Discriminative Reasoning with Sparse Event Representation for Document-level Event-Event Relation Extraction

Changsen Yuan[1], **Heyan Huang**[1,*], **Yixin Cao**[2], **Yonggang Wen**[3]
[1]Beijing Institute of Technology, Beijing, China
[2]Singapore Management University, Singapore
[3]Nanyang Technological University, Singapore
{yuanchangsen, hhy63}@bit.edu.cn
caoyixin2011@gmail.com, ygwen@ntu.edu.sg

Abstract

Document-level Event-Event Relation Extraction (DERE) aims to extract relations between events in a document. It challenges conventional sentence-level task (SERE) with difficult long-text understanding. In this paper, we propose a novel DERE model (SENDIR) for better document-level reasoning. Different from existing works that build an event graph via linguistic tools, SENDIR does not require any prior knowledge. The basic idea is to discriminate event pairs in the same sentence or span multiple sentences by assuming their different information density: 1) low density in the document suggests sparse attention to skip irrelevant information. Our module 1 designs various types of attention for event representation learning to capture long-distance dependence. 2) High density in a sentence makes SERE relatively easy. Module 2 uses different weights to highlight the roles and contributions of intra- and inter-sentential reasoning, which introduces supportive event pairs for joint modeling. Extensive experiments demonstrate great improvements in SENDIR and the effectiveness of various sparse attention for document-level representations. Codes will be released later.

1 Introduction

Event-Event Relation Extraction (ERE) is the task of identifying the relation between two events in texts. As shown in Figure 1, for any pair of events[1], e.g., (*Services*, *downtime*), it shall make the classifications for which relation type it holds. Clearly, event pairs may be in the same sentence (SERE) (Kadowaki et al., 2019a; Liu et al., 2020; Kadowaki et al., 2019b), or scattered across the entire document (DERE) (Phu and Nguyen, 2021). In practice, DERE can benefit a wider range of applications, such as knowledge graph construction (Chen et al., 2019) and future event forecasting (Hashimoto, 2019) but it remains challenging due to the difficulty in long-text understanding.

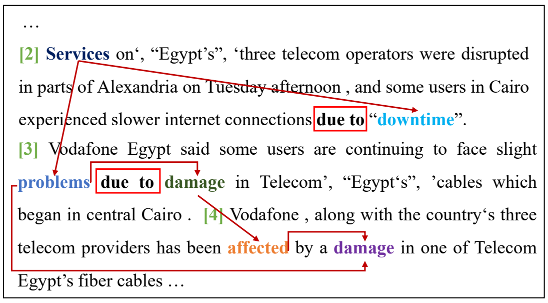

…
[2] **Services** on', "Egypt's", 'three telecom operators were disrupted in parts of Alexandria on Tuesday afternoon , and some users in Cairo experienced slower internet connections **due to** "**downtime**".
[3] Vodafone Egypt said some users are continuing to face slight **problems** **due to** **damage** in Telecom', "Egypt's", 'cables which began in central Cairo . [4] Vodafone , along with the country's three telecom providers has been **affected** by a **damage** in one of Telecom Egypt's fiber cables …

Figure 1: Example of document-level ERE from EventStoryLine. Solid lines denote causal relations; some events and relations are omitted for clarity. The red boxes are indicator words for causal relations.

In this paper, we propose to improve the reasoning ability among events spanning the entire document for DERE. Different from conventional methods, which build an event graph based on linguistic tools (Phu and Nguyen, 2021; Gao et al., 2019; Xu et al., 2023; Zeng et al., 2021), we focus more on the nature of document itself and do not rely on any prior knowledge. To do this, we highlight the following key questions:

- How to capture events' dependence that may be far away?
- Should we treat all event pairs equally considering the essential difference between SERE and DERE?

To address them, we propose a novel DERE model that learns **S**parse **E**ve**N**t representations for **D**iscriminating **I**ntra- and inter-sentential **R**easoning, namely **SENDIR**. Inspired by MAE (He et al., 2022), we observe a different information density between sentences and documents — for an event, most parts of the document are irrelevant, leading to a low information density. By contrast, the sentence has a high density and

[*] Corresponding author
[1] Event is defined as the trigger word in this area.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 16222–16234
July 9-14, 2023 ©2023 Association for Computational Linguistics

usually contains related words as causal indicators. As shown in Figure 1, $problem_3 \xrightarrow{causal} damage_3$ in the third sentence has clear causal word (`due to`)[2], making the prediction much easier. While, for $problem_3 \xrightarrow{causal} damage_4$ across the third and fourth sentences, there is no such pattern. This motivates us the design two modules as follows.

The goal of module 1 is to shorten the dependence distance among events to learn high-quality local and global context representations. The basic idea is to learn event-specific sentence embeddings as local features. Based on that, we further utilize sparse self-attention globally to skip irrelevant information. We have defined various types of attention masks to reflect specific dependence among sentences. In addition to conventional random sparse attention (Tay et al., 2021), we have also explored Narrative, Flashback, Global→, Global←, and Banded attention, to reflect specific language bias according to human writing habits. We name this module sparse event representation learning, while these sparse attentions shall also benefit other long-text understanding tasks.

Module 2 aims to discriminate intra- and inter-sentential reasoning to help difficult cross-sentence events with relatively easy within-sentence-level events. As shown in Figure 1, it is easy to predict ($problem_3 \xrightarrow{causal} damage_3$) with high confidence, which is part of the path $problem_3 \xrightarrow{causal} damage_3 \xrightarrow{causal} affected_4 \xrightarrow{causal} damage_4$ for prediction of ($problem_3 \xrightarrow{causal} damage_4$). Thus, for each event pair, we take the outputs of module 1 as intra-sentential features. We then enhance them with selected supportive event pairs to constitute possible reasoning chain, and utilize Gated Attention Unit (GAU) (Hua et al., 2022) to conduct inter-sentential reasoning. Finally, we combine these two types of features with varying weights to differentiate their confidence and roles in ERE. Thus, we can improve the prediction of event pairs across sentences without hurting the performance of the pairs within a sentence.

We summarize the contributions as follows:

- We propose to discriminate intra- and inter-sentential reasoning considering the essential difference between SERE and DERE.
- We propose a novel DERE model SENDIR without any prior knowledge or external tools.
- Experimental results on three public datasets demonstrate the effectiveness of SENDIR. Further studies also verify our proposed sparse attentions and discriminative reasoning in long-text understanding.

2 Related Work

2.1 Sentence-level ERE

Early ERE methods focus on SERE and exploit various textual features to represent the relation, such as syntactic features (Venkatachalam et al., 2021; Ning et al., 2019), causal patterns (Riaz and Girju, 2010; Hidey and McKeown, 2016), and statistical features of causal information (Mirza et al., 2014; Hu et al., 2017; Tan et al., 2022). Following the success of pre-trained language models (PLMs) (Devlin et al., 2019; Liu et al., 2019), recent works tend to enhance PLMs with external knowledge, so that SERE models can obtain high-quality contextualized event representations for classification. Hashimoto (2019) exploited the cause and effect entities in Wikipedia and the multilingual inter-wiki links as weak supervision. Zuo et al. (2021a) introduced external causal statements and adapted a contrastive transfer strategy to incorporate them into a target model. Cao et al. (2021) utilized ConceptNet (Speer et al., 2017) to learn latent structure of event causal relation, and Zuo et al. (2021b, 2020) designed a knowledge-guided method to generate new samples based on several knowledge sources, such as WordNet (Miller, 1998) and VerbNet (Schuler, 2005). Although the SERE has achieved great success, events usually scatter the entire document in real scenarios. Therefore, DERE has attracted more and more research attention.

2.2 Document-level ERE

Compared with SERE, DERE has a wider range of applications but is more challenging due to the difficult long-text understanding. Thus, researchers tend to consider event-event structures for global reasoning. Gao et al. (2019) used integer linear programming (ILP) to model causal information by designing constraints and modifying the objective function to encourage causal structures and discourage the opposite. To build the connections among the events, Phu and Nguyen (2021) designed various document-level graphs and used graph convolutional networks to learn structure-preserved features. Chen et al. (2022) proposed to build an event pair relational graph and converted DERE to

[2] We use subscript to denote the sentence index.

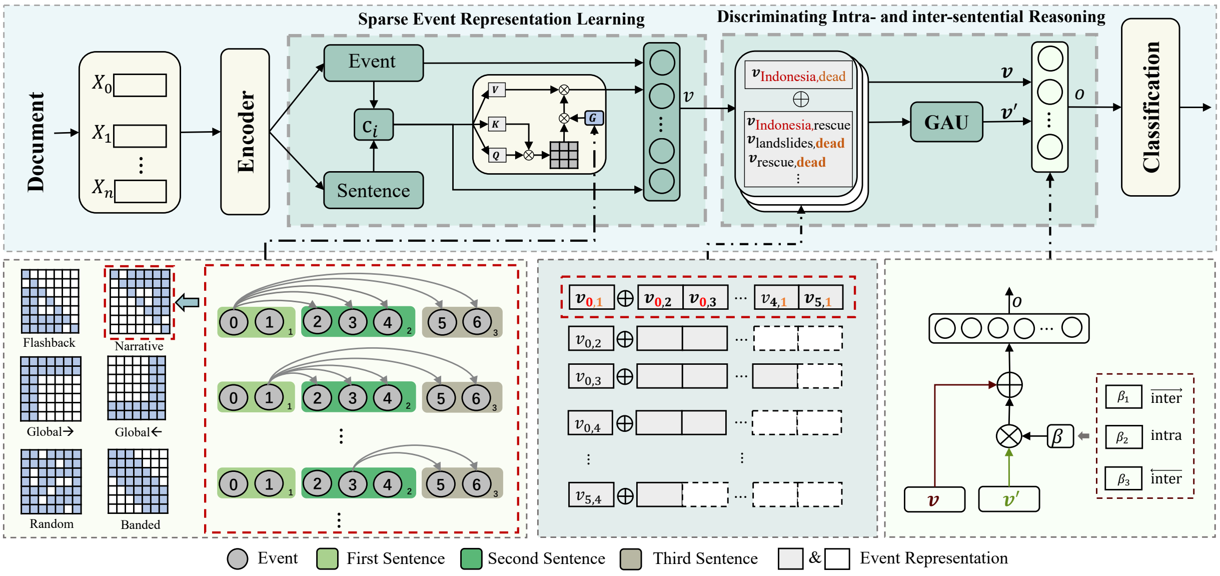


Figure 2: Architecture of proposed model. The white dotted box is zero embedding, which can supplement matrices.

the node classification, which captures the global interactions and alleviates the spurious correlation between events. Man et al. (2022) proposes to model the important context sentences to identify relations. However, to capture long-distance information, these approaches typically construct an additional document-level graph to assist global reasoning. The graph introduces unnecessary noise and decreases efficiency. Instead, we explore the information in documents without requiring prior knowledge or external tools. Certainly, our method can be further improved by incorporating structural knowledge. We leave it in the future.

3 Methodology

SENDIR aims to learn high-quality event representations to facilitate both intra- and inter-sentential reasoning. As shown in Figure 2, our framework has four main components: **Encoder** to encode a document into vector, **Sparse Event Representation Learning** (SER) that further learns event representations based on document embeddings, **Discriminating Intra- and inter sentential Reasoning** (DIR) that conducts joint inference based on each pair of event representations, and **Classification** to make final predictions.

3.1 Encoder

We utilize the BERT (Devlin et al., 2019) with Bi-LSTM to encode the document for long documents (more than 512 tokens). Given a document D with n sentences and N events, $D = [X_1, X_2, \ldots, X_n]$, and the sentence ($X_i = [x_1, x_2, \ldots, x_l]$) contains l words, the encoder is expressed as follows:

$$\begin{cases} \mathbf{H} = \text{Bi-LSTM}([\boldsymbol{s}_1, \ldots, \boldsymbol{s}_i, \ldots, \boldsymbol{s}_n]) \\ \boldsymbol{s}_i = \text{BERT}(\mathbf{X}_i) \end{cases}, \quad (1)$$

where $\mathbf{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{n*l}]$ is the embedding of token, and $\boldsymbol{h}_i \in \mathbb{R}^d$. For event $\boldsymbol{e}_{i,p}$, where i denotes i^{th} event and p denotes the index of sentence, we define its embeddings as $\boldsymbol{e}_{i,p} = \boldsymbol{h}_k$, if the event mention word is x_k, the position of the event in the document is k.

3.2 Sparse Event Representation Learning

SER explores various types of attention to capture long-distance dependence between sentences for high-quality document representation, which will be used to enhance the event representation. In specific, SER first learns event-specific sentence embedding as the local context. Based on them, we then apply sparse self-attention to skip irrelevant information as global contexts. Particularly, we introduce various types of long-distance dependency assumptions. Finally, we define event representations based on local and global contexts. We highlight the following differences and advantages of SER from the previous document representation: (1) The global contexts taking sentences as nodes shortens the token-level distance between events. (2) Well-designed sparse attention mechanism will bring useful language bias that further alleviates the difficulty of modeling long-distance dependence.

Event-specific Sentence Embedding. Given event and its sentence embeddings (i.e., $\boldsymbol{e}_{i,p}$ and $\boldsymbol{s}_p$), we use the event as a query and compute event-specific sentence embedding $\boldsymbol{c}_i$ as follows:

$$\begin{cases} \boldsymbol{w}_i = \text{Softmax}(\boldsymbol{s}_p \boldsymbol{e}_{i,p}) \\ \boldsymbol{c}_i = \boldsymbol{w}_i \boldsymbol{s}_p \end{cases}, \tag{2}$$

where $\boldsymbol{c}_i$ is defined as local context of event $e_{i,p}$, and $\boldsymbol{w}_i$ is the trainable parameters.

Sparse Attention Mask. We apply self-attention technique to fuse information from the above event-specific sentence embeddings $\mathbf{C}^1 = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_N] \in \mathbb{R}^{N \times d}$ as the global context. Note that N is the number of events rather than sentences, because each event has its own sentence embedding, even if two events are located in the same sentence.

Inspired by Child et al. (2019), we design the sparse mask matrix to deal with the long-distance dependency issue. We have designed six types of masks: Global→, Global←, Random, Banded, Narrative, and Flashback. Note that the assumptions behind them are not always true and we will give a discussion later. Their intuitive impression is shown in Figure 2. Formally, we define the mask matrix as $\mathbf{G} \in \mathbb{R}^{N \times N}$, where each element $\mathbf{G_{i,j}} \in \{0, 1\}$ denotes if the information of the j^{th} event in the q^{th} sentence $e_{j,q}$ can be seen by the i^{th} event in the p^{th} sentence $e_{i,p}$. We define the following masks:

- **Global→.** We assume the events in the first several sentences (e.g., the first two sentences) are core topics of the document and should see all of the other events:

$$G_{i,j} = \begin{cases} 1, & if\ p < 3,\ or\ q < 3 \\ 0, & otherwise \end{cases}. \tag{3}$$

- **Global←.** We assume the events in the last several sentences (e.g., the last two sentences) are conclusion topics of the document and should see all of the other events:

$$G_{i,j} = \begin{cases} 1, & if\ p > N-3,\ or\ q > N-3 \\ 0, & otherwise \end{cases}. \tag{4}$$

- **Random.** Random sparse masks are usually used to increase the ability of non-local interactions—we randomly sample 20% matrix element as 0, and others are 1.

- **Banded.** We assume that the related information is narrowed down into neighbor sentences only. That is, each event can only see the events in neighbor sentences:

$$G_{i,j} = \begin{cases} 1, & if\ |p - q| < 3 \\ 0, & otherwise \end{cases}. \tag{5}$$

- **Narrative.** We assume that the events are mostly described in narrative order, so that the former event can see the latter one:

$$\mathbf{G}_{i,j} = \begin{cases} 1, & if\ q - p > 0 \\ 0, & otherwise \end{cases}. \tag{6}$$

- **Flashback.** We assume that events are sequentially written, and thus the latter one should see the former one:

$$\mathbf{G}_{i,j} = \begin{cases} 1, & if\ p - q > 0 \\ 0, & otherwise \end{cases}. \tag{7}$$

Discussion of sparsity assumptions. Due to the complexity of language, we have designed the above six types of attention masks to capture different linguistic biases. Although these masks have patterns suitable for specific settings, they also have unsuitable cases where the underlying assumptions shall fail. However, all of them are designed based on our core assumption — the information density in documents is lower than that in sentences. Thus, we capture long-distance dependence via the sparse attention mechanism. In the experiment (Section 4.7), we demonstrate that even the random attention mask can improve document-level performance, and other types of attention masks (e.g., Narrative) have achieved further improvements by introducing additional language bias.

Event Representation. Given local context $\boldsymbol{C}$ and attention mask G, we now use self-attention (Vaswani et al., 2017) to obtain global context, which is further combined with local context as event representations. The global context $\mathbf{C}' = [\boldsymbol{c}_1^{'}, \boldsymbol{c}_2^{'}, \ldots, \boldsymbol{c}_N^{'}] \in \mathbb{R}^{N \times d}$ can be computed as follows:

$$\mathbf{C}' = \boldsymbol{\alpha}\mathbf{C} \tag{8}$$

where $\boldsymbol{\alpha}$ is the sparse self-attention matrix and each element is defined as follows:

$$\begin{cases} att_{i,j} = \frac{(\boldsymbol{c}_i \mathbf{W}_i)(\boldsymbol{c}_j \mathbf{W}_j)^T}{\sqrt{d}} * G_{i,j} \\ \alpha_{i,j} = \frac{exp(att_{i,j})}{\sum_{z \in N-1} exp(att_{i,z})} \end{cases}, \tag{9}$$

where d is the dimension of hidden states for scaling, $\mathbf{W}_i$ and $\mathbf{W}_j$ are the trainable parameters. To capture different representation subspaces, we also use multi-head attention (Vaswani et al., 2017).

Now, we define the sparse contextualized event representation as follows:

$$\boldsymbol{e}_i^{'} = \mathrm{ReLU}([\boldsymbol{e}_i, \boldsymbol{c}_i, \boldsymbol{c}_i^{'}]\mathbf{W}_e), \tag{10}$$

where $\mathbf{W}_e \in \mathbb{R}^{3d\times d}$ is the trainable parameters, and $\boldsymbol{e}_i^{'} \in \mathbb{R}^d$. Given each pair of events (e_i, e_j), we define their representation as follows[3]:

$$\boldsymbol{v}_{i,j} = \mathrm{ReLU}([\boldsymbol{e}_i^{'} + \boldsymbol{e}_j^{'}, |\boldsymbol{e}_i^{'} - \boldsymbol{e}_j^{'}||]\mathbf{W}_c), \tag{11}$$

where $\mathbf{W}_c \in \mathbb{R}^{2d\times d}$ is the trained parameters.

3.3 Discriminating Intra- and inter-sentential Reasoning

Section 3.2 defines event pair representations based on local and global contexts. In this section, DIR takes them as intra-sentential features, indicating that they have not considered the event pairs in other sentences to form a reasoning chain. To further obtain inter-sentential features for each pair of events, we first select supportive event pairs for each event pair and use GAU (Hua et al., 2022) for information fusion. Then, we combine two types of features with different weights to differentiate two types of reasoning — the relatively easy intra-sentential tasks can help cross-sentence relation identification without the loss of performance.

First, instead of using all event pairs as supports, we assume that only the pairs sharing at least one common event can contribute to the reasoning chain. For example, given four events (a, b, c, d) to predict the relation between (a, c), the event pairs (a, b) and (b, c) are clearly related, but (b, d) is clearly irrelevant. By stacking more layers, our proposed model can implicitly deal with longer reasoning chains, as any length of chains can be decomposed into several shorter chains. Take the chain $a \rightarrow b \rightarrow c \rightarrow d$ for predicting event pair (a, d) as an example, we indeed will use $a \rightarrow c$ and $c \rightarrow d$ as supportive evidence, while $a \rightarrow c$ shall be supported as illustrated above. This is, the longer chain has been decomposed into two sub-chains, modeled using multiple layers, and optimized jointly. Based on the assumption, we build a set of supportive event pairs for query (e_i, e_j), $\mathbf{T}^1 = [\boldsymbol{v}_{i,j}, \boldsymbol{v}_{i,1}, \ldots, \boldsymbol{v}_{N,j}]$. Note that we include the representations of query event pair $\boldsymbol{v}_{i,j}$. Next, we utilize GAU to conduct reasoning over the supportive set as follows:

$$\begin{cases} \mathbf{T}^2 = (\mathbf{U} \odot \mathbf{AV})\mathbf{W}_o \\ \mathbf{U} = \mathbf{T}^1\mathbf{W}_u, \mathbf{V} = \mathbf{T}^1\mathbf{W}_v, \mathbf{Z} = \mathbf{T}^1\mathbf{W}_z \\ \mathbf{A} = (\mathrm{ReLU}((\mathbf{ZW}_q)(\mathbf{ZW}_k)^T + b))^2 \end{cases}, \tag{12}$$

where $\mathbf{W}_o, \mathbf{W}_u, \mathbf{W}_v, \mathbf{W}_z, \mathbf{W}_q, \mathbf{W}_k$, and b are the trainable parameters, and $\mathbf{T}^2 = [\boldsymbol{v}_{i,j}^{'}, \boldsymbol{v}_{i,1}^{'}, \ldots, \boldsymbol{v}_{N,j}^{'}]$ are the output event pair representations enhanced by reasoning chains. We take $\boldsymbol{v}_{i,j}^{'}$ as the inter-sentential reasoning features of query (e_i, e_j).

Now, we are to combine two types of features with different weights. The basic idea is that event pairs within the same sentence is relatively easy to predict with high confidence, e.g., causal indicator words (e.g., *due to* or *lead to*) in Figure 1 provide clear patterns. We thus leverage intra-sentential features to facilitate the event pairs spanning different sentences. To avoid the harm of easier predictions, we assign higher weights to intra-sentential features if the event pair is within the same sentence. By contrast, we assign higher weights to inter-sentential features for events from different sentences to highlight the inter-sentential reasoning. Finally, the query event pair representations for relation between (e_i, e_j) are defined as follows:

$$\boldsymbol{o} = \begin{cases} \boldsymbol{v}_{i,j} + \beta_1\boldsymbol{v}_{i,j}^{'}, & if\ p - q < 0(\overrightarrow{inter}) \\ \boldsymbol{v}_{i,j} + \beta_2\boldsymbol{v}_{i,j}^{'}, & if\ p - q = 0(intra) \\ \boldsymbol{v}_{i,j} + \beta_3\boldsymbol{v}_{i,j}^{'}, & if\ p - q > 0(\overleftarrow{inter}) \end{cases}, \tag{13}$$

where $\beta_1, \beta_2, \beta_3$ are the weights that highlight different type of features for different event distribution. We determine these hyper-parameters by the heuristic experiments. Note that we separate the two cases of events in different sentences: $\overrightarrow{inter}$ and $\overleftarrow{inter}$. $\overrightarrow{inter}$ indicates that the event pairs are located in narrative order, while $\overleftarrow{inter}$ denotes a flashback order. The separation considers the different attention assumptions in Section 3.2, which captures various language biases. In experiments, we indeed found such a bias that $\overleftarrow{inter}$ has a high probability of negative relation, we thus set $\beta_3 = 0$ and $\beta_1 = 0.8, \beta_2 = 0.2$ heuristically, but this may be varying in different scenarios.

3.4 Classification

Given the final representation of an event pair, we use the linear function to predict relations as follows:

[3] For clarity, we delete the index of the sentence.

Model (%)	**Intra-sentence**			**Inter-sentence**			**Intra+Inter**		
	P	R	F1	P	R	F1	P	R	F1
LSIN (Cao et al., 2021)	47.9	58.1	52.5	-	-	-	-	-	-
KnowDis (Zuo et al., 2020)	39.7	66.5	49.7	-	-	-	-	-	-
LR+ (Gao et al., 2019)	37.0	45.2	40.7	25.2	48.1	33.1	27.9	47.2	35.1
LIP (Gao et al., 2019)	38.8	52.4	44.6	35.1	48.2	40.6	36.2	49.5	41.9
KMMG (Liu et al., 2020)	41.9	62.5	50.1	-	-	-	-	-	-
LearnDA (Zuo et al., 2021b)	42.2	69.8	52.6	-	-	-	-	-	-
RichGCN (Phu and Nguyen, 2021)	49.2	63.0	55.2	39.2	45.7	42.2	42.6	51.3	46.6
ERGO (Chen et al., 2022)	49.7	**72.6**	59.0	**43.2**	48.8	45.8	**46.3**	50.1	48.1
SENDIR	**65.8**	66.7	**66.2**	33.0	**90.0**	**48.3**	37.8	**82.8**	**51.9**

Table 1: Main results on the EventStoryLine. The best results are in **bold** and the second-best results are underlined. Intra-sentence denotes that the event pair is in the same sentence, and inter-sentence denotes that the event pair is in different sentences.

$$\hat{y} = \sigma(\boldsymbol{o}\mathbf{W} + b), \tag{14}$$

where $\mathbf{W}$ and b are the trainable parameters, $\hat{y}$ is the probability of being positive, and σ is an activation function. For training, we adopt cross-entropy as the loss function: $\mathcal{L} = -\sum_{e_i,e_j}(1 - y)\log(1 - \hat{y}) + y\log(\hat{y})$ (y is the golden). We use dropout to prevent overfitting.

4 Experiments

4.1 Datasets and Metrics

To demonstrate the performance of our model, we evaluate the model on two domains three datasets. EventStoryLine[4] (Mostafazadeh et al., 2016) and Causal-TimeBank[5] (Mirza and Tonelli, 2014) are event causal relation extraction (RE) dataset, and MATRES[6] (Ning et al., 2018) is event temporal RE dataset. And we use Precision (P), Recall (R), and F1-score (F1) as evaluation metrics.

EventStoryLine annotates 258 documents, 22 topics, $4,316$ sentences, $5,334$ event mentions, $7,805$ intra-sentential event pairs, and $46,521$ inter-sentential event pairs. Following (Gao et al., 2019), we put them in order based on their topic IDs. **Causal-TimeBank** (Causal-TB) annotates 184 documents, $6,813$ events, and $7,608$ event pairs. **MATRES** annotates 275 documents for four temporal relations, i.e., BEFORE, AFTER, EQUAL, and VAGUE.

4.2 Parameter Settings

We choose the most widely used BERT-base as basic PLMs, to avoid exhaustive parameter tuning. The learning rate is set to $1e^{-5}$ for pre-training and $1e^{-4}$ for others. We optimize our model with AdamW. We conduct the grid search to tune hyperparameters: the size of embedding is in $\{64; 128; 256; 512; \mathbf{768}\}$, where bold font denotes the best setup. The batch size for pre-trained model is set to 2. The dropout rate is set to 0.4.

4.3 Baseline

We compare SENDIR with state-of-the-art methods for Event Causal RE and Event Temporal RE.

Event Causal RE. (1) **KMMG** (Liu et al., 2020) that proposes a mention masking generalization and use external knowledge databases; (2) **KnowDis** (Zuo et al., 2020) that investigates a data augmentation to solve the data lacking; (3) **LSIN** (Cao et al., 2021) that employ ConceptNet to capture the latent causal relational structure; (4) **LearnDA** (Zuo et al., 2021b) that augments data to solve the data lacking; (5) **LR+** and **LIP** (Gao et al., 2019) that models rich causal structures via designing constraints and objection function; (6) **RichGCN** (Phu and Nguyen, 2021) that builds multi-level graphs to capture structure-preserved features; (6) **ERGO** (Chen et al., 2022) that designs an event relational graph and converts the event causal identify to a node classification framework.

Event Temporal RE. (1) **DEER** (Han et al., 2020) that constructs many training samples to simulate the machine reading comprehension for event temporal understanding; (2) **SMTL** (Ballesteros et al.,

[4] https://github.com/tommasoc80/EventStoryLine

[5] https://github.com/paramitamirza/CATENA/tree/master/data

[6] https://github.com/qiangning/MATRES

Model (%)		P	R	F1
Causal-TB	**KnowDis**	36.6	55.6	44.1
	RichGCN	39.7	56.5	46.7
	LearnDA	41.9	**68.0**	51.9
	ERGO	58.4	60.5	59.4
	SENDIR	**65.2**	57.7	**61.2**
MATRES	**DEER**	-	-	79.3
	SMTL	-	-	81.6
	TIMERS	81.1	84.6	82.3
	SCS-EERE	78.8	**88.5**	**83.4**
	SENDIR	**81.2**	85.6	83.3

Table 2: The results on the Causal-TB and MATRES.

2020) that extract the important text based on event pairs to identify relations. (3) **TIMERS** (Mathur et al., 2021) that uses rhetorical discourse features, temporal arguments, and syntactic features to extract the relation information. (4) **SCS-EERE** (Man et al., 2022) that seeks to identify the most important context sentences to identify the temporal relation.

4.4 Overall Performance

Table 1 and 2 show the overall performance on EventStoryLine, Causal-TB, and MATRES, respectively. We can see that: **(1)** SENDIR achieves better F1 scores on EventStoryLine and Causal-TB, it also has a competitive result on MATRES, which demonstrates the effectiveness and generalization ability of our model. **(2)** On the MATRES, SENDIR is slightly lower than SCS-EERE. Because event temporal RE is particularly sensitive to direction between events. **(3)** All models perform better on intra-sentence than inter-sentence in Table 1. This is consistent with our claim that intra-sentence is easier to identify. **(4)** Particularly, SENDIR has much higher precision on intra-sentence. Because the discriminative reasoning scheme alleviates the negative impacts of more difficult cross-sentence reasoning. **(5)** On inter-sentence setting, the improvements are mainly from higher recall. We attribute this to the enhanced long-distance modeling ability and the supportive query set — it tends to find relation clues from broader contexts and other event pairs.

4.5 Ablation Study

To further analyze SENDIR, we also conduct an ablation analysis to illustrate the effectiveness of our main modules. We show the results of the ablation study in Table 3[7].

Model (F1 %)	Intra	Inter	Intra+Inter
SENDIR	66.2	**48.3**	**51.9**
w/o **SER**	60.9	46.1	49.5
w/o **Sparse Att**	61.6	48.1	50.9
w/o **Event Repre**	**66.9**	46.2	50.2
w/o **DIR**	64.0	44.9	48.5
w/o **Selection**	66.0	46.5	50.5
w/o **Weight** (β)	63.9	45.7	49.3

Table 3: An ablation study for our model on the EventStoryLine.

SER. We examined the impacts of SER in Table 3. **w/o Sparse Att**, **w/o Event Repre**, and **w/o SER** denote that we gradually remove the key designs in Section 3.2: remove sparse attention mask, use local context only as event representations, and remove the entire SER module. **(1) w/o SER.** The performance becomes sharply poor without SER, especially on intra-sentence. Specifically, the experimental results are reduced by 5.3%/2.2%/2.4% F1 on intra-sentence, inter-sentence, and intra+inter. This demonstrates that SER can capture high-quality document representation, and intra-sentence event pairs rely more on high quality event representation, so intra-sentence drops more. **(2) w/o Sparse Att.** The experimental results are reduced. Specifically, the sparse attention mask has a more significant impact on intra-sentence, and this sparsity is the key to improving the quality of the document representation. We will do further analysis of multiple attention masks in Section 4.7. **(3) w/o Event Repre.** The main performance is the decline in inter-sentence, while intra-sentence is even slightly up. We attribute the reason that the inter-sentence relies more on global context and the intra-sentence relies more on local context. Moreover, the number of inter-sentence is much more than intra-sentence. Therefore, when inter-sentence drops, results of inter+inter drops, even though intra-sentence arises.

DIR. In Table 3, we can also observe the following insights. Note that **w/o DIR**, **w/o Selection**, and **w/o Weight** denote the deletion of our entire DIR module, taking all other event pairs as supports, and regarding intra- and inter-sentential reasoning the same by using the same weight 1. **(1) w/o DIR.** Compared with SER, removed DIR mainly brings a

[7] Intra and inter indicate intra-sentence and inter-sentence setting, respectively.

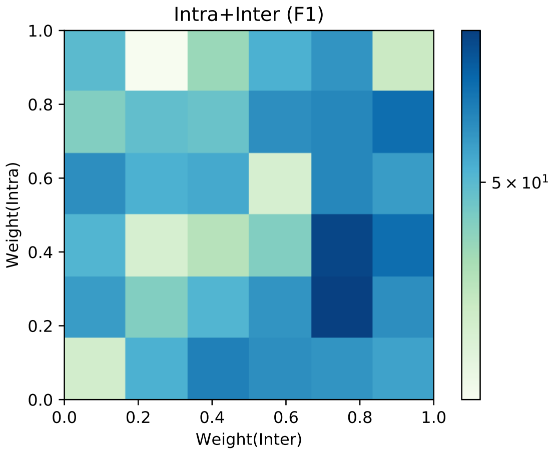


Figure 3: Impacts of the weight (β) on the EventStoryLine. Weight (Inter) is the weight of $\overrightarrow{inter}$.

decrease on inter-sentence, which is consistent with the method's motivation and validates the DIR's effectiveness. **(2) w/o Selection.** The experimental results are reduced. The major reason is that considering all event pairs as candidates will increase the noisy information. **(3) w/o Weight.** The drop on intra-sentence is very significant. Because of the lack of distinction in cross-sentence reasoning, the over-reasoning of simple event pairs leads to poor results. And using different weights facilitates simple tasks and global enhancements. We will follow up with a detailed discussion in Section 4.6.

4.6 Effects of Weights

As shown in Figure 3, we show the performance of Intra+Inter with different values for β_1 ($\overrightarrow{inter}$) and β_2 ($intra$). The darker the color, the higher the F1 value. We did not include β_3 due to the serious bias in datasets ($\beta_3 = 0$ always leads to better results), which makes further investigation trivial. The darker the color, the higher the F1 value. We can find that: **(1)** The scores in the bottom right part is better than in other parts, where the weight of inter-sentence increases and the weight of intra-sentence decreases. This agrees with our assumption that inter-sentential event pairs rely more on cross-sentence reasoning than intra-sentential event pairs. **(2)** In general, the performance of the right part is better than the left. There are two reasons: first, as the weight of inter-sentence increases, we highlight more cross-sentence reasoning, which improves the performance of inter-sentence; second, there are far more inter-sentence than intra-sentence, and the improvements of inter-sentential results will significantly improve the overall experimental performance. **(3)** The sub-diagonal elements indicate that intra-sentence and inter-sentence have the same weight. Clearly, the performance is unsatisfactory. The reason is that inter-sentential event pairs are usually difficult to predict and have lower confidence than intra-sentential event pairs. Thus, treating them equally negatively impacts the easier sentence-level tasks.

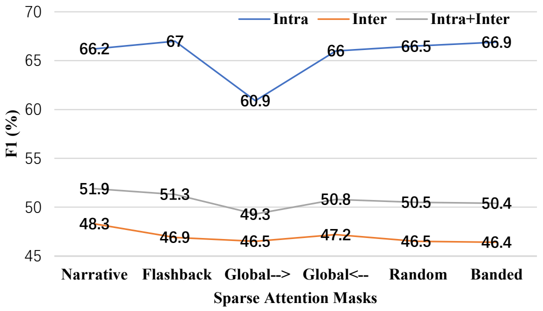


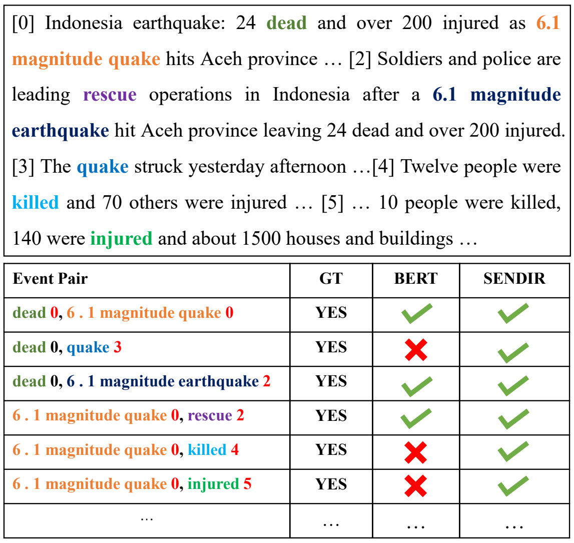
Figure 4: Impacts of Sparse Attention Mask on the EventStoryLine.

4.7 Effects of Sparse Attention Mask

To investigate the impacts of various sparse attention masks on the SER, we report the results using different sparse attention masks: Narrative, Flashback, Global→, Global←, Random, and Banded. From the Figure 4, we can find that: **(1)** On intra-sentence, these sparse attention masks have similar results except for Global→. This result is consistent with the previous results that event pairs rely more on local contexts rather than long-distance dependence on global contexts. **(2)** Random is unexpectedly good, indicating there is numerous redundant information in the document, and the sparse mask matrix can mitigate the effect of noise. **(3)** Narrative has achieved the best performance, which reflects a language bias from human writing habits — always talk about the main topic at first.

4.8 Case Study

As shown in Figure 5, we present a case study to better understand the effect of SENDIR compared to the baseline model (i.e., BERT). We can notice that: BERT and SENDIR can identify the intra-sentential event pair ($dead \xrightarrow{causal} 6.1\ magnitude\ quake$) and some easy inter-sentential event pairs ($dead \xrightarrow{causal} 6.1\ magnitude\ earthquake$ and $6.1\ magnitude\ quake \xrightarrow{causal} rescue$). However, for some complex event pairs that span multiple sentences, SENDIR can employ DIR to infer



[0] Indonesia earthquake: 24 dead and over 200 injured as 6.1 magnitude quake hits Aceh province … [2] Soldiers and police are leading rescue operations in Indonesia after a 6.1 magnitude earthquake hit Aceh province leaving 24 dead and over 200 injured. [3] The quake struck yesterday afternoon …[4] Twelve people were killed and 70 others were injured … [5] … 10 people were killed, 140 were injured and about 1500 houses and buildings …

Event Pair	GT	BERT	SENDIR
dead 0, 6 . 1 magnitude quake 0	YES	✓	✓
dead 0, quake 3	YES	✗	✓
dead 0, 6 . 1 magnitude earthquake 2	YES	✓	✓
6 . 1 magnitude quake 0, rescue 2	YES	✓	✓
6 . 1 magnitude quake 0, killed 4	YES	✗	✓
6 . 1 magnitude quake 0, injured 5	YES	✗	✓
…	…	…	…

Figure 5: The case study of our proposed SENDIR and BERT models on EventStoryLine, where GT denotes ground truth. Red numbers are the sentence numbers.

causal relations, but BERT fails, such as $dead \xrightarrow{causal} quake$ and 6.1 $magnitude\ quake \xrightarrow{causal} injured$. These demonstrate the significant advantages of SENDIR in dealing with inter-sentential event pairs to identify hard-to-identify causal relations by cross-sentence reasoning.

5 Conclusion

In this paper, we exploit a novel Discriminative Reasoning with Sparse Event Representation for DERE. It can learn high-quality event representation and facilitate inter-sentential reasoning for document-level understanding. Experimental results show that our method is effective and significantly better than competitive baselines, improving inter-sentence cases without harming intra-sentence event pairs. The extensive analysis also provides interesting insights about various language biases for sparse long-text representation learning. In the future, we will combine various sparse assumptions for high-quality document representations and incorporate graph reasoning.

6 Limitation

The limitations of SENDIR include the following two points: (1) It has not extended to document-level entity-centric relations tasks. Our work is event-centric, and future work extends it with entity-centric cases. Document-level entity-centric RE needs to consider multiple mentions of an entity and different relations in different directions of the same entity pair. (2) It does not bring in external commonsense knowledge. Knowledge can be used to enrich events and improve the accurate ERE.

Acknowledgements

We appreciate the comments from anonymous reviewers which will help further improve our work.This study is supported by National Natural Science Foundation of China (No.U19B2020), Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant, and Beijing Institute of Technology Southeast Academy of Information Technology.

References

Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen R. McKeown, and Yaser Al-Onaizan. 2020. Severing the edge between before and after: Neural architectures for temporal ordering of events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5412–5417. 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 ACL/IJCNLP*, pages 4862–4872. Association for Computational Linguistics.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: event relational graph transformer for document-level event causality identification. *CoRR*, abs/2204.07434.

Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. Uhop: An unrestricted-hop relation extraction framework for knowledge-based question answering. In *Proceedings of NAACL-HLT*, pages 345–356.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. 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 NAACL-HLT*, pages 1808–1817. Association for Computational Linguistics.

Rujun Han, Xiang Ren, and Nanyun Peng. 2020. DEER: A data efficient language model for event temporal reasoning. *CoRR*, abs/2012.15283.

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from wikipedia. In *Proceedings of EMNLP-IJCNLP*, pages 2986–2997. Association for Computational Linguistics.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2022. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of ACL*. The Association for Computer Linguistics.

Zhichao Hu, Elahe Rahimtoroghi, and Marilyn A. Walker. 2017. Inference of fine-grained event causality from blogs and films. In *Proceedings of the Events and Stories in the News Workshop@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 52–58. Association for Computational Linguistics.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc V. Le. 2022. Transformer quality in linear time. *CoRR*, abs/2202.10447.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019a. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of EMNLP-IJCNLP*, pages 5815–5821. Association for Computational Linguistics.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019b. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of EMNLP-IJCNLP*, pages 5816–5822, Hong Kong, China. Association for Computational Linguistics.

Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of IJCAI*, 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. *CoRR*, abs/1907.11692.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11058–11066. AAAI Press.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad I. Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. TIMERS: document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 524–533. Association for Computational Linguistics.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of EACL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2097–2106. ACL.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James F. Allen, and Lucy Vanderwende. 2016. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events, EVENTS@HLT-NAACL 2016, San Diego, California, USA, June 17, 2016*, pages 51–61. Association for Computational Linguistics.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6202–6208. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1318–1328. Association for Computational Linguistics.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of NAACL-HLT*, pages 3480–3490. Association for Computational Linguistics.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Proceedings of ICSC*, pages 361–368. IEEE Computer Society.

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

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*, pages 4444–4451. AAAI Press.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The causal news corpus: Annotating causal relations in event sentences from news. *arXiv preprint arXiv:2204.11714*.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2021. Synthesizer: Rethinking self-attention for transformer models. In *Proceedings of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 10183–10192. PMLR.

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.

Kritika Venkatachalam, Raghava Mutharaju, and Sumit Bhatia. 2021. SERC: syntactic and semantic sequence based event relation classification. In *33rd IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2021, Washington, DC, USA, November 1-3, 2021*, pages 1316–1321. IEEE.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2023. Document-level relation extraction with path reasoning. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(4):1–14.

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. SIRE: separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 524–534. Association for Computational Linguistics.

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 the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2162–2172. 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 ACL/IJCNLP*, pages 3558–3571. 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 COLING*, pages 1544–1550. International Committee on Computational Linguistics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section 5 and Section 6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.