Demonstration Retrieval-Augmented Generative Event Argument Extraction

Shiming He, Yu Hong*, Shuai Yang, Jianmin Yao, Guodong Zhou
School of Computer Science and Technology, Soochow University, Suzhou, China
{smhelpai, tianxianer, shytostu}@gmail.com, {jyao, gdzhou}@suda.edu.cn

Abstract

We tackle Event Argument Extraction (EAE) in the manner of template-based generation. Based on our exploration of generative EAE, it suffers from several issues, such as multiple arguments of one role, generating words out of context and inconsistency with prescribed format. We attribute it to the weakness of following complex input prompts. To address these problems, we propose the demonstration retrieval-augmented generative EAE (DRAGEAE), containing two components: event knowledge-injected generator (EKG) and demonstration retriever (DR). EKG employs event knowledge prompts to capture role dependencies and semantics. DR aims to search informative demonstrations from training data, facilitating the conditional generation of EKG. To train DR, we use the probability-based rankings from large language models (LLMs) as supervised signals. Experimental results on ACE-2005, RAMS and WIKIEVENTS demonstrate that our method outperforms all strong baselines and it can be generalized to various datasets. Further analysis is conducted to discuss the impact of diverse LLMs and prove that our model alleviates the above issues.

Keywords: Event Argument Extraction, Demonstration Retrieval, Template-based Generation

1. Introduction

Event argument extraction (EAE) is a fundamental and challenging part of event extraction (EE). It aims to discover arguments for each predefined role (Doddington et al., 2004; Ahn, 2006). For example, in Figure 1, given that the word "*hired*" in context triggers a `Start-Position` event, EAE model is required to extract arguments (e.g., *London*) for each specific role (e.g., **Place**).

Event type: `Start-Position`			
Context: Word from London that the tabloid the daily mirror just hired Peter Arnett to be its correspondent in Baghdad.			
Person	Peter Arnett	**Place**	London
Entity	tabloid	-	-

Figure 1: Example of event argument extraction.

Significant efforts are devoted by researchers to advance EAE from feature-based models (Ahn, 2006; Grishman, 2010; Hong et al., 2011) to recent deep learning-driven methods (Chen et al., 2015; Nguyen et al., 2016; Liu et al., 2018; Wang et al., 2019; Du and Cardie, 2020; Wang et al., 2021; Ma et al., 2022; Hsu et al., 2023). Most of EAE methods formulate the task as the paradigm of classification. They aim to map the argument candidate to role space (Chen et al., 2015; Liu et al., 2018; Wang et al., 2019, 2021; Ma et al., 2022). Besides, recent generation-based EAE methods reformulate extraction as structural generation and achieve substantial progress (Li et al., 2021; Huang et al., 2022; Liu et al., 2022b; Hsu et al., 2022, 2023). They both rely on a prompt which explicitly introduces event knowledge and defines the output format.

Based on our observations, models of this paradigm suffer from several issues, including multiple arguments of the same role, generating words out of context and outputs inconsistent with required format. We attribute this to the insufficient elicitation of the ability in generative models to understand input-output mapping. Motivated by demonstration selection of in-context learning (ICL) (Brown et al., 2020; Chung et al., 2022; Chen et al., 2022; Rubin et al., 2022; Li et al., 2023), we expect to retrieve good examples (input-output pairs) to elicit the analogical capability of generative models.

In this paper, we explore to retrieve informative demonstrations for generative EAE model to better make predictions. We propose a framework DRAGEAE (Demonstration Retrieval-augmented Generative Event Argument Extraction). It contains two essential components: demonstration retriever (DR) and event knowledge-injected generator (EKG). Following previous works (Li et al., 2021; Hsu et al., 2022; Liu et al., 2022b; Hsu et al., 2023), the basic input x of EKG is a context and a human-written prompt containing role interaction-based event declaration and formatted template. The target output y comes from the template filled with corresponding arguments. Therefore, the generator better captures both prior event knowledge and cross-role dependencies. Combined with the generator, the retriever is our main contribution. Given a training example (x, y) and a set of candidate demonstrations (examples) from training data, we use large language models (LLMs) to rank them according to the probability of generating ground-truth y conditioned on x and each candidate. The candi-

* Corresponding Author.

LREC-COLING 2024, pages 4617–4625
20-25 May, 2024.

date resulting in a higher probability is supposed to be more informative (Rubin et al., 2022; Fu et al., 2023; Li et al., 2023). Subsequently, the retriever is trained by LLM ranking feedback, thus inheriting the ability to determine high-quality demonstration. Unlike previous studies that use off-the-shelf LLMs to infer results (Rubin et al., 2022; Li et al., 2023; Wang et al., 2023), we finetune the generator with retrieved examples. And this is expected to enhance the analogical capability.

Experimental results show that DRAGEAE outperforms all strong baselines on ACE-2005 and achieves new State-of-The-Art (SoTA) with an F1-score of 74.8%. Additional experiments on RAMS and WIKIEVENTS verify the generalization and compatibility of DRAGEAE. We also conduct ablation studies to evaluate each module. Further analysis reveals the impact of diverse LLMs and the distribution changes of errors.

2. Related Work

Event argument extraction witnesses the thriving iterations of NLP technology. Traditional EAE methods (Ji and Grishman, 2008; Hong et al., 2011; Li et al., 2013) heavily rely on manual rules and feature engineering. Benefiting from deep learning, modern EAE methods shift towards the refinement of both neural network architectures and optimization objectives (Chen et al., 2015; Nguyen et al., 2016; Liu et al., 2018; Wang et al., 2019; Du and Cardie, 2020; Li et al., 2021; Ma et al., 2022; Hsu et al., 2023). This enables the automatic recognition of event-related features and their dependencies, boosting the performance to a higher level. We categorize these works to two different paradigms: classification and generation.

Classification-based methods involve locating the argument span and mapping it into role space. Commonly, they leverage auxiliary syntactic structures (Liu et al., 2018; Pouran Ben Veyseh et al., 2020) and semantic associations (e.g., entity, trigger, relation) (Chen et al., 2015; Sha et al., 2018; Ding et al., 2022; Lin et al., 2020) through sophisticated networks. Based on powerful pretrained language models (PLMs), Wang et al. (2021) designs a contrastive pretraining objectives to learn event knowledge and their semantic structures from large-scale unsupervised data.

By contrast, **generation-based** models are more end-to-end and flexible. They translate extraction to structured generation dependent on constrained decoding (Lu et al., 2021; Paolini et al., 2021) and template-based conditions (discrete or continuous prompts) (Li et al., 2021; Liu et al., 2022b; Hsu et al., 2022, 2023). Li et al. (2021) focuses on document-level event argument extraction and proposes a set of informative templates to capture long-range dependencies. Huang et al. (2022) explore language-agnostic templates to transfer event knowledge in the zero-shot cross-lingual scenario. Benefiting from templates, generative EAE models better exploit event knowledge (e.g., cross-role dependencies or event descriptions) to unleash intrinsic capability – generate anything.

3. Approach

The overall architecture of DRAGEAE (see this link) contains an event knowledge-injected generator (EKG) and a demonstration retriever (DR). We will introduce the details of model designs and training.

3.1. Event Knowledge-Enhanced Generator

We reformulate EAE as template-based conditional generation, following (Li et al., 2021; Hsu et al., 2022). The event extraction dataset defines a set of event types $\Omega = \{\tau_i\}_{i=1}^{|\Omega|}$ and each $\tau \in \Omega$ corresponds to a set of argument roles Φ_τ.

Basically, each input x of an example (x, y) contains the context sentence s and the event knowledge prompt $\mathcal{P}$, denoted as $x = s \oplus \mathcal{P}$ where $\oplus$ refers to text concatenation. Given an event type τ, the trigger w_t and the argument role set Φ_τ, the prompt $\mathcal{P}_{\tau,w_t}$ includes the following two components. (1) **Event type constraint** follows the pattern of "*In the τ event triggered by w_t*". For EAE task, we use golden event type and trigger. (2) **Descriptive template** summaries the τ event with argument roles and indicates the output format.

For example, in Figure 1, the *Start Position* event contains the role set $\Phi_\tau = \{\textbf{Person}, \textbf{Entity}, \textbf{Place}\}$. The prompt $\mathcal{P}$ refers to "*In the Start-Position event triggered by hired, **Person** stared working at **Entity** organization in **Place**.*" [1]. The ground-truth sequence y is processed to follow the template format by replacing role labels with correct arguments[2]. Therefore, the output should be "***Peter Arnett** started working at **tabloid** organization in **London***". Mathematically, the event knowledge-injected generator g is formulated as $p_g(y|x) = p_g(y|s, \mathcal{P}_{\tau,w_t})$.

3.2. Demonstation Retriever

Provided the training set $\mathcal{D}_{train} = \{(x^i, y^i)\}_{i=1}^n$, the goal of DR is to search high-quality demonstration $d' = (x', y') \in \mathcal{D}_{train}$, with input x as the query.

[1] To save time and effort in template design and ensure a fair comparison, we reuse templates from previous works (Li et al., 2021; Liu et al., 2022b).

[2] If there exists no argument for a role, replace the role label with "`None`". If a role is related to multiple arguments, concatenate them with "`and`".

3.2.1. Retriever Architecture

Following Rubin et al. (2022), DR is based on the bi-encoder architecture which can be initialized from any BERT-like models (Devlin et al., 2019). The input x (query) of a training example and the demonstration d' are encoded separately by the siamese encoder E_{DR}. Then, their relevance is calculated by cosine similarity: $rel(x, d') = \frac{E_{DR}(x)^\top E_{DR}(d')}{\|E_{DR}(x)\|\|E_{DR}(d')\|}$, where $E_{DR}(\cdot)$ represents the output vector of built-in special token `[CLS]` encoded by E_{DR} and $\|\cdot\|$ denotes Euclidean norm. We adopt the bi-encoder due to efficiency and effectiveness.

3.2.2. Learning from LLM

Given the LLM $\hat{g}$, input x of a training example (x, y) and a set of demonstration candidates $\Upsilon \subset \mathcal{D}_{train}$, we rank all candidates based on the conditional probability $p_{\hat{g}}(y|x, d_i)$ generated by $\hat{g}$. For d_i, the input of $\hat{g}$ is "d_i `[SEP]` x". Considering the inference cost of LLMs, we only construct a candidate set $\Upsilon = \{d_i\}_{i=1}^{m}$ [3] with top-m ranked demonstrations recalled by the retriever before finetuning. If $p_{\hat{g}}(y|x, d_i)$ is the k-th largest among Υ, we define $r(d_i) = k$. A higher rank of d_i implies that it is more informative for target reasoning. Consequently, we obtain the train set for DR, $\mathcal{D}_{DR} = \{(x^i, y^i, \Upsilon_i, \{r(d_j)|d_j \in \Upsilon_i\}^m)\,|(x^i, y^i) \in \mathcal{D}_{train}\}_{i=1}^{n}$. To enhance DR with the scoring capability of LLM, we regard the ranking of all candidates as supervised signals. Specifically, we propose to leverage the partial orders of ranking and minimize the loss function:

$$\mathcal{L}_r = \sum_{\substack{r(d_i)<r(d_j) \\ \wedge\ d_i,d_j\in\Upsilon\ \wedge\ i\neq j}} \log\left(1 + e^{rel(x,d_j)-rel(x,d_i)}\right). \tag{1}$$

Since learning to rank candidates is also a metric learning problem, following Karpukhin et al. (2020); Rubin et al. (2022), we additionally use the contrastive loss with in-batch negative samples:

$$\mathcal{L}_c = -\log \frac{e^{rel(x,d^+)}}{e^{rel(x,d^+)} + \sum_{d\in\mathcal{N}} e^{rel(x,d)}}, \tag{2}$$

where d^+ denotes the rank-1 candidate of the input x of current training example. $\mathcal{N}$ is the negative set of all rank-1 candidates of other training examples in the same mini-batch. This method aims to optimize the representation space to achieve improved alignment and uniformity (Wang and Isola, 2020), and meanwhile benefits the learning of ranking function. At last, the total loss function of DR is a combination of both the two objectives:

$$\mathcal{L}_{DR} = \mathcal{L}_r + \mathcal{L}_c \tag{3}$$

[3] We set m as 20 to maximize the utilization of GPU memory.

3.3. Demonstration Retrieval-Augmented Generation

The two essential components of DRAGEAE is trained separately. The finetuned DR is directly applied to demonstration selection. During training of EKG, given the input x of an example (x, y), we use DR to calculate the relevance over all training examples and select top-1[4] ranked demonstrations d^*. Let the final input "`[CLS]` d^* `[SEP]` x `[SEP]`" and the grounded output sequence y, the model is optimized by minimizing negative log-likelihood loss:

$$\mathcal{L}_g = -\sum_{i=1}^{|y|} \log p_g\left(y_i|d^*, x, y_{<i}\right) \tag{4}$$

where y_i is the i-th token of sequence y and $y_{<i}$ denotes the subsequence before the i-th position.

During inference, DRAGEAE generates sequences in an autoregressive manner with beam search. To parse predicted arguments, we employ regular expression and string matching.

4. Experiments

4.1. Experimental Settings

Datasets. We apply our model on three representative datasets of Event Argument Extraction, including the most popular ACE-2005 (Doddington et al., 2004),the recent RAMS (Ebner et al., 2020) and WIKIEVENTS (Li et al., 2021). All experiments are conducted on ACE-2005, while RAMS and WIKIEVENTS are left to the study on generalization of the demonstration retriever.

The English part of ACE-2005 we use is a sentence-level datasets. It contains 599 documents with human annotations of events. We follow the preprocessing from Wadden et al. (2019), keeping 33 event types and 22 argument roles.

RAMS is a document-level dataset for EAE task, annotated with 139 event types and 65 roles. Specifically, each document consists of 5 sentences, while arguments of one event are scattered throughout the document. We follow the official data splits from Ebner et al. (2020).

WIKIEVENTS is a more realistic document-level dataset for EAE task, collected from English Wikipedia articles. It provides 246 documents with 50 event types and 59 argument roles. The core idea of this dataset is to extract more informative arguments. The official data splits are provided by Li et al. (2021).

Evaluation. For evaluation, we report metrics of precision (**P**), recall (**R**) and F1-score (**F1**) of argument classification, following previous works (Wad-

[4] The number is tuned on development set.

Hyperparameters	DR	DRAGEAE
learning rate	5×10^{-5}	10^{-4}
weight decay	0.01	0.01
warm-up ratio	0.1	0.1
batch size	64	40
epoch	60	20
max input/output length	128/-	512/80
beam size	-	4

Table 1: Hyperparameters

den et al., 2019; Lin et al., 2020). An argument is correctly extracted only when both its offset and role label match the ground truth exactly.

Implementation Details. The LLM we use to rank candidates is text-davinci-003[5], whose temperature is set to 0. Since we train two components with AdamW (Loshchilov and Hutter, 2019), all common hyperparameters are shown in Table 1. The retriever DR is initialized with SBERT (Reimers and Gurevych, 2019). We finetune DR with early stopping based on training loss. For DRAGEAE, we use T5-large of huggingface implementation[6] as the backbone. All hyperparameters are tuned on development set. Each experiment is conducted on 2 NVIDIA A100 40GB GPUs. The experimental results we report are the average of 5 random seeds from $\{1, 10, 42, 100, 1000\}$.

Baselines. We compare DRAGEAE with other state-of-the-art models of two paradigms: classification-based and generation-based methods. **DyGIE++** (Wadden et al., 2019) integrates entity and relation features to recognize arguments. **EEQA** (Du and Cardie, 2020) transforms the extraction task to an extractive question answering task. It locates the predicted argument for each role-specific question via pointer networks. **BART-Gen** (Li et al., 2021) designs descriptive templates for each event type to capture role interactions and long-range dependencies. It aims to generate the filled template and parse results in role slots. **X-GEAR** (Huang et al., 2022) performs cross-lingual generative extraction with language-agnostic templates. It demonstrates the potential of large generative models to deal with extraction tasks. **AMPERE** (Hsu et al., 2023) encodes AMR auxiliary signals into template-based generative models to improve semantic learning.

4.2. Main Results

The overall performance of DRAGEAE is presented in Table 2. Our model outperforms all previous methods and achieves the state-of-the-art. We conduct t-test to verify the statistical significance and the p-value is about 0.000617 (< 0.05) when compared to AMPERE. It indicates that we obtain a significant improvement.

Noticeably, when compared with classification-based models (e.g., DyGIE++ and EEQA), the performance of generation-based methods is obviously superior. This is attributed to the knowledge-intensive human-written template. We assume that the template effectively exploits the large-scale prior knowledge derived from pretraining. Besides, our model still beats generation-based competitors and exceeds 1.4%, 0.2% and 0.8% in precision, recall and F1-score, respectively. On precision, DRAGEAE demonstrates more improvement than other template-based generation baselines. It indicates our retriever provides valuable and informative demonstrations, as well as assists in learning prompt knowledge and complex input structures.

Model	P	R	F1
DyGIE++ (Wadden et al., 2019)	61.4	55.9	58.5
EEQA* (Du and Cardie, 2020)	67.9	63.0	65.4
BART-Gen* (Li et al., 2021)	67.8	65.6	66.7
X-GEAR (Huang et al., 2022)	69.2	73.6	71.3
AMPERE (Hsu et al., 2023)	72.3	75.8	74.0
DRAGEAE (ours)	**73.7**	**76.0**	**74.8**

Table 2: Main results (%) on ACE-2005 test set. Value in **Bold** represents the best performance. $*$ denotes the result reported in original paper.

Model	RAMS	WIKIEVENTS
BART-Gen (Li et al., 2021)	48.2	64.8
BART-Gen w/ DR	50.3	65.5
DRAGEAE	**54.9**	**69.2**

Table 3: Results (F1) of generalization study on different datasets. DR is only trained on ACE-2005.

4.3. Generalization on Diverse Datasets

We study the generalization of our method on diverse datasets and take into consideration BART-Gen for comparison. The results shown in Table 3 demonstrate that DRAGEAE surpasses BART-Gen on both RAMS and WIKIEVENTS by 6.7% and 4.4%, respectively. We additionally apply DR to BART-Gen (i.e., BART-Gen w/ DR) and this also produces the apparent improvement.

It is noteworthy that our experiments here are grounded on the DR trained only on ACE-2005. In other words, we only finetune the generator on RAMS and WIKIEVENTS. Accordingly, the results prove that our proposed method is highly adaptive and generalized. Therefore, we assume that our proposed DR is compatible with other generative extractors, especially the ones which rely on the

[5]https://platform.openai.com/docs/api-reference/completions

[6]https://huggingface.co/t5-large

Model	P	R	F1
DRAGEAE (full)	**73.7**	**76.0**	**74.8**
-w/o Event type constraint	71.1	75.2	73.1
-w/o Descriptive template	73.3	71.7	72.5
GEAE (-w/o DR)	**70.6**	**73.9**	**72.2**
-w/o Event type constraint	67.7	71.8	69.7
-w/o Descriptive template	69.9	67.0	68.4

Table 4: Results of ablation studies for each component on ACE-2005 test set.

LLM variants	P	R	F1
text-davinci-003 (175B)	73.7	**76.0**	**74.8**
Flan-T5-XXL (11B)	**76.8**	72.5	74.6
LLaMA (7B)	72.3	75.1	73.7

Table 5: Performance of DRAGEAE enhanced by different retrievers trained by diverse LLM rankings.

stability of output patterns and the correctness of constituents within the event-specific templates.

5. Analysis

5.1. Ablation Studies

To investigate the effectiveness of our designs, we conduct ablation studies by removing each design and present the results in Table 4. For fine-grained control of variables, we firstly modify the full **DRAGEAE** to the variant **GEAE** by removing the demonstration retriever.

From the results in Table 4, for both DRAGEAE and GEAE, it is evident that removing descriptive template consistently leads to a larger performance (both recall and F1 score) drop than removing event type constraint. Because the descriptive template encompasses semantics of role labels and inter-role dependencies, facilitating a better comprehension of the event. Meanwhile, it establishes the prescribed output format and asks the model to substitute role labels with corresponding arguments. Moreover, for performance fluctuations in response to prompt changes, DRAGEAE demonstrates the more enhanced stability. We suppose that DR provides informative examples which steer the generator in the right direction.

5.2. LLM Variants

We evaluate the impact of ranking signals derived from LLMs of varying scales. Specifically, at the stage of training DR, we use a range of LLMs, like LLaMA (Touvron et al., 2023) and Flan-T5-XXL (Chung et al., 2022), to rank the candidates and employ the rankings as supervised signals. The EAE results of DRAGEAE combined with different demonstration retrievers supervised by various LLMs feedback are shown in Table 5. There is a trend that as the size of LLM for ranking candidates is scaled up, a corresponding improvement in the performance of DRAGEAE is observed. Despite the significant gap in parameters between Flan-T5-XXL and text-davinci-003, the performance gains obtained by the larger model are modest.

Error Type	DRAGEAE	GEAE
Incompletion	8%	12%
Out of Context	2%	4%
Wrong Format	2%	14%

Table 6: Distribution of three main error types in DRAGEAE and GEAE.

5.3. Error Analysis

To perform an error analysis, we randomly sample 50 examples and examine the predictions of of DRAGEAE and GEAE. The statistical distribution of errors is presented in Table 6.

Typically, we identify three predominant error types. (1) **Incompletion** refers to the situation where not all gold arguments are predicted when multiple arguments is related to the same role. (2) **Out of Context** refers to the generation of arguments not found within the context. (3) **Wrong format** refers to generating sentences whose patterns are inconsistent with prescribed formats, rendering the post-processing unfeasible. Compared with GEAE, our full model DRAGEAE effectively alleviates the three main errors, especially the wrong-format error. It demonstrates the enhanced analogical ability of DRAGEAE to capture template patterns, which benefits from the input-output mappings of retrieved demonstrations.

6. Conclusion

In this paper, we explore the demonstration retrieval augmentation for generation-based event argument extraction and propose the model called DRAGEAE. To provide informative and high-quality demonstrations, we finetune a retriever with a bi-encoder architecture, which is supervised by LLM feedback. Empirically, our model outperforms all baselines on three datasets, demonstrating both generalization capability and compatibility. We conduct thorough ablation studies to explore the effectiveness of each component. Besides, we analyze the influence of different ranking LLMs and compare the distribution changes of errors. In the future, we plan to further apply our work in more challenging document-level event extraction.

7. Limitations

Although our proposed model DRAGEAE achieves the superior performance compared to previous studies, some limitations of our work do exist. First, our model is based on descriptive event templates which is a heavy cost of time and human-labor. Without the templates, it is hard to apply our method into diverse datasets. Second, in this paper, we only experiment on the setting of T5-large as the generator. We leave the investigation on the generalization of various models in future. Moreover, the core idea of our method is to retrieve informative demonstration to facilitate in-context learning. We believe there is still a huge room for improvement of our proposed retriever. And this will contribute to the few-shot capability of current LLMs.

8. Ethical Considerations

Our work relies on the two open-source retriever and generator which are pretrained on a large-scale corpus with potential bias. Therefore, the outputs of our model may have the risk of such prior and unexpected bias. We suggest users should carefully check the generated contents before application.

9. Acknowledgement

The research is supported by National Science Foundation of China (62376182, 62076174). We also sincerely thank the anonymous reviewers for their insightful and helpful suggestions.

10. Reference

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, Seattle, United States. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 167–176. The Association for Computer Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Ding, Chunming Hu, Kai Sun, Samuel Mensah, and Richong Zhang. 2022. Explicit role interaction network for event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3475–3485, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content

extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

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.

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.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ralph Grishman. 2010. The impact of task and corpus on event extraction systems. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1127–1136, Portland, Oregon, USA. Association for Computational Linguistics.

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.

I-Hung Hsu, Zhiyu Xie, Kuan-Hao Huang, Prem Natarajan, and Nanyun Peng. 2023. AMPERE: AMR-aware prefix for generation-based event argument extraction model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10976–10993, Toronto, Canada. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4633–4646, Dublin, Ireland. Association for Computational Linguistics.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, Ohio. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. 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.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and 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, Online. Association for Computational Linguistics.

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.

Xiao Liu, Heyan Huang, Ge Shi, and Bo Wang. 2022b. Dynamic prefix-tuning for generative template-based event extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5216–5228, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Brussels, Belgium. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. 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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774, Dublin, Ireland. Association for Computational Linguistics.

Thien Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

Amir Pouran Ben Veyseh, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Graph transformer networks with syntactic and semantic structures for event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3651–3661, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena,

Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5916–5923. AAAI Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh 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.

Liang Wang, Nan Yang, and Furu Wei. 2023. Learning to retrieve in-context examples for large language models.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019. HMEAE: Hierarchical modular event argument extraction. 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 5777–5783, Hong Kong, China. Association for Computational Linguistics.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. CLEVE: Contrastive Pre-training for Event Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6283–6297, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.