DAPrompt: Deterministic Assumption Prompt Learning for Event Causality Identification

Abstract

Event Causality Identification (ECI) aims at determining whether there is a causal relation between two event mentions. Conventional prompt learning designs a prompt template to first predict an answer word and then maps it to the final decision. Unlike conventional prompts, we argue that predicting an answer word may not be a necessary prerequisite for the ECI task. Instead, we can first make a deterministic assumption on the existence of causal relation between two events and then evaluate its rationality to either accept or reject the assumption. The design motivation is to try the most utilization of the encyclopedia-like knowledge embedded in a pre-trained language model. In light of such considerations, we propose a deterministic assumption prompt learning model, called DAPrompt, for the ECI task. In particular, we design a simple deterministic assumption template concatenating with the input event pair, which includes two masks as predicted events' tokens. We use the probabilities of predicted events to evaluate the assumption rationality for the final event causality decision. Experiments on the EventStoryLine corpus and Causal-TimeBank corpus validate our design objective in terms of significant performance improvements over the state-of-the-art algorithms.

1 Introduction

Event Causality Identification (ECI) is to detect whether there exists a causal relation between two event mentions in a document. Fig. 1 illustrates an example of event mention and causality annotations in an accident topic document in the widely used Event StoryLine Corpus (ESC), in which eleven event pairs are annotated with causal relation, including both the intra-sentence and cross-sentence causalities. The ECI task is to identify a causal relation between two event mentions. Causality identification is of great importance for many Natural Language Processing (NLP) applications, such as question answer (Sui et al., 2022), information extraction (Xiang and Wang, 2023), and etc.

Some recent deep learning-based methods design sophisticated neural models to learn a kind of contextual semantic representation for each event,

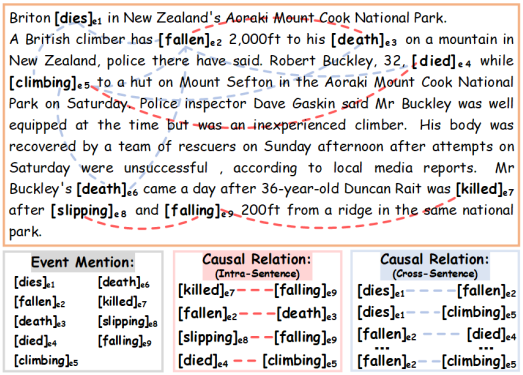


Figure 1: Illustration of event causality annotation for an accident topic document in the ESC corpus. Event mentions are annotated as one or more words in a raw sentence, and causal relation annotations can exist in intra-sentence or cross-sentence event mentions.

such as the Rich Graph Convolutional Network (RichGCN) (Phu and Nguyen, 2021), Event Relational Graph Transformer (ERGO) (Chen et al., 2022), Graph-based Event Structure Induction model (GESI) (Fan et al., 2022). Although these graph neural networks can effectively learn contextual semantics as events' or event pairs' representations, they have ignored to utilize some external commonsense knowledge, like *earthquake causes tsunami*, to augment causality detection.

External knowledge bases can be employed to provide external causal knowledge for augmenting causality identification. For example, the *ConceptNet* (Speer et al., 2017) contains abundant graph-structured knowledge, in which each node represents a concept and each edge corresponds to a semantic relation between concepts. Liu et al. (2020) and Cao et al. (2021) both use such knowledge triplets in the ConceptNet to boost representation learning. Moreover, the *FrameNet* knowledge base (Baker et al., 1998), as well as the *WordNet* (Miller, 1995) and *VerbNet* (Schuler, 2006) lexical knowledge base have also been used to obtain external causal knowledge for the ECI task (Zuo et al., 2021a, 2020).

Although external knowledge bases can provide abundant information, how to extract appropriate knowledge triplets for the ECI task is not easy to implement, not to mention their encoding and fusion into task-specific events' representations. Recently,

the *pre-train, prompt, and predict* paradigm (Liu et al., 2023) (viz., *prompt learning*) based on a Pre-trained Language Model (PLM) has been successfully applied in many NLP tasks. The successes can be contributed to the task transformation via carefully designed templates and answers, so as to well utilizing the encyclopedic linguistic and event causal knowledge embedded within a PLM during model training.

For the ECI task, the basic idea of prompt learning is to design a prompt template such as (... <e_1> `[MASK]` <e_2> ...), and an answer space such as {`cause`, `so`, `not caused`, ...}. The template as a sentence is input into a PLM to output the mask token representation for its classification to an answer word, which is then further mapped into a causal relation. The recent DPJL model (Shen et al., 2022) designs such a prompt template together with two derivative templates to augment representation learning for the mask token, which just achieved the new state-of-the-art performance of the intra-sentence event causality identification on the commonly used ESC corpus (Caselli and Vossen, 2017).

We argue that the performance of such a conventional prompt learning is heavily dependent on the designed prompt templates and selected answer words. On the one hand, the manually-designed templates could be sensitive to its consisting words, as even synonyms (especially nouns/adj. words) could have subtle semantic differences that may impact on template quality. So a good template might need to try different combinations of composing synonyms. On the other hand, it is still a kind of *implicit inference task* that transforms the ECI task into the prediction and mapping of some pre-selected answer word in the PLM vocabulary. As each answer word may also be kind of synonyms and with subtle semantic differences, it is often a big workload for selecting answers.

Unlike conventional prompts, we argue that predicting an answer word may not be a necessary prerequisite for the ECI task. Instead, we can first assume that a causal relation does exist between two input events and then evaluate the rationality of such an assumption by directly predicting the input events. As such, we do not need to search for a well-designed prompt template as well as carefully selected answer words. Furthermore, predicting the input events from the raw sentences could better utilize a PLM for its powerful capability of learning contextual semantic representations, as well as utilizing some encyclopedic linguistic and event causal knowledge embedded within a PLM.

Motivated from such considerations, we propose a novel *deterministic assumption prompt learning* model, called DAPrompt, for the ECI task. Specifically, we first design a simple deterministic assumption template which includes two mask tokens for predicting the input events. We concatenate the two raw event sentences and the assumption template as an input sentence into a PLM. The objective is to predict the input events via the two masks for evaluating the rationality of the deterministic causal assumption. If the likelihood of correctly predicting the input events is larger than a decision threshold, then we accept the assumption and identify the existence of a causal relation. Experiment results show that our proposed DAPrompt significantly outperforms the state-of-the-art algorithms, in terms of much higher F1 score in all intra-sentence, cross-sentence, and overall event causality identifications[1].

2 Related Work

Graph-based Causality Identification: The graph-based approaches first construct a graph and model the ECI as either a graph-based node classification or edge prediction problem. Some have applied graph neural networks for learning event node representations from document-level contextual semantics (Phu and Nguyen, 2021; Cao et al., 2021; Fan et al., 2022). For example, Phu and Nguyen (2021) models diverse connections in between words of a document, like positional connection, syntactic dependency and etc., for the graph construction. They use a graph convolutional network to learn the event mention nodes' representations, and identify causalities through event node pair classification.

Instead of node classification, some studies formalize the ECI task as a graph-based edge prediction problem (Zhao et al., 2021; Chen et al., 2022). For example, Zhao et al. (2021) initialize event nodes' embeddings from a document-level encoder based on the PLM, and use a graph inference mechanism to update the graph for causal edge prediction. Chen et al. (2022) build an event relational graph where each node denotes a pair of events and propose a graph transformer model to capture potential causal chains among nodes. These approaches, however, only exploit contex-

[1]Source codes will be released after the anonymous review.

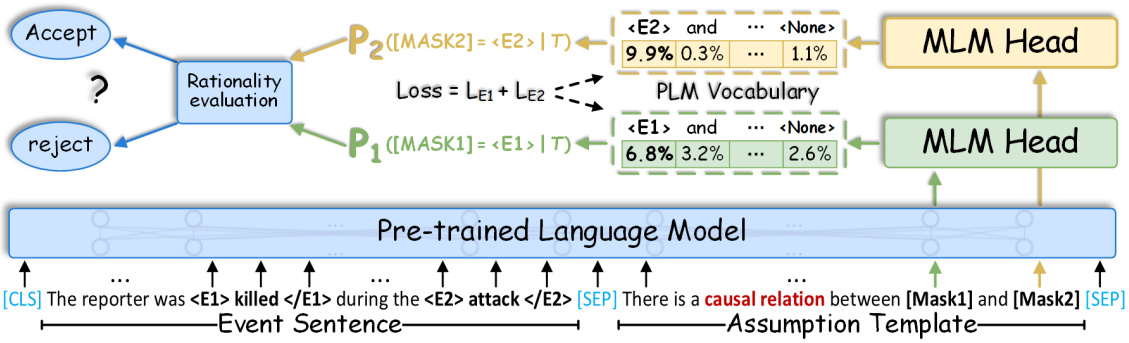


Figure 2: Illustration of our DAPrompt model.

tual and semantic information in a document for causal relation classification.

Knowledge-boosted Causality Identification: As those existing knowledge bases store a large amount of structured information, some studies directly exploit them for data expansion and augmentation in model training (Zuo et al., 2020, 2021b). Besides data expansion, some studies try to discover the causal patterns from external knowledge bases to implement a kind of knowledgeable event causality inference (Liu et al., 2020; Zuo et al., 2021a; Cao et al., 2021). For example, Liu et al. (2020) propose to mine a kind of event-agnostic and context-specific patterns from the ConceptNet to enhance the ability of their model for previously unseen cases. Cao et al. (2021) encode some graph-structured knowledge from the ConceptNet, including descriptive graph knowledge and relational path knowledge, and performs event causality reasoning based on these induced knowledge.

Prompt Learning Paradigm: With the emergence of large-scale PLMs like the BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and etc., the prompt learning has become a new paradigm for many NLP tasks, which uses the probability of text in PLMs to perform a prediction task and has achieved promising results (Seoh et al., 2021; Wang et al., 2021; Xiang et al., 2022). A few studies have applied the prompt learning via designing appropriate prompt templates (Shen et al., 2022; Liu et al., 2020). For example, Shen et al. (2022) use a masked language model as main prompt to predict the causality between event pair. They further design two derivative prompt task to leverage potential causal knowledge in PLM for explicit causality identification based on the causal cue word detection. Liu et al. (2020) use an event mention masking generalization mechanism to encode some event causality patterns for causal relation reasoning.

The proposed DAPrompt is also based on prompt learning paradigm, but it designs a novel prompting style of first deterministic assumption and next rationality evaluation.

3 The Proposed DAPrompt Model

We first make a deterministic assumption on the existence of causal relation between two events in a document. Our DAPrompt identifies event causality by evaluating the rationality of a deterministic assumption. Specifically, we design a prompt template for a deterministic assumption to predict two input events, and use the probabilities of the correctly predicted events to determine whether to accept or reject the assumption, so as to making a final decision on event causality. Fig. 2 illustrates the proposed DAPrompt model.

3.1 Prompt Templatize

The full prompt template T contains two constructed sentences T_1 and T_2 that are concatenated with a `[SEP]` token as the input sentence to a PLM.

The T_1, called the *event sentence*, is designed for predicting two *virtual event tokens* (VETs) `<E1>` and `<E2>`, each representing one of the input events. The design consideration is from the fact that the event mentions in different raw sentences usually consist of much different vocabulary words, not to mention having different lengths. We need to simplify and regulate their representations. We admit that using only two virtual tokens to represent diverse events is a bold attempt. Yet it provides an efficient way to link input events with the mask tokens in our assumption template.

We note that although event mentions are normally annotated by a few words of a raw sentence, their representation learning should include the full

sentence for better capturing contextual semantics. Let $S_1 = (v_1, ..., [e_1, ..., e_m], ..., v_n)$ denote one raw sentence containing the annotated event mention $[e_1, ..., e_m]$, where v_is and e_js are all vocabulary words. We insert the VET <E1> and </E1> before and after $[e_1, ..., e_m]$ respectively to transform a raw sentence. The event sentence T_1 consists of a prefix token [CLS] and the two transformed sentences. Note that if two event mentions are within one raw sentence, we directly insert the VET tokens into the raw sentence to construct T_1. Fig. 2 illustrates such an example of T_1 with one raw sentence containing two event mentions.

The T_2 is our *assumption template*, which designs a deterministic statement of the causal relation between two mask tokens, that is,

T_2 = There is a causal relation between [MASK1] and [MASK2].

The two mark tokens are used to respectively predict the virtual event tokens. Let $\mathcal{V}$ denote the PLM vocabulary. The mask token [MASK1] is used to predict a word from $\mathcal{V}'_1 = \mathcal{V} \cup \{E_1\}$, and the [MASK2] is used to predict a word from $\mathcal{V}'_2 = \mathcal{V} \cup \{E_2\}$. Recall that a virtual event token is used to represent one event mention. So if both mask tokens can be correctly predicted as the corresponding virtual event token, then the deterministic causal assumption can be accepted, that is, there does exist a causal relation between the two events. The assumption template is suffixed with a separate [SEP] token.

3.2 Answer Prediction

We predict a mask token as one of the words in the enriched PLM vocabulary $\mathcal{V}'$. Two Masked Language Model (MLM) classifiers are adopted each for estimating the probability of a mask token as a vocabulary word:

$$P(\texttt{[MASK]} = v \in \mathcal{V}' \mid T). \quad (1)$$

Note that the two MLM classifiers are initially identical, which is pre-trained by the PLM. They will be separately fine-tuned during our model training. In each MLM classifier, a softmax layer is applied on the prediction scores of all words for the probability normalization.

We are mainly interested in the following two probabilities: $P_1(\texttt{[MASK1]} = \texttt{<E1>} \mid T)$ and $P_2(\texttt{[MASK2]} = \texttt{<E2>} \mid T)$. Each can be regarded as the likelihood of an input event appearing in the deterministic assumption template and will be used in our rationality evaluation.

3.3 Rationality Evaluation

We use the sum of P_1 and P_2 as a joint decision variable for *rationality evaluation* of the deterministic assumption, that is,

$$f(T) = \begin{cases} Accept, & if \quad P_1 + P_2 \geq \rho \\ Reject, & if \quad P_1 + P_2 < \rho \end{cases} \quad (2)$$

where ρ is the *joint decision threshold* and $\rho \in [0, 2]$ as $P_1, P_2 \in [0, 1]$. If $P_1 + P_2 \geq \rho$, which suggests that the two masks are much likely to be the input events, then we accept the deterministic assumption of the existence of a causal relation between two events; Otherwise, we reject the assumption and the two input events are not with a causal relation. We note that we use a simple sum operation for $f(T)$, as we have no prior knowledge about which event is harder to predict.

3.4 Training Strategy

In the training phase, we use the <E1> and <E2> token as the positive label, if there is indeed a causal relation between two input events; While the virtual word <None> initialized by all other words is used as negative label for both [MASK] token prediction, if the causal relation assumption is incorrect. We tune the PLM parameters as well as the two MLM classifier parameters based on these labels, and compute a cross entropy loss as a MLM classifier loss $\mathcal{L}_1$ ($\mathcal{L}_2$):

$$\mathcal{L} = -\frac{1}{K} \sum_{k=1}^{K} \mathbf{y}^{(k)} \log(\hat{\mathbf{y}}^{(k)}) + \lambda \|\theta\|^2, \quad (3)$$

where $\mathbf{y}^{(k)}$ and $\hat{\mathbf{y}}^{(k)}$ are the answer label and predicted answer of the k-th training instance, respectively. λ and θ are the regularization hyperparameters. The overall loss of our DAPrompt is as follows:

$$\mathcal{L}_{DAPropmt} = \mathcal{L}_1 + \mathcal{L}_2. \quad (4)$$

We use the AdamW optimizer (Loshchilov and Hutter, 2019) with $L2$ regularization for model training.

4 Experiments Settings

4.1 Datasets

EventStoryLine (Caselli and Vossen, 2017) contains 22 topics and 258 documents from various

news web-sites. There are in total 5,334 event mentions in the ECS dataset. A total number of 5,655 event pairs are annotated with causal relations, among which 1,770 causal relations are from intra-sentence event pairs and 3,855 causal relations are from cross-sentence event pairs. Following the standard data splitting (Gao et al., 2019), we use the last two topics as the development set, and conduct 5-fold cross-validation on the remaining 20 topics. The average results of precision, recall, and F1 score are adopted as performance metrics.

Causal-TimeBank (Mirza and Tonelli, 2014) contains 184 documents from English news articles and 7,608 annotated event pairs. A total of 318 event pairs are annotated with causal relations, among which 300 causal relations are from intra-sentence event pairs and only 18 causal relations are from cross-sentence event pairs. Following the standard data splitting (Liu et al., 2020), we employ a 10-fold cross-validation evaluation and the average results of precision, recall, and F1 score are adopted as performance metrics. Following (Phu and Nguyen, 2021), we only conduct intra-sentence event causality identification experiments on CTB, as the number of cross-sentence event causal pairs is quite small.

4.2 Competitors

We compare our DAPrompt with the following competitors: ILP (Gao et al., 2019) uses integer linear programming to identify causal relations; RichGCN (Phu and Nguyen, 2021) uses a graph convolutional network to learn a document context-augmented representation of event-pairs; GESI (Fan et al., 2022) builds an event co-reference graph, ERGO (Chen et al., 2022) builds an event relational graph, CHEER (Chen et al., 2023) builds a heterogeneous event interaction graph and SENDIR (Yuan et al., 2023) constitutes a reasoning chain to identify event causal relations; KnowDis (Zuo et al., 2020), KnowMMR (Liu et al., 2020), LearnDA (Zuo et al., 2021b) and ECLEP (Pu et al., 2023) all use external knowledge to mine some event causality patterns; CauSeRL (Zuo et al., 2021a) adopts a contrastive strategy to transfer learned external causal statements; LSIN (Cao et al., 2021) uses a graph induction model learn external structural and relational knowledge; DPJL (Shen et al., 2022) leverages two derivative prompt tasks to identify causality; CF-ECI (Mu and Li, 2023) estimates context-keywords bias and event-pairs bias for causality counterfactual reasoning.

4.3 Parameter Setting

We implement the PLM models with their 768-dimension base version provided by the HuggingFace transformers[2] (Wolf et al., 2020), and run PyTorch [3] framework with CUDA on NVIDIA GTX 3090 GPUs. We set the mini-batch size to 16, the learning rate to 1e-5, the determine threshold ρ to 0.6, and all trainable parameters are randomly initialized from normal distributions. As the positive and negative samples are unbalanced, we adopt a random negative sampling with probability of 0.2 on the training dataset.

5 Results and Analysis

5.1 Overall Results

Table 1 and Table 2 compare the overall performance between our DAPrompt and the competitors on both ESC and CTB corpus. The competitors in Table 1 have reported all intra-sentence, cross-sentence, and overall results on ESC dataset; While the competitors in Table 2 has only reported the intra-sentence results on ESC dataset, respectively.

The first observation is that the ILP cannot obviously outperform the other competitors in Table 1. This might be attributed to the use of some graph-based neural networks, operating on the document-level graph structure with large-scale trainable parameters to augment event representation learning. Indeed, graph-based neural networks have been proven to be effective for many NLP tasks (Piao et al., 2022). We can also observe that the improvement of intra-sentence causality identification is more significant than that of cross-sentence. This might be attributed to the use of pre-trained language model for event node encoding, which can capture the semantic interaction between two events in a sentence.

The second observation is that the DPJL adopting the prompt learning paradigm can significantly outperform the other competitors in Table 2. The outstanding performance can be attributed to the task transformation for directly predicting a PLM vocabulary word, other than fine-tuning a downstream task-specific neural model upon a PLM. Although these competitors have used some kind of extra knowledge, such as lexicon knowledge and relational knowledge, from large-scale external

[2]https://github.com/huggingface/transformers
[3]pytorch.org

Model	Causal-TimeBank			EventStoryLine								
	Intra-Sentence			Intra-Sentence			Cross-Sentence			Overall		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
ILP (NAACL, 2019)	-	-	-	38.8	52.4	44.6	35.1	48.2	40.6	36.2	49.5	41.9
RichGCN (NAACL, 2021)	39.7	56.5	46.7	49.2	63.0	55.2	39.2	45.7	42.2	42.6	51.3	46.6
GESI (SIGIR, 2022)	-	-	-	-	-	50.3	-	-	49.3	-	-	49.4
ERGO (COLING, 2022)	62.1	61.3	61.7	57.5	72.0	63.9	51.6	43.3	47.1	48.6	53.4	50.9
CHEER (ACL, 2023)	56.4	69.5	62.3	56.9	69.6	62.6	45.2	52.1	48.4	49.7	53.3	51.4
SENDIR (ACL, 2023)	65.2	57.7	61.2	65.8	66.7	66.2	33.0	90.0	48.3	37.8	82.8	51.9
Our DAPrompt	66.3	67.1	**65.9**	64.5	73.6	**68.5**	59.9	59.3	**59.0**	61.4	63.7	**62.1**

Table 1: Overall results of comparison models for event causality identification on both ESC and CTB corpus.

Model	Causal-TimeBank			EventStoryLine		
	P	R	F1	P	R	F1
KnowDis (COLING,2020)	42.3	60.5	49.8	39.7	66.5	49.7
KnowMMR (IJCAI,2020)	36.6	55.6	44.1	41.9	62.5	50.1
CauSeRL (ACL,2021)	43.6	68.1	53.2	41.9	69.0	52.1
LSIN (ACL,2021)	51.5	56.2	53.7	47.9	58.1	52.5
LearnDA (ACL,2021)	41.9	68.0	51.9	42.2	69.8	52.6
DPJL (COLING,2022)	63.6	66.7	64.6	65.3	70.8	67.9
CF-ECI (ACL,2023)	50.5	59.9	54.8	47.1	66.4	55.1
ECLEP (ACL,2023)	50.6	63.4	56.3	49.3	68.1	57.1
Our DAPrompt	66.3	67.1	**65.9**	64.5	73.6	**68.5**

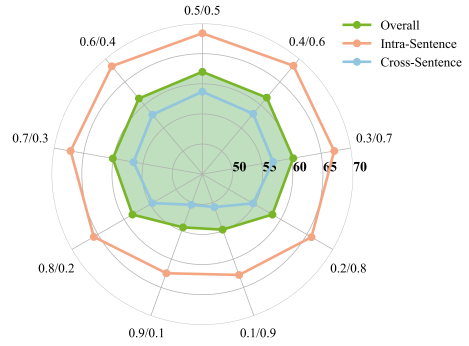
Table 2: Comparison of intra-sentence prediction results on the ESC and CTB corpus.

knowledge bases, the prompt learning model can better enjoy the encyclopedic linguistic knowledge embedded in a PLM during the model training.

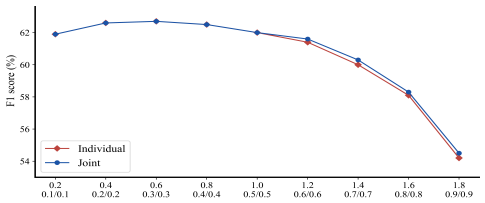
Finally, our DAPrompt (using DeBERTa as the PLM) has achieved significant performance improvements over all competitors in terms of much higher F1 score with all intra-sentence, cross-sentence, and overall event causality identification on both ESC and CTB datasets. We attribute its outstanding performance to our task transformation of evaluating the rationality of a deterministic assumption: We do not need to predict an unknown relation between events, no matter what kind of relations could be. Instead, we only need to evaluate the causal rationality via a deterministic assumption between two input events.

Decision Threshold: To examine the effectiveness of different decision threshold strategies, we conduct experiments on both individual threshold and joint threshold with different threshold values. The joint threshold strategy is that we use a joint decision variable $P_1 + P_2$ and a joint decision threshold ρ. The individual threshold strategy is that we use two *individual decision thresholds* ρ_1 and ρ_2 for P_1 and P_2, respectively. If $P_1 \geq \rho_1$ and $P_2 \geq \rho_2$, we accept the deterministic assumption that a causal relation exists between two events.

Fig. 3 (a) plots the performance of our DAPrompt (DeBERTa) using individual decision threshold in rationality evaluation on the ESC corpus. Each corner of the radar map represents a decision threshold ratio for two events, and the closer a point to the corner, the better performance of identifying event causality. It can be observed that DAPrompt achieves the best performance when the discrimination threshold is set equally for both events, i.e. (0.5/0.5); While DAPrompt suffers from an imbalance discrimination threshold setting, such as (0.1/0.9), (0.9/0.1), and etc. This indicates that the rationality of both events may be significant for identifying the causal relation between them. As we have no prior knowledge about the importance of each event, we simply sum their probabilities for rationality evaluation.



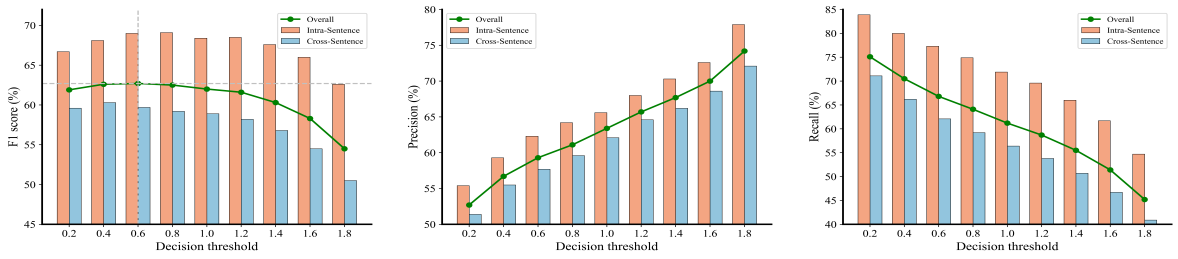
(a) Individual thresholds (Event-1/Event-2)



(b) Joint thresholds

Figure 3: Performance comparison between using individual threshold and joint threshold on the ESC corpus.

Fig. 3 (b) compares the overall performance of DAPrompt (DeBERTa) between using equal individual decision threshold and the joint decision threshold on the ESC corpus. It can be ob-



(a) F1 Score (b) Precision (c) Recall

Figure 4: Performance comparison of using different decision thresholds on the ESC corpus.

served that DAPrompt achieves nearly the same F1 score within a large range of the decision threshold (the range [0.2, 1.0] in the figure) using these two kinds of decision threshold settings. Yet the performance of DAPrompt using equal individual decision threshold cannot outperform the joint decision threshold when the decision threshold is set in the range of [1.0, 1.8]. This can be attributed to the flexibility of using a joint decision threshold, allowing two events to be identified as having a causal relation, even if one event has slightly lower rationality but the other event has higher rationality. For such considerations, we adopt the joint decision threshold in our DAPrompt.

Fig. 4 plots the performance of our DAPrompt against using different joint decision thresholds in rationality evaluation on the ESC corpus. We can observe that our DAPrompt achieves the best overall performance in terms of the F1 score when the discrimination threshold is set to 0.6. Yet the overall performance does not change much within a large range of the decision threshold (the range $[0.2, 1.0]$ in the figure). This suggests the wide applicability of our model for its not much sensitive to the decision threshold.

We can also observe from Fig. 4 that our DAPrompt suffers from either a very large or very small value of the discrimination threshold. Indeed, a small decision threshold relaxes the requirement for correctly predicting the input events, which thus admits too many assumed causal relations to be accepted. As such, the recall is high yet the precision is small. By contrast, a large decision threshold tightens the event prediction requirement, which only allows those event predictions with high confidence to accept a deterministic assumption. As such, the precision is high yet the recall is small. From our experiments, we suggest to take an empirical setting around 0.6 for the decision threshold.

5.2 Ablation Study

Pre-trained Language Model: In the prompt learning, using different PLMs may impact on the task performance. Table 3 compares the results of our proposed DAPrompt on the ESC corpus adopting the most representative PLMs, including BERT (Devlin et al., 2019) proposed by Google, RoBERTa (Liu et al., 2019) proposed by Facebook, ERNIE (Sun et al., 2019) proposed by Baidu, and DeBERTa (He et al., 2021) proposed by Microsoft.

Model	Intra-	Cross-	Overall
DAPrompt (BERT)	68.1	58.5	61.6
DAPrompt (RoBERTa)	68.3	58.1	61.3
DAPrompt (ERNIE)	68.1	56.9	60.7
DAPrompt (DeBERTa)	68.5	59.0	62.1

Table 3: Experiment results of using different PLM.

We can observe that our DAPrompt with all four PLMs has achieved better performance than the competitors. Even most of the competitors have used an advanced PLM like RoBERTa and BERT, to train an elaborate downstream task model or by adopting the prompt learning paradigm. This again validates the design objective of our deterministic assumption prompt learning, which pre-assumes the existence causal relation and next evaluates the assumption rationality, other than directly predicting the existence of causal relation between two events. We can also observe that using different PLMs do result in some performance variations and finally the DAPrompt (DeBERTa) has achieved the best performance. As such, we implement the remaining ablation experiments with DeBERTa.

Conventional Prompt Learning: To compare our DAPrompt with conventional prompt model, we conduct experiments on a conventional prompt model with different prompt designs for ablation study. Prompt is a conventional prompt model with discrete template and some answer words for

Model	Causal-TimeBank			EventStoryLine								
	Intra-Sentence			Intra-Sentence			Cross-Sentence			Overall		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Prompt	52.1	51.4	51.2	63.9	66.8	65.1	52.9	46.0	48.9	56.7	52.5	54.2
Prompt + VA	58.7	51.7	54.2	59.9	73.2	65.7	49.9	52.3	50.3	53.3	58.8	55.3
Prompt + CT	58.2	51.7	53.4	61.7	69.8	64.8	53.3	50.7	49.7	56.1	56.7	54.7
Prompt + VA + CT	55.9	56.4	55.9	62.0	70.3	65.5	52.5	50.8	51.0	55.7	56.9	55.8
DAPrompt w/ SiM	60.7	57.1	58.6	56.6	56.3	55.7	57.3	54.8	55.7	57.3	55.1	55.7
DAPrompt w/ ShM	64.6	59.1	61.3	59.4	75.1	66.0	56.2	65.1	59.5	57.3	68.2	61.6
DAPrompt w/ ET	22.3	12.1	14.6	60.5	42.7	49.6	39.2	38.6	38.6	44.3	39.9	41.7
DAPrompt full (ours)	66.3	67.1	**65.9**	64.5	73.6	**68.5**	59.9	59.3	**59.0**	61.4	63.7	**62.1**

Table 4: Experiment results of ablation study on both ESC corpus and CTB corpus.

prediction. Prompt+Virtual Answer (VA) uses virtual answer words in the conventional prompt model. Prompt+Continuous Template (CT) uses continuous template in the conventional prompt model. Prompt+VA+CT uses both virtual answer words and continuous template in the conventional prompt model

The first group of Table 4 presents the results using conventional prompt learning models. It is observed that the Prompt cannot outperform the Prompt+VA and Prompt+CT that use more representative virtual answer words for prediction and continuous template for automatically prompt template searching, respectively. The Prompt+VA+CT combining both the virtual answer and continuous template achieves better performance compare with the other conventional prompt models.

Although these conventional prompt learning models have employed some advanced techniques, viz. the virtual answer and continuous template, they still cannot outperform our DAPrompt learning. This again validates our new design style of deterministic assumption first and rationality evaluation next, rather than the conventional style of predicting an answer word first and mapping it to some relation.

Module ablation study: To examine the effectiveness of different modules in our DAPrompt, we design the following ablation study. DAPrompt w/ Single Mask (SiM) uses one mask with an event mention to predict the other event for rationality evaluation. DAPrompt w/ Shared MLM (ShM) uses one MLM head for answer prediction of two masks. DAPrompt w/ Event Tokens (ET) uses the probability of predicted event mention for rationality evaluation.

The second group of Table 4 presents the results of ablation modules. We observe that none of them can outperform the full DAPrompt model. This, however, is not unexpected. The DAPrompt w/ Sim misses one event's rationality for causality assumption evaluation; While the causal relation is between two events, thus both rationalities of two predicted events are useful for the assumption evaluation. The DAPrompt w/ ShM ignores the impact between two event predictions with one MLM classifier.

Besides, the inferior performance of the DAPrompt w/ ET may be attributed to the large number of different event description words in the dataset, leading to an unbalance answer label set and inadequate training process. From our statistics, the 5,334 and 6,813 annotated event mentions in ESC and CTB corpus are described by totally 1,656 and 2,045 different words or phrases respectively, and some of them contain very few instances. On the other hand, this also validates our design of using virtual event tokens of <E1> and <E2> to for events' representations.

6 Conclusion

This paper has designed a novel style of prompt learning for event casualty identification, that is, first deterministic assumption and next rationality evaluation, with the considerations of how to best utilize the encyclopedia-like knowledge embedded in a language model. We first assume the existence of causal relation between events and design a deterministic assumption template concatenating with the input event pair to predict event' tokens. We next use the probabilities of correctly predicted input events to evaluate the assumption rationality for the final event causality decision. Experiments on the ESC and CTB corpus validate our design objective in terms of significant performance improvements over all competitors and achieving the new state-of-the-art performance.

Limitation

• Considering the input length constraint of PLMs, the prompt template only contains event sentences; While the document-level semantics are not fused into the assumption evaluation for the ECI task.
• For fair comparison with the competitors, our DAPrompt does not identify the direction of event causalities. We will investigate this in future work.

Ethics Statement

This paper has no particular ethic consideration.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, page 86–90, Quebec, Canada.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. 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*, pages 4862–4872, Virtual.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop@ACL 2017*, pages 77–86, Vancouver, Canada.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: event relational graph transformer for document-level event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128, Gyeongju, Republic of Korea.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. Cheer: Centrality-aware high-order event reasoning network for document-level event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, page 10804–10816, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186, Minneapolis, MN, USA.

Chuang Fan, Daoxing Liu, Libo Qin, Yue Zhang, and Ruifeng Xu. 2022. Towards event-level causal relation identification. In *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1828–1833, Madrid, Spain.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. 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*, pages 1808–1817, Minneapolis, MN, USA.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations*, pages 1–23, Virtual Event, Austria.

Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3608–3614, Virtual, Japan.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):195:1–195:35.

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 preprint*, arXiv:1907.11692(1):1–13.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, pages 1–18, New Orleans, LA, USA.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 2097–2106, Dublin, Ireland.

Feiteng Mu and Wenjie Li. 2023. Enhancing event causality identification with counterfactual reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, page 967–975, Toronto, Canada. Association for Computational Linguistics.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490, Online.

Yinhua Piao, Sangseon Lee, Dohoon Lee, and Sun Kim. 2022. Sparse structure learning via graph neural networks for inductive document classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11165–11173, Virtual.

Ruili Pu, Yang Li, Suge Wang, Deyu Li, Jianxing Zheng, and Jian Liao. 2023. Enhancing event causality identification with event causal label and event pair interaction graph. In *Findings of the Association for Computational Linguistics: ACL 2023*, page 10314–10322, Toronto, Canada. Association for Computational Linguistics.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough. 2021. Open aspect target sentiment classification with natural language prompts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6311–6322, Punta Cana, Dominican Republic.

Shirong Shen, Heng Zhou, Tongtong Wu, and Guilin Qi. 2022. Event causality identification via derivative prompt joint learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2288–2299, Gyeongju, Republic of Korea.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 4444–4451, San Francisco, California, USA.

Yuan Sui, Shanshan Feng, Huaxiang Zhang, Jian Cao, Liang Hu, and Nengjun Zhu. 2022. Causality-aware enhanced model for multi-hop question answering over knowledge graphs. *Knowledge Based Systems*, 250(1):108943–108943.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *arXiv preprint*, arXiv:1904.09223(1):1–8.

Chengyu Wang, Jianing Wang, Minghui Qiu, Jun Huang, and Ming Gao. 2021. Transprompt: Towards an automatic transferable prompting framework for few-shot text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2792–2802, Punta Cana, Dominican Republic.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 1:1–34.

Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022. ConnPrompt: Connective-cloze prompt learning for implicit discourse relation recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911, Gyeongju, Republic of Korea.

Changsen Yuan, Heyan Huang, Yixin Cao, and Yonggang Wen. 2023. Discriminative reasoning with sparse event representation for document-level event-event relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, page 16222–16234, Toronto, Canada. Association for Computational Linguistics.

Kun Zhao, Donghong Ji, Fazhi He, Yijiang Liu, and Yafeng Ren. 2021. Document-level event causality identification via graph inference mechanism. *Information Science*, 561(1):115–129.

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. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 2162–2172, Virtual.

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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3558–3571, Virtual.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1544–1550, Barcelona, Spain.