ERGO: Event Relational Graph Transformer for Document-level Event Causality Identification

Meiqi Chen[1], **Yixin Cao**[2], **Kunquan Deng**[3],
Mukai Li[4], **Kun Wang**[4], **Jing Shao**[4], **Yan Zhang**[1]*
[1] Peking University [2] Singapore Management University
[3] Beihang University [4] SenseTime Research
meiqichen@stu.pku.edu.cn

Abstract

Document-level Event Causality Identification (DECI) aims to identify event-event causal relations in a document. Existing works usually build an event graph for global reasoning across multiple sentences. However, the edges between events have to be carefully designed through heuristic rules or external tools. In this paper, we propose a novel **E**vent **R**elational **G**raph Transf**O**rmer (ERGO) framework[1] for DECI, to ease the graph construction and improve it over the noisy edge issue. Different from conventional event graphs, we define a pair of events as a node and build a complete event relational graph without any prior knowledge or tools. This naturally formulates DECI as a node classification problem, and thus we capture the causation transitivity among event pairs via a graph transformer. Furthermore, we design a criss-cross constraint and an adaptive focal loss for the imbalanced classification, to alleviate the issues of false positives and false negatives. Extensive experiments on two benchmark datasets show that ERGO greatly outperforms previous state-of-the-art (SOTA) methods (12.8% F1 gains on average).

1 Introduction

Event Causality Identification (ECI) is the task of identifying if the occurrence of one event causes another in text. As shown in Figure 1, given the text "*... the outage$_2$ was caused by a terrestrial break in the fiber in Egypt ...*", where "*outage$_2$*" and "*break*" are event triggers, an ECI model should predict if they have a causal relation or not. Discovering causal relationships not only helps to deeply understand how the world progresses, but also is an important goal of empirical research in various areas, such as machine reading comprehension (Berant et al., 2014), question answering (Oh et al., 2016), future event forecasting (Hashimoto, 2019),

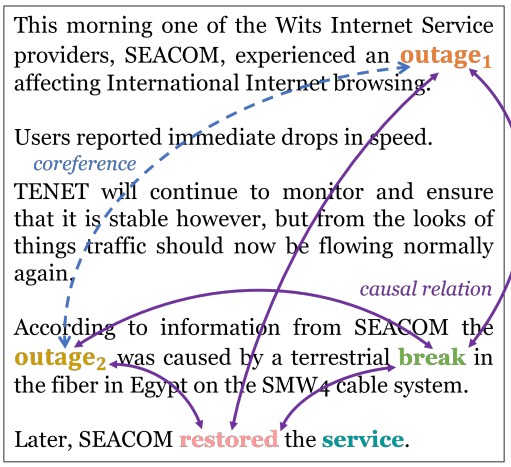

Figure 1: Example of DECI. Solid purple lines denote target causal relations.

and event knowledge graph construction (Ma et al., 2022b).

Causality is usually implicit in natural language (Copley and Martin, 2014), especially when events scatter in a document, a.k.a. Document-level ECI (DECI). Recent methods typically construct an event graph to assist the global inference across multiple sentences, where nodes are events and edges are their relations, such as linguistic dependency or adjacent contexts (Gao et al., 2019; Zhao et al., 2021). However, there are two major issues. First, the edges heavily rely on external tools and heuristic rules, which are not always reliable and may introduce noise (Tran Phu and Nguyen, 2021). Second, it is like a chicken-egg problem — to identify (causal) relations between events, you need to extract their relations to build the graph first.

In this paper, we propose a novel **E**vent **R**elational **G**raph Transf**O**rmer (ERGO) framework for DECI, which doesn't require any external tools and can effectively alleviate the noise issue. Different from conventional event graphs, the basic idea is to build an event relational graph that naturally converts ECI into a node classification problem, where each node denotes a pair of events,

*Corresponding author.
[1]https://github.com/chenmeiqii/ERGO.git

Proceedings of the 29th International Conference on Computational Linguistics, pages 2118–2128
October 12–17, 2022.

and all edges among nodes are initialized to capture potential causal chains, following the assumption "preserving transitivity of causation" (Paul et al., 2013). If event A causes event B and event B causes event C, then we have event A causes event C. That is, if the node of (A, C) receives positive predictions (i.e., causal relation) from nodes (A, B) and (B, C), it is positive, too. By contrast, if either/both of nodes (A, B) and (B, C) are negative (i.e., no relation), it is not necessary that (A, C) is negative, either. To this end, we leverage a graph transformer to model the graph and assign a lower weight to such uninformative edges, paying more attention to other paths or its own textual contexts.

Although the proposed graph directly models all event pairs for causality identification, it poses a great challenge of false positive and false negative issues. First, most of the event pairs have no causal relations. That is, negative nodes are dominant, and the imbalanced classification will easily confuse the model into false-negative predictions. Second, to ease the graph construction, we assume that all nodes can pass information with each other via the complete graph structure. While there are many spurious correlations between events, which can be incorrectly propagated to neighbor nodes, leading to severe false positives. For example, “treatment” and “death” frequently co-occur in the same document, but there is no causality between them since it is not “treatment” that causes “death”.

To address the above issues, we further design a *criss-cross constraint* and an adaptive focal loss. The criss-cross constraint simplifies the paths of each pair of events for global inference. Instead of a complete graph, we assume that there is an edge between two nodes, only if the two pairs of events share at least one event. Clearly, if they have no common event, there must be no direct causal effect between them. The adaptive focal loss re-weights positive and negative samples to tackle the imbalance issue. On the one hand, we leverage a weighting factor to balance two classes' training. On the other hand, we also introduce a scaling factor to focus more on difficult samples.

Our contributions can be summarized as follows:

- We propose to build an event relational graph without using any external tools to capture causal transitivity.
- We propose a novel framework ERGO that further alleviates false positive and false negative issues for DECI.
- Extensive experiments on two benchmark datasets indicate that ERGO greatly outperforms previous SOTA methods (12.8% F1 gains on average). We have also conducted both quantitative and qualitative analysis to better understand key components of ERGO. Furthermore, detailed error analysis provides insights into our approach and the task.

2 Related Work

ECI has attracted much attention in recent years. In terms of text corpus, there are mainly two types of methods: Sentence-level ECI (SECI) and DECI.

In the first research line, early methods usually design various features tailored for causal expressions, such as lexical and syntactic patterns (Riaz and Girju, 2013, 2014a,b), causality cues or markers (Riaz and Girju, 2010; Do et al., 2011; Hidey and McKeown, 2016), statistical information (Beamer and Girju, 2009; Hashimoto et al., 2014), and temporal patterns (Riaz and Girju, 2014a; Ning et al., 2018). Then, researchers resort to a large amount of labeled data to mitigate the efforts of feature engineering and to learn diverse causal expressions (Hu et al., 2017; Hashimoto, 2019). To alleviate the annotation cost, recent methods leverage Pre-trained Language Models (PLMs, e.g., BERT (Devlin et al., 2019)) for the ECI task and have achieved SOTA performance (Kadowaki et al., 2019; Liu et al., 2020; Zuo et al., 2020). To deal with implicit causal relations, Cao et al. (2021) incorporate the external knowledge from ConceptNet (Speer et al., 2017) for reasoning, which achieves promising results. Zuo et al. (2021a) learn context-specific causal patterns from external causal statements and incorporate them into a target ECI model. Zuo et al. (2021b) propose a data augmentation method to further solve the data lacking problem.

Along with the success of sentence-level natural language understanding, many tasks are extended to the entire document, such as relation extraction (Yao et al., 2019), natural language inference (Yin et al., 2021), and event argument extraction (Ma et al., 2022a). A concurrent and relevant work is (Tan et al., 2022), which also leverages focal loss for entity relation extraction. The difference is that the focal loss in (Tan et al., 2022) is used to make long-tail (positive) classes contribute more to the overall loss, while the focal loss in our ERGO tackles the imbalance issue of DECI task by focusing more on difficult samples. We further leverage a

weighting factor in the focal loss to balance two classes' training, which is not considered in (Tan et al., 2022). Moreover, in Section 4.6, we have given a more detailed analysis of the impact of adaptive focal loss on the DECI task.

Compared with SECI, DECI not only aggravates the lack of clear causal indicators but also poses a new challenge of cross-sentence inference. Gao et al. (2019) use Integer Linear Programming (ILP) to model the global causal structures; Zhao et al. (2021) proposes a document-level context-based graph inference mechanism to capture interaction among events; RichGCN (Tran Phu and Nguyen, 2021) constructs document-level interaction graphs and uses Graph Convolutional Network (GCN, Kipf and Welling (2017)) to capture relevant connections. However, the construction of the aforementioned global structure or graph requires sophisticated feature extraction or tools, which may introduce noise and mislead the model (Tran Phu and Nguyen, 2021). Compared with them, we formulate DECI as an efficient node classification framework, which could capture the global interactions among event pairs automatically, as well as alleviate the imbalanced and noisy issues.

3 Methodology

The goal of our proposed framework ERGO is to capture potential causal chains for document-level reasoning. There are three main components: (1) **Document Encoder** to encode the document and obtain contextualized representations of events as the inputs for the following components; (2) **Event Relational Graph Transformer** that models causal chain for global inference by building a handy event relational graph, where node features are from the Document Encoder and enhanced through propagation over the graph; and (3) **Classification with Adaptive Focal Loss** to predict if a node of event pair has causal relation or not based on their enhanced node features, with considering the imbalance issue.

3.1 Document Encoder

Given a document $\mathcal{D} = [x_t]_{t=1}^{L}$ (can be of any length L), the document encoder aims to output the contextualized document and event representations. We leverage a Pre-trained Language Model (PLM) as a base encoder to obtain the contextualized embeddings. Following conventions, we add special tokens at the start and end of $\mathcal{D}$ (e.g., "[CLS]" and "[SEP]" of BERT (Devlin et al., 2019)), and insert additional special tokens "<t>" and "</t>" at the start and end of all the events to mark event positions. Then, we have:

$$H = [h_1, h_2, ..., h_L] = \text{Encoder}([x_1, x_2, ..., x_L]), \quad (1)$$

where $h_i \in \mathbb{R}^d$ is the embedding of token x_i. We use the embedding of token "[CLS]" to represent the document and the embeddings of token "<t>" to represent the events.

In this paper, we choose pre-trained BERT (Devlin et al., 2019) and Longformer (Beltagy et al., 2020) as encoders for comparison. We handle documents longer than the limits of PLMs as follows.

BERT for Document Encoder To handle documents that are longer than 512 (BERT's original limit), we leverage a *dynamic window* to encode the entire document. Specifically, we divide $\mathcal{D}$ into several overlapping spans according to a specific step size and input them into BERT separately (details can be found in Section 4.2). Then, we find and average all the embeddings of token "[CLS]" or "<t>" of different spans to represent the whole document or each event, respectively.

Longformer for Document Encoder Longformer introduces a localized sliding window based attention mechanism (the default window size is 512) with little global attention to reduce computation and extend BERT for long documents. In our implementation, we apply its efficient local and global attention pattern. Specifically, we use global attention on the "<s>" token (Longformer uses "<s>" and "</s>" as the special start and end tokens, corresponding to BERT's "[CLS]" and "[SEP]"), and local attention on other tokens, which could build full sequence representations. The maximum document length allowed by Longformer is 4096, which is suitable for most documents. Therefore, we directly take the embedding of token "<s>" as document representation and embedding of token "<t>" as event representation.

3.2 Event Relational Graph Transformer

In this section, we first introduce how to construct the event relational graph, including the criss-cross constraint. Then, based on it, we leverage a Relational Graph Transformer (RGT) to capture the high-order interaction among event pairs and obtain enhanced event pair representations for the final classification.

3.2.1 Event Relational Graph Construction

Given all the events of document $\mathcal{D}$, we construct an event relational graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ is the set of edges. We highlight the following differences of $\mathcal{G}$ from previous event graphs. First, for each node in $\mathcal{V}$, it refers to a different pair of events in $\mathcal{D}$, instead of a single event. Our motivation is to learn the relation of relations between events, i.e., the logic of causal transitivity, for higher-order reasoning. Second, for edges $\mathcal{E}$, we do not require any prior relations between events. Instead, we add all edges between any two nodes into $\mathcal{E}$. Thus, $\mathcal{G}$ is initialized as a complete graph.

Criss-cross Constraint. To simplify the graph structure and alleviate the negative impacts of false positives propagation, we introduce a criss-cross constraint. It assumes that there is an edge between two nodes, *only* if the two corresponding event pairs share at least one event. The basic idea behind this is that if two pairs of events have no common event, there must be no *direct* causal effect between them. Still, they can have causal interactions if there are some mediator events, and such causality takes effects conditioned on the mediator. For example in Figure 1, (1) the causality information of (*restore*, *service*) has no effect on predicting the causal relation of (*outage*$_1$, *break*). (2) the causality of (*outage*$_2$, *restored*) has a transitive effect on predicting the causal relation of (*outage*$_1$, *break*) if we know that (*restored*, causes, *break*) and (*outage*$_1$, *outage*$_2$) is coreference[2]. Note that the criss-cross constraint is not posed over the graph directly, which is different for each event pair. In Section 4.5, we show that such a simple and intuitive constraint brings considerable performance gains compared with using a complete graph.

3.2.2 Relational Graph Transformer

Node Embedding Initialization For global inference, we first initialize node feature vectors with event pair node embedding, which is based on the contextualized event embeddings by Equation (1). Formally, for event pair (e_1, e_2) and the corresponding contextual embeddings (h_{e_1}, h_{e_2}), their event pair node embedding is initialized by:

$$v_{e_{1,2}}^{(0)} = [h_{e_1} \| h_{e_2}], \tag{2}$$

where $\|$ denotes concatenation, 0 indicates the initial state for the following neural layers.

The event pair node embeddings represent the implicit relational information between two events, which enables us to integrate event pair representation learning and causal chain inference seamlessly, without any prior knowledge or tools. Clearly, better initial features of nodes will provide more discriminative signals from local textual contexts for classification. On the other hand, structural reasoning further improves the discriminative ability of node features by considering all event pairs globally, such that confident prediction shall help others via causality transitivity.

Each RGT layer l is closed to the transformer architecture proposed in (Vaswani et al., 2017). It takes a set of node embeddings $\mathbf{v}^{(l-1)} \in \mathbb{R}^{N \times d_{\text{in}}}$ as input, and outputs a new set of node embeddings: $\mathbf{v}^{(l)} \in \mathbb{R}^{N \times d_{\text{out}}}$, where N is the number of event pairs, d_{in} and d_{out} are the dimensions of input and output embeddings.

To better exploit the relational information from each neighbor to predict the causal relation of an event pair node i, we perform a shared self-attention mechanism to measure the importance of neighbor j to i:

$$\text{att}_{ij} = \frac{(v_i \mathbf{W}_q)(v_j \mathbf{W}_k)^T}{\sqrt{d_k}}, \tag{3}$$

where d_k is the hidden size, $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d_{\text{in}} \times d_k}$ are parameter weight matrices, $\sqrt{d_k}$ is a scaling factor (Vaswani et al., 2017). Thus, negative and uninformative nodes are expected to assign lower attention weights.

Then we normalize att_{ij} across all choices of j using a softmax function to make the importance more comparable:

$$\alpha_{ij} = \text{softmax}_j(\text{att}_{ij}) = \frac{\exp(\text{att}_{ij})}{\sum_{z \in \mathcal{N}_i} \exp(\text{att}_{iz})}, \tag{4}$$

where $\mathcal{N}_i$ are all the first order neighbors of node i.

To aggregate relational knowledge from the neighborhood information, we compute a weighted linear combination of the embeddings :

$$v_i^{(l)} = \sum_{j \in \mathcal{N}_i} \alpha_{ij} (v_j \mathbf{W}_v), \tag{5}$$

where $\mathbf{W}_v \in \mathbb{R}^{d_{\text{in}} \times d_k}$ is the parameter weight matrix. We also perform multi-head attention to

[2]In the datasets, coreference events have similar surface forms and thus can be implicitly captured by PLMs. We leave further coreference modeling in the future work.

jointly attend to information from different representation subspaces. Finally, the output embedding of node i can be represented as:

$$v_i^{(l)} = \Big(\Big\Vert_{c=1}^{C} \sum_{j \in \mathcal{N}_i} \alpha_{ij}(v_j \mathbf{W}_v) \Big) \mathbf{W}_o, \quad (6)$$

where $\|$ denotes concatenation, C is the number of heads. $\mathbf{W}_o \in \mathbb{R}^{C\dot{d}_k \times d_{\text{out}}}$ is the parameter weight matrix. By simultaneously computing embeddings for all the event pair nodes, a node embedding matrix $\mathbf{v}^{(l)} \in \mathbb{R}^{N \times d_{\text{out}}}$ is obtained. By stacking multiple layers, RGT could reach high-order connectivity and capture complex interactions.

Note that our framework is flexible to almost arbitrary Graph Neural Networks (GNNs). Here we leverage RGT for its powerful expressiveness. We also report results with GCN in Section 4.5.

3.3 Classification with Adaptive Focal Loss

Remember that we formulate DECI as a node classification task, which predicts the label of each node as either a positive or negative class. However, the number of negative samples during training far exceeds that of positives, leading to an imbalanced classification problem. What is worse, the dominant negatives contain many spurious correlations between events ("treatment" and "death" example in Section 1). How can we know the difficulties of sample prediction, so that ERGO can penalize them to alleviate false negatives for better performance?

To address this problem, we leverage an adaptive loss function for training, following focal loss (Lin et al., 2017). Specifically, we reshape the loss function to down-weight easy samples and thus focus on hard ones. Formally, a modulating factor is added to Cross-Entropy (CE) loss, with a predefined focusing hyper-parameter $\gamma \geq 0$, which is defined as:

$$\mathcal{L}_{\text{FL}} = -\sum_{e_i, e_j \in \mathcal{D}} (1 - p_{e_{i,j}})^{\gamma} \log(p_{e_{i,j}}). \quad (7)$$

where $p_{e_{i,j}}$ is the predicted probability of whether there is a causal relation between events e_i and e_j. $p_{e_{i,j}}$ is defined as follows:

$$p_{e_{i,j}} = \text{softmax}\left(\left[v_{e_{i,j}} || h_{[CLS]} \right] \mathbf{W}_p \right), \quad (8)$$

where $\mathbf{W}_p$ is the parameter weight matrix, $\|$ denotes concatenation. Here we concatenate embeddings of $h_{[CLS]}$ (of BERT) or $h_{<s>}$ (of Longformer) to each node in order to incorporate the global document representation for classification.

This scaling factor, $(1 - p_{e_{i,j}})^{\gamma}$, allows us to efficiently train on all event pairs by encouraging the model to focus on difficult samples, reducing false-negative predictions. For example, when a sample is misclassified and $p_{e_{i,j}}$ is small, the modulating factor is near 1, and the loss is unaffected. As $p_{e_{i,j}} \to 1$, the factor goes to 0 and the loss for well-classified examples is down-weighted. Therefore, the focusing parameter γ smoothly adjusts the rate at which easy examples are down-weighted. When $\gamma = 0$, $\mathcal{L}_{\text{FL}}$ is equivalent to CE loss, and with the increase of γ, the influence of the modulating factor also increases. We will give further discussion in Section 4.6.

Besides, we use an α-balanced variant of the focal loss, which introduces a weighting factor α in $[0, 1]$ for class "positive" and $1 - \alpha$ for class "negative". The value of α is related to the ratio of positive and negative samples. The final adaptive focal loss $\mathcal{L}_{\text{FL}_b}$ can be written as:

$$\mathcal{L}_{\text{FL}_\text{b}} = -\sum_{e_i, e_j \in \mathcal{D}} \alpha_{e_{i,j}} (1 - p_{e_{i,j}})^{\gamma} \log(p_{e_{i,j}}). \quad (9)$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our proposed method on two widely used datasets, EventStoryLine (version 0.9) (Caselli and Vossen, 2017) and Causal-TimeBank (Mirza, 2014).

EventStoryLine contains 22 topics, 258 documents, 5,334 events, 7,805 intra-sentence and 62,774 inter-sentence event pairs (1,770 and 3,885 of them are annotated with causal relations respectively). Following Gao et al. (2019) and (Tran Phu and Nguyen, 2021), we group documents according to their topics. Documents in the last two topics are used as the development data, and documents in the remaining 20 topics are employed for a 5-fold cross-validation.

Causal-TimeBank contains 184 documents, 6,813 events, and 318 of 7,608 event pairs are annotated with causal relations. Following (Liu et al., 2020), we employ a 10-fold cross-validation evaluation. Note that the number of inter-sentence event pairs in Causal-TimeBank is quite small (i.e., only 18 pairs), following (Tran Phu and Nguyen, 2021), we only evaluate ECI performance for intra-sentence event pairs on Causal-TimeBank.

Evaluation Metrics For evaluation, we adopt Precision (P), Recall (R), and F1-score (F1) as evaluation metrics, same as previous methods to ensure comparability.

4.2 Implementation Details

We implement our method based on Pytorch. We use uncased BERT-base (Devlin et al., 2019) or Longformser-base (Beltagy et al., 2020) as the document encoder. For the BERT-base document encoder, we set the dynamic window size to 256, and divide documents into several overlapping windows with a step size 32. We optimize our model with AdamW (Loshchilov and Hutter, 2019) using a learning rate of 0.00002 with a linear warm-up for the first 8% steps. We apply dropout (Srivastava et al., 2014) between layers and clip the gradients of model parameters to a max norm of 1.0. We perform early stopping based on the F1 score on the development set. We tune the hyper-parameters by grid search based on the development set performance: heads $C \in \{1, 2, \mathbf{4}, 8\}$ for the relational graph transformer model, dropout rate $\in \{0.1, \mathbf{0.2}, 0.3\}$, focusing parameter $\gamma \in \{0, 1, \mathbf{2}, 3\}$, and weighting factor $\alpha \in \{0.65, \mathbf{0.75}, 0.85\}$.

4.3 Baselines

We compare our proposed ERGO with various state-of-the-art SECI and DECI methods.

SECI Baselines (1) **KMMG** (Liu et al., 2020), which proposes a mention masking generalization method and use extenal knowledge databases. (2) **KnowDis** (Zuo et al., 2020), a knowledge enhanced distant data augmentation method to alleviate the data lacking problem. (3) **LSIN** (Cao et al., 2021), which constructs a descriptive graph to leverage external knowledge and has the current SOTA performance for intra-sentence ECI. (4) **LearnDA** (Zuo et al., 2021b), which uses knowledge bases to augment training data. (5) **CauSeRL** (Zuo et al., 2021a), which learns context-specific causal patterns from external causal statements for ECI.

DECI Baselines (1) **OP** (Caselli and Vossen, 2017), a dummy model that assigns causal relations to event pairs. (2) **LR+** and **LIP** (Gao et al., 2019), feature-based methods that construct document-level structures and use various types of resources. (3) **BERT (our implement)** a baseline method that leverages dynamic window and event marker techniques. (4) **RichGCN** (Tran Phu and Nguyen, 2021), which constructs document-level interaction graph and uses GCN to capture relevant connections. RichGCN has the current SOTA performance for inter-sentence ECI.

Model	EventStoryLine			Causal-TimeBank		
	P	R	F1	P	R	F1
OP	22.5	**98.6**	36.6	-	-	-
LR+	37.0	45.2	40.7	-	-	-
LIP	38.8	52.4	44.6	-	-	-
KMMG[○]	41.9	62.5	50.1	36.6	55.6	44.1
KnowDis[○]	39.7	66.5	49.7	42.3	60.5	49.8
LSIN[○]	47.9	58.1	52.5	51.5	56.2	53.7
LearnDA[○]	42.2	69.8	52.6	41.9	68.0	51.9
CauSeRL[○]	41.9	69.0	52.1	43.6	**68.1**	53.2
BERT[○]	47.8	57.2	52.1	47.6	55.1	51.1
RichGCN[○]	49.2	63.0	55.2	39.7	56.5	46.7
ERGO[○]	49.7	72.6	59.0	58.4	60.5	59.4
ERGO[◇]	**57.5**	72.0	**63.9**	**62.1**	61.3	**61.7**

Table 1: Model's intra-sentence performance on EventStoryLine and Causal-TimeBank, the best results are in **bold** and the second-best results are underlined. [○] and [◇] denote models that use pre-trained BERT-base and Longformer-base encoders, respectively. Overall, our ERGO outperforms previous SOTA models (with a significant test at the level of 0.05).

4.4 Overall Results

Since some baselines are evaluated only on EventStoryLine, the baselines used for EventStoryLine and Causal-TimeBank are different. Some baselines can not handle the inter-sentence scenarios in EventStoryLine. Thus we report the results of intra- and inter- sentence settings separately.

4.4.1 Intra-sentence Evaluation

From Table 1, we can observe that:

(1) ERGO outperforms all the baselines by a large margin on both datasets. Compared with SOTA methods, ERGO-BERT$_{\text{BASE}}$ achieves 6.9% improvements of F1-score on EventStoryLine, and 10.6% on Causal-TimeBank. This demonstrates the effectiveness of ERGO.

(2) The feature-based method OP achieves the highest Recall on EventStoryLine, which may be due to simply assigning causal relations by mimicking the textual order of presentation. This leads to many false positives and thus a low Precision.

(3) The usage of PLMs boosts performance. Using Longformer$_{\text{BASE}}$ as the encoder, ERGO achieves better results than ERGO-BERT$_{\text{BASE}}$, which also achieves new SOTA results. The reason may be: 1) Longformer continues pre-training from RoBERTa (Liu et al., 2019), which has been

Model	Inter-sentence			Intra + Inter		
	P	R	F1	P	R	F1
OP	8.4	**99.5**	15.6	10.5	**99.2**	19.0
LR+	25.2	48.1	33.1	27.9	47.2	35.1
LIP	35.1	48.2	40.6	36.2	49.5	41.9
BERT[○]	36.8	29.2	32.6	41.3	38.3	39.7
RichGCN[○]	39.2	45.7	42.2	42.6	51.3	46.6
ERGO[○]	43.2	48.8	45.8	46.3	50.1	48.1
ERGO[◇]	**51.6**	43.3	**47.1**	**48.6**	53.4	**50.9**

Table 2: Model's inter and (intra+inter)-sentence performance on EventStoryLine.

Model	Intra	Inter	Intra + Inter
ERGO[○]	59.0	45.8	48.1
ERGO_1[○]	56.6	43.5	45.6
ERGO_2[○]	56.2	41.8	44.6
ERGO_3[○]	58.3	43.6	47.3
ERGO[◇]	**63.9**	**47.1**	**50.9**
ERGO_1[◇]	61.3	44.7	47.1
ERGO_2[◇]	60.7	43.1	46.3
ERGO_3[◇]	62.6	45.9	49.1

Table 3: F1 Results of Ablation study on EventStoryLine, where ERGO_1 denotes ERGO w/ a complete graph, ERGO_2 denotes ERGO w/ GCN, ERGO_3 denotes ERGO w/o the focal factor.

found to outperform BERT on many tasks; 2) Longformer leverages an efficient local and global attention pattern, which is beneficial to capture longer contextual information for inference.

4.4.2 Inter-sentence Evaluation

From Table 2, we can observe that:

(1) ERGO greatly outperforms all the baselines under both inter- and (intra+inter)-sentence settings, especially in terms of Precision. This demonstrates that our ERGO can make better document-level inference via the event relational graph, while alleviating the negative impacts of false positives.

(2) The overall F1-score of inter-sentence setting is much lower than that of intra-sentence, which indicates the challenge of document-level ECI.

(3) The BERT baseline performs well on intra-sentence event pairs. However, it performs much worse than LIP, RichGCN, and ERGO under inter-sentence settings, which indicates that a document-level structure or graph is helpful to capture the global interactions for prediction.

4.5 Ablation Study

To analyze the main components of ERGO, we have the following variants, as shown in Table 3:

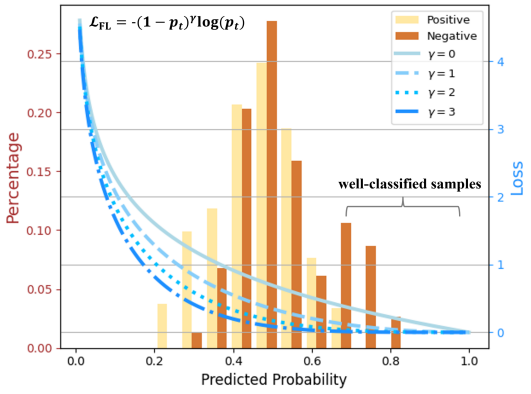

Figure 2: Distribution histogram of predicted probabilities of positive and negative event pairs and the visualized loss with focal parameter $\gamma = \{0, 1, 2, 3\}$.

(1) **w/ a complete graph**, which connects all the nodes in the event relational graph (without the criss-cross constraint mentioned in Section 3.2.1). Compared with the full ERGO model (both $\text{BERT}_{\text{BASE}}$ and $\text{Lonformer}_{\text{BASE}}$), ERGO (w/ a complete graph) clearly decreases the performance, which demonstrates the effectiveness of the handy design of criss-cross constraints.

(2) **w/ GCN**, which replaces the RGT in Section 3.2.2 with a well-known GNN model, GCN. It can be seen that **(i)** ERGO (w/ GCN) also performs better or competitive than other baselines. This indicates that our framework is flexible to other GNNs, and the main improvement comes from our new formulation of the ECI task. **(ii)** the full ERGO model clearly outperforms ERGO (w/ GCN), which validates the effectiveness of our RGT model.

(3) **w/o focal factor**, which sets the focusing parameter $\gamma = 0$ (in Section 3.3) and thus makes the focal loss degenerate into standard CE loss. Compared with the full ERGO model, ERGO (w/o focal factor) also decreases performance. This highlights the effectiveness of penalizing hard samples via an adaptive focal loss in the ECI task.

4.6 Dealing with the Imbalance Issue

In Figure 2, we show the distribution histogram of the predicted probability after the first training epoch for positive and negative samples, respectively (denoted by the bars). The predicted probability of x-axis reflects the difficulty of samples (i.e., the lower, the harder), and the curves denote loss — how much penalization on the corresponding samples during learning. From the histogram, we can find: **(1)** the model is less confident about

Teen Arrested$_1$ in Shooting of Hero Brooklyn Mom
Posted Oct 26, 2011 8:57 AM CDT
An 18-year-old gang member has **confessed** to killing a pregnant mom , who **died** on Friday as she **shielded** a group of children from bullets, but insisted he "did not mean to shoot the ladies," sources tell the New York Daily News.
In addition to Zurana Horton-who was a mother of 12-another mom and an 11-year-old girl were wounded by rooftop sniper Andrew Lopez, who told police his dozen rounds were intended for members of a rival gang.
Lopez has been **charged** with **murder**; his two half-brothers, 17 and 22, were also arrested$_2$.

No.	Event Pair	GT	BERT	ERGO
1	(Shooting, killing)	Yes	*No*	Yes
2	(killing, arrested$_2$)	Yes	*No*	Yes
3	(Shooting, Arrested$_1$)	Yes	Yes	Yes
4	(Arrested$_1$, arrested$_2$)	No	*Yes*	No
5	(Shooting, wounded)	Yes	*No*	Yes
6	(wounded, arrested$_2$)	No	No	No
7	(Shooting, arrested$_2$)	Yes	*No*	Yes
…	…	…	…	…

Figure 3: The case study of BERT baseline and our proposed ERGO, where "GT" denotes the ground truth class, and the right two columns are the output of BERT and ERGO (italic red color means wrong prediction). The thickness of arrows represents the size of attention values, and the bold green arrows show a possible reasoning path.

positives than negatives, i.e., the left-of-center distributed bars of positives. This matches our intuition that the imbalance issue brings a great challenge of false-negative predictions to ECI. **(2)** we visualize focal loss with γ values $\in \{0, 1, 2, 3\}$. The top solid blue curve ($\gamma = 0$) can be seen as the standard CE loss. As γ increases, the shape of focal loss moves to the bottom left corner. That is, the learning of ERGO pays more attention to hard samples. In practice, we find $\gamma = 2$ works best on both datasets, indicating that there is a balance between the focus on simple and difficult samples.

4.7 Case Study

In this section, we conduct a case study to further illustrate an intuitive impression of our proposed ERGO. As shown in Figure 3, We notice that: **(1)** BERT is good at sentence-level ECI (e.g., No.3 event pair), but fails at more complex cross-sentence cases (e.g., No.1, 2, 4, 5, 7). **(2)** By contrast, ERGO can make correct predictions by modeling the global interactions among event pairs. **(3)** Figure 3 shows 3 causal patterns that ERGO could cover: **(i) Transitivity** (No.1, 2, 7 event pairs): knowing both (*Shooting*, *killing*) and (*killing*, *arrested*$_2$) have causal relations, we could infer that (*Shooting*, *arrested*$_2$) has a causal relation; **(ii) Implicit Coreference Assistance** (No.3, 4, 7 event pairs) : Given that (*Shooting*, *Arrested*$_1$) has a causal relation and (*Arrested*$_1$, *arrested*$_2$) is coreference, we could infer that (*Shooting*, *arrested*$_2$) has a causal relation, even if the causal relation of (*Arrested*$_1$, *arrested*$_2$) is implicitly modeled. We attribute this to PLMs

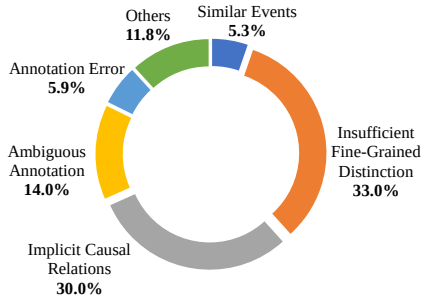

Figure 4: Statistics of Error Types.

that tend to capture coreference relations, such as similar tokens. A piece of supporting evidence is that BERT incorrectly predicts the coreferenced No.4 event pair with a causal relation. **(iii) Deconfounding Negatives** (No.5, 6, 7 event pairs): Knowing (*Shooting*, *wounded*) has a causal relation, although (*wounded*, *arrested*$_2$) does not has a causal relation, it is still possible that (*Shooting*, *arrested*$_2$) has a causal relation through other paths. Correspondingly, as shown in the bottom right, both (*Shooting*, *wounded*) and (*wounded*, *arrested*$_2$) are assigned with very low attention weights, blocking the propagation over these uninformative paths, to avoid the negative confounders contaminating causal transitivity.

4.8 Remaining Challenges

We randomly sample 20 documents of different topics from EventStoryLine, which contains 170 event pairs whose causal relations cannot be correctly predicted by our model. As shown in Figure 4, we manually categorize these pairs into different types and discuss the remaining challenges:

Insufficient Fine-Grained Distinction and Need to Extract Temporal Information (33%) For example, in the following document:

"...Dubai experienced a slight 'tremor' today, ***after*** *a more serious earthquake in Southern Iran, resulting in the evacuation of Emirates Towers and a few other scrapers..."*

The "*tremor*" happens in "*Dubai*" and the "*earthquake*" happens in "*Southern Iran*", they are two different events identified by the temporal indicator "*after*". ERGO incorrectly predicts that there is a causal relation in (*earthquake*, *evacuation*). Future work could consider joint extraction of causal and temporal relations within the document.

Events with Similar Semantics (5.3%) Take the following document as an example:

"...Kenneth Dorsey says the woman accused of killing two co-workers and critically injuring a third at the Kraft plant in Northeast Philly is a good person. And so were the two women she's accused of gunning down with a .357 Magnum, just minutes after she'd been suspended and escorted from the building..."

ERGO incorrectly predicts that there is a causal relation between "*killing*" and "*gunning down*". The reason is that "*killing*" and "*gunning down*" are actually coreference, which suggests a future direction in exploring related tasks.

Implicit Causal Relations (30%) ERGO still fails at many implicit causal relations. For example, the causal relation between "*killing*" and "*suspended*" in the aforementioned document. This is mainly because there are insufficient events for global reasoning and hard negatives bring noise. Clearly, commonsense reasoning will be helpful in this case, since "*suspended*" is an unexpected change that may bring some negative emotions.

Ambiguous Annotation (14%) This type denotes that ambiguous causality within some event pairs. For example, in the following document:

"... A Texas inmate escaped from a prison van near Houston after pulling a gun on two guards who were transporting him between prisons..."

We can think there is a causal relation between "*escaped*" and "*transporting*" because if there is no "*transporting*", the "*inmate*" will have no chance to "*escape*". However, we can also think that there is no causal relation between them because it is not "*transporting*" that directly causes "*escape*".

Finally, as shown in Figure 4, our statistics show that the other errors have to do with annotation errors (5.9%) and more complicated issues that cannot be categorized clearly ("Others", 11.8%).

5 Conclusion

In this paper, we regard DECI as a node classification task by constructing an event relational graph. We propose a novel Event Relational Graph Transformer (ERGO) framework that could capture potential causal chains and mitigate the false positive and false negative issues for DECI. Extensive experiments show great improvements of ERGO under both intra- and inter-sentence settings on two widely used benchmarks. Further analysis provides insights into our approach and the DECI task. In the future, we will consider introducing commonsense reasoning and auxiliary tasks to discover more reliable causality.

Acknowledgments

This work was supported by National Key Research and Development Program of China under Grant No. 2018AAA0101902, NSFC under Grant No. 61532001, the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant, and the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–441. Springer.

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

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. 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. Association for Computational Linguistics.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th ACL and the 11th IJCNLP (Volume 1:*

Long Papers), pages 4862–4872, Online. Association for Computational Linguistics.

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

Bridget Copley and Fabienne Martin. 2014. *Causation in grammatical structures*, volume 52. Oxford University Press.

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

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

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

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from Wikipedia. In *Proceedings of EMNLP-IJCNLP 2019*, pages 2988–2999, Hong Kong, China. Association for Computational Linguistics.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd ACL (Volume 1: Long Papers)*, pages 987–997, Baltimore, Maryland. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th ACL (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Zhichao Hu, Elahe Rahimtoroghi, and Marilyn Walker. 2017. 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. Association for Computational Linguistics.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of 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. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th ICLR, 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3608–3614. ijcai.org.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022a. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774, Dublin, Ireland. Association for Computational Linguistics.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, and Jing Shao. 2022b. MMEKG: Multi-modal event knowledge graph towards universal representation across modalities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 231–239, Dublin, Ireland. Association for Computational Linguistics.

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

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

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. 2016. A semi-supervised learning approach to why-question answering. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3022–3029. AAAI Press.

Laurie Ann Paul, Ned Hall, and Edward Jonathan Hall. 2013. *Causation: A user's guide*. Oxford University Press.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *2010 IEEE Fourth International Conference on Semantic Computing*, pages 361–368. IEEE.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 21–30, Metz, France. Association for Computational Linguistics.

Mehwish Riaz and Roxana Girju. 2014a. 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. Association for Computational Linguistics.

Mehwish Riaz and Roxana Girju. 2014b. 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. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of ACL*.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Kun Zhao, Donghong Ji, Fazhi He, Yijiang Liu, and Yafeng Ren. 2021. Document-level event causality identification via graph inference mechanism. *Information Sciences*, 561:115–129.

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

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021b. LearnDA: Learnable knowledge-guided data augmentation for event causality identification. In *Proceedings of the 59th ACL and the 11th IJCNLP (Volume 1: Long Papers)*, pages 3558–3571, Online. Association for Computational Linguistics.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th COLING*, pages 1544–1550, Barcelona, Spain (Online). International Committee on Computational Linguistics.