PPAT: Progressive Graph Pairwise Attention Network for Event Causality Identification

Anonymous authors
Paper under double-blind review

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 help identify the implicit event causality. However, they ignore 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 causality, and then infers the more implicit inter-sentence causality based on the SECI result. We construct a sentence boundary event relational graph, and PPAT leverages a novel pairwise attention, 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 on two well-established benchmark datasets show that our model achieves state-of-the-art performance (5.5% F1 gains on EventStoryLine and 4.5% F1 gains on Causal-TimeBank).

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 causality 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).

ECI is composed of two parts: sentence-level ECI (SECI) (Liu et al., 2020; Zuo et al., 2021a) 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. One of the great challenges of ECI is how to identify the implicit causal relations, which are always expressed with multiple sentences and without clear causal cues (Cao et al., 2021). To address this problem, recent studies construct event graphs and apply graph neural networks as the reasoning module to infer the implicit event causality (Tran Phu & Nguyen, 2021; Chen et al., 2022). For example, ERGO (Chen et al., 2022), the recent state-of-the-art (SOTA) method, uses a graph transformer to enable event interaction, and reasons on an event relational graph (ERG), where each node represents an event pair and contains its relational information. There is an edge between two nodes only if they share one event.

Although recently proposed reasoning models have achieved some success in ECI, their reasoning process is unnatural, as they reason intra- and inter-sentence event causality in the same time. We observe that most intra-sentence event causality is easy to identify with explicit causality cues, while inter-sentence event causality is more implicit and needs to be inferred from intra-sentence event causality. Take Figure 1 as an example, the causality of intra-sentence event pair "(*quake*, *injured*)" could be identified easily with the causality indicator "*left*". Based on the 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 addition to the unnatural reasoning process, the graph neural networks used as the reasoning model also need improvement. For example, ERGO (Chen et al., 2022) simply aggregates the representations of neighborhood nodes (i.e., event pairs), ignoring the reasoning chains among these

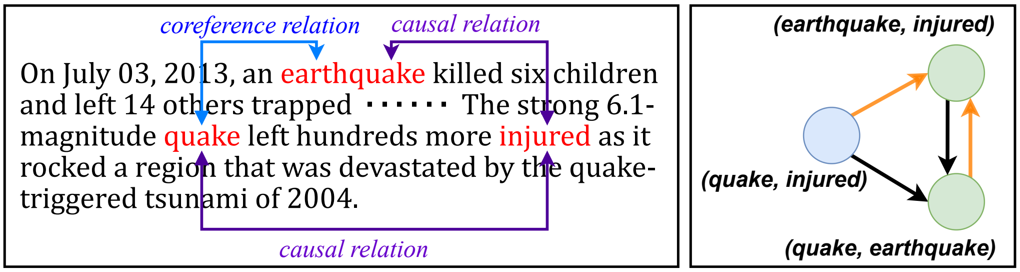

coreference relation
causal relation
On July 03, 2013, an earthquake killed six children and left 14 others trapped ······ The strong 6.1-magnitude quake left hundreds more injured as it rocked a region that was devastated by the quake-triggered tsunami of 2004.
causal relation
(earthquake, injured)
(quake, injured)
(quake, earthquake)

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.

neighbors. In Figure 1, 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: $Cause(quake, injured) \wedge Coreference(earthquake, quake) \rightarrow Cause(earthquake, injured)$. The reasoning model should regard the premise node pair as a whole part to aggregate neighbors at a reasoning chain level.

To address the two problems mentioned above, we propose a novel Progressive Graph Pairwise Attention Network (PPAT) for reasoning event causality on the Sentence boundary Event Relational Graph (SERG). Same as ERG, each node of SERG denotes an event pair, and 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. Two basic ideas of SERG are: (i) only two nodes that share an event can have direct influence on each other. (ii) intra-sentence causality reasoning does not depend on inter-sentence nodes.

PPAT provides effective global reasoning upon three aspects: (1) **Progressive reasoning strategy**: PPAT reasons progressively on SERG, as it first predicts the intra-sentence causality, and then reasons the inter-sentence causality based on the previous SECI prediction. The progressive reasoning strategy takes the dependence of inter-sentence causality on intra-sentence causality into consideration. (2) **Pairwise attention**: PPAT applies a novel graph pairwise attention network, which aggregates neighbors at a reasoning chain level instead of node level. A reasoning chain corresponds to a premise node pair. Pairwise attention can introduce interaction between the target node and its premise node pairs, thus attending to the possible reasoning chains and inferring the target causality. (3) **Causality-guided training strategy**: Since node representations on each 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. We are the first to capture the dependence of inter-sentence causal reasoning on intra-sentence causality.
- We propose a pairwise attention mechanism, which is a simple but effective approach to attend to reasoning chains on the graph for causality propagation.
- Extensive experiments on two benchmark datasets show that PPAT significantly outperforms previous SOTA methods. The results show the effectiveness of our proposed method.

2 Related Work

Early feature-based methods for SECI mainly focus on designing better causality features or using external resources to improve the performance, including the lexicon of causality indicators (Mirza,

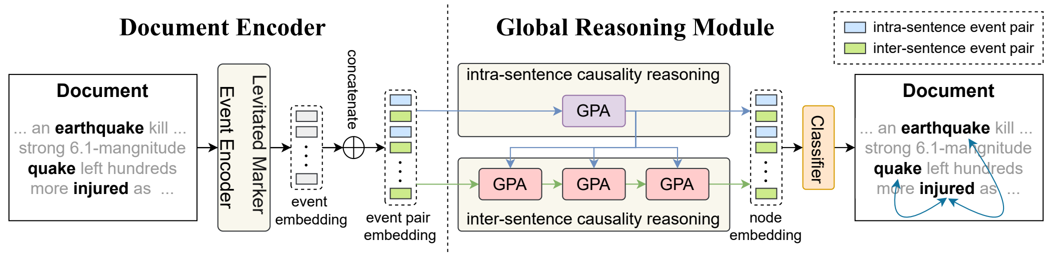

Document Encoder
Global Reasoning Module
intra-sentence event pair
inter-sentence event pair
Document
... an earthquake kill ... strong 6.1-mangnitude quake left hundreds more injured as ...
Levitated Marker Event Encoder
event embedding
concatenate
event pair embedding
intra-sentence causality reasoning
GPA
GPA
GPA
GPA
inter-sentence causality reasoning
node embedding
Classifier
Document
... an earthquake kill ... strong 6.1-mangnitude quake left hundreds more injured as ...

Figure 2: Overview of our proposed PPAT. The Document Encoder gives initial event pair representations, and the Global Reasoning Module updates these representations through graph pairwise attention network (GPA). Finally, the causality links are inferred from node classification.

2014; Hidey & McKeown, 2016), temporal patterns (Mirza, 2014; Ning et al., 2018), event semantics (Riaz & Girju, 2014a;b), 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; Tran Phu & Nguyen (2021) leverage several NLP tools (e.g., dependency parser) and external corpus for building event graphs, and then use graph convolutional network (Kipf & Welling, 2017) for reasoning. ERGO (Chen et al., 2022) achieves 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, no need for sophisticated graph design, external NLP tools or external knowledge.

3 Methods

The overview of our model is shown in Figure 2, which is composed of two modules: (1) **Document Encoder** encodes event mentions of a given document with levitated markers, and outputs initial event pairs representations; (2) **Global Reasoning Module** iteratively updates representations of intra- and inter-sentence event pairs, and then predicts causality based on the learned representations.

3.1 Document Encoder

Given a document $\mathcal{D} = \{w_0, w_1, \cdots, w_{L_\mathcal{D}}\}$ (can be of any length of $L_\mathcal{D}$) 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 & 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:

$$S = \texttt{[CLS]}, w_0, w_1 \cdots event_i \cdots w_{L_\mathcal{D}} \cdots t_1^i, t_2^i \cdots$$

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 $event_i$, $L_\mathcal{D}$ 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 concatenate operation, $H(*)$ 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 Global Reasoning Module

3.2.1 Progressive reasoning strategy

To infer the high-order relation of event pairs in the document $\mathcal{D}$, one of the key points in PPAT is the progressive reasoning strategy. We build a sentence boundary relational event graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Each node in $\mathcal{V}$ represents a unique event pair. Each node has a directed edge to another node when their corresponding event pairs share one event. Moreover, intra-sentence nodes only have directed edges to the intra-sentence nodes in the same sentence. With the initial node representations from the document encoder as input, PPAT first reasons sentence-level causality with a 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 3.2.2.

Based on the node representations, we use a linear classifier to predict the causality of nodes in each layer. The predicted causality possibility of $(event_i, event_j)$ in the l-th layer, denoted as p_{ij}^l, is calculated as follows:

$$p_{ij}^l = \text{softmax}(v_{ij}^l \mathbf{W}_c) \tag{3}$$

where $\mathbf{W}_c$ is the parameter weight matrix in the linear classifier.

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:

$$n_{ij}^{l+1} = v_{ij}^l \oplus p_{ij}^l \oplus a_{ij} \tag{4}$$

where 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. Before the first step of reasoning (i.e., $l = 0$), the node embedding is initialized by:

$$v_{ij}^0 = r_{ij} \tag{5}$$

where r_{ij} is the raw event pair representation from the document encoder.

3.2.2 Graph pairwise attention

Another key point of PPAT is the pairwise attention which introduces reasoning chain-level attention. When $(event_i, event_j)$ is target node to be reasoned, its premise node pairs are defined as $((event_i, event_k), (event_j, event_k))$, where $0 \leq k < N, k \neq i \neq j$. Then we perform a pairwise self-attention mechanism to measure the importance of each premise node pair for the target node:

$$\text{atten}_{ij,k} = \frac{(n_{ij}\mathbf{W}_q)((n_{ik} \oplus n_{jk})\mathbf{W}_k)^T}{\sqrt{d}}, \tag{6}$$

where n_{ij} is the input node embedding of $(event_i, event_j)$ described in Section 3.1, $\mathbf{W}_q, \mathbf{W}_k$ are parameter weight matrices, $\sqrt{d}$ is a scaling factor and d is the hidden size (Vaswani et al., 2017).

Then we normalize the attention coefficients with softmax function:

$$\alpha_{ij,k} = \text{softmax}_{ij}(\text{atten}_{ij}) = \frac{mask_{ij,k} \exp(\text{atten}_{ij,k})}{\sum_{z \in \mathcal{N}_{ij}^-} mask_{ij,z} \exp(\text{atten}_{ij,z})}, \tag{7}$$

where $\mathcal{N}_{ij}^{-}$ is the event mention set that does not contain $event_i$ 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:

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

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:

$$v_{ij}^{l} = \Big(\Big\|_{c=1}^{C} \sum_{k \in \mathcal{N}_{ij}^{-}} \alpha_{ij,k}((n_{ik} \oplus n_{jk})\mathbf{W}_v) \Big) \mathbf{W}_o, \tag{9}$$

where $\|$ and $\oplus$ are both concatenation operation, C is the number of heads, $\mathbf{W}_o$ is the parameter weight matrix. In Appendix A.2, we show a fast pairwise attention algorithm in our implementation.

3.3 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 (i.e., most of the event pairs have no causal relations):

$$\mathrm{FL}(p, y) = -\beta(y((1-p)^{\gamma} \log(p)) + (1-y)(p^{\gamma} \log(1-p))) \tag{10}$$

where p is the predicted possibility and y is the golden 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}):

$$\mathcal{L}_m = \sum_{(i,j) \in \mathcal{M}} \mathrm{FL}(p_{ij}^{L}, y_{ij}) \tag{11}$$

where $\mathcal{M}$ is the event pair set, L is the number of layers, y_{ij} is the ground truth label.

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 3 and calculate the focal loss as follows:

$$\mathcal{L}_c = \sum_{0 \leq l \leq L-1} (\lambda^{l} \sum_{(i,j) \in \mathcal{M}^{l}} \mathrm{FL}(p_{ij}^{l}, y_{ij})), \tag{12}$$

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). L is the number of layers in PPAT. PPAT's final loss is given by:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_c \tag{13}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our PPAT on two benchmark datasets: EventStoryLine (version 0.9) (Caselli & Vossen, 2017) and Cauasl-TimeBank (Mirza, 2014).

EventStoryLine contains 258 documents in 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.

Causal-TimeBank contains 183 documents, 6811 events, 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; Tran Phu & 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. We optimize our model with AdamW 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 sentence-level dynamic window to encode the entire document. The window length is 5 sentences for BERT and 7 sentences for Longformer, and the shift step is 2 sentences, so we set the max length of input tokens at 320 for BERT and 512 for Longformer. We train the model with 128 epochs for EventStoryLine and 32 for Causal-TimeBank. 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. Training on EventStoryLine and Causal-TimeBank takes 2 hours and 1.5 hours respectively.

4.3 Baselines

4.3.1 SECI baseline

We compare our 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) proposes a knowledge-enhanced data augmentation method to tackle data lacking problem. (3) **LSIN** (Cao et al., 2021) proposes 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) proposes a self-supervised method to learn context-specific causal patterns from external causal statements.

4.3.2 ECI baseline

We compare our PPAT with the following ECI methods, which can handle both SECI and DECI: (1) **OP** (Caselli & 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. (3) **RichGCN** (Tran Phu & Nguyen, 2021) proposes a document-level event interaction graph built with various NLP tools and heuristic rules, and uses a graph convolutional network (GCN) to capture relevant connections. (4) **ERGO** (Chen et al., 2022) proposes event relational graph and 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

In Table 1, we report the overall results on EventStoryLine and Causal-TimeBank. In Table 2, We break down the results on EventStoryLine 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) From Table 1 and Table 2, our two versions of PPAT both outperform all baselines on two benchmarks in SECI, DECI and overall ECI. Compared with ERGO (Longformer-base), the previous SOTA method, PPAT (BERT-base) achieves the best F1 score (+1.4 on SECI, +4.9 on DECI and +5.5 on ECI) on EventStoryLine; PPAT (Longformer-base) achieves the best F1 score (+4.5 on SECI) on Causal-TimeBank. The improvement demonstrates the effectiveness of PPAT.

(2) From Table 2, on the EventStoryLine, although PPAT (Longformer-base) has competitive SECI performance with PPAT (BERT-base), it performs worse than PPAT (BERT-base) on DECI. The

Table 1: Main result on EventStoryLine and Causal-TimeBank. The best results are in **bold**, † denotes models that apply Longformer encoders. Causal-TimeBank only supports SECI task, in which setting LR+ and LIP cannot solve. EventStoryLine contains inter-sentence event pairs with causal relations (i.e., DECI task), which SECI baselines in Section 4.3.1 cannot handle.

Model	**EventStoryLine (SECI+DECI)**			**Causal-TimeBank (SECI)**		
	P	R	F1	P	R	F1
OP (Caselli & Vossen, 2017)	10.5	99.2	19.0	3.0	40.7	5.5
LR+ (Gao et al., 2019)	27.9	47.2	35.1	-	-	-
LIP (Gao et al., 2019)	36.2	49.5	41.9	-	-	-
KMMG (Liu et al., 2020)	-	-	-	36.6	55.6	44.1
KnowDis (Zuo et al., 2020)	-	-	-	42.3	60.5	49.8
LSIN (Cao et al., 2021)	-	-	-	51.5	56.2	52.9
LearnDA (Zuo et al., 2021b)	-	-	-	41.9	68.0	51.9
CauSeRL (Zuo et al., 2021a)	-	-	-	43.6	68.1	53.2
RichGCN (Tran Phu & Nguyen, 2021)	42.6	51.3	46.6	39.7	56.5	46.7
ERGO (Chen et al., 2022)	46.3	50.1	48.1	58.4	60.5	59.4
ERGO (Chen et al., 2022)†	48.6	53.4	50.9	62.1	61.3	61.7
PPAT **(ours)**	56.8±1.8	56.0±1.1	**56.4**±0.3	62.5±2.2	62.4±2.4	62.4±1.1
PPAT **(ours)**†	52.9±3.0	56.3±1.1	54.5±1.0	67.9±1.7	64.6±0.3	**66.2**±0.7

Table 2: SECI and DECI on EventStoryLine. The best results are in **bold**, † denotes model with Longformer encoders. SECI baselines listed in Section 4.3.1 cannot handle DECI task.

Model	**EventStoryLine (SECI)**			**EventStoryLine (DECI)**		
	P	R	F1	P	R	F1
OP (Caselli & Vossen, 2017)	10.5	99.2	19.0	3.0	40.7	5.5
LR+ (Gao et al., 2019)	22.5	98.6	36.6	8.4	99.5	15.6
LIP (Gao et al., 2019)	38.8	52.4	44.6	35.1	48.2	40.6
KMMG (Liu et al., 2020)	41.9	62.5	50.1	-	-	-
KnowDis (Zuo et al., 2020)	39.7	66.5	49.7	-	-	-
LSIN (Cao et al., 2021)	47.9	58.1	52.5	-	-	-
LearnDA (Zuo et al., 2021b)	42.2	69.8	52.6	-	-	-
CauSeRL (Zuo et al., 2021a)	41.9	69.0	52.1	-	-	-
RichGCN (Tran Phu & Nguyen, 2021)	49.2	63.0	55.2	39.2	45.7	42.2
ERGO (Chen et al., 2022)	49.7	72.6	59.0	43.2	48.8	45.8
ERGO (Chen et al., 2022)†	57.5	72.0	63.9	51.6	43.3	47.1
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
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

reason might be: (i) PPAT has introduced document-level interaction via graph pairwise attention network, so the ability of Longformer to encode longer text does not show much advantage. (ii) The global attention pattern in Longformer could be ineffective in inter-sentence causality reasoning.

(3) From Table 1, on the Causal-TimeBank, PPAT (Longformer-base) achieves better performance than PPAT (BERT-base). A possible reason is that the performance in SECI-only setting mainly depends on the encoder, and Longformer has been found to outperform BERT in various NLP tasks. Since the sentence-level event interaction can be introduced through the encoder, reasoning might be unnecessary for SECI, and 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 3 to analyse the effectiveness of components in PPAT.

(1) **PPAT (w/o pairwise attention)** infers 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

Table 3: F1-score of ablation study on EventStoryLine.

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

shows that pairwise attention can improve the inter-sentence event causality reasoning. The performance of SECI does not decrease much. A possible reason is that additional reasoning might be unnecessary for SECI, since the sentence-level event interaction has been introduced via encoder.

(2) **PPAT (w/o progressive reasoning)** reasons the inter- and intra-sentence event pairs 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, which shows that progressive reasoning strategy is beneficial to document-level causality reasoning. This also 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

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 teen's Brooklyn neighborhood as …

No.	**Event Pair**		**Golden**	**ERGO**	**PGCAT**
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

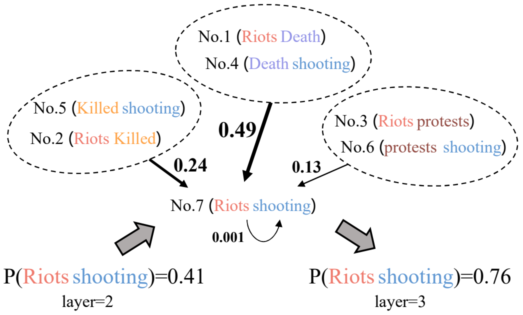
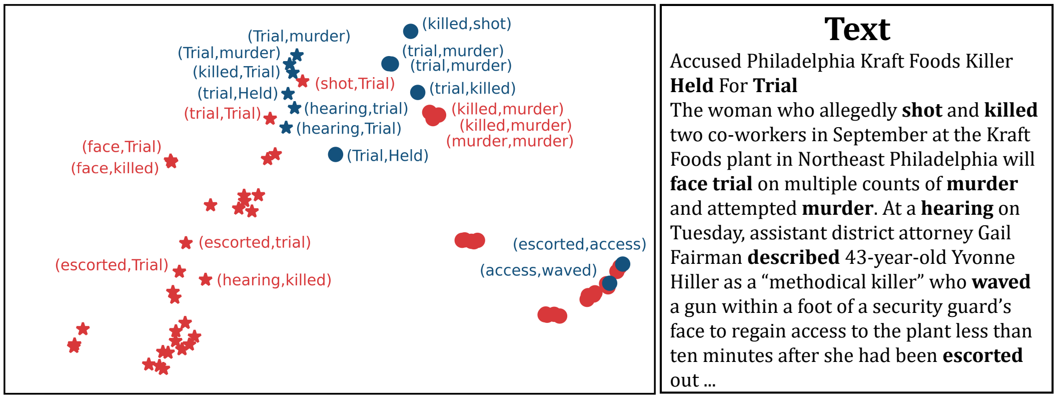


Figure 3: Case study of ERGO (Longformer-base) and our PPAT (BERT-base). The text above is the original document, where event are in **bold**. We focus on the five colored events and show the results of ERGO and PPAT in the table (left), where the correct predictions are in green and the wrong ones are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

In this section, we conduct a case study shown in Figure 3 to compare between 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 wrongly takes coreference as causality (No.3).



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 nodes are inter-sentence event pairs and the circle-shaped nodes are intra-sentence event pairs. Event mentions in text are in **bold**.

In contrast, PPAT can identify the case No.7’s causality via effective global reasoning module. In Figure 3, we visualize the importance of each premise node pair for reasoning the causality of No.7. After reasoning, the predicted causality possibility increases from 0.41 to 0.76, which shows that: (i) PPAT infers inter-sentence event causality based on intra-sentence event causality as expected. (ii) PPAT can infer 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 possible reasoning chain is *coreference transitivity pattern*. Previous work (Chen et al., 2022) has shown PLMs could identify 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 in Figure 3 shows PPAT can perform 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, we present the visualization of event pair representations to further explore the representation learning ability of PPAT. We choose the output event pair representations of the global reasoning module and then visualize them with t-SNE method (Van der Maaten & Hinton, 2008). More visualization can be found in Appendix A.1 about the change of event pair representations after passing each layer. From Figure 4, we observe that: (i) There is an obvious gap between inter-sentence event pairs (i.e., star-shaped nodes) and intra-sentence event pairs (i.e., circle-shaped nodes), which shows that PPAT treats intra- and inter-sentence event pairs differently as expected. (ii) Most causal event pairs’ representations are gathered at the top of the figure, showing the effectiveness of representation learning in the global reasoning module. (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. The event pairs that cannot help reasoning (e.g. “hearing” and “killed”) are far away from causal event pairs. This shows our reasoning module can utilize available relational information for learning good event pair causality 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 causality. Our PPAT achieves SOTA performance on the two widely-used benchmarks. We conduct extensive experiments and case studies to analyse PPAT’s effectiveness. Future work may include extending PPAT to identification of other event relations, especially the implicit relations in need of reasoning.

REFERENCES

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL https://arxiv.org/abs/2004.05150.

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)*, pp. 1499–1510, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1159. URL https://aclanthology.org/D14-1159.

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)*, pp. 4862–4872, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.376. URL https://aclanthology.org/2021.acl-long.376.

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*, pp. 77–86, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2711. URL https://aclanthology.org/W17-2711.

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. *arXiv preprint arXiv:2204.07434*, 2022.

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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

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*, pp. 294–303, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1027.

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)*, pp. 1808–1817, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1179. URL https://aclanthology.org/N19-1179.

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)*, pp. 2988–2999, 2019.

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)*, pp. 1424–1433, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1135. URL https://aclanthology.org/P16-1135.

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*, pp. 52–58, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2708. URL https://aclanthology.org/W17-2708.

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)*, pp. 5816–5822, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1590. URL `https://aclanthology.org/D19-1590`.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=SJU4ayYgl`.

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*, pp. 894–908, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.69. URL `https://aclanthology.org/2021.naacl-main.69`.

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)*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Jian Liu, Yubo Chen, and Jun Zhao. Knowledge enhanced event causality identification with mention masking generalizations. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3608–3614. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/499. URL `https://doi.org/10.24963/ijcai.2020/499`. Main track.

Paramita Mirza. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pp. 10–17, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-3002. URL `https://aclanthology.org/P14-3002`.

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)*, pp. 2278–2288, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1212. URL `https://aclanthology.org/P18-1212`.

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. doi: 10.1609/aaai.v30i1.10388. URL `https://ojs.aaai.org/index.php/AAAI/article/view/10388`.

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)*, pp. 161–170, Philadelphia, PA, U.S.A., June 2014a. Association for Computational Linguistics. doi: 10.3115/v1/W14-4322. URL `https://aclanthology.org/W14-4322`.

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)*, pp. 48–57, Gothenburg, Sweden, April 2014b. Association for Computational Linguistics. doi: 10.3115/v1/W14-0707. URL `https://aclanthology.org/W14-0707`.

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*, pp. 3480–3490, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.273. URL `https://aclanthology.org/2021.naacl-main.273`.

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

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 (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

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*, pp. 764–777, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1074. URL `https://aclanthology.org/P19-1074`.

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*, pp. 50–61, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.5. URL `https://aclanthology.org/2021.naacl-main.5`.

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*, pp. 1544–1550, 2020.

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*, pp. 2162–2172, 2021a.

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)*, pp. 3558–3571, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.276. URL `https://aclanthology.org/2021.acl-long.276`.

A Appendix

A.1 More Representation visualization

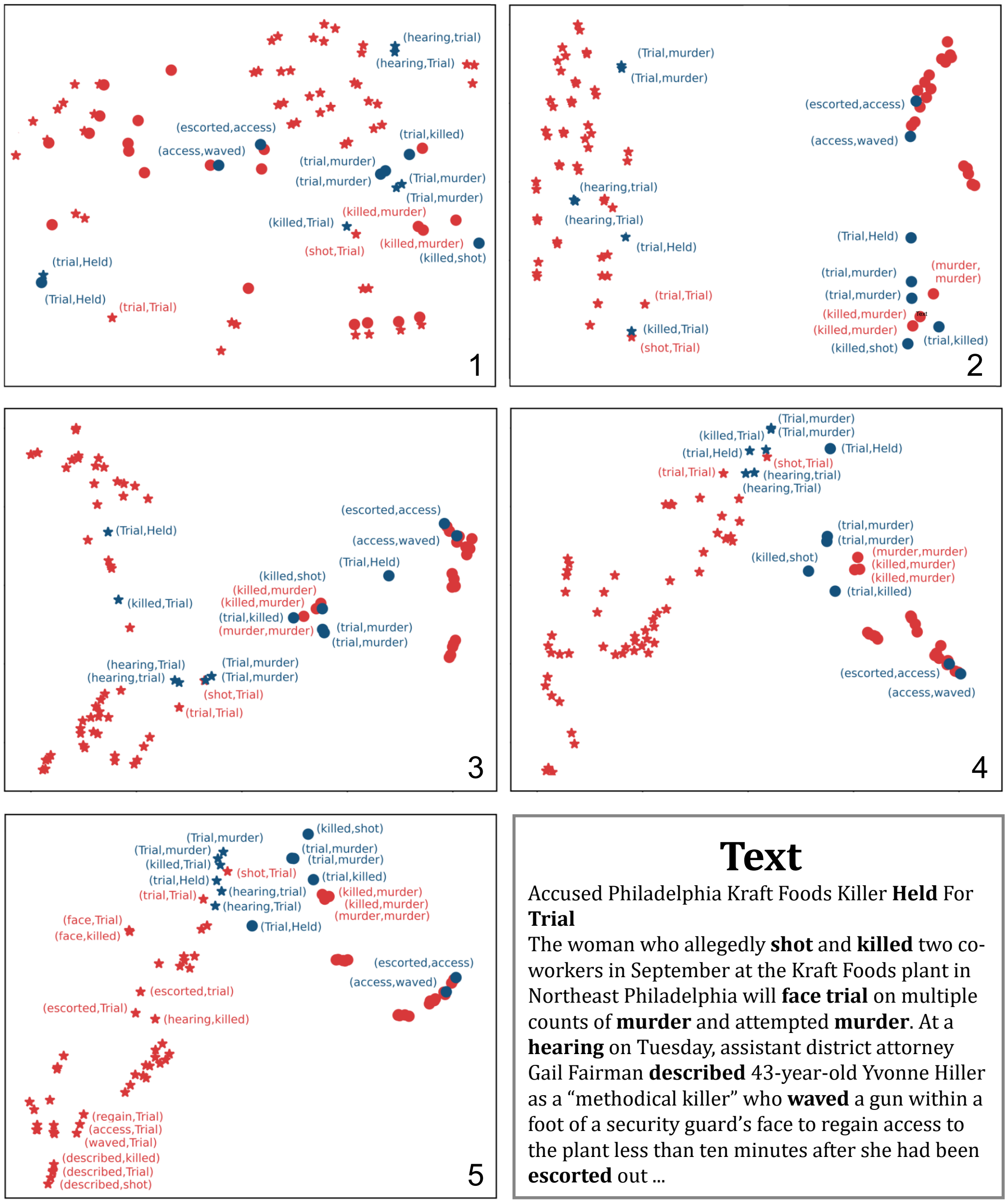
To further explore the reasoning process of graph pairwise attention, we present the node representation visualization after each layer in Figure 5. The first subfigure is the initial node representations from the document encoder, and the rest of subfigures are the node representation after one layer of intra-sentence reasoning and three layers of inter-sentence reasoning. From the subfigures, we can observe that:

(1) In the second subfigure (i.e., after intra-sentence reasoning), intra- and inter-sentence nodes are divided into two clusters, and most intra-sentence causal nodes are gathered together. Inter-sentence causal nodes are scattered since their representations are not updated yet.

(2) During the inter-sentence reasoning in the last three subfigures, intra- and inter-sentence nodes are still clustered into two groups, but the causal nodes are getting closer to each other. This shows that the reasoning model has the ability to learn causality representation.

A.2 Fast Pairwise Attention Algorithm

For different nodes as the target node, their premise neighborhood nodes are different. The original pairwise attention algorithm in Section 3.2.2 is slow in computation, because the concatenation



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 5: Nodes representation visualization after updating in each layer. The blue nodes have causal relations and the red ones do not. The star-shaped nodes are inter-sentence event pairs and the circle-shaped nodes are intra-sentence event pairs. Event mentions in text are in **bold**.

operations in Equation 6 and Equation 8 are difficult to compute parallelly, so we propose a mathematically equivalent algorithm (see Algorithm 1) to avoid the concatenation operation and compute the pairwise attention faster.

Each row of attention score matrix (i.e. S^1 and S^2) denotes a target node and each column denotes one node of a premise node pair. The goal of the function $PremiseNodeSwitch(*)$ is to switch the elements in the same premise node pair in attention score matrix.

Algorithm 1: Fast pairwise attention computation

```
Data: Adjacency matrix of SERG, $A$; Initial node embedding, $E_i$; Two attention key matrix,
      $\mathbf{W}_k^1$ and $\mathbf{W}_k^2$; Two attention value matrix, $\mathbf{W}_v^1$ and $\mathbf{W}_v^2$; Two attention query matrix,
      $\mathbf{W}_q^1$ and $\mathbf{W}_q^2$; Output matrix, $\mathbf{W}_o$
Result: Node embedding output by pairwise attention network, $E_o$
$S^1 = \frac{(E_i \mathbf{W}_q^1)(E_i \mathbf{W}_k^1)^T}{\sqrt{d}}$ ;
$S^2 = \frac{(E_i \mathbf{W}_q^2)(E_i \mathbf{W}_k^2)^T}{\sqrt{d}}$ ;
$\hat{S}^1 = S^1 + PremiseNodeSwitch(S^2)$ ;
$\hat{S}^2 = S^2 + PremiseNodeSwitch(S^1)$ ;
$M^1 = ColumnSoftmax(\hat{S}^1 \times A)$ ;
$M^2 = ColumnSoftmax(\hat{S}^2 \times A)$ ;
$V^1 = \mathbf{W}_v^1 E_i$ ;
$V^2 = \mathbf{W}_v^2 E_i$ ;
$E_o = \mathbf{W}_o((M^1 V^1) + (M^2 V^2))$ ;
return $E_o$ ;
```