Event Causality Identification via Derivative Prompt Joint Learning

Shirong Shen[1] and **Heng Zhou**[1] and **Tongtong Wu**[1] and **Guilin Qi**[1*]
[1]School of Computer Science and Engineering, Southeast University, China
{ssr, zhouheng2020, wutong8023, gqi}@seu.edu.cn

Abstract

This paper studies event causality identification, which aims at predicting the causality relation for a pair of events in a sentence. Regarding event causality identification as a supervised classification task, most existing methods suffer from the problem of insufficient annotated data. In this paper, we propose a new derivative prompt joint learning model for event causality identification, which leverages potential causal knowledge in the pre-trained language model to tackle the data scarcity problem. Specifically, rather than external data or knowledge augmentation, we derive two relevant prompt tasks from event causality identification to enhance the model's ability to identify explicit and implicit causality. We evaluate our model on two benchmark datasets and the results show that our model has great advantages over previous methods.

1 Introduction

Event causality identification (ECI) is an important natural language processing (NLP) task, which aims at identifying causality between events in sentences. Event causality identification supports a variety of NLP applications, e.g., machine reading comprehension (Berant et al., 2014) and event prediction (Radinsky et al., 2012). (Berant et al., 2014; Radinsky et al., 2012). As shown in Figure 1, an ECI model identifies the causalities in sentences S_1 and S_2: (i) **practice** $\overset{cause}{\longrightarrow}$ **won** in S_1; (ii) **attack** $\overset{cause}{\longrightarrow}$ **killed** in S_2. The causality between events in a sentence mainly contains two types: *explicit causality* and *implicit causality*. For instance, the causality **practice** $\overset{cause}{\longrightarrow}$ **won** in S_1 is an explicit causality, which is triggered by the explicit cue words in the sentence. ECI models can take causal cue words as the shortcut for explicit causality identification. As a contrast, the causality

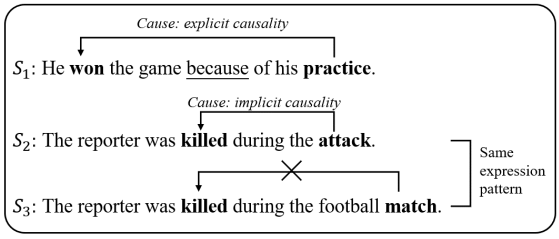

Figure 1: Examples of different causalities. S_1 contains *explicit causality* between **practice** and **won**. S_2 contains *implicit causality* between **killed** and **attack**. S_3 has the same patterns with S_2, but it does not contain any causality.

attack $\overset{cause}{\longrightarrow}$ **killed** in S_2 is an *implicit causality*, because none of explicit cue words is mentioned in S_2. And as shown in Figure 1, comparing S_2 and S_3, we may not always derive the existence of causality from two highly similar expressions without explicit cue words. In other words, *Implicit causality* must be inferred from the semantics and contexts of events.

Most existing methods regard ECI as a classification task, and train customized ECI models on annotated data by supervised learning (Cheng and Miyao, 2017; Choubey and Huang, 2017). However, the large-scale annotated datasets of ECI are relatively hard to collect, referring to that the so far largest dataset EventStoryLine (Caselli and Vossen, 2017) only contains 258 documents and 4316 sentences. Therefore, ECI models are challenged by the data scarcity problem in supervised learning. To address this problem, various methods have been proposed to leverage either the augmented dataset (Hashimoto, 2019) or external knowledge (Liu et al., 2020; Zuo et al., 2021b,a). Hashimoto (2019) exploit weakly supervised method to construct ECI datasets. Liu et al. (2020) and Zuo et al. (2021a) attempt to introduce external structure knowledge to identify causality. However, the model may fail to capture the differences between the explicit causality and

* Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2288–2299
October 12–17, 2022.

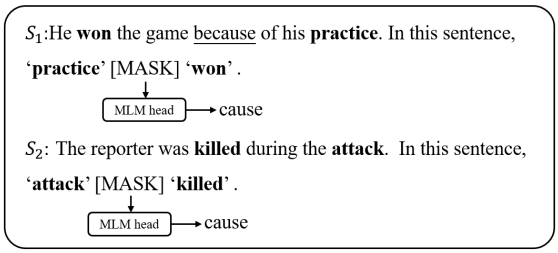

Figure 2: ECI can be converted into the form of predicting `[MASK]` by PLM.

implicit causality without modelling these different types respectively, especially when the ECI model is trained with only insufficient annotated dataset.

In this paper, we propose a new **D**rivative **P**rompt **J**oint **L**earning (**DPJL**) method for ECI, which identifies different causalities effectively without incorporating either more annotated instances or external knowledge. Firstly, as shown in Figure 2, we introduce a new prompt-based learning paradigm to ECI, i.e. converting ECI into a language modelling format and using pre-trained language model (PLM) to identity causalities. Since PLMs accumulated abundant knowledge (Jawahar et al., 2019; Yenicelik et al., 2020; Brown et al., 2020) through the self-supervised training on large-scale corpora, such a prompt-based paradigm may elicit the potential ECI ability of PLM to remedy the scarcity of annotated data. Then, we design two derivative prompt tasks for ECI to identify the explicit causality and implicit causality: (i) Causal cue Word Detection (CWD), which aims to detect the causal cue word of event pairs in a sentence; and (ii) Causal Event Detection (CED), which aims to detect the cause or effect of an event in a sentence. Intuitively, CWD is a straightforward way to identify *explicit causality*, and CED is helpful to identify *implicit causality* by capturing the semantic relevancy between the contextual events. Finally, given the above intuitions, we further propose a joint learning method for event causality identification enhanced by the tasks of CWD and CED. Note that the training data of derivative tasks are generated from ECI, without the cost of additional human annotation. Our contributions are summarized as following:

- We introduce a new prompt-based paradigm to ECI, and we propose a new derivative prompt joint learning method which remedies the problem caused by the scarcity of annotated data. As far as we know, this is the first time to use prompt-based method for ECI.
- We propose two new derivative tasks in the joint learning method, i.e., the causal cue word detection and causal event detection, to strengthen the ability of an ECI model in identifying the explicit causality and implicit causalities respectively. Note that, rather than using more human-annotation, the proposed two derivative tasks leverage the annotated instances modified from the dataset of ECI for training.
- We conduct extensive experiments on two benchmark datasets of ECI, in which our proposed method DPJL achieves the state-of-the-art performance with at least 11 percent F1 improvement on both benchmarks.

2 Related Work

Event Causality Identification Event Causality Identification (ECI) is a crucial information extraction task. Early causal identification methods include rule-based methods (Mirza et al., 2014; Riaz and Girju, 2013; Do et al., 2011) and statistics-based methods (Beamer and Girju, 2009). Recently, some benchmarks on the event causality have been released, e.g. Causal-TimeBank (Mirza and Tonelli, 2014), EventStoryLine Corpus (Caselli and Vossen, 2017) and BECAUSE (Dunietz et al., 2015). Based on these annotated datasets, a number of supervised learning-based methods of ECI have emerged (Kruengkrai et al., 2017; Kadowaki et al., 2019). However, the scale of labeled data in most datasets is relatively small. To solve this problem, Hashimoto (2019) exploited weakly supervised method to construct ECI datasets. Some methods introduce additional knowledge to strengthen the ECI model (Liu et al., 2020; Zuo et al., 2021b,a). Zuo et al. (2020) improved the performance of ECI with distantly supervised labeled training data. These methods introduce the pre-trained language model to generate the high-quality text coding required by the ECI model. But these methods ignore the potential ability of pre-training language model to identify the causality between events, and may fail to capture the differences between the explicit and implicit causalities in a low-resource scenario without modelling these two types of causalities respectively.

Prompt-based learning Recently, pre-trained language models like GPT (Radford and Narasimhan, 2018) , BERT (Devlin et al., 2019) , RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2020) can capture rich knowledge (Jawahar et al., 2019; Yenice-

lik et al., 2020) from massive unlabeled corpora. But there is a big gap between pre-training objectives and fine-tuning objectives, that is, downstream tasks still need to build task-specific models after PLMs, and use task-specific annotated data to fine-tune them. To solve this problem, prompt-based method (Brown et al., 2020) converts the downstream task into the same form as pre-training task. To better build task prompts, automatic search of discrete prompts (Gao et al., 2021), gradient-guided search (Shin et al., 2020) and continuous prompts (Li and Liang, 2021) are successively proposed. To the best of our knowledge, there is no work that uses prompt-based method for ECI task. However, only using prompt-based methods lacks the modeling of different causalities mentioned above.

3 Methodology

In this section, we first introduce problem definition of event causality identification. Then we will show the overview of our proposed model. After that, we introduce a prompt-based ECI method to elicit knowledge from PLMs. Then, we describe the details of two derivative prompt tasks and the joint learning method of derivative prompts. Finally, we introduce the training and prediction process of our model.

3.1 Problem Definition

Given $x = (S, (e_s, e_t))$ as an instance of ECI task, where S is a sentence and (e_s, e_t) is an event pair in S. $\mathcal{Y}$ is the set of causal labels indicating whether there is a causality between event pairs. For an instance x, the purpose of an ECI model is to predict the causal label $y \in \mathcal{Y}$ between e_s and e_t. The traditional approaches formulate ECI as a binary classification problem. In order to learn the feature of different causalities better, we set $\mathcal{Y} : \{Cause, Causedby, NA\}$, which respectively mean that e_s causes e_t, e_s is caused by e_t and there is no causality between (e_s, e_t). The output of our ECI model is a ternary vector corresponding to the probabilities of the three labels. In order to unify with the previous method, *Cause* or *Causedby* both indicate that there is a causality between events.

3.2 Overview

The overview of our approach is shown in Fig. 3. We convert the ECI task into a mask language prediction task, and use an excellent pre-trained masked language model (MLM) named RoBERTa (Liu et al., 2019) to encoder the input sequence and output prediction results. The reason we use MLM is that MLM can make good use of context information and we can flexibly design prompt templates for it. We design two derivative prompt tasks to make our model capture the different abilities of identifying different kinds of causalities. The prompts for all tasks are spliced after the input sentence as the input of RoBERTa. Finally, we use RoBERTa's MLM head to make predictions through a joint learning method.

3.3 Prompt-based Event Causality Identification

For a given instance $x = \{S, (e_s, e_t)\}$, the key to converting ECI task into MLM task is to construct an appropriate prompt template $\mathcal{T}_{ECI}(x)$ and determine the label words $\mathcal{V}$. $\mathcal{T}_{ECI}(x)$ spliced after the input sentence S is used to prompt the PLM to predict the causality between event pair (e_s, e_t). $\mathcal{V}$ refers to a set of words in the vocabulary of PLM that corresponds to the labels of the ECI task. A `[MASK]` token is placed into $\mathcal{T}_{ECI}(x)$ for PLM to fill the label words. There may be many kinds of templates as shown in Figure 2 for ECI, it is not sure which one is most suitable for ECI task. So we add some new learnable tokens to one template to make it dynamically adapt to the task during the model's training process. Since the words in PLM vocabulary may fail to represent the abundant semantic knowledge in causal labels, so we use three virtual words corresponding to three labels form $\mathcal{V}$ as in the previous work (Li and Liang, 2021). Finally, the prompt template and label words for ECI is formalized as follows:

$$\mathcal{T}_{ECI}(x) : In\ this\ sentence, \text{`}e_s\text{'}\ \texttt{<c>}\ \texttt{[MASK]}\ \texttt{</c>}\ \text{`}e_t\text{'}.\ \texttt{[SEP]}$$
$$\mathcal{V} : \{\texttt{Cause}, \texttt{Causedby}, \texttt{NA}\}$$

where `<c>`, `</c>` and virtual words in V are the new learnable tokens added into the vocabulary of PLM, `[SEP]` is the token indicating the end of the sentence. Each new token has an embedding with the same size as the embeddings of original word in the dictionary. Using `Cause` and `Causedby` allows the model to learn the directional features of the causality. We use an injective mapping $\mathcal{M} : \mathcal{Y} \rightarrow \mathcal{V}$ to connect causal labels to label words, each causal label is mapped to a label word with the same name. Then we expand

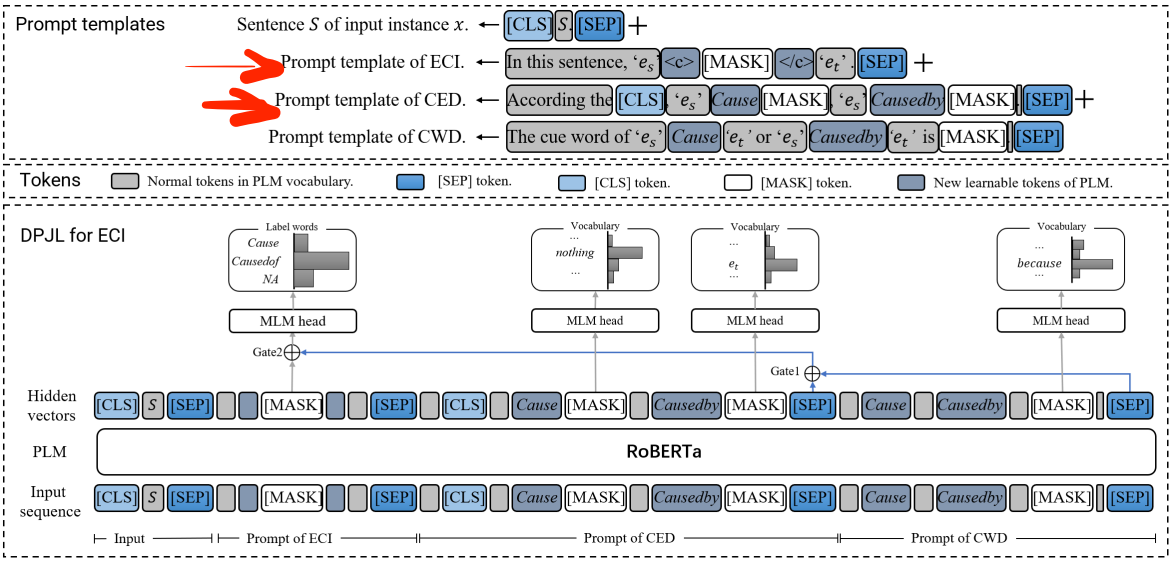

Figure 3: Overview of our derivative prompt joint learning (DPJL). The dashed box above shows the prompt template for each task. The dashed box in the middle shows the correspondence between different token and color blocks. The dashed box below shows the overall framework of DPJL.

ECI : tuong tu detecting prerequisite relation , vay con thieu moi cach giai quyet task 1 thoi

the MLM head layer of PLM with the $\mathcal{V}$ and use the probability distribution over $\mathcal{V}$ at the `[MASK]` position as the probabilities of causal labels. Formally, for an instance x, we first splice `[CLS]` and `[SEP]` on both sides of S to fit the input form of RoBERTa. The probability of its causal label $p(y|x)$ is:

$$p_{MLM\_head}(\texttt{[MASK]} = \mathcal{M}(y|S' + \mathcal{T}_{ECI}(x)) \quad (1)$$

where p_{MLM_head} represents the probability distribution predicted by MLM head layer, '+' means sequence splicing, $S' =$ `[CLS]` S `[SEP]`.

3.4 Derivative Prompt Tasks

We design two derivative prompt tasks for ECI to elicit the abilities to identify explicit causality and implicit causality from PLM. For explicit causality, ECI model need the ability to detect causal cue words of given event pair. For implicit causality, ECI model need the ability to comprehensively analyze event semantics and context. So we design the following two derivative tasks, that is: (i) **Causal cue word detection (CWD)**: Given an instance $x = (S, (e_s, e_t))$, CWD aims to detect the cue word in S which triggers the causality between (e_s, e_t); (ii) **Causal event detection (CED)**: Given an instance $x = (S, e_s)$ where e_s is an event within sentence S, CED aims to detect the event in S which has a causality with e_s.

To elicit the corresponding abilities from PLM, we also set prompts for the two derivative prompt tasks as follows:

$\mathcal{T}_{CWD}(x)$: *The cue word of* 'e_s' `Cause` 'e_t' *or* 'e_s' `Causedby` 'e_t' *is* `[MASK]`. `[SEP]`

$\mathcal{T}_{CED}(x)$: *According the* `[CLS]`, 'e_s' `Cause` `[MASK]`, 'e_s' `Causedby` `[MASK]`. `[SEP]`

The targets of the two tasks are the specific words in the input sentence. If a `[MASK]` has no corresponding answer in the S, its target word is *nothing*, and if the answer word is longer than one word, the target is the first token of answer. The training data of derivative tasks are generated from the original dataset annotated for ECI, please refer to Section 4.2 for more details. We can splice the prompts of derivative task behind the input sentence and use the MLM head layer to predict the probability distribution over PLM's vocabulary at different masked positions as the results of these tasks. Because of the huge vocabulary space, it is possible for PLM to generate words that are not included in sentences. So that, we constrain the candidate vocabulary to the $\mathcal{V}_S$ by setting logits of extraneous words to -inf, where $\mathcal{V}_S$ contains *nothing* and the tokens in S.

3.5 Joint Learning

In this section, we will introduce the joint learning method for derivative tasks and ECI. Firstly,

as shown in Figure 3, we reuse the label words `Cause` and `Causedby` in the prompts of derivative tasks and we concatenate all prompts after S' as the input of RoBERTa (i.e. $S' + \mathcal{T}_{ECI}(x) + \mathcal{T}_{CED}(x) + \mathcal{T}_{CWD}(x)$), all tasks will be predicted and trained simultaneously. On the one hand, the three tasks can share semantics with each other by PLM. On the other hand, learnable label word embedding can be trained during the training of the derivative tasks. Secondly, we set up two gate units as the highway for building the association ECI and derivative tasks. Although the language model can share contextual information, due to the large parameter scale of it, the annotated samples of ECI may not be enough to build the association between ECI and derivative tasks. Specifically, we use the `[SEP]` tokens in $\mathcal{T}_{CED}(x)$ and $\mathcal{T}_{CED}(x)$ to represent the overall semantics of the two derivative prompts. Then we use a gate unit to integrate the semantic information of CWD and CED, and tune the hidden features of the `[MASK]` corresponding to ECI through another gate unit as follows.

$$\begin{aligned} g_1 &= \sigma(W_g^1[\mathbf{h}_{[SEP]}^{CWD}; \mathbf{h}_{[SEP]}^{CED}) \\ \widetilde{\mathbf{h}}_{[SEP]}^{CED} &= (1-g_1)\mathbf{h}_{[SEP]}^{CWD} + g_1\mathbf{h}_{[SEP]}^{CED} \\ g_2 &= \sigma(W_g^2[\mathbf{h}_{[MASK]}^{ECI}; \widetilde{\mathbf{h}}_{[SEP]}^{CED}]) \\ \widetilde{\mathbf{h}}_{[MASK]}^{ECI} &= (1-g_1)\mathbf{h}_{[MASK]}^{ECI} + g_1\widetilde{\mathbf{h}}_{[SEP]}^{CED} \end{aligned} \tag{2}$$

where $\mathbf{h}_{[SEP]}^{CWD}$ and $\mathbf{h}_{[SEP]}^{CED}$ means the hidden features of `[SEP]` in CWD prompts and CED prompts, $\mathbf{h}_{[MASK]}^{ECI}$ is the hidden feature of the `[MASK]` corresponding to ECI, each hidden feature is generated by RoBERTa. W_g^1 and W_g^2 are trainable parameters for gate units, σ is the sigmoid activation function maps variables between $(0, 1)$. $\widetilde{\mathbf{h}}_{[MASK]}^{ECI}$ will replace $\mathbf{h}_{[MASK]}^{ECI}$ as input to MLM head and predict the probability of label words via equation (1). This method effectively builds the link between ECI and derivative tasks without affecting the encoding process of PLM.

3.6 Training and Prediction

We perform supervised training on three tasks simultaneously. We use the cross entropy function to calculate the losses of all tasks, multiply the losses of derived tasks by 0.1, and add them to the losses of ECI as the objective function. In CED task, given a sample may have multiple golden answers, we calculate the averaged cross-entropy loss for each predicted answer. In addition, for each input instance, we have a 10% probability of filling the corresponding `[MASK]` positions with the correct answers of the derivative tasks to speed up the training process. In prediction stage, the target positions in the derivative tasks are all `[MASK]` tokens, and we only predict the probability distribution of the label words of the ECI task by Eq (1).

4 Experiments

Our experiments aim to verify (1) whether the prompt-based method can improve the generalization of the ECI model, and (2) whether the joint learning of derivative prompts can enhance the model's ability to identify different causalities. Our source code is available on Github[1]

4.1 Dataset and Metrics

We perform our method on two main benchmarks, including: EventStoryLine v0.9 (ESC) (Caselli and Vossen, 2017), which contains 258 documents, 4316 sentences, and 1770 causal event pairs; Causal-TimeBank (CTB) (Mirza and Tonelli, 2014) which contains 184 documents, 6813 events, and 318 causal event pairs. Same as previous methods (Gao et al., 2019; Zuo et al., 2021b,a), we use the last two topics of ESC as the development set for two datasets, and conduct 5-fold and 10-fold cross-validation on ESC and CTB respectively. For evaluation, we adopt Precision (P), Recall (R), and F1-score (F1) as evaluation metrics. All the results are the average of three independent experiments.

4.2 Experimental Settings

Training Details. In implementations, we use the RoBERTa[2] with an open pre-trained parameters[3] for our method, which has 12-layers, 768-hiddens, and 12-heads. Each of new tokens added in RoBERTa have 768-dimensional embedding. The size of W_g^1 and W_g^2 are 1536×1. We set the learning-rate of pre-trained parameters and new parameters as 1e-5 and 1e-4 respectively. We use the `causal signal` given in the annotated datasets of ECI as the cue word to construct the training data of CWD, and use the causal event pairs in the annotated datasets of ECI to construct the training data of CED. We adopt a negative sampling rate of 0.5 for training our model, and the batch size for training is 16. And we apply the early stop and AdamW gradient strategy to optimize all models.

[1] If the paper is accepted, a link to the code repository will be published.

[2] https://pytorch.org/hub/pytorch_fairseq_roberta/

[3] https://huggingface.co/roberta-base/tree/main

Compared Methods. We compare our methods with previous state-of-the-art works. For ESC, we prefer the following methods: **LSTM** (Cheng and Miyao, 2017), a dependency path based sequential model. **Seq** (Choubey and Huang, 2017), a sequence model with human designed feaures. **LR+** and **ILP** (Gao et al., 2019), ECI models adopt document structure. For CTB, we prefer the following methods: **RB** (Mirza and Tonelli, 2014), a rule-based system for ECI. **DD** (Mirza and Tonelli, 2014), a data driven machine learning based system. **VR-C** (Mirza, 2014), a verb rule based model with data filtering and gold causal signals enhancement.

In addition, we also compare SOTA methods based on pre-trained language models and introducing external knowledge: **MM** (Liu et al., 2020), the BERT-based SOTA method with mention masking generalization. **KnowDis** (Zuo et al., 2020), a distantly supervised method for ECI. **LearnDA** (Zuo et al., 2021b), a learnable knowledge-guided data augmentation method for ECI. **CauSeRL** (Zuo et al., 2021a), a self-supervised representation learning enhanced ECI method. For a fair comparison, we set up two baseline models based on RoBERTa: **RoBERTa-base**, a RoBERTa-base baseline, we use a linear classifier after RoBERTa for ECI, the input of the classifier is the hidden feature of target events. **Prompt-base**, a prompt-based baseline, our basic proposed ECI method mentioned in Section 3.3.

4.3 Main Results

Table 1 and Table 2 show the experimental results on ESC and CTB respectively. From these results:

Methods	P	R	F1
LSTM(Cheng and Miyao, 2017)	34.0	41.5	37.4
Seq(Choubey and Huang, 2017)	32.7	44.9	37.8
LR+(Gao et al., 2019)	37.0	45.2	40.7
ILP(Gao et al., 2019)	37.4	55.8	44.7
MM(Liu et al., 2020)	41.9	62.5	50.1
KnowDis(Zuo et al., 2020)	39.7	66.5	49.7
LearnDA(Zuo et al., 2021b)	42.2	69.8	52.6
CauSeRL(Zuo et al., 2021a)	41.9	69.0	52.1
RoBERTa-base(**ours**)	40.8	64.7	50.0*
Prompt-base(**ours**)	53.6	68.3	60.0*
DPJL(**ours**)	65.3	70.8	**67.9***

Table 1: Main results on ESC. * denotes a significant test at the level of 0.05.

(1) Our DPJL method outperforms all other ECI methods, and achieves the best F1 on both datasets, 67.9% on ESC and 64.6% on CTB respectively.

Methods	P	R	F1
RB(Mirza and Tonelli, 2014)	36.8	12.3	18.4
DD(Mirza and Tonelli, 2014)	67.3	22.6	33.9
VR-C(Mirza, 2014)	69.0	31.5	43.2
MM(Liu et al., 2020)	36.6	55.6	44.1
KnowDis(Zuo et al., 2020)	42.3	60.5	49.8
LearnDA(Zuo et al., 2021b)	41.9	68.0	51.9
CauSeRL(Zuo et al., 2021a)	43.6	68.1	53.2
RoBERTa-base(**ours**)	40.3	58.2	47.6*
Prompt-base(**ours**)	49.7	69.4	57.9*
DPJL(**ours**)	63.6	66.7	**64.6***

Table 2: Main results on CTB. * denotes a significant test at the level of 0.05.

Specifically, DPJL outperforms the previous SOTA method by at least 10 percentage points. It illustrated that prompt-based methods with derivative prompts joint learning can effectively elicit causal knowledge in PLMs, thus greatly improve the performance of the ECI model.

(2) The experimental results of KnowDis, LearnDA, and CauSeRL show that the introduction of different external knowledge and the method of introducing external knowledge can affect the performance of the ECI model. We note that the performance of Prompt-base and DPJL is higher than that of other knowledge-enhanced methods. It shows that eliciting causal knowledge from PLMs is more beneficial to ECI than the previous approach of introducing external knowledge. The reason may be that previous methods do not fill the gap between external knowledge and true causal representation well, and prompt-based method can directly convert the underlying causal knowledge of PLMs into the ability of causal identification.

(3) RoBERTa-base outperforms the methods without RoBERTa, which illustrates the superiority of RoBERTa. RoBERTa-base is not as good as LearnDA and CauSeRL, which illustrates that simply fine-tuning PLMs cannot completely cover the knowledge required for ECI. Prompt-base and DPJL outperform RoBERTa-base, indicating that prompt-based methods can elicit the potential of PLM to solve ECI better than fine-tuning.

(4) Comparing DPJL and Prompt-base, we notice that DPJL is significantly better than Prompt-base. It is illustrated that joint derivative prompts can elicit more useful knowledge from PLMs for ECI. In addition, the improvement of DPJL of ESC is more obvious than that of CTB. The reason is that ESC has more labeled data, which is beneficial

for training the label word embeddings in derivative prompts joint learning.

4.4 Ablation Experiment

To illustrate the effect of label words reuse and gate units in DPJL, we set up ablation experiments. Different experimental settings are indicated with subscripts, where $Full$ represents the full method of DPJL, $w/o.$ and $w/.$ in subscript mean *with* and *without* respectively, $reuse$ and $gate$ mean label words reuse and gate units of this paper. The ablation results are shown in Table 3 and Table 4. In addition, to verify the generalizability of each derivative prompt task in our method, we adopt two more additional datasets for further ablation studies, i.e., EventStoryLine v1.5 (ESC v1.5) (Caselli and Inel, 2018) [4] and BECAUSE (Dunietz et al., 2017) [5]. The specific experimental results are shown in the Appendix A.

Methods	P	R	F1	Δ
Prompt-base	53.6	68.3	60.0*	-
DPJL$_{w/o.reuse-w/o.gate}$	55.5	68.9	61.4*	+1.4
DPJL$_{w/.reuse-w/o.gate}$	59.9	69.3	64.5*	+4.5
DPJL$_{w/o.reuse-w/.gate}$	62.2	68.8	65.3*	+5.3
DPJL$_{Full}$	65.3	70.8	**67.9***	**+7.9**

Table 3: Ablation results on ESC. * denotes a significant test at the level of 0.05. **Δ** means the points higher than Prompt-base.

Methods	P	R	F1	Δ
Prompt-base	49.7	69.4	57.9*	-
DPJL$_{w.o./reuse-w/o.gate}$	50.2	70.4	58.6*	+0.7
DPJL$_{w./reuse-w/o.gate}$	52.1	71.2	60.1*	+2.2
DPJL$_{w.o./reuse-w/.gate}$	62.5	63.6	63.0*	+5.1
DPJL$_{Full}$	63.6	66.7	**64.6***	**+6.7**

Table 4: Ablation results on CTB. * denotes a significant test at the level of 0.05. **Δ** means the points higher than Prompt-base.

Effect of Derivative Prompts. Comparing Prompt-base with DPJL$_{w/o.reuse-w/o.gate}$, despite simply splicing the derivative prompt after the sentence, the model performance has improved. It illustrates that the derivative prompts contain the knowledge for identifying causality and this knowledge can assist Prompt-based ECI through context. In the same way, all the methods using derivative prompts are better than Prompt-base, indicating that adding derivative prompts is an effective method to improve prompt-based ECI model.

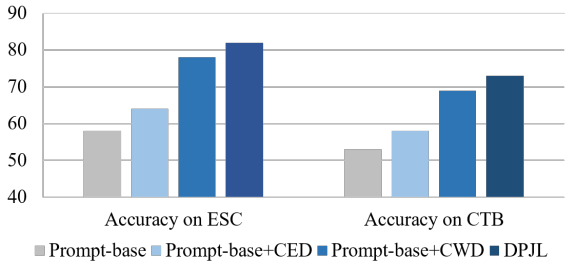

Figure 4: The accuracy of explicit set.

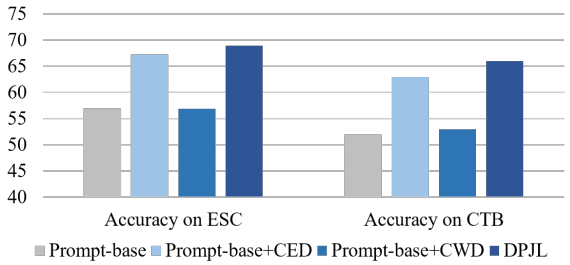

Figure 5: The accuracy of implicit set.

Effect of Gate Units. On both datasets, the methods with gate units outperform the methods without gate units whether the label words are reused in derivative prompts or not. This proves that gate units can provide a highway linking the ECI model and derivative prompts, which can better utilize the information in derivative prompts to improve the performance of causality identification in the case of a small number of training samples.

Effect of Label Word Reuse. On both ESC and CTB, the methods with label words reuse outperform the methods without label words reuse whether with or without gate units. It illustrates that reuse label words in derivative prompts can use the encoding process of the RoBERTa to strengthen the learning of the label word embeddings, and at the same time, the label words can obtain the semantic features of implicit and explicit causalities through derivative prompts. It is worth noting that the effect of label word reuse on CTB is not as obvious as that on ESC, which may be due to the fact that there are fewer training samples in CTB which cannot fully train the word embeddings of label words.

[4] http://github.com/tommasoc80/EventStoryLine
[5] http://github.com/duncanka/BECAUSE

	Samples	Characteristics	Prompt-base	& CED	& CWD	DPJL
1	Iraq said it **invaded** Kuwait <u>because</u> of **disputes** over oil and money.	Simple causality	✓	✓	✓	✓
2	Fans and family **mourn** her **passing** , but Williams had a long , full life	Implicit causality	×	✓	×	✓
3	The general **strike** was <u>staged</u> as a **protest** against a new round of draconian austerity measures.	New cue word/Semantic association	×	✓	✓	✓
4	Mr. Potach notes older , more traditional groups like the Ku Klux Klan are also **opening** new chapters , <u>thanks</u> in part to their ability to **use** new technologies like the Internet .	New cue word	×	×	✓	✓

Figure 6: Case study. Bold words are target events, and underlined words indicate causal cue words. & CED and & CWD represent CED and CWD used in training, respectively.

4.5 Effect of Derivative Tasks on Implicit and Explicit Causalities

To illustrate the effect of derivative tasks on implicit and explicit causalities respectively, we divide the test data into implicit set and explicit set according to whether the test data contains causal cue words. The samples in implicit set contain implicit causalities and the samples in explicit set contain explicit causalities. Then we limit the kinds of derivative prompts in the model, where Prompt-base+CWD means we only use the prompts of CWD, Prompt-base+CED represents our method only using the prompts of CED. All methods with derivative prompts are consistent with DPJL except for the different types of derived tasks. We report the accuracy of these methods on both test sets.

As shown in Figure 4, joint CWD can significantly improve the accuracy of the model on explicit set, which proves that CWD can effectively elicit the ability of PLM to detect causal cue words, thereby improving the performance of the model in identifying explicit causalities. A similar phenomenon also occurs in Figure 5, the CED can effectively improve the performance of prompt-based ECI model on implicit set. It can also be found that, joint CED also helps to identify explicit causality to a certain extent, because CED not only enhances the understanding of event semantics, but also enhances the understanding of underlying causal expressions in context.

4.6 Case Study

In order to visually demonstrate the effectiveness of each of our derivative prompt joint learning method and the effect of each derivative task, we conducted a case study. As shown in Figure 6, case 1 is a simple sample of causality with a causal cue word. All methods can correctly identify the causality between target event pair. Case 2 and case 3 show that the CED task can effectively elicit the causal semantic knowledge of events in PLM, thus improving the effect of ECI between related events. However, in case 4, there is no strong semantic relationship between **opening** and **use**, and the causal cue words <u>thanks</u> did not appear in the training set, so the method only using CED can't correctly identify the causality between the event pairs. Case 3 and case 4 shows that CWD can elicit PLM's ability to identify causal cues, and then make the model show good generalization ability when new cue words appear. However, the method only with CWD can't correctly identify the causality in case 2, which shows that only using CWD cannot extract implicit causality well. DPJL can correctly extract all causality, which shows that our proposed method can strengthen the effect of ECI model extraction to identify explicit causality and implicit causality at same time by joint two derivative cue learning tasks.

Finally, the experiments verify that (1) the prompt-based method can effectively improve the generalization ability of the ECI model by eliciting causal knowledge in PLMs, and (2) the joint learning of derivative prompts can strengthen the model's ability to identify different causalities.

5 Conclusion and Future Work

In this paper, we first introduced a new prompt-based paradigm to event causality identification and proposed a new derivative prompt joint learning method, i.e, DPJL. The proposed method adopts two new derivative tasks, i.e., the causal cue word detection and causal event detection, to strengthen the ability of an ECI model in identifying the explicit causality and implicit causality respectively. The experimental results demonstrate that the proposed method achieves the state-of-the-art performance with at least 11 percent F1 improvement on both of two well-known benchmarks, i.e., EventStoryLine and Causal-TimeBank. Additionally, the

detailed analysis suggests the effectiveness of joint-learned prompt-based derivative tasks on performance improvement in downstream tasks. In the future, we will try knowledge-enhanced methods to construct both of derivative tasks and data for an ECI model, which may fill the gap between knowledge and samples.

6 Acknowledgement

Research in this paper was partially supported by the Natural Science Foundation of China (U21A20488).

References

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *CICLing*.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *EMNLP*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Tommaso Caselli and Oana Inel. 2018. Crowdsourcing storylines: Harnessing the crowd for causal relation annotation. In *EventStory@Coling*.

Tommaso Caselli and P. Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *NEWS@ACL*.

Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional lstm over dependency paths. In *ACL*.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *EMNLP*.

Jesse Dunietz, Lori S. Levin, and Jaime G. Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *LAW@NAACL-HLT*.

Jesse Dunietz, Lori S. Levin, and Jaime G. Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *LAW@ACL*.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *NAACL*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723.

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from wikipedia. In *EMNLP*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL*.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *EMNLP*.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *AAAI*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *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)*, abs/2101.00190.

Jian Liu, Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *IJCAI*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Paramita Mirza. 2014. Extracting temporal and causal relations between events. *ArXiv*, abs/1604.08120.

Paramita Mirza, R. Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *EACL 2014*.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *COLING*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *WWW*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *SIGDIAL Conference*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Eliciting knowledge from language models using automatically generated prompts. *ArXiv*, abs/2010.15980.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does bert capture semantics? a closer look at polysemous words. In *BLACKBOXNLP*.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021a. Improving event causality identification via self-supervised representation learning on external causal statement. *ArXiv*, abs/2106.01654.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021b. Learnda: Learnable knowledge-guided data augmentation for event causality identification. In *ACL/IJCNLP*.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. Knowdis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *COLING*.

A Appendix

We set up external ablation experiments to test the effect of different derivative tasks on EventStoryLine v1.5 (ESC v1.5) (Caselli and Inel, 2018) and BECAUSE (Dunietz et al., 2017). ESC v1.5 is an updated version of ESC v0.9, which contains 1,204 sentences and 7,778 event pairs, covering 22 news topics. The corpus is annotated by experts and crowd (Caselli and Inel, 2018). BECAUSE contains a total of 5380 sentences, and 1803 causal event pairs. The experimental setting is the same as that in Section 4.

We compare our proposed DPJL with the following methods: **RoBERTa-base**, a RoBERTa-base baseline which uses a linear classifier after RoBERTa for ECI. The input of the classifier is the hidden feature of target events; **Prompt-base**, a prompt-based baseline, which is our basic proposed ECI method mentioned in section 3.3; **Prompt-base+CED**, the method of joint learning with prompt-base method and CED; **Prompt-base+CWD**, the method of joint learning with prompt-base method and CWD. To show the real effect of CWD and CED, we use the label words reuse and gate units of this paper in **Prompt-base+CED** and **Prompt-base+CWD**. The experimental results on ESC v1.5 and BECAUSE are shown in Table 5 and Table 6.

Methods	P	R	F1	Δ
RoBERTa-base	53.6	64.3	59.3*	-
Prompt-base	64.0	64.6	64.3*	+5
Prompt-base+CWD	66.5	67.9	67.2*	+7.9
Prompt-base+CED	64.4	**70.4**	67.3*	+8.0
DPJL$_{Full}$	**76.9**	67.5	**71.9***	**+12.6**

Table 5: Experimental results on ESC v1.5. * denotes a significant test at the level of 0.05. **Δ** means the points higher than RoBERTa-base.

Methods	P	R	F1	Δ
RoBERTa-base	50.0	52.6	51.3*	-
Prompt-base	53.7	52.9	53.3*	+2
Prompt-base+CWD	52.9	**56.2**	54.5*	+3.2
Prompt-base+CED	**61.5**	52.6	56.7 *	+5.4
DPJL$_{Full}$	58.8	55.6	**57.1***	**+5.8**

Table 6: Experimental results on BECAUSE. * denotes a significant test at the level of 0.05. **Δ** means the points higher than RoBERTa-base.

DPJL achieves the highest F1 score on both datasets, which demonstrates the consistent effectiveness of our method. Prompt-based is superior to RoBERTa-base, which shows that prompt learning can better elicit causal knowledge in PLM than simple fine-tunig method in ECI task. The performance of Prompt-base+CWD and Prompt-base+CED is better than that of Prompt-based, which shows that joint both derivative tasks improve the model's ability to elicit the ability of ECI from PLM. Prompt-base+CED outperforms Prompt-base+CWD on BECAUSE, this may be because BECAUSE pays more attention to evaluating the ability of the model to identify implicit causality, while CED can help the ECI model to enhance the ability to identify the causal semantic association between events in PLM by predicting causal events. DPJL combines two kinds of derivative tasks, and the performance exceeds that of using only one kind of derivative task, which shows that both derivative tasks are meaningful to ECI and can complement each other.

The experimental results on these two datasets are consistent with the experimental results in the main paper. The results show that our method can adapt to more datasets, and further verifies the effectiveness of DPJL on ECI task.