Optimizing Low-Resource Zero-Shot Event Argument Classification with Flash-Attention and Global Constraints Enhanced ALBERT Model

Tongyue Sun[1], Jiayi Xiao[2]
School of Engineering and Informatics, University of Sussex, Brighton, UK[1]
International Business School Suzhou, Xi'an Jiaotong-Liverpool University, Suzhou, China[2]
Management School, University of Liverpool, Liverpool, UK[2]

Abstract*—Event Argument Classification (EAC) is an essential subtask of event extraction. Most previous supervised models rely on costly annotations, and reducing the demand for computational and data resources in resource-constrained environments is a significant challenge within the field. We propose a Zero-Shot EAC model, ALBERT-F, which leverages the efficiency of the ALBERT architecture combined with the Flash-Attention mechanism. This novel integration aims to address the limitations of traditional EAC methods, which often require extensive manual annotations and significant computational resources. The ALBERT-F model simplifies the design by factorizing embedding parameters, while Flash-Attention enhances computational speed and reduces memory access overhead. With the addition of global constraints and prompting, ALBERT-F improves the generalizability of the model to unseen events. Our experiments on the ACE dataset show that ALBERT-F outperforms the Zero-shot BERT baseline by achieving at least a 3.4% increase in F1 score. Moreover, the model demonstrates a substantial reduction in GPU memory consumption by 75.1% and processing time by 33.3%, underscoring its suitability for environments with constrained resources.

Keywords*—*Artificial intelligence; natural language processing; event argument classification; zero-shot learning; flash-Attention; global constraints; low-resource

I. Introduction

Event Argument Classification (EAC) is a crucial part of event understanding and event argument extraction, embodying the complexity and importance of this interdisciplinary field [1, 2]. This domain, which integrates natural language processing (NLP) and knowledge representation, is dedicated to converting narrative event descriptions and their relational dynamics into a structured form of knowledge. As shown in Fig. 1, for a trigger word "paid" in a "Transfer-Money" event, it has several argument spans (e.g., "O'neal"). By determining the roles of these arguments (e.g., identifying "O'neal" as the "Giver"), this structured knowledge enables us to better understand events and use them for knowledge reasoning and automated decision support, benefiting applications such as biomedical research and question answering recommendation systems.

In the domain of event argument classification, a prevalent strategy has been the manual annotation of domains and patterns. Although effective, this approach necessitates significant labeling efforts for model training. This method also presents challenges in transferring knowledge across different application domains and scaling to new datasets. The laborious nature of annotation incurs significant costs. To mitigate this, some EAC models have turned to few-shot learning [3–6], which, despite its potential, is sensitive to the selection of examples and requires costly, task-specific training, limiting its practicality.In contrast, zero-shot EAC models have been introduced, leveraging label semantic understanding or prompt learning strategies [7–10]. Although existing methods perform well when dealing with events similar to the training data, they may not achieve the expected results when faced with significantly different new events. Some studies have attempted to improve performance in zero-shot and few-shot learning scenarios by integrating Large Language Models (LLMs) [2, 11], but there is still a considerable gap compared to models that have been specifically fine-tuned. Moreover, the operation of LLMs requires a significant amount of computational resources, which may limit their potential for application in resource-constrained environments. Therefore, tackling the efficiency constraints inherent to Zero-Shot EAC tasks in resource-scarce environments has become a formidable obstacle.

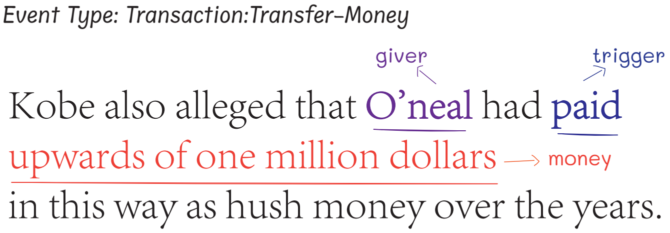
Event Type: Transaction:Transfer-Money
giver
trigger
Kobe also alleged that O'neal had paid
upwards of one million dollars
money
in this way as hush money over the years.

Fig. 1. An example of EAC. The arrows indicate the trigger and argument types respectively.

To address the challenges in the field of event argument classification, we propose a Zero-Shot model tailored for low-resource scenarios. This model integrates an ALBERT architecture [12] optimized with Flash-Attention [13] and is enhanced by global constraints with prompting, aiming to improve the performance of zero-shot EAC tasks.

The ALBERT model mitigates the issues of excessive parameters and inefficiency by simplifying its design in the BERT [12, 14]. Furthermore, global constraints provide critical supervisory guidance to our model, as a manifestation of domain knowledge [2]. This guidance is particularly crucial in zero-shot learning environments with a scarcity of fully annotated data, as it enables the model to better understand and generalize the relations between event arguments. To further accommodate the constrained resources in low-resource scenarios, we propose ALBERT-F, a solution that optimizes the ALBERT model using a Flash Attention module. Flash Attention leverages efficient upper-level storage computational units to reduce access to the slower lower-level storage, thereby maintaining performance while significantly enhancing the model's resource utilization efficiency [13].

Through a series of experiments, we have validated the effectiveness of our proposed method. Specifically, our approach achieved at least a 3.4% increase in F1 score on the ACE dataset compared to the Zero-shot BERT baseline model, with a 75.1% reduction in GPU memory consumption and a 33.3% reduction in processing time. Furthermore, the introduction of Flash Attention resulted in a further 5.1% reduction in GPU memory consumption and an 11.1% decrease in processing time compared to the original ALBERT model. These results not only demonstrate the significant advantage of our method in reducing resource consumption but also confirm its effectiveness in enhancing performance.

Subsequent sections present our experimental setup, results, and a comparative analysis with existing models. We conclude with a discussion on the implications of our findings, limitations and avenues for future research.

II. Related Works

A. Event Extraction

Event Extraction (EE) is one of the most fundamental tasks in information extraction, which can be further divided into four subtasks: trigger identification, trigger classification, argument identification, and argument classification [1, 15–18].

Traditional event extraction relies heavily on feature engineering, which poses its central challenge [1, 18]. However, these methods often encounter limitations when dealing with deep or complex nonlinear patterns. In recent years, some advanced works based on supervised learning have attracted attention due to their two main advantages: first, the applicability of their embedding representations to large-scale datasets; second, the combination of automated feature extraction with specific deep architectures, which effectively captures more intricate nonlinear patterns [19–23].

In the task of information extraction, models can identify the actions in sentences and their corresponding participants by defining constraints [24]. One of the applications of constraint modeling in NLP is in syntactic analysis, where it is used to represent that an object must satisfy general or very specific properties to exclude those that do not belong to the structure of the languag [25]. Particularly in zero-shot scenarios, constraint modeling can provide useful indirect supervision to the model, thereby further improving its performance [26].

Nevertheless, the inherent limitations of supervised learning may impact the model's generalization capabilities across different domains. Moreover, the demand for computational resources and specialized skills (including constraint modeling), along with the reliance on a substantial amount of manually annotated data, become bottlenecks in their practical application.

B. Few-Shot Learning for EE

Few-shot learning methods have garnered widespread attention in the domain of event extraction, and the majority of current research is concentrated on the task of event identification within the context of Few-shot Event Detection (FSED) [1]. These approaches are dedicated to achieving accurate predictions for specific tasks with minimal training samples, such as one-shot, five-shot, etc. By leveraging prior knowledge, transfer learning, or meta-learning strategies, few-shot learning endeavors to surmount the challenge of data scarcity and enhance the model's generalization capability on novel tasks [3–6].

The DEGREE model [6] excels at synthesizing events from a text segment into coherent, naturally constructed sentences that conform to a pre-established template, aided by manually curated prompts. By integrating the semantic essence of labels with the collective intelligence across sub-tasks, DEGREE discerns interdependencies among entities, thereby reducing the volume of training data required. Many previous works on event extraction (EE) necessitate extensive annotations for model training [6, 8, 23], which incurs high costs due to the labor-intensive nature of annotation and poses challenges in scaling to new domains. While DEGREE refines a pre-existing generative language model [27], the output it generates may reflect the characteristics of the corpus from which it was trained. Although infrequent, there is a possibility that the model might produce sentences that are malevolent, mendacious, or prejudiced, thus raising ethical concerns [28, 29].

For classification tasks, LoLoss [4] is employed for training few-shot learning models based on the matching information of examples within the support set.

$$L(x, S) = L_{\text{query}}(x, S) + \lambda \cdot L_{\text{aux}}(S) \tag{1}$$

The components of this equation are as follows: $L(x, S)$ represents the total loss, which is contingent upon the model parameters x and the training samples S. $L_{\text{query}}(x, S)$ refers to the query loss, which assesses the discrepancy between the model's predictions for the query set and their corresponding true labels. λ is a hyperparameter that modulates the trade-off between the query loss and the auxiliary loss within the total loss calculation. $L_{\text{aux}}(S)$ is the auxiliary loss, which leverages the internal matching information of examples within the support set to provide additional training signals. It not only matches the query examples with those in the support set but also further matches the examples among themselves within the support set, thereby providing additional training signals for the model.

The scarcity of samples in long-tail categories increases the complexity of classification in few-shot learning tasks. To overcome this challenge, the Multi-Level Matching and Aggregation Network (MLMAN) [3] employs a hierarchical matching and aggregation strategy. This strategy comprehensively analyzes the support vectors and query vectors at different levels, capturin local features integrating global contextual information, thereby enhancing the classification accuracy of long-tail category samples. Adaptive Attentional Network for Few-Shot Knowledge Graph Completion (FAAN) [5] employs a minimal set of reference samples to adeptly predict and discern connections and relations. These reference relation triples are adaptively encoded within a transformer network through the application of embeddings and an attention mechanism, ensuring precise alignment with the query. FAAN's adaptive encoding of entities and reference pairs significantly enhances the performance of traditional knowledge graph embedding methods, particularly for long-tail relations that are characterized by a paucity of samples.

Nonetheless, constrained by a limited sample size, the model is susceptible to overfitting, with an exacerbated risk in scenarios characterized by class imbalance. This propensity may compromise the model's capacity for robust generalization. Furthermore, the necessity for supplementary computational resources or the adoption of intricate model architectures could potentially restrict the practical applicability of these models.

C. Zero-shot Learning for EE

In the context of lacking prior knowledge and labeled data, existing research tends to adopt preset event information frameworks or experience-based strategies to achieve effective classification of unknown event types [7–10]. Similarly, the zero-shot contrastive learning strategy also emphasizes the use of unlabeled data during the training phase to cultivate features that can distinguish between different categories [2]. Although these methods still have a significant gap compared to supervised methods, they offer an insightful perspective and suggest possible directions for improvement in event extraction under resource-constrained environments.

The event extraction task was conceptualized by Huang et al. [7] as a "grounding" problem, wherein it is encapsulated within a structured ontology that delineates event mentions and their respective types. Semantic similarity measures are harnessed for the purpose of prediction. A transferable neural architecture was proposed by Huang et al., one that capitalizes on manually annotated event patterns alongside a modest subset of previously encountered types. This architecture was adept at transferring knowledge from known types to the extraction of novel types, thereby enhancing the scalability of event extraction and conserving human resources. Further exploration into transfer learning methodologies for novel events was undertaken by Lyu et al. [9] They reframed the event extraction challenge within the contexts of textual entailment (TE) and question answering (QA), advocating for the direct application of pre-trained TE/QA models.

Although these models have demonstrated exceptional performance on standard benchmark tests, they have not yet realized the anticipated generalization effect when applied to the event extraction dataset. Nonetheless, they offer an insightful vision and suggest a possible direction for improvement in event extraction within very low-resource environments.

Lin et al. [2] proposed a Global Constraint Regularization Module that standardizes predictions through three types of global constraints: cross-task constraints, cross-parameter constraints, and cross-event constraints. They utilized a method that combines global constraints with prompting, employing the large language model GPT-J [30], which makes it possible to effectively perform event parameter classification without any annotations or task-specific training. Chen et al. [11], also employing large language models for research, utilized a large language model as an expert annotator for event extraction. Strategically incorporating sample data from the training dataset into the prompts, the researchers ensure that the generated samples from the language model align with the data distribution of the benchmark dataset. This enables the creation of an augmented dataset to supplement the existing benchmarks, alleviating challenges of data imbalance and scarcity, thus enhancing the performance of fine-tuned models. However, existing open-source large language models often require expensive hardware configurations and substantial computational resources [31]. Furthermore, the utility of these models is limited by the fact that most current hardware was developed prior to the emergence of large-scale models, potentially rendering it inadequate for the computational demands of such models during inference. This limitation is particularly pronounced in low-resource settings, where specialized hardware is required to facilitate efficient inference processes for large models [32].

To address low-resource scenarios, we propose a Zero-shot EAC model that incorporates global constraints and prompt, coupled with ALBERT-F. This approach aims to enhance the performance of Zero-shot EAC tasks in resource-constrained environments.

III. Methodology

Our model comprises two distinct modules. As shown in Fig. 2, the first is the prompting module, which is tasked with the generation of several new passages and the subsequent evaluation of their quality. During this creation process, the model integrates candidate role with prefix prompts that contain information regarding the event type and trigger. These candidates are connected to the target parameter range by embedding them within the passages through a cloze prompt. Subsequently, the model employs an ALBERT model optimized with Flash-Attention (ALBERT-F) to score the newly generated passages. Without the need for manual annotation, the initial prediction is the role with the highest prompting score. The second module is the global constraint regularization module, wherein the model regularizes the predictions through three types of global constraints. These are based on domain knowledge related to inter-task, inter-parameter, and inter-event relationships within the event-related context.

A. ALBERT-F

Before delving into the two primary modules, we provide a overview of ALBERT-F. By substituting the attention mechanisms across all modules, the primary structure of our ALBERT network is depicted in Fig. 3.

Flash-attention is designed to expedite the computation of attention mechanisms and curtail memory usage [13]. It leverages the knowledge of the memory hierarchy of underlying hardware, such as the memory architecture of GPUs, to enhance computational speed and reduce the overhead of memory access. By using statistical measures and altering

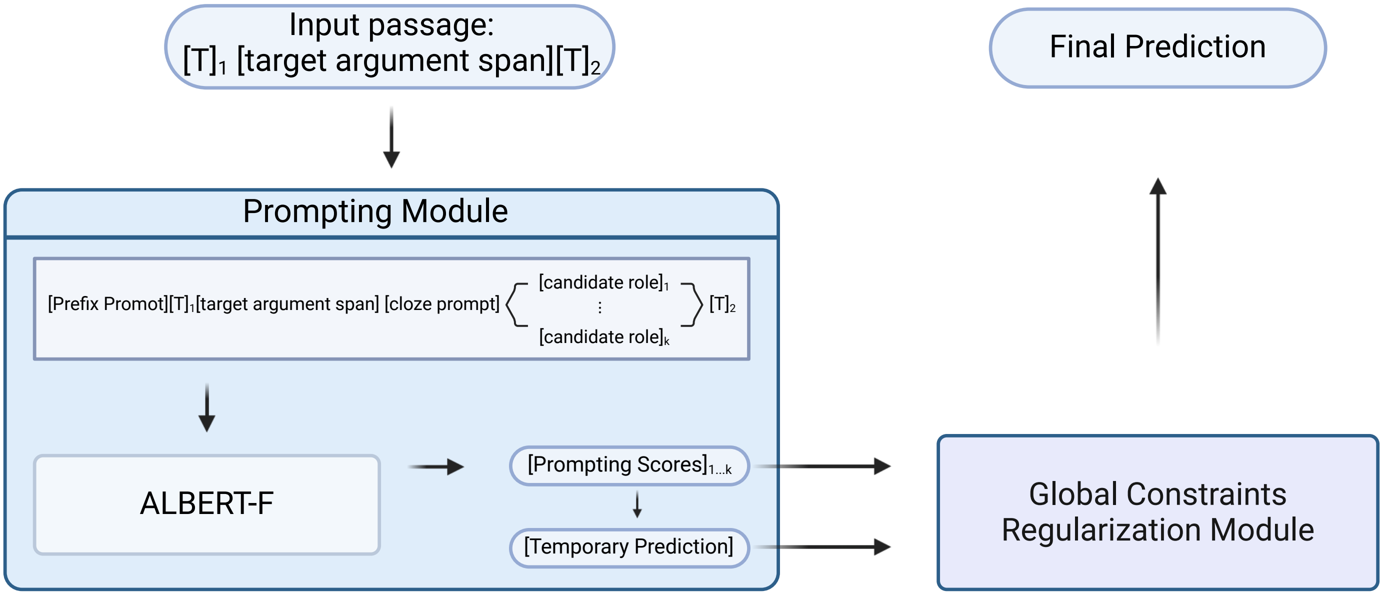

Fig. 2. Model summary, illustrated with the prediction of a single argument span. $[T]_1$ represents the segment of the input text preceding the span, while $[T]_2$ denotes the segment that follows. The variable k signifies the total count of potential roles associated with the event type.

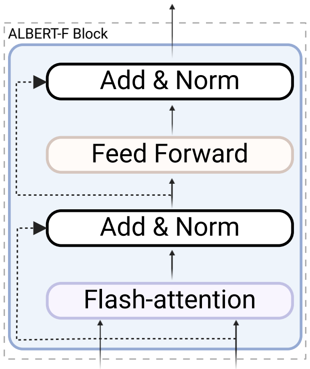

Fig. 3. The main structure of the ALBERT-F module. Enhancing model efficiency by introducing flash-attention to reduce computational resource consumption.

the computation sequence of the attention mechanism, Flash-attention computes in chunks rather than approximates, effectively reducing complexity. The outcomes of Flash-attention are entirely equivalent to those of the native attention mechanism [13, 33].

Increasing the size of a Pre-trained Language Model (PLM) typically enhances its inferential capabilities; however, once the model reaches a certain magnitude, it encounters limitations imposed by the memory capacity of GPUs/TPUs [33]. Consequently, ALBERT implements factorized embedding parameterization, which decomposes the embedding matrix. Instead of directly projecting one-hot encoded vectors into a hidden vector space of dimension H, the vectors are first projected into a lower-dimensional embedding vector space and then into the hidden vector space. This decomposition significantly reduces the number of embedding parameters and results in a more uniform distribution.

In ALBERT, the parameters of the fully connected layers and the attention layers are shared, meaning that ALBERT retains the deep multi-layer connections, but the parameters between layers are identical [12]. Consequently, ALBERT-F optimizes ALBERT using Flash-attention to reduce the model's computational resource consumption, yielding more satisfactory results in low-resource scenarios where computation and memory are constrained.

B. *Prompting Module*

In this section, we primarily elucidate the prompting module. Given a passage, we initially append a prefix prompt at the onset, which encapsulates information pertaining to the event type and the scope of the trigger. Such a prompt serves to guide ALBERT-F in: (1) accurately capturing the correlation between the input text and the event-related associations; (2) possessing a clear trigger awareness capability. In accordance with the definitions of events and triggers [17], we have formulated the following prefix prompt:

- "This is a $[P_1]$ event whose occurrence is most clearly expressed by $[P_2]$."

Where the first and second pairs of square brackets are placeholders for the event type(P_1) and trigger span (P_2), respectively.

For each candidate role, the module inserts a cloze test prompt subsequent to the target parameter range, with the role filling the slot of the prompt. The cloze prompt employs a hypernym extraction pattern "M and any other []", wherein "M" denotes the parameter range, and the square brackets serve as placeholders for the candidate roles. Such prompts harness the linguistic and common-sense knowledge stored within the ALBERT-F to assist in identifying which candidate role is the most plausible [2, 34].

For each novel passage, we apply ALBERT-F to compute the language modeling loss. The prompting score for the corresponding paragraph is determined by the negative value of the loss, where a more negative loss indicates a higher score, reflecting greater plausibility as assessed by ALBERT-F. Since the scoring process for each candidate role is independent of other candidate roles, we implement the steps for different candidate roles in parallel. This parallel implementation significantly enhances the efficiency of our model.

C. Global Constraints Regularization Module

This module regularizes predictions through global constraints. We refer to and leverage the constraint strategy proposed by Lin et al. [2], employing Event Argument Entity Typing (EAET) as an auxiliary task, which aims to categorize arguments into their contextually relevant entity types. These constraints provide our model with a global understanding of event arguments.

By utilizing the label dependencies between EAC and the auxiliary task, our model can glean global information about event arguments from the auxiliary task. Concurrently, to adapt to applications in low-resource scenarios, our model limits the number of specific arguments for certain or all events.

IV. Experiments

We initially present the experimental setup, the baselines for comparison, and certain implementation details. Subsequently, we demonstrate and analyze the results of the experiments. We then conduct an analysis of computational resources and processing duration. Finally, we perform an error analysis.

A. Settings

We utilize the ACE (2005-E +) dataset [23, 35] as the basis for our experiments. The ACE dataset encompasses a total of 33 event types and 22 roles. We preprocess all events, as is done in the work of Lin et al. [23], to retain only the event subtypes when applicable. Since our approach is zero-shot, for each dataset, we consolidate all splits into a single test set, following the preprocessing in the study by Lyu et al. [9].

We evaluate using the F1 score, as proposed by Ji and Grishman [36], employing the ALBERT-F model as the foundational model for our module implementation. We run our experiments on a single NVIDIA RTX 4000 Ada GPU.

B. Main Results

We report results that are compared with several existing zero-shot methods, including those by Huang et al. (2018) [7], Liu et al. (2020) [8], Zhang et al. (2021) [10], and the current state-of-the-art zero-shot approach by Lin et al. (2023) [2]. In our comparisons, we also evaluated the work of Lin et al., where we compared two PLM (Pre-trained Language Model) bases: Bert-large-uncased [12] with a parameter count of 330 million, and GPT-6J [30] with a parameter count of 6 billion.

From Table I, we have the following observations: Compared to all zero-shot baselines, our model has demonstrated superior performance in the Settings category. Specifically, our model has achieved an F1 score on the ACE dataset that surpasses the best non-large model zero-shot baseline by 3.4% (Lin et al., 2023 [2]). This represents a significant gap. Such substantial performance improvement can be attributed to several factors: (1) the prefix and cloze prompts effectively guide the PLM to capture the input's event-related perspectives and triggers; (2) the global constraint regularization incorporates global information and domain knowledge into the inference process; (3) our model has effectively enhanced the inferential capabilities of the EAC (Event Argument Classification) task.

TABLE I. Comparison F1 Score of Different Models on the ACE 2005 E+ Dataset, the Best Performance of Non-LLM is Marked in Bold Font

Model	Year	ACE 2005 E+
Lin et al. [2] (GPT-6J)	2023	66.1
Liu et al. [8]	2020	46.1
Lyu et al. [9]	2021	47.8
Zhang et al. [10]	2021	53.6
Lin et al. [2] (BERT-Large)	2023	58.2
Ours	2024	**61.6**

Compared to the state-of-the-art (SOTA) results based on large models, there remains a significant performance gap for our model. Specifically, Lin et al. achieved a 4.5% higher score on the ACE dataset than our model. The advantage of models with ample resources over our zero-shot method is even more pronounced. This may be due to the fact that our model's parameter count (60M) is only 1% of the SOTA model's parameter count (6B). Based on the theoretical knowledge presented in Section III-A, there is still a partial performance gap between our EAC model and those utilizing Large Language Models (LLMs).

C. Comparison Between Different Prefix Prompts

In this section, we conduct experiments on the ACE (2005 E+) dataset to compare the effectiveness of using different prefix prompts within the model. We compare the following prefix prompts with those mentioned in Section III-B:

1) "[P1] most accurately represents the occurrence of this [P2]."
2) "The event type is [P1] and the trigger is [P2]."

TABLE II. Performance of Different Prefix Prompts

Prefix Prompt	F1 Score
Prefix(0)	61.6
Prefix(1)	61.3
Prefix(2)	60.8

From Table II, it can be observed that the prompts described in Section III-B are the most effective, which may be attributed to the fact that the prefix prompts are not only based on the definitions of events and triggers [17], but also possess a naturally fluent expression [2].

D. Computational Resource Analysis

The state-of-the-art (SOTA) model based on GPT-J, with its substantial parameter count of 6 billion, excels in resource-intensive tasks but also implies a significant demand for computational resources, making it generally unsuitable for low-resource scenarios. Therefore, in this section, we primarily compare the version implemented within the framework of Lin et al. [2] using BERT-Large with our model.

In contrast, the BERT-Large model used by Lin et al. has a parameter count of 334 million, whereas our model has a parameter count of only 60 million, significantly reducing the model's storage and computational requirements. Runtime and GPU memory usage are key indicators for gauging the feasibility of models in practical applications. We independently ran each model five times to calculate their average resource consumption.

TABLE III. THE RESOURCE CONSUMPTION OF EACH MODEL, WITH NUMERICAL VALUES REPRESENTING THE AVERAGE DURATION AND GPU MEMORY USAGE OVER FIVE INDEPENDENT RUNS

Model	Paramaters	Run Time	GPU Memory
Lin et al. (BERT-Large)	334M	1.2h	2151MiB
Ours. (w/o Flash-att)	60M	0.9h	563MiB
Ours. (Flash-att)	60M	0.8h	534MiB

As shown in Table III, the model of Lin et al. requires 1.2 hours to complete training or inference, while our model (without Flash-Attention) and (with Flash-Attention) only requires 0.9 hours and 0.8 hours, respectively. Compared to the BERT-Large-based model and the model without Flash-Attention, the time consumption is reduced by 33.3% and 11.1%, respectively, indicating that our model can provide faster processing speeds while maintaining a smaller parameter size.

The model of Lin et al. (BERT-Lager) [2] requires 2151 MiB of GPU memory, whereas our model significantly reduces this demand, with the version without Flash-Attention requiring 563 MiB and the Flash-Attention version further reducing to 534 MiB. This indicates that our model, while maintaining a smaller parameter size, has reduced GPU memory usage by 75.1% and 5.1%, respectively, making it more suitable for operation in resource-constrained environments.

The results indicate a substantial improvement in F1 score and a significant reduction in resource consumption. We attribute these improvements to the synergistic effect of Flash-Attention and global constraints within our model. However, we also acknowledge potential limitations, such as the model's generalizability to other domains and the need for further adaption to enhance its robustness.

E. Discussion

The ALBERT-F model, without the implementation of Flash-Attention, exhibits an average runtime of 0.9 hours, which is further reduced to 0.8 hours with the integration of Flash-Attention. This is a significant reduction compared to the 1.2 hours required by the BERT-Large model. Concurrently, the GPU memory consumption is markedly decreased from 2151 MiB for the BERT-Large to 563 MiB for the ALBERT-F model without Flash-Attention, and an additional reduction to 534 MiB is achieved with the utilization of Flash-Attention. These results indicate that the ALBERT-F model substantially diminishes resource consumption while maintaining performance, making it particularly suitable for scenarios with limited computational resources.

The fusion of global constraints and prompting strategies enhances the model's generalizability to unknown events, rendering it more competitive in zero-shot learning tasks. This characteristic implies that in practice, the model can make reasonable predictions for new event types even without specific training data, which is invaluable in situations where data is scarce or difficult to annotate. However, despite the ALBERT-F model demonstrating advantages in multiple aspects, its limitations in handling complex event structures and long-distance dependencies remain a subject worthy of investigation. Complex events often involve multi-layered nested semantic relationships, and long-distance dependencies require the model to capture associations between words that are distant in the text. The model's performance may be compromised in such cases, as traditional attention mechanisms may not effectively span long sequences to capture crucial information. Therefore, future research could focus on developing more advanced attention mechanisms or model architectures to strengthen the model's comprehension of complex events.

V. CONCLUSION

In conclusion, we propose a ALBERT-F model for zero-shot EAC that employs global constraints and prompting. Compared to previous works, our model has a significantly lower parameter count, which not only reduces storage requirements but also potentially mitigates the risk of model overfitting. Additionally, it offers advantages in terms of run time, implying faster iterations and adaptation to new data. In terms of GPU memory usage, our model is substantially suitable for operation on devices with limited memory. These advantages make our model particularly appealing in resource-constrained environments.

VI. LIMITATIONS

In this section, we summarize the limitations of our work as follows:

1) Expressiveness: Although the ALBERT-F model demonstrates exceptional resource efficiency, it may not match the robust expressiveness of large language models in certain complex natural language understanding tasks. Large language models typically excel in handling intricate linguistic structures and long-distance dependencies due to their substantial parameter count and deeper network architectures.

2) Domain-specific performance: In certain domains or tasks, large language models may exhibit superior performance due to exposure to a more diverse range of texts during their pre-training phase. While the ALBERT-F model possesses strong zero-shot learning capabilities, it may require additional domain adaptation to achieve optimal results with specialized terminology and concepts in specific fields.

3) Scalability: Although the ALBERT-F model shows significant optimization in resource consumption, whether it can maintain these advantages when dealing with larger datasets or more complex tasks, or if further adjustments to the model structure and parameters are needed, remains a subject that necessitates further research and validation.

Acknowledgment

We sincerely appreciate the collaboration and insightful discussions with all authors, which greatly contributed to this work. Our thanks also extend to the anonymous reviewers for their thorough and thoughtful evaluations, which have helped us refine our approach and presentation of the findings.

References

[1] Q. Li, J. Li, J. Sheng, S. Cui, J. Wu, Y. Hei, H. Peng, S. Guo, L. Wang, A. Beheshti *et al.*, "A survey on deep learning event extraction: Approaches and applications," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[2] Z. Lin, H. Zhang, and Y. Song, "Global constraints with prompting for zero-shot event argument classification," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 2527–2538.

[3] Z.-X. Ye and Z.-H. Ling, "Multi-level matching and aggregation network for few-shot relation classification," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2872–2881.

[4] V. D. Lai, F. Dernoncourt, and T. H. Nguyen, "Exploiting the matching information in the support set for few shot event classification," in *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24*. Springer, 2020, pp. 233–245.

[5] J. Sheng, S. Guo, Z. Chen, J. Yue, L. Wang, T. Liu, and H. Xu, "Adaptive attentional network for few-shot knowledge graph completion," *arXiv preprint arXiv:2010.09638*, 2020.

[6] I.-H. Hsu, K.-H. Huang, E. Boschee, S. Miller, P. Natarajan, K.-W. Chang, and N. Peng, "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*, 2022, pp. 1890–1908.

[7] L. Huang, H. Ji, K. Cho, I. Dagan, S. Riedel, and C. Voss, "Zero-shot transfer learning for event extraction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2160–2170.

[8] J. Liu, Y. Chen, K. Liu, W. Bi, and X. Liu, "Event extraction as machine reading comprehension," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 1641–1651.

[9] Q. Lyu, H. Zhang, E. Sulem, and D. Roth, "Zero-shot event extraction via transfer learning: Challenges and insights," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 322–332.

[10] H. Zhang, H. Wang, and D. Roth, "Zero-shot label-aware event trigger and argument classification," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1331–1340.

[11] R. Chen, C. Qin, W. Jiang, and D. Choi, "Is a large language model a good annotator for event extraction?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 772–17 780.

[12] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2019.

[13] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 344–16 359, 2022.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] B. M. Sundheim, "Overview of the fourth message understanding evaluation and conference," in *Proceedings of the 4th conference on Message understanding - MUC4 '92*, Jan 1992.

[16] R. Grishman and B. Sundheim, "Message understanding conference-6," in *Proceedings of the 16th conference on Computational linguistics*, Jan 1996.

[17] R. Grishman, D. Westbrook, and A. Meyers, "Nyu's english ace 2005 system description," *Ace*, vol. 5, no. 2, 2005.

[18] W. Xiang and B. Wang, "A survey of event extraction from text," *IEEE Access*, vol. 7, pp. 173 111–173 137, 2019.

[19] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 167–176.

[20] Y. Chen, S. Liu, S. He, K. Liu, and J. Zhao, "Event extraction via bidirectional long short-term memory tensor neural networks," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 15th China National Conference, CCL 2016, and 4th International Symposium, NLP-NABD 2016, Yantai, China, October 15-16, 2016, Proceedings 4*. Springer, 2016, pp. 190–203.

[21] L. Sha, F. Qian, B. Chang, and Z. Sui, "Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[22] D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi, "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)*, Jan 2019.

[23] Y. Lin, H. Ji, F. Huang, and L. Wu, "A joint neural model for information extraction with global features," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jan 2020.

[24] H. Wang, M. Chen, H. Zhang, and D. Roth, "Joint constrained learning for event-event relation extraction," *arXiv preprint arXiv:2010.06727*, 2020.

[25] P. Blache, "Constraints, linguistic theories, and natural language processing," in *International Conference on Natural Language Processing*. Springer, 2000, pp. 221–232.

[26] K. Ganchev, J. Graça, J. Gillenwater, and B. Taskar, "Posterior regularization for structured latent variable models," *The Journal of Machine Learning Research*, vol. 11, pp. 2001–2049, 2010.

[27] T. Hagendorff, "Mapping the ethics of generative ai: A comprehensive scoping review," *arXiv preprint arXiv:2402.08323*, 2024.

[28] X. Fang, S. Che, M. Mao, H. Zhang, M. Zhao, and X. Zhao, "Bias of ai-generated content: an examination of news produced by large language models," *Scientific Reports*, vol. 14, no. 1, p. 5224, 2024.

[29] L. Acion, M. Rajngewerc, G. Randall, and L. Etcheverry, "Generative ai poses ethical challenges for open science," *Nature Human Behaviour*, vol. 7, no. 11, pp. 1800–1801, 2023.

[30] B. Wang and A. Komatsuzaki, "GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model," https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

[31] G. Bai, Z. Chai, C. Ling, S. Wang, J. Lu, N. Zhang, T. Shi, Z. Yu, M. Zhu, Y. Zhang *et al.*, "Beyond efficiency: A systematic survey of resource-efficient large language models," *arXiv preprint arXiv:2401.00625*, 2024.

[32] S. Zeng, J. Liu, G. Dai, X. Yang, T. Fu, H. Wang, W. Ma, H. Sun, S. Li, Z. Huang *et al.*, "Flightllm: Efficient large language model inference with a complete mapping flow on fpgas," in *Proceedings of the 2024 ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, 2024, pp. 223–234.

[33] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," *arXiv preprint arXiv:2307.10169*, 2023.

[34] H. Dai, Y. Song, and H. Wang, "Ultra-fine entity typing with weak supervision from a masked language model," 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)*, Jan 2021. [Online]. Available: http://dx.doi.org/10.18653/v1/2021.acl-long.141

[35] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, "The automatic content extraction (ace) program tasks, data, and evaluation," *Language Resources and Evaluation,Language Resources and Evaluation*, May 2004.

[36] H. Ji and R. Grishman, "Refining event extraction through cross-document inference," *Meeting of the Association for Computational Linguistics,Meeting of the Association for Computational Linguistics*, Dec 2008.