RoSE: A Role Correlation Structure-Enhanced Model for Multi-Event Argument Extraction

Geting Huang, Jilong Zhang, Kai Zhou, Zhang Yi, Xiuyuan Xu*

Sichuan University, Chengdu, Sichuan, China
huanggt@stu.scu.edu.cn, 2024223045117@stu.scu.edu.cn,
zhoukaics@stu.scu.edu.cn, zhangyi@scu.edu.cn, xuxiuyuan@scu.edu.cn

Abstract

Event co-occurrences have been proven effective for event argument extraction (EAE) in previous studies; however, few have considered intra- and inter-event role correlation. Since roles vary among different event types, event structure heterogeneity and overlap pose significant challenges to EAE. To address this issue, we propose a **Ro**le Correlation **S**tructure-**E**nhanced model for Multi-Event Argument Extraction (**RoSE**), capable of capturing both heterogeneity and overlap of event structures through modeling role correlation. The proposed RoSE model employs a joint context-prompts input, **ro**le-centric **g**raph-guided **e**ncoder (RoGE), and **ro**le-specific **i**nformation **f**usion (RoIF). The RoGE is designed to enhance the intra- and inter-event role correlation between prompts and their corresponding event contexts. The RoIF module utilizes intra-event role information to improve multi-event arguments extraction. Extensive experiments on four widely-used benchmarks (RAMS, WikiEvents, MLEE, and ACE05) demonstrate that our proposed approach achieves state-of-the-art performance, validating the effectiveness of incorporating both intra- and inter-event role correlation.

Code — https://github.com/huanggeting/RoSE

Introduction

Event Argument Extraction (EAE) is a key task within Information Extraction (Hobbs and Riloff 2010; Grishman 2015; Lou et al. 2023), identifying event arguments and role types for a given text or document (Li et al. 2023; He, Hu, and Tang 2023; Hei et al. 2025). By extracting critical information, EAE is closely related to natural language processing applications (Liu, Min, and Huang 2021; Berant et al. 2014; Zhang, Chen, and Bui 2020; Li et al. 2020). Figure 1a illustrates an EAE instance. For the event type *Life.Die*, triggered by "killed", the EAE task is to extract arguments: "checkpoint" as the *Place*, "soldiers" as the *Victim*. EAE involves heterogeneous role sets for different events, with overlap due to co-referential entities across roles: for example, the Victim in Life.Die and Target in Conflict.Attack are consistent and point to the argument "soldiers". This demonstrates that role-centric variations among events bring structural heterogeneity and overlap.

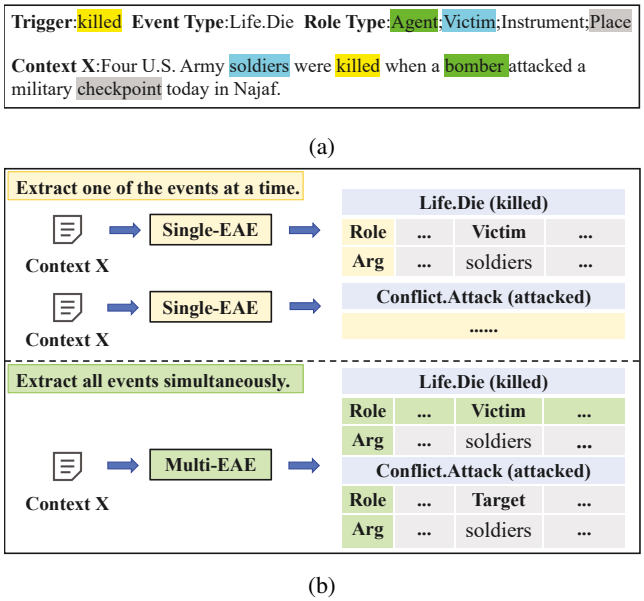

Figure 1: (a) presents the Event Argument Extraction (EAE) task. (b) illustrates the distinction between Single-EAE and Multi-EAE methods.

Recently, substantial progress has been reported on EAE. According to the workflow of event processing (Lin and Chen 2021; Ma et al. 2022; He, Hu, and Tang 2023; Liu et al. 2024; Hei et al. 2025), EAE methods can be categorized into two types: Single-EAE and Multi-EAE, as shown in Figure 1b. Single-EAE processes the context one event at a time, suffering from the following issues: numerous iterations leading to inefficient extraction; isolating each event but overlooking the role correlation among events. Multi-event argument extraction (Multi-EAE) extracts all event arguments in parallel. Benefiting from pre-trained language models (PLMs), Multi-EAE strengthens the correlation between prompts and context (He, Hu, and Tang 2023; Liu et al. 2024). However, it still faces the challenge that general prompts cannot capture the structural heterogeneity and overlap of different event types.

As shown in the above example, both structural heterogeneity and overlap fundamentally arise from the role-

*Corresponding author.

centric variation and intersections during the specific event argument extraction process. To tackle this issue, this paper proposes a **Ro**le correlation **S**tructure-**E**nhanced model for multi-EAE (**RoSE**), by injecting role correlation into the encoder and fusing the role-specific information during argument span prediction.

The RoSE model employs a joint context-prompts input, a role-centric graph-based encoder (RoGE), and a role-specific information fusion module (RoIF). We design the RoGE module to capture the role-centric heterogeneity and overlap among multi-events by injecting intra- and inter-event role correlation into the Transformer encoder. The RoIF utilizes role-specific information devised from RoGE to improve multi-event arguments extraction. Experiments on ACE05, RAMS, WikiEvents, and MLEE show that RoSE not only sets new state-of-the-art results but also validates that the explicit modeling of role correlation successfully mitigates both event heterogeneity and overlap. Our contributions can be summarized as follows:

- We introduce RoSE, a novel role correlation structure-enhanced model for Multi-EAE task, capable of capturing both heterogeneity and overlap of event structures.
- We propose a role-centric graph-based encoder (RoGE) to enhance the intra- and inter-event role correlation between prompts and their corresponding event contexts, and devise role-specific information fusion (RoIF) to improve multi-event arguments extraction.
- We conduct extensive experiments on four widely-used benchmarks (ACE05, RAMS, WikiEvents, and MLEE), demonstrating that RoSE achieves state-of-the-art performance while addressing the structural heterogeneity and overlap.

Related Work

Single-EAE

Previous works have generally classified the methods of Single-EAE into three categories: (1) Classification-based EAE, which casts EAE as a semantic role labeling problem (O'Gorman 2019; Zhang, Strubell, and Hovy 2022), or as a textual entailment (Lyu et al. 2021; Sainz et al. 2022) and a Question Answering (QA) task (Du and Cardie 2020; Lu et al. 2023; Hong and Liu 2024). (2) Sequence-to-sequence method, which casts EAE as a sequence generation task to extract all arguments (Hsu et al. 2022; Li, Ji, and Han 2021; Lu et al. 2021; Liu et al. 2022; Lin et al. 2025), applying special decoding strategies or guiding pretrained language models with prompts to produce conditional sequences. (3) Prompt-based method, which leverages slotted prompts and adopts prompt tuning to identify the start and end spans of arguments in a slot-filling manner (Ma et al. 2022; Li et al. 2023; Luo and Xu 2023; Zhang et al. 2024). However, all of them consider only one event at a time, ignoring the role correlation among events.

Multi-EAE

Limited by the inefficient inference of Single-EAE, existing studies (Wu, Zhang, and Li 2022; He, Hu, and Tang 2023; Liu et al. 2024) focus on Multi-EAE to extract multiple events in a document simultaneously, which allows for mutual interaction between different events. TabEAE (He, Hu, and Tang 2023) constructs the slotted table input based on given trigger(s) and prompt(s) and generates the results. To address the time-consuming problem in TabEAE, DEEIA (Liu et al. 2024) proposes a dependency-guided encoding method to solve the information complexity of multiple event prompts.

However, these Multi-EAE methods overlook the structural heterogeneity and overlap across events in a document. Specifically, events exhibit structural heterogeneity due to the varying role sets (Xu et al. 2021) associated with different events in the real world. Additionally, events of different types may share the same role entities, resulting in structural overlaps that are not explicitly captured (Chen et al. 2023; Hei et al. 2025). Therefore, we propose a **Ro**le correlation **S**tructure-**E**nhanced model for multi-EAE (**RoSE**), by injecting role-correlation into the encoder and fusing the role-specific information during argument span prediction to capture structural heterogeneity and overlap among multi-events.

Methodology

Task Definition

We first define the Multi-EAE task. Given an input instance represented as $\big(X, \{e_i\}_{i=1}^{K}, \{t_i\}_{i=1}^{K}, \{R^{(e_i)}\}_{i=1}^{K}\big)$, where $X = (x_0, x_1, \ldots, x_{N-1})$ denotes the context with N words, K is the number of events, and e_i is the i-th event type. The trigger $t_i \subseteq X$ is a span indicating the occurrence of event e_i, and $R^{(e_i)}$ is the set of argument roles defined for that event e_i. The goal is to extract a set of argument spans $A^{(e_i)}$ for each event e_i, which satisfies $\forall a^{(r)} \in A^{(e_i)},\ \big(a^{(r)} \subseteq X\big) \wedge \big(r \in R^{(e_i)}\big)$, where r denotes the associated role. Most existing EAE methods operate under the Single-EAE framework, applicable when $K = 1$.

Joint Context-Prompts Input

Most Single-EAE approaches typically associate each event instance with a single extraction prompt (Ma et al. 2022; Hsu et al. 2022; Hei et al. 2025). Recent works (He, Hu, and Tang 2023; Peng et al. 2024; Liu et al. 2024) demonstrated that pre-trained language models can simultaneously handle multiple prompts.

Therefore, following the designation in (Liu et al. 2024), we concatenate trigger-aware context $\hat{\boldsymbol{X}}$ and multiple event prompts $\hat{\boldsymbol{P}}$ as the final input of the model:

$$\begin{aligned} \hat{\boldsymbol{X}} &= x_0 \ \ldots \ \langle t_0\rangle\, t_0\, \langle /t_0\rangle \ \ldots \ \langle t_i\rangle\, t_i\, \langle /t_i\rangle \ \ldots \ x_{N-1}, \\ \hat{\boldsymbol{P}} &= \langle e_0\rangle E_0 \langle /e_0\rangle P_0 \ldots \langle e_i\rangle E_i \langle /e_i\rangle P_i \ldots . \end{aligned} \tag{1}$$

Given an input text $\boldsymbol{X} = x_0, x_1, \ldots, x_{N-1}$ and a set of event triggers, we insert a pair of trigger markers $(\langle t_i\rangle, \langle /t_i\rangle)$ around each trigger word $t_i \in \boldsymbol{X}$, where i denotes its order of appearance in the context. P_i is the prompt for the i-th event and E_i is the description of event type e_i, and we also insert a pair of type marker $(\langle e_i\rangle, \langle /e_i\rangle)$ around each type description E_i.

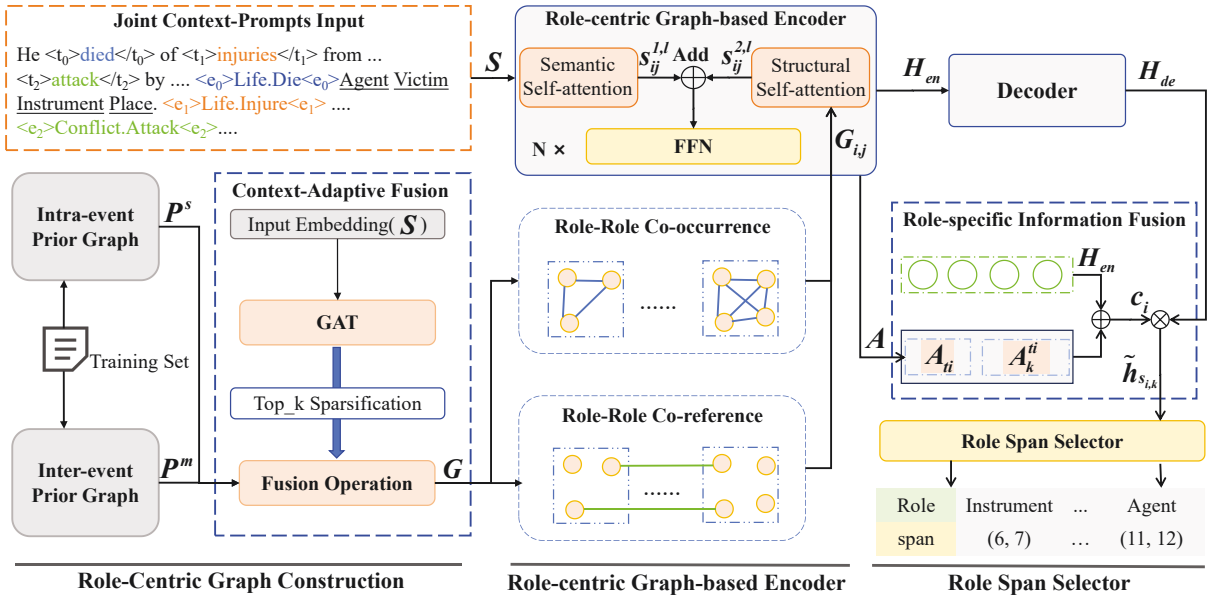

Figure 2: The architecture of the proposed RoSE model.

Role-centric Graph Construction

Prior Graph for Multi-Event

- **Intra-event Co-occurrence** The heterogeneity of events is primarily reflected in the diversity of their role sets. Additionally, the relationships among roles within each event further highlight this heterogeneity. To capture these relationships, we construct a local adjacency matrix, where the nodes represent roles within a specific event. As shown in Figure 2, we compute the co-occurrence frequency between roles for a given event type e in the training data, which serves as the matrix values:

$$\boldsymbol{P}^{s}_{r_i,r_j} = \frac{c(e, r_i, r_j)}{c(e, r_i)}, \tag{2}$$

where $\boldsymbol{P}^{s}_{r_i,r_j}$ denotes the co-occurrence edge weight between roles r_i and r_j in the matrix, $c(e, r_i, r_j)$ represents the frequency of co-occurrence between roles r_i and r_j within event e, and $c(e, r_i)$ indicates the frequency of the occurrence of role r_i within the event in the training set.

- **Inter-Event Co-reference** Since different roles across events may share the same argument entity within a document, we argue that this role correlation reflects the structural overlap between events. To capture this correlation, we construct a global adjacency matrix, where roles from all event types are represented as nodes. Specifically, we define the role-role co-reference as the edge weights in the graph:

$$\boldsymbol{P}^{m}_{r_i,r_j} = \frac{c(r_i, r_j)}{c(r_i)}. \tag{3}$$

Here, $c(r_i, r_j)$ denotes the number of arguments shared between roles r_i and r_j, and $c(r_i)$ represents the number of arguments associated with role r_i in the training set.

Context-Adaptive Fusion Rather than relying solely on the prior graph constructed by role correlation from the training set, we introduce a Graph Attention Network (GAT) (Veličković et al. 2018) and adopt a top-k sparsification(with k = 8) to complement the prior graph by learning soft, data-driven structural interactions among tokens. As shown in Figure 2, we define the process as follows:

$$\boldsymbol{G} = (1 - \gamma) \cdot \boldsymbol{GAT}(\boldsymbol{S}) + \gamma \cdot \boldsymbol{P}. \tag{4}$$

Here, $\boldsymbol{GAT}(\boldsymbol{S})$ denotes the learned attention-based structure from the input $\boldsymbol{S}$, and $\boldsymbol{P}$ is the prior graph combining $\boldsymbol{P}^s$ and $\boldsymbol{P}^m$. γ is a learnable fusion coefficient that balances the contributions of the data-driven GAT-learned structure and the prior role-correlation graph.

Role-centric Graph-guided Encoder

To integrate the role-centric graph into the base transformer, we innovatively design a graph-guided encoder, inheriting the architecture of the Transformer (Vaswani et al. 2017) encoder. Different from the original transformer, we adopt a dual self-attention mechanism: Semantic Self-attention and Structural Self-attention. Semantic Self-attention can capture contextual information among input tokens, and Structural Self-attention encodes role correlation guided by external graphs.

In each self-attention layer l, for the input token representation $h_i, h_j \in R^d$, we can get semantic attention scores $s^{1,l}_{ij}$ and structural attention scores $s^{2,l}_{ij}$:

$$\begin{cases} s^{1,l}_{ij} = \dfrac{(\boldsymbol{h}^l_i \boldsymbol{Q}^l)(\boldsymbol{h}^l_j \boldsymbol{K}^l)^T}{\sqrt{d_k}}, \\ s^{2,l}_{ij} = \dfrac{(\boldsymbol{h}^l_i \boldsymbol{Q}^l)(\boldsymbol{W}_{ij})(\boldsymbol{h}^l_j \boldsymbol{K}^l)^T}{\sqrt{d_k}} \cdot \boldsymbol{G}_{ij}, \end{cases} \tag{5}$$

where $\boldsymbol{Q}^l, \boldsymbol{K}^l \in R^{d \times d_k}$, d_k is the dimension of each attention head, $\boldsymbol{W}_{ij}$ is a trainable parameter, and $\boldsymbol{G}_{ij}$ is the token representation h_i, h_j dependency derived from the role-centric graph. We then compute the final attention score by adding the semantic and structural scores:

$$s_{ij}^l = s_{ij}^{1,l} + s_{ij}^{2,l}. \tag{6}$$

We apply the dual self-attention mechanism mentioned above to all layers of the Roberta-large encoder. By feeding joint input S into the role-centric graph-based encoder Encoder_s, we can obtain the encoding. And we decode the encoding with a simple decoder to obtain the decoding:

$$\begin{aligned} \mathbf{A}; \mathbf{H}_\text{en} &= \text{Encoder}_\text{s}(\boldsymbol{S}), \\ \mathbf{H}_\text{de} &= \text{Decoder}(\mathbf{H}_\text{en}), \end{aligned} \tag{7}$$

where $\mathbf{A} \in R^{H \times L \times L}$ is the multi-head attention matrix and $\mathbf{H}_\text{en}, \mathbf{H}_\text{de} \in R^{L \times d}$. H is the attention head numbers and L is the length of the input sequence $\boldsymbol{S}$.

Role-specific Information Fusion

We hope the model can make use of the event structural heterogeneity and overlap information when extracting the specific event arguments. Therefore, we design a role-specific information fusion module through fusing context information and specific role information. We consider using triggers and roles to capture the heterogeneity information between the target event and the context.

Specifically, we consider using triggers and roles to capture the heterogeneity information between the target event and the context. For the k-th role slots $s_{i,k}$ to be predicted for the i-th event, we first get $\mathbf{A}_{t_i} \in R^L$ and $\mathbf{A}_k^{t_i} \in R^L$ from $\mathbf{A}$, corresponding to the trigger t_i and the k-th role slot in the prompt respectively. Then for the role slot $s_{i,k}$, we obtain the context-enhanced vector $\boldsymbol{c}_i \in R^d$:

$$\begin{aligned} \boldsymbol{p}_k &= \text{softmax}\big(\mathbf{A}_{t_i} \cdot \mathbf{A}_k^{t_i}\big), \\ \boldsymbol{c}_i &= \mathbf{H}_\text{en}^T \boldsymbol{p}_k, \end{aligned} \tag{8}$$

where $\boldsymbol{p}_k \in R^L$ is the attention product for trigger t_i, role slots s_i. In detail, we use a gated fusion to incorporate the context and role representations. Given $\boldsymbol{c}_i$ and $\boldsymbol{h}_{s_{i,k}} \in R^d$ deriving from decoder output, we calculate the gate vector $\boldsymbol{g}_i$ with trainable parameters $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, then we get fused role representations $\tilde{h}_{s_{i_k}}$:

$$\begin{aligned} \boldsymbol{g}_i &= \text{sigmoid}(\boldsymbol{W}_1 \boldsymbol{h}_{s_{i,k}} + \boldsymbol{W}_2 \boldsymbol{c}_i), \\ \tilde{\boldsymbol{h}}_{s_{i,k}} &= \boldsymbol{g}_i \odot \boldsymbol{h}_{s_{i_k}} + (1 - \boldsymbol{g}_i) \odot \boldsymbol{c}_i. \end{aligned} \tag{9}$$

Role Span Selector

After obtaining roles representation $\tilde{\boldsymbol{h}}_{s_{i,k}}$ for the i-th event, we follow (Ma et al. 2022) and transform these representations into a set of span selectors $\{\boldsymbol{\Phi}_{s_{i,k}}^\text{start}, \boldsymbol{\Phi}_{s_{i,k}}^\text{end}\}$:

$$\begin{aligned} \boldsymbol{\Phi}_{s_{i,k}}^\text{start} &= \tilde{\boldsymbol{h}}_{s_{i,k}} \circ \boldsymbol{w}_\text{start}, \\ \boldsymbol{\Phi}_{s_{i,k}}^\text{end} &= \tilde{\boldsymbol{h}}_{s_{i,k}} \circ \boldsymbol{w}_\text{end}, \end{aligned} \tag{10}$$

where $\boldsymbol{w}_\text{start}, \boldsymbol{w}_\text{end} \in R^d$ are learnable parameters and $\circ$ represents element-wise multiplication. Then $(\boldsymbol{\Phi}_{s_{i,k}}^\text{start}, \boldsymbol{\Phi}_{s_{i,k}}^\text{end})$ is responsible for determining the span of k-th role slot for the i-th event in the context.

We adopt the loss function in (He, Hu, and Tang 2023; Liu et al. 2024) informed by Bipartite Matching Loss (Carion et al. 2020), which predicts the argument span through the Hungarian algorithm (Kuhn 1955).

Experiments

Implementation Details

We adopt the RoBERTa as our PLM, implement RoSE with Pytorch and run the experiments on a NVIDIA GeForce RTX 4090 GPU. Following the comparison conducted by (He, Hu, and Tang 2023; Liu et al. 2024), we use the initial 17 layers of RoBERTa as the encoder and the subsequent 7 layers as the decoder. We randomly initialize the cross-attention module in the decoder and assign it a learning rate 1.5 times higher than that of the remaining parameters. The model is optimized using the AdamW optimizer (Loshchilov and Hutter 2017), along with a linear learning rate scheduler. We use the prompts proposed in DEEIA (Liu et al. 2024).

Datasets

We evaluate our model on 4 EAE datasets, including ACE05 (Doddington et al. 2004), RAMS (Ebner et al. 2020), WikiEvents (Li, Ji, and Han 2021) and MLEE (Pyysalo et al. 2012). ACE05 is a sentence-level dataset, while the others are document-level datasets. ACE05, RAMS, and WikiEvents are drawn primarily from news articles, whereas MLEE is sourced from the biomedical domain. Moreover, nested events occur frequently in MLEE but are rare in the other three datasets.

Evaluation Metrics

Following previous works (Ma et al. 2022; Wang et al. 2025; Hei et al. 2025), we measure the performance with two metrics: (1) argument identification (Arg-I), requiring the predicted argument span of an event to fully match any golden arguments of the event; (2) argument classification (Arg-C), requiring the predicted argument of an event to both fully match boundary and role type. For all datasets, we adopt F1-score (F1) as the evaluation metric.

Baselines

We compare our RoSE with the following models, all of which evaluate the EAE performance on both sentence-level and document-level datasets: EEQA (Du and Cardie 2020), BART-Gen (Li, Ji, and Han 2021), PAIE (Ma et al. 2022), TabEAE (He, Hu, and Tang 2023), DEEIA (Liu et al. 2024), DEGAP (Wang et al. 2025), ERCL (He et al. 2025). Details about the baselines are listed in Appendix A.4. For the Multi-EAE baselines, we adopt PAIE-multi and TabEAE-multi in (He, Hu, and Tang 2023; Liu et al. 2024).

Main Results

Table 1 presents the overall performances of compared baselines and RoSE across all datasets. Based on the experiment results, we can observe that: (1) The performance of RoSE outperforms that of Single-EAE baselines. This

Scheme	Model	PLM	ACE05		RAMS		WikiEvents		MLEE	
			Arg-I	Arg-C	Arg-I	Arg-C	Arg-I	Arg-C	Arg-I	Arg-C
Single-EAE	EEQA (2020)	BART	70.5	68.9	48.7	46.7	56.9	54.5	68.4	66.7
	EEQA (2020)	RoBERTa	72.1	70.4	51.9	47.5	60.4	57.2	70.3	68.7
	BART-Gen (2021)	BART	69.9	66.7	51.2	47.1	66.8	62.4	71.0	69.8
	PAIE (2022)	BART	75.7	72.7	56.8	52.2	70.5	65.3	72.1	70.8
	PAIE (2022)	RoBERTa	76.1	73.0	57.1	52.3	70.9	65.5	72.5	71.4
	TabEAE (2023)	RoBERTa	77.2	75.0	57.3	52.7	71.4	66.5	72.0	71.3
	DEGAP (2024)	RoBERTa	76.6	74.4	58.5	54.2	72.2	67.1	74.0	73.4
	ERCL (2025)	BART	76.2	73.7	57.4	53.1	70.9	65.5	–	–
Multi-EAE	PAIE-multi	BART	-	-	55.9	50.9	67.2	61.7	71.3	69.5
	TabEAE-multi	RoBERTa	75.9	73.4	56.7	51.8	71.1	66.0	75.1	74.2
	DEEIA	RoBERTa	76.3	74.1	58.0	53.4	71.8	67.0	75.2	74.3
	RoSE (ours)	RoBERTa	**77.5**	**75.8**	**58.6**	**54.4**	**72.9**	**67.5**	**75.7**	**74.6**

Table 1: Main results on four benchmarks. Both RoBERTa and BART here are of large-scale (with 24 Transformer layers). Bold indicates the best experimental result.

demonstrates that our approach fully utilizes the heterogeneity information when extracting specific event's arguments from the context. (2) Compared with Multi-EAE baselines, our RoSE model outperforms the SOTA model by 1.2 Arg-I and 1.7 Arg-C on ACE05, by 0.6 Arg-I and 1.0 Arg-C on RAMS, by 1.1 Arg-I and 0.5 Arg-C on WiKiEvents, by 0.5 Arg-I and 0.3 Arg-C on MLEE. These results indicate that RoSE effectively exploits the overlap of structures when extracting all events' arguments while maintaining high efficiency. (3) Compared with DEEIA and TabEAE-multi, we observe that the improvement of our RoSE is around 1.2-2.4 F1 on ACE05, around 0.6-2.6 F1 on RAMS, around 0.5-1.8 F1 on WiKiEvents and around 0.3-0.6 F1 on MLEE. Therefore, the improvement of RoSE on ACE05, RAMS and WiKiEvent is more pronounced compared to MLEE. We hypothesize that the MLEE dataset contains less role-role coreference relationship than other datasets.

Model	ACE05	RAMS	WikiEvents	MLEE
RoSE	**75.8**	**54.4**	**67.5**	**74.6**
w/o RoGE	72.6	52.4	64.8	72.0
w/o RG	74.6	52.9	65.3	73.4
w/o RG_{intra}	74.8	53.0	66.3	73.8
w/o RG_{inter}	75.2	52.9	65.9	74.0
w/o CAF	74.0	53.0	65.8	73.8
w/o RoIF	74.2	53.1	65.6	73.8

Table 2: Ablation results on four datasets. Strict argument classification F1 scores (Arg-C) are reported. Bold indicates the best experimental results. The reported results are averaged from 6 different random seeds.

Ablation Study

Without Role-centric Graph-based Encoder (RoGE). We replace the role-centric graph-based Encoder(RoGE) module with a vanilla transformer encoder (Vaswani et al. 2017). As shown in Table 2, the performance reduction illustrates that RoGE module effectively provides role-centric structural guidance for multi-EAE.

- **Without Role-Role correlation.** To explore the role-role co-occurrence RG_{intra} and role-role co-reference RG_{inter} effectiveness in help model capture heterogeneity and overlap, we remove the role-role co-occurrence or role-role co-reference when prior graph construction. As in Table 2, the performance reduction for four datasets illustrates that both intra-event co-occurrence and inter-event co-reference relations can contribute to the model to capture the heterogeneity and overlap through the role-centric relation.

  And we can also find that different datasets exhibit varying dependencies on intra-event co-occurrence and inter-event co-reference. As shown in the Table 2, removing specific relation leads to different degrees of performance degradation, indicating that each dataset relies on different structural cues to varying extents.

- **Without Context-adaptive Fusion(CAF)** We further explore the effectiveness of context-adaptive fusion module (CAF). We replace the prior graph weights with zero, and attain the result in Table 2. Moreover, we remove the CAF module via directly adopting the hard prior graph to guide the graph-based encoder. As is shown in Table 2, it is observed that solely fusing the prior knowledge or context-adaptive dependency is less effective than combine them together.

Without Role-specific Information Fusion (RoIF). The performance of four datasets has significantly declined. This indicates that our RoIF module can provide beneficial role-specific information that contains event structural heterogeneity and overlap when extracting arguments.

Model	ACE05		RAMS		WikiEvents		MLEE	
	E = 1 [185]	E > 1 [218]	E = 1 [114]	E > 1 [251]	E = 1 [587]	E > 1 [284]	E = 1 [175]	E > 1 [2025]
PAIE-multi	-	-	51.75	48.94	65.01	60.18	-	-
TabEAE-multi	73.38	73.45	52.87	50.82	67.30	65.32	81.13	73.60
DEEIA	73.90	73.92	53.84	52.76	**67.49**	66.57	81.65	73.42
RoSE	**75.14**	**74.30**	**54.28**	**53.95**	66.72	66.95	**85.00**	**73.85**
RoSE-RG$_{intra}$	72.82	72.82	53.45	53.63	64.65	**68.81**	84.35	72.50
RoSE-RG$_{inter}$	74.32	71.78	53.32	52.82	66.43	64.93	84.26	72.86

Table 3: Performance comparison across datasets with varying event counts

Analysis

Analysis of Role Correlation

To evaluate our model's effectiveness in multi-event scenarios, we partition each test set by event number and perform separate evaluations. As shown in Table 3, RoSE mostly outperforms existing multi-event extraction methods in both single-event and multi-event settings. This demonstrates that our proposed model, capable of capturing event heterogeneity and overlap, strongly benefits event argument extraction.

We further ablate the two forms of role correlation. Removing intra-event co-occurrence leads to degraded performance in both single- and multi-event cases, indicating that our model addresses the event heterogeneity issue by modeling intra-event role correlation. Moreover, removing inter-event co-reference causes a more pronounced drop under multi-event conditions, confirming that our model solves the structural overlap problem by injecting inter-event role correlation.

Effect Analysis on Event Numbers

To further evaluate the effectiveness of our model, we divide the instances from datasets into different groups based on the event number. As illustrated in Figure 3, as the event number increases, we observe a decreasing trend in the performance of all models. However, RoSE maintains a smaller performance drop and outperforms others, especially on samples with four events, where competing models degrade significantly. All results above show the superiority of RoSE in capturing the event correlation among multiple events.

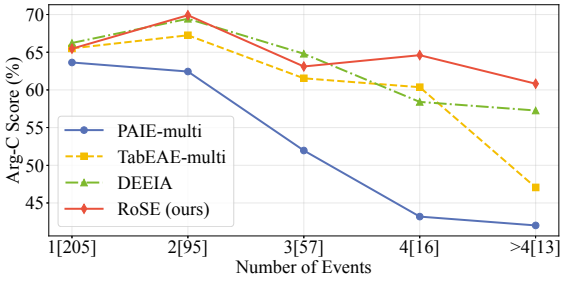

Figure 3: The averaged performance of the PAIE, TabEAE, DEEIA, RoSE models on samples with different event numbers in the WikiEvents dataset.

Analysis of Role-specific Information Fusion

To explore whether the role-specific information fusion module is effective for RoSE to extract the event arguments, we visualize the attention of RoSE and RoSE-RIF when extracting the event arguments, as shown in Figure 4. We find that when extracting Places for Conflict.Attack events, RoSE and RoSE-RoIF learn universal event knowledge through the self-attention mechanism, which makes the two variants pay higher attention to places, e.g., 'Iraq' and 'Fallujah'. Since RoSE also learns the role heterogeneity and overlap features, it is more capable of explicitly recognizing the roles of 'Fallujah', which is more accurate in judging the boundaries of the event argument.

Context: War in Iraq The US is currently engaged in major fighting in Fallujah, Iraq.

Event Type: Conflict.Attack **Trigger**: fighting **Role**: Place

(a)

Context: War in Iraq The US is currently engaged in major fighting in Fallujah, Iraq.

Event Type: Conflict.Attack **Trigger**: fighting **Role**: Place

(b)

Figure 4: Subfigure (a) is visualization of attentive weights without RoIF from an example in ACE05. Subfigure (b) is visualization of attentive weights with RoIF from an example in ACE05.

Compare with Large Language Models

Large language models (LLMs), empowered by their strong generalization and learning capabilities, have been widely applied across various tasks, such as the text-classification (Li, Liu, and Jiang 2023; Lee et al. 2024), dialogue systems (Mishra et al. 2023; Ye et al. 2025) and graph tasks (Ge et al. 2023; Xu et al. 2024; Lan et al. 2025a,b; Zhou et al. 2025b), including the Information Extraction tasks (Peng et al. 2023; Goel et al. 2023). To assess their effectiveness on event argument extraction (EAE), we specifically selected recent state-of-the-art LLM-based methods for comparison (Zhou et al. 2025a). The experimental results are summarized in Table 5. The experiment is

Method	Error Category					
	Wrong Span	Partial-less	Partial-more	Overlap	Miss	Over Extract
PAIE-multi	46	16	6	1	170	111
DEEIA	38	8	10	1	158	81
RoSE-RG	36	8	10	0	127	98
RoSE-RoIF	30	10	10	0	126	101
RoSE	27	10	14	0	116	86

Table 4: Error Analysis on WikiEvents test set.

Model	RAMS		WikiEvents	
	Arg–I	Arg–C	Arg–I	Arg–C
ChatGPT	46.2	40.4	42.4	42.4
ChatGLM	50.4	45.8	60.3	58.6
LLaMA	46.4	38.2	64.3	51.0
RoSE	58.6	54.4	72.9	67.5

Table 5: Comparison with large language model.

conducted on three prominent large language models: ChatGPT3.5, ChatGLM2-6B (Du et al. 2022) and LLaMA2-7b-chat (Touvron et al. 2023). The results in Table 5 reveal a clear performance gap between LLMs and supervised models in the EAE task. Although LLMs offer considerable flexibility, they often underperform and incur substantial inference overhead; in contrast, our method achieves higher accuracy, greater efficiency, and lower computational cost on multi-EAE.

Error Analysis

To explore the effectiveness of our method, we further conduct error analysis. We analyze all prediction errors on the WikiEvents test set and categorize them into five classes. A Wrong span refers to a predicted span that has no overlap with the gold span. A Partial match occurs when the predicted span and the gold span overlap partially, meaning one is a proper subset of the other. An Overlap indicates a non-partial but non-exact match, where the two spans intersect but do not fully contain one another. Over-extraction describes the case where a span is predicted despite the absence of a corresponding gold span.

As shown in Table 4, compared to the Multi-EAE baseline PAIE-multi and DEEIA, our model reduces the number of errors to 247 less than 358 of PAIE-multi and 292 of DEEIA. We find that RoSE reduces the Wrong Span error and Miss error, indicating the effectiveness of RoSE in capturing the structural relationships of events.

Case Study

We conduct the case study to further explore the effect of our proposed modules in multi-EAE to explore the RoGE and RoIF module effectiveness in dealing with the complex document EAE. As shown in Figure 5. First, with the RoGE and RoIF, our model can refine semantic role boundaries and correctly omit distracting entities, such as "several people" and "man", which indicates RoGE and RoIF provide more detailed heterogeneity and overlap information from role correlation.

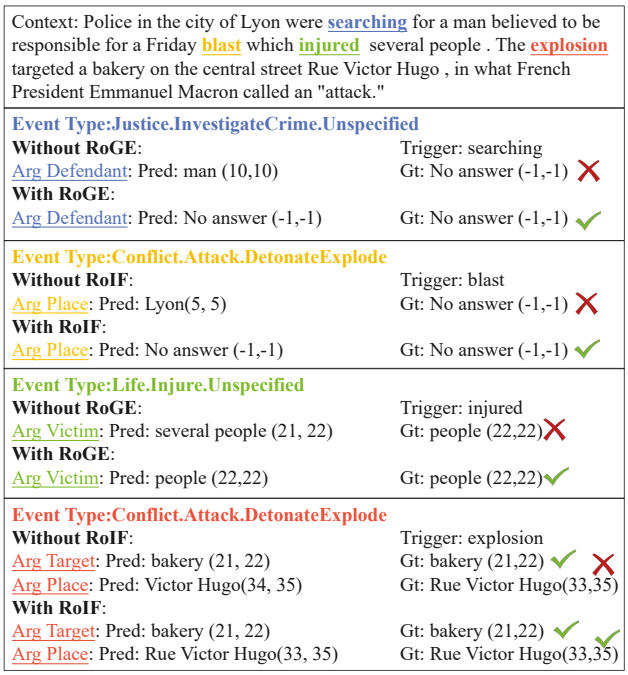

Figure 5: A multi-event test case from WikiEvents.

Conclusion

In this paper, we propose a RoSE model for Multi-EAE, which overcomes the structural heterogeneity and overlap via modeling the role correlation. The proposed Role-centric Graph-based Encoder (RoGE) module and Role-specific Information Fusion (RoIF) module effectively utilize the heterogeneity and overlap from a role-correlation perspective. Our extensive experiments on four public sentence-level and document-level datasets illustrate the superiority of our model in performance and efficiency. As we adopt a prior graph from a statistical view, this limits the application of our model to the scenes where the prior information difficult to construct. To address this, we will look into the area of automatic graph construction during the training process in the future.

Acknowledgments

This work was supported in part by the Major Program of the National Natural Science Foundation of China under Grant 62495064, in part by the Key Research and Development Program of the Department of Science and Technology of the Tibet Autonomous Region under Grant XZ202402ZY0003, in part by the Innovation Research Team Program of the Science and Technology Department of Sichuan Province Grant 2024NSFTD0051, in part by in part by the Industrial Chain Collaborative Innovation Project of Science and Technology under Grant 2025-XT00-00018-GX, and in part by the Clinical Medical Research Promotion Program of China Medical Foundation under Grant 2024CMFA10.

References

Berant, J.; Srikumar, V.; Chen, P.-C.; Vander Linden, A.; Harding, B.; Huang, B.; Clark, P.; and Manning, C. D. 2014. Modeling biological processes for reading comprehension. In *Proc. of EMNLP 2014*, 1499–1510.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, Q.; Yang, K.; Guo, X.; Wang, S.; Liao, J.; and Zheng, J. 2023. Joint Overlapping Event Extraction Model via Role Pre-Judgment with Trigger and Context Embeddings. *Electronics*, 12(22): 4688.

Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S. M.; and Weischedel, R. M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, 837–840. Lisbon.

Du, X.; and Cardie, C. 2020. Event Extraction by Answering (Almost) Natural Questions. In *Proc. of EMNLP 2020*, 671–683.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proc. of ACL 2022*, 320–335.

Ebner, S.; Xia, P.; Culkin, R.; Rawlins, K.; and Van Durme, B. 2020. Multi-Sentence Argument Linking. In *Proc. of ACL 2020*, 8057–8077.

Ge, J.; Subramanian, S.; Darrell, T.; and Li, B. 2023. From Wrong To Right: A Recursive Approach Towards Vision-Language Explanation. In *Proc. of EMNLP 2023*, 1173–1185.

Goel, A.; Gueta, A.; Gilon, O.; Liu, C.; Erell, S.; Nguyen, L. H.; Hao, X.; Jaber, B.; Reddy, S.; Kartha, R.; et al. 2023. Llms accelerate annotation for medical information extraction. In *machine learning for health (ML4H)*, 82–100. PMLR.

Grishman, R. 2015. Information extraction. *IEEE Intelligent Systems*, 30(5): 8–15.

He, S.; Du, W.; Peng, X.; Wei, Z.; and Li, X. 2025. Multi-hierarchical error-aware contrastive learning for event argument extraction. *Knowledge-Based Systems*, 309: 112889.

He, Y.; Hu, J.; and Tang, B. 2023. Revisiting Event Argument Extraction: Can EAE Models Learn Better When Being Aware of Event Co-occurrences? In *Proc. of ACL 2023*.

Hei, Y.; Sheng, J.; Guo, S.; Wang, L.; Li, Q.; Liu, J.; Liu, Y.; and Tiwari, P. 2025. RCEAE: A Role Correlation-enhanced Model for Event Argument Extraction. *Neurocomputing*, 129504.

Hobbs, J. R.; and Riloff, E. 2010. Information Extraction. *Handbook of natural language processing*, 15: 16.

Hong, Z.; and Liu, J. 2024. Towards Better Question Generation in QA-based Event Extraction. In *Proc. of ACL Findings 2024*, 9025–9038.

Hsu, I.-H.; Huang, K.-H.; Boschee, E.; Miller, S.; Natarajan, P.; Chang, K.-W.; and Peng, N. 2022. DEGREE: A Data-Efficient Generation-Based Event Extraction Model. In *Proc. of NAACL 2022*, 1890–1908.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Lan, T.; Chen, N.; Yi, Z.; Xu, X.; and Zhu, M. 2025a. Domain Generalization for Pulmonary Nodule Detection via Distributionally-Regularized Mamba. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 152–162. Springer.

Lan, T.; Yi, Z.; Xu, X.; and Zhu, M. 2025b. LooBox: Loose-box-supervised 3D Tumor Segmentation with Self-correcting Bidirectional Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 8077–8086.

Lee, J.-H.; Hahn, J.; Seo, H.-T.; Park, J.; and Han, Y.-S. 2024. SuperST: Superficial Self-Training for Few-Shot Text Classification. In *Proc. of LREC-COLING 2024*, 15436–15447.

Li, H.; Cao, Y.; Ren, Y.; Fang, F.; Zhang, L.; Li, Y.; and Wang, S. 2023. Intra-event and inter-event dependency-aware graph network for event argument extraction. In *Proc. of EMNLP Findings 2023*, 6362–6372.

Li, M.; Zareian, A.; Lin, Y.; Pan, X.; Whitehead, S.; Chen, B.; Wu, B.; Ji, H.; Chang, S.-F.; Voss, C.; et al. 2020. GAIA: A fine-grained multimedia knowledge extraction system. In *Proc. of ACL 2020*, 77–86.

Li, R.; Liu, C.; and Jiang, D. 2023. Efficient dynamic feature adaptation for cross language sentiment analysis with biased adversarial training. *Knowledge-Based Systems*, 279: 110957.

Li, S.; Ji, H.; and Han, J. 2021. Document-Level Event Argument Extraction by Conditional Generation. In *Proc. of NAACL 2021*, 894–908.

Lin, J.; and Chen, Q. 2021. Poke: A prompt-based knowledge eliciting approach for event argument extraction. *arXiv preprint arXiv:2109.05190*.

Lin, X.; Lyu, S.; Wang, X.; Chen, Q.; and Chen, H. 2025. Generation-Augmented and Embedding Fusion in Document-Level Event Argument Extraction. In *Proc. of COLING 2025*, 4078–4084.

Liu, J.; Min, L.; and Huang, X. 2021. An overview of event extraction and its applications. *arXiv preprint arXiv:2111.03212*.

Liu, W.; Zhou, L.; Zeng, D.; Xiao, Y.; Cheng, S.; Zhang, C.; Lee, G.; Zhang, M.; and Chen, W. 2024. Beyond Single-Event Extraction: Towards Efficient Document-Level Multi-Event Argument Extraction. In *Proc. of ACL Findings 2024*, 9470–9487.

Liu, X.; Huang, H.; Shi, G.; and Wang, B. 2022. Dynamic Prefix-Tuning for Generative Template-based Event Extraction. In *Proc. of ACL 2022*, 5216–5228. Association for Computational Linguistics (ACL).

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lou, J.; Lu, Y.; Dai, D.; Jia, W.; Lin, H.; Han, X.; Sun, L.; and Wu, H. 2023. Universal information extraction as unified semantic matching. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, 13318–13326.

Lu, D.; Ran, S.; Tetreault, J.; and Jaimes, A. 2023. Event Extraction as Question Generation and Answering. In *Proc. of ACL 2023*.

Lu, Y.; Lin, H.; Xu, J.; Han, X.; Tang, J.; Li, A.; Sun, L.; Liao, M.; and Chen, S. 2021. Text2Event: Controllable Sequence-to-Structure Generation for End-to-end Event Extraction. In *Proc. of ACL-IJCNLP 2021*, 2795–2806.

Luo, L.; and Xu, Y. 2023. Context-aware prompt for generation-based event argument extraction with diffusion models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1717–1725.

Lyu, Q.; Zhang, H.; Sulem, E.; and Roth, D. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proc. of ACL-IJCNLP 2021*, 322–332.

Ma, Y.; Wang, Z.; Cao, Y.; Li, M.; Chen, M.; Wang, K.; and Shao, J. 2022. Prompt for Extraction? PAIE: Prompting Argument Interaction for Event Argument Extraction. In *Proc. of ACL 2022*, 6759–6774.

Mishra, N.; Sahu, G.; Calixto, I.; Abu-Hanna, A.; and Laradji, I. 2023. Llm aided semi-supervision for efficient extractive dialog summarization. In *Proc. of EMNLP Findings 2023*, 10002–10009.

O'Gorman, T. J. 2019. *Bringing together computational and linguistic models of implicit role interpretation*. University of Colorado at Boulder.

Peng, J.; Yang, W.; Wei, F.; and He, L. 2024. Prompt for extraction: Multiple templates choice model for event extraction. *Knowledge-based systems*, 289: 111544.

Peng, R.; Liu, K.; Yang, P.; Yuan, Z.; and Li, S. 2023. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data. *arXiv preprint arXiv:2308.03107*.

Pyysalo, S.; Ohta, T.; Miwa, M.; Cho, H.-C.; Tsujii, J.; and Ananiadou, S. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18): i575–i581.

Sainz, O.; Gonzalez-Dios, I.; de Lacalle, O. L.; Min, B.; and Agirre, E. 2022. Textual Entailment for Event Argument Extraction: Zero-and Few-Shot with Multi-Source Learning. In *Proc. of NAACL Findings 2022*, 2439–2455.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

Wang, G.; Liu, D.; Nie, J.-Y.; Wan, Q.; Hu, R.; Liu, X.; Liu, W.; and Liu, J. 2025. DEGAP: Dual Event-Guided Adaptive Prefixes for Templated-Based Event Argument Extraction with Slot Querying. In *Proc. of ACL 2025*, 7598–7613.

Wu, X.; Zhang, J.; and Li, H. 2022. Text-to-Table: A New Way of Information Extraction. In *Proc. of ACL 2022*, 2518–2533.

Xu, R.; Liu, T.; Li, L.; and Chang, B. 2021. Document-level Event Extraction via Heterogeneous Graph-based Interaction Model with a Tracker. In *Proc. of ACL-IJCNLP 2021*. Association for Computational Linguistics.

Xu, Y.; He, S.; Chen, J.; Wang, Z.; Song, Y.; Tong, H.; Liu, G.; Zhao, J.; and Liu, K. 2024. Generate-on-Graph: Treat LLM as both Agent and KG for Incomplete Knowledge Graph Question Answering. In *Proc. of EMNLP 2024*, 18410–18430. Association for Computational Linguistics (ACL).

Ye, G.; Zhao, H.; Zhang, Z.; and Jiang, Z. 2025. UniDE: A multi-level and low-resource framework for automatic dialogue evaluation via LLM-based data augmentation and multitask learning. *Information Processing & Management*, 62(3): 104035.

Zhang, G.; Zhang, H.; Wang, Y.; Li, R.; Tan, H.; and Liang, J. 2024. Hyperspherical multi-prototype with optimal transport for event argument extraction. In *Proc. of ACL 2024*, 9271–9284.

Zhang, T.; Chen, M.; and Bui, A. A. 2020. Diagnostic prediction with sequence-of-sets representation learning for clinical events. In *International Conference on Artificial Intelligence in Medicine*, 348–358. Springer.

Zhang, Z.; Strubell, E.; and Hovy, E. 2022. Transfer learning from semantic role labeling to event argument extraction with template-based slot querying. In *Proc. of EMNLP 2022*, 2627–2647.

Zhou, J.; Shuang, K.; Wang, Q.; Qian, B.; and Guo, J. 2025a. Bi-directional feature learning-based approach for zero-shot event argument extraction. *Information Processing & Management*, 62(5): 104199.

Zhou, K.; Chen, N.; Yi, Z.; and Xu, X. 2025b. SA-Seg: Annotation-Efficient Segmentation for Airway Tree Using Saliency-Based Annotation. *IEEE Transactions on Medical Imaging*, 44(11): 4156–4170.