Incorporating Schema-Aware Description into Document-Level Event Extraction

Zijie Xu[1] , **Peng Wang**[1,2*] , **Wenjun Ke**[1,2] , **Guozheng Li**[1] , **Jiajun Liu**[1] ,
Ke Ji[1] , **Xiye Chen**[3] and **Chenxiao Wu**[1]
[1]School of Computer Science and Engineering, Southeast University
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
[3]Nanjing University of Finance and Economics
{zijiexu, pwang, kewenjun, gzli, jiajliu, keji}@seu.edu.cn

Abstract

Document-level event extraction (DEE) aims to extract the structured event information from a given document, facing two critical challenges: (1) event arguments always scatter across sentences (*arguments-scattering*); (2) multiple events can co-occur in one document (*multi-event*). Most recent studies mainly follow two simplified settings to ease the challenges: one simplifies DEE with the no-trigger-words design (NDEE), and the other focuses on event argument extraction (DEAE), a sub-task of DEE. However, the former excludes trigger extraction and suffers from error propagation in the sub-tasks. The latter relies heavily on the gold triggers as prerequisites and struggles to distinguish multiple arguments playing the same role in different events. To address the limitations above, we propose a novel joint trigger and argument extraction paradigm SEELE to enhance the DEE model via incorporating **S**ch**E**ma-awar**E** descriptions into Document-**L**evel **E**vent extraction. Specifically, the schema-aware descriptions are leveraged from two aspects: (1) guiding the attention mechanism among event-aware tokens across sentences, which relieves *arguments-scattering* without error propagation; (2) performing the fine-grained contrastive learning to distinguish different events, which mitigates *multi-event* without gold triggers. Extensive experiments show the superiority of SEELE, achieving notable improvements (2.1% to 9.7% F1) on three NDEE datasets and competitive performance on two DEAE datasets. Our code is available at https://github.com/TheoryRhapsody/SEELE.

1 Introduction

Event extraction (EE) [Deng *et al.*, 2015] aims to extract event triggers and corresponding arguments from natural language texts, facilitating various downstream applications, such as information retrieval [Li *et al.*, 2023], recommender systems [Lu *et al.*, 2016], and question answering [Liu *et al.*, 2023] in finance, healthcare, and law industries.

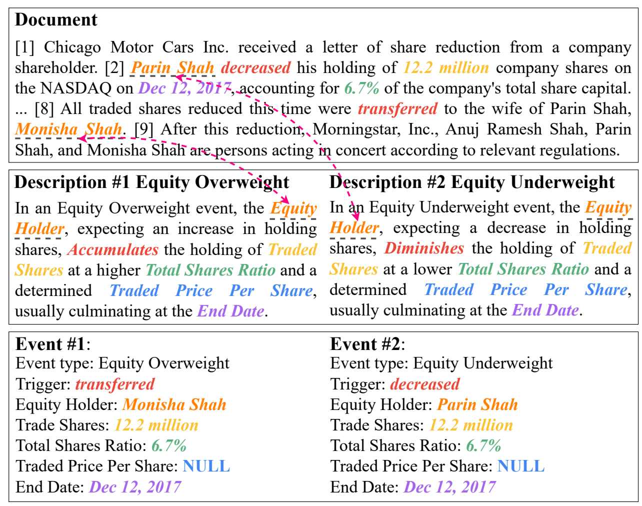

Figure 1: An example of a document containing *Equity Overweight* and *Equity Underweight* events and the corresponding schema-aware descriptions. [·] denotes the sentence numbering.

In real-world scenarios, events are usually distributed in several sentences, which raises two critical challenges for document-level event extraction (DEE) [Zheng *et al.*, 2019]: (1) *Arguments-scattering*, an event contains a cluster of arguments that may scatter across multiple sentences. As shown in Figure 1, the arguments of Event #2, *Equity Overweight*, are scattered in sentences [2] and [8]. Therefore, it is vital to model long-distance dependency among these arguments. (2) *Multi-event*, a document is likely to contain various events. For example, the document in Figure 1 includes two events, *Equity Overweight* and *Equity Underweight*, sharing several common arguments. Thus, identifying each event and which arguments participate in the same event is non-trivial.

Although enormous efforts have been devoted to DEE, most follow two simplified settings to ease the above challenges. One is *No-trigger-words DEE* (NDEE) [Zheng *et al.*, 2019; Xu *et al.*, 2021; Wang *et al.*, 2023], which reformalizes DEE as a table-filling task [Li *et al.*, 2024] with the no-trigger-words design. NDEE allows decomposing DEE

*Corresponding author

into a set of relatively simple sub-tasks to ease the *arguments-scattering* challenge: (1) extracting entities in the document as the candidate arguments; (2) modeling long-distance dependency among candidate arguments; (3) detecting the types of events and classifying the roles of the candidate arguments for each detected event type. However, errors exist in entity extraction [Xu *et al.*, 2021], leading to error propagation in each sub-task. The other is *Document-level Event Argument Extraction* (DEAE) [Li *et al.*, 2021; Xu *et al.*, 2022; Ma *et al.*, 2022], which aims to extract arguments of the target event with a given gold trigger. Since the triggers are given as prerequisites, there is no need to identify which events occurred in the document, alleviating the *multi-event* challenge. However, DEAE methods are trained to extract the arguments for one target event at a time, neglecting the correlation and discrimination of different events. Thus, DEAE methods struggle to distinguish different arguments that participate in multiple events and play the same argument role. As shown in Figure 1, *Parin Shah* and *Monisha Shah* are both *Equity Holder* but participate in different events. In addition, both NDEE and DEAE are inappropriate to extract triggers and arguments simultaneously in practical applications.

Intuitively, the event schema predefines each event type and corresponding argument roles, which can instruct DEE models to learn the scattering event information and the characteristics of different events [Shang *et al.*, 2024]. Therefore, we propose a novel learning paradigm **SEELE** for joint document-level event trigger and argument extraction (JDEE) via schema-aware descriptions constructed based on the event schema. Figure 1 demonstrates an example of the *Equity Overweight* and *Equity Underweight* events with the corresponding schema-aware descriptions. Note that the schema-aware description is natural language text containing the complete label information of the specific event type, which can serve as the guideline for understanding the DEE task.

Specifically, to tackle the *arguments-scattering* challenge, each schema-aware description is applied to query the tokens semantically relevant to the event conforming to this schema, namely event-aware tokens. We devise a description-guided attention mechanism to model the long-distance dependency among these event-aware tokens adaptively. However, the description may query some distracting entities. For example, the entities *Morningstar, Inc.* and *Anuj Ramesh Shah* in Figure 1 are also equity holders but do not participate in any event. Thus, we further assign a trainable query to filter event-distracting tokens via a gradient reverse layer (GRL) [Ganin and Lempitsky, 2015]. To relieve the *multi-event* challenge, each schema-aware description serves as the prototype of a specific event type to enhance the semantic discrimination of different events via fine-grained span-level contrastive learning. For *Parin Shah* and *Monisha Shah* in Figure 1, their representations are encouraged to be similar to the corresponding argument role Equity Holder in Description #1 and Description #2, respectively, but to be dissimilar from each other. Finally, for effective and efficient JDEE, we devise a novel event complete graph decoding strategy, where the event trigger and all arguments within each event are connected via undirected edges, naturally forming a complete graph. Then, all events can be decoded from the graph in a parallel process.

In summary, the main contributions of this work are threefold:

- We argue that schema knowledge is vital to understanding the structural event information from the document and propose a new JDEE paradigm that leverages schema-aware descriptions to enhance DEE models.
- We devise a new description-guided attention mechanism to model the long-distance dependency and a fine-grained contrastive learning to distinguish various events, alleviating two challenges of DEE.
- We conduct extensive experiments on five DEE benchmarks, demonstrating that our model achieves notable improvements (2.1% $\sim$ 9.7% F1) on NDEE benchmarks and competitive performance on DEAE benchmarks.

2 Related Work

No-trigger-words DEE. Early studies [Zheng *et al.*, 2019] formulate DEE as an event schema table-filling task where an event is defined by an event type and a set of arguments without triggers. To fully utilize all entities in the document, Doc2EDAG [Zheng *et al.*, 2019] and GIT [Xu *et al.*, 2021] adopt an entity-based directed acyclic graph. DE-PPN [Yang *et al.*, 2021] extracts all events in parallel. ReDEE [Liang *et al.*, 2022] further models the relation information between entities. PTPCG [Zhu *et al.*, 2022] proposes a graph-based decoding strategy. ProCNet [Wang *et al.*, 2023] decodes events based on Hausdorff distance minimization. IPGPF [Huang *et al.*, 2023] proposes an iteratively parallel generation method with the pre-filling strategy. However, these NDEE methods follow a pipeline paradigm, leading to error propagation.

Document-level Event Argument Extraction. DEAE is a challenging sub-task of DEE. The span-based methods predict the argument role for candidate spans in the document. For instance, TSAR [Xu *et al.*, 2022] introduces an AMR parser to identify candidate spans and devises an AMR interaction graph. PAIE [Ma *et al.*, 2022] and TabEAE [He *et al.*, 2023] leverage slotted prompts to obtain candidate spans for each argument role. The generation-based methods formulate DEE as a sequence-to-sequence task. For example, BART-Gen [Li *et al.*, 2021] leverages generative PLMs to generate arguments of the target event. However, most of them ignore the correlation and discrimination of multiple events.

3 Methodology

3.1 Preliminaries

Task Definition. For DEE, an event schema $\mathcal{S}_t$ consists of a event type t and an argument role set $R_t = \{r_i\}_{i=1}^{R}$. Given a document $D = \{w_i\}_{i=1}^{|D|}$, it contains a set of mentioned events $\mathcal{E}$. For each event $e \in \mathcal{E}$ corresponds to a specific schema $\mathcal{S}_t$, DEE aims to detect its event type t, extract the trigger $trg \in D$ and the corresponding argument set $\{(arg, r_i)\}$, where $arg \in D$ is a text span representing an argument and r_i is the role that arg plays. To unify DEE scenarios with (or without) triggers, we treat the event trigger as a special argument, i.e., adding the pair (trg, t) to $\{(arg, r_i)\}$, where the event type t is the special argument role of the event trigger.

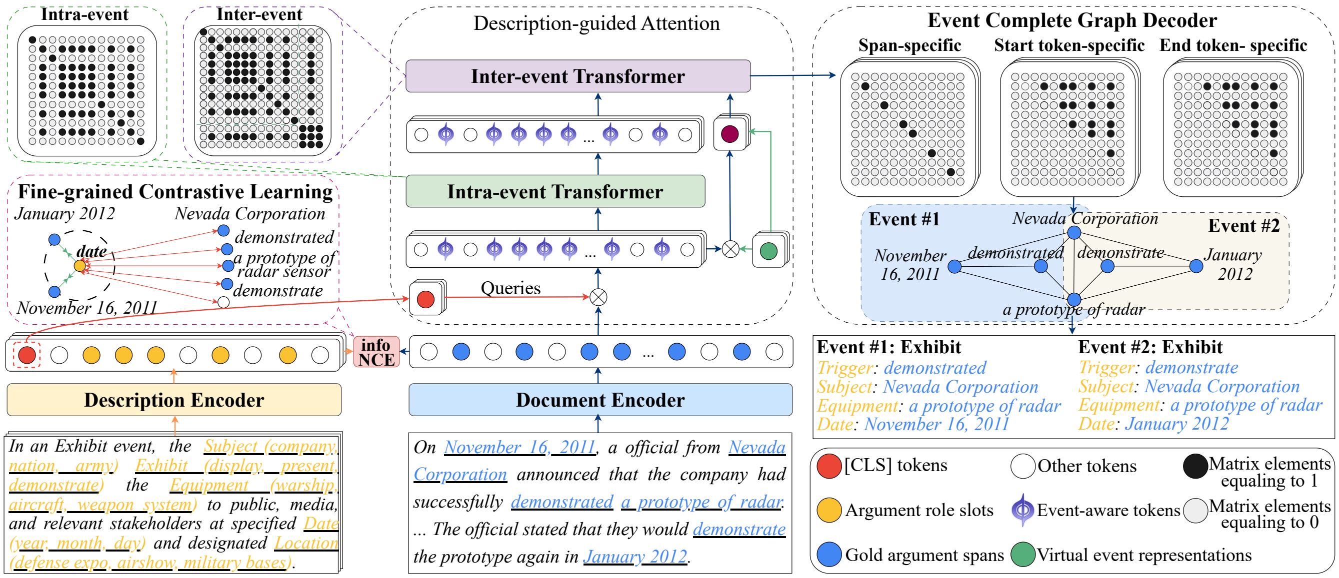

Figure 2: The overall architecture of SEELE, where underlined spans colored in yellow are argument role slots in the descriptions and underlined spans colored in blue are gold arguments in the given document containing two *Exhibit* events, $\otimes$ is the similarity computation.

Schema-aware Description Construction. A description $\mathcal{D}_t$ corresponds to an event schema $\mathcal{S}_t$, covering the argument role set R_t of the specific event type t, allowing the DEE model to capture the interactions among various argument roles. For example in Figure 2, given t as *Exhibit*, R_t as $\{$*Subject, Equipment, Date, Location*$\}$, the schema-aware description is constructed as follows:

In an Exhibit event, the Subject (company, nation, army) Exhibits (displays, presents, demonstrates) the Equipment (warship, aircraft, weapon system) to the public, media, and relevant stakeholders at specified Date (year, month, day) and designated Location (defense expo, airshow, military base).

$\mathcal{D}_t$ contains complete label information of $\mathcal{S}_t$, where the underlined spans serve as argument role slots. Each slot consists of two parts: the argument role r_i itself and several referential words in $(\cdot)$, which are the possible entity types of the arguments play r_i or the synonyms of t. The referential words are heuristically determined according to the statistics in the training dataset, aiming to enrich the semantics of labels.

3.2 Architecture Overview

The overall architecture is illustrated in Figure 2, containing three main components: description-guided attention mechanism, fine-grained contrastive learning, and event complete graph decoder. Given a document D and the schema-aware descriptions $\{\mathcal{D}_t\}_{t=1}^{T}$, SEELE first separately encodes D and $\mathcal{D}_t$ into fixed-length embedding sequences $\mathbf{H}_D = \{\mathbf{h}_i\}_{i=1}^{|D|}$ and $\mathbf{X}_t = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}_t|}$, where $\mathbf{h}_i \in \mathbb{R}^d$, $\mathbf{x}_i \in \mathbb{R}^d$ and d is the hidden state dimension. Then the [CLS] token from $\mathcal{D}_t$ is adopted to query the event-aware tokens for modeling the intra-event and inter-event long-distance dependency. Simultaneously, fine-grained contrastive learning aligns gold argument spans in D and argument role slots in $\mathcal{D}_t$ at span level, enhancing the representations of gold arguments. Finally, the event complete graph decoder forms predicted arguments into the complete graph and decodes each event from the graph.

3.3 Description-guided Attention Mechanism

Event-aware Token Querying. For each $\mathcal{D}_t$, we apply the [CLS] token embedding $\mathbf{x}_{[\text{CLS}]}$ of $\mathcal{D}_t$ as a query, document embedding $\mathbf{H}_D$ as the keys to obtain the event-aware tokens semantically relevant to the event conforming to schema $\mathcal{S}_t$:

$$\tilde{M}_t = \text{sigmoid}\left(\frac{\mathbf{x}_{[\text{CLS}]} \cdot \mathbf{H}_D{}^{\top}}{\sqrt{d}}\right) \tag{1}$$

where sigmoid $(\cdot)$ is a dot scaled similarity function to estimate the semantic similarity, and $\tilde{M}_t$ denotes the similarity matrix. We use a threshold Θ_1 to separate the event-aware tokens and derive a mask matrix M_t from $\tilde{M}_t$:

$$M_{t,i} = \begin{cases} 1, & \tilde{M}_{t,i} > \Theta_1 \\ -\infty, & Otherwise \end{cases} \tag{2}$$

Distracting Token Discarding. Considering that a document naturally contains more information than a single sentence, some distracting tokens or entities can be queried and mislead the DEE task. Intuitively, the tokens having negative effects on event extraction are probably distracting tokens. Therefore, we employ a learnable parameter $\mathbf{x}_q \in \mathbb{R}^d$ as the query of distracting tokens and optimize $\mathbf{x}_q$ via a separate event type classification task with a gradient reverse layer (GRL) [Ganin and Lempitsky, 2015]:

$$\mathbf{H}_q = \text{softmax}\left(\tfrac{\mathbf{x}_q \cdot \mathbf{H}_D{}^{\top} \odot M_q}{\sqrt{d}}\right) \cdot \mathbf{H}_D \tag{3}$$

$$P_q = \text{softmax}\left(\text{GRL}\left(\mathbf{H}_q\right) \cdot W_q + b_q\right) \tag{4}$$

$$\mathcal{L}_{ce} = \text{CrossEntropy}\left(Y_D, P_q\right) \tag{5}$$

where M_q is another mask matrix to filter the distracting tokens, $\odot$ is the Hadamard product, $\mathbf{H}_q \in \mathbb{R}^d$ is the distracting feature queried by $\mathbf{x}_q$, $W_q \in \mathbb{R}^{d \times T}$ and $b_q \in \mathbb{R}^T$ are the weights and biases of the event classifier, Y_D is the gold event type label. The GRL layer changes the gradient sign during backpropagation, which enables $\mathbf{x}_q$ to pay more attention to distracting tokens by gradient ascent to increase $\mathcal{L}_{ce}$.

Intra-event Attention. To model the long-distance dependency among event-aware tokens of one event, we apply an intra-event transformer layer to achieve the intra-event representation of the document D: $\mathbf{H}_D^{intra} = \mathrm{Transformer}^{intra}(\mathbf{H}_D)$. We devise a masked-attention matrix M^{intra} based on the two mask matrices M_t and M_q mentioned above to guarantee that two tokens $\mathbf{h}_i$ and $\mathbf{h}_j$ can only attend to each other when they are not distracting but event-aware tokens of one event:

$$\mathrm{Attention}(\mathcal{Q},\mathcal{K},\mathcal{V}) = \mathrm{softmax}\left(\frac{\mathcal{Q}\cdot\mathcal{K}^\top \odot M^{intra}}{\sqrt{d}}\right)\cdot\mathcal{V} \tag{6}$$

$$M_{i,j}^{intra} = \begin{cases} 1, & M_{t,i} = M_{t,j} = 1, M_{q,i} = M_{q,j} = -\infty \\ -\infty, & Otherwise \end{cases} \tag{7}$$

where $\mathcal{Q}$, $\mathcal{K}$, $\mathcal{V}$ refers to query, key, and value matrix of the intra-event transformer layer.

Inter-event Attention. To further model the correlation of different events, we employ T learnable parameters $\{\mathbf{v}_t\}_{t=1}^T$, and each of them serve as a virtual event representation under a specific schema $\mathcal{S}_t$ to integrate event-aware tokens into event-level features $\mathbf{H}_T = \{\mathbf{H}_t\}_{t=1}^T$:

$$\mathbf{H}_t = \mathrm{softmax}\left(\frac{\mathbf{v}_t\cdot{\mathbf{H}_D}^\top \odot M_t}{\sqrt{d}}\right)\cdot\mathbf{H}_D \tag{8}$$

We concatenate intra-event representation $\mathbf{H}_D^{intra}$ with $\mathbf{H}_T$ and apply another Transformer layer to perform information propagation among different events, denoted as $\mathbf{H}_D^{inter}, \mathbf{H}_T^{inter} = \mathrm{Transformer}^{inter}\left(\left[\mathbf{H}_D^{intra};\mathbf{H}_T\right]\right)$, where $\mathbf{H}_D^{inter}$ is the inter-event representation of document D. We devise another masked-attention matrix M^{inter} to ensure token $\mathbf{h}_i$ can only attend to the token $\mathbf{h}_j$ related to it. The constraints are as follows: (1) token $\mathbf{h}_i$ and $\mathbf{h}_j$ are event-aware tokens of one event; (2) token $\mathbf{h}_i$ and $\mathbf{h}_j$ are different event-level features; (3) token $\mathbf{h}_j$ is an event-level feature $\mathbf{H}_t$ and token $\mathbf{h}_i$ is the event-aware token integrated into $\mathbf{H}_t$.

3.4 Fine-grained Contrastive Learning

Paralleling the description-guided attention, a fine-grained contrastive learning is applied to align each argument in D with the corresponding argument role slot in $\mathcal{D}_t$, enhancing the semantic discrimination of various arguments.

We denote a text span representation in the document embedding $\mathbf{H}_D$ as $[\mathbf{h}_\alpha : \mathbf{h}_\beta]$, a contiguous embedding sequence with a start token $\mathbf{h}_\alpha$ and an end token $\mathbf{h}_\beta$. A gold argument representation of a specific argument role r_i is then denoted as $[\mathbf{h}_{r,\alpha} : \mathbf{h}_{r,\beta}]$. For each argument role slot in $\mathcal{D}_t$, we mean-pool its embedding sequence from $\mathbf{X}_t$ to obtain the slot representation $\mathbf{s}_r$ of argument role r_i. Intuitively, the text spans having the same start/end tokens of the gold argument $[\mathbf{h}_{r,\alpha} : \mathbf{h}_{r,\beta}]$ are potentially related to $\mathbf{s}_r$. Therefore, the optimization goal is to push the slot representation $\mathbf{s}_r$ close to the gold start token $\mathbf{h}_{r,\alpha}$ and the gold end token $\mathbf{h}_{r,\beta}$ (positive tokens), and far away from other distracting tokens (negative tokens). Specifically, we apply two linear layers to map $\mathbf{s}_r$ into the start/end representations $\boldsymbol{s}_{r,\alpha}$ and $\boldsymbol{s}_{r,\beta}$ as the anchors and devise the fine-grained infoNCE loss:

$$\mathcal{L}_n^{start} = -\log\frac{\exp\left(\mathbf{h}_{r,\alpha}\cdot\boldsymbol{s}_{r,\alpha}/\tau\right)}{\sum_{\mathbf{u}_\alpha\in\mathcal{U}_{r,\alpha}\cup\mathbf{h}_{r,\alpha}}\exp\left(\mathbf{u}_\alpha\cdot\boldsymbol{s}_{r,\alpha}/\tau\right)} \tag{9}$$

$$\mathcal{L}_n^{end} = -\log\frac{\exp\left(\mathbf{h}_{r,\beta}\cdot\boldsymbol{s}_{r,\beta}/\tau\right)}{\sum_{\mathbf{u}_\beta\in\mathcal{U}_{r,\beta}\cup\mathbf{h}_{r,\beta}}\exp\left(\mathbf{u}_\beta\cdot\boldsymbol{s}_{r,\beta}/\tau\right)} \tag{10}$$

$$\mathcal{L}_n = \left(\mathcal{L}_n^{start} + \mathcal{L}_n^{end}\right)/2 \tag{11}$$

where τ is a temperature hyper-parameter, $\mathcal{U}_{r,\alpha}$ and $\mathcal{U}_{r,\beta}$ are two sets of negative tokens that are not the start/end tokens of the gold arguments with the argument role r_i.

3.5 Event Complete Graph Decoder

Considering that events are structured information, it is vital to identify which arguments participate in the same event. Thus, we devise an event complete graph to explicitly model the event structure and decode each event from the graph.

Specifically, we utilize three Global Pointer Networks (GPs) [Su *et al.*, 2022] to extract each argument (or trigger) from the document and construct the event complete graph simultaneously, without any error propagation. Given the inter-event document representation $\mathbf{H}_D^{inter}$, the first GP is span-specific, adopted to identify which text span $[\mathbf{h}_\alpha : \mathbf{h}_\beta] \in \mathbf{H}_D^{inter}$ is a gold argument of a specific argument role r_i. The second GP is start token-specific, which aims to identify and connect two start tokens of any two gold arguments within one event. Similarly, the third GP is applied for two end tokens of any two gold arguments within one event. Compared with entities, event arguments are much sparser in a document. To optimize these three GPs, we apply Circle loss [Sun *et al.*, 2020] to alleviate label imbalance of gold arguments and negative spans:

$$\begin{aligned}\mathcal{L}_c = \log(1 + &\sum_{[\mathbf{h}_\alpha:\mathbf{h}_\beta]\in\mathcal{P}} \mathrm{e}^{-s([\mathbf{h}_\alpha:\mathbf{h}_\beta])}) \\ + \log(1 + &\sum_{[\mathbf{h}_\alpha:\mathbf{h}_\beta]\in\mathcal{N}} \mathrm{e}^{s([\mathbf{h}_\alpha:\mathbf{h}_\beta])})\end{aligned} \tag{12}$$

$$\mathcal{L}_{gp} = \left(\mathcal{L}_c^{span} + \mathcal{L}_c^{start} + \mathcal{L}_c^{end}\right)/3 \tag{13}$$

where $s(\cdot)$ is the score function in GP [Su *et al.*, 2022], $\mathcal{P}$ denotes the set of gold argument spans, $\mathcal{N}$ denotes the set of negative spans that are not arguments.

In the inference phase, all arguments extracted by the first GP are connected via the second and third GP according to whether they participate in the same event, forming the event complete graph $\mathcal{G}_\mathcal{E} = (V, L)$. $V = \{([\mathbf{h}_\alpha : \mathbf{h}_\beta], r_i)\}$ is the extracted argument node set and $L = \{t\}_{t=1}^T$ is the edge set defined by the event type. Each event e in the document D corresponds to a complete subgraph $\mathcal{G}_e^{sub} \in \mathcal{G}_\mathcal{E}$. We apply Bron-Kerbosch algorithm [Bron and Kerbosch, 1973] to search each complete subgraph $\mathcal{G}_\mathcal{E}^{sub}$ until all events and their corresponding argument set $\{(arg, r_i)\}$ are decoded from $\mathcal{G}_\mathcal{E}$.

3.6 Overall Optimization

We apply a multi-task joint training strategy by combining the three training losses mentioned above, and the overall training objective is a weighted sum of all losses:

$$\mathcal{L} = \lambda_1\mathcal{L}_{ce} + \lambda_2\mathcal{L}_n + \mathcal{L}_{gp} \tag{14}$$

where λ_1 and λ_2 are applied to balance the optimization.

4 Experiments

4.1 Experiment Setup

Dateset. We conduct trigger extraction and argument extraction on five DEE datasets divided into two settings. **NDEE Benchmarks**: (1) ChFinAnn dataset [Zheng *et al.*, 2019] is a widely used financial dataset without trigger annotations. (2) DuEE-Fin dataset [Han *et al.*, 2022] is another classic financial DEE dataset. (3) FNDEE [1] is a recent military news dataset. **DEAE Benchmarks**: (1) RAMS [Ebner *et al.*, 2020] is a typical DEAE news dataset containing news articles from Reddit. (2) WikiEvents [Li *et al.*, 2021] is another commonly used DEAE dataset based on English Wikipedia articles. We follow the official train/dev/test split. Since DuEE-Fin has not released gold labels in the test set and the online evaluation does not cover trigger extraction, we follow previous work [Liang *et al.*, 2022; Wang *et al.*, 2023] that uses the development set as the test set and split 500 documents from the training set as the development set.

Evaluation Metric. For trigger extraction, we adopt the criteria defined in previous work [Li *et al.*, 2013]: A trigger is correctly extracted if its span and event type match those of a gold trigger. For argument extraction on NDEE benchmarks, we follow previous studies [Zheng *et al.*, 2019; Xu *et al.*, 2021]: For each predicted event, the most similar gold event is matched without replacement. Then, an argument in the predicted event is correctly extracted if its span and argument role match those of an argument in the gold event. For argument extraction on DEAE benchmarks, we follow previous work [Li *et al.*, 2021; He *et al.*, 2023] and adopt the typical DEAE metrics: (1) Argument identification (Arg-I), where an argument is correctly identified if its span and event type match any gold argument. (2) Argument classification (Arg-I), where an argument is correctly classified if its argument role is also correct. The metrics above use micro-averaged precision (P), recall (R), and F1 scores (F1).

Baselines. For fair and strictly consistent comparison, we compare SEELE with strong baselines in three categories. **Trigger Extraction Baselines**: To verify the performance of trigger extraction, we make comparisons with four trigger extraction baselines: **BERT-CRF** [Lafferty *et al.*, 2001], **EEQA** [Du and Cardie, 2020], **MLBiNet** [Lou *et al.*, 2021], and **EDM3** [Anantheswaran *et al.*, 2023]. **NDEE Baselines**: To evaluate argument extraction on NDEE benchmarks, several competitive models are taken into consideration: **Doc2EDAG** [Zheng *et al.*, 2019], **DE-PPN** [Yang *et al.*, 2021], **GIT** [Xu *et al.*, 2021], **PTPCG** [Zhu *et al.*, 2022], **ReDEE** [Liang *et al.*, 2022], **ProCNet** [Wang *et al.*, 2023], and **IPGPF** [Huang *et al.*, 2023]. **DEAE Baselines**: To assess argument extraction on DEAE benchmarks, we further choose several state-of-the-art DEAE baselines: **BART-Gen** [Li *et al.*, 2021], **EEQA** [Du and Cardie, 2020], **CUP** [Lin *et al.*, 2022], **PAIE** [Ma *et al.*, 2022], **TSAR** [Xu *et al.*, 2022], **TabEAE** [He *et al.*, 2023].

[1] https://www.datafountain.cn/competitions/638/datasets

Model	DuEE-Fin			FNDEE			RAMS			WikiEvents		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT-CRF	**95.4**	86.3	90.6	**83.8**	68.2	75.2	73.4	68.4	70.8	**71.4**	60.0	65.2
EEQA	88.7	93.1	90.8	77.4	80.1	78.7	71.0	73.5	72.2	70.3	69.2	69.7
MLBiNet	88.8	91.4	90.1	74.0	77.4	75.7	70.7	65.7	68.5	62.6	53.9	58.0
EDM3	87.7	**94.3**	90.9	73.6	80.0	76.5	71.6	71.0	71.3	60.8	60.6	60.7
SEELE	90.3	93.6	**91.9**	76.7	**87.1**	**81.6**	**73.7**	**78.4**	**75.6**	70.6	**77.2**	**73.8**

Table 1: Main results of trigger extraction on four datasets. The bold scores indicate the best results and underlined scores indicate the second best results.

Setting	Model	EF	ER	EU	EO	EP	Overall
NDEE	Doc2EDAG†	70.2	87.3	71.8	75.0	77.3	78.8
	DE-PPN†	73.5	87.4	74.4	75.8	78.4	79.9
	GIT†	73.4	90.8	74.3	76.3	77.7	80.3
	PTPCG	71.4	91.6	71.5	72.2	76.4	79.4
	ReDEE†	74.1	90.7	75.3	78.1	80.1	81.9
	ProCNet†	75.7	**93.7**	76.0	72.0	81.3	83.0
	IPGPF‡	73.6	93.0	76.1	74.8	80.9	81.3
JDEE	SEELE	**78.8**	92.0	**77.7**	**82.4**	**83.1**	**85.1**

Table 2: Main results (F1) of argument extraction of 5 event types on ChFinAnn. **Overall** denotes the overall performance of all event types. † means the results from [Wang *et al.*, 2023]. ‡ means the results from [Huang *et al.*, 2023].

4.2 Main Results

Table 1 illustrates the main results of trigger extraction on four benchmarks. Though SEELE is designed for JDEE, it consistently outperforms other trigger extraction baselines by 1.0% $\sim$ 4.1% F1, including the document-level event detection methods MLBiNet and EDM3. Compared with the second-best EEQA, SEELE leverages schema-aware descriptions to enhance trigger extraction, which is more effective than the natural questions devised in EEQA.

Table 2 demonstrates the main results of argument extraction on ChFinAnn, a classical NDEE dataset without trigger annotations. SEELE achieves an absolute improvement in the overall performance by 2.1% F1, compared with other NDEE baselines. For specific event types, the F1 scores are significantly improved by 3.1% and 4.3% on Equity Freeze (EF) and Equity Overweight (EO) events, respectively. Most events of these two types involve more sentences than others, aggravating the *arguments-scattering* problem. Such improvement verifies that schema knowledge contributes to modeling interactions of arguments scattered across the document.

As shown in Table 3, SEELE achieves state-of-the-art performance on most event types in DuEE-Fin and surpasses the previous best IPGPF in the overall performance by 4.0% F1. It can be seen that most NDEE methods decrease notably compared with the results on ChFinAnn (shown in Table 2), possibly due to more argument roles in DuEE-Fin (up to 92). On the contrary, SEELE exhibits distinct improvements concerning the SR, EP, CP, and SC events by 7.4% $\sim$ 7.9% F1. It demonstrates the effectiveness of schema-aware descriptions in enhancing the semantic discrimination of various argument roles. In addition, the latest state-of-the-art IPGPF delivers the best results on FI, GB, and BG events.

Setting	Model	CL	SR	EO	EA	FI	ER	EP	CP	GB	EL	BG	BI	SC	Overall
NDEE	Doc2EDAG†	51.5	50.7	52.9	50.6	63.2	83.7	61.7	64.8	40.1	78.3	51.2	60.0	51.2	63.4
	DE-PPN†	23.0	32.9	31.3	41.3	36.3	67.8	36.3	42.1	21.4	62.7	25.8	50.7	23.4	45.1
	GIT†	55.1	58.8	71.2	56.6	68.5	86.4	71.3	66.4	44.7	77.6	45.0	58.8	53.8	67.8
	PTPCG	39.8	58.3	46.0	51.3	66.1	87.7	69.0	62.2	39.0	76.0	46.4	62.3	47.5	66.0
	ReDEE†	56.7	68.2	56.6	58.9	76.7	90.6	77.8	75.0	53.4	81.2	49.9	72.2	56.6	74.4
	ProCNet†	60.5	69.3	68.2	69.8	79.0	89.2	76.9	77.4	63.5	85.0	50.0	76.0	56.9	75.6
	IPGPF	65.7	68.9	74.8	71.8	**85.5**	94.6	63.6	74.3	**74.0**	85.7	**64.9**	75.3	53.6	76.8
JDEE	SEELE	**66.5**	**76.7**	**75.5**	**72.1**	79.1	**94.8**	**85.3**	**84.8**	62.0	**86.2**	58.0	**76.6**	**64.8**	**80.8**

Table 3: Main results (F1) of argument extraction of 13 event types on the DuEE-Fin. † means the results from [Wang *et al.*, 2023].

Model	EP	MA	DE	SU	AC	EH	CO	IN	Overall
Doc2EDAG	53.4	26.2	30.5	40.2	53.3	47.9	44.0	50.6	43.9
DE-PPN	50.8	25.8	30.6	44.4	42.6	46.2	35.4	32.8	43.0
GIT	54.2	27.6	41.0	44.9	51.8	58.2	43.1	56.6	48.5
PTPCG	51.1	23.8	54.5	60.9	38.6	44.3	40.8	70.7	52.3
ReDEE	56.6	40.4	50.2	62.4	56.3	63.2	61.0	74.0	56.8
ProCNet	53.9	44.7	55.8	64.7	54.3	67.7	50.6	62.6	57.3
IPGPF	51.7	47.2	49.0	64.5	59.0	67.3	61.9	72.2	56.3
SEELE	**67.4**	**49.8**	**66.0**	**69.2**	**60.8**	**71.2**	**62.1**	**86.6**	**67.0**

Table 4: Main results (F1) of argument extraction of 8 event types on the FNDEE.

Model	ChFinAnn	DuEE-Fin	FNDEE
Doc2EDAG	89.0†	82.6	67.5
DE-PPN	89.8	81.6	67.2
GIT	89.1†	82.3	68.7
PTPCG	88.8	82.7	68.5
ReDEE	90.4	84.8	69.2
ProCNet	89.7	83.4	69.0
IPGPF	92.2	86.8	69.7

Table 5: Results (F1) for entity extraction of NDEE models on the three DEE benchmarks. † means the results from [Xu *et al.*, 2021].

Setting	Model	RAMS		WikiEvents	
		Arg-I	Arg-C	Arg-I	Arg-C
DEAE	EEQA†	48.7	46.7	56.9	54.5
	BART-Gen†	51.2	47.1	66.8	62.4
	CUP	55.8	51.5	69.4	65.6
	TSAR†	57.0	52.1	71.1	65.8
	PAIE†	56.8	52.2	70.5	65.3
	TabEAE†	57.3	**52.7**	71.4	66.5
JDEE	SEELE	**57.6**	52.1	**73.8**	**67.3**

Table 6: Main results of argument extraction on RAMS and WikiEvents datasets. † means the results from [He *et al.*, 2023].

The average number of these events is less than 50, resulting in less-than-satisfactory results. However, IPGPF devises an iterative generation method to utilize historical results with the pre-filling strategy, which eases the inadequate samples.

Note that SEELE consistently outperforms NDEE baselines on all event types in FNDEE and has significant improvement in the overall performance by 9.7%, as illustrated in Table 4. Such results are attributed to the error propagation in NDEE methods. All NDEE methods adopt CRF-based entity extraction to obtain candidate arguments. Thus, the performance is subject to the accuracy of entity extraction. FNDEE contains massive nested entities compared with ChFinAnn and DuEE-Fin, while the CRF-based approaches are not applicable to identifying the span boundary of these nested entities. In addition, the average length of all arguments in FNDEE is more than eight characters, and around 500 arguments contain over 30 characters, making it non-trivial to extract the candidate arguments. As shown in Table 5, the entity extraction performance of the NDEE models drops heavily on FNDEE. The errors of entity extraction are more than 30%, severely influencing the subsequent argument role classification and leading to unfavorable results.

Table 6 reports the main results of argument extraction conducted on RAMS and WikiEvents. Though the gold triggers specify which events are mentioned in the document for DEAE baselines, SEELE demonstrates competitive performance on RAMS and surpasses the latest state-of-the-art TabEAE on WikiEvents by 0.8% F1 score. It is worth noting that in RAMS, over 90% of the documents consist of only one event, which somewhat limits SEELE's performance since SEELE is specially designed to tackle the *multi-event*.

4.3 Further Analysis

Arguments-scattering. To verify the effectiveness of SEELE in extracting the scattering arguments, we follow the previous work [Xu *et al.*, 2021] dividing ChFinAnn into four equal-sized subsets I/II/III/IV with growing average number of sentences involved in the events. Meanwhile, we follow previous work [Ma *et al.*, 2022; Xu *et al.*, 2022] that split RAMS into five subsets according to the sentence distance between the argument and trigger (i.e., -2∼2). Figure 3 (left) demonstrates that SEELE outperforms the strongest NDEE baselines on all subsets, especially on the most challenging IV, by 4.4% F1. Figure 3 (right) shows that SEELE significantly improves the ability to extract the scattering arguments, yielding 3.9% and 8.3% improvement when the arguments are two sentences before and behind the triggers. It indicates that SEELE proficiently captures the long-distance dependency via the description-guided attention mechanism, which mitigates the *arguments-scattering* challenge.

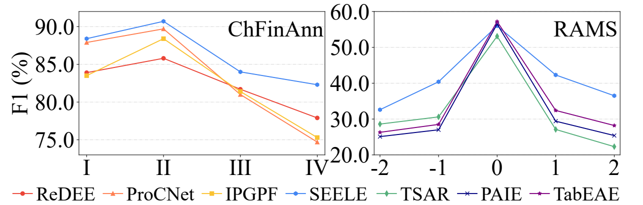

Figure 3: Comparison with the strongest baselines on ChFinAnn (left) and RAMS (right) in *arguments-scattering* scenario.

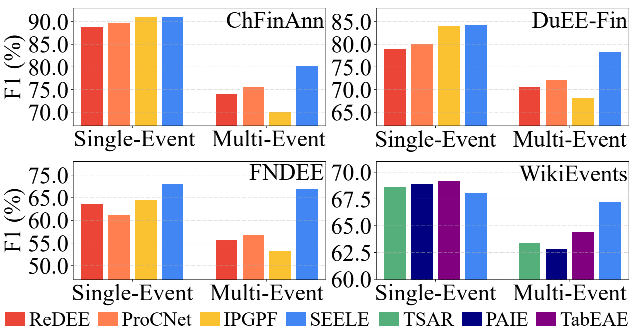

Figure 4: Comparison with the strongest baselines on four benchmarks in *multi-event* scenario.

Multi-event. To assess SEELE's performance in the *multi-event* scenario, we divide the test set of four benchmarks into the single-event set containing documents with only one event and the multi-event set containing multiple events. As shown in Figure 4, SEELE demonstrates competitive performance on the single-event set and dramatically outperforms all strongest baselines on the multi-event set, improving about 5%, 6%, 11% and 3% on ChFinAnn, DuEE-Fin, FNDEE, and WikiEvents, respectively. In addition, we find that SEELE is inferior to DEAE baselines on the single-event set of WikiEvents. It can be attributed to the fact that DEAE baselines do not predict redundant events (False-Positive) with given gold triggers, providing some advantage to the performance of DEAE baselines. Nevertheless, SEELE achieves state-of-the-art performances in *multi-event* scenario.

Same-Role. In *multi-event* scenario, multiple arguments participating in different events can play the same argument role. To evaluate SEELE's ability to distinguish the same-role arguments across different events, we split the test set of WikiEvents according to the number of arguments playing the same role. As demonstrated in Figure 5, SEELE achieves significant improvement ranging from 7.6% to 16.8%, especially in the most challenging set where the number of arguments is four or even more. Compared with the previous best TabEAE that models the event correlation via table generation, SEELE proves more effective in accurately extracting multiple arguments of the same role via fine-grained contrastive learning.

Ablation Study. We investigate the effectiveness of all key components in SEELE by removing each in turn. (1) **Intra-event Attention** (**Intra**), the transformer layer directly performs interaction among event-aware tokens from different events without intra-event attention. (2) **Inter-event Attention** (**Inter**), we only model the interactions of event-aware tokens within the same event. (3) **Token Discarding** (**TD**), the potential event-distracting tokens are retained in the intra-event and inter-event attention modules. (4) **Contrastive Learning** (**CL**), we drop the contrastive learning in the training phase. (5) **Schema-aware Description** (**SD**). We remove all components related to the schema-aware descriptions.

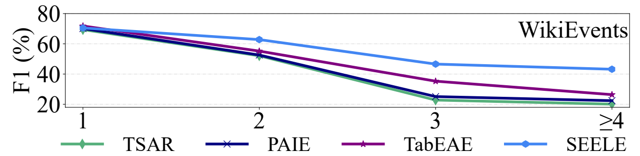

Figure 5: Comparison with the strongest baselines on WikiEvents in *same-role* scenario.

Model	ChFinAnn	DuEE-Fin		FNDEE		RAMS		WikiEvents	
	Arg	Trg	Arg	Trg	Arg	Trg	Arg	Trg	Arg
SEELE	85.1	91.9	80.8	81.6	67.0	75.6	52.1	73.8	67.3
w/o **Intra**	84.0	90.6	78.1	80.5	64.8	74.3	49.5	72.6	65.3
w/o **Inter**	83.8	91.0	79.2	81.2	64.4	75.2	51.6	73.1	64.8
w/o **TD**	83.5	91.4	79.5	80.2	65.3	74.7	50.3	72.9	66.1
w/o **CL**	82.9	84.6	75.6	74.3	63.0	69.3	47.8	66.9	61.6
w/o **SD**	80.9	85.2	72.1	72.8	56.7	68.9	45.4	65.5	61.2

Table 7: Ablation study results (F1) on five benchmarks. Trg and Arg denote trigger extraction and argument extraction.

Table 7 demonstrates the ablation experiments. We summarize the following observations: (1) Intra-event and inter-event attention modules have a vital effect on all the benchmarks, especially on ChFinAnn, where *arguments-scattering* issue widely exists. (2) The irrelevant token discarding module improves the performance by 0.9% to 1.8% on the five benchmarks. The results indicate that the irrelevant information in the context indeed influences the model's performance. (3) Significant impact lies in fine-grained span-level contrastive learning, and its absence results in a noticeable decrease in both trigger extraction and argument extraction, particularly on RAMS and WikiEvents. These benchmarks have more event types and argument roles than others, exacerbating the *multi-event* challenge. It demonstrates the significance of fine-grained contrastive learning, which aligns the unstructured documents with structured schema to enhance semantic discrimination of various events. (4) Removing all components related to the event description further causes a significant decline. It indicates the superiority of the SEELE, taking full advantage of the schema knowledge to boost the model's performance in complicated DEE scenarios.

5 Conclusion

In this work, we propose a new JDEE paradigm that incorporates schema-aware descriptions into DEE. We devise a novel description-guided attention mechanism and fine-grained contrastive learning, which mitigate two critical challenges of DEE. Extensive experiments on five benchmarks demonstrate the state-of-the-art performance and generalization ability of our method in various scenarios.

Acknowledgments

We thank the reviewers for their insightful comments. This work was supported by National Science Foundation of China (Grant Nos.62376057) and the Start-up Research Fund of Southeast University (RF1028623234). All opinions are of the authors and do not reflect the view of sponsors.

References

[Anantheswaran *et al.*, 2023] Ujjwala Anantheswaran, Himanshu Gupta, Mihir Parmar, Kuntal Kumar Pal, and Chitta Baral. Edm3: Event detection as multi-task text generation. *arXiv preprint arXiv:2305.16357*, 2023.

[Bron and Kerbosch, 1973] Coen Bron and Joep Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, sep 1973.

[Deng *et al.*, 2015] Jingsheng Deng, Fengcai Qiao, Hongying Li, Xin Zhang, and Hui Wang. An overview of event extraction from twitter. In *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 251–256. IEEE, 2015.

[Du and Cardie, 2020] Xinya Du and Claire Cardie. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[Ebner *et al.*, 2020] Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[Han *et al.*, 2022] Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. Duee-fin: A large-scale dataset for document-level event extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 2022.

[He *et al.*, 2023] Yuxin He, Jingyue Hu, and Buzhou Tang. Revisiting event argument extraction: Can eae models learn better when being aware of event co-occurrences? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

[Huang *et al.*, 2023] Guanhua Huang, Runxin Xu, Ying Zeng, Jiaze Chen, Zhouwang Yang, and Weinan E. An iteratively parallel generation method with the pre-filling strategy for document-level event extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[Lafferty *et al.*, 2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

[Li *et al.*, 2013] Qi Li, Heng Ji, and Liang Huang. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013.

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

[Li *et al.*, 2023] Guozheng Li, Peng Wang, and Wenjun Ke. Revisiting large language models as zero-shot relation extractors. In *Findings of EMNLP*, 2023.

[Li *et al.*, 2024] Guozheng Li, Wenjun Ke, Peng Wang, Zijie Xu, Ke Ji, Jiajun Liu, Ziyu Shang, and Qiqing Luo. Unlocking instructive in-context learning with tabular prompting for relational triple extraction. *arXiv preprint arXiv:2402.13741*, 2024.

[Liang *et al.*, 2022] Yuan Liang, Zhuoxuan Jiang, Di Yin, and Bo Ren. Raat: Relation-augmented attention transformer for relation modeling in document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

[Lin *et al.*, 2022] Jiaju Lin, Qin Chen, Jie Zhou, Jian Jin, and Liang He. Cup: Curriculum learning based prompt tuning for implicit event argument extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

[Liu *et al.*, 2023] Jiajun Liu, Peng Wang, Ziyu Shang, and Chenxiao Wu. Iterde: an iterative knowledge distillation framework for knowledge graph embeddings. In *AAAI*, 2023.

[Lou *et al.*, 2021] Dongfang Lou, Zhilin Liao, Shumin Deng, Ningyu Zhang, and Huajun Chen. Mlbinet: A cross-sentence collective event detection network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[Lu *et al.*, 2016] Di Lu, Clare Voss, Fangbo Tao, Xiang Ren, Rachel Guan, Rostyslav Korolov, Tongtao Zhang, Dongang Wang, Hongzhi Li, Taylor Cassidy, Heng Ji, Shih-fu Chang, Jiawei Han, William Wallace, James Hendler, Mei Si, and Lance Kaplan. Cross-media event extraction and recommendation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016.

[Ma *et al.*, 2022] Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. Prompt for extraction? paie: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

[Shang *et al.*, 2024] Ziyu Shang, Wenjun Ke, Nana Xiu, Peng Wang, Jiajun Liu, Yanhui Li, Zhizhao Luo, and Ke Ji. Ontofact: Unveiling fantastic fact-skeleton of llms via ontology-driven reinforcement learning. In *AAAI*, 2024.

[Su *et al.*, 2022] Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*, 2022.

[Sun *et al.*, 2020] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[Wang *et al.*, 2023] Xinyu Wang, Lin Gui, and Yulan He. Document-level multi-event extraction with event proxy nodes and hausdorff distance minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

[Xu *et al.*, 2021] Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[Xu *et al.*, 2022] Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang, and Zhifang Sui. A two-stream amr-enhanced model for document-level event argument extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

[Yang *et al.*, 2021] Hang Yang, Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Taifeng Wang. Document-level event extraction via parallel prediction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[Zheng *et al.*, 2019] Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019.

[Zhu *et al.*, 2022] Tong Zhu, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Min Zhang. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.