et al. [11] transformed the joint identification task of event time series and causality into an integer linear programming problem and used deep learning technology to solve the problems existing in causality identification.

Riccomagno and Smith [12] proposed the chain event graph model, which is a discrete Bayesian network model and provides a flexible and highly scalable framework. The model can be used to express and analyze the meaning of causal hypothesis and strengthen the causal reasoning ability of the model through the interactive calculation of causal correlation generated in the basic network. Acharya and Lee [13] proposed an incremental causality network model to assist in inferring causality by learning time priority. The model infers causality by using an incremental Bayesian network called incremental hill climbing Monte Carlo. In addition, the authors also propose a two-layer causal network, which can realize the causal analysis of event flow without prior knowledge.

*2.2. Event Timing Relation Identification.* The early research on the temporal relation of events paid more attention to the various semantic features contained in the text itself. Marcu and Echiabi [14] paired words in order and took it as a feature of temporal relation to realize the discovery of temporal relation. With the establishment and development of TimeML (Time Markup Language) tagging system and the emergence of time series corpora such as TimeBank, more and more researchers began to extract event temporal relations from high-quality time series corpora such as

TimeBank. Mani et al. [15] used event attributes to construct feature vectors based on TimeBank labeled corpus, including event type, posture, shape, polarity, and tense, and used maximum entropy classifier to identify temporal relations. On the basis of Chambers et al. [16], Mani et al. [17] further combined semantic features such as part of speech and syntactic tree structure and extracted lexical and morphological features from WordNet, so as to greatly expand the feature space, which is conducive to the classifier to fully learn the temporal features between events.

In recent years, the global optimization method based on graph model has been widely used in many tasks, such as event identification and event timing relation identification. Chambers and Jurafsky [18] used integer linear programming method to improve the experimental performance on English temporal relation corpus. Li et al. [19] mined multiple document-level constraints derived from Chinese event semantics and used the integer linear programming method to globally optimize the classifier results, which significantly improves the recognition performance of event timing relation in Chinese text. Xu et al. [4] proposed an event timeline framework based on joint reasoning; that is, the events in the article form a complete event chain according to the order of their occurrence, then used the integer linear programming model to optimize the event chain, and add the event homonymy information to the model, which further improves the recognition ability of the model to the temporal relation.

The existing event causality identification methods mainly focus on explicit causality. However, due to the lack of explicit causality indicators in some texts, the model cannot accurately obtain the causal semantic features between events and identify the implicit causality in the text. For the identification of event timing relation, the existing research mainly constructs the event chain through the dependent path between events or the sequence of events in the text and then optimizes the event chain through global reasoning, integer linear programming, and other methods, so as to distinguish the event sequence and synchronous timing. However, the sequence of events in some texts is quite different from that in the text. When the generated event chain is different from the beginning and end nodes of the real time chain and the intermediate nodes are also misplaced, the existing methods can only optimize some nodes of the event chain. It is difficult to optimize the nodes with a large span, such as the beginning and end nodes, resulting in the model that cannot accurately distinguish the sequence of events and synchronous timing relation in the text.

### 3. Model

We propose a joint event relation identification model based on multiscale CNNs and sharing strategy. The overall architecture is shown in Figure 3. Firstly, the initial semantic representation of the text is obtained by BERT [20]. The context semantic features and multiscale local semantic features of the text are obtained through Bi-GRU [21] and multiscale convolution neural network [22], respectively. The multiscale CNN obtains the local semantic features of the text

with different granularity by setting different convolution kernel sizes. Then, the context semantic features and multiscale local semantic features are fused to fully obtain the rich semantic information in the text. Based on the fused semantic information, encoders and decoders of event causality and temporal relation are constructed, respectively, to amplify the causal and temporal features implied in the semantic features of the text itself, and three different shared parameter strategies are used to realize the correlation between causal and temporal features, so that temporal and causal relations can provide additional semantic information for each other's accurate identification. Finally, the event relation classifier is used to recognize the event causality and temporal relation.

*3.1. Context and Multiscale Local Semantic Feature.* For text context semantic feature acquisition, we use Bi-GRU to extract text features and obtain text context semantic features through forward and backward GRU networks, respectively. The specific calculation method of GRU network semantic status update is as follows:

$$\begin{aligned} z_t &= \sigma(W_z \bullet [h_{t-1}, x_t]), \\ r_t &= \sigma(W_r \bullet [h_{t-1}, x_t]), \\ \tilde{h}_t &= \tanh(W \bullet [r_t * h_{t-1}, x_t]), \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t, \end{aligned} \quad (1)$$

where  $h_{t-1}$  is the contextual semantic information of the  $t - 1$  word in the text,  $x_t$  is the initial semantic representation of the  $t$ th word in the text,  $h_t$  is the contextual semantic information of the  $t$ th word in the text, and  $\sigma$  is the activation function.  $z_t$  is the update door,  $r_t$  is the reset door, and  $W_z$  and  $W_r$  is the weight calculated by the two gates, respectively. For the update gate, when its value is larger, it means that the more text context semantic information is retained at present, and the less text context semantic information is retained in the previous sequence step. For the reset gate, the smaller its value is, the more context semantic information of the previous sequence step will be discarded, and the more semantic features of the current input word will be retained.

The text initial embedding is used as the input of Bi-GRU. The Bi-GRU network is composed of two GRU in different directions, which learn the contextual semantic features of the text from the front and back, respectively:

$$\begin{aligned} \overrightarrow{h}_n &= \overrightarrow{\text{GRU}}(\overrightarrow{h}_{n-1}, x_n, \theta_{\text{GRU}}), \\ \overleftarrow{h}_n &= \overleftarrow{\text{GRU}}(\overleftarrow{h}_{n-1}, x_n, \theta_{\text{GRU}}), \\ h_n &= \overrightarrow{h}_n \oplus \overleftarrow{h}_n, \end{aligned} \quad (2)$$

where  $\overrightarrow{h}_n$  and  $\overleftarrow{h}_n$ , respectively, represent the hidden layer semantic representation of forward and backward GRU when the sequence step size is  $n$ ;  $\theta_{\text{GRU}}$  is the network

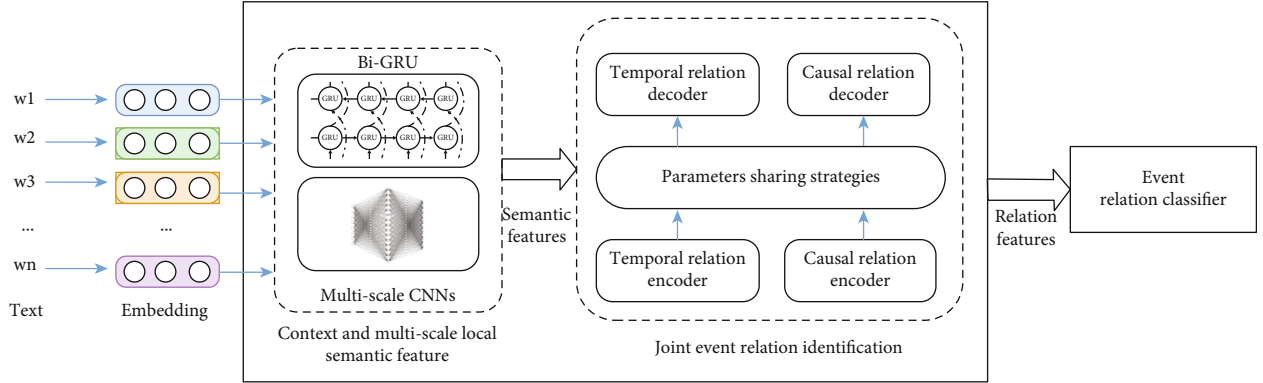


FIGURE 3: The overall architecture of the proposed joint event relation identification model.

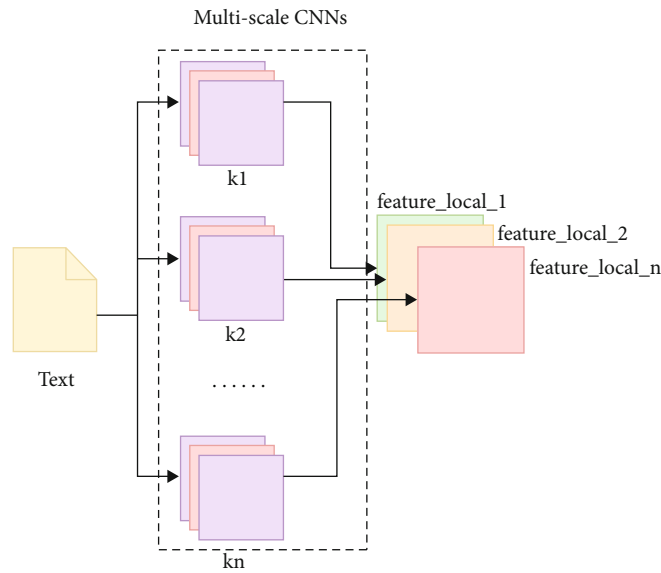


FIGURE 4: The feature learning mechanism of multiscale CNNs.

TABLE 1: Experimental results.

Model	Precision	Recall	F1
CNN-GRU-CRF	0.521	0.438	0.476
Attention-LSTM	0.493	0.469	0.481
Joint identification method	0.628	0.542	0.582
Encoder sharing	0.631	0.549	0.587
Decoder sharing	0.649	0.548	0.594
Both sharing	0.675	0.581	0.624

parameter of GRU; and  $x_n$  represents the initial semantic representation of the  $n$ th word in the text.

In order to fully mine the local semantic features of different granularity in the text, this section constructs a multiscale convolution neural network and sets different convolution kernel sizes. The feature learning mechanism of multiscale convolution neural network is shown in Figure 4. Given the text, the embedded initial semantic representation of the text is obtained through the BERT model

and used as the input of multiscale convolutional neural network. Firstly, the convolution kernel set  $K$  of multiscale convolution neural network is defined, as shown in equation (3) below, where  $k_i$  represents the number of convolution kernels and  $n$  represents the number of convolution kernels.

$$K = \{k_1, k_2, k_3, \dots, k_n\}. \quad (3)$$

The initial semantic representation of the text is input into the convolution kernel  $k_i$  to carry out convolution operation to obtain the local feature of words in the text, as shown in the following equation:

$$\text{Feature}_{\text{local}_i} = \sigma(W_i \text{embedding}_{t:t+j-1} + b_i) \quad (4)$$

where  $\text{embedding}_{t:t+j-1}$  is the embedded representation of the input word vector and  $j$  is the convolution kernel  $k_i$  window size;  $W_i$  and  $b_i$  is the weight and bias of convolution layer corresponding to different convolution kernel sizes in

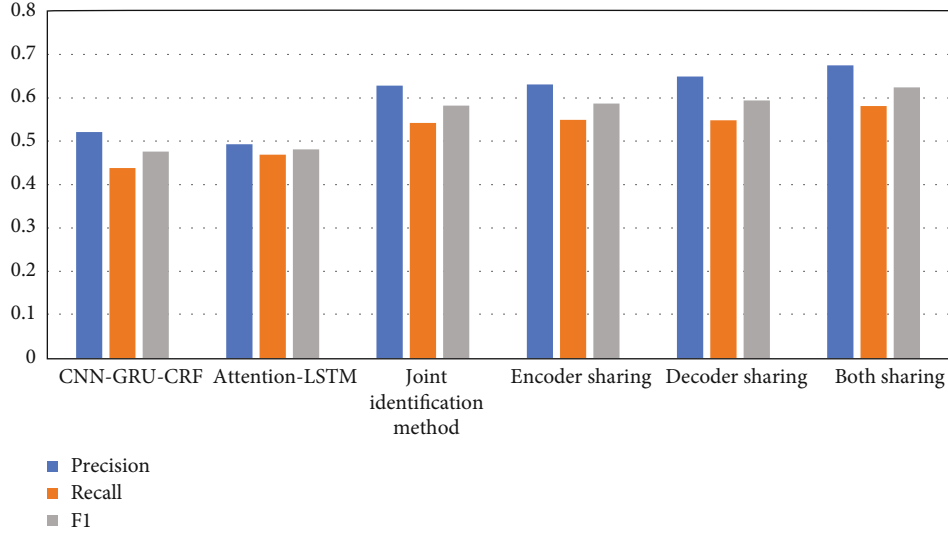


FIGURE 5: Comparison of experimental results.

TABLE 2: Ablation experimental results.

Model	Precision	Recall	F1
MCNN	0.428	0.394	0.410
GRU	0.399	0.384	0.391
MCNN-GRU	0.575	0.519	0.546
MCNN-GRU-sharing	0.642	0.573	0.606

multiscale convolution neural network, which is a learnable parameter; and  $\sigma(\cdot)$  is the activation function.

Using equation (4) above, make all convolution kernels in the convolution kernel set act on the initial semantic representation of the text, and the local semantic features of words in the text with different granularity can be obtained, as shown in the following equation:

$$\text{Feature}_{\text{local}} = \{\text{feature}_{\text{local}_1}, \text{feature}_{\text{local}_2}, \dots, \text{feature}_{\text{local}_n}\}. \quad (5)$$

Because different convolution kernels can form local semantic features of text with different granularity, if these local semantic features are directly spliced, the dimension of local semantic features of words will be too high. Therefore, this section through the local semantic  $\text{Feature}_{\text{local}}$  performs the maximum pooling operation to reduce the dimension of the text while retaining the local semantic features of different granularity, as shown in the following equation:

$$\max_{\text{local}} = \text{maxpooling}(\text{feature}_{\text{local}}). \quad (6)$$

The local features with different granularity of word vectors in the text are maximally pooled, and the output dimension is fixed through the full connection layer, and finally, the multiscale local semantic feature representation of the

central word vector is obtained, as shown in the following equation:

$$F_{\text{local}} = \{\max_{\text{local}_1}, \max_{\text{local}_2}, \dots, \max_{\text{local}_n}\}. \quad (7)$$

Repeat the convolution process of the above central word vector, and scan the whole text sequence with the convolution set  $K$  to obtain the multiscale local semantic features of the text, as shown in the following equation:

$$\text{Feature}_{\text{local}} = \{F_{\text{local}_1}, F_{\text{local}_2}, F_{\text{local}_3}, \dots, F_{\text{local}_n}\}. \quad (8)$$

**3.2. Relation Coding and Decoding.** We build encoders and decoders of event causality and temporal relation, respectively. The event relation encoder is used to learn the semantic feature representation of event timing and causality in the text, and the event relation decoder is used to correspond the learned semantic feature representation of event relation with event relation coding.

**3.2.1. Event Relation Encoder.** We use Bi-LSTM [23] as the temporal and causal semantic feature encoder. This is because LSTM is mainly used to learn long-term dependency problems, which can well model and represent the dependency in the text, and because LSTM introduces memory unit, it can automatically update and selectively forget the dependency features in the text.

Fuse the context and multiscale local semantic features obtained in Section 3.1, and input them into the Bi-LSTM network to obtain the temporal or causal features in the text. The specific calculation method is as follows:

$$\begin{aligned} \overrightarrow{\text{rel}}\mathcal{H}_n &= \overrightarrow{\text{LSTM}}(\overrightarrow{\text{rel}}\mathcal{h}_{n-1}, [h_n; F_{\text{local}}], \theta_{\text{LSTM}}), \\ \overleftarrow{\text{rel}}\mathcal{H}_n &= \overleftarrow{\text{LSTM}}(\overleftarrow{\text{rel}}\mathcal{h}_{n-1}, [h_n; F_{\text{local}}], \theta_{\text{LSTM}}), \\ \text{rel}\mathcal{H}_n &= \overrightarrow{\text{rel}}\mathcal{H}_n \oplus \overleftarrow{\text{rel}}\mathcal{H}_n, \end{aligned} \quad (9)$$



where  $\overrightarrow{\text{rel}}_n$  and  $\overleftarrow{\text{rel}}_n$ , respectively, represent the hidden layer representation of event relation semantics in the text of forward and backward LSTM at time  $n$ ,  $\theta_{\text{LSTM}}$  is the network parameter of LSTM, and  $[h_n; F_{\text{local}}]$  indicates the splicing and fusion of context semantic features and multiscale local semantic features of text.

**3.2.2. Event Relation Decoder.** The decoder is used to apply the learned event relation semantic feature  $\text{rel}_n$  converted into the hidden layer representation of event relation, which can be simply expressed by the following equation:

$$\text{rel} = \text{decoder}(\text{rel}_n). \quad (10)$$

We use a full connection layer as the event relation decoder. As shown in equation (11) below, we map the event relation semantic features learned by the encoder to the event relation label representation space, so as to establish association with the event relation label, where  $W$  and  $B$  are the weight and offset of the full connection layer, respectively, which are learnable parameters, and  $\sigma(\cdot)$  is the activation function.

$$\text{rel} = \sigma(W\text{rel}_n + b). \quad (11)$$

**3.3. Joint Event Relation Identification.** Considering the correlation between timing and causality, there is often time series between events with causality, while there must be no causality between events with synchronous time series. Therefore, we take temporal relation identification and causality identification as two subtasks:  $\text{task}_{\text{time}}$  and  $\text{task}_{\text{cause}}$ . There is no sequence between the two subtasks, but they are carried out at the same time. For the interaction between the two subtasks, we choose three different parameter sharing strategies: (1) sharing the coding layer, (2) shared the decoding layer, and (3) share the encoding layer and decoding layer.

**3.3.1. Sharing the Coding Layer.** Firstly, the event temporal relation identification task and the causal relation identification task, respectively, use their respective event relation encoders to process the semantic features of the text and obtain their respective event relation semantic features, as shown in equations (12) and (13) below:

$$\text{rel}_{\text{time}} = \text{Encoder}_{\text{time}}([h_n; F_{\text{local}}]), \quad (12)$$

$$\text{rel}_{\text{cause}} = \text{Encoder}_{\text{cause}}([h_n; F_{\text{local}}]), \quad (13)$$

where the event relation encoder is the Bi-LSTM encoder introduced in Section 3.2,  $h_n$  is context semantic feature, and  $F_{\text{local}}$  is a multiscale local semantic feature. Then, the two subtasks share their own event relation coding layer states, respectively, and splice them with their own coding layer states to generate joint relation semantic features, as shown in equations (14) and (15) below.

$$\text{rel}'_{\text{time}} = \text{rel}_{\text{time}} \oplus \text{rel}_{\text{cause}}, \quad (14)$$

$$\text{rel}'_{\text{cause}} = \text{rel}_{\text{cause}} \oplus \text{rel}_{\text{time}}. \quad (15)$$

Finally, the semantic features of event joint relation are decoded by using their respective event relation decoders to obtain the decoded event relation representation.

**3.3.2. Shared the Decoding Layer.** Similarly, the event temporal relation encoder and causality encoder are used to process the semantic features of the text to obtain their respective event relation semantic features  $\text{rel}_{\text{time}}$  and  $\text{rel}_{\text{cause}}$ , as shown in equations (12) and (13) above.

When decoding, the event timing relation and causality are decoded by their respective event relation decoders, as shown in equations (16) and (17) below.

$$d_{\text{time}} = \text{Decoder}_{\text{time}}(\text{rel}_{\text{time}}), \quad (16)$$

$$d_{\text{cause}} = \text{Decoder}_{\text{cause}}(\text{rel}_{\text{cause}}). \quad (17)$$

Then, the two subtasks share their own decoding layer states and splice them with their own decoding layer states, as shown in equations (18) and (19) below.

$$d'_{\text{time}} = d_{\text{time}} \oplus d_{\text{cause}}, \quad (18)$$

$$d'_{\text{cause}} = d_{\text{cause}} \oplus d_{\text{time}}. \quad (19)$$

In the final classification of event relation, the full connection layer is used to predict the event relation of the spliced decoding layer state, as shown in equations (20) and (21) below, where  $W_{\text{time}}$ ,  $b_{\text{time}}$  and  $W_{\text{cause}}$ ,  $b_{\text{cause}}$  are the weight and offset of the full connection layer of the two subtasks, respectively, and  $\sigma(\cdot)$  is the activation function.

$$d'_{\text{time}} = d_{\text{time}} \oplus d_{\text{cause}}, \quad (20)$$

$$d'_{\text{cause}} = d_{\text{cause}} \oplus d_{\text{time}}. \quad (21)$$

**3.3.3. Share the Encoding Layer and Decoding Layer.** After passing through the respective event relation encoders, the semantic feature representation of the respective event relation is obtained. The two subtasks share their respective coding layer states and splice them to generate the semantic representation of joint relation  $\text{rel}'_{\text{time}}$  and  $\text{rel}'_{\text{cause}}$ , as shown in equations (14) and (15) above.

Then, each event relation decoder is used for decoding to obtain the decoded event relation hidden layer representation  $d_{\text{time}}$  and  $d_{\text{cause}}$ . Then, the two subtasks share their own decoding layer states and splice them with their own decoding layer states, as shown in equations (18) and (19) above.

Similarly, when classifying the event relation, the full connection layer is used to predict the event relation of the spliced decoding layer state, as shown in equations (20) and (21) above.

## 4. Experiments

In this part, we give the experimental results of the proposed model. We first describe the constructed dataset. Then, we

introduce the relevant settings of the experiment and the baseline of our comparison. The experimental results show that the proposed model is improved in the joint identification of event timing and causality.

**4.1. Dataset.** According to the needs of event sequence and causality extraction task and the relation between events which has directionality, it is necessary to mark the head event and tail event of the event relation in the text. We use “< E1 > event\_1 < / E1 >” to represent the head event, use “< E2 > event\_2 < /E2 >” to represent the tail event, and use “rel” to represent the relation between the head event “event\_1” and tail event “event\_2”. For event relations, we define the following relation types: before, after, meanwhile, cause-effect, effect-cause, and other.

**4.2. Experimental Settings.** The hyperparameters of the model we use are set as follows: in the initial vectorization representation stage, set the word vector dimension output by the BERT pretraining language model to 762. In the stage of using Bi-GRU to obtain the semantic features of text context, set the number of layers of Bi-GRU to 2 and its dimension to 512. The multiscale convolution neural network is used to obtain the multiscale local semantic features of the text, and the convolution set  $K$  is set as

$$K = \{5,10,15,20\}. \quad (22)$$

In the event timing and causality feature coding stage, set the number of layers of Bi-LSTM to 2 and its dimension to 512. In the event timing and causality feature decoding stage, set the number of network layers of the full connection layer to 1 and its dimension to 256. In the event timing and causality prediction stage, set the number of network layers of the full connection layer to 1, and its dimension is the total number of event relation labels 7. In all parts using activation functions, except that the last step of event relation prediction uses softmax activation function, the rest uses relu activation function. The normal distribution with standard deviation  $N(0,0.01)$  is adopted for all parameters in the model to initialize the parameters. The batch size during training is 16. The back propagation algorithm is used for learning, and the Adam optimizer is used for optimization training. LR is set to 0.0001.

**4.3. Baseline and Evaluation Metrics.** In order to verify the effectiveness of our proposed model, three comparison models are selected for comparison:

**CNN-GRU-CRF:** Zheng proposed an event causality identification model based on double-layer CNN-GRU-CRF, which regards event causality identification as a sequence annotation task. In this method, CNN and Bi-GRU are used to obtain the local semantic features and contextual features of the text, respectively, and fuse them. Then, CRF is used to obtain the dependency rules between event relation tags, determine the final prediction tag sequence, and complete the identification of event causality.

**Attention-LSTM:** Zhang proposed an event timing relation recognition model combining self-attention mechanism

and neural network. Taking the shortest dependent path sequence of event sentences as the input of the model, first use the nonlinear sublayer (CNN or RNN) to preliminarily semantically encode the input of the model, then use the self-attention network layer to capture the global information in the output of the nonlinear layer, and finally use a softmax layer to classify the event timing relation.

**Joint identification method:** Zhang proposed a joint identification model of event timing and causality based on neural network. This method takes the dependent path sequence as the input of the model, takes the event timing identification as the main task, and takes the event causality identification as the auxiliary task. The correlation between causality and temporal relation is realized through parameter sharing. Finally, the relation classifier is used to classify timing and causality, respectively.

In addition, since we proposed three different parameter sharing strategies in the joint event relation identification stage, we also carried out experiments on three different parameter sharing strategies.

**4.4. Quantitative Results.** For the task of event timing and causality identification in the text, the accuracy, recall, and F1 are used to evaluate the performance of the model. Table 1 shows the specific values of the experimental results of the comparison model and the three models based on the parameter sharing strategy. Figure 5 more vividly shows the accuracy, recall, and F1 of the event relation extraction by different models. It can be seen that the model proposed in this paper has been significantly improved. It can be seen from the data in the table that compared with the comparison model, the event relation joint identification model based on multiscale convolution neural network and sharing strategy has improved in various evaluation indexes. By comparing the experimental data of three shared parameter models, it can be found that the performance of the model based on shared coding and decoding layer is relatively better, while the performance of the model only shared coding layer is relatively weak.

By comparing with two single event relation identification models (CNN-GRU-CRF and attention-LSTM model in the table), it is found that the joint event relation identification model based on multiscale convolution neural network and sharing strategy proposed in this paper considers the correlation between event timing and causal relation, integrates event timing characteristics with event causal characteristics, and learns the correlation characteristics, which makes the model perform better in implicit causality identification and sequential, synchronous, and temporal differentiation tasks. Compared with the joint identification model (joint identification method in the table), it is found that although this method also considers the correlation between event timing and causality, it does not well model and represent the semantic features of the text itself in the previous feature construction stage, resulting in the lack of some semantic features of the text, so the performance of this model is relatively poor. Through the experimental results of the three proposed parameter sharing models, it is found that the strategy of sharing coding and decoding

layer comprehensively considers the different semantic information of event relation semantic features in the coding stage and decoding stage and integrates the event relation features, so it can well correlate the event timing relation and causality and play the role of auxiliary prediction.

In addition, in order to highlight the importance of each module in our model, we also conducted ablation experiments. The experimental results are shown in Table 2.

The number of layers and parameter settings of each model in the table are the same as those of the previous experiments. MCNN indicates that only multiscale local semantic features of text are obtained. GRU indicates that only context semantic features are obtained. MCNN-GRU acquires both features. MCNN-GRU-sharing also adds a parameter sharing policy. From the experimental results, it can be seen that compared with the context semantic features, the multiscale local semantic features obtained by MCNN are more important for the identification of event relations. By combining the two semantic features and fully modeling the semantic features of the text, the accuracy of event relation recognition can be further improved. The introduction of sharing strategy can fully consider the correlation characteristics between event timing and causality and play an auxiliary role in prediction.

## 5. Conclusion

Aiming at the difficulty of identifying the implicit causal relation of events in text and distinguishing the sequential and synchronous temporal relation, this paper proposes a joint event relation identification model based on multiscale convolutional neural network and sharing strategy. The context sequence semantic features and multiscale local semantic features of the text are obtained through Bi-GRU and multiscale CNNs, respectively, and then, the context sequence semantic features and multiscale local semantic features are fused to make full use of the rich semantic information in the text. Then, the coders and decoders of causality and temporal relation are constructed, respectively, to amplify the causality and temporal features implied in the semantic features of the text itself. In addition, we use three different shared parameter strategies to realize the correlation between temporal features and causal features, so that temporal features and causal relations can provide additional semantic features for each other's accurate prediction. We evaluated the proposed method on the constructed dataset. The results show that the performance of this method is better than that of the previous methods in the joint identification of event timing and causality.

## Data Availability

The data that support the findings of this study are available from the first author upon reasonable request.

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

## Acknowledgments

This work was supported by the National Key Research and Development Program of China (no. 2021YFC3300602).

## References

- [1] Z. Li, Q. Li, X. Zou, and J. Ren, "Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings," *Neurocomputing*, vol. 423, pp. 207–219, 2021.
- [2] S. Zhao, T. Liu, S. Zhao, Y. Chen, and J. Y. Nie, "Event causality extraction based on connectives analysis," *Neurocomputing*, vol. 173, pp. 1943–1950, 2016.
- [3] Q. Do, W. Lu, and R. Dan, "Joint inference for event timeline construction," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 677–687, 2012.
- [4] Y. Xu, R. Jia, and L. L. Mou, "Improved relation classification by deep recurrent neural networks with data augmentation," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1461–1470, 2016.
- [5] R. M. Kaplan and G. Berry-Rogghe, "Knowledge-based acquisition of causal relationships in text," *Knowledge Acquisition*, vol. 3, no. 3, pp. 317–337, 1991.
- [6] C. S. G. Khoo and J. Kornfilt, "Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing," *Literary and Linguistic Computing*, vol. 13, no. 4, pp. 177–186, 1998.
- [7] S. Bethard, W. Corvey, and S. Klingenstein, "Building a corpus of temporal-causal structure," in *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp. 908–915, Marrakech, Morocco: European Language Resources Association, 2008.
- [8] S. Bethard and J. H. Martin, "Learning semantic links from a corpus of parallel temporal and causal relations," in *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*, pp. 177–180, Columbus, Ohio, USA: Association for Computational Linguistics, 2008.
- [9] N. Mostafazadeh, A. Grealish, and N. Chambers, "CaTeRS: causal and temporal relation scheme for semantic annotation of event structures," in *Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation*, pp. 51–61, San Diego, California: Association for Computational Linguistics, 2016.
- [10] P. Mirza and S. Tonelli, "CATENA: causal and temporal relation identification from natural language texts," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 64–75, Osaka, Japan: Association for Computational Linguistics, 2016.
- [11] Q. Ning, Z. Feng, and H. Wu, "Joint reasoning for temporal and causal relations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 2278–2288, Melbourne, Australia: Association for Computational Linguistics, 2018.
- [12] E. Riccomagno and J. Q. Smith, "The causal manipulation of chain event graphs," 2007, preprint arXiv: 0709.3380.
- [13] S. Acharya and B. S. Lee, "Incremental causal network construction over event streams," *Information Sciences*, vol. 261, pp. 32–51, 2014.
- [14] D. Marcu and A. Echihiabi, "An unsupervised approach to recognizing discourse relations," in *Proceedings of the 40th*



- Annual Meeting of the Association for Computational Linguistic*, pp. 368–375, Philadelphia, Pennsylvania, USA, 2002.
- [15] I. Mani, M. Verhagen, B. Wellner et al., “Machine learning of temporal relations,” in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 753–760, Sydney, Australia, 2006.
  - [16] N. Chambers, S. Wang, and D. Jurafsky, “Classifying temporal relations between events,” in *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 173–176, Prague, Czech Republic, 2007.
  - [17] I. Mani, B. Wellner, and M. Verhagen, “Three approaches to learning TLINKS in time ML,” *Computer Science Department*, 2007.
  - [18] N. Chambers and D. Jurafsky, “Jointly combining implicit constraints improves temporal ordering,” in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 698–706, Honolulu, Hawaii, USA, 2008.
  - [19] P. Li, Q. Zhu, and G. Zhou, “Global inference to Chinese temporal relation extraction,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1451–1460, Osaka, Japan, 2016.
  - [20] K. Cho, B. Van Merriënboer, and C. Gulcehre, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, 2014.
  - [21] J. Chung, C. Gulcehre, and K. H. Cho, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, preprint arXiv:1412.3555.
  - [22] Y. Wang, D. Ning, and S. Feng, “A novel capsule network based on wide convolution and multi-scale convolution for fault diagnosis,” *Applied Sciences*, vol. 10, no. 10, p. 3659, 2020.
  - [23] B. A. Pearlmutter, “Learning state space trajectories in recurrent neural networks,” *Neural Computation*, vol. 1, no. 2, pp. 263–269, 1989.