MAVEN-ARG: Completing the Puzzle of All-in-One Event Understanding Dataset with Event Argument Annotation

Anonymous ACL submission

Abstract

Understanding events in texts is a core objective of natural language understanding, which requires detecting event occurrences, extracting event arguments, and analyzing inter-event relationships. However, due to the annotation challenges brought by task complexity, a large-scale dataset covering the full process of event understanding has long been absent. In this paper, we introduce MAVEN-ARG, which augments MAVEN datasets with event argument annotations, making the first all-in-one dataset supporting event detection, event argument extraction (EAE), and event relation extraction. As an EAE benchmark, MAVEN-ARG offers three main advantages: (1) a **comprehensive schema** covering 162 event types and 612 argument roles, all with expert-written definitions and examples; (2) a **large data scale**, containing $98,591$ events and $290,613$ arguments obtained with laborious human annotation; (3) the **exhaustive annotation** supporting all task variants of EAE, which annotates both entity and non-entity event arguments in document level. Experiments indicate that MAVEN-ARG is quite challenging for both fine-tuned EAE models and proprietary large language models (LLMs). Furthermore, to demonstrate the benefits of an all-in-one dataset, we preliminarily explore a potential application, future event prediction, with LLMs. MAVEN-ARG and our baseline codes will be publicly released.

1 Introduction

Conveying information about events is a core function of human languages (Levelt, 1993; Pinker, 2013; Miller and Johnson-Laird, 2013), which highlights *event understanding* as a major objective for natural language understanding and a foundation for various downstream applications (Ding et al., 2015; Li et al., 2018a; Goldfarb-Tarrant et al., 2019; Huang et al., 2019; Wang et al., 2021a). As illustrated in Figure 1, event understanding is typically organized as three information extraction tasks (Ma

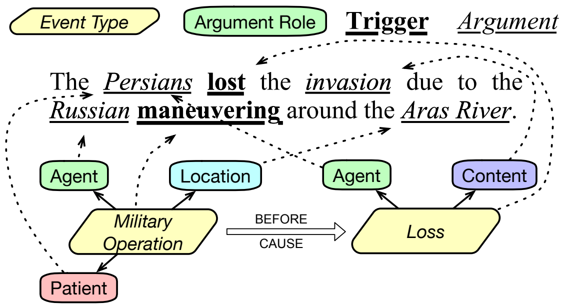

Figure 1: Illustration for the overall event understanding, consisting of event detection, event argument extraction, and event relation extraction tasks.

et al., 2021; Peng et al., 2023b): event detection (ED), which detects event occurrences by identifying event triggers and classifying event types; event argument extraction (EAE), which extracts event arguments and classifies their argument roles; event relation extraction (ERE), which analyzes the coreference, temporal, causal, and hierarchical relationships among events.

Despite the importance of event understanding, a large-scale dataset covering all the event understanding tasks has long been absent. Established sentence-level event extraction (ED and EAE) datasets like ACE 2005 (Walker et al., 2006) and TAC KBP (Ellis et al., 2015, 2016; Getman et al., 2017) do not involve event relation types besides the basic coreferences. RAMS (Ebner et al., 2020) and WikiEvents (Li et al., 2021) extend EAE to the document level but do not involve event relations. ERE datasets are mostly developed independently for coreference (Cybulska and Vossen, 2014), temporal (Chambers et al., 2014; Ning et al., 2018), causal (Mirza et al., 2014; Mostafazadeh et al., 2016b; Caselli and Vossen, 2017), and subevent (Hovy et al., 2013; Glavaš and Šnajder, 2014) relationships and do not cover event arguments. Given annotation challenges from task complexity, these datasets often cover only thousands of events. Due to the inconsistent event schemata and data, these datasets cannot be unified. This status quo hinders the development of end-to-end event understanding methods and limits the potential for event-based downstream applications.

MAVEN (Wang et al., 2020) is the largest human-annotated ED dataset, with a high-coverage event schema for general-domain events. Based on it, Wang et al. (2022) further annotates the first unified ERE dataset MAVEN-ERE, which covers all four types of event relationships and has a massive scale with more than one million event relations. Building on the sustained efforts of these works over years, we complete the puzzle of an all-in-one event understanding dataset in this work. We construct MAVEN-ARG, which provides exhaustive event argument annotations based on MAVEN.

Beyond finishing an all-in-one event understanding dataset, three main advantages of MAVEN-ARG make it a valuable EAE benchmark. (1) **Comprehensive Event Schema.** The original MAVEN schema only defines event types but without argument roles. We engage experts to enhance MAVEN schema with argument roles and to write detailed definitions for them, which help annotators and can also serve as task instructions for prompting large language models. The resulting event schema contains 162 event types, 612 argument roles, and 14,655 words of definitions, which well cover general-domain events. (2) **Large Data Scale.** MAVEN-ARG comprises 107,507 event mentions, 290,613 event arguments, and 129,126 entity mentions, all of which are human annotated. To our knowledge, this makes it the largest EAE dataset currently available. (3) **Exhaustive Annotation.** The development of EAE has seen many variations in task settings, including annotating only the topic event (Ebner et al., 2020; Tong et al., 2022) of a document or all fine-grained events (Walker et al., 2006), annotating event arguments at the sentence level (Walker et al., 2006) or document level (Ebner et al., 2020; Li et al., 2021), and limiting event arguments to entities (Walker et al., 2006; Li et al., 2021) or including non-entity arguments (Grishman and Sundheim, 1996; Parekh et al., 2023). MAVEN-ARG adopts the most exhaustive annotation. We annotate event arguments for all fine-grained events at the document level, covering both entity and non-entity arguments. This enhances the dataset's utility for benchmarking and developing a wide range of EAE methods.

In the experiments, we reproduce several recent state-of-the-art EAE models as baselines and also evaluate large language models with in-context learning. Experimental results show that they can only achieve at most 40% F1 scores, which is far from promising. It indicates that MAVEN-ARG is quite challenging and more research efforts are needed to develop practical EAE methods. Furthermore, to demonstrate the advantage of an all-in-one event understanding dataset for enabling sophisticated event-based applications, we conduct a preliminary exploration of *future event prediction*. We sample causally related event chains from MAVEN-ARG and prompt LLMs to predict future events, including their types and arguments. Experiments show that while most of the predictions are reasonable, they seldom align with the actual future. We encourage future work to further explore this application and hope MAVEN-ARG can help improve EAE and develop diverse event-based applications.

2 Dataset Construction

2.1 Event Schema Creation

The event schema of MAVEN (Wang et al., 2020) covers a broad range of general-domain events and has a well-defined hierarchical structure. To enable event argument annotation based on MAVEN, one author and two engaged linguistic experts devoted three years to manually designing argument roles for MAVEN schema. Each argument role is accompanied by informative text definitions that are easy to understand, and each event type is provided with detailed annotation examples. An example in shown in appendix A.1. This not only helps annotators understand their tasks but also can prompt LLMs to perform EAE via in-context learning. To ensure quality, the argument role design for each event type is reviewed by at least one expert.

Our event schema creation involves the following steps: (1) Initially, to reduce annotation difficulty, we invite ten ordinary annotators, who are without dedicated study on event semantics, to review the event type schema and a portion of the data. Based on their feedback, we deleted 6 event types that are similar to others and renamed 4 event types for clarity. (2) The basic schema is constructed from a simplification and modification of FrameNet (Baker et al., 1998). The *frame elements* in FrameNet are widely considered akin to argument roles (Aguilar et al., 2014; Parekh et al., 2023), but they are often too complex for ordinary annotators to comprehend since FrameNet is pri-

Dataset	#Event Type	#Argument Role
ACE 2005	33	36
DocEE	59	356
WikiEvents	50	59
RAMS	139	65
MEE	16	23
GENEVA	115	220
MAVEN-ARG	162	612

Table 1: Event schema statistics of MAVEN-ARG compared with other datasets.

marily constructed for linguistic experts (Aguilar et al., 2014). Therefore, for each event type, we manually select frame elements related to describing events and suitable for annotation as MAVEN-ARG argument roles from their FrameNet equivalents, and we rewrite the definitions and examples. (3) Extending argument roles based on the 5W1H analysis (What, Where, When, Why, Who, How) for describing events (Karaman et al., 2017; Hamborg et al., 2019). Temporal and causal relations from event relation extraction describe When and Why, while the event type describes What. We primarily refer to Who (participants), Where (locations), and How (manners, instruments, etc.) to design argument roles. (4) Considering the hierarchical structure. When designing subordinate types, we inherit and refine the argument roles of their superordinate types. (5) Sampling data to check if any event argument is missing.

Schema Statistics After the schema design, the final MAVEN-ARG schema contains 162 event types, 612 unique argument roles, and $14,655$ words of definitions. Taking inspiration from semantic role labeling (Fillmore, 1976; Banarescu et al., 2013), we tend to let the argument roles sharing the same semantic role use the same name but distinguish them with different textual definitions. For instance, we do not use `Killer` for the `Killing` event type and use `Attacker` for the `Attack` event type. Instead, we use `Agent` to denote them both but write different definitions for them. This is to encourage the knowledge transfer between EAE for different event types. Therefore, 612 is the number of argument roles with unique definitions, and there are 143 unique names for all the argument roles. Table 1 compares the event schema size of MAVEN-ARG with existing EAE datasets, including ACE 2005 (Walker et al., 2006), DocEE (Tong et al., 2022), WikiEvents (Li et al., 2021), RAMS (Ebner et al., 2020), MEE (Pouran Ben Veyseh et al., 2022), and GENEVA[1] (Parekh et al., 2023). We can observe that MAVEN-ARG has the largest event schema, which more comprehensively covers the broad range of diverse events and will help develop more generalizable methods.

2.2 Entity Annotation

The mainstream task setting for EAE (Walker et al., 2006; Li et al., 2021) confines event arguments to entities, which reduces the task's complexity to some extent and provides more definite and standardized extraction results. Hence, before annotating event arguments, we annotate entities for the $4,480$ MAVEN documents. We follow the task definition and guidelines of a recent named entity recognition benchmark Few-NERD (Ding et al., 2021), but we only annotate coarse-grained entity types, including `Person`, `Organization`, `Location`, `Building`, `Product`, `Art`, and `MISC`. To deliver more unambiguous EAE results and reduce the argument annotation difficulty, we also annotate entity coreference, which means judging whether multiple entity mentions refer to the same entity. During entity annotation, we engage 47 annotators, including 8 senior annotators selected during the annotation training. Each document is annotated by three independent annotators and further checked by one senior annotator. The final annotation results are aggregated via majority voting. If the senior annotator judged the accuracy of a document's annotation to be below 90%, the document will be returned to the three first-stage annotators for re-annotation. To check data quality, we calculate Fleiss' kappa (Fleiss, 1971) to measure the inter-annotator agreements. The result for entity recognition is 73.2%, and for entity coreference is 78.4%, both indicating high consistency.

2.3 Event Argument Annotation

Based on the event detection annotations of MAVEN and event coreferences of MAVEN-ERE, we conduct event argument annotations. For multiple coreferent event mentions (triggers), only one of them is displayed during annotation to reduce annotation overhead. Once the annotator selects an event trigger, the corresponding argument roles for its event type are displayed on the annotation interface, along with definitions and examples. This ensures that annotators do not have to memorize the lengthy event schema. To annotate an event

[1] GENEVA has a larger "full ontology" but is without data. Here we compare with its schema actually used in dataset.

Dataset	#Doc.	#Event	#Trigger	#Arg.	#Entity	#Entity Mention	Fine-grained Event	Doc. Level	Entity Arg.	Non-Entity Arg.
ACE 2005	599	4,090	5,349	9,683	45,486	59,430	✓	✗	9,683	✗
DocEE	27,485	27,485	-	180,528	-	-	✗	✓	✗	180,528
WikiEvents	246	3,951	-	5,536	13,937	33,225	✓	✓	5,536	✗
RAMS	3,993	9,124	-	21,237	-	-	✗	✓	✗	21,237
MEE	13,000	17,642	-	13,548	-	190,592	✓	✓	13,548	✗
GENEVA	-	7,505	-	12,269	-	36,390	✓	✗	8,544	3,725
MAVEN-ARG	4,480	98,591	107,507	290,613	83,645	129,126	✓	✓	116,024	174,589

Table 2: Statistics of MAVEN-ARG compared to existing widely-used EAE datasets. "Doc." is short for "Document" and "Arg." is short for "Argument". "-" denotes not applicable due to lack of document structure or corresponding annotations. "Fine-grained Event" means annotating all the events rather than only one topic event for a document. "Doc. Level" means annotating arguments within the whole document rather than only the sentence containing the trigger. For multilingual datasets, we only compare with its English subset.

argument, annotators can either choose an entity from the whole document or select a continuous textual span; once an entity mention is selected, all of its coreferent entity mentions are automatically selected. Annotators have the option to report errors in the event type annotation of a trigger, which allows for the discarding of that trigger. In the annotation process, approximately 4% of triggers are discarded. We employ 202 annotators, including 71 senior annotators selected during annotation training and 33 experts with rich annotation experiences. The annotation is divided into three phases. Each document is first annotated by an ordinary annotator, and then modified by a senior annotator. Finally, an expert will check whether the annotation accuracy reaches 90%. If not, the document's annotation will be returned to the second phase. To measure data quality, we randomly sample 100 documents and conduct the three-phrase annotation for them twice with different annotator groups. The Fleiss' kappa is 68.6%, which indicates a satisfactory level of annotation agreement. More annotation details are shown in appendix A.

3 Data Analysis

3.1 Data Statistics

Table 2 shows the main statistics of MAVEN-ARG compared with various existing EAE datasets. Appendix B.1 further shows the statistics of different splits. We can observe that MAVEN-ARG has two advantages: (1) MAVEN-ARG has the largest data scale, surpassing previous datasets by several times. This ensures that even for long-tail event types, MAVEN-ARG has sufficient data to fully train and stably evaluate EAE models. (2) The exhaustive annotation of MAVEN-ARG makes it the only dataset that covers all settings of EAE task. MAVEN-ARG includes complete annotations of entity and event coreference and annotates both entity and non-entity arguments for all fine-grained events at the document level. This allows MAVEN-ARG to support the evaluation of all variants of EAE methods and the development of comprehensive event understanding applications.

3.2 Data Distribution

We present the distributions of the annotated entity and event arguments of MAVEN-ARG in Figure 2. Argument roles with the same name across different event types are merged for presentation clarity. We observe that: (1) The distribution of entity types is generally similar to that of Few-NERD (Ding et al., 2021), demonstrating sufficient diversity. (2) The three most frequent basic argument roles (`Agent`, `Patient`, and `Location`) account for over 60% of event arguments. This highlights their ubiquity and encourages knowledge transfer among different event types in EAE methods. (3) Event arguments exhibit a highly long-tailed distribution. The 136 argument roles counted as "Others", each constituting less than 1.5%, collectively accounts for 27.8% of event arguments. The long-tailed distribution of MAVEN-ARG poses a significant challenge to model generalizability.

3.3 Trigger-argument Distance

We analyze the distribution of trigger-argument distances in Figure 3. For events with multiple coreferent triggers and entity arguments with multiple entity mentions, the distance is calculated between the nearest trigger-argument pairs. The overall average trigger-argument distance is 37.8. From Figure 3, we observe that while the majority of event

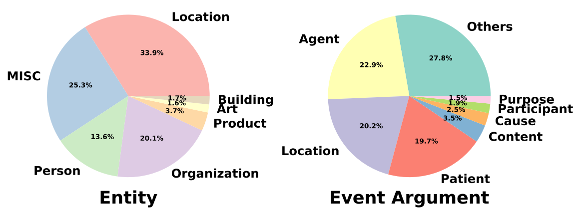

Figure 2: MAVEN-ARG entity and event argument distributions. For clarity, only the top event argument roles are shown and the others are summed up in "Others".

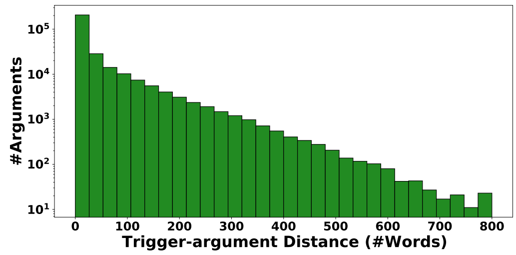

Figure 3: Distribution of distances between triggers and arguments in MAVEN-ARG.

arguments are located near their triggers, which is natural for human writing, a substantial number of arguments are situated far from their triggers, with the furthest exceeding 800 words. This data characteristic challenges the ability of EAE methods to capture long-distance dependencies.

4 Experiment

4.1 Experimental Setup

Models To assess the challenge of MAVEN-ARG, we evaluate multiple advanced methods. For fine-tuned EAE models, we implement several state-of-the-art models, including **DMBERT** (Wang et al., 2019), **CLEVE** (Wang et al., 2021b), **BERT+CRF** (Wang et al., 2020), **EEQA** (Li et al., 2020), **Text2Event** (Lu et al., 2021), and **PAIE** (Ma et al., 2022). These methods cover all the mainstream EAE modeling paradigms (Peng et al., 2023c). Their detailed descriptions and implementations are introduced in appendix C.1.

We also evaluate large language models (LLMs) with in-context learning on MAVEN-ARG. Specifically, we select two advanced LLMs, **GPT-3.5** (OpenAI, 2022) and **GPT-4** (OpenAI, 2023), and evaluate them with 2-shot in-context learning. Here 2-shot means using full annotations of two documents as demonstrations. Considering time and cost constraints, we sample 50 documents from the test set for experimentation. We employ the gold trigger evaluation approach (Peng et al., 2023c) to directly assess their EAE performance.

Evaluation Metric Considering that MAVEN-ARG covers non-entity argument annotations, traditional evaluation metrics (Peng et al., 2023c) designed only for entity arguments are no longer applicable. By taking each argument role as a question to the document, we propose to view EAE as a **multi-answer question answering** task[2] and adopt its evaluation metrics (Rajpurkar et al., 2016; Amouyal et al., 2022; Yao et al., 2023), including **bag-of-words F1** and **exact match (EM)**.

Conventional evaluation calculates the micro average over all the entity and event mentions, which we dub it as **mention-level** evaluation. Considering that real-world applications only require the accurate prediction for one of all the coreferent mentions, we propose to consider entity (Li et al., 2021) and event coreference in evaluation. Specifically, for **entity coreference level** evaluation, an entity argument is considered as predicted correctly if one of its mentions is predicted correctly. For **event coreference level** evaluation, an argument is considered as predicted correctly if it is predicted correctly for one of the coreferent triggers.

4.2 Experiment Results of Fine-tuned Models

The results of fine-tuned EAE models are shown in Table 3, and we have the following observations:

(1) Existing state-of-the-art EAE models exhibit moderate performance on MAVEN-ARG, which is significantly worse than their results on existing datasets (Peng et al., 2023c). This indicates that MAVEN-ARG is challenging and there is a need for increased efforts in developing practical event understanding models. (2) The BERT+CRF and PAIE models exhibit the best performance, potentially attributable to their ability to model rich interactions between different event arguments. (3) The previous top-performing classification-based models (DMBERT and CLEVE) (Peng et al., 2023c) perform poorly on MAVEN-ARG, which is due to their inability to handle non-entity arguments. Therefore, future research necessitates more flexible approaches to tackle the complex and real-world scenario in MAVEN-ARG. (4) Text2Event notably underperforms. This is potentially due to the intensive annotations of MAVEN-ARG, i.e., a high volume of events and argument annotations within a single document, making generating all events

[2]A single role may correspond to multiple argument spans.

Model	#Params	Mention Level				Entity Coref Level				Event Coref Level			
		P	R	F1	EM	P	R	F1	EM	P	R	F1	EM
DMBERT	110M	19.7	19.7	19.7	19.5	12.5	12.4	12.4	12.3	11.8	11.8	11.8	11.6
CLEVE	355M	22.1	22.1	22.1	22.0	13.2	13.2	13.2	13.0	12.3	12.2	12.2	12.1
BERT+CRF	110M	31.7	31.4	30.9	27.0	33.5	32.8	32.2	27.1	32.3	31.8	31.2	26.3
EEQA	110M	21.4	19.5	19.6	15.8	24.5	22.9	22.8	18.8	23.7	22.2	22.1	18.1
Text2Event	770M	12.9	12.9	12.7	11.3	12.5	12.4	12.1	10.4	10.8	10.7	10.5	9.0
PAIE	406M	**37.2**	**36.2**	**35.6**	**30.3**	**42.3**	**41.1**	**40.5**	**34.4**	**42.1**	**41.0**	**40.3**	**34.3**

Table 3: Experimental results (%) of existing state-of-the-art fine-tuned EAE models on MAVEN-ARG.

and arguments at once difficult. It indicates that generating complex structured outputs remains a major challenge for generation models (Peng et al., 2023a), requiring further exploration.

4.3 Experiment Results of LLMs

The results of LLMs with in-context learning are presented in Table 4, revealing that while LLMs with in-context learning are competitive compared to some fine-tuned EAE models, they still fall significantly short of the state-of-the-art. This is consistent with previous findings, suggesting that existing LLMs with in-context learning perform notably worse on specification-heavy information extraction tasks (Peng et al., 2023a; Li et al., 2023; Han et al., 2023). The LLMs' bag-of-words F1 scores are notably higher than their exact match scores, suggesting that the LLMs' predictions tend to be free-format and do not strictly match human annotations (Han et al., 2023).

One possible reason for the suboptimal performance is that LLMs cannot easily understand the schema from their names. Therefore, we conduct experiments with more informative prompts by incorporating definitions for each used argument role into the prompt, which are high-quality instructions used for guiding human annotators during data annotation. The results of these enhanced prompts are also shown in Table 4 (`w/ definition`). There is an obvious but marginal improvement after adding definitions, possibly due to the LLMs' limitations in understanding long contexts (Shaham et al., 2022; Peng et al., 2023a; Liu et al., 2023).

4.4 Analysis on Trigger-Argument Distance

As shown in Figure 3, MAVEN-ARG provides document-level annotations, covering data with varying trigger-argument distances. We conduct an analytical experiment on the impact of trigger-argument distance to model performance. Specifically, we break down the predictions and annotations in the test set by their trigger-argument distances and evaluate how the performance changes along with different distances. The experimental results are shown in Figure 4, which demonstrate that models perform significantly worse on samples with longer trigger-argument distances. This aligns with previous findings in document-level relation extraction regarding the distance between entity pairs (Ru et al., 2021). It suggests that modeling long-distance dependencies between triggers and arguments remains a challenge for existing EAE models. Future research can leverage MAVEN-ARG to explore advanced methods for handling long-distance trigger-argument instances.

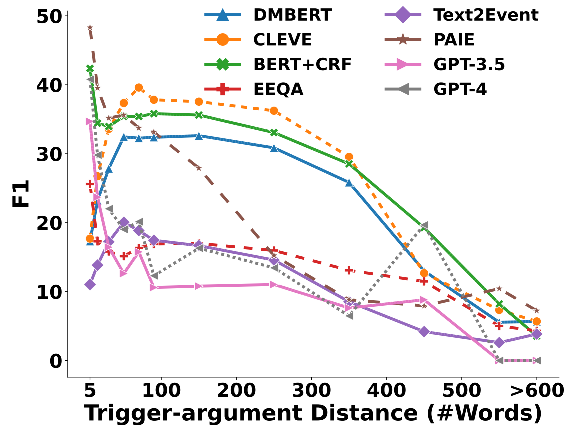

Figure 4: Mention-level F1 (%) of models on data with varying trigger-argument distances, i.e., the number of words between an event argument and its trigger.

4.5 Analysis on Entity and Non-Entity Arguments

MAVEN-ARG provides comprehensive annotations, including both entity and non-entity arguments. We analyze the performance breakdown of investigated EAE models on these two types of arguments. The results are presented in Table 5, which reveals that EAE models generally perform better on non-entity arguments. The possible reason may be that there are more non-entity arguments in MAVEN-ARG and non-entity arguments are often presented in a

Model	Mention Level				Entity Coref Level				Event Coref Level			
	P	R	F1	EM	P	R	F1	EM	P	R	F1	EM
GPT-3.5	21.3	20.9	19.9	14.3	24.5	25.1	23.4	16.8	24.4	24.8	23.2	16.9
w/ definition	21.8	21.7	20.6	15.2	25.0	25.8	24.1	17.8	24.9	25.4	23.9	17.9
GPT-4	25.6	27.2	25.1	17.9	28.9	31.7	28.7	20.2	27.9	30.5	27.6	19.5
w/ definition	**27.2**	**28.7**	**26.6**	**19.1**	**30.5**	**33.3**	**30.3**	**21.3**	**29.8**	**32.3**	**29.5**	**21.1**

Table 4: Experimental results (%) of LLMs with 2-shot in-context learning on MAVEN-ARG.

Model	Entity		Non-Entity	
	F1	EM	F1	EM
DMBERT	19.7	19.5	–	–
CLEVE	22.1	22.0	–	–
BERT+CRF	17.8	18.5	19.4	24.0
EEQA	6.2	5.6	17.5	13.9
Text2Event	5.5	5.2	1.6	1.1
PAIE	20.3	19.2	37.6	30.4

Table 5: Mention-level results (%) of EAE models on entity and non-entity arguments. Classification-based models, e.g., DMBERT and CLEVE, are not applicable to non-entity arguments.

looser form, making it easier for the models to learn the patterns and extract them. An exception is observed for the generation-based model Text2Event, which exhibits poorer performance on non-entity arguments. This may be because non-entity arguments are typically longer, which are harder to generate at once. It suggests that further exploration is needed to investigate how to effectively handle EAE with generation methods.

5 Future Event Prediction Demonstration

MAVEN-ARG, in conjunction with MAVEN and MAVEN-ERE, creates the first all-in-one event understanding benchmark, which covers the full process of ED, EAE, and ERE. Beyond serving as an evaluation benchmark for these tasks, an all-in-one event dataset naturally enables a variety of event-based applications, especially considering the recent advances brought by LLMs. Here we preliminarily explore an application case, future event prediction, as a demonstration.

Predicting future events based on causality can help decision-making, which is of self-evident importance. Therefore, since the early script learning (Schank and Abelson, 1975; Mooney and DeJong, 1985), future event prediction has continually attracted research interest (Chambers and Jurafsky, 2008; Jans et al., 2012; Granroth-Wilding and Clark, 2016; Hu et al., 2017; Chaturvedi et al., 2017; Li et al., 2018b; Lee and Goldwasser, 2019; Zhao, 2021). However, due to the lack of high-quality event resources, the evaluation of future event prediction often compromises by merely predicting verbs and subjects (Chambers et al., 2014), predicting according to textual order (Jans et al., 2012), or selecting story endings (Mostafazadeh et al., 2016a; Chaturvedi et al., 2017). The MAVEN series of datasets, with annotations of complete event structures and rich causal relations, may aid in predicting future events in real-world scenarios.

Model	Reasonable (%)	Matched (%)
GPT-3.5	92.7	7.8
GPT-4	95.2	12.2

Table 6: Future event prediction results (%), averaged over 2 evaluators and 3 prompts. **Reasonable** denotes the rate of predictions judged as reasonable to happen next. **Matched** denotes the rate of predictions matched with the actual future events.

Experiment Setup We sample 100 event chains, each consisting of 3-5 events, from the training and validation sets. In each chain, preceding events cause the subsequent ones. Events are described in a structured JSON format, containing event type, event trigger, and event arguments. For each event chain, we hold out the last event and input the remaining incomplete chain into two proprietary LLMs, GPT-3.5 and GPT-4 (OpenAI, 2023), requiring them to predict the next occurring event. These LLMs are prompted with detailed task instructions and 5 demonstration event chains. To minimize the influence of the demonstrations, predictions are made independently three times under different demonstrations. More experimental details are shown in appendix D. We employ manual evaluation, with two experts engaged to judge (1) whether the prediction is reasonable, and (2) whether the prediction matches the actual future event.

Experimental Results Experimental results are shown in Table 6. From these, we can see that the powerful LLMs can produce highly reasonable event predictions. However, their predictions seldom align with the actual future, making them not directly helpful. These observations suggest that using LLMs for future event prediction is promising, but there remain topics to explore on how to build practical future event prediction systems with LLMs. For instance, using retrieval-augmented methods may help LLMs access more timely evidence when making future predictions. As a preliminary attempt, the experiments demonstrate how our all-in-one event understanding dataset can assist in conveniently building and evaluating event-based applications. We hope that future works can explore using the MAVEN series datasets to build diverse applications.

6 Related Work

Event Argument Extraction Datasets Since the early MUC datasets (Grishman and Sundheim, 1996), event argument extraction (EAE) as a part of event extraction has received widespread attention. To reduce task complexity and provide standardized extraction results, the ACE datasets (Doddington et al., 2004) are designed with a schema covering 33 event types, limiting event argument annotation to entities within the same sentence as the trigger. ACE 2005 (Walker et al., 2006) has been the most widely used dataset for a long time, and the practice of ACE has been broadly adopted. Rich ERE (Song et al., 2015) expands ACE schema to 38 event types and constructs the TAC KBP datasets (Ellis et al., 2014, 2015, 2016; Getman et al., 2017). MEE (Pouran Ben Veyseh et al., 2022) follows the ACE schema to build a multilingual dataset. With the advancement of NLP methods, some works break some of the constraints of ACE task definition to construct more practical datasets. RAMS (Ebner et al., 2020), WikiEvents (Li et al., 2021), and DocEE (Tong et al., 2022) extends the annotation scope to the whole documents. However, RAMS and DocEE only annotate one topic event per document, ignoring fine-grained events within documents. MAVEN (Wang et al., 2020) and GENEVA (Parekh et al., 2023) both construct high-coverage general event schemata with over 100 event types. MAVEN supports only event detection. GENEVA extends event arguments to cover non-entity spans but focuses on testing the generalizability rather than developing practical EAE methods. Its data are repurposed from FrameNet (Baker et al., 1998) examples, which are individual sentences without document structure. MAVEN-ARG meticulously designs 612 unique argument roles for MAVEN schema and conducts large-scale exhaustive annotation, which annotates both entity and non-entity arguments for fine-grained events at the document level.

Event Argument Extraction Methods Traditional EAE methods primarily involve (1) Classification-based methods (Chen et al., 2015a, 2017; Sha et al., 2018; Wadden et al., 2019; Wang et al., 2019; Lin et al., 2020; Wang et al., 2021b; Zhou and Mao, 2022): employing text encoders like CNN (Krizhevsky et al., 2012) and BERT (Devlin et al., 2019), followed by an information aggregator, such as dynamic multi-pooling mechanism (Chen et al., 2015a), to obtain role-specific representations for classification. (2) Sequence labeling methods (Nguyen et al., 2016; Yang and Mitchell, 2017; Nguyen et al., 2021; Peng et al., 2023c): mainly adopting the conditional random field (CRF) (Lafferty et al., 2001) as the output layer to model structured dependencies between different arguments. Recently, increasing attention has been paid to transforming EAE into a question-answering task, transferring question-answering capabilities to boost EAE (Liu et al., 2020; Du and Cardie, 2020; Li et al., 2020; Ma et al., 2022; Lu et al., 2023; Nguyen et al., 2023). Additionally, some research focuses on using generation models to directly generate structured outputs containing events and their arguments (Lu et al., 2021; Li et al., 2021; Lu et al., 2022; Ren et al., 2023; You et al., 2022, 2023; Hsu et al., 2022, 2023; Zhang et al., 2023), which has been becoming increasingly important with the advance of large language models.

7 Conclusion and Future Work

We introduce MAVEN-ARG, an event argument extraction dataset with comprehensive schema, large data scale, and exhaustive annotation. Experiments indicate that MAVEN-ARG is quite challenging for both fine-tuned EAE models and proprietary large language models. Together with MAVEN and MAVEN-ERE, MAVEN-ARG completes an all-in-one dataset covering the entire process of event understanding. An application case of future event prediction demonstrates how an all-in-one dataset can enable broad event-based applications. In the future, we will explore constructing multilingual resources under this framework and developing practical EAE methods with MAVEN-ARG.

Limitations

(1) MAVEN-ARG currently includes only English corpus, which limits its potential applications and coverage for diverse linguistic phenomena. In future work, we will try to support more languages under our framework and we also encourage community efforts in developing multilingual event understanding benchmarks. (2) MAVEN-ARG, along with MAVEN (Wang et al., 2020) and MAVEN-ERE (Wang et al., 2022), exclusively supports mainstream event understanding tasks. However, these datasets do not cover more broad event-related tasks such as event factuality identification (Qian et al., 2019, 2022) and event salience identification (Liu et al., 2018). We encourage future explorations in building more challenging and diverse tasks and applications on top of MAVEN data. (3) While previous research has found that LLMs perform poorly on specification-heavy tasks (Peng et al., 2023c; Han et al., 2023; Li et al., 2023) including the EAE task, there is no in-depth exploration of effective LLM-based approaches addressing the EAE task in this paper. We leave the exploration of how to better leverage LLMs for EAE tasks in future work.

Ethical Considerations

In this section, we discuss the ethical considerations of this work: (1) **Intellectual property.** The MAVEN dataset is released under the CC BY-SA 4.0 license[3]. The MAVEN-ERE is shared under GPLv3[4] license and the original Wikipedia corpus is shared under the CC BY-SA 3.0 license[5]. The usage of these data in this work strictly adheres to the corresponding licenses and intended use. (2) **Intended use.** MAVEN-ARG is an event argument extraction dataset. Researchers and practitioners can utilize MAVEN-ARG to train and evaluate models for event argument extraction, thereby advancing the field of event understanding. (3) **Potential risk control.** MAVEN-ARG is constructed based on publicly available data. We believe that the underlying public data has been adequately desensitized and anonymized. The event argument annotation does not involve judgments about social issues and thus we believe MAVEN-ARG will not involve additional risks. To avoid unfair comparisons caused by mismatched evaluation implementations (Peng et al., 2023c) and potential cheating behaviors, the event argument annotations of MAVEN-ARG test set will not be publicly released. Instead, following previous works (Rajpurkar et al., 2016; Wang et al., 2020, 2022), we will maintain an online judgment system with a leaderboard, allowing users to submit predictions and obtain evaluation results. (4) **Worker Treatments** are discussed in appendix A.2. (5) **AI assistant.** The writing of this paper was assisted by ChatGPT in rephrasing some sentences.

[3] https://creativecommons.org/licenses/by-sa/4.0/

[4] https://www.gnu.org/licenses/gpl-3.0.html

[5] https://creativecommons.org/licenses/by-sa/3.0/

References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53.

Samuel Joseph Amouyal, Ohad Rubin, Ori Yoran, Tomer Wolfson, Jonathan Herzig, and Jonathan Berant. 2022. QAMPARI: : An open-domain question answering benchmark for questions with many answers from multiple paragraphs. *CoRR*, abs/2205.12665.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of ACL-COLING*, pages 86–90.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-HLT*, pages 789–797.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of EMNLP*, pages 1603–1614.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically Labeled Data Generation for Large Scale Event Extraction. In *Proceedings of ACL*, pages 409–419.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015a. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL-IJCNLP*, pages 167–176.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015b. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL*, pages 167–176.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of LREC*, pages 4545–4552.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of ACL*, pages 3198–3213.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Proceedings of IJCAI*.

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 LREC*.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of EMNLP*, pages 671–683.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of ACL*, pages 8057–8077.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2015. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *TAC*.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2016. Overview of Linguistic Resources for the TAC KBP 2016 Evaluations: Methodologies and Results. In *TAC*.

Joe Ellis, Jeremy Getman, and Stephanie M Strassel. 2014. Overview of linguistic resources for the TAC KBP 2014 evaluations: Planning, execution, and results. In *TAC*.

Charles J Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie Strassel. 2017. Overview of linguistic resources for the tac kbp 2017 evaluations: Methodologies and results. In *TAC*.

Goran Glavaš and Jan Šnajder. 2014. Event graphs for information retrieval and multi-document summarization. *Expert systems with applications*, 41(15):6904–6916.

Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: an interactive system for open-domain story generation. In *Proceedings of NAACL: Demonstrations*, pages 89–97.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the AAAI*, volume 30.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *Proceedings of COLING*.

Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5w1h: A universal system for extracting main events from news articles. In *Proceedings of INRA@RecSys*, CEUR Workshop Proceedings, pages 35–43.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by ChatGPT? An analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Proceedings of Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28.

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 NAACL-HLT*, pages 1890–1908.

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 ACL*, pages 10976–10993.

Linmei Hu, Juanzi Li, Liqiang Nie, Xiao-Li Li, and Chao Shao. 2017. What happens next? future subevent prediction using contextual hierarchical lstm. In *Proceedings of AAAI*, volume 31.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of EMNLP-IJCNLP*, pages 2391–2401.

Bram Jans, Steven Bethard, Ivan Vulic, and Marie-Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings EACL*, pages 336–344.

Çagla Çig Karaman, Serkan Yaliman, and Salih Atilay Oto. 2017. Event detection from social media: 5w1h analysis on big data. In *Proceedings of SIU*, pages 1–4. IEEE.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of NeurIPs*, pages 1106–1114.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289.

I-Ta Lee and Dan Goldwasser. 2019. Multi-relational script learning for discourse relations. In *Proceedings of ACL*, pages 4214–4226.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*, pages 3045–3059.

Willem JM Levelt. 1993. *Speaking: From intention to articulation*. MIT press.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Findings of EMNLP*, pages 829–838.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of NAACL-HLT*, pages 894–908.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018a. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of IJCAI*, pages 4201–4207.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018b. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of IJCAI*.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of ACL*, pages 7999–8009.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of EMNLP*, pages 1641–1651.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Zhengzhong Liu, Chenyan Xiong, Teruko Mitamura, and Eduard Hovy. 2018. Automatic event salience identification. In *Proceedings of EMNLP*, pages 1226–1236.

Di Lu, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2023. Event extraction as question generation and answering. In *Proceedings of ACL*, pages 1666–1688.

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 ACL-IJCNLP*, pages 2795–2806.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of ACL*, pages 5755–5772.

Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. EventPlus: A temporal event understanding pipeline. In *Proceedings of NAACL: Demonstrations*, pages 56–65.

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 ACL*, pages 6759–6774.

George A Miller and Philip N Johnson-Laird. 2013. Language and perception. In *Language and Perception*. Harvard University Press.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19.

Raymond J Mooney and Gerald DeJong. 1985. Learning schemata for natural language processing. In *Proceedings of IJCAI*, pages 681–687.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61.

Chien Nguyen, Hieu Man, and Thien Nguyen. 2023. Contextualized soft prompts for extraction of event arguments. In *Findings of ACL 2023*, pages 4352–4361.

Minh Van Nguyen, Tuan Ngo Nguyen, Bonan Min, and Thien Huu Nguyen. 2021. Crosslingual transfer learning for relation and event extraction via word category and class alignments. In *Proceedings of EMNLP*, pages 5414–5426.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL-HLT*, pages 300–309.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of ACL*, pages 1318–1328.

OpenAI. 2022. Introducing ChatGPT.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of ACL*, pages 3664–3686.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, et al. 2023a. When does in-context learning fall short and why? a study on specification-heavy tasks. *arXiv preprint arXiv:2311.08993*.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023b. Omnievent: A comprehensive, fair, and easy-to-use toolkit for event understanding. pages 508–517.

Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu, and Weixing Shen. 2023c. The devil is in the details: On the pitfalls of event extraction evaluation. In *Findings of ACL*.

Steven Pinker. 2013. *Learnability and Cognition, new edition: The Acquisition of Argument Structure*. MIT press.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. MEE: A novel multilingual event extraction dataset. In *Proceedings of the EMNLP*, pages 9603–9613.

Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2019. Document-level event factuality identification via adversarial neural network. In *Proceedings of NAACL-HLT*, pages 2799–2809.

Zhong Qian, Heng Zhang, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2022. Document-level event factuality identification via machine reading comprehension frameworks with transfer learning. In *Proceedings of COLING*, pages 2622–2632.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

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 ACL*, pages 293–306.

Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. 2021. Learning logic rules for document-level relation extraction. In *Proceedings of EMNLP*, pages 1239–1250.

Roger C Schank and Robert P Abelson. 1975. Scripts, plans, and knowledge. In *Proceedings of IJCAI*, volume 75, pages 151–157.

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 AAAI*, pages 5916–5923.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of EMNLP*, pages 12007–12021.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of NAACL-HLT*, pages 3970–3982.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of EMNLP-IJCNLP*, pages 5783–5788.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium*, 57.

Shichao Wang, Xiangrui Cai, Hongbin Wang, and Xiaojie Yuan. 2021a. Incorporating circumstances into narrative event prediction. In *Findings of EMNLP*, pages 4840–4849.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, et al. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of EMNLP*, pages 926–941.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of EMNLP*, pages 1652–1671.

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 EMNLP-IJCNLP*, pages 5777–5783.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021b. CLEVE: Contrastive Pre-training for Event Extraction. In *Proceedings of ACL-IJCNLP*, pages 6283–6297.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*, pages 38–45.

Bishan Yang and Tom M. Mitchell. 2017. Leveraging knowledge bases in lstms for improving machine reading. In *Proceedings of ACL*, pages 1436–1446.

Zijun Yao, Yantao Liu, Xin Lv, Shulin Cao, Jifan Yu, Juanzi Li, and Lei Hou. 2023. KoRC: Knowledge oriented reading comprehension benchmark for deep text understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11689–11707.

Huiling You, David Samuel, Samia Touileb, and Lilja Øvrelid. 2022. EventGraph: Event extraction as semantic graph parsing. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 7–15.

Huiling You, Lilja Vrelid, and Samia Touileb. 2023. JSEEGraph: Joint structured event extraction as graph parsing. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 115–127.

Kaihang Zhang, Kai Shuang, Xinyue Yang, Xuyang Yao, and Jinyu Guo. 2023. What is overlap knowledge in event argument extraction? APE: A cross-datasets transfer learning model for EAE. In *Proceedings of ACL*, pages 393–409.

Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Comput. Surv.*, 54(5).

Hanzhang Zhou and Kezhi Mao. 2022. Document-level event argument extraction by leveraging redundant information and closed boundary loss. In *Proceedings of NAACL-HLT*, pages 3041–3052.

Appendices

A Data Collection Details

A.1 Annotation Instruction

As introduced in § 2.1, we create a detailed event schema for both defining the task and instructing the annotators. We present the annotation instructions for the event type `Incident` in Table 7, including its argument schema and annotation examples. The overall event schema will be released along with the dataset. To support the highly customized annotation process designed for us, we developed a new online annotation platform. A screenshot for the annotation platform is shown in Figure 5 to help understand the annotation operations.

A.2 Annotation Coordination

We employ annotators (including senior annotators and expert annotators) from multiple commercial data annotation companies. 61% of them are female and 39% of them are male. Annotators for entity and event argument annotation have no overlap since we cooperated with different companies for the two annotation tasks. The experts involved in schema creation are invited by the authors through personal connections. All the workers are fairly paid with agreed salaries and workloads. All employment is under contract and in compliance with local regulations. The overall annotation cost, including annotating entities and event arguments as well as developing and maintaining annotation platforms, is about $85,000$ USD.

B Additional Data Statistics

B.1 Data Split Statistics

The detailed statistics of different data splits of MAVEN-ARG are shown in Table 8.

B.2 Differences with Predecessors

MAVEN-ARG inherits the efforts of previous works MAVEN (Wang et al., 2020) and MAVEN-ERE (Wang et al., 2022). MAVEN supports the event detection task by annotating event triggers and event types, along with a preliminary version of event coreferences. MAVEN-ERE supports the event relation extraction task by annotating event coreference, temporal, causal, and hierarchical relations. MAVEN-ARG completes the all-in-one event understanding dataset by adding the annotations of event arguments, which supports the event argument extraction task. The construction of MAVEN-ERE and MAVEN-ARG involves fixing or ignoring the erroneous and ambiguous annotations of event triggers and coreference clusters in MAVEN, which results in minor statistical differences shown in Table 9.

C EAE Experimental Details

C.1 Fine-tuning Implementation Details

Here we provide brief descriptions of the fine-tuning-based models involved in our experiments. (1) **DMBERT** (Wang et al., 2019) utilizes BERT (Devlin et al., 2019) as the text encoder and a dynamic multi-pooling mechanism (Chen et al., 2015b) on top of BERT to aggregate argument-specific features and map them onto the distribution in the label space. (2) **CLEVE** (Wang et al., 2021b) is an event-oriented pre-trained language model, which is pre-trained using contrastive pre-training objectives on large-scale unsupervised data and their semantic structures. (3) **BERT+CRF** (Wang et al., 2020) is a sequence labeling model, which leverages BERT as the backbone and the conditional random field (CRF) (Lafferty et al., 2001) as the output layer to model the structural dependencies of predictions. (4) **EEQA** (Li et al., 2020) is a span prediction model, which formulates event extraction as a question-answering task and outputs start and end positions to indicate triggers and arguments. (5) **Text2Event** (Lu et al., 2021) is a conditional generation model, which proposes a sequence-to-structure paradigm and generates structured outputs containing triggers and corresponding arguments with constrained decoding. (6) **PAIE** (Ma et al., 2022) adopts prompt tuning (Lester et al., 2021) to train two span selectors for each argument role in the provided prompt and conduct joint optimization to find optimal role-span assignments. We adopt the same backbones with their original papers for all EAE models in our experiments. We employ pipeline evaluation as suggested by Peng et al. (2023c). Specifically, for PAIE, we conduct EAE experiments based on the triggers predicted by CLEVE. For the other models, the EAE experiments are based on the triggers extracted by corresponding models.

We implement the EAE models using code from the official repositories of OmniEvent (Peng et al., 2023b), PAIE (Ma et al., 2022), and Text2Event (Lu et al., 2021). The numbers of parameters of the EAE models are shown in Table 3. All open-source models are downloaded from the

[Incident] Accident, unfortunate event

Event Arguments:

1. Participant : Entities involved in the accident (individuals, institutions, organizations, and even trains, ships, etc.). They can be the ones causing the accident or the ones affected by it. Similar to the combination of Agent and Patient in previous events, but due to the difficulty in distinguishing between Agent and Patient in accidents, they are uniformly labeled as Participants.

2. Location : The location or position where the incident occurred. If the incident involves multiple locations during the process, they should be marked separately.

3. Content : In general, only one annotation is needed, which accurately indicates the content and type of the accident.

4. Loss : The losses caused by accidents can include the number of deaths and injuries, property damage, and so on.

Annotation Examples:

1. British losses were confined to a single man wounded by an **accident** aboard "Crescent" .

2. On 6 June 1982, during the Falkland's war, the British Royal Navy type 42 destroyer engaged and destroyed a [British army gazelle helicopter, serial number "XX377"] Participant + Loss , in a friendly fire **incident**, killing all four occupants .

Table 7: Example annotation instructions for event type `Incident`. Different argument roles are denoted by different background colors. **Triggers** are bolded in red.

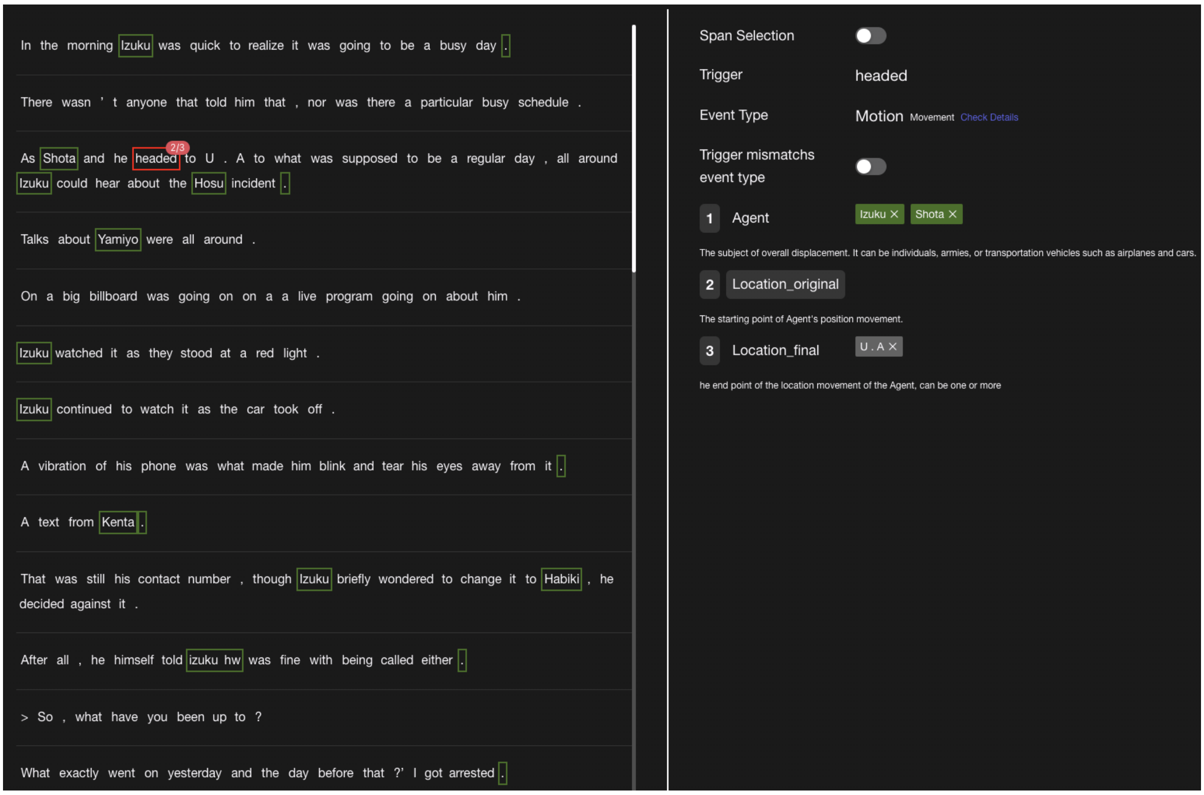

Figure 5: Screenshot for the annotation platform. The trigger "headed" is selected for annotation (in the right panel) and entities are highlighted in green as the options for annotating event arguments.

Dataset	#Doc.	#Event	#Trigger	#Arg.	#Entity	#Entity Mention	Entity Arg.	Non-Entity Arg.
Train	2,913	64,923	70,775	190,479	55,421	86,969	76,882	113,597
Dev	710	15,556	16,996	46,458	12,927	18,806	18,040	28,418
Test	857	18,112	19,736	53,676	15,297	23,351	21,102	32,574

Table 8: Statistics of the data splits of MAVEN-ARG. "Doc." is short for "Document" and "Arg." is short for "Argument".

	#Trigger	#Coreference Cluster
MAVEN	118,732	111,611
MAVEN-ERE	112,276	103,193
MAVEN-ARG	107,507	98,591

Table 9: Statistical differences between MAVEN-ARG and predecessors in number of event triggers and coreference clusters.

HuggingFace Transformers community (Wolf et al., 2020). Each of our fine-tuning experiments is conducted only once, on Nvidia A100 GPUs, consuming approximately 800 GPU hours in total. The hyper-parameters of the model are set based on prior experience and references from previous papers (Lu et al., 2021; Ma et al., 2022; Peng et al., 2023b). All hyper-parameters are shown in Table 10.

C.2 LLM Experimental Details

We access ChatGPT and GPT-4 through the official OpenAI interfaces, namely `gpt-3.5-turbo` and `gpt-4`, respectively. The API access period spans from October 1 to October 31, 2023. The decoding sampling temperature for both models is set to 0. An example of the prompt, input, output, and ground-truth of this experiment are presented in Table 11. Model outputs are automatically extracted and evaluated using the evaluation approach mentioned in § 4.1.

D Event Prediction Experimental Details

The future event prediction experiments (§ 5) were conducted in October and November, 2023. We use OpenAI API endpoints `gpt-3.5-turbo` and `gpt-4` for GPT-3.5 and GPT-4 experiments, specifically. To ensure the consistency among different runs, we set `temperature=0.0`. Detailed instructions and example input and output are shown in Table 12.

E More Experimental Results

In this section, we present more experimental results of using different proportions of training data for training (appendix E.1) and results on entity and non-entity arguments (appendix E.2).

E.1 Analysis on Data Size

The data volume of MAVEN-ARG significantly exceeds that of commonly used datasets. To examine the benefits of increased data scale, we train models on training data of varying sizes and observe their performance on the original test set. The experimental results are shown in Figure 6, which demonstrates that more training data indeed enhances model performance and allows for a comprehensive comparison of different models. The extensive data of MAVEN-ARG make it feasible to train a large language model (LLM) for general event understanding, which we leave as future work. Table 13 shows the detailed experimental results, i.e., mention level, entity coreference level, and event coreference level.

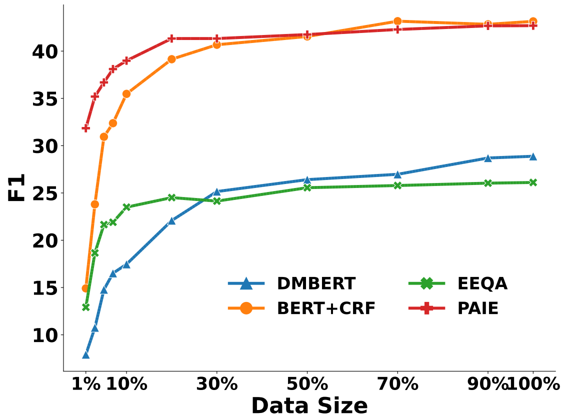

Figure 6: Mention-level F1 scores of investigated EAE models using different proportions of training data on MAVEN-ARG. This experiment adopts the gold trigger evaluation approach (Peng et al., 2023c) and hence Text2Event is not applicable. Due to the computation limitations, CLEVE is not evaluated in this experiment.

E.2 Entity and Non-Entity Arguments

Table 14 presents the overall results on entity and non-entity arguments of MAVEN-ARG. The non-entity arguments do not have coreferential relationship with each other and hence there is no entity coreference level evaluation for them.

	DMBERT	CLEVE	BERT+CRF	EEQA	PAIE	Text2Event
Learning Rate	5×10^{-5}	1×10^{-5}	5×10^{-5}	5×10^{-5}	2×10^{-5}	5×10^{-5}
Weight Decay	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-2}
Batch Size	32	128	64	32	16	8
Epoch	6	5	10	10	–	30
Warmup Rate	0.1	0.1	0.1	0.1	0	0.1

Table 10: Hyper-parameters of fine-tuning EAE models on MAVEN-ARG. PAIE utilizes $10,000$ gradient update steps to optimize the parameters.

PROMPT: Please extract event argument roles and the corresponding mentions for the events marked with <event> and </event> in the text, the possible roles must be chosen from the Roleset. If there are no roles for the event, place output NA.

INPUT: The February 2010 Australian **<event>** cyberattacks **</event>** were a series of denial-of-service attacks conducted by the Anonymous online community against the Australian government in response to proposed web censorship regulations . Operation Titstorm was the name given to the cyber attacks by the perpetrators . They resulted in lapses of access to government websites on 10 and 11 February 2010 . This was accompanied by emails , faxes , and phone calls harassing government offices . The actual size of the attack and number of perpetrators involved is unknown but it was estimated that the number of systems involved ranged from the hundreds to the thousands . The amount of traffic caused disruption on multiple government websites . Australian Telecommunications Minister Stephen Conroy proposed the regulations that would mainly filter sites with pornographic content . Various groups advocating uncensored access to the Internet , along with companies like Google and Yahoo ! , object to the proposed filter . A spokesperson for Conroy said that the actions were not a legitimate form of protest and called it irresponsible . The attacks also drew criticism from other filter protest groups . The initial stage was followed by small in-person protests on 20 February that were called “ Project Freeweb ” . Roleset: [Agent, Location, Patient, Consequence]. The event trigger is: **<event>** cyberattacks **</event>**.

GPT-3.5 OUTPUT: Answer: Patient: the Australian government; Consequence: lapses of access to government websites; Agent: the Anonymous online community

GPT-4 OUTPUT: Answer: Agent: the Anonymous online community; Location: Australia; Patient: Australian government; Consequence: lapses of access to government websites, disruption on multiple government websites.

GROUND-TRUTH: Agent: Anonymous online community; Patient: Australian government; Consequence: lapses of access to government websites on 10 and 11 February 2010; Location: Australian

Table 11: Example prompt, input, output, and ground-truth for the LLMs experiments. Demonstrations are omitted.

PROMPT: You need to complete the event chain prediction task. I will provide you with an event chain consisting of multiple events, where earlier events cause the occurrence of subsequent events. You need to predict what the next most likely event in this chain will be. Events will be provided in a structured JSON format, including event type, core event trigger words, and event elements. I will first give you 5 example chains: ... Now please predict the next event for the following event chain:

INPUT: Event No.1: {"event type": "Attack", "event trigger": ["attack"], "event argument": {"Agent": ["unknown assailant"], "Patient": ["Boris Nemtsov"], "Location": ["Bolshoy Moskvoretsky Bridge"]}} Event No.2: {"event type": "Use_firearm", "event trigger": ["fired"], "event argument": {"Agent": ["unknown assailant"], "Patient": ["Nemtsov"], "Location": ["Bolshoy Moskvoretsky Bridge"], "Instrument": ["Makarov pistol"]}} Event No.3: {"event type": "Bodily_harm", "event trigger": ["hit"], "event argument": {"Agent": ["shots"], "Patient": ["Boris Nemtsov"], "Location": ["Bolshoy Moskvoretsky Bridge"]}}

GPT-3.5 OUTPUT: Based on the provided event chains, the next most likely event in the given event chain is: Event No.4: {"event type": "Criminal_investigation", "event trigger": ["indicted"], "event argument": {"Agent": ["unknown assailant"], "Suspect": ["Boris Nemtsov"], "Incident": ["shooting on Bolshoy Moskvoretsky Bridge"], "Conclusion": ["investigation and charges"]}}

GPT-4 OUTPUT: Event No.4: {"event type": "Death", "event trigger": ["died"], "event argument": {"Agent": ["Boris Nemtsov"], "Cause": ["gunshot wounds"], "Location": ["Bolshoy Moskvoretsky Bridge"]}}

GROUND-TRUTH: Event No.4: {"event type": "Death", "event trigger": ["died"], "event argument": {"Agent": ["Boris Nemtsov"], "Location": ["Bolshoy Moskvoretsky Bridge"]}}

Table 12: Example prompt, input, output, and ground-truth for the future event prediction experiments. Demonstrations are omitted and the JSON strings are unformatted to avoid taking up to much space.

Proportion	Mention Level				Entity Coref Level				Event Coref Level			
	P	R	F1	EM	P	R	F1	EM	P	R	F1	EM
DMBERT												
1%	8.3	7.8	7.9	7.2	10.4	9.5	9.7	8.6	9.4	8.6	8.8	7.7
3%	11.1	10.7	10.8	10.1	11.9	11.2	11.3	10.3	10.2	9.6	9.7	8.8
5%	15.2	14.7	14.8	14.0	14.8	13.9	14.0	12.9	13.1	12.3	12.5	11.5
7%	17.0	16.4	16.5	15.7	17.9	16.8	17.1	15.7	16.4	15.4	15.6	14.4
10%	18.0	17.4	17.5	16.6	17.5	16.4	16.6	15.4	16.0	15.0	15.2	14.1
20%	22.6	22.0	22.1	21.2	21.0	19.8	20.0	18.6	19.3	18.2	18.4	17.1
30%	25.7	25.0	25.2	24.2	23.1	21.7	21.9	20.4	21.5	20.2	20.4	19.0
50%	26.9	26.3	26.4	25.5	23.9	22.7	22.9	21.4	22.1	21.0	21.2	19.8
70%	27.5	26.9	27.0	26.1	24.0	22.7	23.0	21.5	22.1	21.0	21.2	19.9
90%	29.2	28.6	28.7	27.8	24.7	23.5	23.8	22.2	22.9	21.8	22.0	20.6
BERT+CRF												
1%	16.4	14.8	14.9	11.6	21.7	19.6	19.8	15.4	21.1	19.1	19.3	15.1
3%	25.3	24.1	23.8	19.2	31.4	29.6	29.4	23.4	30.4	28.8	28.5	22.9
5%	32.5	31.2	30.9	25.8	38.8	36.8	36.5	29.6	38.0	36.2	35.9	29.2
7%	33.9	32.8	32.4	27.1	41.6	39.7	39.3	31.9	40.7	38.9	38.5	31.3
10%	36.8	36.0	35.5	30.0	43.1	41.8	41.1	33.9	42.3	41.1	40.4	33.4
20%	40.5	39.7	39.1	33.7	46.6	45.2	44.5	37.0	45.7	44.4	43.7	36.4
30%	42.0	41.3	40.7	35.1	47.2	45.9	45.2	37.6	46.4	45.3	44.5	37.0
50%	42.7	42.2	41.5	36.3	48.3	47.2	46.4	39.0	47.3	46.4	45.6	38.3
70%	44.3	43.9	43.2	37.8	49.5	48.5	47.6	40.0	48.6	47.7	46.8	39.4
90%	43.8	43.7	42.8	37.5	47.9	47.5	46.4	39.2	46.9	46.6	45.5	38.5
100%	44.3	43.9	43.1	37.7	48.4	47.5	46.6	39.3	47.4	46.6	45.8	38.7
EEQA												
1%	14.0	13.2	12.9	9.9	16.6	16.3	15.7	12.2	15.7	15.4	14.8	11.5
3%	20.5	18.8	18.7	14.4	23.3	22.1	21.7	17.0	22.4	21.3	20.8	16.4
5%	24.0	21.5	21.7	17.0	27.1	25.0	25.0	20.0	26.1	24.2	24.0	19.2
7%	24.2	21.8	21.9	17.1	27.1	25.2	25.0	20.0	26.4	24.5	24.4	19.5
10%	25.9	23.4	23.5	18.5	29.1	27.0	26.9	21.7	28.3	26.4	26.2	21.1
20%	26.7	24.5	24.5	19.6	29.9	28.2	28.0	22.9	29.1	27.4	27.2	22.2
30%	26.3	24.1	24.1	19.4	29.6	27.8	27.7	22.7	28.6	27.0	26.7	21.9
50%	27.8	25.5	25.6	20.7	31.3	29.4	29.2	24.0	30.3	28.5	28.3	23.2
70%	28.1	25.7	25.8	20.9	31.4	29.5	29.3	24.2	30.5	28.7	28.5	23.5
90%	28.2	26.0	26.0	21.0	31.5	29.8	29.6	24.4	30.7	29.1	28.8	23.7
100%	28.3	26.1	26.1	21.1	31.6	29.9	29.6	24.5	30.8	29.2	28.9	23.7
PAIE												
1%	33.6	32.7	31.8	25.3	39.3	38.4	37.4	30.1	39.1	38.3	37.3	30.0
3%	37.0	36.0	35.2	28.6	43.2	42.0	41.1	33.7	43.1	42.0	41.1	33.7
5%	38.6	37.4	36.7	30.0	45.9	44.4	43.6	35.8	46.2	44.8	43.9	36.1
7%	39.8	39.0	38.1	31.5	45.8	45.0	43.9	36.5	46.1	45.3	44.2	36.7
10%	40.6	40.0	39.0	32.4	46.9	46.2	45.1	37.6	47.2	46.6	45.4	37.9
20%	43.2	42.1	41.3	34.7	49.5	48.3	47.4	39.9	49.8	48.6	47.7	40.1
30%	43.2	42.1	41.3	34.7	49.5	48.3	47.4	39.9	49.8	48.6	47.7	40.1
50%	43.4	42.6	41.8	35.3	49.9	49.0	48.0	40.6	50.3	49.4	48.4	40.9
70%	44.0	43.0	42.3	35.8	50.7	49.5	48.7	41.3	51.2	50.1	49.1	41.7
90%	44.4	43.4	42.7	36.2	51.3	49.9	49.1	41.8	51.5	50.3	49.4	42.0
100%	44.5	43.4	42.7	36.3	50.8	49.4	48.7	41.4	51.1	49.8	49.0	41.7

Table 13: Experimental results (%) of the EAE models using different proportions of training data of MAVEN-ARG. In this experiment, we adopt the gold trigger evaluation approach (Peng et al., 2023c).

Model	#Params	Mention Level				Entity Coref Level				Event Coref Level			
		P	R	F1	EM	P	R	F1	EM	P	R	F1	EM
Entity Argument													
DMBERT	110M	19.7	19.7	19.7	19.5	12.5	12.4	12.4	12.3	11.8	11.8	11.8	11.6
CLEVE	355M	22.1	22.1	22.1	22.0	13.2	13.2	13.2	13.0	12.3	12.2	12.2	12.1
BERT+CRF	110M	18.6	18.5	18.5	17.8	12.7	12.6	12.6	12.0	12.0	11.8	11.8	11.3
EEQA	110M	6.3	6.2	6.2	5.6	9.1	9.1	9.0	8.3	8.5	8.5	8.4	7.7
Text2Event	770M	5.5	5.6	5.5	5.2	4.0	4.0	4.0	3.7	3.2	3.2	3.1	2.9
PAIE	406M	20.4	20.5	20.3	19.2	21.0	21.1	20.9	19.8	20.0	20.1	19.9	18.9
Non-Entity Argument													
BERT+CRF	110M	24.8	24.8	24.0	19.4	–	–	–	–	25.3	25.3	24.5	19.6
EEQA	110M	18.9	17.6	17.5	13.9	–	–	–	–	18.6	17.4	17.2	13.7
Text2Event	770M	1.7	1.7	1.6	1.1	–	–	–	–	1.5	1.6	1.5	1.1
PAIE	406M	39.4	38.4	37.6	30.4	–	–	–	–	38.7	37.8	36.9	29.7

Table 14: Experimental results (%) of existing state-of-the-art fine-tuned EAE models on entity and non-entity arguments of MAVEN-ARG. Classification-based models, e.g., DMBERT and CLEVE, are inapplicable to non-entity arguments.