Contents lists available at ScienceDirect

Knowledge-Based Systems

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

KEPT: Knowledge Enhanced Prompt Tuning for event causality identification

Jintao Liu [a,b,c,d], Zequn Zhang [a,b,*], Zhi Guo [a,b], Li Jin [a,b], Xiaoyu Li [a,b], Kaiwen Wei [a,b,c,d], Xian Sun [a,b,c]

[a] *Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China*
[b] *Key Laboratory of Network Information System Technology(NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China*
[c] *University of Chinese Academy of Sciences, Beijing 100190, China*
[d] *School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China*

ARTICLE INFO

Article history:
Received 31 May 2022
Received in revised form 18 October 2022
Accepted 20 October 2022
Available online 2 November 2022

MSC:
00-01
99-00

Keywords:
Event causality identification
Prompt tuning
Attention mechanism
Knowledge representation learning

ABSTRACT

Event causality identification (ECI) aims to identify causal relations of event mention pairs in text. Despite achieving certain accomplishments, existing methods are still not effective due to the following two issues: (1) the lack of causal reasoning ability, imposing restrictions on recognizing implicit causal relations; (2) the significant gap between fine-tuning and pre-training, which hinders the utilization of pre-trained language models (PLMs). In this paper, we propose a novel **K**nowledge **E**nhanced **P**rompt **T**uning (KEPT) framework for ECI to address the issues mentioned above. Specifically, this method leverages prompt tuning to incorporate two kinds of knowledge obtained from external knowledge bases (KBs), including background information and relational information, for causal reasoning. To introduce external knowledge into our model, we first convert it to textual descriptions, then design an interactive attention mechanism and a selective attention mechanism to fuse background information and relational information, respectively. In addition, to further capture implicit relations between events, we adopt the objective from knowledge representation learning to jointly optimize the representations of causal relations and events. Experiment results on two widely-used benchmarks demonstrate that the proposed method outperforms the state-of-the-art models.

1. Introduction

Event causality identification (ECI) is an important yet challenging language understanding task that requires identifying causal relations of event pairs in text. For example, given the sentence "Many buildings were *destroyed* in the sudden *earthquake*"., an ECI system should have the ability to recognize the causal relations between the mentioned events, i.e., *earthquake* $\xrightarrow{cause}$ *destroyed*. This technique exhibits a wide range of application value in natural language processing, including machine reading comprehension [1], process extraction [2], future event forecasting [3,4], and why-question answering [5–7], etc.

Many studies have concentrated on this task, ranging from early feature-based methods [4,8–10] to the recent deep learning approaches [11–13]. Despite achieving promising performance, current methods are limited in that they mainly adopt a fine-tuning paradigm for the ECI task, which has a significant gap with pre-training, restricting the use of task-related knowledge in pre-trained language models [14]. As shown in Fig. 1, the goal of pre-training is usually to predict the masked words in cloze-style form, while fine-tuning for the ECI task is formalized as a classification task with additional fully connected layers. This significant difference may prevent models from reaching their full potential. To this end, prompt tuning has emerged to narrow the gap of objective forms between pre-training and fine-tuning [15–17]. With appropriate task-specific prompt templates and label words, prompt tuning could control the model behavior to adapt to the task.

Although employing prompt tuning could promote the use of task-related knowledge, the model still struggles in recognizing causality, especially in the case of implicit causal relations. For instance, in the sentence given above, there is no explicit causal indicator in the sentence, imposing challenges to predicting the causal relation between *earthquake* and *destroyed*. Recently, several works have pointed out that leveraging useful information from external knowledge bases (e.g., ConceptNet [18]) can benefit causal reasoning. Their main motivation is to make the model recognize causality based on prior knowledge like a human. Liu

[*] Corresponding author at: Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: zqzhang1@mail.ie.ac.cn (Z. Zhang).

https://doi.org/10.1016/j.knosys.2022.110064
0950-7051/© 2022 Elsevier B.V. All rights reserved.

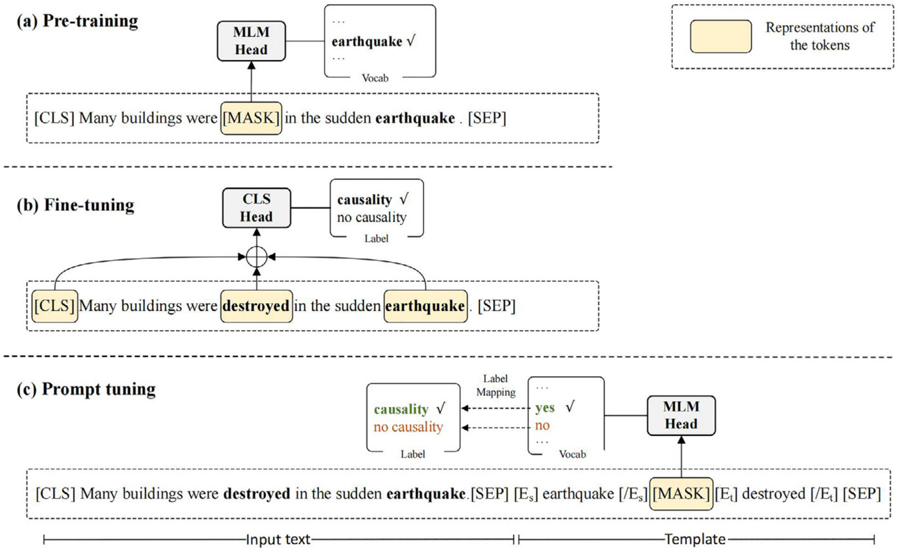
(a) Pre-training
MLM Head
...
earthquake √
...
Vocab
[CLS] Many buildings were [MASK] in the sudden earthquake . [SEP]
Representations of the tokens
(b) Fine-tuning
CLS Head
causality √
no causality
Label
[CLS] Many buildings were destroyed in the sudden earthquake . [SEP]
(c) Prompt tuning
Label Mapping
causality √
no causality
Label
...
yes √
no
...
Vocab
MLM Head
[CLS] Many buildings were destroyed in the sudden earthquake.[SEP] [Es] earthquake [/Es] [MASK] [Et] destroyed [/Et] [SEP]
Input text
Template

Fig. 1. An illustration of (a) masked language model (MLM) pre-training, (b) fine-tuning for event causality identification, and (c) standard prompt tuning for event causality identification. The bold words in the sentence are mentioned events.

et al. [11] proposed an approach that retrieves related triples of the event and integrates them with the text. This idea enriches the background information of the event, but it ignores relational information and is prone to introduce noise. Cao et al. [13] proposed a Graph-based network to model descriptive knowledge and relational knowledge, which has achieved satisfactory results. However, it only uses the information of events on the relational path for causal reasoning and underutilizes the relational information. Therefore, we seek to improve causal reasoning ability by introducing background information for the event and relational information between events at the same time. On the one hand, the background information can make the model understand the properties of each event, contributing meaningful knowledge for the ECI task; on the other hand, the relational information can help extract causal clues between the events, which provides evidence for causal reasoning effectively.

Recently, several methods have been proposed to use the idea of knowledge representation learning (e.g., the translation-based method) to capture relations between entities, such as TME [19] and TransRel [20]. They feed the input text to the encoder to obtain the representations of the entities from corresponding positions and then directly predict relations by translation from head entity to tail entity. Chen et al. [21] proposed KnowPrompt for relation extraction, which utilizes knowledge constraints to synergistically optimize the representations of template words and answer words. Nevertheless, these methods only use entity embeddings to predict relations and rarely consider background information of the entity, thus hard to capture implicit relations between entities. Based on this observation, specific to the ECI task, we expect to incorporate background information into the event and jointly learn the representation of event and causal relation.

In this paper, we propose a novel **K**nowledge **E**nhanced **P**rompt **T**uning (KEPT) framework for ECI to address the issues mentioned above. Specifically, our method has the following three main steps: (1) We first obtain background information of the event and relational information between the events from ConceptNet. Take Fig. 2 as an example, given the mentioned event pair (i.e., *earthquake* and *destroyed*), we can retrieve related triples for *earthquake* and *destroyed*, and the triples that on the relational path (i.e., the red edges) between *earthquake* and *destroyed*. Then we transform these triples into textual descriptions, forming background information of the event and relational information between the events. (2) To mitigate the influence of irrelevant knowledge on text, we design a selective attention module to fuse relational information. And we utilize prompt tuning for the ECI task by creating prompt templates based on the corresponding event pairs and identifying causal relations according to the predicted answer words. (3) To capture potentially important features from external knowledge, we propose an interactive attention module to inject background information into the event. Then we borrow the idea from knowledge representation learning to jointly optimize the representation of causal relation and event, which can obtain a more expressive causal relation representation and assist in extracting implicit relation.

In experiments, we compare our KEPT with other baseline models on two widely-used benchmarks, including EventStoryLine and Causal-TimeBank. The results illustrate that our method achieves state-of-the-art performance. Besides, our model performs well under low-resource conditions, demonstrating that KEPT can solve the data scarcity problem. Moreover, we conduct cross-topic adaption and unseen events prediction to evaluate the generalization ability of our model.

To summarize, our contributions are as follows:

- We propose a **K**nowledge **E**nhanced **P**rompt **T**uning (KEPT) method, which utilizes prompt tuning to incorporate background information and relational information for causal reasoning. To the best of our knowledge, this is the first work to adopt prompt tuning for the ECI task.
- To improve causal reasoning capability of the model, we design an interactive attention mechanism and a selective attention mechanism to fuse background information and relational information, respectively.
- We adopt the objective from knowledge representation learning to jointly optimize the representations of events and causal relations, which can further capture implicit relations between events.
- Experiments on EventStoryLine and Causal-TimeBank demonstrate that KEPT outperforms all the other baseline methods and improves F1-score by a large margin, which indicates the effectiveness of our model.

The remaining sections of this paper are structured as follows. Section 2 summarizes related works. Section 3 gives preliminaries

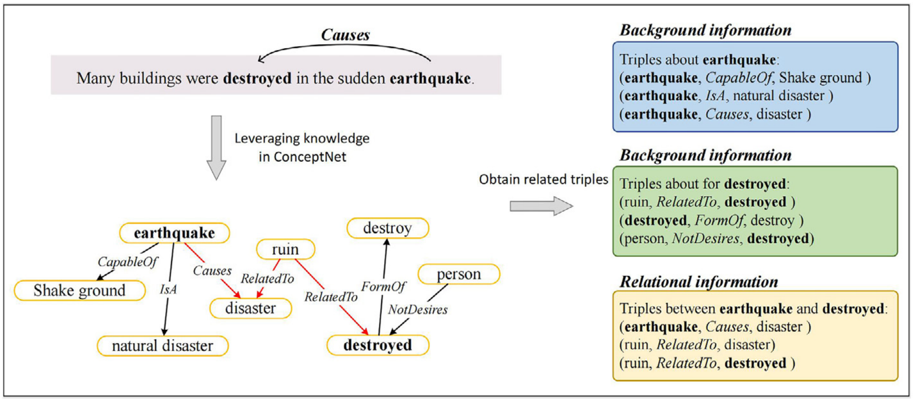
Causes
Many buildings were destroyed in the sudden earthquake.
Leveraging knowledge in ConceptNet
earthquake
CapableOf
Shake ground
IsA
natural disaster
Causes
ruin
RelatedTo
disaster
RelatedTo
destroy
FormOf
person
NotDesires
destroyed
Obtain related triples
Background information
Triples about earthquake:
(earthquake, CapableOf, Shake ground)
(earthquake, IsA, natural disaster)
(earthquake, Causes, disaster)
Background information
Triples about for destroyed:
(ruin, RelatedTo, destroyed)
(destroyed, FormOf, destroy)
(person, NotDesires, destroyed)
Relational information
Triples between earthquake and destroyed:
(earthquake, Causes, disaster)
(ruin, RelatedTo, disaster)
(ruin, RelatedTo, destroyed)

Fig. 2. An example of leveraging background information and relational information in ConceptNet for the ECI task. The bold words in the sentence denote candidate events. The red lines represent the relational path between "earthquake" and "destroyed".

about fine-tuning and prompt tuning of PLMs. Section 4 describes the architecture of our model in detail. Section 5 shows the experiment setup, experiment results, analysis, and discussion. Section 6 is the conclusion of this paper.

2. Related work

2.1. Event causality identification

The ECI task aims to extract causal relations between event pairs in text and has attracted increasing attention. For datasets, a few works have released benchmarks for event causality identification. Mirza et al. [22] proposed Causal-TimeBank corpus by annotating causal relations in TempEval-3 corpus. Mostafazadeh et al. [23] constructed a corpus with temporal and causal relations annotated. Caselli et al. [24] raised the EventStoryLine corpus based on the former annotated dataset [23]. Early methods for ECI mainly focus on syntactic and lexical features [10, 25–27], causality patterns [4,28], statistical causal clues [8,29–31], temporal patterns [9,32]. Recently, deep learning methods have been adopted to solve ECI problems [33–36] and achieved state-of-the-art performance. Gao et al. [10] identified causal relations between events by modeling document-level structures. Zuo et al. [37] exploited distantly supervised labeled data to achieve data augmentation. Liu et al. [11] proposed a BERT-based model with mention masking generalization and external knowledge. Zuo et al. [12] released a learnable knowledge-guided data augmentation method, which leverages a dual learning framework to iteratively generate new samples and identify causality. Cao et al. [13] utilized external structural knowledge to construct a descriptive graph and relational graph for recognizing event causality. Nevertheless, existing works mainly solve the ECI task in a fine-tuning paradigm and have difficulty in dealing with implicit causal relations.

2.2. Prompt tuning

Prompt tuning has been proposed to bridge the gap between objective forms in pre-training and fine-tuning. This paradigm is inspired by GPT-3 [38] and achieves superior performance on many mainstream NLP tasks, including text classification [16, 39], sentiment analysis [40], entity typing [41], information extraction [42–44], and so on. Early methods adopted artificially designed prompt templates [45], which built the foundation for the development of prompt tuning. Schick and Schütze [16] proposed a PET approach that used manually designed templates for text classification under a semi-supervised setting. Hu et al. [46] focused on incorporating external knowledge into the verbalizer to improve and stabilize prompt tuning. Later, a series of works studied automatic discrete prompts generation [17,47] to avoid labor-intensive prompt design. Gao et al. [48] first proposed the automatic generation of prompt templates and answer words. Jiang et al. [15] proposed a method based on mining and paraphrasing to generate high-quality and diverse prompt templates automatically. In addition, continuous prompts [49,50] have been proposed, which use randomly initialized embeddings to replace discrete embeddings. Li and Liang [51] fixed the parameters of pre-trained language models and optimized continuous task-specific embeddings. Liu et al. [52] proposed a P-tuning method that applied trainable continuous prompt vectors and performed better than fine-tuning. To the best of our knowledge, no method has introduced prompt tuning to the ECI task.

2.3. Knowledge representation learning

Many works have been conducted to represent real-world entities and relations in a low-dimensional space. Bordes et al. [53] proposed the translation-based model TransE, which learns the representation of the relation and entity by treating the relation as the translation from head entity to tail entity. TransH [54], TransR [55], and TransD [56] break the translation-based limitation and can model complex relations. ConvE [57] and ConvKB [58] utilize convolutional neural network (CNN) to extract important features. Besides, there is a line of works that employ language models for knowledge representation learning and achieve significant performance [59–61]. Moreover, graph-based methods are also adopted to incorporate entities and relations in heterogeneous graphs [62,63]. In this work, we leverage the idea of TransE to jointly optimize the representations of event and causal relation.

3. Preliminaries

In this section, we give some important preliminaries about fine-tuning and prompt tuning before introducing our method. A classification dataset can be represented as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X}$ is the set of input samples and $\mathcal{Y}$ is the set of classification labels. Each input sample $x_{in} \in \mathcal{X}$ consists of some tokens $x_{in} = \{w_1, w_2, \ldots, w_n\}$ and corresponds to a label $y \in \mathcal{Y}$.

3.1. Fine-tuning of PLMs

For classification task, given a pre-trained language model $\mathcal{M}$, standard fine-tuning approach first converts the input sample $x_{in} = \{w_1, w_2, \ldots, w_n\}$ into a sequence formalized as $[CLS] \oplus$

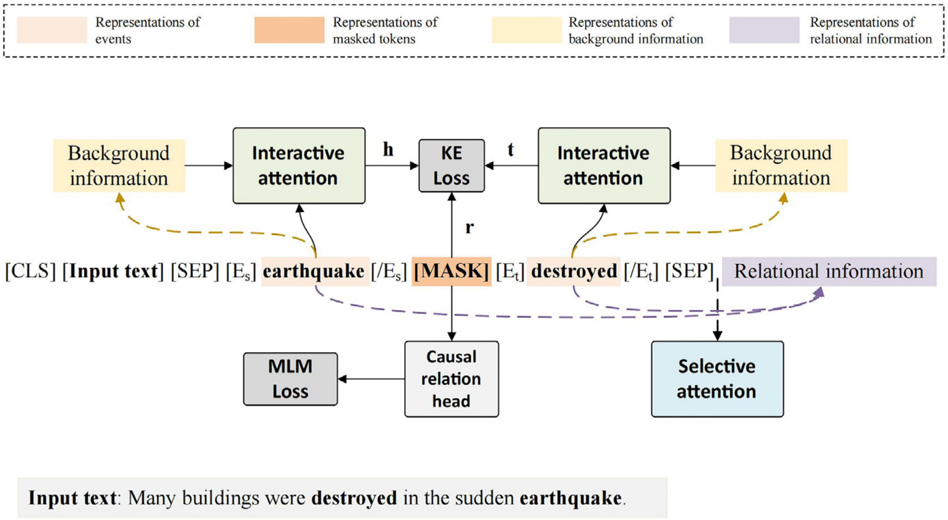
Representations of events
Representations of masked tokens
Representations of background information
Representations of relational information
Background information
Interactive attention
h
KE Loss
t
Interactive attention
Background information
r
[CLS] [Input text] [SEP] [E$_s$] earthquake [/E$_s$] [MASK] [E$_t$] destroyed [/E$_t$] [SEP]
Relational information
MLM Loss
Causal relation head
Selective attention
Input text: Many buildings were destroyed in the sudden earthquake.

Fig. 3. The overview of our proposed KEPT for event causality identification. In the example, the golden dashed line represents obtaining background information for the event and the purple dashed line represents obtaining relational information between the event pair. The input sequence is the concatenation of input text, template, and relational information. The input sequence and background information are encoded with the pre-trained language model, which is not drawn for the convenience of display. The relational information is incorporated into the model via the selective attention mechanism. The background information is injected into the event through the interactive attention mechanism.

$x_{in} \oplus [SEP]$. Here $[CLS]$ and $[SEP]$ are special tokens. $[CLS]$ is added in front of every input sequence, which is often used for downstream classification tasks. $[SEP]$ is utilized to separate different sequences. Then $\mathcal{M}$ is utilized to encode the sequence into contextualized hidden vectors $H = \{H_{[CLS]}, H_{w_1}, H_{w_2}, \ldots, H_{w_n}, H_{[SEP]}\}$. We can obtain the probability distribution over the label set $\mathcal{Y}$ by training a classification head, $p(y|x_{in}) = softmax(W_o H_{[CLS]})$, where W_o is a randomly initialized parameter matrix, $H_{[CLS]}$ is the representation of $[CLS]$. W_o and $\mathcal{M}$ are optimized by minimizing the cross entropy loss over $p(y|x_{in})$ on $\mathcal{X}$.

3.2. Prompt tuning of PLMs

Given a pre-trained language model $\mathcal{M}$, we first construct a prompt template x_t including $[MASK]$, and a set of answer words $\mathcal{V}$. The prompt template is connected with the sample x_{in} to form the input sequence x_s, and $[MASK]$ marks the masked position for $\mathcal{M}$ to predict. Besides, a verbalizer $v(\cdot) : \mathcal{Y} \rightarrow \mathcal{V}$ is built to map class label $y \in \mathcal{Y}$ to answer word $v(y) \in \mathcal{V}$. We can obtain probability distribution over $\mathcal{Y}$ as follows:

$$p(y|x_{in}) = p([MASK] = v(y)|x_s) = \frac{exp(W_{v(y)} \cdot H_{[MASK]})}{\sum_{y' \in \mathcal{Y}} exp(W_{v(y')} \cdot H_{[MASK]})} \quad (1)$$

where $W_{v(\cdot)}$ reuses the weights in pre-training process and $H_{[MASK]}$ is the representation of $[MASK]$. For example, in a binary sentiment classification task, we construct a template $x_t = It\ was\ [MASK]$ and set $v(y = positive) \longrightarrow great$ and $v(y = negative) \longrightarrow terrible$. The input sequence can be formalized as $x_s = [CLS] \oplus x_{in} \oplus [SEP] \oplus x_t \oplus [SEP]$. Then we utilize the probability of predicted answer words (*great* or *terrible*) to express the probability of corresponding classes (*positive* or *negative*).

4. Methodology

In this section, we describe the architecture of our model in detail. Given a sentence and its event set, the ECI task aims to extract causal relations between any two events in the event set. We enumerate every two events in the event set and combine them with the sentence to form a sample. The causal label of the candidate event pair serves as the label of the sample, which includes *causality* and *no causality*.

Fig. 3 shows the overview of our proposed model. We first construct a prompt template based on the event pair and obtain background information and relational information from external knowledge bases. Then we concatenate input text, prompt template, and relational information to form an input sequence, which is encoded into contextualized representations via a pre-trained language model. The model outputs causal results by predicting answer words at the masked positions. It is worth noting that to avoid irrelevant knowledge, we design a selective attention mechanism to integrate relational information. To capture potentially important features, an interactive attention mechanism is proposed to inject background information into the event to get knowledge-enriched event representation. We utilize the objective from knowledge representation learning to jointly optimize the representations of events and causal relations, which can further extract implicit relations between events. And the translation between events is used to revise the predicted results.

4.1. Prompt creation

We construct the prompt template and answer words to facilitate the utilization of task-related knowledge. The template contains the event pair and $[MASK]$. For example, given the event *earthquake* and *destroyed*, the prompt template can be designed in a cloze-style paradigm:

$$x_t = [E_s]\ \underline{earthquake}\ [/E_s]\ [MASK]\ [E_t]\ \underline{destroyed}\ [/E_t]. \quad (2)$$

where $[E_s]$, $[/E_s]$ and $[E_t]$, $[/E_t]$ are special tokens indicating the boundaries of source event and target event, respectively; $[MASK]$ represents the masked position for pre-trained language models to predict; the underlined events are slots for source event and target event. We utilize the probability of predicting *yes* or *no* at the masked position to denote the probability of *causality* or *no causality*.

4.2. Knowledge acquisition

4.2.1. Background information obtaining

Given an event pair from a sentence, the model first retrieves related triples for each event from ConceptNet [18] and then

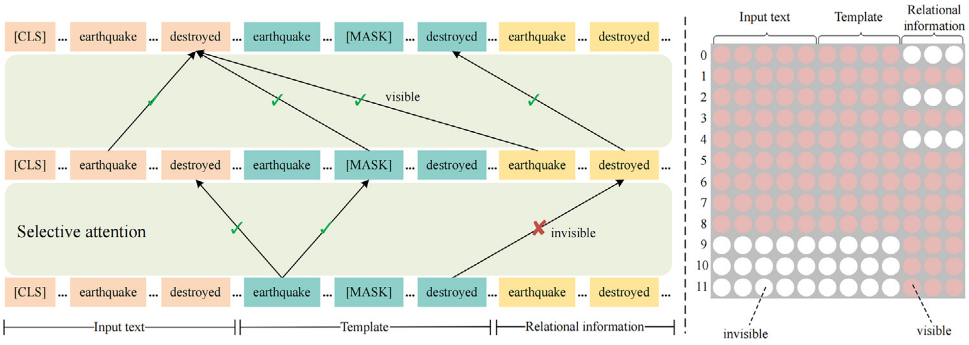

Fig. 4. Illustration of selective attention mechanism. The left shows some typical attention connections among different layers. The right exhibits the attention mask matrix.

transforms them into textual descriptions. We combine these descriptions to form background information for the event.

ConceptNet constructs knowledge as a graph in which each node represents a concept in the form of a natural language word or phrase, and each edge represents a semantic relation. It should be noted that some relations convey little information for causal reasoning, so we prune the corresponding edges in the graph and keep the remaining ones.[1] We treat the event as a concept and match it with the concepts in ConceptNet to obtain the related triples. To make the model better understand background information, these triples are converted into textual descriptions. For example, the background information of the event *earthquake* is "earthquake causes disaster, earthquake is a natural disaster, earthquake is capable of shake ground". In this way, we get rich background information of the event, which can provide important features for identifying causality.

4.2.2. Relational information obtaining

To get relational information, our model retrieves related triples between the event pair and then converts them into textual descriptions. We refer to the concatenation of these descriptions as relational information.

Given a pair of events, we can retrieve the path between the events from ConceptNet as the relational path. The distance from the source concept to the target concept represents the relevance between the two concepts to a certain extent. The closer distance means the two events have stronger relevance [13,64]. Therefore, we seek to find the shortest path between two events with the NetworkX toolkit.[2] When obtaining multiple shortest paths, we randomly select one to avoid redundant information. After getting the relational path, we extract related triples from it and transform them into textual descriptions. For example, the relational information of the event pair *earthquake* and *destroyed* can be represented as "earthquake causes disaster, ruin is related to disaster, ruin is related to destroyed". There are two exceptional cases, one is that the length of the shortest path is greater than L, which means the two events are less relevant; the other is no relational path exists between the event pair. In both cases, we consider that there is no relation between the two events in the knowledge base. With relational information between events, the model can extract causal clues from external knowledge effectively.

4.3. Encoding module

BERT [65] is a bidirectional Transformer architecture [66], which is used to encode the input sequence into contextualized real-valued vectors and has achieved significant performance in many NLP tasks. We first concatenate the input text, prompt template, and relational information as input sequence, which can be represented as:

$$x_s = [CLS] \oplus x_{in} \oplus [SEP] \oplus x_t \oplus [SEP] \oplus x_r \oplus [SEP] \tag{3}$$

where x_{in} denotes tokens of input text, x_t denotes tokens of prompt template, x_r denotes tokens of relational information, $[CLS]$ and $[SEP]$ are special tokens, $\oplus$ represents the concatenation operation. Then we leverage the pre-trained BERT model to encode the input sequence as follows:

$$H^s = BERT(x_s) \tag{4}$$

where $H^s \in \mathbb{R}^{n \times d}$ is the representation of the input sequence, n is the number of input sequence tokens including $[CLS]$ and $[SEP]$, d is the hidden dimension of BERT.

Another pre-trained BERT model is adopted to encode background information. The parameters of the BERT that encodes the input sequence are shared with this BERT model. We can obtain the representation of background information H^b from the last layer of the BERT model in a similar way. It is worth noting that we fix the parameters of this BERT model when training and only utilize the representation of background information for the subsequent process.

4.4. Knowledge infusion

4.4.1. Selective attention mechanism

As studied by [67,68], the fusion of external knowledge may alter the textual meanings of the original input text, thus leading to performance decay. To mitigate the influence of irrelevant knowledge, we design a selective attention mechanism in which the relational information is only visible to the template and events. The attention mask matrix is defined as:

$$M_{(i,j)} = \begin{cases} 0 & w_i, w_j \in x_{in} \oplus x_t \\ 0 & w_i, w_j \in x_r \\ 0 & w_i \in x_r \ and \ w_j \in x_t \\ 0 & w_i \in x_r \ and \ w_j \in E \\ -\infty & otherwise \end{cases} \tag{5}$$

where w_i, w_j are any two tokens from input sequence x_s, 0 means there exists attention from token w_i to token w_j, $-\infty$ means there is no attention from token w_i to token w_j, E denotes the tokens of events from input text. The illustration of the selective attention mechanism is shown in Fig. 4. We can see that the input text and

[1] We remain the following relations: IsA, Causes, RelatedTo, CapableOf, HasProperty, MannerOf, UsedFor, HasSubevent, PartOf, Entails, UsedFor, CreatedBy, MadeOf, and Desires.

[2] https://networkx.org

template do not affect relational information. And the relational information can affect the template and events from the input text, but not other tokens.

4.4.2. Interactive attention mechanism

To exploit potentially important features from external knowledge, we propose an interactive attention mechanism to incorporate background information. Motivated by [66], we leverage multi-head attention to perform interaction and fusion between the event and its background information. Taking the representation of event H_e^s in the template as queries, the representation of background information H^b as keys and values, and the inputs of the ith head of multi-head attention can be computed as:

$$Q_i = H_e^s W_i^q,\ \ K_i = H^b W_i^k,\ \ V_i = H^b W_i^v \tag{6}$$

where $W_i^q \in \mathbb{R}^{d\times d_k/m}$, $W_i^k \in \mathbb{R}^{d\times d_k/m}$ and $W_i^v \in \mathbb{R}^{d\times d_v/m}$ are the projection matrix for the ith head, m is the number of heads. When the event contains more than one token, we perform max-pooling on the representation of each token of the event. The ith head of multi-head attention can be calculated as below:

$$head_i = softmax(\frac{Q_i K_i^T}{\sqrt{d_v/m}})\ V_i \tag{7}$$

where $softmax(\cdot)$ denotes the softmax activation function. Then we concatenate each head and project the result to the output space:

$$H_e^k = Concat(head_1;\ \ldots\ ; head_m)\ W_M \tag{8}$$

where $W_M \in \mathbb{R}^{d_v\times d}$ is the projection matrix, $Concat(\cdot)$ represents the concatenation function, $H_e^k \in \mathbb{R}^d$ represents the knowledge-enriched event representation.

4.5. Causal relation representation

To further capture implicit relations between events, we adopt TransE [53] to jointly optimize the representations of causal relation and event. Given a triple (h, r, t), TransE encodes entities and relations into a uniform low-dimensional space and regards the embeddings of relation $\mathbf{r}$ as translation from head entity vector $\mathbf{h}$ to tail entity vector $\mathbf{t}$ as follows:

$$\mathbf{h} + \mathbf{r} \approx \mathbf{t} \tag{9}$$

We concatenate the knowledge-enriched event representation H_e^k and the original event representation H_e^s and project it to relation space. Then we use the result as event embeddings and the representation of $[MASK]$ as relation. $\mathbf{h}$, $\mathbf{r}$, $\mathbf{t}$ are set as:

$$\begin{aligned}\mathbf{h} &= Concat(H_{e_s}^s, H_{e_s}^k)\ W_h\\ \mathbf{r} &= H_{[MASK]}^s\\ \mathbf{t} &= Concat(H_{e_t}^s, H_{e_t}^k)\ W_t\end{aligned} \tag{10}$$

where $W_h, W_t \in \mathbb{R}^{2d\times d}$ are projection matrices; e is instantiated as e_s or e_t, standing for source event or target event, respectively. We can obtain the causal relation representation by computing translation between event representations $\hat{\mathbf{r}} \approx \mathbf{t} - \mathbf{h}$. And the causal relation can be predicted by calculating which answer word has a higher similarity to $\hat{\mathbf{r}}$, serving as an auxiliary task to revise the prediction results. We leverage the loss in [69] as the training objective:

$$\mathcal{L}_k = -\log\sigma(\gamma - d_r(\mathbf{h},\mathbf{t})) - \sum_{i=1}^{c}\frac{1}{c}\log\sigma(d_r(\mathbf{h}_i',\mathbf{t}_i') - \gamma) \tag{11}$$

where $\sigma(\cdot)$ is sigmoid function, (h_i', r, t_i') represents negative sample, γ is margin, $d_r(\cdot)$ is scoring function defined as:

$$d_r(\mathbf{h},\mathbf{t}) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_p \tag{12}$$

where p is the norm defined as 2. For negative sampling, we use the embeddings of correct answer words as relation and replace the head entity or tail entity with a random entity from other samples to build corrupted triples.

4.6. Model training and predicting

For prompt tuning, the model predicts label words with the $[MASK]$'s representation $H_{[MASK]}^s$. The probability distribution over each label word can be calculated as in Eq. (1). The training objective is to minimize:

$$\mathcal{L}_m = -\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} y\log p([MASK] = v(y)|x_s) \tag{13}$$

where $\mathcal{X}$ is the set of training samples, $v(\cdot)$ is the verbalizer.

The final training objective function of our method can be calculated as follows:

$$\mathcal{L} = \mathcal{L}_m + \alpha\mathcal{L}_k \tag{14}$$

where α is a weight coefficient. If the predicted label word is *yes* and the translation between event representations has a higher similarity to the embedding of *yes*, we consider there is a causal relation between events.

5. Experiments

5.1. Experimental settings

5.1.1. Datasets

We evaluate our method on two benchmark datasets, including EventStoryLine (ESL) [24] and Causal-TimeBank (CTB) [22]. Table 1 gives the statistics of the two mentioned datasets. The EventStoryLine dataset contains 258 documents in 22 topics, 5334 events, and 7805 event pairs, 1770 of which are annotated with causal relations. The percentage of causal event pairs is 22.7%. The Causal-TimeBank dataset contains 184 documents, 6813 events, and 7608 event pairs, 318 of which are annotated with causal relations. The percentage of causal event pairs is only 4.2%. For EventStoryLine, we use the documents in the last two topics as development set, and the rest documents are conducted for a 5-fold cross-valuation evaluation, the same as the prior split in [10,11]. For Causal-TimeBank, we employ a 10-fold cross-validation evaluation, the same as the previous method [11].

5.1.2. Evaluation metrics

For evaluation, we adopt Precision (P), Recall (R), and F1-score (F1) as evaluation metrics, which can be calculated as follows:

$$\begin{aligned}P &= \frac{|A^g \cap A^p|}{|A^p|}\\ R &= \frac{|A^g \cap A^p|}{|A^g|}\\ F1 &= \frac{2\cdot P\cdot R}{P+R}\end{aligned} \tag{15}$$

where A^g denotes the set of all causal samples in the dataset, A^p represents the set of predicted results that contain causal relation. A predicted result is correct if and only if the causal result precisely matches a causal sample.

Table 1
Statistics of EventStoryLine and Causal-TimeBank. "Causal Pair" means the number of event pairs containing causal relation.

Dataset	Document	Topic	Event	Event pair	Causal pair
EventStoryLine	258	22	5334	7805	1770
Causal-TimeBank	184	–	6813	7608	318

5.1.3. Implementation details

In our implementations, all experiments are conducted on NVIDIA RTX 3090 GPU with Pytorch. We use CONCEPTNET 5.7[3] as external knowledge base to obtain background information and relational information. We utilize the cased BERT-base architecture from HuggingFace's Transformers library,[4] which has 12-layers and 768-hidden, to encode the input sequence. The number of heads m is 5, and the weight coefficient α is 1.0. The length of the longest relational path L is 5. The learning rate is initialized as $3e-5$, and the batch size is set to 12. We apply the early stop and AdamW gradient strategy [70] to optimize model parameters. The gradients are calculated by back-propagation. Due to the sparseness of positive examples, we adopt a negative sampling rate of 0.5 for the training process.

5.1.4. Baselines

In order to prove the effectiveness of this work, we compare the proposed method with the state-of-the-art models as follows:

For EventStoryLine, we choose the following baselines for comparison:

- **OP:** proposed by [24], a method that assigns causal relation to every pair of event mentions.
- **LSTM:** proposed by [71], a sequential model that employs a dependency path to model context between event mentions for causality identification.
- **Seq:** proposed by [72], a sequential model for classifying temporal relations between events and is applied to ECI.
- **LR+ and LIP:** proposed by [10], document-level methods that model document causal structure for ECI.
- **KnowDis:** proposed by [37], a distant supervised data augmentation method for this task.
- **KMMG:** proposed by [11], a BERT-based model with mention masking generalization and external knowledge.
- **LSIN:** proposed by [13], a novel method that utilizes external knowledge to construct a descriptive graph and relational graph for ECI.
- **LearnDA:** proposed by [12], which creates a learnable knowledge-guided data augmentation method for this task.
- **Fine-tuning:** a baseline method implemented on BERT by ourselves, which employs a fine-tuning paradigm for ECI.
- **Prompt tuning:** a baseline method that adopts a standard prompt tuning paradigm for ECI.

For Causal-TimeBank, we prefer the following methods for comparison:

- **RB:** proposed by [73], which raises a rule-based method for ECI.
- **DD:** proposed by [73], a data-driven approach based on machine learning.
- **VR-C:** proposed by [32], a method with data filtering and enhanced causal signals based on verb rules.

Table 2
Overall performance compared to the state-of-the-art methods on the EventStoryLine dataset. P, R, and F1 denote precision (%), recall (%) and F1-score (%). The best results are denoted in bold.

Methods	P	R	F1
OP [24]	22.5	**98.6**	36.6
LSTM [71]	34.0	41.5	37.4
Seq [72]	32.7	44.9	37.8
LR+ [10]	37.0	45.2	40.7
LIP [10]	38.8	52.4	44.6
KnowDis [37]	39.7	66.5	49.7
KMMG [11]	41.9	62.5	50.1
LSIN [13]	47.9	58.1	52.5
LearnDA [12]	42.2	69.8	52.6
Fine-tuning	36.3	59.7	45.2
Prompt tuning	38.0	75.6	50.6
KEPT (Ours)	**50.0**	68.8	**57.9**

Table 3
Overall performance compared to the state-of-the-art methods on the Causal-TimeBank dataset. P, R, and F1 denote precision (%), recall (%), and F1-score (%). The best results are denoted in bold.

Methods	P	R	F1
RB [73]	36.8	12.3	18.4
DD [73]	67.3	22.6	33.9
VR-C [32]	**69.0**	31.5	43.2
KnowDis [37]	42.3	60.5	49.8
KMMG [11]	36.6	55.6	44.1
LSIN [13]	51.5	56.2	52.9
LearnDA [12]	41.9	**68.0**	51.9
Fine-tuning	40.9	49.1	44.6
Prompt tuning	47.2	54.0	50.4
KEPT (Ours)	48.2	60.0	**53.5**

5.2. Main results

Tables 2 and 3 show the experiment results on different baselines. From the tables, we can draw the following observations:

(1) The proposed method achieves the best F1-score compared to all the baselines on EventStoryLine and Causal-TimeBank, and our method also has relatively high precision and recall, which verifies the effectiveness of our method. For EventStoryLine, our KEPT outperforms the state-of-the-art method (LearnDA) by a margin of 5.3% on F1-score. For Causal-TimeBank, our KEPT outperforms the state-of-the-art method (LSIN) by a margin of 0.6% on F1-score.

(2) Compared with fine-tuning on the BERT model, our method improves the F1-score by 12.7% and 8.9% on EventStoryLine and Causal-TimeBank, respectively, indicating that simply fine-tuning on the BERT model is not enough for ECI. The reason may be that the BERT model has difficulty in fully using the causal knowledge between the event pair in pre-trained language models. And it also suggests that our method leveraging prompt tuning and causal relation representation learning is more beneficial.

(3) Our KEPT model improves the F1-score by 13.3% and 10.3% upon state-of-the-art methods without BERT (LIP and VR-C) on EventStoryLine and Causal-TimeBank, respectively. And the methods with BERT perform better than no-BERT methods. This is because BERT has a stronger potential to mine important features from the text than no BERT methods, and our approach

[3] https://github.com/commonsense/conceptnet5
[4] https://github.com/huggingface/transformers

can further improve the causal reasoning capability of BERT and introduce rich external knowledge for the ECI task.

(4) Compared with prompt tuning, our model achieves 7.3% and 3.1% improvements in terms of F1-score on the two datasets, respectively. The result proves that introducing background information for the event and relational information between events are effective for the ECI task. It also indicates that the selective attention mechanism and interactive attention mechanism can help integrate external knowledge into the model better and assist in enhancing reasoning ability.

(5) With the same external knowledge base (KMMG and LSIN), our method surpasses those models that introduce external knowledge by a large margin on EventStoryLine. We find that KMMG only introduces knowledge for each individual event and LSIN constructs a relational graph based on the events on the relational path. Both of them ignore the relational information that is significant for the ECI task. It also illustrates that our approach of converting related triples into textual descriptions can take full advantage of external knowledge and achieve superior performance.

(6) Our method exceeds data augmentation approaches (KnowDis and LearnDA) by 8.2% and 5.3% on EventStoryLine, and 3.7% and 1.6% on Causal-TimeBank. It is worth noting that data augmentation can alleviate the data shortcoming problem to a certain extent. But these methods are task-specific and suffer from introducing noise, which may hinder the model's performance. Our model can mitigate the data scarcity problem in a different way and avoid labor-intensive data augmentation.

(7) We can observe that the precision is generally lower than the recall among most models. A good explanation is that the number of positive samples is far less than the number of negative samples. Our method achieves the highest F1-score while guaranteeing relatively balanced precision and recall on both datasets.

5.3. Analysis and discussion

5.3.1. Ablation study

In this section, we conduct ablation experiments to explore the contribution of each component to the model. Specifically, we compare our KEPT with the following variant models:

- **w/o background information** stands for removing the background information and interactive attention mechanism.
- **w/o interactive attention** stands for removing the interactive attention mechanism and replacing the knowledge-enriched event representation with the representation of [*CLS*] from background information.
- **w/o relational information** stands for removing the relational information and selective attention mechanism.
- **w/o selective attention** stands for removing selective attention mechanism.
- **w/o causal relation representation** stands for removing the knowledge representation learning of causal relation.

The experimental results on EventStoryLine are presented in Table 4. We can draw the following conclusions:

(1) Our KEPT model outperforms all the other variant methods on F1-score, and the performance of the model becomes worse after removing each component, which proves that all components are valid and essential for our model.

(2) After removing background information and interactive attention mechanism, the performance of KEPT drops 2.8% and 1.9% on F1-score, respectively. This suggests that the background information can provide additional background knowledge for the event, which plays a significant role in the ECI task. The result also illustrates that the interactive attention mechanism can promote interactions between event and background information, and make the model capture important features from external knowledge.

Table 4
Experiment results of ablation study on EventStoryLine dataset. P, R, and F1 denote precision (%), recall (%), and F1-score (%). The best results are denoted in bold. ∇ represents the difference with KEPT on F1-score.

Methods	P	R	F1	∇
KEPT (Ours)	50.0	68.8	**57.9**	–
w/o background information	42.7	77.3	55.1	−2.8
w/o interactive attention	**51.9**	60.8	56.0	−1.9
w/o relational information	40.7	76.7	53.1	−4.8
w/o selective attention	41.3	**80.1**	54.5	−3.4
w/o causal relation representation	43.1	70.5	53.5	−4.4

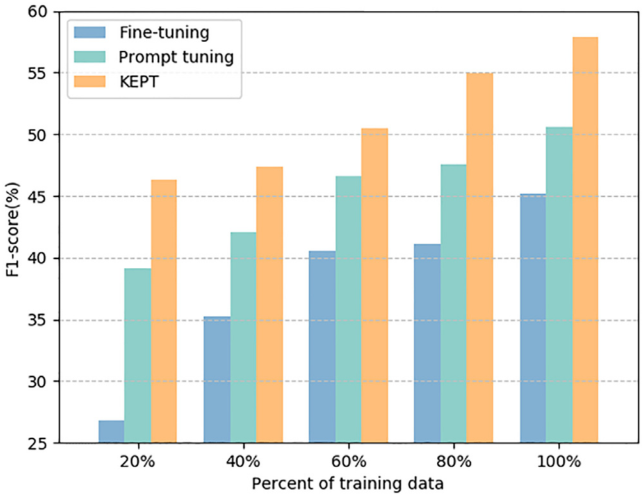

Fig. 5. Effect of training data size on EventStoryLine. Experiments are conducted on 20%, 40%, 60%, 80%, and 100% of training data, respectively.

(3) Without relational information, the performance of KEPT drops 4.8% in terms of F1-score. The result demonstrates that the relational information contains useful causal clues between events that facilitate causal inference, and our model can make full use of the causal clues from ConceptNet, thus boosting the model performance.

(4) We can see the model performance decreases by 3.4% when removing the selective attention mechanism, which reveals that the fusion of relational information introduces noise and changes the textual meanings of the text. We credit the KEPT's improvements to that the selective attention mechanism can mitigate the effect of relational information on the input text and only preserve relevant knowledge for causal reasoning.

(5) Compared with removing causal relation representation learning, our KEPT has better performance, indicating that knowledge representation learning can effectively optimize the representations of causal relation and refine the results predicted by prompt tuning.

5.3.2. Effect of training data size

In order to validate the performance of our model under low-resource conditions, we use fine-tuning, prompt tuning, and KEPT to conduct experiments on different amounts of training data. The experiment results are displayed in Fig. 5.

We can observe that the more training data, the better performance on the three methods, which demonstrates that the scale of training data has a significant effect on the model. Another observation is that the performance of fine-tuning drops

Table 5
Results (F1-score (%)) of unseen events on EventStoryLine. The best results are denoted in bold.

Methods	Both seen	One seen	Both unseen
Fine-tuning	64.5	46.9	36.7
Prompt tuning	71.1	53.8	42.4
KEPT	**73.4**	**56.1**	**48.8**

quickly as the training data decreases, while prompt tuning and KEPT drop slowly. Compared with fine-tuning, prompt tuning has better performance, especially with little training data. We attribute the reason to the fact that prompt tuning can fully use the causal knowledge of the event pair and stimulate the potential of pre-trained language models. The performance difference between prompt tuning and KEPT shows that our method has a stable improvement on F1-score and performs well under low-resource conditions. It also demonstrates that external knowledge is beneficial for ECI. Compared with fine-tuning, our method has a large performance improvement for any percentage of training data. And our method with a small amount of data can achieve comparable performance to fine-tuning with full data, indicating that only a BERT model is not enough for ECI and our KEPT is more effective.

5.3.3. Effect of unseen events

To evaluate the model performance on unseen events, we conduct experiments follow [11] on EventStoryLine. We randomly select 86 documents (1/3 of the total 258 documents) as training set. Then we construct the seen event set that consists of all events in training set. The remaining documents are divided into three groups: (1) **Both seen**, which means both events have appeared in the seen event set and contains 5114 samples; (2) **One seen**, which means only one event has appeared in the seen set and contains 4868 samples; (3) **Both unseen**, which means both events have not appeared in the seen set and contains 2434 samples. We utilize the training set to train our KEPT and test it on Both seen, One seen, and Both unseen, respectively. Experiments are also conducted with fine-tuning and prompt tuning in the same way. The experiment results are reported in Table 5.

We can find that KEPT achieves the best performance on the three test sets. KEPT and prompt tuning outperform fine-tuning by a large margin, which indicates that prompt tuning can stimulate the task-specific knowledge in pre-trained language models and the external knowledge can further improve the causal reasoning ability of the model. Compared with prompt tuning, KEPT improves 6.4% on Both unseen and 2.3% on Both seen in terms of F1-score, which proves that our method can introduce useful knowledge that benefits the ECI task for causal reasoning, especially in the case of unseen events.

5.3.4. Effect of the length of relational path

We explore the performance of our model under different lengths of the longest relational path on EventStoryLine. The length of the longest relational path L is set from $\{3, 4, 5, 6, 7\}$. The results are shown in Fig. 6. It can be observed that the model achieves the best performance when the longest relational path is set to 5. We believe that when the relational path is greater than 5, which means the two events are less relevant, it will incorporate a lot of useless information. When the longest relational path is less than 5, it will cause many event pairs to have no relational path.

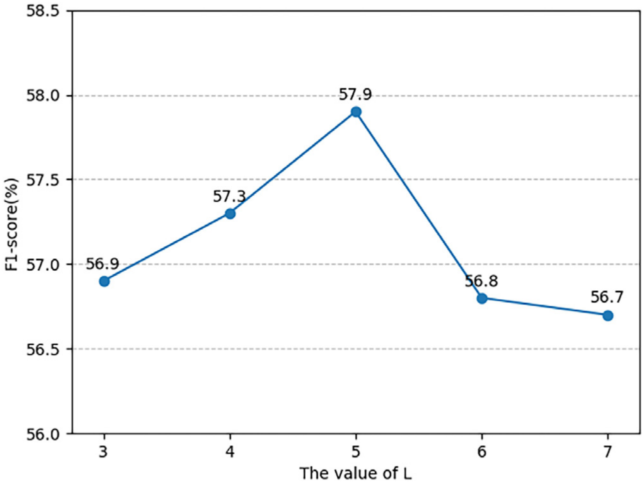

Fig. 6. Results (F1-score (%)) under different lengths of the longest relational path on EventStoryLine.

Table 6
Results under different operations on the knowledge-enriched representation and the original representation. P, R, and F1 denote precision (%), recall (%), and F1-score (%). The best results are denoted in bold.

Operations	P	R	F1
Sum	**51.2**	64.5	57.1
Max	49.7	66.3	56.8
Enriched	50.5	64.1	56.5
Concat (Ours)	50.0	**68.8**	**57.9**

Table 7
Results (F1-score (%)) of cross-topic adaptation on EventStoryLine dataset. "Source" denotes source topic which serves as training set, "Target" denotes target topic which serves as test set, and δ denotes similarity score between source topic and target topic. The best results are denoted in bold.

Setting	Source	Target	δ	LIP	KMMG	KEPT
Low	T8	T35	0%	2.8	44.7	**46.6**
	T13	T12	0%	–	25.1	**26.1**
	T18	T30	0%	–	19.5	**21.8**
Med.	T8	T3	1.7%	6.7	30.9	**36.6**
	T13	T41	0.1%	4.5	28.6	**29.6**
	T18	T35	2.8%	17.1	44.5	**51.1**
High	T8	T19	12.4%	19.4	45.1	**56.0**
	T13	T14	17.1%	27.4	46.0	**53.2**
	T18	T33	27.2%	32.2	49.0	**50.2**

Table 8
Results on SemEval 2010 Task 8. P, R, and F1 denote precision (%), recall (%) and F1-score (%). The best results are denoted in bold.

Methods	P	R	F1
LIP [10]	24.6	21.1	22.8
KMMG [11]	59.4	75.0	66.0
BERT	58.2	72.8	64.7
KEPT (Ours)	**61.8**	**86.4**	**72.1**

5.3.5. Effect of the fusion of the event representations

In this section, we design experiments to prove the effectiveness of the concatenation of the two event representations, which contains the following methods: (1) **Sum**, which stands for summing the two event representations; (2) **Max**, which stands for the maximum of the two event representations; (3) **Enriched**, which stands for only using knowledge-enriched event representation. The experiment results are shown in Table 6. It can be observed that the concatenation achieves the best performance

Table 9
Case study experimented on BERT and KEPT (Ours). "Golden" means the standard causal label, BERT is the result of our implementation. The italic words denote the candidate event pair, ✓ or ✗ denotes there is causality or no causality between the event pair.

Examples	Gloden	BERT	KEPT
In 2004 a massive tremor sparked a *tsunami* that *killed* 170000 people in the province.	✓	✗	✓
The UN refugee agency on friday strongly *condemned* the aerial *bombing* of a temporary refugee camp in South Sudan.	✓	✗	✓
Four bombs were *dropped* within just a few moments — two *landed* inside the camp itself.	✓	✗	✓
Many people were *injured*, but it is difficult to evacuate them due to *traffic jams*.	✗	✓	✗

among all methods, which illustrates the effectiveness of our approach. The reason may be that the other methods cause a certain lack of feature information. And our approach can maximize the retention of useful information.

5.3.6. Cross-topic adaption

To further evaluate the generalization ability of our proposed method, we conduct cross-topic adaption experiments on EventStoryLine. Since different topics usually have different events, we can train our model on a source topic and evaluate it on a target topic to prove the generalization. To ensure a fair comparison, we choose the same source topics and target topics as in [11]. Then we rank the other topics according to their similarity with the source topic. The similarity score between two topics t_1 and t_2 can be defined as $\delta = \frac{E_{t_1} \cap E_{t_2}}{E_{t_1} \cup E_{t_2}}$, where E_t denotes the event set of topic t. We select the topics with the lowest, medium, and highest similarity scores with the source topic as the target topics. We prefer LIP [10], KMMG [11] and proposed KEPT for comparison. The experiment results of cross-topic adaption are shown in Table 7.

From the table, we can observe that LIP performs well when the source and target topic are of high similarity and performs poorly in the case of low similarity. While our KEPT has better performance in both cases. Besides, our KEPT outperforms LIP and KMMG by a large margin over all the cross-topic settings, which demonstrates the excellent generalization ability of our model.

5.3.7. Results on SemEval 2010 Task 8

We conduct experiments to prove the generalization ability of the model on other datasets. Following [11], we first train our model on EventStoryLine. Then the model is tested on the SemEval 2010 Task 8 [74]. It is worth noting that we mainly focus on identifying causal relations and do not take other relations into account. The results are reported in Table 8. It can be observed that the proposed model achieves the best performance, which exhibits great generalization ability. We believe that our model can make full use of the knowledge in pre-trained language models and exploit external knowledge for causal reasoning.

And it can also be observed that the model suffers from low precision and high recall. We attribute this to the fact that the number of positive samples is far less than the number of negative samples. Motivated by [75], to prove our explanations, we conduct experiments on a balanced dataset. We use all the 1003 positive examples and randomly select 1003 negative examples to form a balanced dataset. We train our model on EventStoryLine and test it on this dataset. We can obtain the results: Precision: 78.92%, Recall: 78.76%, and F1-score: 78.84%, which demonstrates that our model can get balanced precision and recall on a perfectly balanced dataset.

5.3.8. Case study

In this section, we present case studies to explore the effectiveness of our model further. Table 9 shows some examples selected from experiment results. We observe that BERT struggles to make the correct predictions when there is no explicit causal clue, but our method performs well in these cases. For example, in the sentence "In 2004 a massive tremor sparked a *tsunami* that *killed* 170,000 people in the province"., BERT thinks there is no causal relation between *tsunami* and *killed*, while our method can obtain the knowledge between *tsunami* and *killed* and give the correct answer. It indicates that BERT has difficulty in understanding the background knowledge of events, while our method can leverage background information and relational information in ConceptNet and fully use knowledge in pre-trained language models to make the right decision.

6. Conclusion

In this paper, we propose a novel **K**nowledge **E**nhanced **P**rompt **T**uning (KEPT) framework for ECI. Different from previous fine-tuning methods, our KEPT employs prompt tuning to incorporate background information and relational information for causal reasoning. To fully exploit external knowledge, our method converts the retrieved triples into textual descriptions and integrates them into the model. We design a selective attention mechanism to mitigate the influence of relational information and propose an interactive attention mechanism to extract useful knowledge from background information. To capture implicit relations from events, we leverage TransE to jointly optimize the representations of events and causal relations. Experiment results on two widely-used benchmarks, EventStoryLine and Causal-Timebank, show that our method achieves state-of-the-art performance. Moreover, KEPT realizes superior performance under low-resource settings and exhibits excellent generalization ability in unseen events and cross-topic adaption. In the future, we plan to extend our work to few-shot learning and apply this method to other related tasks.

CRediT authorship contribution statement

Jintao Liu: Conceptualization, Methodology, Writing – original draft, Software. **Zequn Zhang:** Review & Editing, Validation, Supervision. **Zhi Guo:** Investigation, Supervision. **Li Jin:** Data curation, Supervision. **Xiaoyu Li:** Visualization, Data curation. **Kaiwen Wei:** Review & Editing, Visualization. **Xian Sun:** Data curation, Formal analysis.

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2021YFB3900504) and the National Natural Science Foundation of China (No. 62206267).

References

[1] J. Berant, V. Srikumar, P.-C. Chen, A. Vander Linden, B. Harding, B. Huang, P. Clark, C.D. Manning, Modeling biological processes for reading comprehension, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1499–1510.

[2] A.T. Scaria, J. Berant, M. Wang, P. Clark, J. Lewis, B. Harding, C.D. Manning, Learning biological processes with global constraints, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1710–1720.

[3] K. Radinsky, S. Davidovich, S. Markovitch, Learning causality for news events prediction, in: Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 909–918.

[4] C. Hashimoto, K. Torisawa, J. Kloetzer, M. Sano, I. Varga, J.-H. Oh, Y. Kidawara, Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 987–997.

[5] R. Girju, Automatic detection of causal relations for question answering, in: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, 2003, pp. 76–83.

[6] J.-H. Oh, K. Torisawa, C. Hashimoto, R. Iida, M. Tanaka, J. Kloetzer, A semi-supervised learning approach to why-question answering, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[7] J.-H. Oh, K. Torisawa, C. Kruengkrai, R. Iida, J. Kloetzer, Multi-column convolutional neural networks with causality-attention for why-question answering, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017, pp. 415–424.

[8] Q. Do, Y.S. Chan, D. Roth, Minimally supervised event causality identification, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 294–303.

[9] Q. Ning, Z. Feng, H. Wu, D. Roth, Joint reasoning for temporal and causal relations, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2278–2288.

[10] L. Gao, P.K. Choubey, R. Huang, Modeling document-level causal structures for event causal relation identification, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 1808–1817.

[11] J. Liu, Y. Chen, J. Zhao, Knowledge enhanced event causality identification with mention masking generalizations, in: IJCAI, 2020, pp. 3608–3614.

[12] X. Zuo, P. Cao, Y. Chen, K. Liu, J. Zhao, W. Peng, Y. Chen, Learnda: Learnable knowledge-guided data augmentation for event causality identification, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 3558–3571.

[13] P. Cao, X. Zuo, Y. Chen, K. Liu, J. Zhao, Y. Chen, W. Peng, Knowledge-enriched event causality identification via latent structure induction networks, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4862–4872.

[14] X. Han, W. Zhao, N. Ding, Z. Liu, M. Sun, Ptr: Prompt tuning with rules for text classification, 2021, CoRR arXiv:2105.11259.

[15] Z. Jiang, F.F. Xu, J. Araki, G. Neubig, How can we know what language models know? Trans. Assoc. Comput. Linguist. 8 (2020) 423–438.

[16] T. Schick, H. Schütze, Exploiting cloze-questions for few-shot text classification and natural language inference, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 255–269.

[17] T. Shin, Y. Razeghi, R.L. Logan IV, E. Wallace, S. Singh, AutoPrompt: Eliciting knowledge from language models with automatically generated prompts, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 4222–4235.

[18] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[19] Z. Tan, X. Zhao, W. Wang, W. Xiao, Jointly extracting multiple triplets with multilayer translation constraints, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, (01) 2019, pp. 7080–7087.

[20] H. Huang, Y.-M. Shang, X. Sun, W. Wei, X. Mao, Three birds, one stone: A novel translation based framework for joint entity and relation extraction, Knowl.-Based Syst. 236 (2022) 107677.

[21] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, H. Chen, KnowPrompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction, in: WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022, ACM, 2022, pp. 2778–2788.

[22] P. Mirza, R. Sprugnoli, S. Tonelli, M. Speranza, Annotating causality in the tempeval-3 corpus, in: EACL 2014 Workshop on Computational Approaches To Causality in Language (CAtoCL), Association for Computational Linguistics, 2014, pp. 10–19.

[23] N. Mostafazadeh, A. Grealish, N. Chambers, J. Allen, L. Vanderwende, Caters: Causal and temporal relation scheme for semantic annotation of event structures, in: Proceedings of the Fourth Workshop on Events, 2016, pp. 51–61.

[24] T. Caselli, P. Vossen, The event storyline corpus: A new benchmark for causal and temporal relation extraction, in: Proceedings of the Events and Stories in the News Workshop, 2017, pp. 77–86.

[25] M. Riaz, R. Girju, Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations, in: Proceedings of the SIGDIAL 2013 Conference, 2013, pp. 21–30.

[26] M. Riaz, R. Girju, Recognizing causality in verb-noun pairs via noun and verb semantics, in: Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), 2014, pp. 48–57.

[27] C. Hashimoto, Weakly supervised multilingual causality extraction from Wikipedia, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 2988–2999.

[28] C. Hidey, K. McKeown, Identifying causal relations using parallel wikipedia articles, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1424–1433.

[29] B. Beamer, R. Girju, Using a bigram event model to predict causal potential, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2009, pp. 430–441.

[30] Z. Hu, M. Walker, Inferring narrative causality between event pairs in films, in: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, 2017, pp. 342–351.

[31] Z. Hu, E. Rahimtoroghi, M. Walker, Inference of fine-grained event causality from blogs and films, in: Proceedings of the Events and Stories in the News Workshop, 2017, pp. 52–58.

[32] P. Mirza, Extracting temporal and causal relations between events, in: Proceedings of the ACL 2014 Student Research Workshop, 2014, pp. 10–17.

[33] K. Kadowaki, R. Iida, K. Torisawa, J.-H. Oh, J. Kloetzer, Event causality recognition exploiting multiple annotators' judgments and background knowledge, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5816–5822.

[34] C. Kruengkrai, K. Torisawa, C. Hashimoto, J. Kloetzer, J.-H. Oh, M. Tanaka, Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, (1) 2017.

[35] H. Kayesh, M.S. Islam, J. Wang, Event causality detection in tweets by context word extension and neural networks, in: 20th International Conference on Parallel and Distributed Computing, Applications and Technologies, PDCAT 2019, Gold Coast, Australia, December 5-7, 2019, IEEE, 2019, pp. 352–357.

[36] H. Kayesh, M.S. Islam, J. Wang, A.S.M. Kayes, P.A. Watters, A deep learning model for mining and detecting causally related events in tweets, Concurr. Comput. Pract. Exp. 34 (2) (2022).

[37] X. Zuo, Y. Chen, K. Liu, J. Zhao, KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 1544–1550.

[38] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual, 2020.

[39] N. Zhang, L. Li, X. Chen, S. Deng, Z. Bi, C. Tan, F. Huang, H. Chen, Differentiable prompt makes pre-trained language models better few-shot learners, 2021, CoRR arXiv:2108.13161.

[40] C. Li, F. Gao, J. Bu, L. Xu, X. Chen, Y. Gu, Z. Shao, Q. Zheng, N. Zhang, Y. Wang, et al., Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis, 2021, CoRR arXiv:2109.08306.

[41] N. Ding, Y. Chen, X. Han, G. Xu, P. Xie, H.-T. Zheng, Z. Liu, J. Li, H.-G. Kim, Prompt-learning for fine-grained entity typing, 2021, CoRR arXiv:2108.10604.

[42] X. Chen, N. Zhang, L. Li, X. Xie, S. Deng, C. Tan, F. Huang, L. Si, H. Chen, Lightner: A lightweight generative framework with prompt-guided attention for low-resource ner, 2021, CoRR arXiv:2109.00720.

[43] L. Cui, Y. Wu, J. Liu, S. Yang, Y. Zhang, Template-based named entity recognition using BART, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 1835–1845.

[44] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, H. Chen, Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction, 2021, CoRR arXiv:2104.07650.

[45] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021, CoRR arXiv:2107.13586.

[46] S. Hu, N. Ding, H. Wang, Z. Liu, J. Li, M. Sun, Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification, 2021, CoRR arXiv:2108.02035.

[47] T. Schick, H. Schmid, H. Schütze, Automatically identifying words that can serve as labels for few-shot text classification, in: Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, 2020, pp. 5569–5578.

[48] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 3816–3830.

[49] K. Hambardzumyan, H. Khachatrian, J. May, WARP: Word-level adversarial reprogramming, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4921–4933.

[50] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3045–3059.

[51] X.L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4582–4597.

[52] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, J. Tang, GPT understands, too, 2021, CoRR arXiv:2103.10385.

[53] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, Adv. Neural Inf. Process. Syst. 26 (2013).

[54] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 28, (1) 2014.

[55] Y. Lin, Z. Liu, M. Sun, Y. Liu, X. Zhu, Learning entity and relation embeddings for knowledge graph completion, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[56] G. Ji, S. He, L. Xu, K. Liu, J. Zhao, Knowledge graph embedding via dynamic mapping matrix, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 687–696.

[57] T. Dettmers, P. Minervini, P. Stenetorp, S. Riedel, Convolutional 2d knowledge graph embeddings, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, (1) 2018.

[58] T.D. Nguyen, D.Q. Nguyen, D. Phung, et al., A novel embedding model for knowledge base completion based on convolutional neural network, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 327–333.

[59] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, J. Tang, KEPLER: A unified model for knowledge embedding and pre-trained language representation, Trans. Assoc. Comput. Linguist. 9 (2021) 176–194.

[60] L. Yao, C. Mao, Y. Luo, KG-BERT: BERT for knowledge graph completion, 2019, CoRR arXiv:1909.03193.

[61] B. Wang, T. Shen, G. Long, T. Zhou, Y. Wang, Y. Chang, Structure-augmented text representation learning for efficient knowledge graph completion, 2021.

[62] Y. Lu, H. Lu, G. Fu, Q. Liu, KELM: knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs, 2021, CoRR arXiv:2109.04223.

[63] L. Hu, T. Yang, L. Zhang, W. Zhong, D. Tang, C. Shi, N. Duan, M. Zhou, Compare to the knowledge: Graph neural fake news detection with external knowledge, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 754–763.

[64] B.Y. Lin, X. Chen, J. Chen, X. Ren, KagNet: Knowledge-aware graph networks for commonsense reasoning, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, 2019, pp. 2829–2839.

[65] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long Short Papers), 2019, pp. 4171–4186.

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[67] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, P. Wang, K-BERT: Enabling language representation with knowledge graph, 2020.

[68] H. Ye, N. Zhang, S. Deng, X. Chen, H. Chen, F. Xiong, X. Chen, H. Chen, Ontology-enhanced prompt-tuning for few-shot learning, 2020, CoRR arXiv:2201.11332.

[69] Z. Sun, Z. Deng, J. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019.

[70] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019.

[71] F. Cheng, Y. Miyao, Classifying temporal relations by bidirectional lstm over dependency paths, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017, pp. 1–6.

[72] P.K. Choubey, R. Huang, A sequential model for classifying temporal relations between intra-sentence events, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1796–1802.

[73] P. Mirza, S. Tonelli, An analysis of causality between events and its relation to temporal information, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 2097–2106.

[74] I. Hendrickx, S.N. Kim, Z. Kozareva, P. Nakov, D.Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz, SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals, in: K. Erk, C. Strapparava (Eds.), Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010, The Association for Computer Linguistics, 2010, pp. 33–38.

[75] H. Kayesh, M.S. Islam, J. Wang, Answering binary causal questions using role-oriented concept embedding, IEEE Trans. Artif. Intell. (2022).