Enhancing Event Causality Identification with LLM Knowledge and Concept-Level Event Relations

Ya Su[1], Hu Zhang[1,2]*, Guangjun Zhang[1], Yujie Wang[1], Yue Fan[1], Ru Li[1,2], Yuanlong Wang[1,2]

[1]School of Computer and Information Technology, Shanxi University, Taiyuan, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, China
su_ya6990@163.com, zhanghu@sxu.edu.cn, zgj2866@gmail.com, init_wang@foxmail.com, yuefan24@163.com, {liru,ylwang}@sxu.edu.cn

Abstract

Event Causality Identification (ECI) aims to identify fine-grained causal relationships between events in an unstructured text. Existing ECI methods primarily rely on knowledge-enhanced and graph-based reasoning approaches, but they often overlook the dependencies between similar events. Additionally, the connection between unstructured text and structured knowledge is relatively weak. Therefore, this paper proposes an ECI method enhanced by LLM Knowledge and Concept-Level Event Relations (LKCER). Specifically, LKCER constructs a conceptual-level heterogeneous event graph by leveraging the local contextual information of related event mentions, generating a more comprehensive global semantic representation of event concepts. At the same time, the knowledge generated by COMET is filtered and enriched using LLM, strengthening the associations between event pairs and knowledge. Finally, the joint event conceptual representation and knowledge-enhanced event representation are used to uncover potential causal relationships between events. The experimental results show that our method outperforms previous state-of-the-art methods on both benchmarks, EventStoryLine and Causal-TimeBank.

1 Introduction

Understanding causality like humans is crucial for successful natural language processing (NLP) applications, especially in high-risk fields such as finance and healthcare. The ECI task aims to comprehend fine-grained causal relationships between events in unstructured text. It has broad applications in NLP, including machine reading comprehension (Berant et al., 2014), why-question answering (Oh et al., 2017, 2016), and more. However, the causal relationship between events often lacks explicit causal clues, and complex associations between events across different sentences exist. Additionally, the combinations of different event pairs exhibit various dependency relations within the same sentence. Accurately identifying whether these events have causal relationships remains a crucial challenge to be addressed. As shown in Figure 1(a), the four sentences come from the same document segment, where e_i represents the i-th event. Based on the inputs and event information, the ECI model needs to identify the causal relationship chain as illustrated in Figure 1(b). For example, in the sentence-level ECI task, given the input sentence S_1 and events $\{e_1, e_2, e_3, e_4, e_5, e_6\}$are required to identify five combinations of intra-sentence causal event pairs <$e_1, cuase, e_2$>, <$e_1, cuase, e_6$>, <$e_6, cuase, e_2$>, <$e_6, cuase, e_5$>, <$e_5, cuase, e_2$>.

Existing studies (Zuo et al., 2021; Ding et al., 2024; Wu et al., 2023) usually directly match event mentions with entity information from external knowledge bases. However, these studies have the following shortcomings: **(1) They pay less attention to the dependency relationships between similar events.** As shown in Figure 1(a), event mentions of the same colour indicate that they share similar semantics and belong to the same event concept. For example, the event mentions "following/followed" in the dataset belong to the event concept "follow". An event concept typically includes multiple semantically and formally similar event mentions, which have coreference relationships between these event mentions. However, existing methods tend to focus only on the local context of event mentions within sentences, neglecting the global semantic information of the event concept within the document, leading to inconsistencies when predicting causal relationships between multiple event-mention pairs. **(2) The connection between unstructured texts and structured knowledge has been relatively weak, and the quality of the knowledge is not high.** Most existing methods directly use event mentions in

*Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics, pages 7403–7414
January 19–24, 2025.

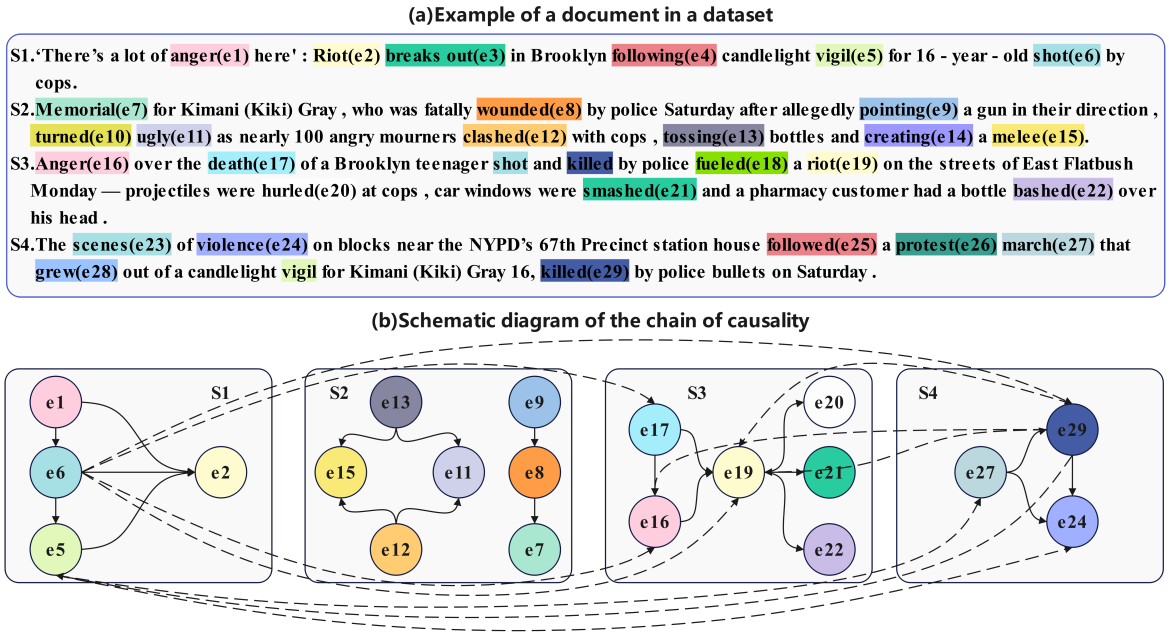

Figure 1: Schematic diagram of event pair interactions in ECI. (a) Example of a document in the dataset that contains 4 sentences with 29 events, different event mentions of the same colour indicate that these event mentions belong to the same event concept. (b) Schematic diagram of causal chains, where solid lines indicate causal interactions of events within sentences and dashed lines indicate causal relationships of events between sentences.

Data set	ESC	CTB
Event Count	5334	6881
Covered Event Count	1547	2202
Coverage	29%	32%

Table 1: Coverage of events in ConceptNet.

the dataset as keywords for matching in external knowledge bases, without considering the coverage of events in the dataset. As shown in Table 1, statistical analysis of the EventStoryLine (ESC) (Caselli and Vossen, 2017) and Cause-TimeBank (CTB) (Mirza and Tonelli, 2014) reveals that about 70% of the events cannot obtain relevant knowledge from external knowledge bases. In addition, commonsense knowledge bases are typically represented in structured triples (head entity, relation, tail entity), but the events and their contexts are unstructured texts, resulting in heterogeneity between them (Bian et al., 2021). Meanwhile, the relationships between entities in knowledge graphs are complex and diverse. In particular, implicit causality often lacks clear causal signal words, and inadequate filtering can introduce noise into the training data.

To address the above limitations, we propose the LKCER method, which introduces the event conceptual information and rich event knowledge to enhance the representations of event pairs. At the same time, it enhances the model's ability to identify causal relationships by distinguishing between different categories of causal relationships. Figure 2 illustrates the overall framework of the LKCER method. First, we construct a concept-level event heterogeneous graph, which links multiple event mentions belonging to the same event concept within the document, thus capturing the global semantic information of event concepts. Next, we propose a multi-scale hybrid matching strategy to align events in the dataset with entity information from ConceptNet (Speer et al., 2017). Meanwhile, we use COMET (Bosselut et al., 2019) to generate structured knowledge related to events and further refine and expand it using the LLM. Finally, we design three joint prompt templates for explicit and implicit causal relationships and predict the probability distributions at masked positions using prompt learning methods. The main contributions of this paper are as follows:

- We constructed a conceptual-level event heterogeneous graph that links the local contextual information of events mentioned within a sentence to the global semantic information of the event concept in the document.
- We enhanced the quality of knowledge and improved the connection between unstructured

text and structured knowledge by filtering and expanding structured knowledge through the inductive reasoning capabilities of LLMs.

- On two ECI benchmark datasets, the LKCER method achieved improvements of 2.3% and 2.7% in F1 scores compared to the SOTA methods, demonstrating superior performance among all baseline methods.

2 Related Work

Event extraction, a key technology for extracting events from unstructured text, has made significant progress in recent years (Wang et al., 2021; Ren et al., 2023; Zhang et al., 2024a). As a crucial downstream task of event extraction, the ECI task initially relied on pattern-matching methods (Ittoo and Bouma, 2011; Hashimoto et al., 2014), which improved performance through lexical and syntactic features (Beamer and Girju, 2009; Riaz and Girju, 2014). With the advancement of deep learning technologies, modern approaches are generally categorized into knowledge enhancement, graph-based reasoning, and prompt adjustment. Recent studies (Zuo et al., 2021; Wu et al., 2023) have begun to explore integrating external knowledge bases (such as ConceptNet and WordNet) into models to enrich event information and enhance model performance. Some graph-based reasoning studies (Tran Phu and Nguyen, 2021; Pu et al., 2023) model ECI as a graph-based node classification or edge prediction problem. However, relying solely on sentence-internal semantic information often fails to capture the deep semantics of events. To address this, some studies (Cao et al., 2021; Ding et al., 2024; Huang et al., 2024) have combined graph-based reasoning with knowledge enhancement, mapping out the relationships between events and various types of knowledge to further enrich event information and improve performance. Additionally, some research (Liu et al., 2023; Shen et al., 2022) has used prompt adjustment to capture implicit causal relationships between events, thereby enhancing ECI effectiveness.

3 Methodology

3.1 Task Definition

Given a document $D = \{w_1, w_2, \ldots, w_k\}$ and its set of event mentions $M = \{m_1, m_2, \ldots, m_n\}$, where D consists of multiple sentences $\{s_1, s_2, \ldots, s_i\}$, and each s_i contains multiple event concepts $E = \{e_1, e_2, \ldots, e_j\}$. As shown in Figure 1(b), event concepts within the same document are often composed of multiple event mentions $e_i = \{m_1, m_2, \ldots, m_k\}$, where m_i may correspond to multiple words in D. These mentions may include synonyms, different forms of words (such as nouns, verbs and so on), and words where case variations do not affect the meaning. The ECI model needs to identify the causal relationships between any two events $(e_i, e_j)(i \neq j)$ in the E based on the input s_i and E. Figure 2 illustrates the framework of the proposed LKCER method, which primarily consists of three components: concept-level event heterograph(Section 3.2), knowledge generation(Section 3.3), joint prompt learning(Section 3.4). The following sections provide a detailed explanation of each component.

3.2 Conceptual-Level Event Heterogeneous Graph

An event concept typically includes multiple semantically and formally similar event mentions, which have coreference relationships between these event mentions. To enhance the model's understanding of event concepts, we construct a heterogeneous graph $\mathcal{G}$ that includes event mention nodes and event concept nodes. As illustrated in Figure 2(a), given an input document D, the synsets from WordNet (Miller, 1995) are used to obtain multiple event mentions corresponding to the event concept in the document, handling related irregular verbs and case variations. As shown in Equations 1, we use the RoBERTa (Liu et al., 2019) as an encoder to extract the hidden contextual representations of the D. As shown in Equations 2, we initialize the node representation h_{m_i} of the event mention m_i using the word sequence vectors $\{h_{w_b}, \ldots, h_{w_a}\}$ corresponding to the event mention. As shown in Equations 3, we apply Mix Pooling (Yu et al., 2014) the vectors of multiple event mention nodes to obtain the representation vector of an event concept node, where λ is a random value of either 0 or 1.

$$h_{w_i} = RoBERTa\,(w_i) \tag{1}$$

$$h_{m_i} = \frac{1}{a-b+1}\sum_{b}^{a}(h_{w_b}, \ldots, h_{w_a}) \tag{2}$$

$$h_{e_i} = \lambda \max_{m_j \in e_i} h_{m_j} + (1-\lambda)\frac{1}{|e_i|}\sum_{m_j=e_i}^{a} h_{m_j} \tag{3}$$

We introduced three types of edges to capture the interactions between sentences and event concepts. **Mention-Mention Edge**(as shown in Figure 2(a1)): multiple event mention nodes belonging to the same event concept are connected to enhance the semantic consistency of the same event. **Mention-Event Edge**(as shown in Figure 2(a2)): the event mention nodes are connected to their corresponding event concept nodes, aggregating into the global semantics of the event concept. **Event-Event Edge**(as shown in Figure 2(a3)): when event mentions within an event concept appear in different sentences, edges are constructed between the event concept nodes in different sentences based on co-occurrence relationships. Finally, RGCN (Chen et al., 2019) is used to transmit semantic information between nodes in the heterogeneous graph, as shown in Equation 4. The attention mechanism fuses the semantics information of multiple event mentions contained in the event concept to obtain the representation vector $h_{e_i}^{m_j}$ of the event concept.

$$h_{e_i}^{m_j} = \sum_{m_j \in M(e_i)} \frac{exp\left(h_{m_j}^T h_{e_i}\right)}{\sum_{m_n \in M(e_i)} \exp\left(h_{m_n}^T h_{e_i}\right)} h_{m_j} \tag{4}$$

3.3 Knowledge Generation

Event matching: As shown in Table 1, events in the benchmark datasets for the ECI task suffer from a low success rate indirectly matching with ConceptNet knowledge base nodes. Therefore, this paper proposes a multi-scale hybrid matching strategy to map events to knowledge base nodes more accurately. Figure 2(b1) shows that the specific matching rules are: **(1) Direct Matching Phase:** Check whether the event description exactly matches a concept in the knowledge base. If the match is successful, it is set as the event concept. **(2) Tokenization Matching Phase:** If the match fails in phase 1 and the length of the event description is no longer than one word, tokenization matching is performed. In this phase, the Unigram Language Model (Kudo, 2018) is used to extract the primary form of the event description, which is then matched with concepts in the knowledge base. If the match is successful, it is set as the event concept. **(3) Word Embedding Matching Phase:** This phase occurs under two conditions: the match fails in phase 1 and the event description is longer than one word, or the match fails in phase 2. In this phase, the cosine similarity between the event description and the embeddings of the concepts in the knowledge base is calculated, and the concept with the highest similarity is selected as the event concept.

Knowledge generation: To generate more accurate and rich event knowledge, we use COMET, which can generate loosely structured open-domain knowledge descriptions with an accuracy of 91.7% on ConceptNet. As shown in Figure 2(b2), we input the event concept and specified relations into COMET to generate event-related knowledge. For example, for the generated knowledge path $\{k_1, k_2, \ldots, k_q, \ldots\}$, where k_p is the q-th knowledge triplet, the template-based transformation algorithm (Bian et al., 2021) is used to convert each triple into sentences $\{s_1, s_2, \ldots, s_q, \ldots\}$ that describing its content, where sentence s_q describes the triplet k_q.

Knowledge filtering and expansion: To further filter and calibrate the generated knowledge and uncover implicit knowledge contained in the LLM, we use commonsense knowledge as prompts and leverage the LLM's learning capabilities for knowledge filtering and expansion. We decompose the problem and use multi-stage prompting (multiple rounds of input-output) to explicitly provide background knowledge, follow-up questions, and intermediate answers in the prompts, guiding the LLM in reasoning and generating the required event causality knowledge. As shown in Figure 2(b3), in the first stage, the event commonsense knowledge $\{s_1, s_2, \ldots, s_q, \ldots\}$ generated by COMET is used as the background knowledge. Then, the LLM filters, summarizes and supplements valuable knowledge for the ECI. In the second stage, document D, containing the event and the knowledge generated in the first step, is re-input into the model to generate event causality knowledge.

3.4 Joint Prompting

In the ECI task, events have different types of causality categories (explicit causality and implicit causality). Based on these distinct causality characteristics, we designed three different prompt templates and labeled words to utilize task-related knowledge effectively. Specifically, as shown in Figure 2(c), given the input sequence $x_{in} = [CLS] \oplus s_i \oplus [SEP]$, the source event e_s and target event e_t form an event pair (e_s, e_t), where $\oplus$ represents the concatenation operation. The formalization of the three prompt templates is shown in Appendix A. $ECI(x_{in})$ represents a prompt template

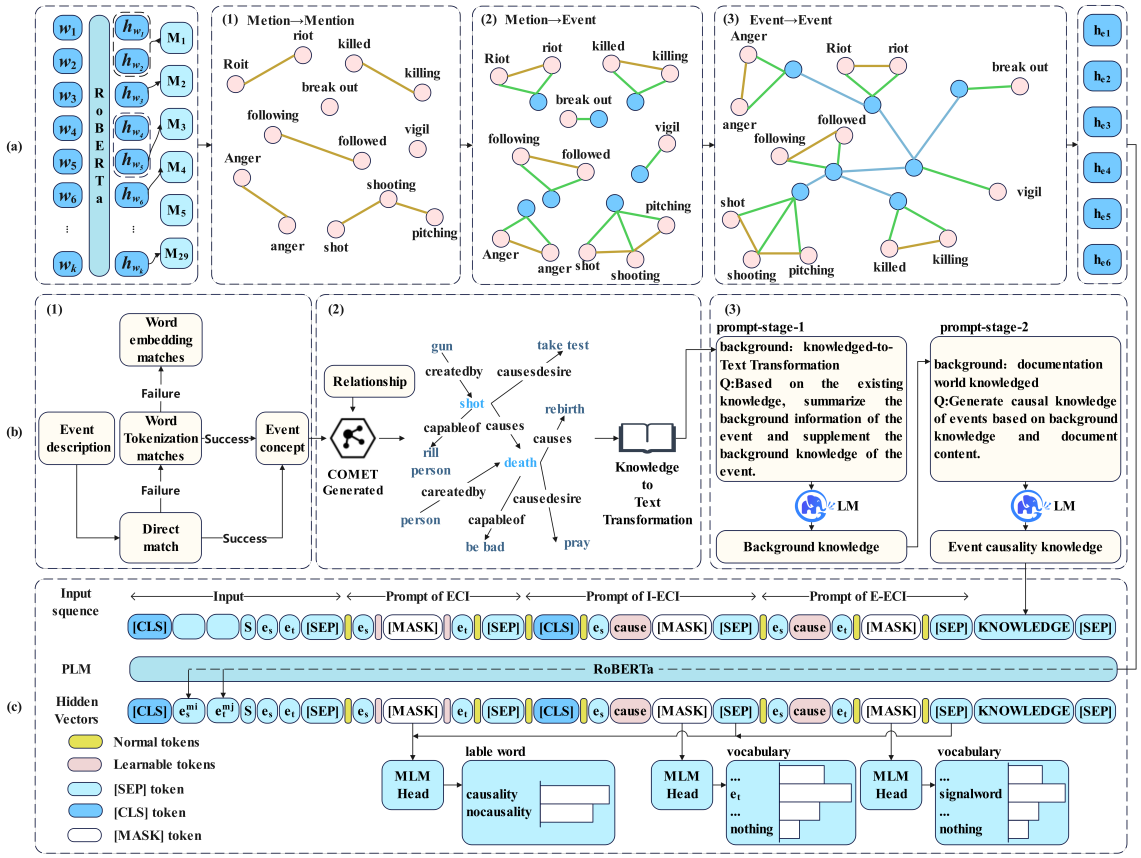

Figure 2: The overall framework of LKCER. (a) represents the construction process of the concept-level event heterogeneous graph, (b) represents the knowledge generation process, and (c) represents the joint prompt learning process.

that does not distinguish between types of causal relationships, and the set of causality label words is defined as V_{ECI}:$\{Causality, noCausality\}$. As shown in Equation 5, the probability distribution of $V(y)$ at position $[MASK]$ is used to determine the probability of the causal label $p(y|x_{in})$, where R is the RoBERTa, and $w_{v(\bullet)}$ reuses the weights from RoBERTa; $E\text{-}ECI(x_{in})$ represents the explicit causality prompt template, which aims to detect whether causal signal words exist in x_{in}. By identifying these signal words, the model determines the explicit causal relationship between events; $I-ECI(x_{in})$ represents the implicit causality prompt template, which aims to detect x_{in} events causally related to the source event e_s. The label words for both $E-ECI(x_{in})$ and $I-ECI\,(x_{in})$ are $V_{I/E-ECI}$:$\{w_1, w_2, ..., w_k, nothing\}$.

$$p\Big(R\Big(\nu(y)|[x_{in}+ECI(x_{in})]\Big)\Big) = \frac{\exp(w_{\nu(y)}h_{[MASK]})}{\sum_{y'\in\gamma}\exp(w_{\nu(y')}h_{[MASK]})} \quad (5)$$

To enrich the causal event information, we incorporate the knowledge-enhanced event representation from Section 3.3 and the event concept representation from Section 3.2. We use RoBERTa's MLM head with the joint prompt learning method to predict the probability distribution at the masked positions, which is then used as the result. These three prompt tasks share semantic information, effectively connecting the ECI task and joint prompt learning.

3.5 Training and Prediction

For any two events $(e_s, e_t)(s \neq t)$ in the given text $D = \{s_1, s_2, ..., s_n\}$, their final feature representation $h_{d_{s,t}}$ includes: (1) sentence s_i feature representation; (2) source event e_s feature and target event e_t feature; (3) multiple event mention representations h_{e_i} contained in the event concept; (4) corresponding generated knowledge representation; (5) prompt templates with learnable labels, where the label words of the prediction model are represented as [MASK]. We use the LSTM model to process the input features. The probability p_{st} of predicting the source event and target event is

as shown in Equation 6, where W_{st} and b_{st} are learnable parameters.

$$p_{st} = softmax\big(W_{st} h_{d_{s,t}} + b_{st}\big) \quad (6)$$

As shown in Equation 7, we use the cross-entropy loss function to obtain $\mathcal{L}_{\text{ECI}(x_{in})}$, where D^* represents the training sample, and $\tilde{p}_{st} \in (0, 1)$ denotes the ground truth label of event pair (e_s, e_t). The losses of $E - ECI(x_{in})$ and $I - ECI\,(x_{in})$ are multiplied by $\lambda \in \big(0, 1\big)$, then added to the ECI loss to obtain the final loss $\mathcal{L}$.

$$\mathcal{L}_{\text{ECI}(x_{in})} = -\sum_{s \in D^*} \sum_{e_s \neq e_t} \tilde{p}_{st} - (1 - \tilde{p}_{st})\, log\,(1 - p_{st}) \quad (7)$$

$$\mathcal{L}_{all} = \mathcal{L}_{ECI(x_{in})} + \lambda(\mathcal{L}_{I-ECI(x_{in})} + \mathcal{L}_{E-ECI(x_{in})}) \quad (8)$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We validate our method on two benchmark datasets: ESC and CTB. The specific information related to the dataset can be found in Appendix B. Following the same splits as previous studies (Shen et al., 2022; Liu et al., 2023), we perform 5-fold cross-validation on ESC and 10-fold cross-validation on CTB. In addition, we use Precision (P), Recall (R), and F1-score (F1) as evaluation metrics.

4.2 Parameter Settings

We use the pre-trained language model RoBERTa-base as the base model to encode the input sequences. The newly added tokens in RoBERTa, such as knowledge and graph features, are all 768-dimensional embeddings. ChatGLM-6B (Zeng et al., 2022) is used as the LLM for generating event causality knowledge. We use AdamW (Loshchilov and Hutter, 2017) as the optimization algorithm for the model. We set the learning rate of the pre-trained parameters to 1e-5 and the learning rate of the newly added parameters to 1e-4. The batch size is set to 6, the number of RGCN (Chen et al., 2019) layers is set to 2.

4.3 Baselines

To demonstrate the effectiveness of this work, we compare our method with previous state-of-the-art models. For the ESC and CTB datasets, we chose the following baselines for comparison. **Feature-based methods:** (1) **Seq** (Choubey and Huang, 2017), a model for partitioning event temporal relationships in ECI. (2) **DD** (Mirza and Tonelli, 2014), a data-driven method. **Knowledge-augmented methods:** (1) **KnowDis** (Zuo et al., 2020), a knowledge-enhanced distantly supervised method for ECI. (2) **LearnDA** (Zuo et al., 2021), a learnable knowledge-guided data augmentation method (3) **LSIN** (Cao et al., 2021), a method that induces structured knowledge into networks to enhance ECI. (4) **DPF** (Huang et al., 2024), a method for integrating task-specific knowledge from commonsense graphs into ECI. **Graph neural network-based methods:** (1) **RichGCN** (Tran Phu and Nguyen, 2021), a GCN-based document-level ECI model. (2) **SemSIn** (Hu et al., 2023), a semantic structure network-based method for ECI. (3) **ECLEP** (Pu et al., 2023), a method that enhances ECI using event pair interaction graphs. (4) **GCKAN** (Ding et al., 2024), a method that enhances ECI using graph contrastive learning. **Prompt-adjusted methods:** (1) **KEPT** (Liu et al., 2023), a knowledge-augmented and prompt-adjusted method for ECI. (2) **DPJL** (Shen et al., 2022), a prompt-adjusted approach for enhancing ECI.

4.4 Main Result

Table 2 shows our experimental results on the ESC and CTB datasets. Experimental results show that our proposed LKCER method outperforms all baseline methods, improving the F1 score by 1.9% and 2.7% over previous SOTA methods on two datasets, respectively. Specifically, the $\text{LKCER}_{\text{single}}$ performs better on the CTB dataset, with an F1 score 3.9% higher than the DPJL method, while the $\text{LKCER}_{\text{muti}}$ performs better on the ESC dataset. This may be because the ESC dataset has more annotated data, which helps identify causal signal words and distinguish between explicit and implicit causal relationships, thereby enabling better training of label word embeddings in the $\text{LKCER}_{\text{muti}}$ method. On the other hand, the limited annotated data in the CTB dataset may cause excessive prompt templates to introduce additional noise. Overall, these results demonstrate that the LKCER method significantly enhances the performance of the ECI task.

5 Analysis

5.1 The Effect of LKCER

To better understand the advantages of the LKCER method, we attribute its performance improvements in the ECI task to the following factors:

Methods	Model	EventStoryLine			Cause-TimeBank		
		P	R	F1	P	R	F1
Feature-based methods	DD(Mirza and Tonelli, 2014)	-	-	-	67.3	22.6	33.9
	Seq (Choubey and Huang, 2017)	32.7	44.9	37.8	-	-	-
Knowledge-augmented methods	KnowDis (Zuo et al., 2020)	39.7	66.5	49.7	42.3	60.5	49.8
	LearnDA(Zuo et al., 2021)	42.2	69.8	52.6	41.9	68.0	51.9
	LSIN(Cao et al., 2021)	47.9	58.1	52.5	51.5	56.2	53.7
	DPF(Huang et al., 2024)	55.9	69.8	62.1	53.7	64.2	58.5
Graph-based methods	RichGCN(Tran Phu and Nguyen, 2021)	49.2	63.0	55.2	39.7	56.5	46.7
	SemSIn(Hu et al., 2023)	50.5	63.0	56.1	52.3	65.8	58.3
	ECLEP(Pu et al., 2023)	49.3	68.1	57.1	50.6	63.4	56.3
	GCKAN(Ding et al., 2024)	50.9	60.6	55.3	52.2	60.7	56.1
Prompt-adjusted methods	KEPT(Liu et al., 2023)	50.0	68.8	57.9	48.2	60.0	53.5
	DPJL(Shen et al., 2022)	65.3	70.8	67.9	63.6	66.7	64.6
$\text{Prompt}_{\text{single}}$	**$\text{LKCER}_{\text{single}}$(ours)**	67.1	69.5	68.1	**64.5**	75.4	**68.5**
$\text{Prompt}_{\text{muti}}$	**$\text{LKCER}_{\text{muti}}$(ours)**	**67.3**	**72.7**	**69.8**	61.0	**76.3**	67.3

Table 2: Experimental Results on the ESC and CTB Datasets(%).

(1) Knowledge-enhanced methods improve the performance of the ECI task effectively by integrating PLMs, compared to purely feature-based methods. However, relying solely on event mentions in the dataset as keywords to match with external knowledge bases not only result in a low matching success rate but also tend to introduce noise, leading to information imbalance between events. Experimental results show that the LKCER method using knowledge generation methods is effective in improving the performance of ECI on $\text{Prompt}_{\text{single}}$ alone. Accurate event knowledge improves the ability to reason about relationships between events.

(2) The LKCER method significantly outperforms graph-based methods, as most previous methods primarily used retrieved external knowledge to construct graphs or introduce additional causal labels to distinguish events, while neglecting the coreference relationships between event mentions within the same document. The LKCER method constructs graphs based on the relationships between events within a document, deeply exploring the global semantic connections between event concepts, thereby enhancing the performance of ECI.

(3) The LKCER method, using only the $\text{Prompt}_{\text{single}}$ module, significantly outperforms the KEPT method, which combines knowledge enhancement with a single prompt. In addition to more precise knowledge matching, the LKCER method also addresses the heterogeneity between structured knowledge and unstructured text. It further leverages LLM to filter and inductively generate causal knowledge, enriching event information.

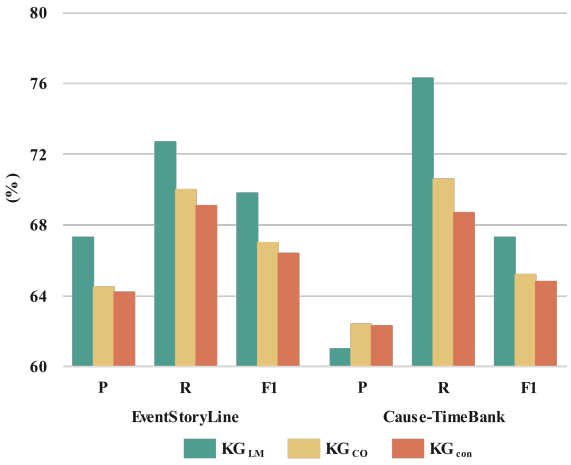

Figure 3: The effect of different types of knowledge on the LKCER method.

5.2 Ablation Study

This section analyzes the contribution of each module in the $\text{LKCER}_{\text{muti}}$ and $\text{LKCER}_{\text{single}}$ methods through ablation experiments, with the results shown in Table 3. We examined the following ablation models: **- CE:** removing the concept-level event heterogeneous graph module. After removing this module, the results show a decline in overall performance, indicating that it helps the model better learn the global semantic connections between event concepts. Additionally, the performance drop is more pronounced in the $\text{LKCER}_{\text{muti}}$ method compared to $\text{LKCER}_{\text{single}}$ method. This

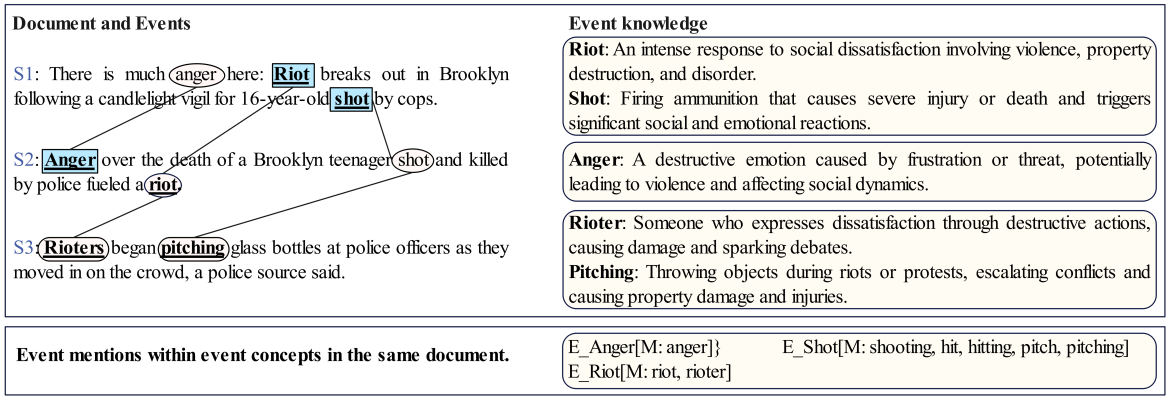

Figure 4: Case Study. The upper part of the figure shows a portion of sentences from the same document, where S_i represents the sentence number. The event pairs that require causal relationship identification are underlined and highlighted in bold black. Boxes represent event concepts, and connected circles represent event mentions with semantic similarity. Repeated knowledge generated from the same event is shown only once on the right side. The lower part of the figure displays all event mentions contained within the event concepts from the same document, where E represents the event concepts and M represents the event mentions.

may be due to the lack of relational information between event concepts, which affects the model's ability to effectively integrate semantic information across multiple prompt templates, thus impacting its performance in capturing complex causal relationships. **- KG:** removing all knowledge. The experimental results demonstrate that the $\text{LKCER}_{\text{muti}}$ method experiences a more significant performance drop on the ESC dataset, while the $\text{LKCER}_{\text{single}}$ method shows a more significant decline in the CTB dataset. This suggests that after removing the knowledge, multiple prompt templates can partially alleviate the problem of limited annotated data in the CTB dataset through diversity in prompts. However, in the more richly annotated ESC dataset, noise and semantic redundancy may reduce the model's ability to capture semantic consistency across templates. This phenomenon further confirms the critical role of rich knowledge matching in event relationship reasoning, particularly in scenarios with limited labeled data, where unstructured knowledge is crucial to improving the performance of ECI models.

5.3 Effect of Different Knowledge Types

To validate the effect of different types of knowledge on the ECI task, we conducted tests using the LKCER method on two datasets while keeping other settings fixed, with the results shown in Figure 3. KG_{LM} refers to using only unstructured causal knowledge generated by LLM; KG_{CO} refers to structured knowledge generated by the COMET model with precise matching; KG_{con} refers to directly using structured knowledge from ConceptNet. The results indicate that although the LKCER method, using ConceptNet knowledge with lower event coverage, achieves some performance improvement compared to the KEPT method within the same knowledge base, the knowledge generated through accurate event matching performs significantly better. Specifically, the unstructured knowledge expanded and filtered by the LLM enables LKCER method to achieve best performance.

Methods	EventStoryLine			Cause-TimeBank		
	P	R	F1	P	R	F1
$\text{LKCER}_{\text{muti}}$	67.3	72.7	69.8	61.0	76.3	67.3
-CE	67.1	67.7	68.6	63.5	71.9	66.1
-KG	65.0	67.3	66.1	62.3	68.7	64.8
$\text{LKCER}_{\text{single}}$	67.1	69.5	68.1	64.5	75.4	68.5
-CE	64.9	70.4	67.3	64.2	72.9	67.9
-KG	66.2	67.3	66.6	62.6	69.0	65.1

Table 3: Ablation Results on the ESC and CTB Datasets(%).

5.4 Case Study

This section demonstrates the effectiveness of the LKCER method and the contribution of each module through a case study. As shown in Figure 4, the three sentences are extracted from the same document in the dataset. The analysis reveals that semantically similar event mentions create connections between the sentences. For example, the event pair to be identified in S1 is (Riot, shot), where "Riot" in S1 is semantically similar to "riot"

in S2. We leverage the related event information in S2 to supplement S1. Additionally, incorporating knowledge generated by the LLM enriches event representations, aiding the model in learning and understanding complex causal relationships between events.

6 Conclusion

We propose an ECI method enhanced by LLM knowledge and conceptual-level event relationships. This method introduces the event concept information and precise event knowledge to enrich the representation of event pairs, while also improving the model's ability to identify causal relationships by distinguishing between different categories of these relationships. The experimental results on two widely used datasets show that our method performs exceptionally well in ECI tasks.

7 Limitations

In this paper, we focus solely on whether a causal relationship exists between given events, without delving into the potential causal features in unannotated data. At the same time, this paper only implements sentence-level ECI, while document-level ECI tasks still face numerous challenges. In addition, LKCER remains a black-box model, thus exploring the interpretability of ECI models, such as self-explaining rationalization (Zhao et al., 2024; Lei et al., 2016) and structured explanation (Fan et al., 2024; Zhang et al., 2024b; Song et al., 2024), represents a promising direction for research. These aspects will serve as critical focuses for our future work.

Acknowledgements

We thank all the anonymous reviewers for their constructive comments and suggestions. This work is supported by the National Natural Science Foundation of China (62176145, 62476161).

References

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '09, page 430–441, Berlin, Heidelberg. Springer-Verlag.

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 *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1499–1510.

Ning Bian, Xianpei Han, Bo Chen, and Le Sun. 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12574–12582.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 4862–4872, Online. Association for Computational Linguistics.

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*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Junjie Chen, Hongxu Hou, Jing Gao, Yatu Ji, and Tiangang Bai. 2019. Rgcn: recurrent graph convolutional networks for target-dependent sentiment analysis. In *International Conference on Knowledge Science, Engineering and Management*, pages 667–675. Springer.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1796–1802, Copenhagen, Denmark. Association for Computational Linguistics.

Ling Ding, Jianting Chen, Peng Du, and Yang Xiang. 2024. Event causality identification via graph contrast-based knowledge augmented networks. *Information Sciences*, 656:119905.

Yue Fan, Hu Zhang, Ru Li, YuJie Wang, Hongye Tan, and Jiye Liang. 2024. FRVA: Fact-retrieval and verification augmented entailment tree generation for explainable question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9111–9128, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. 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)*, pages 987–997, Baltimore, Maryland. Association for Computational Linguistics.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023. Semantic structure enhanced event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10901–10913, Toronto, Canada. Association for Computational Linguistics.

Peixin Huang, Xiang Zhao, Minghao Hu, Zhen Tan, and Weidong Xiao. 2024. Distill, fuse, pre-train: Towards effective event causality identification with commonsense-aware pre-trained model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5029–5040, Torino, Italia. ELRA and ICCL.

R.A. Ittoo and G. Bouma. 2011. Extracting implicit and explicit causal relationships from sparse, domain-specific texts. In *16th International Conference on Applications of Natural Language to Information Systems*, volume 6716 of *Lecture Notes in Computer Science*, pages 52 – 63. Springer. 2011/g.bouma/pub004.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. 2023. Kept: Knowledge enhanced prompt tuning for event causality identification. *Knowledge-Based Systems*, 259:110064.

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.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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 COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. 2016. A semi-supervised learning approach to why-question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Jong-Hoon Oh, Kentaro Torisawa, Canasai Kruengkrai, Ryu Iida, and Julien Kloetzer. 2017. 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*, pages 415–424.

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*, pages 10314–10322, Toronto, Canada. Association for Computational Linguistics.

Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. 2023. Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–306, Toronto, Canada. Association for Computational Linguistics.

Mehwish Riaz and Roxana Girju. 2014. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 161–170, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

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. International Committee on Computational Linguistics.

Junyue Song, Xin Wu, and Yi Cai. 2024. Step feasibility-aware and error-correctable entailment tree generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15296–15308, Torino, Italia. ELRA and ICCL.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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. Association for Computational Linguistics.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. CLEVE: Contrastive Pre-training for Event Extraction. 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)*, pages 6283–6297, Online. Association for Computational Linguistics.

Sifan Wu, Ruihui Zhao, Yefeng Zheng, Jian Pei, and Bang Liu. 2023. Identify event causality with knowledge and analogy. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.

Dingjun Yu, Hanli Wang, Peiqiu Chen, and Zhihua Wei. 2014. Mixed pooling for convolutional neural networks. In *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24-26, 2014, Proceedings 9*, pages 364–375. Springer.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, P. Zhang, Yuxiao Dong, and Jie Tang. 2022. Glm-130b: An open bilingual pre-trained model. *ArXiv*, abs/2210.02414.

Guangjun Zhang, Hu Zhang, YuJie Wang, Ru Li, Hongye Tan, and Jiye Liang. 2024a. Hyperspherical multi-prototype with optimal transport for event argument extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9271–9284, Bangkok, Thailand. Association for Computational Linguistics.

Longyin Zhang, Bowei Zou, and Ai Ti Aw. 2024b. Empowering tree-structured entailment reasoning: Rhetorical perception and LLM-driven interpretability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5783–5793, Torino, Italia. ELRA and ICCL.

Yunxiao Zhao, Zhiqiang Wang, Xiaoli Li, Jiye Liang, and Ru Li. 2024. AGR: Reinforced causal agent-guided self-explaining rationalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 510–518, Bangkok, Thailand. Association for Computational Linguistics.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. 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)*, pages 3558–3571, Online. Association for Computational Linguistics.

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 (Online). International Committee on Computational Linguistics.

A Design of Different Prompt Templates

The details of the three prompt templates are provided in Equations 9, 10, and 11. Among them, $</t_i></t_{cause}>$ are learnable tokens, aimed at enhancing the dynamic adaptability of the model during the training process.

$$\begin{aligned} ECI(x_{in}) = \text{The event}&</t_1>e_s</t_2> \\ &</t_5>\text{[MASK]}</t_6> \\ &\text{the event}</t_3>e_t</t_4> \end{aligned} \tag{9}$$

$$\begin{aligned} E-ECI(x) = \text{The event}&</t_1>e_s</t_2> \\ &</t_{cause}></t_3>e_t</t_4> \\ &\text{by the singnal word of} \\ &</t_{13}>\text{[MASK]}</t_{14}> \end{aligned} \tag{10}$$

$$\begin{aligned} I-ECI(x) = \text{In the Sentence}&[CLS], \\ &\text{event}</t_1>e_s</t_2></t_{causs}> \\ &event</t_9>\text{[MASK]}</t_{10}> \end{aligned} \tag{11}$$

B Datasets

ESC is an event storyline dataset that supports instance-level causality identification tasks. It contains 22 topics and 258 documents. Statistical analysis shows that the dataset includes 5,334 event mentions, which can be aggregated into 3,678 event concepts, forming 54,326 event mention pairs and 34,491 event concept pairs, of which 5,625 event mention pairs and 1,814 event concept pairs have causal relationships. The dataset covers both intra-sentence and cross-sentence causal relationships,

with most of them being implicit causality. The CTB dataset contains 184 documents with 9631 event mention pairs, which can be aggregated into 6,813 event concepts, forming 7,608 event concept pairs, including 318 causal labels.