Identifying while Learning for Document Event Causality Identification

Cheng Liu[1], **Wei Xiang**[2], **Bang Wang**[1,*]
[1]School of Electronic Information and Communications,
Huazhong University of Science and Technology, Wuhan, China
[2]School of Software Engineering,
Huazhong University of Science and Technology, Wuhan, China
{liu_cheng, xiangwei, wangbang}@hust.edu.cn

Abstract

Event Causality Identification (ECI) aims to detect whether there exists a causal relation between two events in a document. Existing studies adopt a kind of *identifying after learning* paradigm, where events' representations are first learned and then used for the identification. Furthermore, they mainly focus on the causality existence, but ignore causal direction. In this paper, we take care of the causal direction and propose a new *identifying while learning* mode for the ECI task. We argue that a few causal relations can be easily identified with high confidence, and the directionality and structure of these identified causalities can be utilized to update events' representations for boosting next round of causality identification. To this end, this paper designs an *iterative learning and identifying framework*: In each iteration, we construct an event causality graph, on which events' causal structure representations are updated for boosting causal identification. Experiments on two public datasets show that our approach outperforms the state-of-the-art algorithms in both evaluations for causality existence identification and direction identification.[1]

1 Introduction

Event Causality Identification (ECI) is the task of identifying whether there exists a causal relation between two events. ECI can facilitate a wide range of practical applications, including knowledge graph construction (Chen et al., 2019; Al-Khatib et al., 2020), question answering (Oh et al., 2017), and information extraction (Xiang and Wang, 2023). The ECI task can be divided into the sentence-level ECI (two events are in the same sentence) and document-level ECI (two events may be in different sentences).

[*]Corresponding author

[1] The source code is available at https://github.com/LchengC/iLIF

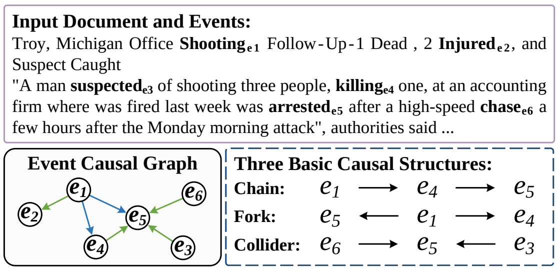

Figure 1: An example of the event causality graph and event structures in the EventStoryLine corpus.

In this paper, we focus on the document-level ECI task, which faces greater challenges due to the requirement of comprehending long texts for cross-sentence reasoning. The traditional feature-based methods (Gao et al., 2019) utilize Integer Linear Programming (ILP) to model the document causal structure. In order to better capture the interactions among events, recent methods (Phu and Nguyen, 2021; Chen et al., 2022, 2023; Fan et al., 2022) usually construct document-level undirected graphs to facilitate cross-sentence causal reasoning. Other methods (Yuan et al., 2023) use sparse attention to address the issue of long-distance dependencies and distinguish between intra- and inter-sentential reasoning.

Modeling the interactions among events has been proven effective for the document-level ECI task, however, almost all existing methods focus on only identifying the existence of causal relation between the event e_i and e_j, yet without considering the causality direction being from e_i to e_j (or from e_j to e_i). In this paper, $e_i \rightarrow e_j$ indicates that "event e_i causes e_j". This may lead to the learning of events' representations towards capturing events' correlations, but correlations may not be directly mapped into causalities (Pearl and Mackenzie, 2018). Furthermore, undirected connections may also lead to incorrect causality identifications, as some properties of causal structures cannot be

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3815–3827
August 11-16, 2024

respected without directionality.

There are three basic causal structures, namely, the chain, fork, and collider (He et al., 2021). Causality identification without directionality cannot well exploit causal structures. As shown in Figure 1, "*Shooting*$_{e1}$"→"*killing*$_{e4}$"→"*arrested*$_{e5}$" is a chain causal structure. For a model considering causality direction, if the two directional causal relations, i.e., $e_1 \rightarrow e_4$ and $e_4 \rightarrow e_5$, can be first identified with high confidence, then this can help to identify the causal relation between e_1 and e_5 due to the causal transmission in the chain structure. We argue that events' causal relation should be with directionality, and considering causal directions could further boost event causality identification.

Besides ignoring directionality, existing solutions for the ECI task adopt a kind of *identifying after learning* paradigm. That is, learning events' representations first via some advanced neural networks, and then identifying causal relations for all event pairs at only one pass. However, it could happen that some causal relations can be easily identified with high confidence. As reported by Yuan et al. (2023), identifying intra-sentence events' causality (two events in a same sentence) is often easier and with better accuracy than identifying inter-sentence events' causality (two events in different sentences). This motivates us to propose a new *identifying while learning* mode for the ECI task. That is, identifying some events' causal relations with high confidence, and then utilizing the directionality and structure of such identified causalities to update events' representations for boosting next round of causality identification.

Motivated from the aforementioned considerations, this paper proposes an *iterative Learning and Identifying Framework* (iLIF) for the document-level event causality identification. For an event e_i, we not only encode its contextual text representation $\mathbf{h}_i$, but also update its causal structure representation $\mathbf{z}_i$ in each iteration. Causality identification is modeled as a classification issue based on the representation $\mathbf{h}_i$ and $\mathbf{z}_i$ of an event pair. Initially, we employ a pretrained language model to encode $\mathbf{h}_i$. In each iteration, we first construct a directed *event causality graph* (ECG) based on the identified causalities, and propose a causal graph encoder to next update $\mathbf{z}_i$ on the ECG. After the termination, we output the directed ECG as the final causality identification results. In order to differentiate the importance of iterations, we design a novel iteration discounted loss function to mitigate the error propagation issue.

We conduct experiments on two public datasets: The EventStoryLine(v0.9) dataset (Caselli and Vossen, 2017) and MAVEN-ERE dataset (Wang et al., 2022) and consider both direction and existence settings for causal relations. We preprocess the EventStoryLine dataset [2] to ensure that each ground truth ECG is a directed acyclic graph (Gopnik et al., 2007). Experiment results validate that our iLIF outperforms the state-of-the-art competitors for the document-level ECI task in evaluations for both causality existence identification and direction identification.

2 Related work

Sentence-level ECI Early methods in feature engineering construct classifiers by searching for effective features, such as connective word categories (Zhao et al., 2016), syntactic features (Pitler et al., 2009), and contextual semantic features (Do et al., 2011). Some studies have employed external knowledge bases or linguistic tools. Cao et al. (2021) induce descriptive knowledge and relation path knowledge from the ConceptNet for reasoning. Liu et al. (2021) enhance model recognition capability by mining context-specific patterns from the ConceptNet. Zuo et al. (2020) design data augmentation methods based on lexical knowledge bases like the WordNet (Fellbaum, 1998) and VerbNet (Schuler, 2005) to generate more training data. Zuo et al. (2021) introduce specific causal patterns and transfer them to the target model using a contrastive transfer learning framework. Hu et al. (2023) utilize an AMR parser (Banarescu et al., 2013) to transform text into semantic graphs, enabling explicit semantic structure modeling and implicit association mining. Shen et al. (2022) leverage the prompt learning paradigm to jointly construct derived templates in order to utilize latent causal knowledge.

Document-level ECI Compared to the sentence-level ECI, document-level ECI holds greater potential for applications but also faces greater challenges, such as weak long-term dependencies and a lack of clear causal indicators. To address these challenges, researchers attempt to construct the topological structure of events for global reasoning. Gao et al. (2019) propose an approach based on Integer Linear Programming (ILP) to model the

[2] See Appendix A for more details.

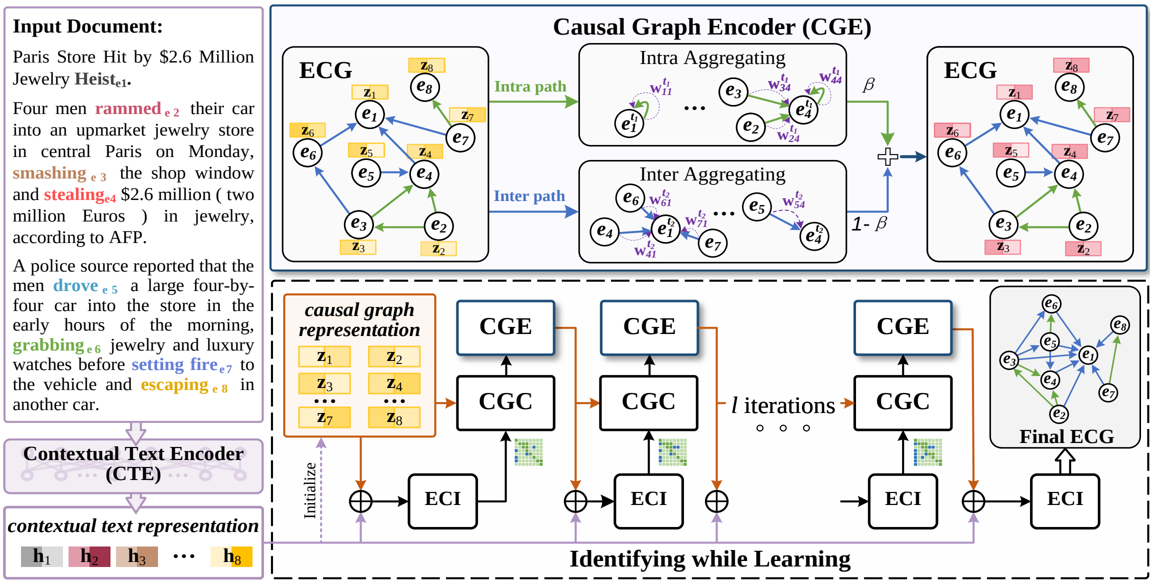

Figure 2: Illustration of the iterative learning and identifying framework (iLIF). Causality identification is based on events' contextual text and causal graph representations. The event causality graph (ECG) is iteratively constructed to update events' causal graph representations. The final ECG contains all identified causal relations as the output.

global causal structure. Phu and Nguyen (2021) construct a document-level event graph and utilize GNN to learn structural features. Chen et al. (2022) build a relational graph and modeled the interactions between event pairs. Furthermore, Chen et al. (2023) introduce prior knowledge such as central events and coreference to construct an event interaction graph for achieving global reasoning. Yuan et al. (2023) propose a method that utilizes sparse attention to learn high-quality representations and distinguish between intra- and inter-sentential reasoning.

To the best of our knowledge, existing methods ignore casuality direction and adopt the identifying after learning paradigm for the ECI task. In this paper, we take care of causality direction and propose an identifying while learning mode.

3 Methodology

We propose an *iterative Learning and Identifying Framework* (iLIF) for document-level event causality identification. The basic idea is to iteratively update events' representations for causality identification by exploiting causal structures on the most recent event causality graph. As shown in Figure 2, the iLIF includes four main module: (1) *Contextual Text Encoder* (CTE); (2) *Causal Graph Encoder* (CGE); (3) *Event Causality Identification* (ECI) (4) *Causality Graph Construction* (CGC).

3.1 Contextual Text Encoder

The CTE module is to encode contextual text information for each event mention. Given a document $\mathcal{D}$ with n sentences and the j-th sentence S_j contains m words. We use a Pretrained Language Model (PLM), say the BERT-base (Devlin et al., 2018), to sequentially encode the n sentences and output the encodings for all the words in the document $\mathcal{D}$. For the i-th event e_i, its *contextual text representation* $\mathbf{h}_i$ is an average of individual token representations of the event trigger words.

3.2 Causal Graph Encoder

Based on the contextual text representation $\mathbf{h}_i$, we can apply a simple neural classifier, such as a multi-layer perceptron, to identify the existence of a causal relation between two events. Although this naive approach is often not with excellent performance, some identified causal relations may be with high confidence. So we can choose some identified causal relations with high confidence to construct a *event causality graph* (ECG), denoted by $\mathcal{G}$. We will discuss how to construct and update $\mathcal{G}$ in the next subsection.

Besides describing the causality of two events, the ECG $\mathcal{G}$ contains more information about causal relations for all events in a document. Furthermore, some causal structures in $\mathcal{G}$ can be used to boost inferring new causal relations (such as the chain/fork causal structure) or to help correcting unreasonble causal relations (such as the collider causal structure). To encode causal direction and structure information from $\mathcal{G}$, we propose to learn a *causal graph representation* for each event node, denoted by $\mathbf{z}_i$ (with potentially different cardinality F).

We construct $\mathcal{G}$ as a heterogeneous directed graph containing one type of event node and two types of directed edges: Intra-sentence and Inter-sentence. We first utilize the self-attention (Vaswani et al., 2017) on the event nodes using a shared attentional mechanism $\mathsf{attention}: \mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}$ to compute the type-specific importance coefficient (Velickovic et al., 2017). For two events e_i and e_j with an edge r_{ij}^t of type t, the type-specific importance coefficient c_{ij}^t of r_{ij}^t is computed as follows:

$$c_{ij}^t = \mathsf{attention}(\mathbf{W}\mathbf{z}_i, \mathbf{W}\mathbf{z}_j\ ; t) \tag{1}$$

where $\mathbf{W}$ is a learnable parameter matrix. We note that the two edge types have different attention network parameters.

For an event node e_i, let $\mathcal{N}_i^t$ denote the set of its neighbors each with an edge e_{ji}^t. That is, $e_j \in \mathcal{N}_i^t$ indicates a directed edge from e_j to e_i. We next compute the weight w_{ji}^t for the edge e_{ji} as follows:

$$w_{ji}^t = \frac{\exp(c_{ji}^t)}{\sum_{k \in \mathcal{N}_i^t} \exp(c_{ki}^t)}. \tag{2}$$

Now, we update the causal graph representation for the event node e_i by the following multi-head attention mechanism as follows:

$$\mathbf{z}_i^t \leftarrow \|_{k=1}^K \sigma\Big(\sum_{j \in \mathcal{N}_i^t} w_{ji}^t \mathbf{W}^k \mathbf{z}_i\Big) \tag{3}$$

where $\sigma(\cdot)$ denotes an element-wise activation function and $\|$ the concatenation operation.

For an event e_i with only one edge type, then its causal graph representation is updated by Equation (3). For an event e_i connected by two types of edges (t_1 for intra-sentence and t_2 for inter-sentence edges), we combine these two types of features with different weights to distinguish their confidence in ECI. Since two events within the same sentence are usually easier to identify, we tend to give more weight to the features related to intra-sentence edges. Finally, we combine the features $\mathbf{z}_i^{t_1}$ and $\mathbf{z}_i^{t_2}$ as follows:

$$\mathbf{z}_i \leftarrow \beta \mathbf{z}_i^{t_1} + (1-\beta)\mathbf{z}_i^{t_2}. \tag{4}$$

event pairs within the same sentence is relatively easier to identify

3.3 Event Causality Identification

After learning the contextual text representation $\mathbf{h}$ and causal graph representation $\mathbf{z}$, we use a Multi-Layer Perceptron (MLP) to output a *causal relation vector* $\mathbf{p}_{ij} \in \mathbb{R}^3$ for two events e_i and e_j as follows:

$$\mathbf{p}_{ij} = \mathsf{softmax}([\mathbf{h}_i \,\|\, \mathbf{h}_j \,\|\, (\mathbf{z}_i - \mathbf{z}_j)]\mathbf{W}), \tag{5}$$

where $\|$ stands for concatenation operation and $\mathbf{W}$ is a learnable parameter matrix. We note that the subtraction of $\mathbf{z}_i$ and $\mathbf{z}_j$ is to emphasize the causal directionality, as they are learned from a directed causal graph. We write $\mathbf{p}_{ij} = (p_{ij}^n, p_{ij}^c, p_{ij}^e)$ with the element $p_{ij}^n/p_{ij}^c/p_{ij}^e$ denoting the probability of a NONE/CAUSE/EFFECT relation between the two events.

3.4 Causality Graph Construction

The CGC module is to construct a document-level *event causality graph* (ECG) in each iteration. The ECG $\mathcal{G} = (\mathcal{N}, \mathcal{R})$ contains events as nodes, and event causal relations as edges. Considering the different information density (Yuan et al., 2023) and recognition difficulty for inter-sentence event pairs and intra-sentence event pairs, we define two types of edges in $\mathcal{R}$: (1) Intra-sentence edges for two events in the same sentence, e.g., the green edge of *rammed* $\to$ *smashing* in Figure 2. (2) Inter-sentence edges for two events in two different sentences, e.g., the blue edge of *drove* $\to$ *stealing* in Figure 2.

To minimize error propagation, we resort to the causal relation vector $\mathbf{p}_{ij}$ by Equation (5) to employ only those identified causal relations with high confidence as edges during iteration. We define ω as the relation confidence threshold. For two events e_i and e_j, if p_{ij}^c is the largest and $p_{ij}^c \geq \omega$, then a directed edge e_{ij} is constructed from e_i to e_j; If p_{ij}^e is the largest and $p_{ij}^e \geq \omega$, then a directed edge e_{ji} is constructed from e_j to e_i. After processing all event pairs' $\mathbf{p}_{ij}$, we construct the ECG $\mathcal{G}$ with an *adjacency matrix* $\mathbf{A}$: $\mathbf{A}_{ij} = 1$, if there exists a directed edge from e_i to e_j; Otherwise, $\mathbf{A}_{ij} = 0$

In our *iterative Learning and Identifying Framework*, the ECG $\mathcal{G}$ is first initialized and then iteratively updated till the termination.

Initialization: We initialize the causal graph representation for each event node as the event contextual text representation, i.e., $\mathbf{z}^{(0)} = \mathbf{h}$, and construct an ECG $\mathcal{G}^{(0)}$ based on the $\mathbf{z}^{(0)}$ and $\mathbf{h}$.

Iteration: In the l-th iteration, we first use the previous ECG $\mathcal{G}^{(l-1)}$ to learn the new causal graph representations $\mathbf{z}^{(l)}$ for all nodes, as described in Section 3.2. We next construct a new ECG $\mathcal{G}^{(l)}$ based on the $\mathbf{z}^{(l)}$ and $\mathbf{h}$.

Termination: To address the scale differences among documents and prevent over iteration, we design an iteration condition that terminates upon meeting either of the following two criteria. (1) The iteration count l reaches a predefined *maximum iteration number* L. For a document with $n < L$ sentences, we set its maximum number of iterations as n. (2) If the *structural difference* between $\mathcal{G}^{(l)}$ and $\mathcal{G}^{(l-1)}$ is less than a predefined threshold:

$$\sum_i \sum_j |\mathbf{A}_{ij}^{(l)} - \mathbf{A}_{ij}^{(l-1)}| \leq \delta. \tag{6}$$

After terminating the iteration, all the directed edges in the final ECG $\mathcal{G}$ are the identified document-level causal relations.

3.5 Training strategy

Intuitively, the ECG $\mathcal{G}^{(l)}$ in each iteration carries different significance for the final ECG. We differentiate the iterations by first calculating an iteration-level loss $\mathcal{L}_l$ for the l-th iteration. We adopt the α-balanced variant of focal loss to address the issue of class imbalance and compute $\mathcal{L}_l$ by

$$\mathcal{L}_l = - \sum_{e_i, e_j \in \mathcal{D}} \alpha^c (1 - \hat{p}_{ij}^c)^\gamma \log(\hat{p}_{ij}^c), \tag{7}$$

where $\hat{p}_{ij}^c$ is the predicted probability of the true class c, α^c is a weighting factor for the true class c. Additionally, γ represents a predefined focusing hyper-parameter.

We note that during the identifying while learning process, if some misidentification happens in an iteration, it may propagate to the later iterations. To penalize such propagations, we emphasize the importance of earlier iterations and introduce a balancing factor inversely proportional to the iteration count in the final loss function. We define the final loss function as follows:

$$\mathcal{L} = \sum_{l=1}^{L_\mathcal{D}} \frac{1}{l} \mathcal{L}_l, \tag{8}$$

where $L_\mathcal{D}$ denotes the actual iteration counts of the document $\mathcal{D}$.

4 Experiments

4.1 Experimental Settings

Datasets Details We evaluate our iLIF on the widely used EventStoryLine (v0.9) dataset and MAVEN-ERE dataset. The EventStoryLine dataset comprises 22 topics, 258 documents, 5,334 event mentions, 1,770 intra-sentence causal event pairs and 3,855 inter-sentence causal event pairs (Caselli and Vossen, 2017). Following (Caselli and Vossen, 2017; Gao et al., 2019), we designate event pairs annotated with 'PRECONDITION' as CAUSE relation, and 'FALLING_ACTION' as EFFECT relation. In both settings, we utilize the last two topics as the development dataset, leaving the remaining 20 topics for 5-fold cross-validation.

The MAVEN-ERE dataset is a large-scale dataset, which comprises 4,480 documents, 103,193 events, and 57,992 causal event pairs. These causal event pairs are annotated as 'CAUSE' or 'PRECONDITION,' both representing CAUSE relations. We randomly reverse half of the event pairs with a CAUSE relation to represent the EFFECT relation. As MAVEN-ERE did not release the test set, following Tao et al. (2023), we use the original development set as the test set. Additionally, we sample 10% of the data from the original training set to form the development set.

implementation details Our method is implemented based on the PyTorch version of Huggingface Transformer (Wolf et al., 2020). We use the uncased BERT-base (Devlin et al., 2018) as the base PLM and fine-tune it during the training process. We optimize our model using AdamW (Loshchilov and Hutter, 2018), with a linear warm-up for the first 10% of steps. The learning rate for the PLM is set to 2e-5, while for other modules, it is set to 1e-4. The batch size is set to 1. More details can be found in Appendix B.

Evaluation Metrics We adopt the commonly used Precision (P), Recall (R), and F1-score (F1) as the evaluation metrics. In the direction evaluation, we calculate the micro-averaged results for Precision, Recall, and F1-score specifically for the CAUSE and EFFECT classes. In the existence evaluation, for the EventStoryLine dataset, we follow the same approach as previous methods to ensure fair comparison (Phu and Nguyen, 2021).

Direction and Existence Settings In the direction setting, we utilize three labels: NONE, CAUSE, EFFECT, which respectively represent the noncausal relation, the cause relation, and the effect relation. The *direction identification* results of competitors are derived by expanding the classification module of the competitors' algorithms from binary

Model	Intra-sentence			Inter-sentence			Intra+Inter		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Causality Direction Identification									
BERT (Devlin et al., 2018)	<u>62.4</u>	32.6	42.8	34.4	30.7	32.4	40.7	31.3	35.4
RoBERTa (Liu et al., 2019)	59.7	38	46.4	31.3	34.2	32.7	37.3	35.5	36.4
LONG (Beltagy et al., 2020)	59.0	40.5	48.0	35.2	30.5	32.7	41.6	33.8	37.3
ERGO (Chen et al., 2022)	58.8	47.6	52.6	36.1	<u>41.2</u>	38.5	41.5	43.3	42.4
SENDIR (Yuan et al., 2023)	56.0	<u>52.6</u>	<u>54.2</u>	<u>38.6</u>	39.4	<u>39.0</u>	<u>43.8</u>	<u>43.7</u>	<u>43.7</u>
LLaMA-2-7B (Gao et al., 2023)	17.5	17	17.2	6.8	19.2	10.0	8.3	18.5	11.5
iLIF (the proposed)	**66.7**	**54.5**	**60.0**	**41.2**	**44.6**	**42.8**	**47.9**	**47.8**	**47.8**
Causality Existence Identification									
BERT (Devlin et al., 2018)	60.4	45.7	52.0	30.6	39.1	34.3	37.2	41.2	39.1
RoBERTa (Liu et al., 2019)	62.7	45.4	52.7	32.7	38.3	35.3	39.7	40.6	40.1
LONG (Beltagy et al., 2020)	47.7	69.3	56.5	26.1	55.6	35.5	31.4	60.0	41.2
ERGO (Chen et al., 2022)	49.7	72.6	59.0	<u>43.2</u>	48.8	45.8	<u>46.3</u>	50.1	48.1
SENDIR (Yuan et al., 2023)	<u>65.8</u>	66.7	<u>66.2</u>	33.0	**90.0**	<u>48.3</u>	37.8	**82.8**	<u>51.9</u>
text-davinci-003 (Gao et al., 2023)	33.2	74.4	45.9	-	-	-	-	-	-
gpt-3.5-turbo (Gao et al., 2023)	27.6	<u>80.2</u>	41.0	-	-	-	-	-	-
gpt-4 (Gao et al., 2023)	27.2	**94.7**	42.2	-	-	-	-	-	-
LLaMA-2-7B (Gao et al., 2023)	26.9	29.3	28.0	10.8	31.9	16.1	13.2	31.1	18.5
iLIF (the proposed)	**76.8**	66.3	**71.2**	**53.5**	<u>65.9</u>	**59.1**	**59.2**	<u>66.1</u>	**62.5**

Table 1: Overall results on the EventStoryLine dataset in both direction and existence evaluation settings: The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>. Intra-sentence indicates that the event pair is within the same sentence, while Inter-sentence indicates that the event pair is in different sentences.

to ternary classification. In the existence setting, we adopt two labels: NONE, CAUSAL, which respectively represent noncausal relation and causal relation. For MAVEN-ERE dataset, we combine the event pairs identified as the CAUSE class and the EFFECT class in the direction experiment into the CAUSAL class event pairs to obtain the *existence identification* result.

4.2 Competitors

We compare iLIF with the following competitors:

PLM-base concatenates two events' contextual text representations and then identifies causality relation using a MLP. We use BERT-base (Devlin et al., 2018), RoBERTa-base (Liu et al., 2019), and Longformer-base (Beltagy et al., 2020) as the PLM.

ERGO (Chen et al., 2022) builds a relational graph to model interactions between event pairs.

SENDIR (Yuan et al., 2023) leverages intra-sentence event pairs to construct a reasoning chain, facilitating inter-sentence causality reasoning.

Large Language Models (LLMs). Gao et al. (2023) conduct zero-shot ECI experiments using OpenAI's official API [3], covering three versions of ChatGPT: text-davinci-003, gpt-3.5-turbo and gpt-4. We also test another popular LLM, the LLaMA2 (Touvron et al., 2023) of Llama-2-7b-chat version. Appendix D reports the designed prompts for the LLMs.

We consider both *existence identification* and *direction identification* for performance evaluations. The existence identification means to only identify the existence of causal relation between two events with causality direction. The direction identification means to correctly identify the causal direction, if existing, between two events.

4.3 Overall Results

Table 1 and Table 2 compare the overall results on the EventStoryLine and MAVEN-ERE dataset, respectively. We note that all the competitors adopt the *identifying after learning* paradigm; While our iLIF adopts the *identifying while learning*. Our iLIF achieves the best overall F1 results (intra+inter) on the two datasets in both existence and direction evaluation. This validates the superiority of our iLIF model with the new yet effective identifying while learning mode for the ECI task. For those LLMs, although they are capable of zero-shot causal reasoning, their performance is much worse than those methods based on the fine-tuned small PLMs. This suggests that LLMs, like ChatGPT and

[3] https://platform.openai.com/

Model	Intra-sentence			Inter-sentence			Intra+Inter		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Causality Direction Identification									
BERT (Devlin et al., 2018)	46.0	48.8	47.4	42.1	45.2	43.6	42.4	45.5	43.9
ERGO (Chen et al., 2022)	62.3	**63.1**	**62.7**	47.8	**59.8**	53.1	48.7	**60.1**	53.8
SENDIR (Yuan et al., 2023)	46.6	44.2	45.4	46.8	43.0	44.8	46.8	43.1	44.9
iLIF (the proposed)	**73.7**	49.7	59.4	**66.3**	47.5	**55.3**	**66.9**	47.6	**55.6**
Causality Existence Identification									
BERT (Devlin et al., 2018)	46.8	50.3	48.5	43.0	46.8	44.8	43.3	47.1	45.1
ERGO (Chen et al., 2022)	63.1	65.3	**64.2**	48.7	**62.0**	54.6	49.6	**62.3**	55.2
SENDIR (Yuan et al., 2023)	51.4	53.6	52.5	51.9	52.8	52.4	51.9	52.9	52.4
text-davinci-003 (Gao et al., 2023)	25.0	75.1	37.5	-	-	-	-	-	-
gpt-3.5-turbo (Gao et al., 2023)	19.9	85.8	32.3	-	-	-	-	-	-
gpt-4 (Gao et al., 2023)	22.5	**92.4**	36.2	-	-	-	-	-	-
iLIF (the proposed)	**74.4**	51.5	60.9	**67.1**	49.2	**56.8**	**67.7**	49.4	**57.1**

Table 2: Overall results on the MAVEN-ERE dataset in both direction and existence evaluation settings.

Model	Direction			Existence		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
iLIF	47.9	47.8	47.8	59.2	66.1	62.5
iLIF w/o Direction	42.3	49.3	45.5	55.9	63.8	59.6
iLIF w/o Heterogeneity	45.5	47.2	46.3	56.4	66.3	61.0
iLIF w/o Iteration	43.2	45.2	44.2	52.8	65.9	58.6

Table 3: Results of ablation study on EventStoryLine.

LLAMA, may not be effective causal reasoners for complex causal reasoning tasks.

Taking a close observation of the two tables, all models perform better on identifying causality in intra-sentence than that in inter-sentence. This is in accordance with the report in Yuan et al. (2023), as comprehending events in one sentence is generally easier with the same sentential context. We note that our iLIF achieves large improvements on the Precision in the intra-sentence causality identification. This can be attributed to our using a high confidence threshold, by which the intra-sentence causal relations are often with high identification confidence; While this also leads to a relatively lower Recall compared with other models. However, using such identified intra-sentence relations with high confidence can help improving inter-sentence identification, as the constructed event causality graph is becoming more confident with the iterations, on which events' causal structure representations can be well learned to further boost inter-sentence causality identification. This is evidenced from the high Precision and F1-score in the inter-sentence identification, and as a result, the overall intra+inter identification of our iLIF performs better than the competitors.

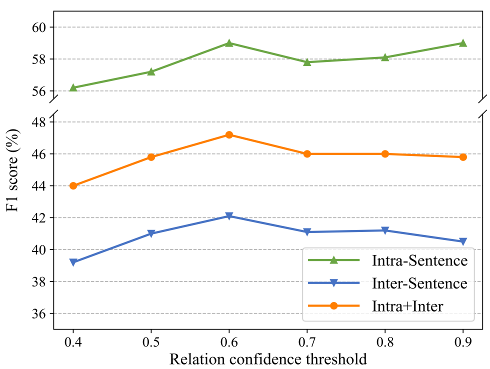

Figure 3: F1 scores on EventStoryLine when using different edge thresholds in the direction setting.

4.4 Ablation Study

Table 3 presents the ablation studies for examining the module functionalities.

(1) iLIF w/o Direction, which removes the directionality of the ECG. The removal of directionality causes a decrease in F1-scores by 4.8% and 4.6% in the two evaluations, respectively. Additionally, the decrease is primarily concentrated in the Precision. These results imply the importance of encoding causal direction and structure information on a directed ECG for learning potential events' interactions. We also observe that removing the direction increases the Recall in the direction evaluation. One possible reason is that the absence of direction increases the number of edges for encoding more interaction information. However, this also introduces more noisy edges, leading to inac-

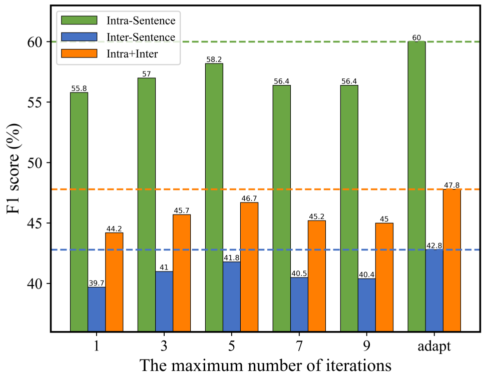

Figure 4: Results of different maximum number of iterations on EventStoryLine in the direction setting.

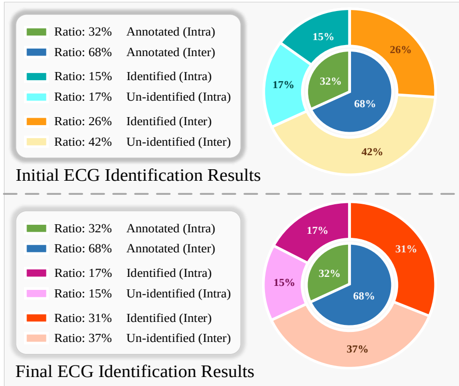

Figure 5: Comparison of causality identification results in the initial ECG and final ECG.

curate causal structure representations and worse F1 scores.

(2) iLIF w/o Heterogeneity, which only set one type of edges in the ECG, without distinguishing the intra-sentence and inter-sentence edges. We observe that removing edge heterogeneity results in a reduction in the Precision and F1-scores. One possible reason is that recognition difficulty does differ in the intra- and inter-sentence cases; While neglecting such differences cannot well utilize more confident intra-sentence relations when encoding on the ECG.

(3) iLIF w/o Iteration, which constructs the ECG only once for learning causal graph representations to output final identifications. As it does not evolve with the re-identified causal relations, it cannot enjoy the potentials of learning more accurate events' causal graph representations in the later iterations, so resulting in decreases of identification performance. The results again validate the effectiveness of our *identifying while learning* mode.

4.5 Parameter Study

Relation Confidence Threshold Figure 3 presents the results on examining the impact of using different relation confidence thresholds. It can be observed that using too low or two high thresholds introduces performance decrease. This is not unexpected. Using too low threshold would introduce more edges in the ECG, yet some may be incorrect ones, leading to inaccurate causal graph representation learning on the ECG. Setting the threshold too high results in a sparse ECG, which restricts the quality of the learned causal graph representation. Experiment results suggest to set 0.6 as the relation confidence threshold.

Number of Iterations Figure 4 presents the results on examining using different numbers of iterations. We can fix the number of iterations or use our termination criterion to adaptively adjust maximum iterations for different documents. As different documents may contain different numbers of sentences, it can be observed that using such fixed setting cannot well adapt to the length of documents. Furthermore, our criterion also enables to terminate iterations for two consecutive ECGs with small topological difference, hence helping to balance the under-fitting or over-fitting issue in the causal representation learning process.

Increasing the number of iterations not only affects the results but also increases the time complexity of the algorithm. We compare the complexity of different algorithms from both temporal and spatial perspectives. Experimental details can be found in Appendix C. The results demonstrate that our algorithm achieves a trade-off between algorithmic complexity and identification performance.

4.6 Exploration Study

Figure 5 compares the identification results of the initial and final ECG on the EventStoryLine testing dataset, which contains 32% intra-sentence and 68% inter-sentence causal relations. Compared with the initial ECG, the final ECG can increase the ratio of correct identification and decrease the ratio of un-identification for both intra-sentence and inter-sentence causal relations. This indicates that causality identification can benefit from the updating of events' representations during the identifying while learning process. Figure 6 presents a case of the identifying while learning process: Compared with the annotated intra-sentence and inter-sentence causality (i.e., the ground truth), the

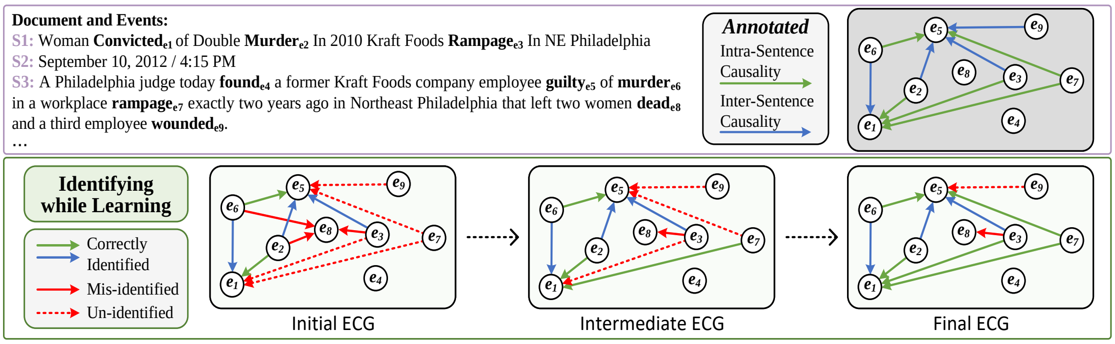

Figure 6: A case of the identifying while learning process on the EventStoryLine dataset. The initial ECG is with 3 mis-identified and 4 un-identified causal relations. After learning and updating events' causal graph representations from a few ECGs, some mis-identified relations can be corrected and some un-identified relations can be identified. At the termination, the final ECG is with only 1 mis-identified and 1 un-identified causal relation.

initially constructed ECG is with 3 mis-identified and 4 un-identified causal relations; While after learning and updating events' causal graph representations from a few ECGs, 2 mis-identified relations can be corrected and 1 un-identified can be correctly identified in an intermediate ECG. At the termination, the final ECG reduces mis-identified relations from 3 to 1 and un-identified from 4 to 1.

5 Conclusion

In this paper, we propose a novel *identifying while learning* mode where the central idea is to iteratively update events' representations for boosting next round of causality identification. Within the proposed *iterative Learning and Identifying Framework*, an event causality graph is constructed in each iteration based on previously identified causal relations with high confidence, which helps to mine events' directed interactions for updating their representations. Experiments on two widely used datasets have validated the superiority of this new working mode.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China (Grant No:62172167). The computation is supported by the HPC Platform of Huazhong University of Science and Technology.

Limitation

We preprocess the EventStoryLine dataset (Caselli and Vossen, 2017) to ensure that each ground truth ECG is a directed acyclic graph (Gopnik et al., 2007). However, We do not guarantee that the final *event causality graph* is a directed acyclic graph. In future work, we plan to introduce the directed acyclic constraint into the causality graph construction process to enhance the practical application effectiveness of the model.

Ethics Statement

Our research work meets the ethics of ACL. The proposed identifying while learning model can identify causal relations among events in a document. However, the algorithm is not perfect and may result in erroneous predictions. Therefore, researchers should not rely solely on the model to make real-world decisions.

References

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7367–7374.

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 the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4862–4872.

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

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. *arXiv preprint arXiv:2204.07434*.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. Cheer: Centrality-aware high-order event reasoning network for document-level event causality identification.

Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. Uhop: An unrestricted-hop relation extraction framework for knowledge-based question answering. *arXiv preprint arXiv:1904.01246*.

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*.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 294–303.

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

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

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

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817.

Alison Gopnik, Laura Schulz, and Laura Elizabeth Schulz. 2007. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Yue He, Peng Cui, Zheyan Shen, Renzhe Xu, Furui Liu, and Yong Jiang. 2021. Daring: Differentiable causal discovery with residual independence. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 596–605.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023. Semantic structure enhanced event causality identification. *arXiv preprint arXiv:2305.12792*.

Jian Liu, Yubo Chen, and Jun Zhao. 2021. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3608–3614.

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

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jong-Hoon Oh, Kentaro Torisawa, Canasai Kruengkrai, Ryu Iida, and Julien Kloetzer. 2017. Multi-column convolutional neural networks with causality-attention for why-question answering. In *Proceedings of the Tenth ACM international conference on web search and data mining*, pages 415–424.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 3480–3490.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

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

Zhengwei Tao, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Chengfeng Dou, Yongqiang Zhao, Fang Wang, and Chongyang Tao. 2023. Seag: Structure-aware event causality generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4631–4644.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat*, 1050(20):10–48550.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, et al. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *arXiv preprint arXiv:2211.07342*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 55(12):1–34.

Changsen Yuan, Heyan Huang, Yixin Cao, and Yonggang Wen. 2023. Discriminative reasoning with sparse event representation for document-level event-event relation extraction. ACL.

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

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.

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

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. Knowdis: Knowledge enhanced data augmentation for event causality detection via distant supervision. *arXiv preprint arXiv:2010.10833*.

A EventStoryLine Dataset Preprocessing

Following the assumption that "provided that pairs of events have a purely causal relationship, that is edges represent causal relations between the events, we will have a directed acyclic graph" (Gopnik et al., 2007). However, in the EventStoryLine dataset, not all document-level *event causality graphs* (ECG) constructed with event nodes and ground truth causal relations as edges are directed acyclic graphs.

A.1 Directed Acyclic Detection

For a document, we can construct an ECG $\mathcal{G}$ with an *adjacency matrix* $\mathbf{A}$: $\mathbf{A}ij = 1$ if there exists a causal relation from e_i to e_j; otherwise, $\mathbf{A}ij = 0$, based on ground truth causal relations. We employ the directed acyclic constraint proposed by NOTEARS (Zheng et al., 2018) on $\mathbf{A}$:

$$\mathcal{H}(\mathbf{A}) = tr(e^{\mathbf{A} \odot \mathbf{A}}) - d = 0 \quad (9)$$

where $tr(\cdot)$ denotes the matrix trace operation, $\odot$ is the Hadamard product, and d is the number of nodes. If the adjacency matrix $\mathbf{A}$ satisfies Equation (9), then $\mathcal{G}$ is a directed acyclic graph.

Finally, we find that 16 documents in the EventStoryLine dataset are not satisfied with this constraint.

A.2 Event Conflict Relation Detection

We note that Caselli and Vossen (2017) only label 2,265 causal relations, and the rest are extended using within-document event co-reference chains. One possible reason is that the expansion process introduces false causal relations. Considering the mutually exclusive relation between co-reference and causality in event relations, we conduct a further analysis of the dataset. We observe that some co-referenced event pairs are incorrectly labeled as causality. Since the EventStoryLine dataset is composed by the Event Coreference Bank+ corpus, we remove causal relations that conflict with co-reference relations in all documents.

A.3 Manual Check

After Event Conflict Relation Detection, we note that there are still 6 documents that do not satisfy the directed acyclic constraint. We invite 3 ECI task researchers to check the causal relations in these 6 documents according to the annotated requirements Caselli and Vossen (2017), aiming to minimize causal loops in the ground truth causal relations. Table 4 presents the statistics of the EventStoryLine v0.9 dataset before and after preprocessing.

Item	Original Size	Preprocessed Size
Topic	22	22
Documents	258	258
Sentences	4,316	4,316
Event Mentions	5,334	5,334
Intra-sentence causal links	1,770	1,751
Cross-sentence causal links	3,855	3,727
The Total causal links	5,625	5,478

Table 4: The EventStoryLine v0.9 dataset

B Implementation Details

Our method is implemented based on the PyTorch version of Huggingface Transformer (Wolf et al., 2020). We use the uncased BERT-base (Devlin et al., 2018) as the base PLM and fine-tune it during the training process. We optimize our model using AdamW (Loshchilov and Hutter, 2018), with a linear warm-up for the first 10% of steps. The learning rate for the PLM is set to 2e-5, while for other modules, it is set to 1e-4. The batch size is set to 1. The number of attention head K is set to 4. The weights β is set to 0.7. The relation confidence threshold ω is set to 0.6. We set the maximum iteration number L to 9 for the EventStoryLine dataset and 19 for the MAVEN-ERE dataset. The structural difference threshold δ is set to 2. For loss function, we set the focusing parameter γ to 2, the weighting factor of CAUSE/EFFECT class α to 0.75, and the weighting factor of NONE class is $1-\alpha$. The MLP is a two-layer fully connected network, in which the activation function is LeakyReLU and the rate of dropout is 0.4.

The MAVEN-ERE primarily comprises lengthy documents with a large number of sentences. Capturing sufficient semantics using a single-sentence encoding approach is challenging. Therefore, for experiments on the MAVEN-ERE dataset, we employed the same encoding method as Chen et al. (2022), leveraging dynamic window and event marker techniques. In addition, in the existence setting, to fulfill the requirement of constructing a directed event causality graph, we set all the positive samples to the CAUSE direction for the EventStoryLine dataset.

Method	Time Per Epoch	GPU Per Batch
SENDIR (Yuan et al., 2023)	2362 s	24 GB
ERGO (Chen et al., 2022)	**131 s**	24 GB
iLIF	823 s	**16 GB**

Table 5: Results of algorithm complexity comparison on the EventStoryLine dataset. Bold numbers represent the smallest cost.

C Algorithm Complexity Comparison

Table 5 compares the average resource consumption of various algorithms on the EventStoryLine (v0.9) dataset, considering both time and space perspectives on an NVIDIA RTX 3090 GPU with 24GB memory.

From the Table, we observe that the directional causal graph generation and iterations for causality updates increase the time complexity of algorithm. However, due to the excellent identification performance of our model, we believe that the time cost is within an acceptable range, especially compared to the Yuan et al. (2023). Additionally, due to the adopted ECG updating approach, our method performs exceptionally well in terms of spatial complexity.

We argue that the key to balancing algorithmic complexity and recognition performance lies in the actual number of iterations. Therefore, we propose termination criterion to achieve the trade-off between algorithm complexity and identification performance.

D LLaMA2 Prompt Details

We evaluate LLaMA2's performance under zero-shot settings. Following previous works (Gao et al., 2023), only the top 20 topics of EventStoryLine dataset are used for evaluation. For the event e_i and event e_j, we design two prompt templates as follows:

- *causality existence prompt*

  Input: {document content}

  Question: is there a causal relationship between " $\{e_i\}$ " and " $\{e_j\}$ " ? Let's think step by step.

- *causality direction prompt*

  Question: is the causal direction between " $\{e_i\}$ " and " $\{e_j\}$ " from " $\{e_i\}$ " to " $\{e_j\}$ "? Let's think step by step.

Demonstration of LLaMA2's reasoning

Input: Lindsay Lohan **" Leaves "** Betty Ford , Checks Into Malibu Rehab First Published : June 13 , 2013 4 : 59 PM EDT Lindsay Lohan has left the Betty Ford Center and is **" moving "** to a rehab facility in Malibu , Calif . , Access Hollywood has confirmed . A spokesperson for The Los Angeles Superior Court confirmed to Access that a judge signed an order yesterday allowing the transfer to Cliffside , where she will continue with her 90 - day court - mandated rehab . Lohan ' s attorney , Shawn Holley , spoke out about the move . " Lindsay is grateful for the treatment she received at the Betty Ford Center . She has completed her course of treatment there and looks forward to continuing her treatment and building on the foundation established at Betty Ford , " Holley said in a statement to Access . The actress checked into the Betty Ford Center in May as part of a plea deal stemming from her June 2012 car accident case .
Question: is there a causal relationship between **" Leaves "** and **" moving "** ? Let's think step by step.
Answer: **Yes**, there is a causal relationship between **" Leaves "** and **" moving "**.
Reasoning Process:
1. Lindsay Lohan left the Betty Ford Center. (Leaves)
2. She is moving to a rehab facility in Malibu. (Moving)
The two events are closely related, as Lindsay Lohan is leaving one rehab facility (Betty Ford Center) to move to another (Cliffside) for her 90-day court-mandated rehab. Therefore, there is a causal relationship between **" Leaves "** and **" moving "**.
Question: is the causal direction between **" Leaves "** and **" moving "** from **" Leaves "** to **" moving "** ? Let's think step by step.
Answer: **Yes**, the causal direction between **" Leaves "** and **" moving "** is from **" Leaves "** to **" moving "**.
Reasoning Process:
The sentence "Lindsay Lohan has left the Betty Ford Center" implies that she is no longer at the Betty Ford Center, which is the cause of her moving to another rehab facility. Therefore, the causal direction is from **" Leaves "** (Lindsay Lohan leaving the Betty Ford Center) to **" moving "** (Lindsay Lohan moving to another rehab facility).

Figure 7: An example of LLaMA2 reasoning ECI task

where we fill the templates with event mentions in the document to reference specific events.

We first input the *causality existence prompt* to obtain the *existence identification* results. If there is a causal relation, we input the *causality direction prompt* for continuous questioning to obtain the *direction identification* results. Figure 7 shows an example of LLaMA2 reasoning ECI task.