Document-level Causal Relation Extraction with Knowledge-guided Binary Question Answering

Zimu Wang, Lei Xia, Wei Wang, Xinya Du
University of Texas at Dallas
{zimu.wang, lei.xia, xinya.du}@utdallas.edu

Abstract

As an essential task in information extraction (IE), Event-Event Causal Relation Extraction (ECRE) aims to identify and classify the causal relationships between event mentions in natural language texts. However, existing research on ECRE has highlighted two critical challenges, including the lack of document-level modeling and causal hallucinations. In this paper, we propose a **Know**ledge-guided binary **Q**uestion **A**nswering (**KnowQA**) method with event structures for ECRE, consisting of two stages: *Event Structure Construction* and *Binary Question Answering*. We conduct extensive experiments under both zero-shot and fine-tuning settings with large language models (LLMs) on the MECI and MAVEN-ERE datasets. Experimental results demonstrate the usefulness of event structures on document-level ECRE and the effectiveness of KnowQA by achieving state-of-the-art on the MECI dataset. We observe not only the effectiveness but also the high generalizability and low inconsistency of our method, particularly when with complete event structures after fine-tuning the models[1].

1 Introduction

Event-Event Causal Relation Extraction (ECRE) is an essential task in information extraction (IE) that aims to identify and classify the causal relationships between event mentions in natural language texts. For example, given a sentence and an event mention pair of interest (***established***, ***bearing***), an ECRE model should recognize the causal relationship between them, i.e., ***established*** $\xrightarrow{cause}$ ***bearing*** (Figure 1). ECRE is regarded as a precondition of various downstream tasks, such as event knowledge graph construction (Ma et al., 2022), future event prediction (Lin et al., 2022), machine reading comprehension (Zhu et al., 2023), and natural language logical/temporal reasoning (Yang et al., 2020, 2023, 2024).

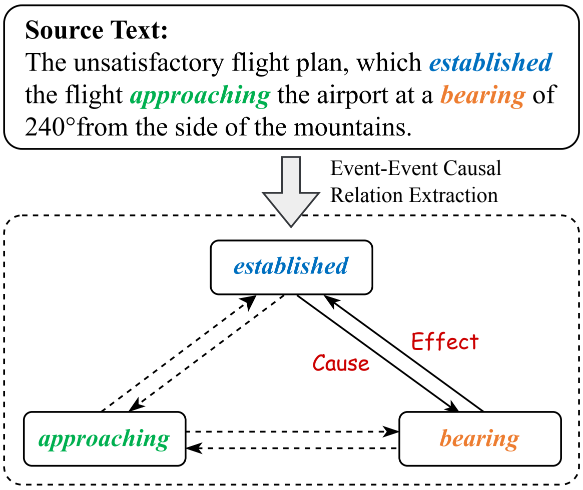
Source Text:
The unsatisfactory flight plan, which established the flight approaching the airport at a bearing of 240°from the side of the mountains.
Event-Event Causal Relation Extraction
established
Cause
Effect
approaching
bearing

Figure 1: Overview of the ECRE process. The dashed lines indicate there are no causal relationships between the event mentions.

Early studies of ECRE mainly focus on identifying the existence of causal relationships, regarding it as a binary classification task and ignoring the directions of these relationships. Researchers have utilized pre-trained language models (PLMs) to model the contexts and improved performance by enriching the event representations and modeling event associations (Tran Phu and Nguyen, 2021; Hu et al., 2023a). Some research has also investigated the potency of large language models (LLMs) on this task (Gao et al., 2023). Recently, the availability of large-scale datasets makes it possible to classify the causal relationships between event mentions (Lai et al., 2022; Wang et al., 2022), which is much more challenging since it necessitates fully understanding the contexts and determining the cause and effect of each event pair while likewise taking additional factors like language varieties into account. In general, existing research on ECRE has highlighted the following two critical challenges:

[1] The source code for this paper is publicly released at https://github.com/du-nlp-lab/KnowQA.

Findings of the Association for Computational Linguistics: EACL 2024, pages 16944–16955
November 12-16, 2024 ©2024 Association for Computational Linguistics

(1) **Lack of Document-level Modeling.** Existing ECRE models typically leverage semantic structures, particularly the Abstract Meaning Representation (AMR) graph (Banarescu et al., 2013), to model event-related contextual information, where the nodes of the graphs represent events, entities, etc., and the edges denote the semantic relationships between them. However, as AMR graphs are built at the sentence level, they are limited in capturing document-level semantics, restricting their capacity to identify implicit and fine-grained information in texts (Tran Phu and Nguyen, 2021; Hu et al., 2023a). Consequently, some approaches that apply AMR graphs for document modeling have not been evaluated for their performance in inter-sentence ECRE (Hu et al., 2023a).

(2) **Causal Hallucinations.** LLMs, such as ChatGPT and GPT-4, have been shown to fall short on the ECRE task and suffer from serious causal hallucination issues by overestimating the existence of causal relationships, largely attributed to reporting biases in natural language, where causal relationships are often described, while the events involved in these relationships are not expressed explicitly (Gao et al., 2023). This phenomenon contributes to low precision and high recall of LLMs on this task, which severely hampers their performance in this field (Gao et al., 2023; Liu et al., 2024). Such challenges are critical to designing more reliable ECRE models, particularly at the document level with appropriate document-level semantic features.

In this paper, we propose a **Know**ledge-guided binary **Q**uestion **A**nswering (**KnowQA**) method to deal with the aforementioned challenges. Unlike previous work that relys heavily on semantic structures, we leverage cross-task knowledge to construct document-level event structures to enrich event information, motivated by the effectiveness of cross-task knowledge in IE (Lin et al., 2020; Jin et al., 2023). We define the ECRE task as consisting of the two subtasks: (1) **Event Causality Identification (ECI)**, which identifies the existence of causal relationships, and (2) **Causal Relation Classification (CRC)**, which classifies the event pairs containing causal relationships into their corresponding relation types. As shown in Figure 2, the overall framework of KnowQA consists of two stages: *Event Structure Construction* and *Binary Question Answering*. In the first stage, we extend the event structures as event mentions, event arguments, and the single-hop relationships of arguments, and we utilize IE models to construct them at the document level. Then, we formulate ECRE as a binary question answering (QA) task with single-turn (for identification) and multi-turn (for identification and classification) strategies with specific relation types, and we incorporate the constructed event structures into the questions.

We conduct comprehensive experiments under zero-shot and fine-tuning settings with LLMs. Experimental results on the MECI (Lai et al., 2022) and MAVEN-ERE (Wang et al., 2022) datasets demonstrate the effectiveness of KnowQA by outperforming the baseline models and achieving the state-of-the-art on the MECI dataset. We also discuss the following benefits of our approach brings: (1) the effectiveness of complete event structures in the ECRE task, particularly at the document level; (2) the efficacy of multi-turn QA under the zero-shot setting and single-turn QA for the identification and multi-turn QA for the classification of causal relationships under the fine-tuning setting; and (3) the high generalizability and low inconsistency of our multi-turn QA strategy, particularly when with complete event structures after fine-tuning the models.

The key contributions of this work are summarized as follows:

- We propose KnowQA, formulating ECRE as a binary QA task with single-turn and multi-turn strategies. To the best of our knowledge, we are the first to utilize QA strategies for ECRE with specific relation types.

- We extend the event structures as event mentions, event arguments, and the single-hop relationships of arguments and validate their effectiveness in the ECRE task.

- We demonstrate the effectiveness of KnowQA, particularly at the document level, and we discuss the high generalizability and low inconsistency of our method.

2 Related Work

Event-Event Causal Relation Extraction. The field of ECRE has been increasingly recognized for its diverse applications; however, research on ECRE has mainly focused on the ECI task that ignores the directions of causal relationships. Early studies on ECI have concentrated on utilizing syntactic patterns (Riaz and Girju, 2013; Gao et al., 2019), statistical event occurrences (Do et al., 2011; Hu and Walker, 2017), and weakly supervised

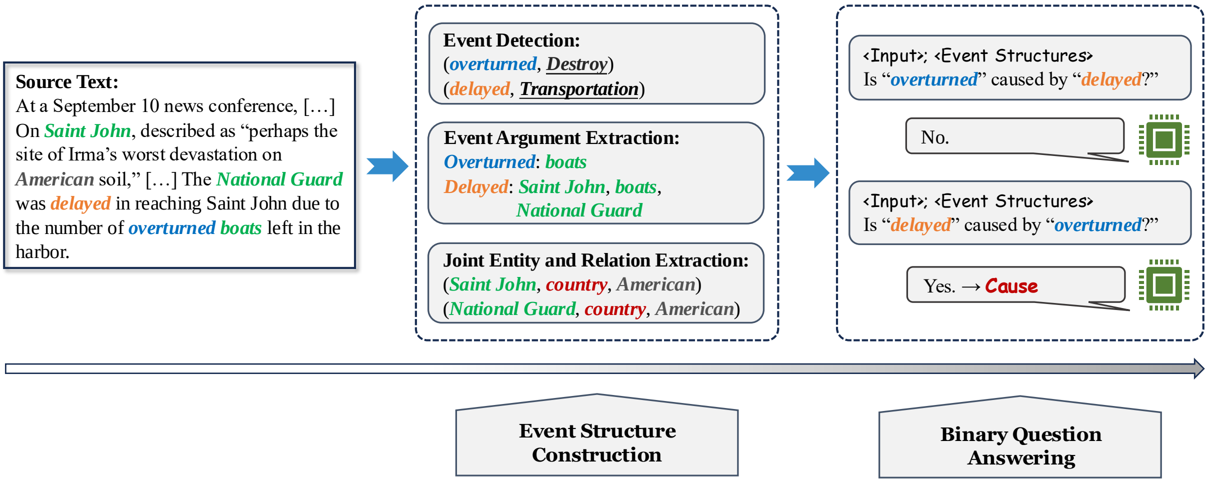
Source Text:
At a September 10 news conference, […] On Saint John, described as "perhaps the site of Irma's worst devastation on American soil," […] The National Guard was delayed in reaching Saint John due to the number of overturned boats left in the harbor.
Event Detection:
(overturned, Destroy)
(delayed, Transportation)
Event Argument Extraction:
Overturned: boats
Delayed: Saint John, boats, National Guard
Joint Entity and Relation Extraction:
(Saint John, country, American)
(National Guard, country, American)
<Input>; <Event Structures>
Is "overturned" caused by "delayed?"
No.
<Input>; <Event Structures>
Is "delayed" caused by "overturned?"
Yes. → Cause
Event Structure Construction
Binary Question Answering

Figure 2: Overall framework of the proposed KnowQA method for ECRE, consisting of two stages: *Event Structure Construction* and *Binary Question Answering*. In the first stage, we utilize IE models to form event structures at the document level. In the second stage, we formulate ECRE as a binary QA task with single-turn and multi-turn strategies and fully leverage the event structures for ECRE predictions.

data (Hashimoto, 2019). Additionally, recent advancements have leveraged PLMs and introduce semantic structures (Tran Phu and Nguyen, 2021; Hu et al., 2023a), external knowledge (Liu et al., 2020b; Cao et al., 2021), and data augmentation (Zuo et al., 2020, 2021) approaches and investigated the potency of ECI with LLMs (Gao et al., 2023). Recently, with the availability of large-scale datasets, some research also focuses on the classification of causal relationships to their corresponding relation types (Deng et al., 2023; Hu et al., 2023b).

However, previous research has struggled to model document-level event-related contextual information comprehensively, as the semantic structures are typically confined to the sentence level. In this paper, we construct document-level event structures and enrich them with event mentions for ECRE predictions.

QA-based IE. Motivated by the effectiveness of natural languages in offering external supervision and mitigating the gap between contexts and tasks, QA-based methods have been extensively studied in the field of IE. Du and Cardie (2020); Liu et al. (2020a) utilize heuristic-based methods for generating questions. Du and Ji (2022); Choudhary and Du (2024) leverage end-to-end deep learning-based methods (Du et al., 2017; Du and Cardie, 2018) for generating QA pairs for representing events. In addition, QA-based methods have also been investigated in temporal relation extraction (Cohen and Bar, 2023) and ECI (Gao et al., 2023), or retrieving useful background knowledge to improve event causality recognition (Kruengkrai et al., 2017; Kadowaki et al., 2019).

However, previous research has not thoroughly leveraged the relation information for supervision considering the misaligned schema between human and LLMs (Peng et al., 2023a) and only discussed the in-context learning (ICL) approach. Unlike the previous work, we formulate ECRE as a binary QA task with event structures to fully utilize the schema, context, and event-associated information. We also fine-tune LLMs to conduct more comprehensive analysis.

3 Methodology

Following the overall framework of KnowQA illustrated in Figure 2, in this section, we introduce each part of the framework, i.e., the *Event Structure Construction* module and the *Binary Question Answering* module, in detail.

3.1 Problem Definition

Following the previous research in EE (Du and Cardie, 2020; Deng et al., 2021), we define our Event-Event Causal Relation Extraction (ECRE) task as two subtasks: **Event Causality Identification (ECI)**, which identifies the existence of causal relationships between event mentions, and **Causal Relation Classification (CRC)**, which classifies the event pairs containing causal relationships into their corresponding relation types. Formally, given a document $D = \{w_1, w_2, \ldots, w_N\}$ that contains multiple sentences (N is the number of words in

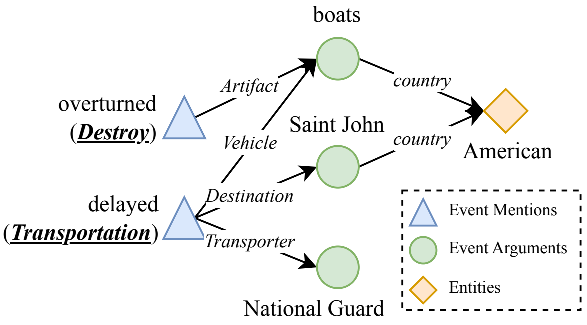

Figure 3: Event structures for the example shown in Figure 2, consisting of event mentions, event arguments, and the single-hop relationships of the arguments.

the document), we use $E = \{e_1, e_2, \ldots\}$ to denote the set of event mentions, $A = \{a_1, a_2, \ldots\}$ for event arguments, and $R = \{r_1, r_2, \ldots\}$ for the single-hop relationships of the arguments. An event mention e_i, an event argument a_j, and a relationship r_k are connected if a_j is an event argument of e_i and r_k includes a_j as either a head or tail entity, and an event mention with its associated entities form an event structure. The event arguments and their single-hop relations are obtained from relevant IE models (Peng et al., 2023b; Li et al., 2021; Eberts and Ulges, 2021).

Our ECRE task aims to predict both the existence of causal relationships and their corresponding relation types. Given the document D and two event mentions of interest e_h and e_t, ECI predicts whether there is a causal relationship between e_h and e_t, and CRC predicts the specific relationship for the pair (e_h, e_t), such as `Cause`, `Effect`, and `Precondition`.

3.2 Event Structure Construction Module

In the first stage of KnowQA, we extract the event arguments and their single-hop relationships to form document-level event structures. We extend the definition proposed by Automatic Content Extraction (ACE), consisting of event mentions and event arguments (Frisoni et al., 2021), with the single-hop relationships of arguments to enrich their information in contexts, as examples shown in Figure 3. An intuitive option is to utilize LLMs for this procedure; however, since they have been found to be insufficient for IE (Li et al., 2023; Peng et al., 2023a), we adopt PLM-based approaches to construct event structures, consisting of three steps: *Event Detection*, *Event Argument Extraction*, and *Joint Entity and Relation Extraction*.

Event Detection. We first conduct event detection to classify the event mentions into pre-defined schema. To ensure the richness of classification, we adopt the KAIROS[2] ontology, a superset of ACE 2005 (Walker et al., 2006) that consists of 50 event types and 59 argument roles[3], to classify the event mentions. We train an event detection model using the CLEVE (Wang et al., 2021) PLM on the WikiEvents dataset (Li et al., 2021) to classify the event mentions to their most likely belonged event type in the KAIROS ontology.

Event Argument Extraction. Afterwards, we follow previous IE toolkits (Wen et al., 2021; Du et al., 2022) to extract the event arguments with BART-Gen (Li et al., 2021), a generative model for document-level event argument extraction (EAE). It formulates EAE as a conditional generation, consisting of the original document and a series of blank event templates with respect to the arugment roles for each event type in the KAIROS ontology, e.g., "*<arg1> damaged <arg2> using <arg3> instrument in <arg4> place*". We adopt the templates defined by Li et al. (2021) to extract arguments for the event mentions.

Joint Entity and Relation Extraction. Finally, we extract the single-hop relationships of the event arguments. To obtain richer relationships, we utilize a joint entity and relation extraction model named JEREX (Eberts and Ulges, 2021) to complete this process, which is a model pre-trained on the DocRED dataset (Yao et al., 2019) with 6 named entity types and 96 relation types. After the entities and their relationships are extracted, we match the entities with the event arguments. We make revisions to the event arguments and entities with higher spans once and select the corresponding head or tail entities with the largest spans.

3.3 Binary Question Answering Module

Following the construction of event structures, we formulate ECRE as a binary QA task with single-turn and multi-turn QA strategies with the constructed event structures. Specifically, single-turn QA is proposed for identifying causal relationships, and multi-turn QA is for identifying and classifying the relationships, adding specific relation types in the questions.

[2] https://www.ldc.upenn.edu/collaborations/current-projects

[3] https://github.com/raspberryice/gen-arg/blob/main/event_role_KAIROS.json

Single-turn QA. In the single-turn QA strategy, we make use of the prompt proposed by previous work (Man et al., 2022; Gao et al., 2023) and incorporate the event structures of the two event mentions into the prompt. It starts with the word "*Input:*", followed by the original text and the event arguments and their relationships obtained from the IE models. Finally, we add a question designed to predict the existence of causal relationships, accompanied with the word "*Answer:*" that ask the model to answer the binary question:

> *Input: {source_text}*
>
> *Arguments of {head_evt}: {head_args}*
>
> *Arguments of {tail_evt}: {tail_args}*
>
> *Argument relationships: {relations}*
>
> *Question: Is there a causal relationship between "{head_evt}" and "{tail_evt}"?*
>
> *Answer:*

In the prompt illustrated above, "*{head_args}*" and "*{tail_args}*" are the arguments of the event mentions, which are listed in sequential following the extraction results. Generally, it appears like (m is the number of arguments of the event mention):

> *<Argument 1>, ..., <Argument m>*

Following Li and Du (2023), we list the argument relationships, denoted as "*{relations}*", with (subject, relation, object) triples. For example (n is the number of triples of the arguments):

> *(<Head 1>, <Relation 1>, <Tail 1>),*
> *...,*
> *(<Head n>, <Relation n>, <Tail n>).*

Multi-turn QA. Considering that LLMs have misaligned schema understanding in IE tasks (Peng et al., 2023a), we construct multi-turn QA prompts based on the specific relation types and regard them as additional supervision for ECRE, for example:

> *Input: {source_text}*
>
> *Arguments of {head_evt}: {head_args}*
>
> *Arguments of {tail_evt}: {tail_args}*
>
> *Argument relationships: {relations}*
>
> *Question: Is "{head_evt}" {relation_type} "{tail_evt}"?*
>
> *Answer:*

	MECI	**MAVEN-ERE**
#Document	438	4,480
#Sentence	2,190	49,873
#Avg. Token/Doc.	146	385
#Event	8,732	112,276
#Evt. Relation	4,100	57,992
#Argument	11,593	290,613
#Arg. Relation	1,751	–

Table 1: Characteristics of the MECI and MAVEN-ERE datasets.

In the prompt illustrated above, "*{relation_type}*" can be "*caused by*" or "*preconditioned by*", following the relation types (`Cause/Effect` and `Precondition`) annotated in ECRE datasets. We iterate the relation types and prompt LLMs in both directions to obtain the causal relationship between each event mention pair.

Intuitively, this QA strategy contains two potential problems: (1) the expression of causal relationship usually varies (e.g., "*cause*" and "*caused by*"), so it is necessary to test on multiple causal expressions to ensure the generalizability of the proposed method; and (2) because we ask the model for a specific event mention pair for multiple times, it may potentially answer positively to all questions, yet the QA process terminates as long as the model receives a positive answer. As a result, we conduct additional analysis in Section 4.6 on the impact of causal expressions and the order of the questions to model performance.

4 Experiments

4.1 Datasets

We conducted experiments on two commonly used document-level ECRE datasets: MECI (Lai et al., 2022) and MAVEN-ERE (Wang et al., 2022). MECI contains the `Cause` and `Effect` relationships, and MAVEN-ERE contains the `Cause` and `Precondition` relationships. During our experiments, we selected the English subset from MECI, and we followed Gao et al. (2023) and Chen et al. (2024) to sample a subset from MAVEN-ERE. The characteristics of the datasets are shown in Table 1, in which the number of event arguments and their relationships were derived from the IE models for the MECI dataset, and we utilized the golden argument annotation from MAVEN-ARG (Wang et al., 2024) for the MAVEN-ERE dataset.

Model	MECI						MAVEN-ERE					
	ECI			CRC			ECI			CRC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
	GPT-3.5											
Single-turn	24.5	**92.3**	38.8	–	–	–	25.8	67.7	37.4	–	–	–
w/ Args.	27.3	80.9	40.8	–	–	–	27.2	**73.8**	**39.8**	–	–	–
w/ Rels.	28.8	72.0	41.2	–	–	–	–	–	–	–	–	–
Multi-turn	32.8	76.3	**45.9**	20.9	**48.7**	29.3	27.1	55.0	36.3	**15.0**	18.6	16.6
w/ Args.	33.2	64.2	43.8	24.1	46.8	31.9	**27.7**	64.4	38.8	13.9	**23.9**	**17.6**
w/ Rels.	**34.9**	59.1	43.9	**25.5**	43.3	**32.1**	–	–	–	–	–	–
	Flan-T5$_{XL}$											
Single-turn	30.2	83.5	44.4	–	–	–	26.0	57.9	35.9	–	–	–
w/ Args.	30.4	**83.9**	44.6	–	–	–	26.4	**66.7**	37.8	–	–	–
w/ Rels.	31.7	83.2	45.9	–	–	–	–	–	–	–	–	–
Multi-turn	**40.3**	64.8	49.7	**35.4**	56.9	**43.6**	28.7	52.2	37.0	17.4	38.4	24.0
w/ Args.	39.0	66.9	49.3	34.0	58.3	42.9	**29.2**	62.1	**39.7**	**18.4**	**45.4**	**26.2**
w/ Rels.	38.8	71.8	**50.4**	33.5	**62.0**	43.5	–	–	–	–	–	–

Table 2: Performance of KnowQA against baselines on the MECI and MAVEN-ERE datasets under the *zero-shot* setting. The best and second-best results for each model are highlighted in **bold** and underlined, respectively.

4.2 Baselines

We compared the performance of KnowQA against the following state-of-the-art baselines from existing ECRE research: (1) **PLM** (Tran Phu and Nguyen, 2021) classifies causal relationships after obtaining event representations; (2) **Know** (Liu et al., 2020b) retrieves related concepts and relations for event mentions from ConceptNet to augment input texts; (3) **RichGCN** (Tran Phu and Nguyen, 2021) enriches the event representations by constructing interaction graphs between essential objects; (4) **ERGO** (Chen et al., 2022) builds event relational graphs to convert ECRE into a node classification problem; (5) **DiffusECI** (Man et al., 2024a) develops a diffusion model to generate causal label representations to eliminate irrelevant components; (6) **HOTECI** (Man et al., 2024b) leverages optimal transport to select the most important words and sentences from full documents; (7) **GIMC** (He et al., 2024) constructs a heterogeneous graph interaction network to model long-distance dependencies between events. We followed the original implementations of the baselines using the XLM-RoBERTa PLM (Conneau et al., 2020).

4.3 Experimental Setup

We conducted experiments under both zero-shot and fine-tuning settings with two LLMs: GPT-3.5 and Flan-T5 (Chung et al., 2024). Specifically, we experimented with GPT-3.5 (`gpt-3.5-turbo-0125`) under the zero-shot setting from its official API[4], and we set the temperature as 0 to stabalize the outputs. We also conducted experiments using Flan-T5$_{XL}$ under zero-shot and Flan-T5$_{Large}$ under fine-tuning settings. During the fine-tuning process, we set the batch size as 4, the gradient accummulation steps as 4, the learning rate as $5e-5$, and the number of epochs as 5, and selected the best validation models to test performance on the test set. The main experiments were conducted on a single GeForce RTX 3090 graphic card. Detailed experimental settings for the event structure construction process are organized in Appendix A.

4.4 Main Results

The main experimental results of KnowQA against baselines under zero-shot and fine-tuning settings are presented in Tables 2 and 3, respectively. From the tables, we have the following observations:

Firstly, under the zero-shot setting, both GPT-3.5 and Flan-T5$_{XL}$ fell short on the ECI task no matter the identification and classification of causal relationships and exhibited high-level causal hallucination issues, which tended to assume the existence of causal relationships between event mentions. It can be observed by the unsatisfied overall perfor-

[4] `https://platform.openai.com/`

Model	ECI			CRC		
	P	**R**	**F1**	**P**	**R**	**F1**
PLM[◇]	56.4	77.8	65.4	45.9	59.7	51.9
Know[◇]	42.4	75.7	54.3	34.3	47.2	39.7
RichGCN[◇]	63.5	**79.2**	70.5	52.5	63.6	57.5
DiffusECI	70.1	68.3	69.2	–	–	–
HOTECI	66.6	67.1	66.8	–	–	–
ERGO	–	–	–	55.0	57.5	56.2
GIMC	–	–	–	61.5	58.4	59.9
	Flan-T5$_{Large}$					
Single-turn	67.3	75.7	71.2	–	–	–
w/ Args.	69.1	75.4	72.1	–	–	–
w/ Rels.	70.2	78.1	**73.9**	–	–	–
Multi-turn	65.5	71.8	68.5	59.6	65.3	62.3
w/ Args.	**71.2**	75.0	73.0	**64.3**	67.7	**66.0**
w/ Rels.	70.8	76.3	73.5	63.2	**68.1**	65.6

Table 3: Performance of KnowQA against baselines on the MECI dataset under the *fine-tuning* setting. The best and second-best results are highlighted in **bold** and underlined, respectively. [◇] denotes that the results are from our reproduction.

Model	ECI		CRC	
	Intra F1	**Inter F1**	**Intra F1**	**Inter F1**
Single-turn	77.8	12.5	–	–
w/ Args.	76.0	**36.0***	–	–
w/ Rels.	**78.7**	34.3	–	–
Multi-turn	74.2	27.4	67.5	25.3
w/ Args.	76.8	31.5	**69.4**	28.3
w/ Rels.	**77.0**	**40.0***	68.9	**34.7***

Table 4: Intra-sentence and inter-sentence performance of ECRE on the MECI dataset with the Flan-T5$_{Large}$ model. * denotes statistical significance ($p < 0.01$).

mance and the low precision and high recall illustrated in the tables. Flan-T5$_{XL}$ performed much better than GPT-3.5, whose performance was close to the fine-tuned Know baseline; it also achieved better performance on the classification of causal relationships, indicating its better reasoning ability compared with GPT-3.5.

Secondly, after fine-tuning Flan-T5$_{Large}$, it outperformed all baselines and achieved state-of-the-art on both the identification and classification of causal relationships, even though its parameter is smaller than Flan-T5$_{XL}$. Besides, the event structures were notably helpful in making better ECRE predictions, and the performance under both zero-shot and fine-tuning settings was correlated with the completeness of event structures, i.e., without event structures $<$ with event arguments $<$ with event arguments and their relationships. Compared with MECI, the event structures were more helpful for MAVEN-ERE, whose model performance, after incorporating them, consistently outperformed the performance without event structures. Nevertheless, the experimental results illustrated in the tables were enough to indicate the high transferability between different IE tasks and the effectiveness of rich information for generative models when conducting reasoning tasks.

Finally, multi-turn QA was more effective in both identifying and classifying causal relationships than single-turn QA under the zero-shot setting and was able to alleviate the causal hallucination issues revealed in LLMs by observing the decreased recall and increased precision values; however, after fine-tuning the LLMs, the multi-turn QA strategy was adept for the classification, while the single-turn QA strategy was effective for the identification of causal relationships. By comparing differences between precision and recall values, it is evident that the differences between them significantly decreased after fine-tuning the models, indicating that the causal hallucination issues reported by LLMs could be solved after fine-tuning them. In this case, multi-turn QA detected more event mention pairs that do not contain causal relationships, but it would also omit more pairs that contain relationships compared with the single-turn QA strategy. The false negative predictions of the questions had a high likelihood of accumulating and decreasing the prediction results when identifying the existence of causal relationships.

4.5 Effectiveness of Event Structures

From the experimental results illustrated in Tables 2 and 3, it is evident that the event structures are helpful for LLMs to make better ECRE predictions by comparing the results with respect to different completeness of event structures. We analyzed the intra-sentence and inter-sentence performance of the models and conducted a case study to further understand the efficacy of event structures.

Intra- and Inter-sentence Performance. Table 4 presents the intra-sentence and inter-sentence performance after fine-tuning the Flan-T5$_{Large}$ model. From the table, we observed that the original Flan-T5$_{Large}$ was not proficient in inter-sentence ECRE (e.g., an F1-score of 12.5 in ECI under the single-turn QA strategy) and had a large gap with its intra-sentence performance. However, after in-

Input: Once that ***happened***, the INS mode would change from "armed" to "capture" and the ***plane*** would ***track*** the flight-planned course from then on. The HEADING mode of the ***autopilot*** would normally be ***engaged*** sometime after take off to comply with vectors from ***ATC***, and then after ***receiving*** appropriate ATC clearance, to guide the plane to intercept the desired INS course line.
Event Mention Pair: (***happened***, ***track***) **Argument(s) of *happened*:** (None) **Argument(s) of *track*:** ***plane*** **Prediction without Event Structure:** `None` **Prediction with Event Structure:** `Cause`
Event Mention Pair: (***receiving***, ***engaged***) **Argument(s) of *receiving*:** ***ATC*** **Argument(s) of *engaged*:** ***autopilot*** **Prediction without Event Structure:** `Cause` **Prediction with Event Structure:** `Effect`

Table 5: Examples of the event mention pairs that can be correctly classified with event structures (***Blue***: event mentions of the first example, ***Orange***: event mentions of the second example, ***Green***: event arguments extracted by IE models).

corporating the event structures, while the improvement of the intra-sentence performance was minor, the inter-sentence performance increased by a significant margin, and its gap with intra-sentence performance also decreased significantly. This indicates the effectiveness of document-level event structure in document-level ECRE predictions.

Case Study. We sampled 50 cases that could not derive correct prediction without event structures but were able to predict correctly with them. We categorized them into two cases: *identifying implicit causal relationships* and *correcting mispredictions*. Table 5 illustrates examples concerning the two cases.

In the first example, the original model could not identify the causal relationship because it was not explicitly expressed, i.e., there was not explicit causal clues, such as "*cause*" and "*because*", between the event mentions; however, with the event argument "plane" of the event mention "track", the model was able to detect the "`Cause`" relationship between the event mentions. In the second example, the original model classified the relationship as "`Cause`" because there was a confusable temporal clue between the event mentions, i.e., "*after*"; however, the temporal clue was for the events "take off" but not "receiving". With the event structures, the model could predict correct relationships between them, i.e., engaged $\xrightarrow{cause}$ receiving, by potentially leveraging the relationships between the two event arguments, e.g., the engagement of autopilot may affect ATC (air traffic control) in some cases.

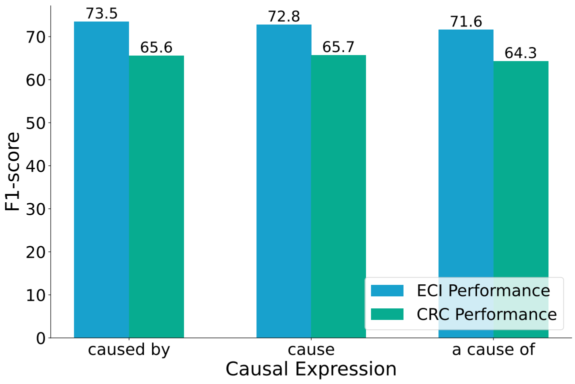

Figure 4: Performance of Flan-T5$_{Large}$ with complete event structures using different causal expressions on the MECI dataset.

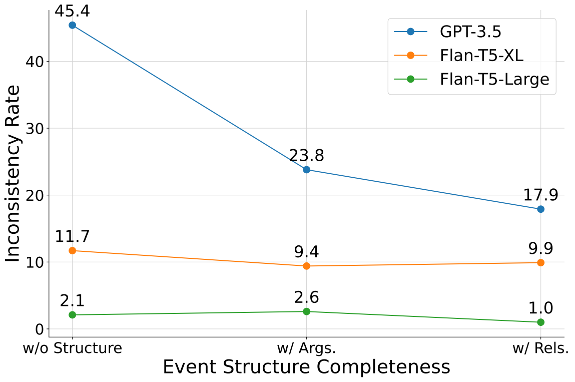

Figure 5: Inconsistency evaluation results of the models on the MECI dataset.

4.6 Additional Analysis

Effects of Causal Expressions. In the main experiments, we formulated our questions with passive voice, i.e., "*caused by*" and "*preconditioned by*". Because the expression of causal relationship usually varies, for a pair of event mentions (e_h, e_t), we conducted additional experiments with two different causal expressions, including (1) "*Does* e_h *cause* e_t*?*" and (2) "*Is* e_h *a cause of* e_t*?*" As shown in Figure 4, our method achieved similar performance when using different causal expressions, which all outperformed the baseline models, indicating our high generalizability; however, the performance using the expression "*Is* e_t *caused by* e_h*?*" achieved the best result, indicating that LLMs are more favored in the passive voice when understanding causal relationships.

Inconsistency Evaluation. To guarantee that the order of the questions does not significantly alter model predictions under the multi-turn QA strategy, we defined an additional evaluation metric called "inconsistency" to evaluate the models. Specifically, we inferred the models with both binary questions with reversed causal directions (e.g., `Cause` and `Effect`) and obtained model predictions. We formulated the "inconsistency" metric as the proportion of the number of event pairs that the model predicts both questions positive to the number of pairs that the model predicts at least one positive, formally:

$$\text{Inconsistency} = \frac{\text{\# of Both Positive}}{\text{\# of At Least One Positive}}. \quad (1)$$

Figure 5 illustrates the results of the inconsistency of the experimented models. It can be observed that the models under the zero-shot setting, particularly GPT-3.5, exhibited a high inconsistency problem. This well explains the reason why the performance difference of GPT-3.5's comprehensible result in the identification of causal relationships but unsatisfactory in classification compared with Flan-T5$_{XL}$: the model had a high likelihood of generating a positive response in the first turn and predicting the relationship as `Effect`. Under the fine-tuning setting, Flan-T5$_{Large}$ had a minimal inconsistency. Notably, the incorporation of event structures could alleviate the inconsistency problem on all models, and the inconsistency rate of Flan-T5$_{Large}$ with complete event structures was only 1.0, indicating that the order of the questions hardly affects the model prediction and the high reliability of our proposed method.

5 Conclusion and Future Work

We introduced KnowQA, a novel method for ECRE that formulates the task as a binary QA task with the utilization of cross-task knowledge in IE, i.e., document-level event structures, consisting of two stages: *Event Structure Construction* and *Binary Question Answering*. Experimental results on the MECI and MAVEN-ERE datasets demonstrated that the effectiveness, high generalizability, and low inconsistency of KnowQA, particularly with complete event structures after fine-tuning the models. In the future, we will test our method with more event relations (e.g., temporal and subevent relations) and more languages to further validate the generalizability of our proposed method.

Limitations

The limitations of KnowQA in the current work are as follows: (1) The event structures were constructed with relevant IE models on ECRE datasets, which were not golden labels. Although they have been proven to be helpful for the ECRE task, the errors from the event structure construction process may have negative impacts on ECRE predictions. (2) We only tested KnowQA with limited LLMs and only on English corpus. Future investigations of KnowQA with other commonly used LLMs (e.g., Llama and Mistral) and languages (e.g., Danish and Spanish) can be conducted in the future to validate the generalizability of KnowQA in more scenarios.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of LAW*, pages 178–186.

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 ACL-IJCNLP*, pages 4862–4872.

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 COLING*, pages 2118–2128.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024. Improving large language models in event relation logical prediction. In *Proceedings of ACL*, pages 9451–9478.

Milind Choudhary and Xinya Du. 2024. QAEVENT: Event extraction as question-answer pairs generation. In *Findings of EACL*, pages 1860–1873.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Omer Cohen and Kfir Bar. 2023. Temporal relation classification using Boolean question answering. In *Findings of ACL*, pages 1843–1852.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Shumin Deng, Shengyu Mao, Ningyu Zhang, and Bryan Hooi. 2023. SPEECH: Structured prediction with energy-based event-centric hyperspheres. In *Proceedings of ACL*, pages 351–363.

Shumin Deng, Ningyu Zhang, Luoqiu Li, Chen Hui, Tou Huaixiao, Mosha Chen, Fei Huang, and Huajun Chen. 2021. OntoED: Low-resource event detection with ontology embedding. In *Proceedings of ACL-IJCNLP*, pages 2828–2839.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of EMNLP*, pages 294–303.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of ACL*, pages 1907–1917.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of EMNLP*, pages 671–683.

Xinya Du and Heng Ji. 2022. Retrieval-augmented generative question answering for event argument extraction. In *Proceedings of EMNLP*, pages 4649–4666.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of ACL*, pages 1342–1352.

Xinya Du, Zixuan Zhang, Sha Li, Pengfei Yu, Hongwei Wang, Tuan Lai, Xudong Lin, Ziqi Wang, Iris Liu, Ben Zhou, Haoyang Wen, Manling Li, Darryl Hannan, Jie Lei, Hyounghun Kim, Rotem Dror, Haoyu Wang, Michael Regan, Qi Zeng, Qing Lyu, Charles Yu, Carl Edwards, Xiaomeng Jin, Yizhu Jiao, Ghazaleh Kazeminejad, Zhenhailong Wang, Chris Callison-Burch, Mohit Bansal, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, Martha Palmer, and Heng Ji. 2022. RESIN-11: Schema-guided event prediction for 11 newsworthy scenarios. In *Proceedings of NAACL-HLT (Demo)*, pages 54–63.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of EACL*, pages 3650–3660.

Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2021. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of EMNLP*, pages 11111–11126.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *Proceedings of NAACL-HLT*, pages 1808–1817.

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from Wikipedia. In *Proceedings of EMNLP-IJCNLP*, pages 2988–2999.

Zhitao He, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, Zhiqiang Zhang, Mengshu Sun, and Jun Zhao. 2024. Zero-shot cross-lingual document-level event causality identification with heterogeneous graph contrastive transfer learning. In *Proceedings of LREC-COLING*, pages 17833–17850.

Zhichao Hu and Marilyn Walker. 2017. Inferring narrative causality between event pairs in films. In *Proceedings of SIGDIAL*, pages 342–351.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023a. Semantic structure enhanced event causality identification. In *Proceedings of ACL*, pages 10901–10913.

Zhilei Hu, Zixuan Li, Daozhu Xu, Long Bai, Cheng Jin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2023b. Protoem: A prototype-enhanced matching framework for event relation extraction. *Preprint*, arXiv:2309.12892.

Xiaomeng Jin, Haoyang Wen, Xinya Du, and Heng Ji. 2023. Toward consistent and informative event-event temporal relation extraction. In *Proceedings of MATCHING*, pages 23–32.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of EMNLP-IJCNLP*, pages 5816–5822.

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. *Proceedings of AAAI*, 31(1).

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022. MECI: A multilingual dataset for event causality identification. In *Proceedings of COLING*, pages 2346–2356.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *Preprint*, arXiv:2304.11633.

Ruosen Li and Xinya Du. 2023. Leveraging structured information for explainable multi-hop question answering and reasoning. In *Findings of EMNLP*, pages 6779–6789.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of NAACL-HLT*, pages 894–908.

Li Lin, Yixin Cao, Lifu Huang, Shu'Ang Li, Xuming Hu, Lijie Wen, and Jianmin Wang. 2022. What makes the story forward? inferring commonsense explanations as prompts for future event generation. In *Proceedings of SIGIR*, page 1098–1109.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of ACL*, pages 7999–8009.

Cheng Liu, Wei Xiang, and Bang Wang. 2024. Identifying while learning for document event causality identification. In *Proceedings of ACL*, pages 3815–3827.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020a. Event extraction as machine reading comprehension. In *Proceedings of EMNLP*, pages 1641–1651.

Jian Liu, Yubo Chen, and Jun Zhao. 2020b. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of IJCAI*, pages 3608–3614.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, and Jing Shao. 2022. MMEKG: Multi-modal event knowledge graph towards universal representation across modalities. In *Proceedings of ACL (Demo)*, pages 231–239.

Hieu Man, Franck Dernoncourt, and Thien Huu Nguyen. 2024a. Mastering context-to-label representation transformation for event causality identification with diffusion models. *Proceedings of AAAI*, 38(17):18760–18768.

Hieu Man, Chien Van Nguyen, Nghia Trung Ngo, Linh Ngo, Franck Dernoncourt, and Thien Huu Nguyen. 2024b. Hierarchical selection of important context for generative event causality identification with optimal transports. In *Proceedings of LREC-COLING*, pages 8122–8132.

Hieu Man, Minh Nguyen, and Thien Nguyen. 2022. Event causality identification via generation of important context words. In *Proceedings of *SEM*, pages 323–330.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023a. When does in-context learning fall short and why? a study on specification-heavy tasks. *Preprint*, arXiv:2311.08993.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023b. OmniEvent: A comprehensive, fair, and easy-to-use toolkit for event understanding. In *Proceedings of EMNLP (Demo)*, pages 508–517.

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 *Proceedings of SIGDIAL*, pages 21–30.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of NAACL-HLT*, pages 3480–3490.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium*, 57.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of EMNLP*, pages 926–941.

Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, Jie Zhou, and Juanzi Li. 2024. MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation. In *Proceedings of ACL*, pages 4072–4091.

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 ACL-IJCNLP*, pages 6283–6297.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of NAACL-HLT (Demo)*, pages 133–143.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2024. Language models as inductive reasoners. In *Proceedings of the EACL*, pages 209–225.

Zonglin Yang, Xinya Du, Erik Cambria, and Claire Cardie. 2023. End-to-end case-based reasoning for commonsense knowledge base completion. In *Proceedings of EACL*, pages 3509–3522.

Zonglin Yang, Xinya Du, Alexander Rush, and Claire Cardie. 2020. Improving event duration prediction via time-aware pre-training. In *Findings of EMNLP*, pages 3370–3378.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of ACL*, pages 764–777.

Jiazheng Zhu, Shaojuan Wu, Xiaowang Zhang, Yuexian Hou, and Zhiyong Feng. 2023. Causal intervention for mitigating name bias in machine reading comprehension. In *Findings of ACL*, pages 12837–12852.

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 ACL-IJCNLP*, pages 3558–3571.

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 COLING*, pages 1544–1550.

A Experimental Setup for Event Structure Construction

Event Detection We trained an event detection using the OmniEvent toolkit (Peng et al., 2023b) on the WikiEvents dataset (Li et al., 2021), and we selected CLEVE (Wang et al., 2021) as the PLM. During the training process, we followed the original experimental settings proposed in OmniEvent[5]. We set the batch size as 40, the learning rate as $1e-5$, and the number of epochs as 30, and we selected the best validation model on the WikiEvents dataset to conduct event detection on our datasets.

Event Argument Extraction & Joint Entity and Relation Extraction We adopted the pre-trained models released by BART-Gen[6] (Li et al., 2021) and JEREX[7] (Eberts and Ulges, 2021) to conduct EAE and joint entity and relation extraction on our datasets, separately.

[5] https://github.com/THU-KEG/OmniEvent/blob/main/config/all-models/ed/tc/roberta-large/cleve.yaml

[6] https://github.com/raspberryice/gen-arg

[7] https://github.com/lavis-nlp/jerex