Open Event Causality Extraction by the Assistance of LLM in Task Annotation, Dataset, and Method

Kun Luo[1,2], Tong Zhou[1,2], Yubo Chen[1,2], Jun Zhao[1,2] and Kang Liu[1,2,3*]
[1]The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Shanghai Artificial Intelligence Laboratory
{luokun2024, tong.zhou}@ia.ac.cn
{yubo.chen, jzhao, kliu}@nlpr.ia.ac.cn

Abstract

Event Causality Extraction (ECE) aims to extract explicit causal relations between event pairs from the text. However, the event boundary deviation and the causal event pair mismatching are two crucial challenges that remain unaddressed. To address the above issues, we propose a paradigm to utilize LLM to optimize the task definition, evolve the datasets, and strengthen our proposed customized **C**ontextual **H**ighlighting **E**vent **C**ausality **E**xtraction framework (CHECE). Specifically in CHECE, we propose an Event Highlighter and an Event Concretization Module, guiding the model to represent the event by a higher-level cluster and consider its causal counterpart in event boundary prediction to deal with event boundary deviation. And we propose a Contextual Event Causality Matching mechanism, meanwhile, applying LLM to diversify the content templates to force the model to learn causality from context to targeting on causal event pair mismatching. Experimental results on two ECE datasets demonstrate the effectiveness of our method.

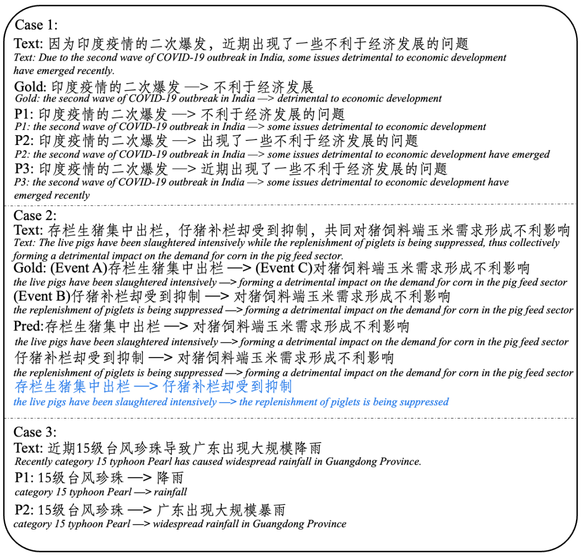
Keywords: Event Extraction, Large Language Model, Knowledge Graph

1. Introduction

Event Causality Extraction (ECE) aims to extract causal relations between event pairs, in which each event is presented as a continuous span within the sentences or documents. Abundant downstream application tasks can be facilitated after extracting event causality from text, including event detection (Weng and Lee, 2011), event prediction (Granroth-Wilding and Clark, 2016) Xu et al. (2020), logical reasoning (Tappin et al., 2020), question answering (Karpukhin et al., 2020), and constructing an event logic graph (Ding et al., 2019) Gao et al. (2022).

Given plain text, an ECE system is responsible for extracting event spans and matching them by causality. Previous works (Yang et al., 2022) (Lyu et al., 2022) (Zhang et al., 2022) Yang et al. (2022) Heindorf et al. (2020) in event causality extraction predominantly employ a two-stage method: event tagging and span-based event causality matching. Much progress (Yang et al., 2022) has been made in this paradigm with the development of pre-trained language models (Devlin et al., 2018). However, two challenges have not caught much attention: Event Boundary Deviation and Event Causality Mismatching.

Event Boundary Deviation: Previous methods struggle to predict causal event boundaries, resulting in redundant or missing words. As shown in Fig 1 Case 1, a typical ECE model makes different predictions P1 to P3 in spans of the effect, but all predictions describe the same event as labeled in GOLD. We explore the origin of the event boundary deviation phenomenon from two perspectives: concluding practical experimental experience and digging deep into the principle of the ECE task.



Case 1:
Text: 因为印度疫情的二次爆发，近期出现了一些不利于经济发展的问题
Text: Due to the second wave of COVID-19 outbreak in India, some issues detrimental to economic development have emerged recently.
Gold: 印度疫情的二次爆发 —> 不利于经济发展
Gold: the second wave of COVID-19 outbreak in India —> detrimental to economic development
P1: 印度疫情的二次爆发 —> 不利于经济发展的问题
P1: the second wave of COVID-19 outbreak in India —> some issues detrimental to economic development
P2: 印度疫情的二次爆发 —> 出现了一些不利于经济发展的问题
P2: the second wave of COVID-19 outbreak in India —> some issues detrimental to economic development have emerged
P3: 印度疫情的二次爆发 —> 近期出现了一些不利于经济发展的问题
P3: the second wave of COVID-19 outbreak in India —> some issues detrimental to economic development have emerged recently

Case 2:
Text: 存栏生猪集中出栏，仔猪补栏却受到抑制，共同对猪饲料端玉米需求形成不利影响
Text: The live pigs have been slaughtered intensively while the replenishment of piglets is being suppressed, thus collectively forming a detrimental impact on the demand for corn in the pig feed sector.
Gold: (Event A)存栏生猪集中出栏 —> (Event C)对猪饲料端玉米需求形成不利影响
the live pigs have been slaughtered intensively —> forming a detrimental impact on the demand for corn in the pig feed sector
(Event B)仔猪补栏却受到抑制 —> 对猪饲料端玉米需求形成不利影响
the replenishment of piglets is being suppressed —> forming a detrimental impact on the demand for corn in the pig feed sector
Pred:存栏生猪集中出栏 —> 对猪饲料端玉米需求形成不利影响
the live pigs have been slaughtered intensively —> forming a detrimental impact on the demand for corn in the pig feed sector
仔猪补栏却受到抑制 —> 对猪饲料端玉米需求形成不利影响
the replenishment of piglets is being suppressed —> forming a detrimental impact on the demand for corn in the pig feed sector
存栏生猪集中出栏 —> 仔猪补栏却受到抑制
the live pigs have been slaughtered intensively —> the replenishment of piglets is being suppressed

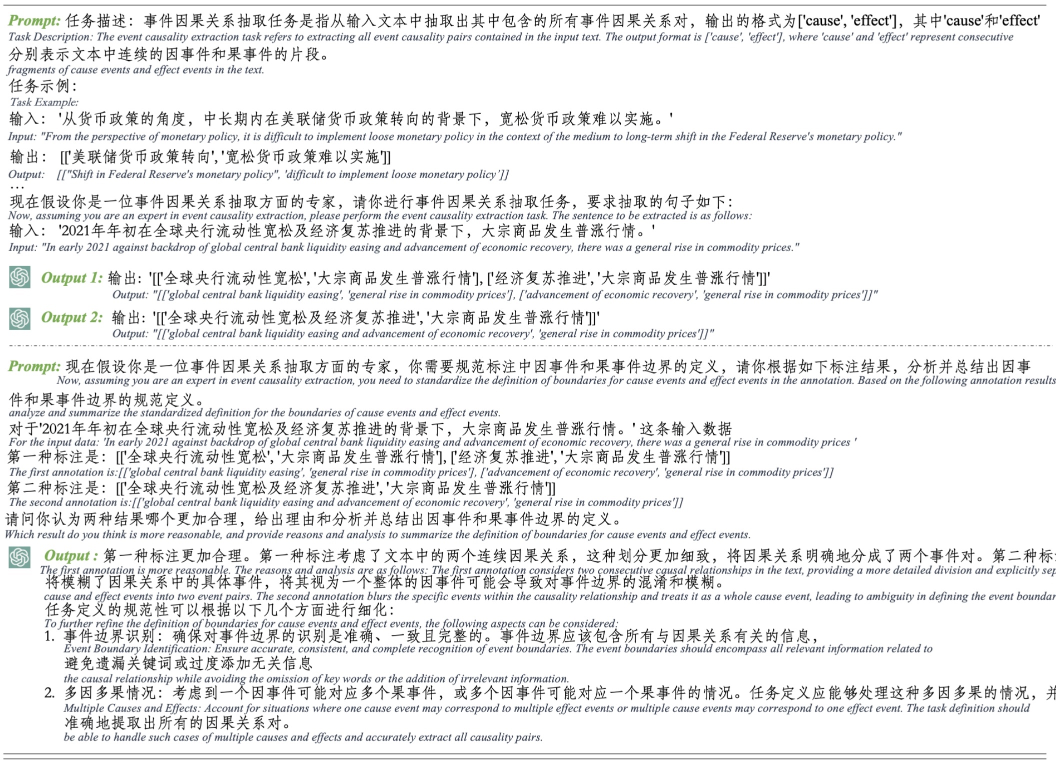
Case 3:
Text: 近期15级台风珍珠导致广东出现大规模降雨
Recently category 15 typhoon Pearl has caused widespread rainfall in Guangdong Province.
P1: 15级台风珍珠 —> 降雨
category 15 typhoon Pearl —> rainfall
P2: 15级台风珍珠 —> 广东出现大规模暴雨
category 15 typhoon Pearl —> widespread rainfall in Guangdong Province

Figure 1: Case study in ECE.

In the process of the case study in preliminary experiments, we observe frequent inconsistent annotations of ECE datasets. As shown in Table

[*] Corresponding author

Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ LREC-Coling 2024, pages 33–44
21 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0



Prompt: 任务描述：事件因果关系抽取任务是指从输入文本中抽取出其中包含的所有事件因果关系对，输出的格式为['cause', 'effect']，其中'cause'和'effect'
Task Description: The event causality extraction task refers to extracting all event causality pairs contained in the input text. The output format is ['cause', 'effect'], where 'cause' and 'effect' represent consecutive
分别表示文本中连续的因事件和果事件的片段。
fragments of cause events and effect events in the text.
任务示例：
Task Example:
输入：'从货币政策的角度，中长期内在美联储货币政策转向的背景下，宽松货币政策难以实施。'
Input: "From the perspective of monetary policy, it is difficult to implement loose monetary policy in the context of the medium to long-term shift in the Federal Reserve's monetary policy."
输出： [['美联储货币政策转向','宽松货币政策难以实施']]
Output: [["Shift in Federal Reserve's monetary policy", 'difficult to implement loose monetary policy']]
…
现在假设你是一位事件因果关系抽取方面的专家，请你进行事件因果关系抽取任务，要求抽取的句子如下：
Now, assuming you are an expert in event causality extraction, please perform the event causality extraction task. The sentence to be extracted is as follows:
输入：'2021年年初在全球央行流动性宽松及经济复苏推进的背景下，大宗商品发生普涨行情。'
Input: "In early 2021 against backdrop of global central bank liquidity easing and advancement of economic recovery, there was a general rise in commodity prices."

Output 1: 输出：'[['全球央行流动性宽松','大宗商品发生普涨行情'], ['经济复苏推进','大宗商品发生普涨行情']]'
Output: "[['global central bank liquidity easing', 'general rise in commodity prices'], ['advancement of economic recovery', 'general rise in commodity prices']]"

Output 2: 输出：'[['全球央行流动性宽松及经济复苏推进','大宗商品发生普涨行情']]'
Output: "[['global central bank liquidity easing and advancement of economic recovery', 'general rise in commodity prices']]"

Prompt: 现在假设你是一位事件因果关系抽取方面的专家，你需要规范标注中因事件和果事件边界的定义，请你根据如下标注结果，分析并总结出因事件和果事件边界的规范定义。
Now, assuming you are an expert in event causality extraction, you need to standardize the definition of boundaries for cause events and effect events in the annotation. Based on the following annotation results, analyze and summarize the standardized definition for the boundaries of cause events and effect events.
对于'2021年年初在全球央行流动性宽松及经济复苏推进的背景下，大宗商品发生普涨行情。'这条输入数据
For the input data: 'In early 2021 against backdrop of global central bank liquidity easing and advancement of economic recovery, there was a general rise in commodity prices '
第一种标注是：[['全球央行流动性宽松','大宗商品发生普涨行情'], ['经济复苏推进','大宗商品发生普涨行情']]
The first annotation is:[['global central bank liquidity easing', 'general rise in commodity prices'], ['advancement of economic recovery', 'general rise in commodity prices']]
第二种标注是：[['全球央行流动性宽松及经济复苏推进','大宗商品发生普涨行情']]
The second annotation is:[['global central bank liquidity easing and advancement of economic recovery', 'general rise in commodity prices']]
请问你认为两种结果哪个更加合理，给出理由和分析并总结出因事件和果事件边界的定义。
Which result do you think is more reasonable, and provide reasons and analysis to summarize the definition of boundaries for cause events and effect events.

Output : 第一种标注更加合理。第一种标注考虑了文本中的两个连续因果关系，这种划分更加细致，将因果关系明确地分成了两个事件对。第二种标注将模糊了因果关系中的具体事件，将其视为一个整体的因事件可能会导致对事件边界的混淆和模糊。
The first annotation is more reasonable. The reasons and analysis are as follows: The first annotation considers two consecutive causal relationships in the text, providing a more detailed division and explicitly separates cause and effect events into two event pairs. The second annotation blurs the specific events within the causality relationship and treats it as a whole cause event, leading to ambiguity in defining the event boundaries.
任务定义的规范性可以根据以下几个方面进行细化：
To further refine the definition of boundaries for cause events and effect events, the following aspects can be considered:
1. 事件边界识别：确保对事件边界的识别是准确、一致且完整的。事件边界应该包含所有与因果关系有关的信息，避免遗漏关键词或过度添加无关信息
   Event Boundary Identification: Ensure accurate, consistent, and complete recognition of event boundaries. The event boundaries should encompass all relevant information related to the causal relationship while avoiding the omission of key words or the addition of irrelevant information.
2. 多因多果情况：考虑到一个因事件可能对应多个果事件，或多个因事件可能对应一个果事件的情况。任务定义应能够处理这种多因多果的情况，并准确地提取出所有的因果关系对。
   Multiple Causes and Effects: Account for situations where one cause event may correspond to multiple effect events or multiple cause events may correspond to one effect event. The task definition should be able to handle such cases of multiple causes and effects and accurately extract all causality pairs.

Figure 2: Prompts and responses for task definition generation.

1, there exists a large proportion of labeling inconsistencies in both typical Chinese and English ECE datasets. These inconsistencies confused the model in event boundary predictions trained in these datasets. However, we argue that labeling mistakes are ineluctable. As the model prediction case shown in Fig 1 Case 1, a causal event expressed in span form with consecutive words exists in multiple reasonable variants. In addition, previous research overlooked the explicit definition and annotation guidelines in event causality extraction, hindering the restoration process in these datasets. To this end, clarifying the ECE task definition and fixing inconsistencies in datasets are the primary goals.

On the other hand, we dig into the reasonable span variants of the event. First, each event composed of a continuous span within the input text exhibits multiple literal forms that depict the event with different emphases. Therefore, modeling the event with a specific span fails to capture its overall perspective. However, previous works employ a particular span on behalf of the event, having an inherent shortage of capturing the entirety. Furthermore, building an association between cause and effect events when predicting event span boundaries is essential. As illustrated in Fig 1 Case 3, for the first prediction, "category 15 Typhoon Pearl" cause "rain" constitutes a reasonable but rough causal event pair when independently predicting event span boundaries. But when considering the interdependence between the causal event pair, for the second prediction given the level and name of the typhoon in the cause event, the effect event should include specific rainfall locations and intensity. However, previous works restricted to predicting causal event span boundaries independently, lacking in the consideration of causal associations.

Event Causality Mismatching: After the extraction of potential causal events, the next step is matching cause and effect events with semantics and knowledge. Previous methods will usually face an inevitable challenge, which is mismatching two events by event span. As illustrated in Fig 1 Case 2, a human always estimates causality between event pairs from two perspectives: semantic information and contextual information. From the semantic perspective, based on their common sense and linguistics knowledge, humans can evaluate event causality based on span. However, the final decision cannot be divorced from contextual information aside from causal events, such as conjunctions, background, and correlations. Unfortunately, previous studies focused on modeling the semantic information inside event pairs, neglecting the crucial role of contextual information outside. This flaw in design could lead to confusion for models when tackling complex causal event pair-matching cases. Fig 1 Case 2 illustrates an example where the model incorrectly predicts a causal relationship between events A and B due to their perceived semantic similarity. However, leveraging contextual information, we can determine that there is no causal relationship between events A and B, but rather that events A and B cause event C, simultaneously.

In this paper, we utilize LLM to optimize the task definition, evolve the datasets, and strengthen our proposed customized event causality extraction framework to address the above issues. We introduce a pattern that applies LLM to conclude a task definition and annotation criteria according to the case of labeling inconsistency. And then automatically fix datasets by LLM based on their viewpoint. Apart from the foundation of the task, we construct a Contextual Highlighting Event Causality Extraction framework (CHECE). Specifically, we propose an Event Highlighter to represent an event independent of a specific span, and an Event Concretization Module to predict a single event boundary based on its causal counterpart. Together deal with the event boundary deviation from these three aspects. To deal with the event causality mismatching problem, we propose a Contextual Event Causality Matching mechanism. And to further ensure the model learns from context correlation, we utilize LLM to diversify the context templates.

The contributions of this paper are as follows:

1) We propose a paradigm to utilize LLM to clarify the event causality extraction task annotation and fix existing datasets. And we release the metrics and datasets to promote the relevant research.
2) To handle the event boundary deviation, we propose an Event Highlighter and an Event Concretization Module, guiding the model to represent the event by a higher-level cluster and consider its causal counterpart in event boundary prediction. To tackle the event causality mismatching, we devise a Contextual Event Causality Matching mechanism and apply LLM to diversify the content templates to force the model to learn causality from context.
3) Experiments on both Chinese and English event causality extraction datasets show our method outperforms state-of-the-art methods, especially in our new metrics.

2. The Annotation Clarification and Dataset of Event Causality Extraction

Due to the frequent inconsistent annotations of ECE datasets and their inevitability, which leads to confusion in the final model, we explore clarifying and aligning the dataset annotation with the assistance of LLM, which is well-aligned with the given annotating requirements and requires much less labor compared with human annotators.

Dataset	Manual Label	After Fix
CFC	85%	92%
FinCR	87%	93%

Table 1: Statistics of labeling accuracy before and after dataset evolution.

2.1. Annotation Clarification by LLM

We clarify the annotation criteria with the assistance of the Large Language model(LLM). Taking into account the presence of inconsistent annotated data, we employ the LLM (specifically, text-divinci-003) to generate multiple predictions, preserving each distinct output. Thanks to the rich knowledge that LLM contains and its great ability to follow given instructions, the LLM is able to analyze which of the various outputs it predicts makes the most sense, thereby establishing the essential attribute that should define the event boundary judgment. As shown in Fig 2, the prompts are organized in the format of the chain of thought (Wang et al., 2022b). Through the above process on several sets of inconsistent annotated data, the annotation criteria of event boundary can be concluded, which can be used to create instructions for the LLM to perform dataset repairment following these standards subsequently.

2.2. Measurement

Event Boundary Deviation arises due to the ambiguous task definition and inconsistent dataset annotations, as well as the inherent multivariate nature of events. We propose Easy F1 to measure the model performance more reasonably. In Easy F1, a predicted causal event span is considered correct if its similarity with the gold span surpasses a predefined threshold. The choice of this threshold can be adapted to the data distribution, and we set it at 80 percent. In the English dataset the similarity is measured by tokens, whereas in the Chinese dataset, the similarity is measured by word segmentations.

2.3. Fix Dataset by LLM

We set the concluded task's definitions into prompts and ask the LLM to repair the dataset as required. In the example of concluded event boundary definition *"causal event should be fine-grained, which means the output span cannot contain more than one event and should include all the words describing the same event"*. We set the obtained definition to the requirements, and give three shots

Dataset	Train	Dev	Test	Pairs	Text Length	Causal Distance	Span Length
CFC	2000	250	250	2.17	41	3.4	9.67
FineCR	12541	1583	1557	1.14	69	8.3	13.38

Table 2: Statistics of two ECE datasets.

of manual repair of data according to the requirements. Then we ask the LLM to determine whether the event boundary in the data meets the definition according to the requirements, if not, it needs to be corrected and explain the reason. The prompt examples are shown in the Appendix.

3. Method

In this section, we first formally define the event causality extraction task and then elaborate on each component of our model. The overall architecture of our ECE framework is shown in Fig 3.

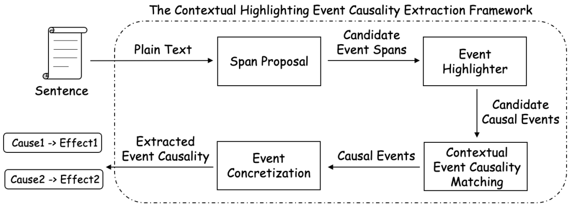


Figure 3: The overall framework of CHECE.

3.1. Problem Definition

The input sentence is $X = \{x_1, x_2, ..., x_n\}$ with n tokens. Let $S = \{s_1, s_2, ..., s_n\}$ be all the possible spans in X. The desired outputs are causal event pairs as $T(X) = \{(c, e)|c, e \in S\}$, where c and e are the cause event and effect event presented as continuous spans in the input text.

The problem is decomposed into two parts, first identifying the candidate cause events and effect events and then assessing causality within event pairs formed by combining all candidate cause events with candidate effect events.

3.2. Span Proposal

Given the input sentence X, to obtain the representation of each token, we use a pre-trained language model (PLM) as our sentence encoder. The output is

$$\{h_1, h_2, \ldots, h_n \mid h_i \in \mathbb{R}^{d\times 1}\} \tag{1}$$

where d is the embedding dimension, and n is the number of tokens.

Then we judge each s_i in S whether it is a causal event span following the previous span-based method (Su et al., 2022), which uses a global scoring matrix that considers the beginning and the end positions of spans to predict all the candidate cause(effect) spans. It's worth noting that the casual event spans predicted by the Span Proposal Model are not exact events, they may be part of the event lacking some boundary components or they may include the event. In other words, these spans reflect different emphases of the event.

With the obtained sentence representation, using two feedforward layers that rely on the begin and end indices of the span:

$$q_i = W_q h_i + b_q \tag{2}$$

$$k_j = W_k h_j + b_k \tag{3}$$

where $q_i \in \mathbb{R}^d, k_j \in \mathbb{R}^d$ denote the vector representations of the start and end positions. The score $p_{i,j}$ indicating the score of span $s[i : j]$ that starts with i being a cause(effect) span is computed as follows:

$$p_{i,j} = \sigma(q_i^\top k_j) \tag{4}$$

where σ is the sigmoid function. Then we set a threshold μ for the predicted score. We consider the span $s[i : j]$ as a candidate cause(effect) span if $p_{i,j}$ exceeds the threshold value.
Class Imbalance Loss is introduced to the training process $\mathcal{L}_s$:

$$\log(1 + \sum_{(q,k)\in P} e^{-p_{q,k}}) + \log(1 + \sum_{(q,k)\in Q} e^{p_{q,k}}) \tag{5}$$

where q, k represent the start and tail indexes of a span, P represents a collection of spans that are considered candidate cause(effect) spans, Q represents a collection of spans that are not candidate cause(effect) spans.

3.3. Event Highlighter

The Event Highlighter aims to build better representations for events. Due to the inherent uncertainty and multivariate nature of events expressed with natural language, employing a single specific span to model the target event yields multiple candidates with varying boundaries for a given event, thereby significantly occupying the search space and inducing model confusion during the matching of event causality. To this end, we propose a cluster-based event highlighter model to catch the overall perspective and significance of events, exploring to model the event at event-level instead of span-level.

After obtaining all the candidate cause spans and effect spans $s[i : j]$ together with their scores $p_{i,j}$, the event highlighter captures the complete picture and emphasis of the target event E. Specifically, there may be multiple candidate spans that have a slight boundary deviation from the target span (e.g., a different adjective), all describing the same target event E but with different emphases. We use the clustering method to gather the spans that describe the same event. First, we evaluate the token similarity between each span and the target event centered on the target event E. If the similarity exceeds the threshold λ, the span is considered to describe the target event and is clustered with the target event. Each target event corresponds to an event cluster $C = \{s_1, s_2, ..., s_k\}$. Where λ is an adjustable hyperparameter and k is the number of all spans describing the same target event obtained after span clustering.

The representation of a single span is acquired as follows.

$$\boldsymbol{h}_{s_i} = Avgpool(\boldsymbol{h}_{START(i):END(i)}) \tag{6}$$

where $Avgpool$ is the average pooling operation(Lin et al., 2013), $START(i)$ and $END(i)$ denote the start and end indices of the candidate span s_i.

Then the Event Highlighter combines all the candidate spans describing the same target event, weighted by their score in the span proposal model to find the most important tokens of the event and see the full event description covering the longest boundary:

$$\boldsymbol{h_E} = \sum_{1}^{k} p_k \boldsymbol{h}_{s_k} \tag{7}$$

where k is the number of all spans describing the same target event obtained after span clustering.

3.4. Contextual Event Causality Matching

The Event Causality matching model aims to take a pair of cause event E_c and effect event E_e as input and predict whether there is a causal relationship. Previous works concatenate the representations of event pairs and put them into the feedforward layer for causality judgment, which only considers the semantic representation. We argue that the explicit use of contextual information plays an important role in causal judgment. As illustrated in Fig 1, only relying on the semantic representation of events, it is easy to mistakenly judge that there is a causal relationship between events A and B due to their semantic similarity. But in fact, combining the context structure information, we can judge that there is no causal relationship between A and B, and it is A and B that cause C together.

To this end, we propose utilizing the semantic information and context information jointly to evaluate the causality of input event pairs. First, the semantic representations of input event pairs are obtained as:

$$\psi_{sem}(E_c, E_e) = W_{sem}[\boldsymbol{h}_{E_c}; \boldsymbol{h}_{E_e}] + b_{sem} \tag{8}$$

where $\boldsymbol{h}_{E_c}$ and $\boldsymbol{h}_{E_e}$ are the event representation obtained in the Event Highlighter model, W_{sem} and b_{sem} are trainable parameters, and $[A; B]$ denotes the concatenation operation.

Next, we turn to obtain the explicit contextual representation. The mask token is used to replace the position of the event's original tokens and sent to the BERT encoder to let the model pay attention to the context information other than the semantics of the specific token. Then the output mask token is used as the contextual representation, rich in context structure information:

$$\psi_{con}(E_c, E_e) = W_{con}[\boldsymbol{m}_{E_c}; \boldsymbol{m}_{E_e}] + b_{con} \tag{9}$$

where $\boldsymbol{m}_{E_c}, \boldsymbol{m}_{E_e}$ are the contextual representation obtained from the mask token, W_{con} and b_{con} are trainable parameters.

With semantic representation and contextual representation, we model the judgment of event causality jointly by combing the two parts of information using a hyperparameter θ.

$$\psi(E_c, E_e) = \theta\psi_{sem}(E_c, E_e) + (1 - \theta)\psi_{con}(E_c, E_e) \tag{10}$$

where $\psi(E_c, E_e)$ is the score for cause event E_c and effect event E_e to be a pair of causal events. Then we set a threshold υ for the predicted score. We consider cause event E_c and effect event E_e to be a pair of causal events if $p_{i,j}$ exceeds the threshold value.

During the training process, the loss is considered as follows:

$$\mathcal{L}_e = - \sum_{E_i \in E_c, E_j \in E_e} \log P\left(R^*_{i,j} \mid E_i, E_j\right) \tag{11}$$

where $R^*_{i,j}$ represents the gold relation type of event pair.

To better utilize the contextual information, we also augment the training data by constructing template training data. Specifically, we replace each causal event pair of the data in the training set with $[cause]$ and $[effect]$ to form a contextual template that preserves only structural information. We store all the templates in a file. Furthermore, we employ ChatGPT[1] to generate causal event pairs and causal templates in order to harvest rich domain knowledge and diverse causal contextual information from LLM. The detailed prompt is shown in Appendix A.1. With the obtained causal event pairs and causal templates, we synthesize extra data during the training process by replacing the $[cause]$ and $[effect]$ in the chosen template from

[1] https://chat.openai.com/chat

the stored file with a random pair of causal events in the training set or the causal event pairs generated by ChatGPT to enhance the model's ability capacity in capturing contextual information and injecting domain knowledge into the model.

3.5. Event Concretization Module

The Event Concretization Module aims to reify the event pairs judged to have causal relations in the last step from the abstract event representation to the concrete event span. In other words, given a pair of input causal event clusters and their representations, the Event Concretization Module needs to output the most suitable cause span and effect span that best represents the target causal event pair as the final extraction result.

Previous works consider the spans that score higher than the preset threshold in the Span Proposal Module as predicted events. There are two possible disadvantages to this practice: First, multiple spans with slight boundary differences are referred to as the same event, thereby disentangling the information inside the event and inevitably introducing subsequent matching errors. Furthermore, this practice cut off the connections between causal event pairs. As illustrated in Figure 2, judging the boundaries of cause or effect events separately ignores the overall connection of causal events and is error-prone.

Given a pair of input cause and effect event cluster $C_c = \{s_1, s_2, ..., s_m\}$ and $C_e = \{s_1, s_2, ..., s_n\}$ with their representations. First, iterate over each span s_i in the cause event cluster C_c and build its connection with the effect event cluster C_e by concatenating their representation and put into a feedforward network:

$$P_{s_i} = \sigma(W_{concre}[\boldsymbol{h}_{s_i}; \boldsymbol{h}_{E_e}] + b_{concre}) \tag{12}$$

where σ is the sigmoid function, $\boldsymbol{h}_{s_i}$ and $\boldsymbol{h}_{E_e}$ are the span representation of s_i and event representation of E_e. We choose the span with the highest score P_{s_i} in the cause event cluster as the final output cause event. During the training process, the loss is considered as follows:

$$\mathcal{L}_c = -\sum_{s_i \in C_c} \log P\left(r^*_{i,j} \mid s_i\right) \tag{13}$$

where $r^*_{i,j}$ represents the gold type of span s_i which means whether the span s_i is the gold span to represent the cause event. Event Concretization for the effect event cluster is conducted in a symmetric way.

3.6. Training Strategy

We adopt a joint training approach, wherein we optimize the combined objective function throughout the training process while sharing the parameters of the BERT encoder. The total loss is the sum of these three parts:

$$\mathcal{L}_{total} = \omega_1\mathcal{L}_s + \omega_2\mathcal{L}_e + \omega_3\mathcal{L}_c \tag{14}$$

Performance might be better by carefully tuning the weight of each sub-loss, but we just assign equal weights for simplicity (i.e., ω_1 = ω_1 = ω_1 = 1).

Dataset	Method	Dev		Test	
		Easy-F1	Hard-F1	Easy-F1	Hard-F1
CFC	BERT-CRF	54.94	42.81	53.29	38.54
	GlobalPointer	59.49	51.18	61.96	53.84
	TP-Linker	62.81	53.39	62.28	53.97
	PL-Marker	63.89	53.06	63.71	54.65
	ChatGPT	-	-	31.39	12.56
	Ours	64.40	53.86	63.81	55.24
	Ours+LLM	**64.88**	**55.09**	**65.63**	**55.73**
FineCR	BERT-CRF	55.12	35.60	54.92	35.58
	GlobalPointer	55.72	39.89	54.76	38.97
	TP-Linker	56.21	40.05	56.60	39.39
	PL-Marker	57.99	40.14	58.75	39.90
	ChatGPT	-	-	17.68	7.62
	Ours	58.85	40.37	59.77	40.21
	Ours+LLM	**59.91**	**40.47**	**60.61**	**40.55**

Table 3: Comparison of our model and other baselines on two event causality extraction datasets. We test ChatGPT with 3-shot task examples and task descriptions. *"Ours+LLM"* means our full model with ChatGPT data augmentation.

4. Experiments

4.1. Datasets and Preprocessing

We conduct experiments on FineCR and CFC (Yang et al., 2022) proposed in Section 2 to verify the effectiveness of our method. FineCR is a widely used dataset in English. The experiments and analysis on it could be regarded as fair comparisons with previous works. CFC is a more challenging dataset with more ambiguous causal event spans and multiple complicated causalities in a single sentence. The detailed statistical information and split information are shown in Table 2.

4.2. Metrics and Parameter Settings

For automatic evaluation, we utilize easy F1 and hard F1 introduced in Section 2. Since previous methods in full tagging paradigm apply token-wise tag F1 score to report the performance, to fairly compare our performance with baselines, we reproduce these methods and report our metrics.

We use bert-base-uncased (Devlin et al., 2018) and chinese-roberta-wwm-ext (Cui et al., 2021) as the base encoders for the English dataset FineCR and the Chinses dataset CFC. The learning rate is set as 3e-5 in the backbone of BERT. We set the max length of the input sentence to 200/75 for

CFC and FineCR. The batch size is set as 16. We train the model for at most 30 epochs and choose the model with the best performance on the dev set to output results on the test set.

4.3. Baselines

We compare our method with the following baselines:

BERT-CRF(Yang et al., 2022): BERT-CRF is a powerful model that combines BERT's contextual understanding with CRF's sequential tagging for accurate squeue tagging.

GlobalPointer(Su et al., 2022): GlobaoPointer is a span-based method using a global scoring matrix that considers the beginning and the end positions of spans with a global view.

TP-Linker(Wang et al., 2020): TP-Linker is a one-stage joint entity and relation extraction model. We use it to extract event spans that have larger granularity than entities thus bringing great challenges to the model.

PL-Marker(Ye et al., 2021): PL-Marker proposed a novel span representation approach to consider the interrelation between the spans (pairs) by strategically packing the markers in the encoder and achieving SOTA performance in the entity and relation extraction task.

ChatGPT: ChatGPT is a large language model developed by OpenAI which has strong zero-shot and few-shot learning abilities. However, it struggles in the difficult task such as causal relation extraction that requires more comprehensive commonsense knowledge and higher logical reasoning ability. We test the model with a task description and three-shot task examples.

4.4. Compared with State-of-the-art Methods

Table 3 shows the results of our method on two event causality extraction datasets. Overall, our method achieves the best performance from these baselines. Indicating our method's effectiveness and advancement. Specifically, compared among full sequence tagging methods, whatever the tagging schema setting, PLMs help them achieve better performances. However, comparing ChatGPT-Gen with other baselines, we can draw a conclusion that the performance of LLM in this task is inferior to supervised training models. It could be the reason that the complex and specific demands in the ECE task hinder the release of LLM's extensive capacity. Transferring methods in joint extraction of entities and relations to ECE, TP-Linker(Wang et al., 2020) and PL-Marker(Ye et al., 2021) achieve higher f1 than full tagging methods. Prove they can model the span representation and span relation better than plain tagging. Our method obtained better performance in easy and hard f1 than TP-Linker and PL-Marker, which struggle to extract events that have larger granularity than entities. It demonstrated the proposed Event Highlighter and Contextual Causal Event Matching is more customized in this task and could deal with Event Boundary Deviation and Event Causality Mismatching.

Method	CFC		FineCR	
	Easy-F1	Hard-F1	Easy-F1	Hard-F1
Ours	**65.63**	**55.73**	**60.61**	**40.55**
w/o event highlighter	60.11	53.37	52.15	38.12
w/o causal event matching	63.91	53.93	58.61	39.27
w/o LLM template	63.81	55.24	59.75	40.21
with 500 Augment Causality	64.08	53.74	-	-
with 1000 Augment Causality	65.63	55.73	-	-
with 1500 Augment Causality	63.43	55.12	-	-

Table 4: Ablation results on the CFC and FineCR test set.

4.5. Ablation Experiments

To investigate the effectiveness of our proposed components in the method, we also perform ablation experiments on the CFC and FineCR datasets. The ablation results are shown in Table 4, indicating that none of these models can achieve a comparable result with our full version. Demonstrate that all those factors contribute a certain improvement to our model.

Specifically, when we discard the whole event highlighter part, and represent an event with a specific span (Ours w/o event highlighter), the performance drops demonstrate the effectiveness of the event highlighter. In Ours w/o causal event matching, we calculate causal pair score only referring to event representation and ignore the contextual information. The suboptimal performance demonstrates the effectiveness of contextual causal event matching.

We employ ChatGPT to generate causal event pairs and incorporate them into the constructed template to synthesize new training data as introduced in Section 3.4, injecting knowledge and contextual information into the model simultaneously. To further explore the effectiveness of event pair augment from LLM, we attempted varying template count N utilized during training. The bottom of Table 4 shows, when n is zero, the easy f1 drop X from the full model, indicates the effectiveness of the event pair augments from LLM. In addition, the model performance could not improve with the increase of N after N is larger than 1000, manifesting that the templates generated by LLM is varies considerably in quality.

5. Related Work

Event Causality Extraction. Previous works use feature-based methods for event causality extraction. (Ittoo and Bouma, 2011) proposes a method for extracting causal pairs by leveraging part-of-speech analysis, syntactic analysis, and causality templates. (Hashimoto et al., 2014) uses semantic relation (between nouns), context, and association features to extract event causalities from the web. In recent years, deep learning techniques employed in causality extraction. (Li et al., 2021) uses the BiLSTM-CRF model as the backbone to extract cause and effect directly, formulating the task in the causality tagging scheme. (Wang et al., 2022a) proposed a model that aims to transform event causality extraction into causal argument extraction, by incorporating both sentence-level and document-level contextual information. Recently, much progress has been made in this task with the strong language modeling capabilities and rich world knowledge of pre-trained language models (PLMs) (Devlin et al., 2018). (Fajcik et al., 2022) used T5 to identify all cause-effect-signal span triplets. (Yang et al., 2022) and (Lyu et al., 2022) use BERT-CRF model in Fine-grained Event Causality Extraction and FinCausal 2022 tasks, resulting in significant advancements.

LLMs Assist Tasks. The capability of Large Language Models (LLMs) like ChatGPT to comprehend user intent and provide reasonable responses has made them extremely popular lately. Recent studies show that the latest LLMs have the ability to do Information Extraction tasks such as Named Entity Recognization(NER), Relation Extraction(RE), and Event Extraction(EE). (Xu et al., 2023) proposed task-related instructions and schema-constrained data generation to enhance LLM's few-shot relation extraction performance. (Tang et al., 2023) used LLM's rich domain knowledge to induce new event schemas. Some works utilize LLM to improve the performance of downstream tasks. (Dai et al., 2023) and (Ubani et al., 2023) leveraged ChatGPT for text data augmentation and synthetic training data generating to induce extensive knowledge.

6. Conclusion

This paper proposes to utilize LLM to generate the definition of event causality extraction tasks and automatically evolve the datasets. Lay the foundation for further research and improvement. We propose a framework called CHECE to deal with two unaddressed problems. Specifically, the Event Highlighter and an Event Concretization Module, guide the model to represent the event by a higher-level cluster and consider its causal counterpart in event boundary prediction to deal with event boundary deviation. And the Contextual Event Causality Matching mechanism forces the model to predict causality from context information to overcome the causal event pair mismatching issue. Meanwhile, we apply LLM to diversify the content templates to enhance this side. Experimental results on two ECE datasets demonstrate the effectiveness of the method.

7. Ackonwledgements

This work is supported by the National Key Research and Development Program of China (No. 2022ZD0160503), the National Natural Science Foundation of China (No. 62176257). This work is also supported by Yunnan Provincial Major Science and Technology Special Plan Projects (No.202202AD080004).

Limitations

Our work is not without limitations. From the LLM side, on the one hand, the best prompt or the chain of thought for the conclusion of task definition by LLM is under-explored. We believe there exists a better way for LLM to generate the definition and further utilize it to evolve the datasets. On the other hand, this paradigm could produce more labeled data from news or documents from the web. Release a larger dataset remains in our future work. From the framework side, although effective our method is slightly complicated. How to address the above two challenges more concisely is a worth exploring topic.

8. Bibliographical References

Wajid Ali, Wanli Zuo, Wang Ying, Rahman Ali, Gohar Rahman, and Inam Ullah. 2023. Causality extraction: A comprehensive survey and new perspective. *Journal of King Saud University-Computer and Information Sciences*, page 101593.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiao Ding, Zhongyang Li, Ting Liu, and Kuo Liao. 2019. Elg: an event logic graph. *arXiv preprint arXiv:1907.08015*.

Martin Fajcik, Muskaan Singh, Juan Zuluaga-Gomez, Esaú Villatoro-Tello, Sergio Burdisso, Petr Motlicek, and Pavel Smrz. 2022. Idiapers@ causal news corpus 2022: Extracting cause-effect-signal triplets via pre-trained autoregressive language model. *arXiv preprint arXiv:2209.03891*.

Jianqi Gao, Hang Yu, and Shuang Zhang. 2022. Joint event causality extraction using dual-channel enhanced neural network. *Knowledge-Based Systems*, 258:109935.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation. *arXiv preprint arXiv:2305.07375*.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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.

Christopher Hidey and Kathleen McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Ashwin Ittoo and Gosse Bouma. 2011. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011, Alicante, Spain, June 28-30, 2011. Proceedings 16*, pages 52–63. Springer.

Xianxian Jin, Xinzhi Wang, Xiangfeng Luo, Subin Huang, and Shengwei Gu. 2020. Inter-sentence and implicit causality extraction from chinese corpus. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*, pages 739–751. Springer.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Christopher SG Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 336–343.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219.

Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400*.

Chenyang Lyu, Tianbo Ji, Quanwei Sun, and Liting Zhou. 2022. Dcu-lorcan at fincausal 2022: Span-based causality extraction from financial documents using pre-trained language models. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 116–120.

Chaveevan Pechsiri and Rapepun Piriyakul. 2010. Explanation knowledge graph construction through causality extraction from texts. *Journal of computer science and technology*, 25(5):1055–1070.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*.

Jialong Tang, Hongyu Lin, Zhuoqun Li, Yaojie Lu, Xianpei Han, and Le Sun. 2023. Harvesting event schemas from large language models. *arXiv preprint arXiv:2305.07280*.

Ben M Tappin, Gordon Pennycook, and David G Rand. 2020. Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences*, 34:81–87.

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*.

Longbao Wang, Li Chen, Zeyu Zhang, Yingchi Mao, Chong Long, and Yican Shen. 2022a. Event causality extraction based on fusion attention. In *2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC)*, pages 1–5. IEEE.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 401–408.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? *arXiv preprint arXiv:2305.01555*.

Linyi Yang, Zhen Wang, Yuxiang Wu, Jie Yang, and Yue Zhang. 2022. Towards fine-grained causal reasoning and qa. *arXiv preprint arXiv:2204.07408*.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2021. Packed levitated marker for entity and relation extraction. *arXiv preprint arXiv:2109.06067*.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt. *arXiv preprint arXiv:2304.05454*.

Yujie Zhang, Rujiang Bai, Ling Kong, and Xiaoyue Wang. 2022. 2sce-4sl: A 2-stage causality extraction framework for scientific literature.

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian-Yun Nie. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950.

9. Language Resource References

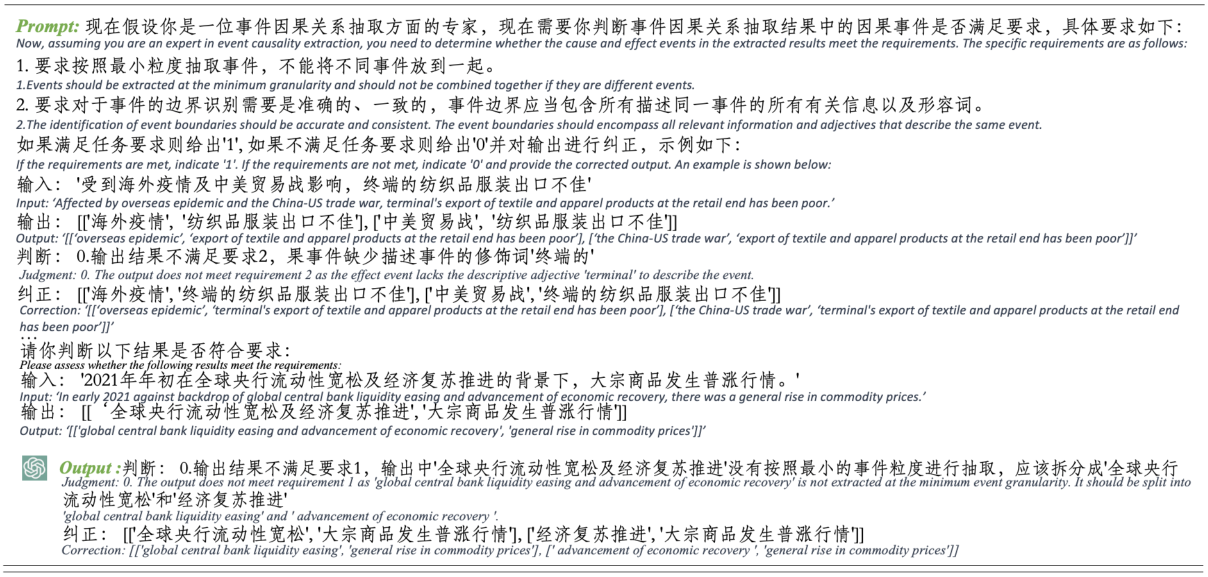
Gao, Jianqi and Luo, Xiangfeng and Wang, Hao. 2022. *Chinese causal event extraction using causality-associated graph neural network*. Wiley Online Library.

Heindorf, Stefan and Scholten, Yan and Wachsmuth, Henning and Ngonga Ngomo, Axel-Cyrille and Potthast, Martin. 2020. *Causenet: Towards a causality graph extracted from the web*.

Xu, Jinghang and Zuo, Wanli and Liang, Shining and Zuo, Xianglin. 2020. *A review of dataset and labeling methods for causality extraction*.

Yang, Linyi and Wang, Zhen and Wu, Yuxiang and Yang, Jie and Zhang, Yue. 2022. *Towards fine-grained causal reasoning and qa*.

A. Appendix



Prompt: 现在假设你是一位事件因果关系抽取方面的专家，现在需要你判断事件因果关系抽取结果中的因果事件是否满足要求，具体要求如下:

Now, assuming you are an expert in event causality extraction, you need to determine whether the cause and effect events in the extracted results meet the requirements. The specific requirements are as follows:

1. 要求按照最小粒度抽取事件，不能将不同事件放到一起。

1.Events should be extracted at the minimum granularity and should not be combined together if they are different events.

2. 要求对于事件的边界识别需要是准确的、一致的，事件边界应当包含所有描述同一事件的所有有关信息以及形容词。

2.The identification of event boundaries should be accurate and consistent. The event boundaries should encompass all relevant information and adjectives that describe the same event.

如果满足任务要求则给出'1', 如果不满足任务要求则给出'0'并对输出进行纠正，示例如下:

If the requirements are met, indicate '1'. If the requirements are not met, indicate '0' and provide the corrected output. An example is shown below:

输入: '受到海外疫情及中美贸易战影响，终端的纺织品服装出口不佳'

Input: ‘Affected by overseas epidemic and the China-US trade war, terminal's export of textile and apparel products at the retail end has been poor.’

输出: [['海外疫情', '纺织品服装出口不佳'], ['中美贸易战', '纺织品服装出口不佳']]

Output: ‘[[‘overseas epidemic’, ‘export of textile and apparel products at the retail end has been poor’], [‘the China-US trade war’, ‘export of textile and apparel products at the retail end has been poor’]]’

判断: 0.输出结果不满足要求2，果事件缺少描述事件的修饰词'终端的'

Judgment: 0. The output does not meet requirement 2 as the effect event lacks the descriptive adjective 'terminal' to describe the event.

纠正: [['海外疫情','终端的纺织品服装出口不佳'], ['中美贸易战','终端的纺织品服装出口不佳']]

Correction: ‘[[‘overseas epidemic’, ‘terminal's export of textile and apparel products at the retail end has been poor’], [‘the China-US trade war’, ‘terminal's export of textile and apparel products at the retail end has been poor’]]’

...

请你判断以下结果是否符合要求:

Please assess whether the following results meet the requirements:

输入: '2021年年初在全球央行流动性宽松及经济复苏推进的背景下，大宗商品发生普涨行情。'

Input: ‘In early 2021 against backdrop of global central bank liquidity easing and advancement of economic recovery, there was a general rise in commodity prices.’

输出: [[‘全球央行流动性宽松及经济复苏推进','大宗商品发生普涨行情']]

Output: ‘[['global central bank liquidity easing and advancement of economic recovery', 'general rise in commodity prices']]’

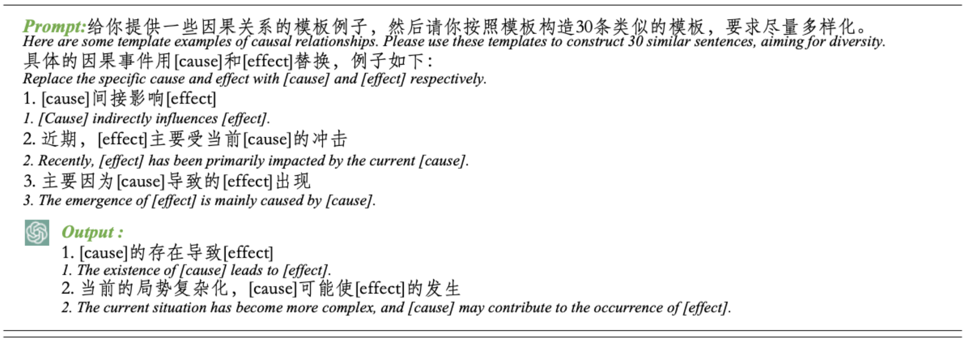
Output :判断: 0.输出结果不满足要求1，输出中'全球央行流动性宽松及经济复苏推进'没有按照最小的事件粒度进行抽取，应该拆分成'全球央行流动性宽松'和'经济复苏推进'

Judgment: 0. The output does not meet requirement 1 as 'global central bank liquidity easing and advancement of economic recovery' is not extracted at the minimum event granularity. It should be split into 'global central bank liquidity easing' and ' advancement of economic recovery '.

纠正: [['全球央行流动性宽松','大宗商品发生普涨行情'], ['经济复苏推进','大宗商品发生普涨行情']]

Correction: [['global central bank liquidity easing', 'general rise in commodity prices'], [' advancement of economic recovery ', 'general rise in commodity prices']]

Figure 4: Prompts and responses for dataset evolution.



Prompt:给你提供一些因果关系的模板例子，然后请你按照模板构造30条类似的模板，要求尽量多样化。

Here are some template examples of causal relationships. Please use these templates to construct 30 similar sentences, aiming for diversity.

具体的因果事件用[cause]和[effect]替换，例子如下:

Replace the specific cause and effect with [cause] and [effect] respectively.

1. [cause]间接影响[effect]

1. [Cause] indirectly influences [effect].

2. 近期，[effect]主要受当前[cause]的冲击

2. Recently, [effect] has been primarily impacted by the current [cause].

3. 主要因为[cause]导致的[effect]出现

3. The emergence of [effect] is mainly caused by [cause].

Output :

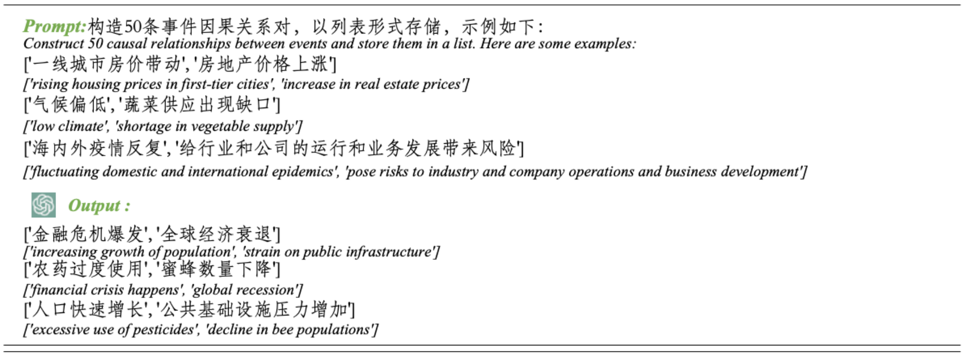
1. [cause]的存在导致[effect]

1. The existence of [cause] leads to [effect].

2. 当前的局势复杂化，[cause]可能使[effect]的发生

2. The current situation has become more complex, and [cause] may contribute to the occurrence of [effect].

Figure 5: Prompts and responses for contextual template generation.



Prompt:构造50条事件因果关系对，以列表形式存储，示例如下:

Construct 50 causal relationships between events and store them in a list. Here are some examples:

['一线城市房价带动','房地产价格上涨']

['rising housing prices in first-tier cities', 'increase in real estate prices']

['气候偏低','蔬菜供应出现缺口']

['low climate', 'shortage in vegetable supply']

['海内外疫情反复','给行业和公司的运行和业务发展带来风险']

['fluctuating domestic and international epidemics', 'pose risks to industry and company operations and business development']

Output :

['金融危机爆发','全球经济衰退']

['increasing growth of population', 'strain on public infrastructure']

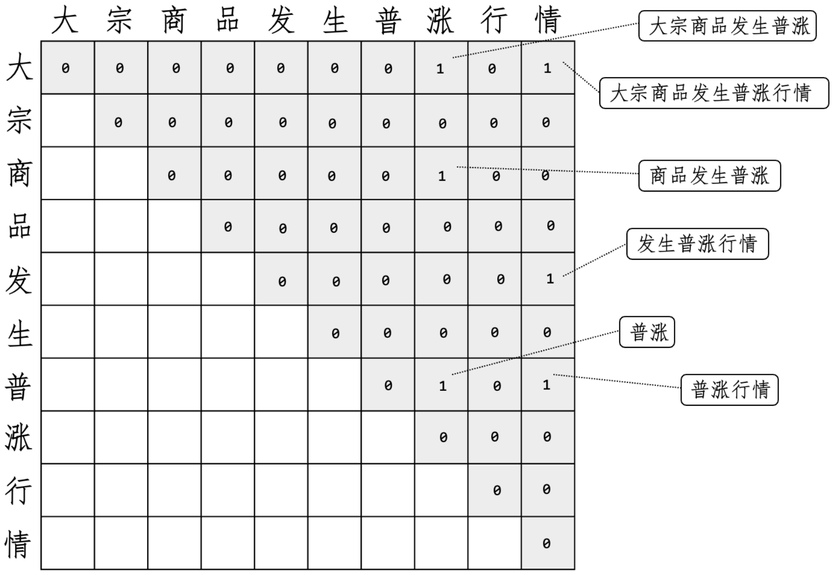
['农药过度使用','蜜蜂数量下降']

['financial crisis happens', 'global recession']

['人口快速增长','公共基础设施压力增加']

['excessive use of pesticides', 'decline in bee populations']

Figure 6: Prompts and responses for event span pair generation.



	大	宗	商	品	发	生	普	涨	行	情
大	0	0	0	0	0	0	0	1	0	1
宗		0	0	0	0	0	0	0	0	0
商			0	0	0	0	0	1	0	0
品				0	0	0	0	0	0	0
发					0	0	0	0	0	1
生						0	0	0	0	0
普							0	1	0	1
涨								0	0	0
行									0	0
情										0

Figure 7: Tagging schema in Global Pointer.