PPAT: Progressive Graph Pairwise Attention Network for Event Causality Identification

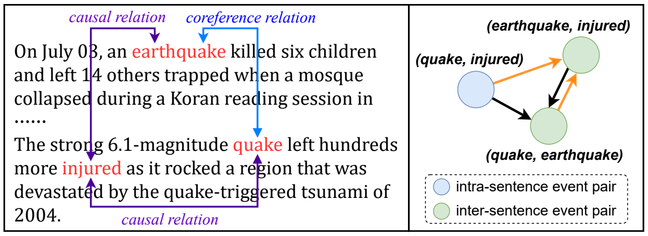
Zhenyu Liu , **Baotian Hu**[*] , **Zhenran Xu** and **Min Zhang**
Harbin Institute of Technology, Shenzhen
liuzhenyuhit@gmail.com, xuzhenran@stu.hit.edu.cn, {hubaotian, zhangmin2021}@hit.edu.cn

Abstract

Event Causality Identification (ECI) aims to identify the causality between a pair of event mentions in a document, which is composed of sentence-level ECI (SECI) and document-level ECI (DECI). Previous work applies various reasoning models to identify the implicit event causality. However, they indiscriminately reason all event causality in the same way, ignoring that most inter-sentence event causality depends on intra-sentence event causality to infer. In this paper, we propose a **P**rogressive graph **P**airwise **A**ttention ne**t**work (PPAT) to consider the above dependence. PPAT applies a progressive reasoning strategy, as it first predicts the intra-sentence event causality, and then infers the more implicit inter-sentence event causality based on the SECI result. We construct a sentence boundary event relational graph, and PPAT leverages a simple pairwise attention mechanism, which attends to different reasoning chains on the graph. In addition, we propose a causality-guided training strategy for assisting PPAT in learning causality-related representations on every layer. Extensive experiments show that our model achieves state-of-the-art performance on three benchmark datasets (5.5%, 2.2% and 4.5% F1 gains on EventStoryLine, MAVEN-ERE and Causal-TimeBank). Code is available at https://github.com/HITsz-TMG/PPAT.

1 Introduction

Event Causality Identification (ECI) seeks to identify the causal relation between two events in text. For example, as shown in Figure 1, in the sentence "*The strong 6.1-magnitude quake left hundreds more injured ...*", the ECI model should identify the causality between "*quake*" and "*injured*". ECI presents the causal structure of text, which is beneficial to a wide range of applications in natural language processing (NLP), including future event forecasting [Hashimoto, 2019], machine reading comprehension [Berant *et al.*, 2014], and question answering [Oh *et al.*, 2016].



On July 03, an earthquake killed six children and left 14 others trapped when a mosque collapsed during a Koran reading session in ......

The strong 6.1-magnitude quake left hundreds more injured as it rocked a region that was devastated by the quake-triggered tsunami of 2004.

Figure 1: Example of ECI and SERG. The purple lines denote target causal relations. The coreference relation assists reasoning, denoted by the blue line. In SERG, the nodes of intra- and inter-sentence event pairs are in blue and green respectively. The orange edges denote a reasoning chain.

ECI consists of two parts: sentence-level ECI (SECI) [Liu *et al.*, 2020] which aims to identify the intra-sentence event causality, and document-level ECI (DECI) [Gao *et al.*, 2019] which aims to identify the inter-sentence event causality. Previous studies [Phu and Nguyen, 2021; Chen *et al.*, 2022] do not explicitly distinguish intra- and inter-sentence event causality and use the same model to learn their representation, yet intra- and inter-sentence causality are expressed differently. Most intra-sentence event causality are explicitly expressed with causal cues in local context. Take Figure 1 as an example, the causality of intra-sentence event pair "(*quake*, *injured*)" could be identified easily with the causality indicator "*left*". However, inter-sentence event causality is more implicitly expressed with multiple sentences, and needs to be inferred from intra-sentence event causality. As shown in Figure 1, based on the above intra-sentence event causality and coreference relation "(*quake*, *earthquake*)", we can propagate the causality via the coreference chain and infer that the event pair "(*earthquake*, *injured*)" also has causality.

In this paper, we aim to address the above issue by presenting a novel Progressive Graph Pairwise Attention Network (PPAT). PPAT applies a **progressive reasoning strategy**, i.e., it first predicts the intra-sentence causality with local context, and then reasons the inter-sentence causality based on the previous SECI prediction, taking the dependence of inter-sentence causality on intra-sentence causality into consideration. For the implementation of progressive reasoning, we construct a Sentence boundary Event Relational Graph (SERG). Each node of SERG denotes an event pair, and

[*] Corresponding author.

two nodes that share one event have two directed edges connecting with each other. Specially, the intra-sentence nodes only connect with the intra-sentence nodes in SERG. Figure 1 shows an example. The intra-sentence node (in blue) does not have edges directed from inter-sentence nodes (in green), while inter-sentence nodes can aggregate information from the intra-sentence node via directed edges. The edges model the interaction between directly related node pairs that share the same event, and encourage PPAT to reason the intra-sentence event causality on the local patterns.

Moreover, different from previous work that uses graph neural networks to simply aggregate representations of neighborhood nodes (i.e., event pairs) [Phu and Nguyen, 2021; Chen *et al.*, 2022], PPAT applies a simple **pairwise attention** mechanism, which aggregates neighbors at a reasoning chain level instead of node level. Take Figure 1 as an example. When the node of "(*earthquake*, *injured*)" is the target node to be reasoned, its two neighbors form a premise node pair if they contain the same event that the target node does not contain, e.g., nodes of "(*quake*, *injured*)" and "(*quake*, *earthquake*)". Then the causality of the target node could be reasoned via the following reasoning chain: $\textit{Cause}(\textit{quake}, \textit{injured}) \wedge \textit{Coreference}(\textit{earthquake}, \textit{quake}) \rightarrow \textit{Cause}(\textit{earthquake}, \textit{injured})$. Therefore, the reasoning model should regard the premise node pair as a whole part and aggregate neighbors at a reasoning chain level. To this end, our proposed pairwise attention mechanism can capture interaction between the target node and its premise node pairs, thus attending to the possible reasoning chains and inferring the target causality.

In addition, we propose a causality-guided training strategy for PPAT. Since node representations on every layer of PPAT will be served as auxiliary information for reasoning on the next layer, it is important for every layer of PPAT to learn causality-related node representations, so we apply an additional loss to provide causality supervision on every layer and assist PPAT to have better reasoning performance.

To summarize, our contributions can be listed as:

- We propose a novel progressive graph pairwise attention network (PPAT), which reasons progressively on the sentence boundary event relational graph. To the best of our knowledge, we are the first to capture the dependence of inter-sentence causal reasoning on intra-sentence causality.
- We propose a pairwise attention mechanism, a simple yet effective approach to attending to reasoning chains on the graph for causality propagation.
- Extensive experiments on three ECI datasets show that PPAT significantly outperforms previous state-of-the-art methods, demonstrating the effectiveness of our method.

2 Related Work

Early feature-based methods for SECI mainly focus on designing better causality features or using external resources to improve performance, including the lexicon of causality indicators [Mirza, 2014; Hidey and McKeown, 2016], temporal patterns [Mirza, 2014; Ning *et al.*, 2018], event semantics [Riaz and Girju, 2014a; Riaz and Girju, 2014b], event co-occurrence [Do *et al.*, 2011; Hu *et al.*, 2017], and weakly supervised data [Hashimoto, 2019]. As Pre-trained Language Models (PLMs) have achieved great success in a wide range of NLP tasks, many SECI work shows promising performance gains based on PLMs [Kadowaki *et al.*, 2019; Liu *et al.*, 2020; Zuo *et al.*, 2020].

In recent years, more and more studies pay attention to document-level NLP tasks, such as event argument extraction [Li *et al.*, 2021] and relation extraction [Yao *et al.*, 2019]. Recent ECI work focuses on global inference: Gao *et al.* [2019] use Integer Linear Programming (ILP) to model global causal structures; RichGCN [Phu and Nguyen, 2021] utilizes several NLP tools (e.g., dependency parser) and external corpus for building event graphs, and uses graph convolutional network [Kipf and Welling, 2017] for reasoning. ERGO [Chen *et al.*, 2022] achieves state-of-the-art (SOTA) performance with a graph transformer on an event relational graph for high-order interaction of event relations. Compared with previous work, our model focuses on reasoning progressively and attending to reasoning chains, with no need for sophisticated graph design, external NLP tools or external knowledge.

3 Methods

As illustrated in Figure 2, the overall architecture consists of two tiers: (1) A document encoder yields event contextual representations, then concatenates the event representations for initial event pair representations. (2) The intra- and inter-sentence pairwise attention layers reason the event pair causality representation progressively, and then a classifier predicts causality based on the learned representations.

3.1 Document Encoder

Given a document $\mathcal{D} = \{w_i\}_{i=1}^{L_\mathcal{D}}$ containing $L_\mathcal{D}$ words with event mention set $\mathcal{N}$ ($|\mathcal{N}| = N$), Document Encoder aims to represent all event pairs. We use BERT [Devlin *et al.*, 2019] and Longformer [Beltagy *et al.*, 2020] respectively as a basic encoder to obtain contextualized embeddings. For the document longer than the length limitation of encoder, we use a *dynamic window* to encode the entire document. Specifically, we divide $\mathcal{D}$ into overlapping spans according to a fixed step and input them to the encoder separately.

We apply the *levitated marker* [Zhong and Chen, 2021] to represent the event mentions in the document. Specifically, for each event mention, we add two marker tokens (i.e., t_1 and t_2) to the end of text. t_1 will share position embedding with the first token of the event mention, and t_2 will share position embedding with the last token of the event mention. By setting the attention matrix, the original document tokens cannot attend to the marker tokens, and each marker pair can only attend to the corresponding event mention tokens. We also insert "`[CLS]`" at the start of document ("`<s>`" for Longformer). The input text for BERT encoder could be written as follows:

$$S = \texttt{[CLS]}, w_1, w_2 \cdots event_i \cdots w_{L_\mathcal{D}} \cdots t_1^i, t_2^i \cdots$$

where w_x denotes the x-th words of document, t_1^i and t_2^i are the levitated markers associated with the event mention

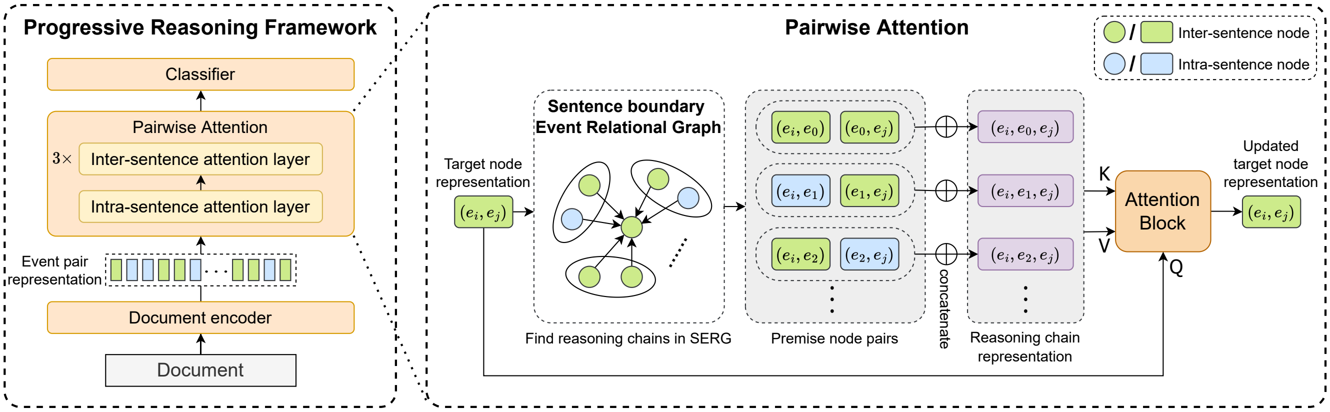

Progressive Reasoning Framework
Classifier
Pairwise Attention
3×
Inter-sentence attention layer
Intra-sentence attention layer
Event pair representation
Document encoder
Document
Pairwise Attention
Inter-sentence node
Intra-sentence node
Target node representation
(e_i, e_j)
Sentence boundary Event Relational Graph
Find reasoning chains in SERG
(e_i, e_0) (e_0, e_j)
(e_i, e_1) (e_1, e_j)
(e_i, e_2) (e_2, e_j)
Premise node pairs
concatenate
(e_i, e_0, e_j)
(e_i, e_1, e_j)
(e_i, e_2, e_j)
Reasoning chain representation
K
V
Q
Attention Block
Updated target node representation
(e_i, e_j)

Figure 2: The overall architecture of our PPAT (left) and the detail of pairwise attention (right). With initial event pair representations from the document encoder, the intra- and inter-sentence pairwise attention layers consecutively update representations. The pairwise attention mechanism finds all possible reasoning chains of the target node in Sentence boundary Event Relational Graph (SERG), uses target node representation as query (Q) and reasoning chain representations as key (K) and value (V) for the attention block.

$event_i$, $L_{\mathcal{D}}$ is the length of document. We use BERT or Longformer to encode S and then obtain the representation of $event_i$, denoted as e_i, as follows:

$$e_i = \frac{H(t_1^i) + H(t_2^i)}{2} \oplus H(\texttt{[CLS]}) \tag{1}$$

where $\oplus$ denotes concatenation, $H(*)$ denotes the contextualized word embedding computed by the encoder. Then the raw representation of the event pair $(event_i, event_j)$, i.e. r_{ij}, can be obtained by the following equation:

$$r_{ij} = e_i \oplus e_j \oplus (e_i * e_j) \tag{2}$$

where e_i and e_j are the representation of $event_i$ and $event_j$ respectively, $*$ means pointwise product.

3.2 Progressive Reasoning Strategy

As stated in Section 1, the intra- and inter-sentence event causality should be reasoned separately. We thus propose a progressive reasoning strategy for learning the high-order relation of event pairs. Furthermore, we build a sentence boundary relational event graph (SERG) $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ for constraining event pair interaction, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges.

With the initial node representations from the document encoder as input, PPAT first reasons sentence-level causality with single layer, and then reasons document-level causality with three layers. The output representations in the last layer is used for causality prediction. Note that the three layers for document-level reasoning share parameters. Here we only introduce the input and output of PPAT on each layer, leaving the details of graph pairwise attention to Section 3.3.

For the node of $(event_i, event_j)$, after updating its representation at layer l, we obtain the input node embedding for the next layer, i.e., n_{ij}^{l+1}, by concatenating causality prediction, a binary intra-sentence marker and the updated node representation in layer l:

$$n_{ij}^{l+1} = v_{ij}^l \oplus p_{ij}^l \oplus a_{ij} \tag{3}$$

where $\oplus$ denotes concatenated operations. v_{ij}^l is the node representation output in the l-th layer. a_{ij} is 1 if $(event_i, event_j)$ is an intra-sentence event pair, otherwise a_{ij} is 0. p_{ij}^l is the predicted causality possibility of $(event_i, event_j)$ in the l-th layer. We use a binary classifier to predict the causality of nodes in each layer.

$$p_{ij}^l = \text{softmax}(v_{ij}^l \mathbf{W}_c) \tag{4}$$

where $\mathbf{W}_c$ is the parameter weight matrix in the linear classifier. Before the first step of reasoning (i.e., $l = 0$), the node embedding is initialized by the raw event pair representation r_{ij} from the document encoder.

3.3 Graph Pairwise Attention

In order to introduce reasoning chain-level information into representation learning, we propose a graph pairwise attention for SERG. As shown in the right part of Figure 2, when $(event_i, event_j)$ is the target node to be reasoned, its premise node pairs are defined as $((event_i, event_k)$, $(event_j, event_k))$, where $0 \le k < N$, $k \ne i \ne j$. Then we perform a pairwise self-attention mechanism to measure the importance of each premise node pair for the target node:

$$\text{atten}_{ij,k} = \frac{(n_{ij}\mathbf{W}_q)((n_{ik} \oplus n_{jk})\mathbf{W}_k)^T}{\sqrt{d}} \tag{5}$$

where n_{ij} is the input node embedding of $(event_i, event_j)$ described in Section 3.2, $\mathbf{W}_q$, $\mathbf{W}_k$ are parameter weight matrices, $\sqrt{d}$ is a scaling factor and d is the hidden size.

Then we normalize the attention coefficients:

$$\begin{aligned} \alpha_{ij,k} &= \text{softmax}_{ij}(\text{atten}_{ij}) \\ &= \frac{mask_{ij,k} \exp(\text{atten}_{ij,k})}{\sum_{z \in \mathcal{N}_{ij}^-} mask_{ij,z} \exp(\text{atten}_{ij,z})} \end{aligned} \tag{6}$$

where $\mathcal{N}_{ij}^-$ is the event mention set without $event_i$ and $event_j$. The attention mask $mask_{ij,k}$ is 1 if node of $(event_i$, $event_j)$ have edges directed from the premise node pair, i.e.,

$(event_i, event_k)$ and $(event_k, event_j)$, otherwise $mask_{ij,k}$ is 0. After obtaining the normalized attention coefficients $\alpha_{ij,k}$, we aggregate relational knowledge from each reasoning chain:

$$v_{ij}^l = \sum_{k \in \mathcal{N}_{ij}^-} \alpha_{ij,k}((n_{ik} \oplus n_{jk})\mathbf{W}_v) \tag{7}$$

where $\mathbf{W}_v$ is the parameter weight matrix.

Following Vaswani *et al.* [2017], we also perform multi-head attention to combine the information from different representation subspaces. The final output embedding of node $(event_i, event_j)$ can be represented as:

$$v_{ij}^l = \Big(\Big\|_{c=1}^{C} \sum_{k \in \mathcal{N}_{ij}^-} \alpha_{ij,k}((n_{ik} \oplus n_{jk})\mathbf{W}_v)\Big)\mathbf{W}_o, \tag{8}$$

where $\|$ and $\oplus$ are both concatenation operation, C is the number of heads, $\mathbf{W}_o$ is the parameter weight matrix.

3.4 Training Objective

Following Chen *et al.* [2022], we adopt the focal loss [Lin *et al.*, 2017] to address the imbalance of positive and negative examples, as most of the event pairs have no causal relations:

$$\mathrm{FL}(\hat{p}) = -\beta(1-\hat{p})^\gamma \log(\hat{p}) \tag{9}$$

where $\hat{p}$ is the predicted possibility of right label, β is a weighting factor to balance the huge number of negative examples. $\gamma(\gamma \geq 0)$ is a focusing parameter.

We calculate the main loss $\mathcal{L}_m$ with the predicted causality possibility at the last layer (i.e., p_{ij}^L):

$$\mathcal{L}_m = \sum_{(i,j) \in \mathcal{M}} \mathrm{FL}(p_{ij}^L) \tag{10}$$

where $\mathcal{M}$ is the event pair set, L is the number of layers.

We adopt a causality-guided training strategy to assist PPAT to learn causality-related representation on each layer. Specifically, we use the predicted causality possibility on each layer p_{ij}^l computed from Equation 4 and calculate the focal loss as follows:

$$\mathcal{L}_c = \sum_{0 \leq l \leq L-1} (\lambda^l \sum_{(i,j) \in \mathcal{M}^l} \mathrm{FL}(p_{ij}^l)), \tag{11}$$

where λ^l is loss weight in the l-th layer. $\mathcal{M}^l$ is the focused event pair set in the l-th layer (in the first layer $\mathcal{M}^l$ is intra-sentence event pair set, otherwise $\mathcal{M}^l$ is inter-sentence event pair set). PPAT's final loss is given by:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_c \tag{12}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate PPAT on three datasets: EventStoryLine (version 0.9) [Caselli and Vossen, 2017], MAVEN-ERE [Wang *et al.*, 2022] and Causal-TimeBank [Mirza, 2014].

EventStoryLine contains 258 documents across 22 topics, 5334 event mentions, 10347 intra-sentence event pairs and 60232 inter-sentence event pairs (1770 and 3885 of them have causal relations respectively). Following previous work [Gao *et al.*, 2019; Chen *et al.*, 2022], we use documents in the last two topics as development set, and employ 5-fold cross-validation on the remaining documents.

MAVEN-ERE contains 3555 documents, 85912 event mentions, 97521 intra-sentence event pairs and 1226168 inter-sentence event pairs (16044 and 47108 of them have causal relations respectively). Since the original test set does not contain gold labels, we divide the development set into a new development set and a new test set, both of which contain 348 documents. As MAVEN-ERE is a relatively new dataset, we reproduce the SOTA method and several strong baselines.

Causal-TimeBank contains 183 documents, 6811 event mentions, 7608 intra-sentence event pairs (300 of them have causal relations). Following previous work [Liu *et al.*, 2020; Chen *et al.*, 2022], we employ 10-fold cross-validation evaluation for intra-sentence event pairs. Note that the number of inter-sentence causal event pairs is quite small (only 20 of 252084 inter-sentence event pairs). Following the above previous work, we only evaluate the performance of SECI on Causal-TimeBank.

Evaluation Metrics We adopt Precision (P), Recall (R) and F1-score (F1) as evaluation metrics, same as previous work [Gao *et al.*, 2019; Phu and Nguyen, 2021; Chen *et al.*, 2022].

4.2 Implementation Details

We employ *BERT-BASE-UNCASED* [Devlin *et al.*, 2019] or *Longformer-base* [Beltagy *et al.*, 2020] as the encoder. The models are optimized with AdamW [Loshchilov and Hutter, 2019] with the learning rate of 1e-5 and weight decay of 0.01. We use the linear warmup with 0.1 warmup ratio. We apply a dynamic window to encode the entire document. The window length is 512 for BERT and 2048 for Longformer, and the shift step is 120 for BERT and 500 for Longformer. We train the model for 128 epochs on EventStoryLine, 64 on Causal-TimeBank and MAVEN-ERE. We choose the best checkpoint on the development set for testing. As token-level attention cannot be set on Longformer, we use the solid marker, i.e. inserting marker tokens before and after the event mention, and set "<s>" and the marker tokens as global tokens. The loss weight λ^l are set as 2, 6, 0.1, 0.3 for l from 0 to 3. We run all the experiments on a single NVIDIA A100 GPU.

4.3 Baselines

SECI baseline We compare PPAT with the following SECI methods: (1) **KMMG** [Liu *et al.*, 2020] leverages external knowledge and proposes a mention masking generalization method for accurate reasoning. (2) **KnowDis** [Zuo *et al.*, 2020] uses a knowledge-enhanced data augmentation method to tackle the data lacking problem. (3) **LSIN** [Cao *et al.*, 2021] uses a descriptive graph induction module for exploiting external structural knowledge. (4) **LearnDA** [Zuo *et al.*, 2021b] proposes a knowledge-guided dual learning method for data augmentation. (5) **CauSeRL** [Zuo *et al.*, 2021a]

Model	**EventStoryLine (SECI)**			**EventStoryLine (DECI)**			**EventStoryLine (Overall)**		
	P	R	F1	P	R	F1	P	R	F1
OP	10.5	99.2	19.0	3.0	40.7	5.5	10.5	99.2	19.0
LR+	22.5	98.6	36.6	8.4	99.5	15.6	27.9	47.2	35.1
LIP	38.8	52.4	44.6	35.1	48.2	40.6	36.2	49.5	41.9
KMMG	41.9	62.5	50.1	-	-	-	-	-	-
KnowDis	39.7	66.5	49.7	-	-	-	-	-	-
LSIN	47.9	58.1	52.5	-	-	-	-	-	-
LearnDA	42.2	69.8	52.6	-	-	-	-	-	-
CauSeRL	41.9	69.0	52.1	-	-	-	-	-	-
BERT	47.8	57.2	52.1	36.8	29.2	32.6	41.3	38.3	39.7
Longformer*	71.7	47.5	57.2	56.1	38.6	45.7	60.9	41.4	49.3
RichGCN	49.2	63.0	55.2	39.2	45.7	42.2	42.6	51.3	46.6
ERGO	49.7	72.6	59.0	43.2	48.8	45.8	46.3	50.1	48.1
ERGO*	57.5	72.0	63.9	51.6	43.3	47.1	48.6	53.4	50.9
PPAT **(ours)**	62.1±1.5	68.8±1.2	**65.3**±1.0	54.0±1.9	50.2±1.4	**52.0**±0.3	56.8±1.8	56.0±1.1	**56.4**±0.3
PPAT **(ours)***	60.7±1.2	70.5±1.7	65.2±0.4	48.9±3.7	49.8±1.6	49.3±1.2	52.9±3.0	56.3±1.1	54.5±1.0

Table 1: Main result on EventStoryLine. The best results are in **bold**, * denotes model with Longformer encoders. SECI baselines listed in Section 4.3 cannot handle DECI task, and thus labeled as "-" in DECI and overall results.

Model	**MAVEN-ERE (SECI)**			**MAVEN-ERE (DECI)**			**MAVEN-ERE (Overall)**		
	P	R	F1	P	R	F1	P	R	F1
BERT	43.7	5.9	10.5	29.4	12.3	17.4	30.8	10.7	15.9
Longformer*	48.4	8.1	14.0	29.8	12.9	18.0	31.9	11.7	17.1
ERGO	42.9	40.9	41.9	27.9	41.7	33.5	30.5	41.5	35.2
ERGO*	40.7	58.1	47.9	28.2	40.8	33.6	31.3	45.1	37.0
PPAT **(ours)**	37.9±4.1	66.7±1.3	47.7±0.4	32.6±0.9	39.3±0.8	**35.6**±0.3	34.2±1.4	46.2±3.5	**39.2**±0.3
PPAT **(ours)***	42.2±1.6	58.4±3.2	**49.0**±1.4	31.2±0.8	41.3±2.1	35.5±0.6	34.1±0.9	45.5±2.0	39.0±0.2

Table 2: Main result on MAVEN-ERE. The best results are in **bold**, * denotes models that apply Longformer encoders.

proposes a self-supervised method to learn context-specific causal patterns from external causal statements.

ECI baseline We compare PPAT with the following ECI methods, which can handle both SECI and DECI: (1) **OP** [Caselli and Vossen, 2017] is a heuristic rule that assigns causal relations to neighboring events. (2) **LR+** and **LIP** [Gao *et al.*, 2019] are feature-based methods to construct document-level structures with various resources. (4) **BERT** [Devlin *et al.*, 2019] is a baseline that consists of the BERT encoder and a linear classifier. (5) **Longformer** [Beltagy *et al.*, 2020] is a baseline that consists of the Longformer encoder and a linear classifier. Due to the lack of reported results, we report the performance of our implementation. (6) **RichGCN** [Phu and Nguyen, 2021] proposes a document-level event interaction graph built with various NLP tools and heuristic rules, and uses a graph convolutional network (GCN) for transitivity. (7) **ERGO** [Chen *et al.*, 2022] proposes an event relational graph and a graph transformer for high-order event relational interaction. On EventStoryLine and Causal-TimeBank, ERGO achieves the current SOTA performance on both SECI and DECI.

4.4 Main Result

We report the main results on EventStoryLine, MAVEN-ERE and Causal-TimeBank in Table 1, 2 and 3 respectively. We break down the results on EventStoryLine and MAVEN-ERE into the SECI setting (i.e., intra-sentence event pairs) and DECI setting (i.e., inter-sentence event pairs). From the results, we have the following observations:

(1) As shown in Table 1, 2 and 3, our two versions of PPAT both outperform all baselines on three benchmarks in all settings. Compared with ERGO (Longformer-base), the previous SOTA method, PPAT (BERT-base) achieves the best F1 score on EventStoryLine (+1.4 on SECI, +4.9 on DECI and +5.5 on ECI) and MAVEN-ERE (+1.1 on SECI, +2.0 on DECI and +2.2 on ECI); PPAT (Longformer-base) achieves the best F1 score on Causal-TimeBank (+4.5 on SECI). The improvement demonstrates the effectiveness of PPAT.

(2) From Table 1 and 2, on the EventStoryLine and MAVEN-ERE, although PPAT (Longformer-base) has competitive SECI performance with PPAT (BERT-base), it performs worse than PPAT (BERT-base) on DECI. The reason might be: (i) PPAT has introduced document-level interaction via graph pairwise attention network, so the ability of Long-

Model	Causal-TimeBank (SECI)		
	P	R	F1
OP	3.0	40.7	5.5
KMMG	36.6	55.6	44.1
KnowDis	42.3	60.5	49.8
LSIN	51.5	56.2	52.9
LearnDA	41.9	68.0	51.9
CauSeRL	43.6	68.1	53.2
BERT	47.6	55.1	51.1
Longformer*	63.6	55.3	59.2
RichGCN	39.7	56.5	46.7
ERGO	58.4	60.5	59.4
ERGO*	62.1	61.3	61.7
PPAT **(ours)**	62.5±2.2	62.4±2.4	62.4±1.1
PPAT **(ours)***	67.9±1.7	64.6±0.3	**66.2**±0.7

Table 3: Main result on Causal-TimeBank. The best results are in **bold**, * denotes models that apply Longformer encoders. Note that Causal-TimeBank only supports SECI task.

Model	SECI	DECI	ECI
PPAT	65.3	52.0	56.4
w/o pairwise attention	64.8	49.1	54.3
w/o progressive reasoning	64.3	44.3	49.9
w/o causality-guided training	63.1	48.6	53.2

Table 4: F1-score of ablation study on EventStoryLine.

former to encode longer text does not show much advantages. (ii) The global attention pattern and simplified local attention in Longformer seem not competent for inter-sentence causality reasoning.

(3) On all datasets, PPAT (Longformer-base) achieves comparable or better performance than PPAT (BERT-base) on SECI. A possible reason is that Longformer can capture more abundant local context for SECI than BERT by extending token length limitation. Since the sentence-level event interaction can be introduced through the encoder, reasoning might be less important for SECI compared with DECI. Simply changing the encoder to a more expressive PLM could boost SECI performance. This also verifies the intuition that DECI is more complex to solve than SECI.

4.5 Ablation Study

We provide an ablation study of PPAT (BERT-base) on the EventStoryLine in Table 4 to analyse the effectiveness of components in PPAT.

(1) **PPAT (w/o pairwise attention)** reasons node embedding via the original attention method of Transformer [Vaswani *et al.*, 2017]. Compared with full version of PPAT, PPAT (w/o pairwise attention) has much poorer ability in identifying the inter-sentence event causality (-2.9 on DECI). It demonstrates that pairwise attention can effectively improve inter-sentence causality reasoning. The performance of SECI has a slight drop. A possible reason is that additional reasoning might be less important for SECI, since the sentence-level event interaction has been introduced via the encoder.

(2) **PPAT (w/o progressive reasoning)** reasons the intra- and inter-sentence event pairs together in the same time on each layer of event relational graph. Compared with removing other components, performance of PPAT (w/o progressive reasoning strategy) decreases the most on DECI and ECI (-7.7 on DECI and -6.5 on ECI). This shows that it would be better to predict inter-sentence causality based on well-reasoned intra-sentence causality representation than reasoning them together. In addition, the performance decrease of SECI verifies our hypothesis when building sentence boundary event relational graph: inter-sentence event relational information is unnecessary for intra-sentence causality reasoning.

(3) **PPAT (w/o causality-guided training)** is trained without causality guided loss on each layer. We see that causality-guided training strategy has significant improvement on both SECI and DECI, which proves that assisting model in learning causality-related representations is universally useful.

4.6 Case Study

In this section, we conduct a case study shown in Figure 3 to compare our PPAT (BERT-base) with current SOTA method, i.e, ERGO (Longformer-base). We also visualize the attention score of a relatively hard case, to explore the reasoning ability of our PPAT.

From the prediction table in Figure 3, we can observe that: although ERGO is good at identifying sentence-level causality (e.g., case No.1 and No.2), it has limitations in reasoning implicit inter-sentence causality. ERGO fails at identifying the case No.7's causality, which can be reasoned from No.1 and No.4 or from No.2 and No.5. ERGO also mistakenly takes coreference as causality (No.3).

In contrast, PPAT correctly identify the case No.7's causality via effective reasoning. As shown in the attention visualization in Figure 3, the predicted causality possibility increases from 0.41 to 0.76 after reasoning, indicating that: (i) PPAT infers inter-sentence event causality based on intra-sentence event causality as expected. (ii) PPAT infers with several transitivity patterns. Specifically, with the causality of "(*Death*, *shooting*)" and "(*Riots*, *Death*)", PPAT could reason that "(*Riots*, *shooting*)" has causal relation via *causality transitivity pattern*. Another reasoning pattern is *coreference transitivity pattern*: Previous work [Chen *et al.*, 2022] has shown PLMs could recognize coreference through similar word semantics, e.g., "(*Riots*, *protests*)". Together with the causality of "(*protests*, *shooting*)", PPAT can reason the causality of "(*Riots*, *shooting*)". In conclusion, the attention visualization indicates PPAT can perform highly effective reasoning with progressive reasoning and pairwise attention.

4.7 Representation Visualization

A good performance on ECI needs good causality representations for each event pair before classifying, so in Figure 4, to further explore the representation learning ability of PPAT, we choose the event pair causality representations from PPAT and then visualize them with t-SNE method [Van der Maaten and Hinton, 2008], observing that: (i) There is an obvious gap

Sunday, March 17, 2013 | 5 : 00 PM
Riots **Erupt** Following Death of Brooklyn Teen Killed By Police
In the week following the fatal **shooting** of 16-year-old Kimani Gray, several **protests** and **riots** have **erupted** in the …

No.	Event Pair		Golden	ERGO	PPAT
1	Riots	Death	Yes	Yes	Yes
2	Riots	Killed	Yes	Yes	Yes
3	Riots	protests	No	Yes	No
4	Death	shooting	Yes	Yes	Yes
5	Killed	shooting	Yes	Yes	Yes
6	protests	shooting	Yes	Yes	Yes
7	Riots	shooting	Yes	No	Yes
8	protests	Killed	Yes	No	Yes

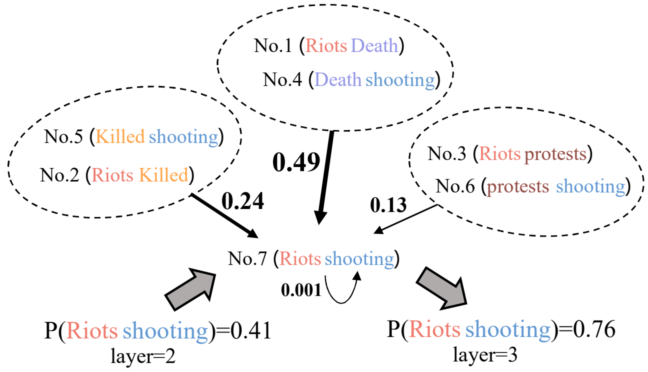
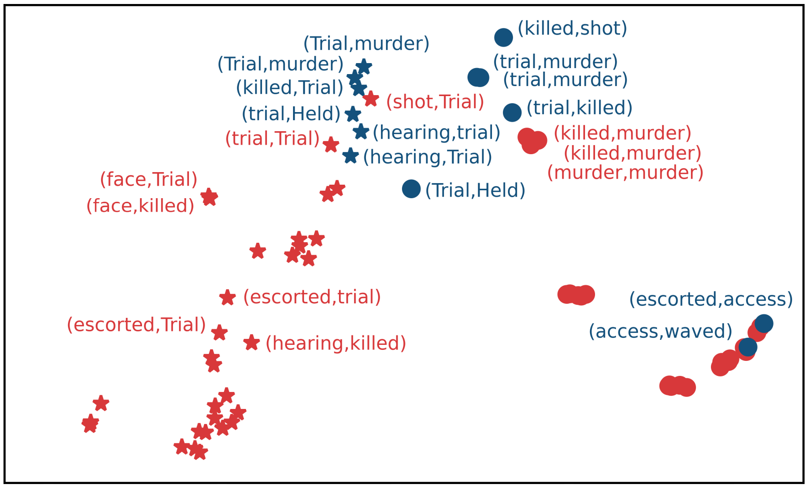


Figure 3: Case study of ERGO (Longformer-base) and our PPAT (BERT-base). The text above is the original document, where events are in **bold**. We focus on the five colored events and show the results of ERGO and PPAT in the table (left), where the correct predictions are in green and the wrong ones are in red. In the graph (right), the two event pairs in a circle denote a reasoning chain, and the graph shows the attention scores of various reasoning chains in the 3rd layer of PPAT when reasoning the No.7 case. The predicted causality possibility P of "(*Riots*, *shooting*)" increases after passing the 3rd layer.



Text

Accused Philadelphia Kraft Foods Killer **Held** For **Trial**
The woman who allegedly **shot** and **killed** two co-workers in September at the Kraft Foods plant in Northeast Philadelphia will **face trial** on multiple counts of **murder** and attempted **murder**. At a **hearing** on Tuesday, assistant district attorney Gail Fairman **described** 43-year-old Yvonne Hiller as a "methodical killer" who **waved** a gun within a foot of a security guard's face to regain access to the plant less than ten minutes after she had been **escorted** out ...

Figure 4: Visualization (left) of event pair representations and the original text (right). The blue nodes have causal relations and the red ones do not. The star-shaped and circle-shaped nodes denote inter- and intra-sentence event pairs respectively. Event mentions in text are in **bold**.

between inter-sentence event pairs (i.e., star-shaped nodes) and intra-sentence event pairs (i.e., circle-shaped nodes), which indicates that PPAT treats intra- and inter-sentence event pairs differently as expected. (ii) Most causal event pairs' representations are gathered together, which brings a lot convenience for later classification, showing the effectiveness of representation learning in the pairwise attention block. (iii) Pairs of semantically similar events (e.g., "killed" and "murder") are close to the causal node cluster, as they might be helpful for reasoning and need to interact with causal event pairs. Meanwhile, the event pairs that cannot help reasoning (e.g. "hearing" and "killed") are far away from causal event pairs. This shows PPAT can utilize available relational information for learning good representations.

5 Conclusion

In this paper, we propose a Progressive Graph Pairwise Attention Network (PPAT), which leverages pairwise attention to capture reasoning chains on the sentence boundary event relational graph. PPAT infers progressively, as it uses SECI results to help reason implicit document-level event causality. Our PPAT achieves SOTA performance on three widely-used ECI datasets with significant improvements. We conduct extensive ablation experiments, case studies and representation visualization to analyse PPAT's effectiveness. Future work may include extending PPAT to identification of other event relations, especially implicit relations in need of reasoning.

Acknowledgments

This work is jointly supported by grants: Natural Science Foundation of China (No. 62006061), Stable Support Program for Higher Education Institutions of Shenzhen (No.GXWD20201230155427003-20200824155011001) and Strategic Emerging Industry Development Special Funds of Shenzhen(No.JCYJ20200109113441941).

References

[Beltagy *et al.*, 2020] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020.

[Berant *et al.*, 2014] Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Cao *et al.*, 2021] Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4862–4872, Online, August 2021. Association for Computational Linguistics.

[Caselli and Vossen, 2017] Tommaso Caselli and Piek Vossen. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada, August 2017. Association for Computational Linguistics.

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

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Do *et al.*, 2011] Quang Do, Yee Seng Chan, and Dan Roth. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[Gao *et al.*, 2019] Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Hashimoto, 2019] Chikara Hashimoto. Weakly supervised multilingual causality extraction from wikipedia. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2988–2999, 2019.

[Hidey and McKeown, 2016] Christopher Hidey and Kathy McKeown. 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, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Hu *et al.*, 2017] Zhichao Hu, Elahe Rahimtoroghi, and Marilyn Walker. Inference of fine-grained event causality from blogs and films. In *Proceedings of the Events and Stories in the News Workshop*, pages 52–58, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[Kadowaki *et al.*, 2019] Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5816–5822, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[Li *et al.*, 2021] Sha Li, Heng Ji, and Jiawei Han. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online, June 2021. Association for Computational Linguistics.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[Liu *et al.*, 2020] Jian Liu, Yubo Chen, and Jun Zhao. Knowledge enhanced event causality identification with mention masking generalizations. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International*

Joint Conference on Artificial Intelligence, IJCAI-20, pages 3608–3614. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[Mirza, 2014] Paramita Mirza. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[Ning *et al.*, 2018] Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Oh *et al.*, 2016] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. A semi-supervised learning approach to why-question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[Phu and Nguyen, 2021] Minh Tran Phu and Thien Huu Nguyen. 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, 2021.

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

[Riaz and Girju, 2014b] Mehwish Riaz and Roxana Girju. Recognizing causality in verb-noun pairs via noun and verb semantics. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 48–57, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Wang *et al.*, 2022] Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *CoRR*, abs/2211.07342, 2022.

[Yao *et al.*, 2019] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy, July 2019. Association for Computational Linguistics.

[Zhong and Chen, 2021] Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June 2021. Association for Computational Linguistics.

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

[Zuo *et al.*, 2021a] Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, 2021.

[Zuo *et al.*, 2021b] Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. LearnDA: Learnable knowledge-guided data augmentation for event causality identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3558–3571, Online, August 2021. Association for Computational Linguistics.