

DEGAP: Dual Event-Guided Adaptive Prefixes for Templated-Based Event Argument Extraction with Slot Querying

Guanghui Wang¹, Dexi Liu^{1*}, Jian-Yun Nie², Qizhi Wan¹,
Rong Hu¹, Xiping Liu¹, Wanlong Liu³ and Jiaming Liu⁴

¹School of Information Management, Jiangxi University of Finance and Economics

²University of Montreal (UdeM), Department of Computer Science and Operations Research (DIRO)

³School of Computer Science and Engineering, University of Electronic Science and Technology of China

⁴Department of Statistics, Division of the Physical Sciences, the University of Chicago

{15279836691, dexi.liu}@163.com

Abstract

Recent advancements in event argument extraction (EAE) involve incorporating useful auxiliary information into models during training and inference, such as retrieved instances and event templates. These methods face two challenges: (1) the retrieval results may be irrelevant and (2) templates are developed independently for each event without considering their possible relationship. In this work, we propose DEGAP to address these challenges through a simple yet effective components: dual prefixes, i.e. learnable prompt vectors, where the instance-oriented prefix and template-oriented prefix are trained to learn information from different event instances and templates. Additionally, we propose an event-guided adaptive gating mechanism, which can adaptively leverage possible connections between different events and thus capture relevant information from the prefix. Finally, these event-guided prefixes provide relevant information as cues to EAE model without retrieval. Extensive experiments demonstrate that our method achieves new state-of-the-art performance on four datasets (ACE05, RAMS, WIKIEVENTS, and MLEE). Further analysis shows the impact of different components.

1 Introduction

As an essential subtask of Event Extraction, Event Argument Extraction (EAE) aims to identify event arguments and their roles for given event mentions in context. For example, in Fig. 1, given a *Life.Die* event with trigger *killed*, EAE models need to extract arguments *soldiers*, *bomber*, and *checkpoint*, and the corresponding roles — *Victim*, *Agent*, and *Place*. In the field of NLP, EAE has always been a challenging task, which requires deep natural language understanding (Hsu et al., 2023).

Consequently, some efforts incorporate ancillary information useful for EAE into models during



Event Type & Template
Life.Die: <u>Agent</u> killed <u>Victim</u> with <u>Instrument</u> at <u>Place</u>
Conflict.Attack: <u>Attacker</u> attacked <u>Target</u> using <u>Instrument</u> at <u>Place</u>

Figure 1: Two examples of EAE and event template from Ma et al. (2022). Each underlined part (i.e., role slot) in the template is named after an argument role.

training and inference, such as augmenting input with retrieved instances (Du and Ji, 2022; Ren et al., 2023; Zhao et al., 2023a) or querying arguments using role slots in event template (Liu et al., 2024; Ma et al., 2022; He et al., 2023; Nguyen et al., 2023). Moreover, in light of advancements in prompt learning (Li and Liang, 2021; Liu et al., 2022b), some studies introduce prefixes (i.e. learnable prompt vectors) into models as additional keys and values to influence the computation of attention (Liu et al., 2022a; Zhang et al., 2023a; Hsu et al., 2023; Nguyen et al., 2023). Despite the empirical success of previous studies, two issues remain, hindering further performance improvements.

First, retrieved instances are not always reliable due to deficiencies in retrieval. Previous studies regard event instances retrieved from training data (Wang et al., 2022) as cues to extract arguments. However, these methods typically use a retriever pre-trained on other tasks with frozen parameters, which may not always retrieve other event instances relevant to the target event¹. Ad-

¹In a context, only a few words are event arguments, while other distracting context can mislead retrieval based on sentence semantic similarity. According to (Ren et al., 2023), on the RAMS dataset (Ebner et al., 2020), only 16.51% of event instances can recall a sample with the same event type.

*Corresponding author: Dexi Liu.

ditionally, these methods often suffer from high computational costs and nontrivial knowledge loss because they independently retrieve and encode event instances, resulting in a poor local optimum. Directly increasing the number of instances to be retrieved is typically inefficient. Recent advancements show that one can *utilize prefixes as efficient parameters to learn information from different corpora and integrate retrieved instances into prefixes as cues for inference* (Zhang et al., 2023b; Liu et al., 2023a). Inspired by this intriguing observation, in this work, we attempt to explore a novel approach to utilize relevant event instance without retrieval. First, we use prefixes to learn the semantic information of event instances. Then, we leverage the learned prefixes as cues to guide the model.

Second, no relation between different event templates is considered. Most of the existing approaches use a separate template as prompt for each target event. However, some event templates for other relevant events may provide useful information to the given event. This potential connection between event templates has been overlooked in the literature (Ma et al., 2022; He et al., 2023). Only (Ren et al., 2023) attempted to retrieve role labels from other events that are relevant to the target event as demonstrations. As shown in Fig. 1, event templates provide rich semantic information about events, and other relevant events often have similar structures and share part of their roles (e.g., *Life.Die* and *Conflict.Attack*) (Li et al., 2023). Intuitively, one can regard relevant event templates as cues and perform reasoning by analogy. This can help EAE models better extract the target event from complex semantics.

In this paper, we aim to address the above challenges. We propose DEGAP (**D**ual **E**vent-**G**uided **A**daptive **P**refixes for event argument extraction). (1) Specifically, we design two types of prefixes: the instance-oriented prefix and the template-oriented prefix. The reason we design separate prefixes for event instances and templates, rather than a shared prefix for both, is that instances and templates contain different types of information about events: an instance is a concrete example of an event, while a template tries to capture the general structure of an event. They play different roles in EAE, thus it is more appropriate to train separate prefixes for them². These instance/template-

²Experimental results confirm that dual prefixes can better learn semantic information from different event instances and templates, and guide model to extract argument (§ 5.2).

specific parameters are used to learn information from different event instances/templates and enable them to interact sufficiently with each other, rather than ignoring the potential connections between them. They are trained and updated together with model parameters, thus avoiding the insufficient retrieval and poor local optimum issues present in retrieval-based EAE methods. (2) Moreover, not all event instances and templates are useful to the target event. We aim to use only those instances and templates that are relevant to the target event (Xu et al., 2022; Liu et al., 2023b). To connect other relevant events to the target event and thus adaptively capture relevant information from the prefix, we design an event-guided adaptive gating mechanism. It utilizes the target event’s trigger and type to score each instance-oriented and template-oriented prefix token via gated weight assignment at each layer, respectively. Finally, these event-guided prefixes are used to guide model for extracting argument.

To comprehensively evaluate our method, we conduct extensive experiments on four benchmark datasets (ACE05, RAMS, WIKIEVENTS, MLEE). Experimental results demonstrate that our method outperforms existing methods across all datasets, illustrating the effectiveness of our approach. Furthermore, we conduct a thorough study to explore the reasons for the performance improvement of our method, confirming the significance of the proposed design and the efficacy of main components.

2 Related Work

Event (argument) extraction with auxiliary information. As the task requires a deep understanding of context, many previous efforts have been put into exploring which auxiliary information is useful for event predictions. One line of studies is incorporating such auxiliary information during training and inference, such as event templates (Li et al., 2021; Hsu et al., 2022; Zeng et al., 2022; He et al., 2023; Liu et al., 2024), natural questions (Liu et al., 2020, 2021; Lu et al., 2023; Uddin et al., 2024), retrieved instances (Du and Ji, 2022; Ren et al., 2023; Zhao et al., 2023a), already predicted events (Huang et al., 2023; Du et al., 2022), and syntactic and semantic information (Huang et al., 2017; Veysel et al., 2020; Ahmad et al., 2021; Xu et al., 2022; Yang et al., 2023). As mentioned in § 1, these methods may lead to irrelevant results due to retrieval deficiencies that may negatively impact the final results. Our approach is based on

the principle of prompt learning.

Prompt learning. Prompt learning is a straightforward and effective method that adjusts PLMs for specific downstream tasks by adding prompts to the input. One line of research applies an automated search to find appropriate discrete prompting words (Wallace et al., 2019; Shin et al., 2020; Wen et al., 2024). An alternative line of research attaches a set of trainable prefix tokens to the model’s attention module as additional keys and values (Li and Liang, 2021; Liu et al., 2022b; Chen et al., 2023). This method has been widely applied in various studies. For example, Liu et al. (2023a) integrates retrieved instances into the prefix as cues for inference. Zhang et al. (2023b) uses prefixes to learn information from different corpora and refine inference. Additionally, some EAE efforts use prefixes to optimize manually designed prompts (Liu et al., 2022a; Nguyen et al., 2023) or to integrate syntactic and semantic information (Hsu et al., 2023).

Inspired by these advancements (Zhang et al., 2023b; Liu et al., 2023a), We explore a novel approach to utilize potentially relevant event information rather than retrieval: using prefix to learn the semantic information of different event instances and templates, then leveraging the learned prefix as cues to guide model. Moreover, previous methods overlook the potential connections between event templates. Some relevant event templates may provide useful information to the target event. As such, we propose an event-guided adaptive gating mechanism, which can effectively connect other relevant events to the target event and thus adaptively capture relevant information from the prefix.

3 Methodology

In this section, we first define the EAE task. Then, we detail how the proposed DEGAP solves the challenges mentioned.

Task Definition Formally, we first define a set of event types $\mathcal{E} = \{e_1, e_2, e_3, \dots\}$ and a corresponding set of roles $R^e = \{r_1, r_2, r_3, \dots\}$ for each event type $e \in \mathcal{E}$. Then, given a context $X = \{x_1, x_2, \dots, x_n\}$ of n words, the event trigger $x_{\text{trig}} \in X$ and the event type e , the task aims to extract all (s, r) pairs for the event, where s is a text span in X , and r is an argument role in R^e .

3.1 Model Architecture

Fig. 2 shows the overall architecture of DEGAP. For a given event type e , we create an event tem-

plate for it (§ 3.1.1). To learn information from event instances and templates, we design instance-oriented prefixes and template-oriented prefixes, respectively (§ 3.1.2). Dual prefixes can better learn semantic information from different event instances and templates compared to a shared prefix (§ 5.2). Then, we design an event-guided adaptive gating mechanism to connect other relevant events to the target event and thus adaptively capture relevant information in the prefix (§ 3.1.3). Subsequently, these event-guided prefixes provide relevant event information as cues to model (§ 3.1.4). Finally, we introduce the span selection, which utilizes role slots in the event template to query context for extracting argument (§ 3.1.5).

3.1.1 Event Template Creation

We create an event template \mathcal{T}_e for the given event type e , utilizing its rich semantic information to guide the model (Hsu et al., 2022). Similar to previous works (Liu et al., 2022a; Zhang et al., 2023b), the event template \mathcal{T}_e consists of two parts: the type part \mathcal{I}_e and the schema part \mathcal{S}_e . The type part \mathcal{I}_e adopts the format “Event type is [Event type]”. For the schema part \mathcal{S}_e , we directly use that from TabEAE (He et al., 2023). Considering the example in Fig. 2, where there is a *Conflict.Attack* event with the event trigger *hit*, the template for this event is: “<s> Event type is Conflict.Attack </s> <s> Attacker (and Attacker) attacked Target (and Target) using Instrument (and Instrument) at Place (and Place) </s>”, where each underlined part (i.e., role slot) is named after an argument role. The brackets indicate that there may be multiple arguments of that role in an event.

3.1.2 Dual Prefixes

To directly and fully benefit from relevant event instances and templates without retrieval as an intermediate step, we construct two types of prefixes: (1) the **Instance-Oriented Prefix** (IOP) and (2) the **Template-Oriented Prefix** (TOP). These instance/template-specific prefix parameters can learn rich semantic information from different event instances/templates and are trained and updated alongside model parameters.

Note that we only design a shared IOP/TOP for all event instances/templates, rather than one for each event instance/template. This approach allows them to interact fully with each other, rather than ignoring the potential connections between them.

To be more specific, at the i -th layer of encoder

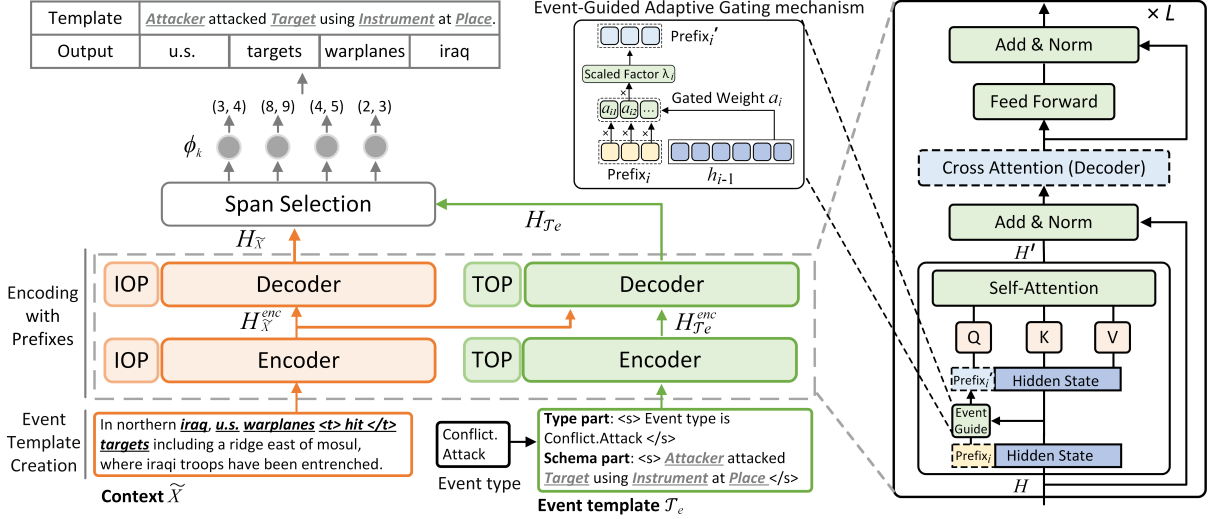


Figure 2: The overview of DEGAP. Here, the two encoders share the same parameters, and the two decoders also share the same parameters. Due to space constraints, the event template here is simplified. When the model input is context, $Prefix_i$ and h_{i-1} refer to $Prefix_i^{ins}$ and h_{i-1}^{trig} , respectively. Conversely, they refer to $Prefix_i^{tem}$ and h_{i-1}^{type} , respectively. **IOP**: Instance-Oriented Prefix. **TOP**: Template-Oriented Prefix.

or decoder, we equip the self-attention block with a set of learnable prefix tokens oriented towards event instances, $Prefix_i^{ins} \in \mathbb{R}^{len_{ins} \times m}$, and event templates, $Prefix_i^{tem} \in \mathbb{R}^{len_{tem} \times m}$. Here, len_{ins} is the length of $Prefix_i^{ins}$, len_{tem} is the length of $Prefix_i^{tem}$, and m corresponds to the hidden dimension. We also attempt to equip the cross-attention block in each layer of the decoder with these prefix tokens, but observe almost no improvement in performance. These additional key and value matrices will be concatenated with the original key and value matrices in the attention block. When calculating dot-product attention, the query at each position will be influenced by these prefix tokens.

3.1.3 Event-Guided Adaptive Gating mechanism

Some relevant events may provide useful cues to the target event and irrelevant event information might hinder model performance. To connect other relevant events to the target event and thus adaptively capture relevant information from the prefix, inspired by Liu et al. (2023b), we utilize the target event’s trigger and type to score each prefix token in $Prefix_i^{ins}$ and $Prefix_i^{tem}$ through gated weight assignment, respectively. For example, in the i -th layer, the weight distribution of tokens in $Prefix_i^{ins}$ is calculated as:

$$a_i^{ins} = \text{sigmoid}(h_{i-1}^{trig} W_i^{ins}), \quad (1)$$

where h_{i-1}^{trig} is the hidden state of the event trigger from the previous layer obtained through average pooling, and $W_i^{ins} \in \mathbb{R}^{m \times len_{ins}}$ corresponds to the parameters to be learned. Here, the prefix weight a_i^{ins} captures the relevance of other event instances to the given instance, and based on this, we can further capture the relevant information within prefix.

Additionally, we also follow Zhang et al. (2023d) to introduce a learnable scaling factor λ_i^{ins} to balance the information provided by the original input and the prefix. We obtain the guided instance-oriented prefix representation $Prefix_i^{ins'}$:

$$Prefix_i^{ins'} = \lambda_i^{ins} a_i^{ins} \odot Prefix_i^{ins}, \quad (2)$$

where \odot is the element-wise multiplication. Similarly, we derive the guided template-oriented prefix representation $Prefix_i^{tem'}$:

$$a_i^{tem} = \text{sigmoid}(h_{i-1}^{type} W_i^{tem}), \quad (3)$$

$$Prefix_i^{tem'} = \lambda_i^{tem} a_i^{tem} \odot Prefix_i^{tem}, \quad (4)$$

where h_{i-1}^{type} is the hidden state obtained through average pooling of the event type from the previous layer, and $W_i^{tem} \in \mathbb{R}^{m \times len_{tem}}$ is learnable parameters. The prefix weight a_i^{tem} captures the relevance of other event templates to the given template.

Finally, these event-guided prefixes can more effectively provide relevant event instances and templates as cues for the target event. The new self-attention module with event information intervention is formalized as:

$$H \leftarrow \text{LayerNorm}(H' + H),$$

$$H' = \text{MHSA}(\text{Prefix}'_i \oplus H)_{|\text{Prefix}'_i|:|\text{Prefix}'_i \oplus H|}, \quad (5)$$

where $\text{MHSA}(\cdot)$ denotes the multi-head self-attention mechanism, Prefix'_i represents either $\text{Prefix}_i^{\text{ins}'}$ or $\text{Prefix}_i^{\text{tem}'}$, H refers to the hidden state, and $b : c$ denotes the slicing operation on the seq_len dimension from b to c .

3.1.4 Encoding with Prefixes

To ensure a fair comparison and to visually demonstrate the effects of our prefixes, we adopt the encoding process from prior work (He et al., 2023) for encoding context instance and event template, and augment it with our prefixes. With both relevant event instances and templates as cues, model can better understand the target event from context.

Given input context $X = \{x_1, x_2, \dots, x_n\}$ and the event trigger x_{trig} , we use text markers $\langle \mathbf{t} \rangle$ and $\langle / \mathbf{t} \rangle$ as special tokens and then insert them before and after x_{trig} in X , respectively.

$$\tilde{X} = \{x_1, x_2, \dots, \langle \mathbf{t} \rangle x_{\text{trig}} \langle / \mathbf{t} \rangle, \dots, x_n\}.$$

Then, we input the processed \tilde{X} into the encoder and the decoder to obtain the context representation $H_{\tilde{X}}$:

$$H_{\tilde{X}}^{\text{enc}} = \text{Encoder}(\text{Prefix}_{\text{enc}}^{\text{ins}}, \tilde{X}), \quad (6)$$

$$H_{\tilde{X}} = \text{Decoder}(\text{Prefix}_{\text{dec}}^{\text{ins}}, H_{\tilde{X}}^{\text{enc}}), \quad (7)$$

where $\text{Prefix}_{\text{enc}}^{\text{ins}}$ and $\text{Prefix}_{\text{dec}}^{\text{ins}}$ are the instance-oriented prefix representations $\text{Prefix}_i^{\text{ins}}$ from the encoder and decoder, respectively.

For the target event template \mathcal{T}_e , we input it into the encoder and the decoder to obtain the template representation $H_{\mathcal{T}_e}$:

$$H_{\mathcal{T}_e}^{\text{enc}} = \text{Encoder}(\text{Prefix}_{\text{enc}}^{\text{tem}}, \mathcal{T}_e), \quad (8)$$

$$H_{\mathcal{T}_e} = \text{Decoder}(\text{Prefix}_{\text{dec}}^{\text{tem}}, H_{\mathcal{T}_e}^{\text{enc}}, H_{\tilde{X}}^{\text{enc}}), \quad (9)$$

where $\text{Prefix}_{\text{enc}}^{\text{tem}}$ and $\text{Prefix}_{\text{dec}}^{\text{tem}}$ are the template-oriented prefix representations $\text{Prefix}_i^{\text{tem}}$ from the encoder and decoder, respectively.

3.1.5 Span Selection

With the template representation $H_{\mathcal{T}_e}$ obtained through Eq. 9, for the k -th role slot in the template, we can obtain its representation h_{s_k} and span selector $\phi_k = \{\phi_k^{\text{start}}, \phi_k^{\text{end}}\}$ (Ma et al., 2022; He et al., 2023):

$$\phi_k^{\text{start}} = h_{s_k} \odot w^{\text{start}},$$

$$\phi_k^{\text{end}} = h_{s_k} \odot w^{\text{end}}, \quad (10)$$

where w^{start} and w^{end} are learnable parameters. Subsequently, the span selector ϕ_k queries the context to extract at most one argument span (s_k, e_k) for role slot k :

$$\text{logit}_k^{\text{start}} = \phi_k^{\text{start}} H_{\tilde{X}} \in \mathbb{R}^n,$$

$$\text{logit}_k^{\text{end}} = \phi_k^{\text{end}} H_{\tilde{X}} \in \mathbb{R}^n,$$

$$\text{score}_k(l, r) = \text{logit}_k^{\text{start}}(l) + \text{logit}_k^{\text{end}}(r),$$

$$(s_k, e_k) = \arg \max_{(l, r) \in \mathcal{C}} \text{score}_k(l, r) \quad (11)$$

where candidate span set $\mathcal{C} = \{(l, r) \mid (l, r) \in n^2, 0 < r - l < \text{MSL}\} \cup \{(0, 0)\}$. MSL denotes the max span length. The loss function is defined as:

$$P_k^{\text{start}} = \text{Softmax}(\text{logit}_k^{\text{start}}),$$

$$P_k^{\text{end}} = \text{Softmax}(\text{logit}_k^{\text{end}}),$$

$$\mathcal{L} = - \sum_X \sum_k [\log P_k^{\text{start}}(s_k) + \log P_k^{\text{end}}(e_k)], \quad (12)$$

where X ranges over all contexts in dataset and k ranges over all role slots in the template for X , and (s_k, e_k) is the golden span for role slot k . For a role slot that does not relate to any argument, it is assigned the empty span $(0, 0)$.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on four benchmark datasets, specifically including ACE05 (Dodgington et al., 2004), RAMS (Ebner et al., 2020), WIKIEVENTS (Li et al., 2021), and MLEE (Pyysalo et al., 2012). ACE05 is a sentence-level dataset, while the other three are document-level datasets. We provide detailed introductions to these datasets in Appendix A.1.

Evaluation Metrics. We adopt two evaluation metrics: (1) Argument Identification F1 score (**Arg-I**): A predicted argument for an event is correctly identified if its boundary matches any golden arguments of the event. (2) Argument Classification F1 score (**Arg-C**): a predicted argument of an event is correctly classified only if its boundary and role type both match the ground truth. The final reported performance is averaged over five random runs.

Implementation details. Please refer to Appendix A.2 for details of implementation.

Category	Model	PLM	ACE05		RAMS		WIKIEVENTS		MLEE	
			Arg-I	Arg-C	Arg-I	Arg-C	Arg-I	Arg-C	Arg-I	Arg-C
Templated-based	BART-Gen (Li et al., 2021)	RoBERTa	69.9	66.7	51.2	47.1	66.8	62.4	71.0	69.8
	DEGREE (Hsu et al., 2022)	BART	75.6	73.0	-	-	-	-	-	-
	SPEAE (Nguyen et al., 2023)	BART	-	-	<u>58.0</u>	<u>53.3</u>	<u>71.9</u>	<u>66.1</u>	-	-
	PAIE (Ma et al., 2022)	BART	75.7	72.7	56.8	52.2	70.5	65.3	72.1	70.8
	TabEAE (He et al., 2023)	RoBERTa	75.5	72.6	57.0	52.5	70.8	65.4	71.9	71.0
Retrieval-based	R-GQA (Du and Ji, 2022)	BART	75.6	72.8	-	-	-	-	-	-
	AHR (Ren et al., 2023)	T5	-	-	54.6	48.4	69.6	63.4	-	-
	DRAAE (Ours)	RoBERTa	<u>76.1</u>	73.2	57.5	53.0	71.0	66.0	<u>72.8</u>	<u>71.7</u>
Prefix-based	DEGAP (Ours)	RoBERTa	76.6	74.4	58.5	54.2	72.2	67.1	74.0	73.4

Table 1: The overall performance of all methods. Models in the PLM column are of large-scale (with 24 Transformer layers). The highest scores are highlighted in bold, and the second-highest scores are underlined.

Baselines. Details about the baselines are listed in Appendix A.3. For a comprehensive comparison, we also conduct a comparative analysis between DEGAP and LLMs in Appendix B.1, showing that DEGAP can outperform the latest LLMs for EAE.

We also compare our method with its variant, the **Dual Retrieval-Augmented Argument Extraction** model (DRAAE), which only retrieves relevant event instances and templates as cues to guide the model, instead of using the prefixes we designed. We provide further implementation details of this model in Appendix A.4.

4.2 Overall Performance

Table 1 presents the performance of all baselines and our method across all datasets. We can conclude:

(1) *Our method outperforms all the baselines and achieves new SOTA results on the four datasets.* Specifically, our method improves the Arg-C F1 score by 1.0%, 0.9%, 0.9%, and 1.9% on ACE05, RAMS, WIKIEVENTS, and MLEE, respectively. This demonstrates the efficacy of our method in handling both sentence-level EAE and document-level EAE tasks.

(2) *Augmenting input with both relevant event instances and templates is beneficial for extracting arguments.* Our DRAAE and DEGAP outperform previous methods, which only utilize event instances or templates to provide insufficient guidance for model. Our approach can fully combine these beneficial cues to better understand the target event from context. In contrast, without these beneficial cues, our base model TabEAE only achieved comparable performance to other baselines.

(3) *Our prefixes can utilize relevant event instances and templates more sufficiently and efficiently to boost performance compared to retrieval.*

As shown in Table 1, Our DEGAP outperforms DRAAE on all datasets. This is because 1) DRAAE uses a pre-trained retriever with frozen parameters to retrieve event instances related to the target event, but fails to find the most relevant events, and 2) DRAAE retrieves and encodes independently, leading to a poor local optimum for each of the steps. In contrast, our prefixes are trained and updated along with model parameters and can fully learn the semantic information of different event instances and templates, avoiding these issues. Moreover, the efficiency analysis in Appendix B.2 indicates that DEGAP has a faster inference speed due to the removal of retrieval.

5 Analysis

In this section, we conduct a comprehensive study to analyze the reasons for the performance improvement of our method, and to validate the significance of our proposed design and the effectiveness of the main components.

5.1 Why is DEGAP Effective?

We explore from two perspectives: 1) domain transfer and 2) handling complex contextual semantics, to investigate whether the performance gains in our method are due to DEGAP providing the model with relevant event instances and event templates as cues. We consider the base model TabEAE, which primarily differs from our model in whether dual prefixes are used.

Domain Transfer. We assume that DEGAP can use relevant event information from the source domain as cues and learn by analogy to better understand new types of events in the target domain. To this end, we examine the model performance when facing domain transfer. Specifically, we select four different domain datasets: ACE05, RAMS,

Model	ACE05	ACE05	ACE05	ACE05	RAMS	RAMS	RAMS	RAMS	WIKI	WIKI	WIKI	WIKI	MLEE	MLEE	MLEE	MLEE	Avg
	↓ ACE05	↓ RAMS	↓ WIKI	↓ MLEE	↓ RAMS	↓ ACE05	↓ WIKI	↓ MLEE	↓ WIKI	↓ ACE05	↓ RAMS	↓ MLEE	↓ MLEE	↓ ACE05	↓ RAMS	↓ WIKI	
TabEAE	72.6	20.1	41.2	5.4	52.5	32.3	20.1	19.6	65.4	48.7	32.6	9.6	71.0	8.5	14.4	9.9	32.7
DEGAP	74.4	23.8	44.8	11.5	54.2	33.9	27.6	25.7	67.1	50.1	34.4	29.3	73.4	21.9	19.1	22.5	38.4

Table 2: Performance (Arg-C F1 score) under different $src \Rightarrow tgt$ settings. We train the model on the src dataset and test it on the tgt dataset. The numbers in the Avg column represent the average of the numbers under all $src \Rightarrow tgt$ settings. WIKIEVENTS is abbreviated as WIKI (the same below).

Model	ACE05		RAMS		WIKI		MLEE	
	S _[185]	M _[218]	S _[587]	M _[284]	S _[114]	M _[251]	S _[175]	M _[2025]
TabEAE	71.2	73.8	52.8	51.6	65.3	65.5	79.3	70.3
DEGAP	72.7 (1.5)	75.8 (2.0)	54.4 (1.6)	54.0 (2.4)	66.5 (1.2)	67.3 (1.8)	81.1 (1.8)	72.8 (2.5)

Table 3: Performance (Arg-C F1 score) on instances with different numbers of events in context. The numbers in square brackets indicate the number of supporting events in the test set. **S**: Single event. **M**: Multiple events.

WIKIEVENTS, and MLEE, train the model on one domain dataset (src), and then test it on all domain datasets (tgt). The results are shown in Table 2. Compared to the results w/o domain transfer, both our method DEGAP and the baseline TabEAE exhibit a sharp decline under different src - tgt transfer settings, where the baseline’s performance even falls below 10% in some cases (e.g., WIKI \Rightarrow MLEE). This indicates that extracting arguments from new types of events is highly challenging for the model. However, encouragingly, our method DEGAP consistently outperforms the baseline in different src - tgt transfer settings. This fully confirms that our hypothesis is correct.

Handling Complex Contextual Semantics. We believe that by fully utilizing relevant event instances and event templates to guide the model, our method excels in identifying the target event from complex contextual semantics. To this end, we explore the ability of models to handle complex semantics from two perspectives: the number of events and the degree of event relevance. **(1) Number of events.** Multiple events in the same context often exhibit complex semantic connections. We measured the performance of the models on instances whose context contains different numbers of events. As shown in Table 3, compared to the baseline TabEAE, our method shows consistent improvements in performance across all datasets, especially in instances that contain multiple events within the same context. **(2) Degree of event relevance.** If two events in the context share arguments, their semantic connection is often more complex. We measured the performance of the model in extracting arguments from overlapping events (events

No.	Model	ACE05	RAMS	WIKI	MLEE
1	DEGAP	74.4	54.2	67.1	73.4
2	DEGAP-CCT	73.6	53.7	66.6	72.3
3	DEGAP-SP	73.9	53.4	66.8	72.6
4	only IOP (w/o TOP)	73.8	52.9	66.4	72.6
5	only TOP (w/o IOP)	73.6	52.7	66.3	72.7
6	w/o DEGAP	72.8	52.0	65.7	71.2
7	DEGAP-TST	73.9	53.0	66.6	72.8

Table 4: Ablations (Arg-C F1 score) on dual prefixes. Note that all models here are equipped with our proposed event-guided adaptive gating mechanism.

with shared arguments). Table 6 shows that our method achieves the best performance across all datasets, particularly gaining greater improvements in handling overlapping events. Experimental results support our conclusion

5.2 Why are Dual Prefixes Necessary?

A core question is why we need to design separate prefixes for event instances and event templates. To explore whether the design of dual prefixes is necessary in our method, we compare DEGAP with two single-prefix variants:

(1) **DEGAP-CCT (Concatenate Context and Template):** Instead of encoding the context and template separately, this variant directly concatenates them and then inputs the combined sequence into the encoder, which is equipped with a shared prefix. For a fair comparison, we use RoBERTa-large as the encoder.

(2) **DEGAP-SP (Shared Prefix):** This variant follows the same encoding process as DEGAP but replaces instance-oriented and template-oriented prefixes with a shared prefix.

Moving forward, we further investigate the effects of instance-oriented prefix (IOP) and template-oriented prefix (TOP). Finally, we consider DEGAP-TST (event-Template-Specific TOP), where each event template has its specific TOP. Therefore, each TOP solely learns and provides semantic information specific to the target event template, leading to insufficient interaction between event templates. Table 4 shows the results.

From the table, we observe that compared to our DEGAP (line 1), the single-prefix variants DEGAP-CCT and DEGAP-SP (lines 2-3) show performance declines to varying degrees across all datasets. This may be because the prefix simultaneously learns the semantic information of event instances and templates, leading to confusion between this information and failing to accurately provide sufficient guidance for model when encoding context and event template. In contrast, dual prefixes separately learn semantic information of event instance and template, which can better provide useful cues for model when encoding context and event template.

Additionally, compared to the baseline w/o DEGAP (line 6), introducing IOP or TOP separately will result in a performance increase on all datasets (lines 4-5), and the performance gains by introducing them at the same time exceed the performance improvement by introducing any one module alone (line 1). This indicates that it is necessary to utilize IOP and TOP to respectively provide the model with relevant event instance information and event template information as cues for inference. Moreover, IOP and TOP can offer complementary guidance to the model from the event mention-level and ontology level, respectively.

Finally, an interesting phenomenon is that despite DEGAP-TST introducing more parameters than DEGAP, its performance declines across all datasets (line 7). This confirms the importance of considering the potential connections between templates, where some relevant templates can usually provide useful information to the target event.

5.3 Is Event-Guided Adaptive Gating mechanism Necessary?

To explore the importance of the Event-Guided Adaptive Gating mechanism (EGAG), we compare DEGAP with the following variants:

(1) **w/o EGAG**: To investigate whether the performance improvement of DEGAP is due to the additional parameters introduced (gating weights and scaling factors) or because event-guided prefixes

No.	Model	ACE05	RAMS	WIKI	MLEE
1	DEGAP	74.4	54.2	67.1	73.4
2	w/o EGAG	73.4	53.3	66.4	72.3
3	DEGAP-S	73.9	53.6	66.8	72.6

Table 5: The study (Arg-C F1 score) of event-guided adaptive gating mechanism.

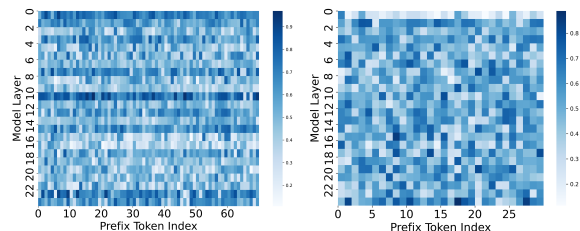


Figure 3: The weight distribution of prefix tokens, which is guided by the target event using the event-guided adaptive gating mechanism.

can better provide model with relevant event information as cues, we designed the variant DEGAP w/o EGAG. For a fair comparison, we expanded the prefix length to 1.5 times to align its parameter count with DEGAP.

(2) **DEGAP-S**: To validate the importance of using the target event to guide prefix, we follow Zhang et al. (2023d) to directly use the sentence representation in the previous layer as h_{i-1}^{trig} in Eq. 1 and h_{i-1}^{type} in Eq. 3, respectively.

Table 5 shows that removing EGAG results in a noticeable decline in Arg-C F1 scores across the four datasets. This phenomenon indicates that utilizing EGAG to connect other relevant events to the target one is very necessary, as it allows our method to further effectively utilize relevant information in the prefix and discard irrelevant information. Additionally, DEGAP-S also shows a decrease in Arg-C F1 scores across the four datasets. This may be because it uses only the first input token “<s>” to calculate h_{i-1}^{trig} and h_{i-1}^{type} , which is insufficient to connect other relevant events to the target event and thus capture relevant information from the prefix. However, compared to the model w/o EGAG, DEGAP-S shows improved performance across all four datasets. This further confirms the importance of utilizing EGAG to connect other relevant events to the target event.

Attention Visualization. We further demonstrate the effects of EGAG visually on the ACE05 dataset. We randomly select a sample and visualize

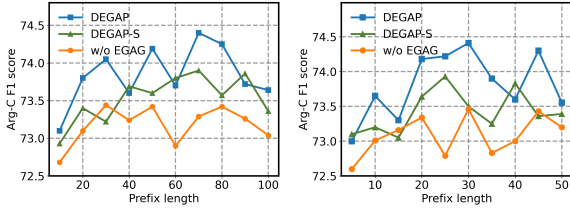


Figure 4: Results under different prefix lengths.

the weight distribution of prefix tokens in IOP and TOP. As shown in Fig. 3, the target event has different levels of attention to the prefixes of different layers. More specifically, even within the same layer, the attention to different prefix tokens varies. This observation confirms that EGAG can adaptively capture the relevance of other event to the given one and further extract relevant information from the prefix tokens at each layer.

Prefix length. We further analyze the impact of prefix length on model performance. Intuitively, a longer prefix implies more trainable parameters, thus enabling it to provide more expressive guidance for the model regarding relevant semantic information. As shown in Fig. 4, as the prefix length increases, the model performance tends to improve, followed by a slight decline. We speculate that this is because too long prefixes provide redundant semantic information, making it difficult for the model to focus on extracting the target event.

6 Conclusion

In this work, we introduce DEGAP. Based on observations of existing research and comprehensive experimental analysis, we show that prefix can effectively learn semantic information from different event instances and templates, and allow them to interact fully. Additionally, we introduce an event-guided adaptive gating mechanism to connect other relevant events to the target event and thus capture relevant information from the prefix. Finally, these event-guided prefixes provide relevant event instances and template information as cues for the target event. Extensive experiments demonstrate the validity of our newly introduced methods. In the future, we plan to integrate other useful auxiliary information into prefixes for solving EAE.

Limitations

We summarize the limitations of our method as follows:

(1) Although experimental results indicate that our method exhibits profound performance, our designed prefixes lack sufficient interpretability compared to methods that directly retrieve relevant event instances and templates to enhance performance (e.g., DEGAP vs DRAAE). Additionally, compared to other methods, our approach requires additional time costs to find suitable prefix lengths for both IOP and TOP.

(2) Due to limited computational resources, we only equipped our prefixes for use on PLMs. In fact, some other works have successfully integrated these efficient prefix parameters into LLMs (e.g., Llama 2 (Touvron et al., 2023)) and demonstrated excellent results (Zhang et al., 2023c). How to effectively combine our prefixes with these LLMs for solving the EAE task remains an open research question.

We leave these issues for future work.

Acknowledgements

This work is supported by the Natural Science Foundation of China (No. 62272206, 62272205, 62462034), the Foundation for Academic and Technical Leaders in Major Disciplines of Jiangxi Province (No. 2021BCJL22041), the Natural Science Foundation of Jiangxi Province (No. 20212ACB202002, 20242BAB25119), and the Science and Technology Research Project from Jiangxi Provincial Department of Education (No. GJJ2200501)

Ethics Statement

This paper focuses on the document-level event argument extraction task, and technically does not involve any ethical concerns. One potential consideration is that practitioners applying this technology should be mindful of possible racial biases present in the datasets used.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: graph attention transformer encoder for cross-lingual relation and event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12462–12470.

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17772–17780.
- Xiang Chen, Lei Li, Shuofei Qiao, Ningyu Zhang, Chuanqi Tan, Yong Jiang, Fei Huang, and Huajun Chen. 2023. One model for all domains: collaborative domain-prefix tuning for cross-domain ner. *arXiv preprint arXiv:2301.10410*.
- George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.
- Xinya Du and Heng Ji. 2022. [Retrieval-augmented generative question answering for event argument extraction](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4649–4666, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Xinya Du, Sha Li, and Heng Ji. 2022. [Dynamic global memory for document-level argument extraction](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5264–5275, 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.
- Yuxin He, Jingyue Hu, and Buzhou Tang. 2023. Revisiting event argument extraction: Can eae models learn better when being aware of event co-occurrences? *arXiv preprint arXiv:2306.00502*.
- I Hsu, Zhiyu Xie, Kuan-Hao Huang, Prem Natarajan, Nanyun Peng, et al. 2023. Ampere: Amr-aware prefix for generation-based event argument extraction model. *arXiv preprint arXiv:2305.16734*.
- I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. [DEGREE: A data-efficient generation-based event extraction model](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.
- Lifu Huang, Heng Ji, Kyunghyun Cho, and Clare R Voss. 2017. Zero-shot transfer learning for event extraction. *arXiv preprint arXiv:1707.01066*.
- 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.
- 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.
- 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.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). 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 4582–4597, Online. Association for Computational Linguistics.
- Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1641–1651.
- Jian Liu, Yufeng Chen, and Jinan Xu. 2021. [Machine reading comprehension as data augmentation: A case study on implicit event argument extraction](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2725, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023a. [RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8404–8419, Toronto, Canada. Association for Computational Linguistics.
- Wanlong Liu, Shaohuan Cheng, Dingyi Zeng, and Qu Hong. 2023b. [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, Shao-huan Cheng, Chen Zhang, Grandee Lee, Malu Zhang, and Wenyu Chen. 2024. Beyond single-event extraction: Towards efficient document-level multi-event argument extraction. *arXiv preprint arXiv:2405.01884*.
- Xiao Liu, Heyan Huang, Ge Shi, and Bo Wang. 2022a. Dynamic prefix-tuning for generative template-based event extraction. *arXiv preprint arXiv:2205.06166*.
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. [P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. 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*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Di Lu, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2023. [Event extraction as question generation and answering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1666–1688, Toronto, Canada. Association for Computational Linguistics.
- 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. *arXiv preprint arXiv:2202.12109*.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.
- 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.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Hanchchol Cho, Jun’ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- 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.
- Ge Shi, Yunyue Su, Yongliang Ma, and Ming Zhou. 2023. A hybrid detection and generation framework with separate encoders for event extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3163–3180.
- Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. [AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.
- Xiangru Tang, Yiming Zong, Yilun Zhao, Arman Cohan, and Mark Gerstein. 2023. Struc-bench: Are large language models really good at generating complex structured data? *arXiv preprint arXiv:2309.08963*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutli Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Hai-Long Trieu, Thy Thy Tran, Khoa NA 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.
- Md Nayem Uddin, Enfa Rose George, Eduardo Blanco, and Steven Corman. 2024. Generating uncontextualized and contextualized questions for document-level event argument extraction. *arXiv preprint arXiv:2404.04770*.
- Amir Pouran Ben Veyseh, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Graph transformer networks with syntactic and semantic structures for event argument extraction. *arXiv preprint arXiv:2010.13391*.
- David Wadden, Ulme Wennberg, Yi Luan, and Hananeh Hajishirzi. 2019. [Entity, relation, and event extraction with contextualized span representations](#).

- In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.
- Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. [Universal adversarial triggers for attacking and analyzing NLP](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.
- Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. *arXiv preprint arXiv:2203.08773*.
- Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36.
- 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.
- 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. *arXiv preprint arXiv:2305.19162*.
- Qi Zeng, Qiusi Zhan, and Heng Ji. 2022. [EA²E: Improving consistency with event awareness for document-level argument extraction](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2649–2655, Seattle, United States. Association for Computational Linguistics.
- Kaihang Zhang, Kai Shuang, Xinyue Yang, Xuyang Yao, and Jinyu Guo. 2023a. [What is overlap knowledge in event argument extraction? APE: A cross-datasets transfer learning model for EAE](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 393–409, Toronto, Canada. Association for Computational Linguistics.
- Kaihang Zhang, Kai Shuang, Xinyue Yang, Xuyang Yao, and Jinyu Guo. 2023b. [What is overlap knowledge in event argument extraction? ape: A cross-datasets transfer learning model for eae](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 393–409, Toronto, Canada. Association for Computational Linguistics.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023c. [Llama-adapter: Efficient fine-tuning of language models with zero-init attention](#). *arXiv preprint arXiv:2303.16199*.
- Zhen-Ru Zhang, Chuanqi Tan, Haiyang Xu, Chengyu Wang, Jun Huang, and Songfang Huang. 2023d. [Towards adaptive prefix tuning for parameter-efficient language model fine-tuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1239–1248, Toronto, Canada. Association for Computational Linguistics.
- Gang Zhao, Xiaocheng Gong, Xinjie Yang, Guanting Dong, Shudong Lu, and Si Li. 2023a. [Demosg: Demonstration-enhanced schema-guided generation for low-resource event extraction](#). *arXiv preprint arXiv:2310.10481*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023b. [A survey of large language models](#). *arXiv preprint arXiv:2303.18223*.
- Hanzhang Zhou, Junlang Qian, Zijian Feng, Hui Lu, Zixiao Zhu, and Kezhi Mao. 2023. [Heuristics-driven link-of-analogy prompting: Enhancing large language models for document-level event argument extraction](#). *arXiv preprint arXiv:2311.06555*.

A Experimental Setup

A.1 Datasets

ACE05 (Doddington et al., 2004)³ is a joint information extraction dataset that compiles news, telephone conversations, and broadcast news, providing annotations for entities, relations, and events in English, Chinese, and Arabic. We utilize only its English event annotations for sentence-level EAE task, adopting the data preprocessing methods used by Wadden et al. (2019). This dataset assumes that events are contained within a single sentence, rather than in a document, which is more typical in real-world scenarios.

RAMS (Ebner et al., 2020)⁴ is a document-level EAE dataset focused on English online news. Each example is a document consisting of five sentences, where the trigger word indicates a pre-defined event type, and their arguments are dispersed throughout the document. Therefore, a major challenge of this dataset is to address the long-distance dependencies between event triggers and arguments.

WIKIEVENTS (Li et al., 2021)⁵ is a document-level EAE dataset. The documents in this dataset are derived from English Wikipedia articles that describe real-world events, followed by the retrieval of related news articles based on reference links. They also annotate coreference links for arguments; however, in this task, we only utilize annotations for conventional arguments. Compared to RAMS, WIKIEVENTS contains a greater number of multi-event documents.

MLEE (Pyysalo et al., 2012)⁶ is a document-level event extraction dataset containing manually annotated summaries of biomedical publications written in English. We follow the preprocessing procedure outlined by Trieu et al. (2020). As the preprocessed data only includes training/test splits, we follow He et al. (2023) to use the training set as the development set. Moreover, overlapping events are more common in MLEE than in the other three datasets.

Data Statistics. Table 7 shows detailed statistics for each dataset.

³<https://catalog.ldc.upenn.edu/LDC2006T06>

⁴<https://nlp.jhu.edu/rams/>

⁵<https://github.com/raspberryyice/gen-arg>

⁶<http://www.nactem.ac.uk/MLEE/>

A.2 Implementation details

We implement DEGAP using PyTorch and conduct experiments on a single NVIDIA GeForce RTX 3090 24G. For simplicity, we randomly initialize the embeddings of the prefix. Following TabEAE’s setting (He et al., 2023), we use the first 17 layers of RoBERTa-large (Liu et al., 2019) to instantiate the encoder, and the weights of the remaining 7 layers to initialize the weight of self-attention layers and feedforward layers of the decoder. We also randomly initialize the weight of cross-attention parts of the decoder. The optimization of our model employs the AdamW optimizer (Loshchilov and Hutter, 2017) equipped with a linear learning rate scheduler. Empirical studies on the selection of the number of encoder/decoder layers can be found in (He et al., 2023).

Most hyperparameters are the same as those in TabEAE (He et al., 2023). For hyperparameters requiring tuning, we conduct grid searches in specific ranges for each dataset, choosing values that maximize the Arg-C F1 on the development set. The search ranges and the final hyperparameter values are listed in Table 8.

A.3 Baselines

We compare our method with the following state-of-the-art models:

- **Bart-Gen** (Li et al., 2021): A conditional generation model that uses a sequence-to-sequence model and pre-defined templates to sequentially generate arguments. We report the results from He et al. (2023).
- **DEGREE** (Hsu et al., 2022): A conditional generation model. In addition to templates, event keywords, and event descriptions are also used as auxiliary information. We report the results from Hsu et al. (2023).
- **AHR** (Ren et al., 2023): A model that utilizes retrieved relevant instances to facilitate sequence generation. The original paper offers three retrieval strategies: Context-Consistency Retrieval, Schema-Consistency Retrieval, and Adaptive Hybrid Retrieval. We report their best-performing method, Adaptive Hybrid Retrieval (AHR).
- **R-GQA** (Du and Ji, 2022): A model that generates arguments given a retrieved demonstration and natural questions. We report the results from the original paper.

Model	ACE05		RAMS		WIKI		MLEE	
	N-O _[319]	Overlap _[84]	N-O _[690]	Overlap _[181]	N-O _[296]	Overlap _[69]	N-O _[1466]	Overlap _[734]
TabEAE	71.1	78.6	51.6	55.6	65.7	64.4	75.4	65.8
DEGAP	72.1 (1.0)	80.5 (1.9)	52.9 (1.3)	59.5 (3.9)	67.3 (1.6)	66.7 (2.3)	76.9 (1.5)	68.5 (2.7)

Table 6: Results (Arg-C F1 score) of extracting overlapping (events with shared arguments) and non-overlapping events. Numbers in square brackets indicate the number of supporting events in the test set. **N-O**: Non-overlapping.

Dataset	ACE05	RAMS	WIKIEVENTS	MLEE
# Event types	33	139	50	23
# Args per event	1.19	2.33	1.40	1.29
# Events per text	1.35	1.25	1.78	3.32
# Events				
Train	4202	7329	3241	4442
Dev	450	924	345	-
Test	403	871	365	2200

Table 7: Statistics of datasets.

- **PAIE** (Ma et al., 2022): A QA-based model that extracts arguments by querying context using role slots in the event template. We report the results from He et al. (2023).
- **SPEAE** (Nguyen et al., 2023): A model based on PAIE that utilizes prefixes to integrate relevant context information into the model and alleviate sub-optimal manual prompts. We report the results from the original paper.
- **TabEAE** (He et al., 2023): Building upon PAIE, it simultaneously extracts multiple events and replaces the base model BART used by PAIE with RoBERTa. The original paper introduces three training-inference paradigms: Multi-Multi, Multi-Single, and Single-Single (consistent with our training-inference paradigm). To ensure a fair comparison and visually demonstrate the effectiveness of our proposed prefix, we report the performance of its Single-Single version.

A.4 The implementation details of DRAAE

We follow (Ren et al., 2023) to use a frozen S-BERT (Reimers and Gurevych, 2019) to retrieve the top- k_{ins} relevant instance \mathbf{d}_{ins} and the top- k_{tem} relevant templates \mathbf{d}_{tem} from training corpus for the input instance and template, respectively. Then, we use RoBERTa-large as an encoder to obtain the retrieved instance embedding $\mathbf{h}_{\mathbf{d}_{ins}}$ and template embedding $\mathbf{h}_{\mathbf{d}_{tem}}$ through the representation of the

first input token:

$$\begin{aligned}\mathbf{h}_{\mathbf{d}_{ins}} &= \text{RoBERTa}(\mathbf{d}_{ins}), \\ \mathbf{h}_{\mathbf{d}_{tem}} &= \text{RoBERTa}(\mathbf{d}_{tem}),\end{aligned}\quad (13)$$

where $\mathbf{h}_{\mathbf{d}_{ins}} = \{\mathbf{h}_{\mathbf{d}_{ins}}^{(1)}, \mathbf{h}_{\mathbf{d}_{ins}}^{(2)}, \dots, \mathbf{h}_{\mathbf{d}_{ins}}^{(k_{ins})}\}$, and $\mathbf{h}_{\mathbf{d}_{tem}} = \{\mathbf{h}_{\mathbf{d}_{tem}}^{(1)}, \mathbf{h}_{\mathbf{d}_{tem}}^{(2)}, \dots, \mathbf{h}_{\mathbf{d}_{tem}}^{(k_{tem})}\}$. These retrieved instance contexts and event templates are incorporated into each transformer layer’s self-attention module.

The new self-attention with retrieved event information is formulated as:

$$\begin{aligned}H &\leftarrow \text{LayerNorm}(H' + H), \\ H' &= \text{MHSA}(\mathbf{h}_{\mathbf{d}} \oplus H)_{|\mathbf{h}_{\mathbf{d}}|:|\mathbf{h}_{\mathbf{d}} \oplus H|},\end{aligned}\quad (14)$$

where $\mathbf{h}_{\mathbf{d}}$ represents either $\mathbf{h}_{\mathbf{d}_{ins}}$ or $\mathbf{h}_{\mathbf{d}_{tem}}$, H is the hidden state.

Finally, we use RoBERTa in Eq. 13 to encode the input context and event template in the same manner as in DEGAP (§ 3.1.4), and further extract argument spans (§ 3.1.5).

B Experimental Analysis

B.1 Performance Comparison with LLMs

In this section, we compare our method with three publicly available LLMs: GPT-3 (Brown et al., 2020) in its text-davinci-003 and GPT-3.5-turbo-instruct versions (Ouyang et al., 2022), and GPT-4 (Achiam et al., 2023).

Hyperparameters	Search Range	ACE05	RAMS	WIKIEVENTS	MLEE
Batch Size	[2, 16]	16	4	4	4
Training Steps	-	10000	10000	10000	10000
Window Size	-	250	250	250	250
Max Span Length	-	10	10	10	10
Learning Rate	-	2e-5	2e-5	2e-5	2e-5
Warmup Ratio	-	0.1	0.1	0.1	0.1
Max Gradient Norm	-	5	5	5	5
Max Encoder Seq Length	-	200	500	500	500
Max Decoder Seq Length	-	250	200	360	360
IOP Length	[10, 100]	70	40	40	70
TOP Length	[5, 50]	30	20	10	40

Table 8: Hyperparameters for DEGAP. The hyperparameters without a search range are set to the same values as in He et al. (2023).

Model	RAMS	
	Arg-I	Arg-C
text-davinci-003 (Brown et al., 2020)	46.2	39.6
GPT-3.5-turbo-instruct (Ouyang et al., 2022)	43.3	37.0
GPT-4 (Achiam et al., 2023)	52.4	44.1
DEGAP (Ours)	58.5	54.2

Table 9: Performance comparison with LLMs. We report their performance from (Zhou et al., 2023).

Table 9 shows that our method achieves improvements of 6.1% and 10.1% in Arg-I and Arg-C F1 scores respectively, emphasizing our method’s significant advantage over LLMs. This makes sense because our model are trained using task-specific data, whereas LLMs can only perform in-context learning through one or two demonstrations. This observation is also consistent with previous findings, suggesting that despite LLMs making substantial progress in many natural language processing tasks (Zhao et al., 2023b; Muennighoff et al., 2022) such as text generation tasks (Qin et al., 2023), further efforts are still needed for them to excel in generating complex structured outputs (e.g., generating structured event information) (Tang et al., 2023; Chen et al., 2024).

B.2 Efficiency Analysis

Table 10 reports the overall inference time for DEGAP (w/o retrieval) and DRAAE (w/ retrieval). It can be observed that DEGAP runs approximately 6-7 times faster than DRAAE on RAMS, WIKIEVENTS, and MLEE, and approximately 15 times faster than DRAAE on ACE05. This is mainly because: 1) DRAAE requires additional

Model	ACE05	RAMS	WIKIEVENTS	MLEE
DRAAE	61.18	256.87	121.76	626.31
TabEAE	3.69	41.35	18.11	71.96
DEGAP	4.36	43.36	19.64	88.60

Table 10: Inference time (second) for different models on test set of four datasets. Experiments are run on one same NVIDIA GeForce RTX 3090.

time costs to retrieve the top- k similar event instances and templates; 2) For the retrieved event instances and templates, DRAAE needs to feed them into the encoder to obtain meaningful embeddings. In contrast, DEGAP does not require retrieval and further encoding. Only equipped with a set of learnable prefixes, our DEGAP can adaptively learn semantic information from different event instances and templates during training, and provide relevant guidance at each model layer during inference. Note that due to the parallelized attention calculation for the introduced prefixes and original input on the GPU, there is a negligible impact on inference speed compared to the base model TabEAE.

B.3 Event Template Variant

As described in § 3.1.1, the event template \mathcal{T}_e consists of two parts: the type part \mathcal{I}_e and the schema part \mathcal{S}_e . To explore the necessity of type part \mathcal{I}_e , following Shi et al. (2023), we use type markers “<Event type>” and “</Event type>” as learnable special tokens, inserting them respectively into the corresponding schema part \mathcal{S}_e as the template for event type e . Additionally, we compute Eq. 3 by averaging the representations of

the type markers.

Model	ACE05	RAMS	WIKIEVENTS	MLEE
w/ type part	74.4	54.2	67.1	73.4
r/ type part	74.2	53.9	67.3	73.1

Table 11: Analysis of the necessity for the Type part. The numbers in the table all refer to the Arg-C F1 scores. **r/**: Replace the type part with type markers.

The experimental results in Table 11 suggest that describing the event type with natural language sentence generally outperforms the use of type markers. Consequently, we retain this setting in our event template \mathcal{T}_e . Note that as our method does not rely on type markers that need to be learned, it can more easily adapt to new event types.