

ATSR: Adaptive Template Selection and Role-Aware Routing for Document-Level Multi-Event Argument Extraction

Yifan Wang¹, Tian Lan¹, Yanglei Gan², Run Lin³,
Leyuan Liu¹, Zufeng Wu^{1*}, Xiaoyue Yan¹, Xiaohua Wu¹

¹University of Electronic Science and Technology of China

²Southwest Minzu University

³College of Computer Science and Technology, Zhejiang University
wyf@std.uestc.edu.cn, {lantian1029, wuzufeng}@uestc.edu.cn

Abstract

Template-based approaches have advanced Event Argument Extraction (EAE) by formulating it as a structured prediction problem. However, existing methods mainly rely on either fixed manually designed templates, which suffer from template under-matching in diverse contexts, or exhaustive enumeration of template permutations, incurring prohibitively high computational cost. Moreover, they often overlook the semantic heterogeneity across events. To address these challenges, we propose **ATSR**, an Adaptive Template Selection and Role-Aware Routing framework for document-level multi-event EAE, integrating an Adaptive Multi-structure Template Selection (AMSTS) module and a Role-Aware Mixture-of-Experts (RAMoE) module. Specifically, AMSTS leverages a set of pre-defined multi-structure templates for each event type and adaptively selects the most compatible one for each event based on learnable weights conditioned on trigger representations. The RAMoE operates over template-aligned slots and routes each role to specialized experts for event semantic modeling. Extensive experiments on three benchmark datasets demonstrate that ATSR achieves state-of-the-art performance with competitive efficiency, while remaining effective in few-shot settings. The source code is available at: <https://github.com/AONE-NLP/ATSR>

1 Introduction

Document-level event argument extraction (EAE) is a crucial task within Information Extraction (Hobbs and Riloff, 2010; Grishman, 2015; Tan et al., 2024), aiming to identify event arguments and assign them to predefined semantic roles in documents containing single or multiple events (Ma et al., 2022; He et al., 2023). As illustrated in Figure 1, given a trigger word “attacked”

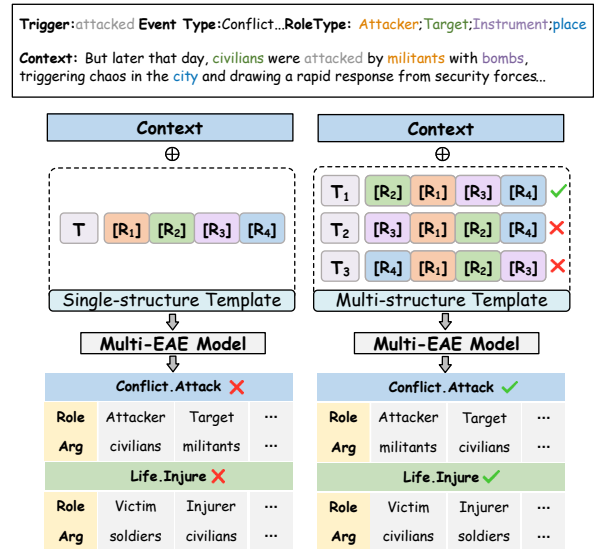


Figure 1: An example of the EAE and a comparison between Single-structure template and our Multi-Structure template selection approach for Multi-EAE.

of event type “*Conflict.Attack*”, the EAE task requires identifying “militants” as Attacker, “civilians” as Target, “bombs” as Instrument, and “city” as Place from the document context.

Existing EAE methods are primarily developed under Single-EAE settings (Ma et al., 2022; Liu et al., 2023; Zhang et al., 2024; Wan et al., 2025), where events are processed independently. Among them, template-based approaches achieve strong performance but remain limited to single-event extraction. Recent studies therefore explore Multi-EAE (He et al., 2023; Liu et al., 2024), which jointly extracts arguments for multiple events within a document more efficiently and captures the interactions among co-occurring events.

Despite these advances, existing Multi-EAE methods still face two challenges. First, as shown in Figure 1, most Multi-EAE methods rely on fixed single-structure templates with predetermined role orderings, such as $[R_1][R_2][R_3][R_4]$,

* Corresponding author.

which fail to capture diverse structural contexts and consequently lead to template under-matching (Nguyen et al., 2023; Peng et al., 2024). Moreover, template-based EAE models are sensitive to template formulations (Lin et al., 2025), where suboptimal role orderings can degrade performance. Second, existing methods overlook the semantic heterogeneity across events. Specifically, events exhibit semantic heterogeneity due to differences in interactions among argument roles and their semantics (Li et al., 2023; He et al., 2023; Zhang et al., 2024). Figure 1 shows that the same argument may correspond to different roles across events, such as “Target” and “Victim”, leading to event overlap. Meanwhile, roles such as “Attacker” and “Injurer” exhibit related yet distinct semantics across event contexts. These findings highlight event semantic heterogeneity, limiting the model’s ability to distinguish event semantics.

To address these challenges, we propose ATSR, an Adaptive Template Selection and Role-Aware Routing framework for document-level multi-event EAE. Specifically, we design an Adaptive Multi-structure Template Selection (AMSTS) module, which selects the optimal template for each event via learnable weights conditioned on trigger representations. We further introduce a Role-Aware Mixture-of-Experts (RAMoE) module that routes role slots to specialized experts for event semantic modeling. We summarize our contributions as follows:

- We propose ATSR, an adaptive template selection and role-aware routing framework for Multi-EAE that addresses template under-matching and event semantic heterogeneity.
- We introduce an Adaptive Multi-structure Template Selection (AMSTS) module and a Role-Aware Mixture-of-Experts (RAMoE) module to improve the effectiveness and reduce inference time for Multi-EAE.
- Extensive experiments on three benchmarks demonstrate that ATSR achieves state-of-the-art performance with competitive efficiency, while remaining effective in few-shot settings.

2 Related Work

2.1 Event Argument Extraction

Existing Single-EAE methods can generally be divided into three categories: (1) Span-based methods, which identify candidate spans and predict their argument roles through span classifica-

tion (Zhang et al., 2020; Xu et al., 2022; Yang et al., 2023; Liu et al., 2023; Shuang et al., 2024). (2) Generation-based methods, which leverage generative PLMs such as BART (Lewis et al., 2020) or T5 (Raffel et al., 2020) to sequentially generate event arguments in a sequence-to-sequence manner (Li et al., 2021; Wei et al., 2021; Huang et al., 2023). (3) Template-based methods, which construct predefined templates for each event and perform argument extraction in a slot-filling manner (Ma et al., 2022; Li et al., 2023; Zhang et al., 2024; Wan et al., 2025). However, these methods are designed for Single-EAE settings, processing each event independently.

Recently, researchers have begun exploring Multi-EAE (Wu et al., 2022; He et al., 2023; Liu et al., 2024), which jointly extracts arguments for multiple events within a document. TabEAE (He et al., 2023) constructs slotted table representations from event triggers and prompts to enable joint modeling of multiple events and capture event co-occurrence. To improve efficiency, DEEIA (Liu et al., 2024) introduces a dependency-guided encoding method to reduce redundancy in multi-event prompts.

2.2 Mixture-of-Experts

Mixture-of-Experts (MoE) is firstly proposed to learn a system composed of several separated modules (Jacobs et al., 1991), which has emerged as an effective paradigm for scaling model capacity by adaptively routing inputs to specialized sub-networks without proportional increases in computational cost (Cai et al., 2025). It has shown strong performance across multiple domains, including computer vision (Riquelme et al., 2021; Chen et al., 2023) and natural language processing (Artetxe et al., 2022; Xue et al., 2024) and time-series forecasting (Shi et al., 2025; Huang et al., 2025). In information extraction, MoE has been explored from multiple perspectives. MoRe (Wang et al., 2022) fuses text-retrieval and image-retrieval predictions via MoE for multi-modal named entity recognition (Gan et al., 2026) and relation extraction (Gan et al., 2025). RBDEF (Li et al., 2022) employs a dual-expert framework for long-tail EAE by separating head and tail roles into different experts. CL4DEE (Liao et al., 2026) employs a MoE adapter architecture for continual document-level event extraction, dynamically expanding expert pools to preserve historical knowledge. In our work, we extend MoE to multi-event

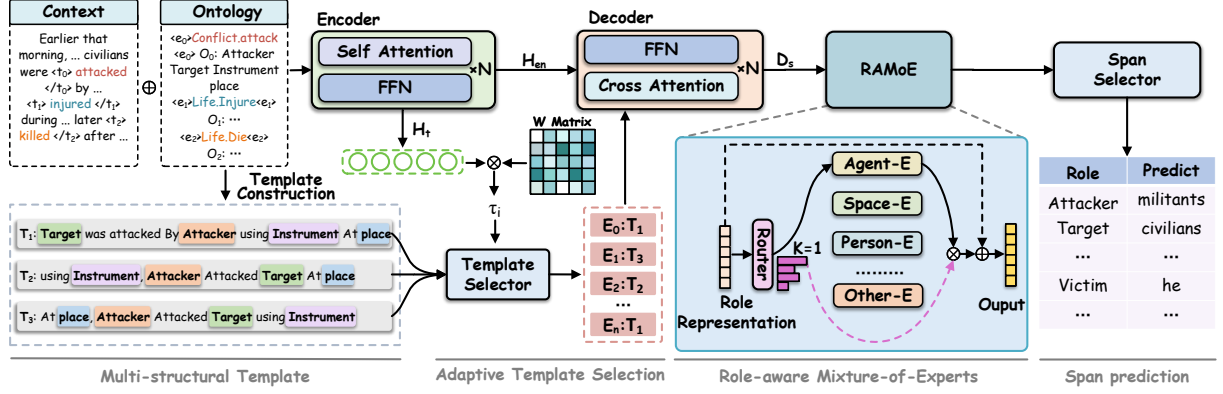


Figure 2: Overview of ATSR. ATSR first encodes the joint context-ontology input, and then adaptively selects suitable templates from pre-constructed multi-structure templates for each event based on trigger-aware representations. The selected template and contextual representations are fed into the decoder, followed by role-aware mixture-of-experts to enhance event semantic modeling for final argument extraction.

argument extraction by introducing a role-aware mixture-of-experts to address event semantic heterogeneity across co-occurring events.

3 Methodology

3.1 Task Definition

In this section, we provide a formal definition of the EAE task. Given an instance $(X, \{t_i\}_{i=1}^N, \{e_i\}_{i=1}^N, \{\mathcal{R}^{(e_i)}\}_{i=1}^N)$, where $X = (w_0, w_1, \dots, w_{S-1})$ denotes the context with S words, N is the number of target events, $t_i \subseteq X$ is the trigger word of the i -th event, e_i is the type of the i -th event, and $\mathcal{R}^{(e_i)}$ is the set of argument roles associated with e_i . The task aims to extract a set of argument-role pairs $\mathcal{A}_i = \{(a, r) \mid a \subseteq X, r \in \mathcal{R}^{(e_i)}\}$ for each event e_i , where a is a textual span serving as the argument of role r . Most previous EAE methods are designed for the single-event setting ($N=1$), while we address the more general multi-event setting where $N \geq 1$.

3.2 Ontology-guided Encoding

Following previous Multi-EAE studies (He et al., 2023; Liu et al., 2024), we construct a unified input sequence by concatenating the trigger-highlighted context and multiple event-specific ontologies, enabling the model to simultaneously process multiple events.

Given an input context X and a set of event triggers, we insert a pair of trigger markers $(\langle t_i \rangle, \langle /t_i \rangle)$ around each trigger mention w_{t_k} to obtain the highlighted context $\hat{X}$:

$$\hat{X} = w_0 \langle t_0 \rangle t_0 \langle /t_0 \rangle \dots \langle t_i \rangle t_i \langle /t_i \rangle \dots w_{S-1} \quad (1)$$

For the i -th event E_i , we adopt an event-specific ontology $\mathcal{O}_i$ consisting of the argument roles associated with the corresponding event type. We prepend each event type with a pair of type markers $(\langle e_i \rangle, \langle /e_i \rangle)$ and concatenate all event ontologies to form the global ontology sequence $\mathcal{O}$:

$$\mathcal{O} = \langle e_0 \rangle E_0 \langle /e_0 \rangle \mathcal{O}_0 \oplus \dots \oplus \langle e_i \rangle E_i \langle /e_i \rangle \mathcal{O}_i \quad (2)$$

The final input sequence is constructed as $[\hat{X}; \mathcal{O}]$ and fed into the encoder to obtain contextualized hidden representations:

$$\mathbf{H}_e = \text{Encoder}([\hat{X}; \mathcal{O}]) \quad (3)$$

3.3 Multi-structure Template Selection

3.3.1 Multi-structure Template Scheme

To better model the structural diversity of events, we construct multiple event templates with different argument organizations for each event type. Given the event ontology description, each template preserves the same argument roles while adopting different structural arrangements through role reordering.

Specifically, for an event type containing role set $\{R_1, R_2, \dots, R_n\}$, we construct multiple template structures by organizing the argument roles in different structural orders. To illustrate, consider the event type “*Conflict.Attack*” with four argument roles: *Attacker*, *Target*, *Instrument*, and *Place*, a canonical ontology can be expressed as: $[R_1]$ attacked $[R_2]$ using $[R_3]$ at $[R_4]$. Alternative structural templates can then be constructed from the same ontology, such as: *Using* $[R_3]$, $[R_1]$ attacked $[R_2]$ at $[R_4]$. Although these templates share the same argument set, they provide diverse

structural organizations through different role arrangements.

3.3.2 Adaptive Multi-structure Template Selection

Given a set of candidate multi-structure templates, our goal is to adaptively select the most suitable template for each event based on the contextual representation of the event trigger.

Trigger-aware Template Weighting. Suppose an event type is associated with n candidate templates. Given the encoded contextual representation $\mathbf{H}_e$, we use the hidden state at the trigger position as the trigger representation $\mathbf{H}_t \in R^d$. To adaptively select a suitable template for different event types, we apply Gumbel-Softmax to obtain a differentiable one-hot selection vector:

$$\tau_i = \text{Gumbel-Softmax}(\mathbf{W}^{temp} \mathbf{H}_t) \quad (4)$$

where $\mathbf{W}^{temp} \in R^{n \times d}$ denotes the learnable template selection matrix and $\tau_i \in \{0, 1\}^n$ is the selection vector for the i -th event, where the index of the non-zero element indicates the selected template. Based on τ_i , the corresponding template embedding $\mathbf{T}_i \in R^{l_t \times d}$ is obtained via the word embedding layer, where l_t is the template sequence length. The selected template embeddings across all events are concatenated to form the global template representation:

$$\mathcal{T}_{all} = \{\mathbf{T}_1; \mathbf{T}_2; \dots; \mathbf{T}_i\} \quad (5)$$

The global template $\mathcal{T}_{all}$ together with the encoded contextual representations $\mathbf{H}_e$ is then fed into the Transformer-based decoder:

$$\mathbf{H}_d = \text{Decoder}(\mathcal{T}_{all}, \mathbf{H}_e) \quad (6)$$

By adaptively selecting suitable template structures for different event types, the proposed module alleviates template under-matching and enables the decoder to better model the argument structures and dependencies among multiple co-occurring events.

3.4 Role-Aware Mixture-of-Experts

Due to the significant semantic heterogeneity across events, we introduce a Role-Aware Mixture-of-Experts (RAMoE) module for enhancing role representations. Specifically, for the m -th role slot $s_{m,i}$ in the i -th event, its slot representation is denoted as $\mathbf{h}_{s_{m,i}} \in R^d$.

The RAMoE module consists of a router $\mathcal{G}$ and a set of E expert networks $\{f_j\}_{j=1}^E$. Given an input slot representation $\mathbf{h}_{s_{m,i}}$, the router computes routing scores over all experts and activates the top- k expert. The outputs of the selected experts are aggregated to obtain the enhanced role slot:

$$\begin{aligned} \mathbf{h}_{s_{m,i}}^* &= \mathbf{h}_{s_{m,i}} + \sum_{j=1}^k \mathcal{G}(\mathbf{h}_{s_{m,i}})_j \cdot f_j(\mathbf{h}_{s_{m,i}}) \\ \mathcal{G}(\mathbf{h}_{s_{m,i}}) &= \text{TopK}\left(\text{Softmax}(\mathbf{W}_r \mathbf{h}_{s_{m,i}}), k\right) \end{aligned} \quad (7)$$

where $\mathbf{W}_r \in R^{E \times d}$ is a learnable parameter.

3.5 Span Selection

After obtaining the final representation $\mathbf{h}_{s_{m,i}}^*$ for each slot within each event, we follow (Ma et al., 2022) and project it into a pair of span selectors $\{\Phi_{s_{m,i}}^{\text{start}}, \Phi_{s_{m,i}}^{\text{end}}\}$:

$$\begin{aligned} \Phi_{s_{m,i}}^{\text{start}} &= \mathbf{h}_{s_{m,i}}^* \odot \mathbf{w}^{\text{start}} \\ \Phi_{s_{m,i}}^{\text{end}} &= \mathbf{h}_{s_{m,i}}^* \odot \mathbf{w}^{\text{end}} \end{aligned} \quad (8)$$

where $\mathbf{w}^{\text{start}}, \mathbf{w}^{\text{end}} \in R^d$ are learnable parameters and $\odot$ denotes element-wise multiplication.

The span selectors predict the start and end positions of the argument span in the input text:

$$\begin{aligned} \text{logit}_{m,i}^{\text{start}} &= \text{Softmax}(\mathbf{H}_e \Phi_{s_{m,i}}^{\text{start}}) \in R^L \\ \text{logit}_{m,i}^{\text{end}} &= \text{Softmax}(\mathbf{H}_e \Phi_{s_{m,i}}^{\text{end}}) \in R^L \\ \text{score}_{m,i}(l, n) &= \text{logit}_{m,i}^{\text{start}}(l) + \text{logit}_{m,i}^{\text{end}}(n) \\ (\hat{s}_{m,i}, \hat{e}_{m,i}) &= \arg \max_{(l, n) \in \mathcal{C}} \text{score}_{m,i}(l, n) \end{aligned} \quad (9)$$

where $\mathbf{H}_e \in R^{L \times d}$ is the encoder output for a length L input, and $(\hat{s}_{m,i}, \hat{e}_{m,i})$ is the predicted argument span. $\mathcal{C} = \{(l, n) \mid (l, n) \in L^2, 0 < n - l \leq l_m\} \cup \{(0, 0)\}$ contains all candidate spans whose length does not exceed the maximum span threshold l_m , along with the empty span $(0, 0)$ for slots with no corresponding argument.

Since multiple arguments may share the same role within a single event, we follow (Carion et al., 2020) and (Ma et al., 2022) and adopt Bipartite Matching Loss to optimally assign golden argument spans to predicted slots during training.

$$\begin{aligned} \mathcal{L}_{\text{eae}} &= - \sum_{i=1}^N \sum_{(\hat{s}_{m,i}, \hat{e}_{m,i}) \in \delta(\mathcal{S}_i)} [\log \text{logit}_{m,i}^{\text{start}}(\hat{s}_{m,i}) \\ &\quad + \log \text{logit}_{m,i}^{\text{end}}(\hat{e}_{m,i})] \end{aligned} \quad (10)$$

where N is the number of target events, $\delta(\mathcal{S}_i)$ denotes the optimal assignment for the i -th event computed via the Hungarian algorithm (Kuhn, 1955).

To avoid routing collapse in the proposed RAMoE module, where only a subset of experts is activated (Shazeer et al., 2017), we adopt an auxiliary load balancing objective from (Fedus et al., 2022) to encourage balanced expert utilization.

$$\begin{aligned}\mathcal{L}_{\text{aux}} &= \frac{1}{|\mathcal{B}|} \sum_{b=1}^{|\mathcal{B}|} E \cdot \sum_{j=1}^E f_{b,j} p_{b,j} \\ f_{b,j} &= \frac{1}{K_b} \sum_{t=1}^{K_b} \mathbf{1}(j \in \text{TopK}(\mathcal{G}(\mathbf{h}_t^b))) \\ p_{b,j} &= \frac{1}{K_b} \sum_{t=1}^{K_b} \mathcal{G}(\mathbf{h}_t^b)_j\end{aligned}\quad (11)$$

where K_b denotes the number of role slots in sample b , $\mathbf{h}_t^b$ denotes the representation of the t -th role slot in sample b , and E is the number of experts. $f_{b,j}$ and $p_{b,j}$ denote the routing frequency and the average routing probability of expert j , respectively.

The final training objective is denoted as:

$$\mathcal{L} = \mathcal{L}_{\text{eae}} + \lambda \cdot \mathcal{L}_{\text{aux}} \quad (12)$$

where λ is the regularization weight.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three widely used Event Argument Extraction (EAE) benchmarks: RAMS (Ebner et al., 2020), WikiEvents (Li et al., 2021), and MLEE (Pyysalo et al., 2012). RAMS and WikiEvents are constructed from news articles and annotated with document-level event structures, while MLEE is collected from the biomedical domain. Detailed statistics of the datasets are reported in Appendix A.1.

Evaluation Metrics. Following prior work (Ma et al., 2022; Liu et al., 2024), we report two evaluation metrics: (1) **Argument Identification F1 (Arg-I)**, where a predicted argument is considered correct if its span boundary exactly matches the gold argument span; (2) **Argument Classification F1 (Arg-C)**, where a prediction is counted as correct only if both the span boundary and the role type are correctly predicted. All results are averaged over five runs with different random seeds.

Baselines. We compare ATSR with representative EAE models, including: (1) Single-EAE methods: **EEQA** (Du and Cardie, 2020), **BART-Gen** (Li et al., 2021), **PAIE** (Ma et al., 2022), **TASR** (Xu et al., 2022), **SCPRG** (Liu et al., 2023), **TabEAE** (He et al., 2023), **AHR** (Ren et al., 2023), **ERCL** (He et al., 2025), **EPIG-EAE** (Wan et al., 2025), **GEMS** (Lin et al., 2025), and **AMOCODE** (He et al., 2026); (2) Multi-EAE methods: **PAIE-multi** (Ma et al., 2022), **TabEAE-multi** (He et al., 2023), and **DEEIA** (Liu et al., 2024).

Detailed descriptions of baselines and implementation details are provided in Appendix A.2 and A.3, and additional comparison with LLM-based methods is presented in Appendix B.1.

4.2 Main Results

Table 1 reports the performance comparison between ATSR and representative baselines on three benchmarks. Overall, ATSR consistently achieves the best performance across all datasets and evaluation metrics. We observe: (1) ATSR consistently outperforms Single-EAE methods. ATSR outperforms all Single-EAE baselines across three benchmarks. This demonstrates that our approach effectively captures the semantic heterogeneity of co-occurring events in document-level argument extraction. (2) Compared with the multi-template baseline GEMS, ATSR improves Arg-C F1 by 4.3, 3.0, and 11.2 on RAMS, WikiEvents, and MLEE, respectively. These results indicate that enumerating template permutations is insufficient for modeling diverse event structures. In contrast, ATSR adaptively selects templates conditioned on trigger semantics, improving argument extraction in multi-event scenarios. (3) ATSR significantly outperforms Multi-EAE methods. Compared with PAIE-multi, TabEAE-multi, and DEEIA, ATSR achieves consistent gains across all datasets. Specifically, compared with TabEAE-multi and DEEIA, ATSR improves Arg-C F1 by around 1.5-2.3 on RAMS, 2.9-3.6 on WikiEvents, and 1.2-1.6 on MLEE. The gains are larger on WikiEvents and MLEE, likely because they contain more complex multi-event instances and greater event semantic heterogeneity.

4.3 Ablation Study

To verify the contribution of each component in our model, we conduct ablation experiments on three benchmarks: RAMS, WikiEvents, and

Scheme	Method	PLM	RAMS		WikiEvents		MLEE	
			Arg-I	Arg-C	Arg-I	Arg-C	Arg-I	Arg-C
Single-EAE	EEQA (2020)	BART-l	48.7	46.7	56.9	54.5	70.3	68.7
	BART-Gen (2021)	BART-l	51.2	48.6	66.8	62.4	71.0	69.8
	PAIE (2022)	BART-l	56.8	52.2	70.5	65.3	72.1	70.8
	TASR (2022)	RoBERTa-l	–	51.2	71.1	65.8	72.6	71.5
	SCPRG (2023)	RoBERTa-l	56.7	52.3	71.3	66.4	–	–
	TabEAE (2023)	RoBERTa-l	57.0	52.5	70.8	65.4	71.9	71.0
	AHR (2023)	T5	54.6	48.8	69.6	63.4	–	–
	ERCL (2025)	BART-l	57.4	53.1	70.9	65.5	–	–
	EPIG-EAE (2025)	BART-l	<u>58.1</u>	<u>53.7</u>	<u>72.2</u>	<u>67.2</u>	–	–
	GEMS (2025)*	T5	56.6	49.7	69.4	65.2	66.5	63.6
	AMCODE (2026)	BART-b	55.6	50.3	70.0	64.8	–	–
Multi-EAE	PAIE-multi (2022)	BART-l	55.9	50.9	67.2	61.7	71.3	69.5
	TabEAE-multi (2023)*	RoBERTa-l	56.3	51.7	71.0	65.3	<u>74.7</u>	<u>73.6</u>
	DEEIA (2024)*	RoBERTa-l	56.9	52.5	69.6	64.6	<u>74.2</u>	<u>73.2</u>
	ATSR(ours)	RoBERTa-l	58.6	54.0	72.7	68.2	75.9	74.8

Table 1: Comparison of EAE performance on the RAMS, WikiEvents, and MLEE test sets. Results are reported in terms of Arg-I and Arg-C F1 scores (%). Methods marked with * are reproduced using their released code and original experimental settings. Bold and underline denote the best and second-best results.

Methods	RAMS		WikiEvents		MLEE	
	Arg-I	Arg-C	Arg-I	Arg-C	Arg-I	Arg-C
ATSR	58.6	54.0	72.7	68.2	75.9	74.8
w/o AMSTS	55.4	50.9	71.0	66.1	73.5	72.8
w/o RAMoE	57.2	52.2	71.1	66.7	74.2	73.1
w/o $\mathcal{L}_{aux}$	56.6	51.8	71.2	65.8	74.3	73.3
ATSR _{rand}	56.4	51.6	71.1	67.1	74.2	73.5
ATSR _{rank}	56.3	52.0	71.7	65.7	74.5	73.5

Table 2: Ablation study results across RAMS, WikiEvents, and MLEE on the impact of each component and different template selection strategies for Arg-C. The best results are marked in **bold**.

MLEE. The results are summarized in Table 2.

Without Adaptive Multi-structure Template Selection (AMSTS). Removing AMSTS results in the most substantial performance degradation across all datasets, with Arg-C drops of 3.1, 2.1, and 2.0 points on RAMS, WikiEvents, and MLEE, respectively. This highlights the pivotal role of adaptive template selection in our framework. By providing diverse and contextually appropriate templates, AMSTS effectively guides the model toward more accurate argument extraction.

Without Role-Aware Mixture-of-Experts (RAMoE). Eliminating RAMoE leads to consistent Arg-C declines of 1.8, 1.5, and 1.7 points across the three datasets. These results demonstrate that the RAMoE mechanism effectively models event semantic heterogeneity across co-occurring events, enabling more accurate

argument extraction for diverse event types.

Without Auxiliary Loss $\mathcal{L}_{aux}$. Excluding $\mathcal{L}_{aux}$ yields Arg-C drops of 3.2, 2.4, and 1.5 points on the three datasets, respectively. This indicates that the auxiliary load-balancing objective is essential for preventing routing collapse and promoting balanced expert utilization, thereby fostering expert specialization across argument types and improving overall extraction performance.

Effect of Template Selection Strategy. We further compare template selection strategies within AMSTS. ATSR_{rank} uses fixed ranking scores, while ATSR_{rand} randomly samples one template per event. As shown in Table 2, both variants underperform the full model. Notably, ATSR_{rank} achieves only 65.7 Arg-C on WikiEvents, even below the random baseline of 67.1, indicating that static ranking is ineffective. In contrast, the full model employs learnable weights for trigger-aware adaptive template selection and consistently achieves better performance across all datasets.

5 Analysis

5.1 Efficiency Analysis

Table 3 reports the inference time on three benchmarks. Compared to PAIE, ATSR reduces inference time by 15.2% and 38.5% on RAMS and MLEE, respectively, demonstrating the efficiency of our Multi-EAE framework. ATSR is also more efficient than TabEAE-multi and slightly faster

Method	Params	Inference Time	
		RAMS	MLEE
PAIE	406.21M	15.86	14.29
PAIE-multi	406.21M	12.53	8.57
TabEAE-multi	383.78M	32.59	30.06
DEEIA	388.12M	14.47	9.26
ATSR (Ours)	392.33M	13.45	8.79

Table 3: Inference time (second) of large-scale models across two dataset test sets, measured on one same Tesla A100 GPU.

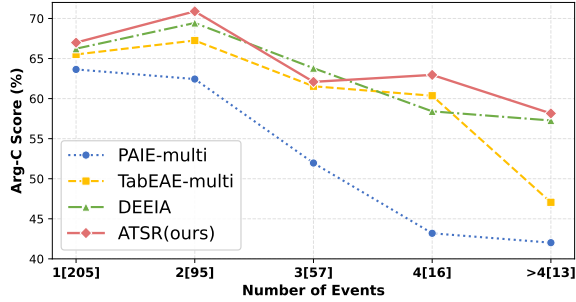


Figure 3: The averaged performance of the PAIE, TabEAE, DEEIA, and ATSR models on samples with different event numbers in the WikiEvents dataset.

than DEEIA with comparable model size, indicating that our model achieves high inference efficiency for the document-level EAE task.

5.2 Effect Analysis on Event Numbers

To evaluate ATSR’s effectiveness under increasing event complexity, we group WikiEvents instances by event count. As shown in Figure 3, all models decline as event numbers increase, reflecting the growing difficulty of modeling densely occurring events and distinguishing event semantic interactions. Nevertheless, ATSR consistently outperforms baselines and shows smaller performance drops, particularly on instances with four or more events, demonstrating its superiority in handling event semantic heterogeneity in complex multi-event scenarios. Results on MLEE are provided in Appendix B.3, along with additional analysis under different event settings in Appendix B.2.

5.3 Analysis on Event Semantics

To further investigate ATSR’s ability in modeling event semantic heterogeneity, we evaluate it on overlapping and non-overlapping multi-event scenarios, where overlapping events share arguments and introduce semantic interference. As shown in Table 4, overlapping events generally pose greater

Model	RAMS		WikiEvents		MLEE	
	N-O [690]	O [181]	N-O [296]	O [69]	N-O [1464]	O [736]
TabEAE	49.7	56.1	65.4	63.3	77.1	52.6
DEEIA	50.9	56.9	65.0	62.7	77.6	51.6
ATSR	53.2	57.6	66.7	71.8	78.2	54.8

Table 4: Results (Arg-C F1) on overlapping and non-overlapping events. Numbers in brackets denote the number of test events. **N-O**: non-overlapping; **O**: overlapping.

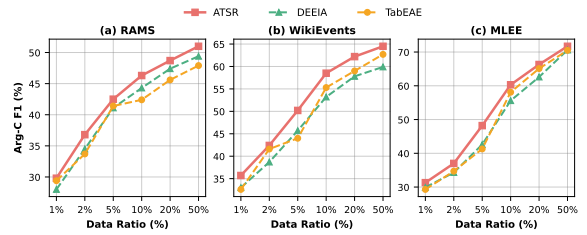


Figure 4: Arg-C F1 scores under different training data ratios on three benchmarks.

challenges for argument extraction, particularly on WikiEvents and MLEE where baselines show notable performance drops. However, ATSR consistently outperforms both TabEAE and DEEIA under both settings across all datasets. Notably, ATSR achieves larger gains on overlapping events, especially on WikiEvents, with over 8 F1 improvement over both baselines, suggesting its capability in handling semantic interference among overlapping events. Overall, these results validate the effectiveness of ATSR in capturing event semantic heterogeneity across co-occurring events.

5.4 Few-shot Setting

We analyze ATSR under limited annotation scenarios. Figure 4 reports its few-shot performance on RAMS, WikiEvents, and MLEE across different training data ratios. ATSR consistently outperforms TabEAE and DEEIA, with larger gains under extremely few-shot conditions (1%–5%). In particular, ATSR achieves substantial improvements on WikiEvents. At the 1% training ratio, ATSR surpasses the strongest baseline by 0.4, 2.7, and 1.4 F1 points on RAMS, WikiEvents, and MLEE, respectively. ATSR maintains a consistent advantage across all three datasets as the training data increases, demonstrating its effectiveness under few-shot settings.

Context: Two **gunmen** entered the **campus** right after the explosion and shot at many Afghan **soldiers** before being gunned down during the clashes...More than 100 **members** of the National Directorate of Security (NDS) were killed in...

Event #1: Conflict.Attack Trigger: shot Role: Attacker;Place;Victim
Event #2: Life.Die Trigger: killed Role: Victim
Event #n: ...

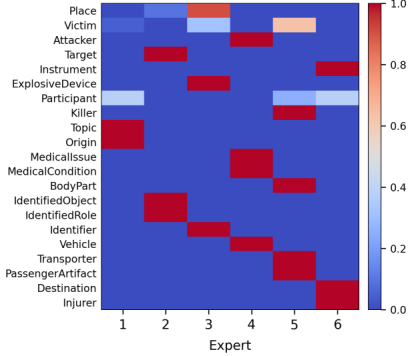


Figure 5: Visualization of expert routing specialization in RAMoE on a WikiEvents example. Rows and columns denote argument roles and experts.

5.5 Effect Analysis on Role-Aware Mixture of Expert

To assess whether RAMoE captures event-specific semantic patterns, we visualize expert assignments over argument roles. As shown in Figure 5, different experts consistently focus on semantically coherent role groups. For instance, Expert 1 concentrates on communication-related roles such as “Topic” and “Origin”, Expert 2 focuses on identification-related roles including “Target”, “IdentifiedObject”, and “IdentifiedRole”, Expert 4 emphasizes medical-related roles such as “MedicalIssue” and “MedicalCondition”, and Expert 5 focuses on conflict agent roles including “Victim”, “Killer”, and “Transporter”. These results indicate that RAMoE effectively routes semantically related roles to specialized experts, which helps model event semantic heterogeneity. Additional visualization results on MLEE are provided in Appendix B.4.

5.6 Sensitivity Analysis

Impact of the numbers of templates. We investigate the effect of varying the number of templates under different data scales. As shown in Figure 6, performance consistently improves as the template count increases from 1 to 3 across all settings. On WikiEvents, Arg-C F1 improves from 64.3 to 68.2 under full-data training, with even larger gains observed in few-shot settings. Beyond 3 templates, however, performance slightly degrades, likely due to increased structural redundancy and degraded template selection quality.

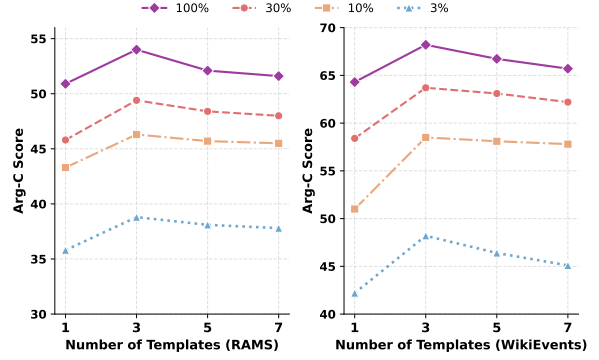


Figure 6: Performance with varying numbers of templates under different training data ratios.

Top- <i>k</i>	RAMS		WikiEvents		MLEE	
	Arg-I	Arg-C	Arg-I	Arg-C	Arg-I	Arg-C
1	58.6	54.0	72.7	68.2	75.9	74.8
2	57.0	52.8	72.2	66.8	74.8	73.8

Table 5: Performance with varying activated experts.

Overall, 3 templates achieve the best trade-off between structural diversity and selection precision.

Impact of the Number of Experts. We analyze the impact of expert count in RAMoE by varying it from 2 to 6, as shown in Figure 7. On RAMS and MLEE, performance peaks at a moderate number of experts, suggesting a trade-off between model capacity and specialization, where too many experts may lead to insufficient training per expert. WikiEvents exhibits a different trend, likely due to greater diversity of event semantics, which benefits from more experts to capture heterogeneous event semantics.

Impact of the Number of Activated Experts. In sparse MoE architectures, the router’s top-*k* selection is typically set to 1 or 2, as larger values negate sparsity benefits and increase computational overhead. Since RAMoE routes role slots rather than tokens, the routed units carry relatively concentrated semantics, making a single expert sufficient to capture the relevant signal. We evaluate performance under Top-1 and Top-2 activation, as shown in Table 5. Ultimately, we chose Top-1 routing, which consistently outperforms Top-2 across all three datasets while incurring lower computational cost.

Impact of the loss weight λ . We further investigate the sensitivity of RAMoE to the load-balancing loss weight λ , with results shown in Ta-



Figure 7: Impact of the number of experts on Arg-C performance across RAMS, MLEE, and WikiEvents.

λ	RAMS		MLEE	
	Arg-I	Arg-C	Arg-I	Arg-C
0.1	58.6	54.0	75.9	74.8
0.01	57.7	53.7	75.4	74.5
0.001	58.1	53.6	75.6	74.6
0.0001	57.4	53.1	75.1	74.3

Table 6: Performance of λ .

ble 6. The performance remains relatively stable across $\lambda \in [0.001, 0.1]$, with only minor variations on both RAMS and MLEE. When λ is further reduced to 0.0001, performance consistently drops across all evaluation metrics, suggesting that an insufficient load-balancing constraint may weaken the effectiveness of expert utilization. These results indicate that the auxiliary load-balancing loss provides a stable regularization effect, while its weight is not overly sensitive within a reasonable range.

5.7 Error Analysis and Case Study

Error Analysis. We conduct error analysis to evaluate ATSR, with details in Appendix B.5. We categorize prediction errors on the WikiEvents test set into five types. As shown in Table 11, compared with PAIE-multi and DEEIA, ATSR reduces errors in most categories, particularly Wrong Span, Miss, and Over-extraction. Specifically, total errors decrease from 350 of PAIE-multi and 313 of DEEIA to 285 of ATSR, indicating ATSR’s effectiveness for multi-event argument extraction.

Case Study. We conduct a case study to further explore the effectiveness of our proposed modules in multi-event argument extraction. As shown in Figure 8, the example contains multiple co-occurring events with overlapping arguments, where semantic heterogeneity across events increases extraction difficulty. Without AMSTS, the model fails to correctly identify key participants such as “I” and “official”, and is misled by irrelevant entities like “members”. In contrast, AMSTS im-

Context: “I have been in touch with the NDS official in the province and they told me that over 100 members of the NDS were killed in the big explosion,” the former provincial official said ... Apologizes to Survivors BOSTON — Moments before a federal judge sentenced him to death ...			
Event Type: Contact.Contact.Correspondence			
Without AMSTS:		Trigger: been in touch	
Participant:	Pred: official	Gt: I	✗
Participant:	Pred: members	Gt: official	✗
With AMSTS:			
Participant:	Pred: I	Gt: I	✓
Participant:	Pred: official	Gt: official	✓
Event Type: Life.Die.Unspecified			
Without AMSTS:		Trigger: killed	
Victim:	Pred: I	Gt: members	✗
With AMSTS:			
Victim:	Pred: members	Gt: members	✓
Event Type: Justice.Sentence.Unspecified			
Without RAMoE:		Trigger: sentenced	
JudgeCourt:	Pred: No answer (-1,-1)	Gt: judge	✗
Defendant:	Pred: judge	Gt: him	✗
With RAMoE:			
JudgeCourt:	Pred: judge	Gt: judge	✓
Defendant:	Pred: him	Gt: him	✓
Event Type: Life.Die.Unspecified			
Without RAMoE:		Trigger: death	
Victim:	Pred: federal judge	Gt: him	✗
With RAMoE:			
Victim:	Pred: him	Gt: him	✓

Figure 8: A multi-event test case from WikiEvents.

proves template matching across events, enabling accurate extraction. Furthermore, with RAMoE, the model captures overlapping arguments such as “him”, demonstrating its effectiveness in capturing event semantic heterogeneity via role experts.

6 Conclusion

In this paper, we propose ATSR for Multi-EAE, which addresses template under-matching and event semantic heterogeneity in existing methods. Specifically, we introduce an Adaptive Multi-structure Template Selection (AMSTS) module to adaptively select compatible templates for different event structures, and a Role-Aware Mixture-of-Experts (RAMoE) module to capture semantic heterogeneity across events. Extensive experiments on three benchmark datasets demonstrate the superiority of ATSR in performance and efficiency, while remaining effective in few-shot settings.

Limitation

Despite the effectiveness of ATSR, several limitations remain. First, the current framework relies on manually designed discrete templates to specify event types and argument roles, which may introduce additional template construction costs in domain-specific scenarios. Second, while ATSR adaptively selects compatible templates for different events, there is still room for exploring more advanced template selection strategies for EAE. Finally, ATSR is currently tailored for document-level EAE benchmarks, and its applicability to broader structured information extraction tasks remains to be explored.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (U2336204), and the Chengdu Industrial Chain Collaborative Innovation Project (2025-XT00-00017-GX)

References

- Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myale Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giridharan Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O’Horo, and 5 others. 2022. [Efficient large scale language modeling with mixtures of experts](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11699–11732, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Aniket Bhattacharyya, Anurag Tripathi, Ujjal Das, Archan Karmakar, Amit Pathak, and Maneesh Gupta. 2025. [Information extraction from visually rich documents using LLM-based organization of documents into independent textual segments](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17241–17256, Vienna, Austria. Association for Computational Linguistics.
- Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.
- Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. 2023. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 17346–17357.
- Gaye Colakoglu, Gürkan Solmaz, and Jonathan Fürst. 2025. [Problem solved? information extraction design space for layout-rich documents using LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17908–17927, Suzhou, China. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Bei Feng, and Zizheng Pan. 2025. [Deepseek-v3 technical report](#). Preprint, arXiv:2412.19437.
- Xinya Du and Claire Cardie. 2020. [Event extraction by answering \(almost\) natural questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. [GLM: General language model pretraining with autoregressive blank infilling](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.
- Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. [Multi-sentence argument linking](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.
- Yanglei Gan, Qiao Liu, Run Lin, Tian Lan, Yuxiang Cai, Xueyi Liu, Changlin Li, and Yan Liu. 2025. [Exploiting instance-label dynamics through reciprocal anchored contrastive learning for few-shot relation extraction](#). *Neural Networks*, 187:107259.
- Yanglei Gan, Yao Liu, Yuxiang Cai, Run Lin, Song Yang, Qiao Liu, Yashen Wang, and Xiaojun Shi. 2026. [Optimizing boundary dynamics for nested named entity recognition via semantic refinement and trimming](#). *Neural Networks*, 196:108218.
- Aaron Grattafiori, Abhimanyu Dubey, and Zhiyu Ma. 2024. [The llama 3 herd of models](#). Preprint, arXiv:2407.21783.
- Ralph Grishman. 2015. Information extraction. *IEEE Intelligent Systems*, 30(5):8–15.

- Song He, Wenli Du, Xin Peng, Zhangpeng Wei, and Xin Li. 2025. [Multi-hierarchical error-aware contrastive learning for event argument extraction](#). *Knowledge-Based Systems*, 309:112889.
- Song He, Wenli Du, Xin Peng, Zhangpeng Wei, and Xin Li. 2026. [Knowledge compensation for event argument extraction via amr-guided dynamic probability gain](#). *Expert Systems with Applications*, 295:128846.
- Yuxin He, Jingyue Hu, and Buzhou Tang. 2023. [Re-visiting 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 (Volume 1: Long Papers)*, pages 12542–12556, Toronto, Canada. Association for Computational Linguistics.
- Jerry R Hobbs and Ellen Riloff. 2010. Information extraction. *Handbook of natural language processing*, 15:16.
- Quzhe Huang, Yanxi Zhang, and Dongyan Zhao. 2023. [From simple to complex: A progressive framework for document-level informative argument extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6129–6140, Singapore. Association for Computational Linguistics.
- Xiaoyu Huang, Weidong Chen, Bo Hu, and Zhen-dong Mao. 2025. Graph mixture of experts and memory-augmented routers for multivariate time series anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 17476–17484.
- Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.
- Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Hao Li, Yanan Cao, Yubing Ren, Fang Fang, Lanxue Zhang, Yingjie Li, and Shi Wang. 2023. [Intra-event and inter-event dependency-aware graph network for event argument extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6362–6372, Singapore. Association for Computational Linguistics.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025a. [Hello again!](#) [LLM-powered personalized agent for long-term dialogue](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5259–5276, Albuquerque, New Mexico. Association for Computational Linguistics.
- Qianlong Li, Chen Huang, Shuai Li, Yuanxin Xiang, Deng Xiong, and Wenqiang Lei. 2025b. [GraphOTTER: Evolving LLM-based graph reasoning for complex table question answering](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5486–5506, Abu Dhabi, UAE. Association for Computational Linguistics.
- Rui Li, Wenlin Zhao, Cheng Yang, and Sen Su. 2022. A dual-expert framework for event argument extraction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1110–1121.
- Sha Li, Heng Ji, and Jiawei Han. 2021. [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*, pages 894–908, Online. Association for Computational Linguistics.
- Junru Liao, Haoran Wang, Weimin Cheng, and Yatian Shen. 2026. [Cl4dee: A continual learning framework for document-level event extraction with mixture-of-experts](#). *Expert Systems with Applications*, 314:131723.
- Run Lin, Yao Liu, Yanglei Gan, Yuxiang Cai, Tian Lan, and Qiao Liu. 2025. [GEMS: Generation-based event argument extraction via multi-perspective prompts and ontology steering](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26392–26409, Vienna, Austria. Association for Computational Linguistics.
- Wanlong Liu, Shaohuan Cheng, Dingyi Zeng, and Qu Hong. 2023. [Enhancing document-level event argument extraction with contextual clues and role relevance](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12908–12922, Toronto, Canada. Association for Computational Linguistics.
- Wanlong Liu, Li Zhou, Dingyi Zeng, Yichen Xiao, Shaohuan Cheng, Chen Zhang, Grandee Lee, Malu Zhang, and Wenyu Chen. 2024. [Beyond single-event extraction: Towards efficient document-level multi-event argument extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9470–9487, Bangkok, Thailand. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

- Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. [Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774, Dublin, Ireland. Association for Computational Linguistics.
- Chien Nguyen, Hieu Man, and Thien Nguyen. 2023. [Contextualized soft prompts for extraction of event arguments](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4352–4361, Toronto, Canada. Association for Computational Linguistics.
- Nicholas Pangakis and Sam Wolken. 2024. [Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels](#). In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 113–131, Mexico City, Mexico. Association for Computational Linguistics.
- Jiaren Peng, Wenzhong Yang, Fuyuan Wei, and Liang He. 2024. [Prompt for extraction: Multiple templates choice model for event extraction](#). *Knowledge-Based Systems*, 289:111544.
- Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun’ichi Tsujii, and Sophia Ananiadou. 2012. [Event extraction across multiple levels of biological organization](#). *Bioinformatics*, 28(18):i575–i581.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. 2023. [Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–306, Toronto, Canada. Association for Computational Linguistics.
- Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. [Scaling vision with sparse mixture of experts](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 8583–8595. Curran Associates, Inc.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.
- Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. 2025. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, volume 2025, pages 34635–34667.
- Kai Shuang, Ji Zhou, Qiwei Wang, and Jinyu Guo. 2024. [Thinking about how to extract: Energizing LLMs’ emergence capabilities for document-level event argument extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5520–5532, Bangkok, Thailand. Association for Computational Linguistics.
- Xin Tan, Ce Zhao, Chengliang Liu, Jie Wen, and Zhanyan Tang. 2024. [A two-stage information extraction network for incomplete multi-view multi-label classification](#). In *AAAI Conference on Artificial Intelligence*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Hai-Long Trieu, Thy Thy Tran, Khoa N A Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. [Deepeventmine: end-to-end neural nested event extraction from biomedical texts](#). *Bioinformatics*, 36(19):4910–4917.
- Qizhi Wan, Tao Liu, Changxuan Wan, Rong Hu, Keli Xiao, and Yuxin Shuai. 2025. [Event pattern-instance graph: A multi-round role representation learning strategy for document-level event argument extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1865–1877, Vienna, Austria. Association for Computational Linguistics.
- Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022. [Named entity and relation extraction with multi-modal retrieval](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5925–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ze Zhong Wang, Xingshan Zeng, Weiwen Liu, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2025. [ToolFlow: Boosting LLM tool-calling through natural and coherent dialogue synthesis](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4246–4263, Albuquerque, New Mexico. Association for Computational Linguistics.

- Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Guo Zhi, and Li Jin. 2021. [Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4672–4682, Online. Association for Computational Linguistics.
- Xueqing Wu, Jiacheng Zhang, and Hang Li. 2022. Text-to-table: A new way of information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2518–2533.
- Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2022. [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*, pages 5025–5036, Seattle, United States. Association for Computational Linguistics.
- Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. 2024. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*.
- Dongil Yang, Minjin Kim, Sunghwan Kim, Beongwoo Kwak, Minjun Park, Jinseok Hong, Woontack Woo, and Jinyoung Yeo. 2025. [LLM meets scene graph: Can large language models understand and generate scene graphs? a benchmark and empirical study](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21335–21360, Vienna, Austria. Association for Computational Linguistics.
- Yuqing Yang, Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. [An AMR-based link prediction approach for document-level event argument extraction](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12876–12889, Toronto, Canada. Association for Computational Linguistics.
- Guangjun Zhang, Hu Zhang, YuJie Wang, Ru Li, Hongye Tan, and Jiye Liang. 2024. [Hyperspherical multi-prototype with optimal transport for event argument extraction](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9271–9284, Bangkok, Thailand. Association for Computational Linguistics.
- Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2025. [Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision](#). In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 20322042, New York, NY, USA. Association for Computing Machinery.
- Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy. 2020. [A two-step approach for implicit event argument detection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7479–7485, Online. Association for Computational Linguistics.
- Ji Zhou, Kai Shuang, Qiwei Wang, Bing Qian, and Jinyu Guo. 2025. [Bi-directional feature learning-based approach for zero-shot event argument extraction](#). *Information Processing Management*, 62(5):104199.
- Yue Zuo, Yuxiao Fei, Wanting Ning, Jiayi Huang, Yubo Feng, and Lishuang Li. 2025. [Rule-guided extraction: A hierarchical rule optimization framework for document-level event argument extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21155–21171, Suzhou, China. Association for Computational Linguistics.

Dataset	RAMS	WikiEvents	MLEE
# Event types	139	50	23
# Events per text	1.25	1.78	3.32
# Args per event	2.33	1.40	1.29
# Events			
Train	7329	3241	4442
Dev	924	345	-
Test	871	365	2200

Table 7: Dataset statistics for RAMS, WikiEvents, and MLEE.

A Experimental Setup

A.1 Profile of Datasets

We conduct experiments on three widely used document-level EAE benchmarks.

RAMS (Ebner et al., 2020) is a document-level EAE dataset containing 9,124 annotated events collected from English online news articles. Since the dataset is originally annotated in an event-wise manner, where each event corresponds to an individual instance, we follow previous work (He et al., 2023) to aggregate events occurring within the same context into multi-event instances using a sliding-window strategy. We adopt the original train/dev/test data split.

WikiEvents (Li et al., 2021) is a document-level EAE dataset constructed from English Wikipedia articles and their associated news reports. The dataset additionally provides co-reference annotations for event arguments. Following prior studies, we only use the exact argument annotations in our experiments.

MLEE (Pyysalo et al., 2012) is a document-level biomedical event extraction dataset consisting of manually annotated abstracts from English biomedical publications. We follow the preprocessing procedure introduced by (Trieu et al., 2020). Since the preprocessed dataset only provides train/test splits, we employ the training split as the development set following previous work.

Statistics: The detailed dataset statistics of three datasets are listed in Table 7.

A.2 Baseline

We compare ATSR with several more advanced baselines.

EEQA (Du and Cardie, 2020) formulates EAE as a machine reading comprehension task by con-

structing role-specific questions for argument extraction.

BART-Gen (Li et al., 2021) casts EAE as a sequence-to-sequence generation problem conditioned on event templates and document contexts.

PAIE (Ma et al., 2022) adopts a prompt-based slot-filling framework with role-specific span selectors for argument extraction.

TASR (Xu et al., 2022) introduces a two-stream AMR-enhanced framework to incorporate semantic graph information into document-level EAE.

SCPRG (Liu et al., 2023) aggregates contextual information through PLM attention weights and enhances span representations via role interaction encoding.

TabEAE (He et al., 2023) extends PAIE for multi-event argument extraction. Following the original setting, we report the Multi-Multi variant for fair comparison under the Multi-EAE setting.

AHR (Ren et al., 2023) improves argument extraction by leveraging retrieval-enhanced contextual representations. The original paper proposes three retrieval strategies. We report the results of their best-performing strategy.

DEEIA (Liu et al., 2024) is a Multi-EAE framework that proposes dependency-guided encoding and event-specific information aggregation to better integrate contextual event semantics for simultaneous multi-event argument extraction.

ERCL (He et al., 2025) introduces multi-hierarchical error-aware contrastive learning with knowledge-free negative samples to improve span selection accuracy for EAE.

EPIG-EAE (Wan et al., 2025) constructs a dual-layer event graph to model role semantics at pattern and instance levels for span-based argument extraction.

GEMS (Lin et al., 2025) models argument dependencies with prompt permutation and incorporates ontology-guided cross-attention for structured argument extraction.

AMCODE (He et al., 2026) mines valid AMR entities as argument clues and applies role-specific dynamic gains to constrain decoding for event argument extraction.

A.3 Implementation Details

We implement ATSR with PyTorch and conduct experiments on a Tesla A100 GPU. Following previous EAE studies (He et al., 2023; Liu et al., 2024), we adopt RoBERTa-large (Liu et al., 2019)

Hyperparameters	RAMS	Wiki	MLEE
Training Steps	10000	10000	10000
Warmup Ratio	0.1	0.1	0.2
Learning Rate	2e-5	3e-5	3e-5
Max Gradient Norm	5	5	5
Batch Size	4	4	4
Context Window Size	250	250	250
Max Span Length	10	10	10
Max Encoder Seq Length	500	500	500
Max Prompt Length	210	360	360
Number of templates*	3	3	3
λ^*	0.1	0.075	0.1
Expert*	4	6	4
Top-k*	1	1	1

Table 8: Hyperparameter settings. * denotes tuned hyperparameters in our experiments.

as the backbone model. We use an encoder-decoder architecture, where the encoder consists of the first 17 layers of RoBERTa-large and the decoder is initialized by reusing the self-attention and feed-forward parameters from the remaining 7 layers, following the setting in (He et al., 2023). Most hyperparameter settings follow those in DEEIA (Liu et al., 2024), while task-specific hyperparameters are manually tuned. Detailed hyperparameter settings are shown in Table 8.

B Experimental Analysis

B.1 Comparison with LLMs

Large language models (LLMs) have demonstrated strong performance across a wide range of tasks, including text classification (Zhang et al., 2025; Pangakis and Wolken, 2024), graph modeling (Li et al., 2025b; Yang et al., 2025), dialogue systems (Wang et al., 2025; Li et al., 2025a), and information extraction (Bhattacharyya et al., 2025; Colakoglu et al., 2025). To compare ATSR against recent LLM-based approaches (Zhou et al., 2025), we include ChatGPT-3.5, LLaMA2-7B-Chat (Touvron et al., 2023), fine-tuned variants of LLaMA2-7B-Chat and ChatGLM2-6B (Du et al., 2022), as well as the recent HERO framework (Zuo et al., 2025) based on LLaMA-3-8B-Instruct (Grattafiori et al., 2024) and DeepSeek-V3 (DeepSeek-AI et al., 2025). As shown in Table 10, despite their strong generalization ability, LLM-based methods still lag behind supervised approaches on EAE. Even after fine-tuning, these models perform substantially worse than ATSR on both RAMS and WikiEvents across all evaluation metrics. This

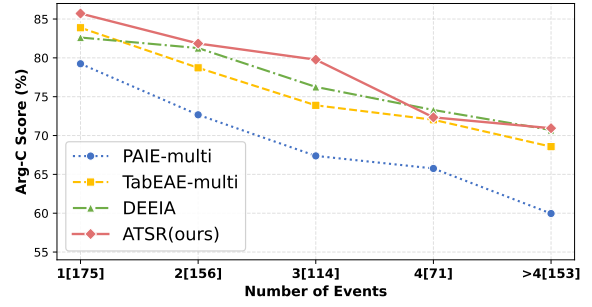


Figure 9: The averaged performance of the PAIE, TabEAE, DEEIA, and ATSR models on samples with different event numbers in the MLEE dataset.

gap can be attributed to the difficulty of applying LLMs to structured prediction tasks that require precise span extraction and role classification. In contrast, ATSR achieves consistently superior performance without relying on large-scale language model pretraining, highlighting the effectiveness of the proposed ATSR for document-level event argument extraction.

B.2 Analysis under Different Event Settings

To evaluate the effectiveness of ATSR under different event settings, we partition test sets into single-event and multi-event subsets. As shown in Table 9, ATSR achieves consistent performance across both, validating each component’s contribution. Removing either AMSTS or RAMoE leads to performance degradation, confirming their roles in template matching and event semantic modeling.

B.3 Effect Analysis on Event Numbers in MLEE

As illustrated in Figure 9, all models exhibit a consistent performance decline as the number of events increases, reflecting the growing complexity of modeling longer contexts with more event semantic interactions. Among the baselines, PAIE-multi degrades most significantly as the number of events increases, while TabEAE-multi and DEEIA exhibit relatively smaller performance drops. In contrast, ATSR maintains a consistent advantage over all baselines across different event-number groups, demonstrating its superiority in handling multi-event extraction under increasing event complexity.

Model	RAMS		WikiEvents		MLEE	
	# Ev = 1 [587]	# Ev > 1 [284]	# Ev = 1 [114]	# Ev > 1 [251]	# Ev = 1 [175]	# Ev > 1 [2025]
PAIE-multi	51.75	48.94	65.01	60.18	-	-
TabEAE-multi	52.88	50.11	65.04	66.19	84.74	73.24
DEEIA	54.74	49.59	61.32	67.26	82.05	72.76
ATSR	55.94	50.32	67.92	68.30	85.71	74.01
w/o AMSTS	53.75	50.00	66.31	66.08	84.06	72.06
w/o RAMoE	53.65	49.52	65.59	67.25	82.62	72.44

Table 9: Arg-C F1 performance comparison under single-event and multi-event settings.

Model	RAMS		WikiEvents	
	Arg-I	Arg-C	Arg-I	Arg-C
ChatGPT	46.2	40.4	42.4	42.4
LLaMA	24.8	19.4	24.7	22.7
LLaMA _{fine-tune}	50.9	45.8	60.3	58.6
ChatGLM _{fine-tune}	46.4	38.2	64.3	51.0
HERO _{Llama3}	43.5	37.2	-	-
HERO _{DeepSeek-V3}	50.5	44.0	-	-
ATSR (ours)	58.6	54.0	72.7	68.2

Table 10: Comparison with large language models on RAMS and WikiEvents.

B.4 Visualization of Experts on MLEE

Figure 10 visualizes the expert assignment distributions on a representative MLEE instance. Compared with WikiEvents, MLEE exhibits more fine-grained biomedical event semantics, yet RAMoE still shows clear and consistent expert specialization patterns. As shown in the heatmap, Expert 1 primarily handles causal-spatial semantics such as “Cause” and “ToLoc”, Expert 2 specializes in origin-related semantics such as “FromLoc”, Expert 3 focuses on instrumental-spatial semantics including “Instrument” and “Site”, and Expert 4 concentrates on core biomedical process semantics such as “Theme” and “Participant”. These results suggest that RAMoE consistently routes semantically related roles to the same expert even in highly specialized biomedical domains, demonstrating its ability to capture coherent semantic patterns across domains with diverse event semantics.

B.5 Error Analysis

To further evaluate the effectiveness of ATSR, we conduct a detailed error analysis on the WikiEvents test set. We categorize all prediction errors into five types. Wrong Span refers to cases where the predicted span and the gold

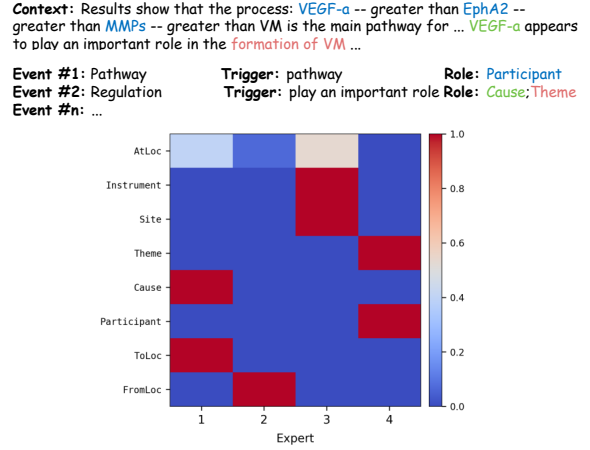


Figure 10: Visualization of expert routing specialization in RAMoE on a MLEE example. Rows and columns denote argument roles and experts.

span do not overlap. Partial denotes cases where the predicted span is a proper subset of the gold span or vice versa. Overlap occurs when the predicted and gold spans intersect but neither fully contains the other. Miss refers to cases where a gold span exists but no prediction is made, while Over-extraction describes cases where a span is predicted without a corresponding gold annotation. As shown in Table 11, ATSR reduces the total number of errors to 285, compared with 350 for PAIE-multi and 313 for DEEIA, demonstrating its overall effectiveness. The most notable reductions are observed in the Miss and Over-extraction categories, indicating that ATSR more accurately captures the boundaries of event argument semantics. Regarding ablation variants, removing AMSTS mainly increases Miss and Over-extraction errors, suggesting that adaptive template selection helps better align event contexts with suitable candidate templates. Removing RAMoE mainly increases Partial and Miss errors, confirming that role-aware mixture-of-experts better models semantic heterogeneity across events. Overall, these

Method	Error Category					
	Wrong Span	Partial-less	Partial-more	Overlap	Miss	Over-extraction
PAIE-multi	46	16	6	1	170	111
DEEIA	38	12	13	1	159	90
ATSR	34	12	8	0	148	83
w/o AMSTS	36	11	9	0	156	97
w/o RAMoE	37	14	11	1	168	86

Table 11: Error Analysis on WikiEvents test set.

results demonstrate that the two modules complementarily address template under-matching and event semantic heterogeneity in Multi-EAE.