DEIE: Benchmarking Document-level Event Information Extraction with a Large-scale Chinese News Dataset

Yubing Ren[1,2], Yanan Cao[1,2*], Hao Li[1,2], Yingjie Li[1,2],
Zixuan Ma[3], Fang Fang[1,2], Ping Guo[1,2], Wei Ma[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China,
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China,
[3] University of Pennsylvania
renyubing@iie.ac.cn

Abstract

A text corpus centered on events is foundational to research concerning the detection, representation, reasoning, and harnessing of online events. The majority of current event-based datasets mainly target sentence-level tasks, thus to advance event-related research spanning from sentence to document level, this paper introduces DEIE, a unified large-scale document-level event information extraction dataset with over 56,000+ events and 242,000+ arguments. Three key features stand out: large-scale manual annotation (20,000 documents), comprehensive unified annotation (encompassing event trigger/argument, summary, and relation at once), and emergency events annotation (covering 19 emergency types). Notably, our experiments reveal that current event-related models struggle with DEIE, signaling a pressing need for more advanced event-related research in the future. DEIE is now available at https://github.com/Lilice-r/DEIE.

Keywords: event research, unified annotation, resource, event-driven, information extraction

1. Introduction

The world is the totality of facts, not of things (Wittgenstein, 1994). In daily life, facts manifest as concrete events. Countless interrelated events constitute the real world, and these events conform to the common cognitive rules. Intuitively, when events occur, quick and precise identification or analysis is vital for governments, relief agencies, affiliated organizations, etc. With the recent surge in global emergencies, the urgency to efficiently extract relevant details from the deluge of news reports and keenly monitor these events' unfolding trajectories is more pressing than ever.

Most existing event-based datasets, such as ACE2005 (Christopher Walker and others, 2005) and KBP2017 (Stephanie Strassel and others, 2016), primarily target sentence-level event extraction. This process involves extracting the event trigger and its corresponding arguments from an individual sentence. The CEC (Fu et al., 2010) dataset, being auto-annotated, encompasses 332 emergency news documents, which is insufficient to capture the vast array of real-world events. However, it's noteworthy that in practical contexts, a significant portion of event components span multiple sentences. This observation has led us to adopt "document" as the descriptive term to represent comprehensive event information.

To promote document-level event-related research, various datasets have been proposed to capture precise event information. For instance, MUC-4 dataset (McLean, 1992) encompasses 1,700 documents addressing 4 event types, and these types are close to each other and confined to the terrorist attack topic. RAMS (Ebner et al., 2020) narrows the context to a mere 5 sentences, which poses a challenge to encapsulating document-level event details. WikiEvents (Li et al., 2021) offers only 246 annotated documents, of which a scant 22% incorporate cross-sentence argument annotations. ChFinAnn (Zheng et al., 2019) is restricted to the financial domain, featuring 5 event types and 35 argument types. DocEE (Tong et al., 2022), although large-scale, is geared toward a one-event-per-document situation, rendering it less adaptable to multi-event scenarios. Furthermore, the existing event-related dataset is limited to support only a single kind of event information and overlooks the unified annotation with other event summaries or relation information, etc. In summary, existing event-based corpus annotation fails in the following aspects: the small scale of data, limited coverage of domain, multi-event adaptability, and uncomprehensive event knowledge. Therefore, it is urgent to develop a manually labeled, unified large-scale dataset to accelerate the research in document-level event information extraction.

In response to the need for such large-scale and high-quality datasets, we introduce DEIE - a Unified Large-scale Document-level Event Information Extraction dataset. This resource meticulously annotates 20,000 documents, spanning 64 event types derived from publicly available Chinese news reports. With such abundant event data — 242k event arguments, 56k event summaries, and 35k

* Yanan Cao is corresponding author

LREC-COLING 2024, pages 4592–4604

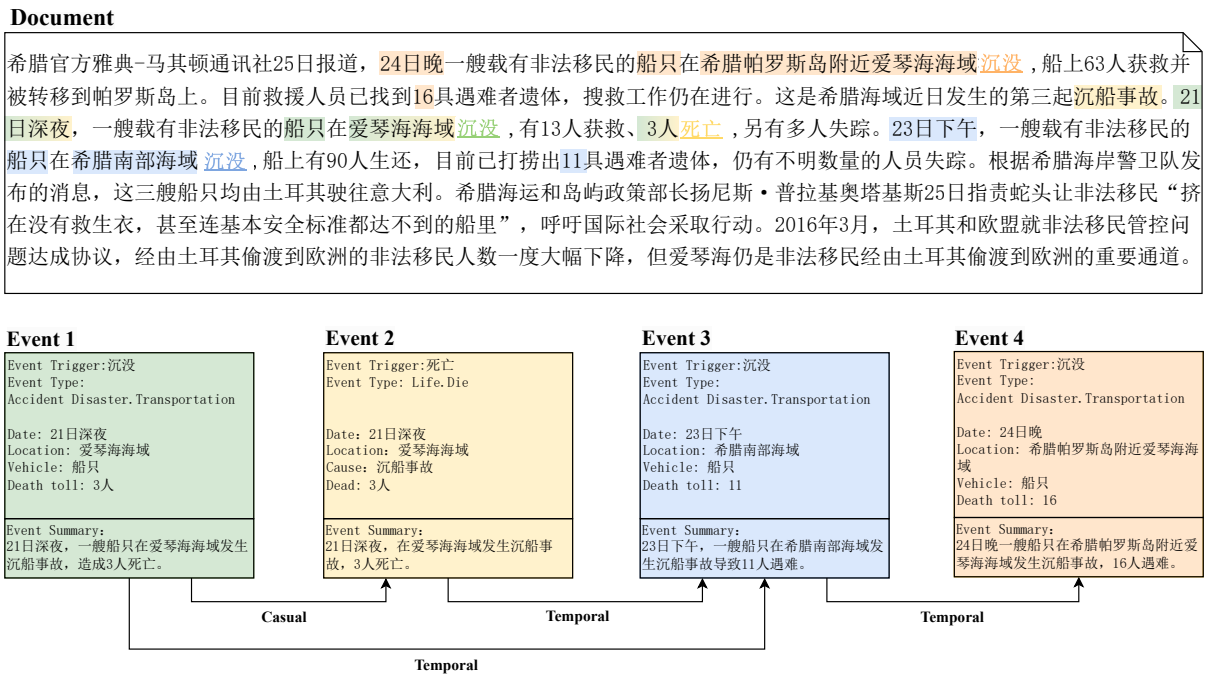

Figure 1: An accident disaster example from DEIE. Each document annotates multiple events and their relations. Each event contains an event trigger, type, arguments, and summary. We employ varied color schemes to differentiate between events, with the underlined “tokens” marking their specific triggers.

event-event relations — DEIE facilitates diverse studies, including Document-level Event Extraction, Event Causality Identification, Event Temporal Relation Extraction, and Event Summary Generation. The significance of DEIE can be distilled into three primary contributions: (1) **Large-scale Manual Annotation.** DEIE offers precise annotations for 56,031 events, accompanied by 243,287 arguments, tailored for intricate document-level event tasks. (2) **Comprehensive Unified Annotation.** To our understanding, DEIE pioneers as the first event-centric benchmark that concurrently incorporates event arguments, summaries, and relations. This integration not only propels research in document-level event information extraction but also aids in building an event-driven knowledge base. (3) **Emergency Event Annotation.** DEIE covers 19 emergency types, thus providing detailed insights into emergency situations, enabling faster and more informed decision-making during crises. Furthermore, DEIE can better cope with realistic multi-event-per-document scenarios, with each document annotating a minimum of two events.

Leveraging DEIE as the benchmark, we undertake extensive experiments to assess the state-of-the-art natural language processing (NLP) technologies across tasks such as document-level event argument extraction, event relation extraction, and event summary generation. Through rigorous testing across these three tasks, our analyses reveal that the introduced dataset poses distinct challenges. Current event-related models exhibit suboptimal performance, particularly in event argument extraction and temporal relation extraction. Thus calling for more research efforts for event-related studies in the future.

2. Related Datasets

Sentence-level Event Extraction Dataset. The Automatic Content Extraction (ACE2005) dataset includes 599 documents, spanning 8 event types and 33 subtypes. The Text Analysis Conference (TAC) has introduced three benchmark datasets (Stephanie Strassel and others, 2014, 2016): TAC-KBP 2015, TAC-KBP 2016, and TAC-KBP 2017, featuring 9, 8, and 8 event types, along with 38, 18, and 18 event subtypes respectively. Chinese Emergency Corpus (CEC) (Fu et al., 2010) is a Chinese dataset pertinent to emergencies, which provides 332 documents covering 5 categories. MAVEN (Wang et al., 2020) concentrates on annotating event triggers and contains 168 types within 11,832 sentences. These datasets have served as the foundation for the development of numerous models aimed at improving sentence-level event extraction, resulting in significant achievements in this field (Wang et al., 2019a,b; Yang et al., 2019; Wadden et al., 2019; Tong et al., 2020; Wang et al., 2021; Lu et al., 2021; Liu et al., 2022).

Document-level Event Extraction Dataset. Several datasets offer cross-sentence event argument

annotations. The MUC-4 dataset (McLean, 1992), centered on Latin American terrorism news articles, comprises 1,700 documents that span 4 event types and 5 argument types. RAMS (Ebner et al., 2020) includes 3,993 annotated documents sourced from news, featuring 139 event types and 65 roles. WikiEvents (Li et al., 2021) presents 246 annotated documents across 50 event types and 59 roles, with each document capturing multiple events. ChFinAnn (Zheng et al., 2019) is narrowly defined, featuring 5 event types and 35 argument types within the Chinese financial domain. DuEE-Fin (Han et al., 2022) is an expansive human-annotated Chinese financial dataset, boasting 13 event types. Datasets like Cancer Genetics, EPM, GENIA2011, GENIA2013, Pathway Curation, and MLEE (Pyysalo et al., 2013; Ohta et al., 2011; Van Landeghem et al., 2013) are tailored exclusively for the biological domain. Researchers have made a lot of progress in this field (Zhang et al., 2020; Xu et al., 2021; Huang and Jia, 2021; Ren et al., 2022; Ma et al., 2022; Xu et al., 2022; Du and Cardie, 2020; Liu et al., 2021b; Wei et al., 2021; Li et al., 2021; Ren et al., 2023; Li et al., 2023). In summary, these datasets are often confined to specific domains, possess limited scale, or lack unified event information annotations.

Event Relation Dataset. Adhering to the TimeML specification (Pustejovsky et al., 2003a, 2010), temporal relation datasets such as TimeBank (James Pustejovsky and others, 2006) and TempEval (Verhagen et al., 2009, 2010; UzZaman et al., 2013) have been constructed. Yet, these works often grapple with low annotation concordance and efficiency challenges. Leveraging temporal insights, causal relation datasets (Do et al., 2011; Mirza et al., 2014; Mostafazadeh et al., 2016; Dunietz et al., 2017; Caselli and Vossen, 2017; Tan et al., 2022) have emerged. These resources seldom integrate different relation types within a single dataset. A few datasets (Caselli and Vossen, 2017; Ning et al., 2018; Wang et al., 2022) annotate both temporal and causal relations concurrently.

3. DEIE Construction

Our primary objective is to assemble a unified large-scale dataset to propel research in document-level event information extraction. In the ensuing sections, we will detail the construction of the event schema, outline the data collection process, and describe the crowdsourced labeling process.

3.1. Event Schema Construction

Conventional event frameworks, like FrameNet (Baker, 2014), predominantly focus on routine activities such as sleeping and shopping. To devise an event schema with comprehensive coverage for general-domain events suitable for our dataset, we organize the event types into a hierarchical event type schema. This comprises 12 primary types (e.g. Accident Disaster, Life) further subdivided into 64 secondary types (e.g. Earthquake, Marry). Notably, our schema emphasizes emergency events of significant public interest, including instances like Social Security.Attack, Natural Disaster.Weather, and Public Health.Epidemic. Such events often necessitate descriptions spanning multiple sentences, as they are of substantial societal importance and cannot be adequately conveyed at the sentence level. All event types are listed in Table 1.

Primary Event Type	Secondary Event Type
Accident Disaster	Fire, Transportation, Ecological Pollution, Industry and Mining, Public Facility
Financial Transaction	Stock Decline, Stock Rally, Sell, Rise in Price, Reduce the Price, Acquisition, Financing, Invest, Listing
Public Health	Epidemic, Food Safety
Communicative Action	Gratitude, Threat, Apologize, Meet, Contact
Product Action	Recall, Removal, Manufacture, Launch, Check-out
Movement	Transport, Trip, Deliver
Competition	Competiton, Promote, Suspend
Justice	Arrest, Sue, Investigation, Fine, Jail, Tip-off, Release
Social Security	Attack, Procession, Demonstrate, Hijack, Explosion, Religion, Economy
Life	Die, Be-Born, Marry, Divorce, Enrollment, Price-winning, Ill
Organizational Action	Dismiss, Elect, Layoff, Opening, Closing, Job-cut
Natural Disaster	Flood and Drought, Weather, Earthquake, Geology, Biohazard

Table 1: 12 primary event types and 64 secondary event types of DEIE.

Following the hard/soft news category framework presented in Lehman-Wilzig and Seletzky (2010), we establish a comprehensive collection of 64 distinct event types. Within this schema, we define 19 emergency event types from hard news and 45 common event types from soft news. To construct the schema for each event type, we identified the most commonly occurring non-empty slots within the Baidu[1] infobox of events. These frequently appearing slots provide essential information for describing events of that specific type and are. In total, our established schema consists of 158 unique event argument types, correlated to the 64 event types. This implies that, on average, each event type incorporates around 6 distinct event argument types. Figure 2 exemplifies several of the event argument types we have defined.

[1] https://baike.baidu.com/

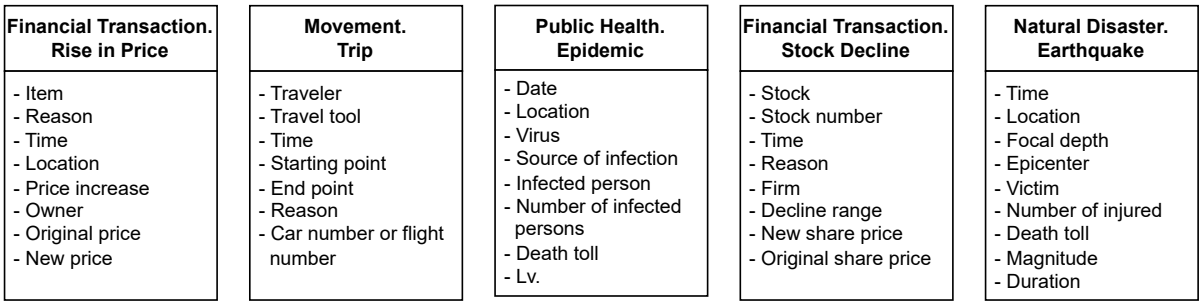

Figure 2: Five examples of event schema in DEIE.

3.2. Data Collection

News serves as the primary source for current events, and as a result, our emphasis lies in identifying events within news articles. Our data collection process involved gathering Chinese web pages and employing a reliable business tool to identify web pages that included mentions of events, with a majority originating from news websites. To gather sufficient events, we provide the keywords (defined by domain experts) of each event type to the annotation company. This approach helped alleviate the long-tail distribution of events in our dataset, meanwhile ensuring diversity.

To compile a timeline of events, we opted to consider news that occurred between 2013 and 2023. To ensure a balanced article length, we filtered out articles shorter than 120 tokens or exceeding 1200 tokens. Then to match the multi-event scenario, we further moved out documents that reported fewer than two events. In the end, we gathered a total of 20,000 news documents containing multiple events from the Internet.

3.3. Crowdsourced Labeling

We cooperate with an annotation company to hire human annotators. To ensure the quality of the dataset, we provide principle guidelines and dedicated training to the annotators. Following an extensive training program spanning three weeks, we handpicked 30 proficient annotators. We decompose the overall annotation task into multiple sequential stages, which reduces competence requirements for annotators. Simultaneously, we enlisted the participation of four experts to conduct two rounds of annotation checking, ensuring the quality and accuracy of the annotations.

Stage 1: Event Classification

This stage aims to extract the event triggers and types in one document. Typically, an event trigger is identified as a verb or noun that serves as an event indicator. We initiated the process by providing 100 expertly annotated documents as references for the annotators. Then we assigned two independent annotators to evaluate each candidate event. If their assessments differed, a third annotator was consulted to make the final judgment.

Stage 2: Event Information Unified Annotation

Given the complexity of event information annotation, we provided annotators with precise definitions and labeling guidelines for event arguments, event summaries, and event-event relations to maintain annotation accuracy and consistency. For each event, annotators are required to extract event arguments from the whole document and determine their argument type, then compose a concise summary of the event, and finally identify event-event relation (causal or temporal). After stage 1, each article will be labeled with the event type. To streamline the second stage, which demands attention to intricate event details, we divided the annotators into six distinct groups, with each group focusing exclusively on one of the six event types. For the event argument with multiple mentions in the document, we will label the mention closest to the trigger. For event summary, we use the shortest natural language to cover all event arguments. We further label the cause event & effect event of the causal relation and the prior event & subsequent event of the temporal relation.

Stage 3: Data Quality

Crowdsourced annotation proceeds in batches and each batch undergoes two rounds of meticulous quality checks. Only after successfully passing these quality assessments are the instances incorporated into the final version of our dataset. It's worth noting that after annotating independently for each task, the annotators collaborate, compare their annotations, and discuss any discrepancies to reach a consensus and finalize the data.

First-round Checking. Once a batch of crowdsourced annotations is finished, it undergoes evaluation by all authors to ensure compliance with our annotation standards. Instances that do not meet the quality criteria are returned for revision,

Dataset	#EventType	#ArgType	#Doc	#Event	#Arg	#Sum	#Rel
MUC-4 (McLean, 1992)	4	5	1,700	1,700	2,641	-	-
WikiEvents (Li et al., 2021)	50	59	246	3,951	5,536	-	-
RAMS (Ebner et al., 2020)	139	65	3,993	9,124	21,237	-	-
ChFinAnn (Zheng et al., 2019)	5	6	32,040	48,000	289,871	-	-
DuEE-Fin (Han et al., 2022)	13	92	11,900	15,850	81,632	-	-
DocEE (Tong et al., 2022)	59	356	27,485	27,485	180,528	-	-
DEIE (Ours)	64	158	20,000	56,031	243,287	56,031	35,285

Table 2: Statistics of Document-level Event Extraction datasets. (ArgType: event argument type, Doc: document, Arg: event argument, Rel: event-event relation, Sum: event summary)

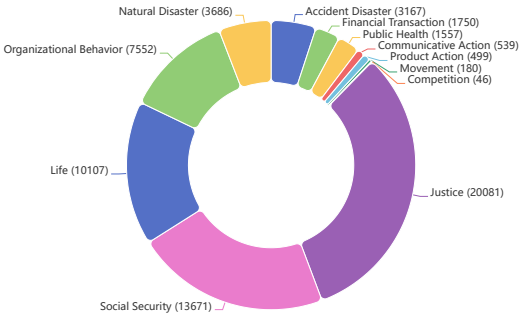

Figure 3: The primary event types (most coarse-grained) and their number of events.

Primary Event Type	Secondary Event Type	Percentage
Justice	Arrest	18.9%
Life	Die	17.5%
Social Security	Economy	11.7%
Justice	Sue	9.6%
Social Security	Attack	5.4%
Natural Disaster	Earthquake	2.9%

Table 3: Top 6 secondary event types in DEIE.

along with detailed reasons for rejection. This iterative process continues until the acceptance rate reaches the threshold of 90%.

Second-round Checking. Following the first-round checking, each batch of annotated instances that meets the criteria is subjected to a dual check by four experts. In this phase, the experts conduct a random evaluation of 30% of the instances and return any unqualified instances to the experts, providing reasons for rejection. Additionally, minor adjustments to the annotation standards may occur during this phase. This iterative process continues until the acceptance rate reaches the level of 95%.

We measure the inter-annotator agreements (IAA) of the event argument/summary/relation annotation between two annotators with Cohen's Kappa (Cohen, 1960). After first-round checking, the results of event argument/summary/relation annotations are 81.2% / 89.7% / 90.3%. After second-round checking, the results of event argument/summary/relation annotations are 89.6% / 93.2% / 96.1%. These results show that although the general domain event annotation is difficult, DEIE's quality is satisfactory.

4. Dataset Statistics and Analysis

4.1. Data Size

In total, DEIE labels **20,000** documents, encompassing **56,031** valid events, **243,287** event arguments, **56,031** event summaries, and **35,285** event-event relations. We show the main statistics of DEIE and compare them with some existing widely-used document-level event datasets in Table 2, including MUC-4, WikiEvents, RAMS, ChFinAnn, DuEE-Fin and DocEE. Compared with existing document-level event datasets, DEIE significantly increases the data scale of all the event-related tasks by providing the most events. Particularly notable is the number of event arguments in DEIE — an astounding two orders of magnitude greater than MUC-4. Furthermore, DEIE introduces large-scale event summaries and event-event relations, absent in competing datasets. When compared to DocEE, DEIE might have fewer documents but boasts a significantly higher number of events (56k versus 27k). This disparity arises from the one-event-per-document annotation of DocEE, while DEIE ensures that each article carries a minimum of two events in the crowdsourcing process. Hopefully, the large-scale unified event dataset can accelerate joint research on document-level event-related tasks.

4.2. Event Distribution

Figure 3 illustrates the distribution of the primary event types in DEIE. We can observe that DEIE also faces the challenge of inherent data imbal-

ance. However, given the large-scale nature of DEIE, **91.7%** and **66.7%** of primary event types have more than 100 and 500 event instances, respectively. Compared with existing datasets like DocEE (only 30.5% event types have more than 500 instances), DEIE significantly mitigates the issue of data scarcity, facilitating the advancement of various event-related applications. Our intention with DEIE is for it to serve as a realistic representation of real-world event data, which naturally follows a long-tail distribution. Hence, we refrained from implementing data augmentation or rebalancing strategies during dataset construction, preserving the real-world distribution in DEIE.

To support future exploration of handling the long-tail challenge, we have devised a hierarchical event schema. This structure aims to facilitate the transfer of knowledge from primary event types to those in the long-tail secondary types. Table 3 presents the top six secondary event types alongside their respective percentages: Justice.Arrest (18.9%), Life.Die (17.5%), Social Security.Economy (11.7%), Justice.Sue (9.6%), Social Security.Attack (5.4%), Natural Disaster.Earthquake (2.9%).

4.3. Event Summary and Relation Statistic

The DEIE dataset annotates **56,031** valid event summaries for the event summary generation task. On average, the summaries have a length of 33 tokens. Each event entry in DEIE has a summary and a content description, which can be seen as naturally annotated data for training a model that generates a summary based on a content description.

Moreover, DEIE also captures inter-event information, marking **35,285** event-event relations, of which 22,264 are causal and 13,021 are temporal. While relation extraction has been extensively studied in NLP, there is a scarcity of unified resources available for event-event relation extraction, making this aspect of DEIE particularly valuable and unique. DEIE plays a pivotal role in addressing this gap and providing a foundation for research in document-level event relation extraction. The inclusion of event-event relations opens up new possibilities, such as the application of knowledge inference techniques.

4.4. Emergency Event Annotation

The structured processing and semantic analysis of large volumes of data on emergency events hold significant importance in enabling their assessment and prediction. For instance, when a natural disaster occurs, relief organizations can consider event analysis as one of the ways to capture different types of information, such as the severity of the incident, the number of victims, and how to provide aid

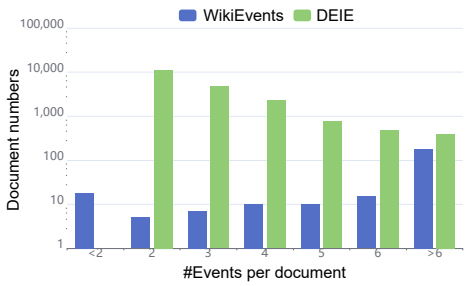

Figure 4: Document numbers containing different numbers of events of two datasets.

	Train	Dev	Test
#Document	15,987	1,996	2,017
#Event	44,670	5,671	5,690
#Argument	193,641	24,821	24,825
#Summary	44,670	5,671	5,690
#Relation	28,101	3,665	3,519

Table 4: The statistics of splitting DEIE.

for them. Despite its value and significance, quick and accurate identification of unspecified emergencies remains an under-explored area. Motivated by this, DEIE annotates **22,081** emergency events, constituting **39.4%** of the total events. This encompasses four primary event types: Social Security (24.4%), Natural Disaster (6.6%), Accident Disaster (5.7%), and Public Health (2.8%), further divided into 19 secondary event types.

4.5. Multi-event Scenario

DEIE closely aligns with real-world multi-event scenarios, and event extraction models can benefit from identifying correlations between these multiple events to classify event types accurately. While the multi-event has been explored in existing document-level event datasets — such as WikiEvents, where 86.88% of documents cover multiple events — it's notable that in DEIE, such scenarios are ubiquitous with **100%** of documents encompassing multiple events. This heightened complexity presents a challenging opportunity for event-related research.

In Figure 4, we compare DEIE's documents containing different numbers of events with WikiEvents. We can observe that due to DEIE's broader scope covering general domain events, its multi-event scenarios are more intricate. There's an exponential surge in the number of documents containing multiple events in DEIE when compared to WikiEvents. Moreover, as more event types are defined in DEIE, the association relations between event types will be much more complex than on WikiEvents. We hope DEIE can facilitate event research, particularly in modeling inter-event correlations.

Method	Arg-I			Arg-C			Head-C		
	P	R	F1	P	R	F1	P	R	F1
BERT-CRF (Shi and Lin, 2019)	50.9	33.2	40.2	46.2	30.0	36.4	50.2	32.7	39.6
EEQA (Du and Cardie, 2020)	54.3	30.7	39.2	53.5	30.2	38.6	58.2	32.9	42.0
BART-GEN (Li et al., 2021)	55.4	44.4	49.3	54.3	43.5	48.3	60.1	48.2	53.5
PAIE (Ma et al., 2022)	59.7	47.9	53.2	58.8	47.2	52.4	65.4	52.4	58.2

Table 5: The overall event argument extraction performance of various models on DEIE.

Method	ROUGE-1	ROUGE-2	ROUGE-L
BART (Lewis et al., 2020)	76.91	64.91	72.33
CPT (Shao et al., 2021)	77.51	65.70	73.20

Table 6: The overall event summary generation performance of various models on DEIE.

Method	Temporal			Casual		
	P	R	F1	P	R	F1
BERT (Ethayarajh, 2019)	55.4	36.5	44.0	79.1	66.2	72.1
MAVEN-ERE (Wang et al., 2022)	47.2	45.6	46.4	79.2	81.8	80.5
Relative-Time (Wen and Ji, 2021)	56.8	35.6	43.8	-	-	-
CF-ECI (Mu and Li, 2023)	-	-	-	77.5	65.7	71.1

Table 7: The overall event relation extraction performance of various models on DEIE.

5. Experiments

In this section, we present the challenges posed by DEIE through a series of comprehensive experiments involving state-of-the-art (SOTA) models. Our experiments encompass document-level event argument extraction, event summary generation, and event relation extraction tasks.

Dataset Division. Table 4 presents the statistics of splitting DEIE. For each event type, we randomly select 80% of the data as the training set, 10% of the data as the validation set, and the remaining 10% of the data as the test set.

5.1. Document-level Event Argument Extraction

Baselines. For strictly consistent comparison, we choose four representative state-of-the-art models: (1) **BERT-CRF** (Shi and Lin, 2019), which uses a BERT-based BIO-styled sequence labeling model; (2) **EEQA** (Du and Cardie, 2020), the first question-answer-based model designed for sentence-level event argument extraction; (3) **BART-GEN** (Li et al., 2021), which formulates the task as a seq2seq task and uses BART to generate corresponding arguments in a predefined format; and (4) **PAIE** (Ma et al., 2022), which defines a prompt tuning paradigm for event argument extraction.

Evaluation Metrics. Our results are reported as F1 score of argument identification (**Arg-I**) and argument classification (**Arg-C**). For the WikiEvents dataset, we follow Li et al. (2021) to additionally evaluate argument head F1 score (**Head-C**).

- **Arg-I**: An event argument is considered correct if its offsets match any of the argument mentions.
- **Arg-C**: An event argument is correctly classified if its offset and role type match the ground truth.
- **Head-C**: only considers the matching of the head-word of an argument.

For the predicted argument, we find the nearest matched string to the golden trigger as the predicted offset. As an event type often includes multiple roles, we use micro-averaged role-level scores as the final metric.

Overall Performance. We present the event argument extraction results in Table 5, indicating that document-level event argument extraction remains a challenging task in DEIE. Several key observations can be distilled from the data. Primarily, the proficiency of the models on DEIE across the three metrics - Arg-I, Arg-C, and Head-C - is notably suboptimal, with the best average performance not surpassing 54.6 F1. This indicates considerable challenges in our dataset for future research. In addition, the current state-of-the-art model, PAIE, surpasses BART-GEN in overall performance, particularly distinguishing itself in the Head-C metric with a substantial 4.7 F1 improvement. The failure of existing baselines may be due to two reasons: (1) Unlike sentence-level tasks, document-level event argument extraction emphasizes the model's proficiency in handling long texts, necessitating that the model comprehensively processes the full text before determining the argument type of a span. (2) Current baselines suffer from inferior capability in semantic understanding, which may lead to failing to distinguish arguments of similar events or mistaking unrelated entities for event arguments. Finally, we find that BART-based methods (i.e., BART-GEN, PAIE) generally outperform BERT-based ones (i.e., BERT-CRF, EEQA) for document-level event argument extraction models. Future research can thus focus on BART to develop better event argument extraction models in DEIE.

5.2. Document-level Event Summary Generation

Baselines. Since document-level event summary generation has not been explored in the past, we choose two neural models as baselines and report their performances on DEIE. (1) **BART$_{base}$** (Lewis et al., 2020), a widely-used sequence-to-sequence model for generation tasks. (2) **CPT** (Shao et al., 2021), a Chinese language understanding and generation model.

Evaluation Metrics. Despite the fact that only single reference summaries are available in benchmark evaluations, we are able to evaluate summary quality using automatic metrics based on lexical similarity (Lin, 2004). We report the full-length F1 score of the ROUGE-1, ROUGE-2, and ROUGE-L metrics with the Porter stemmer option[2].

Overall Performance. Table 6 reports the F1 scores of ROUGE-1, 2, and L for two models. The results show that the event summary annotations within DEIE render it a compelling benchmark for evaluating event summary systems. The experiments discussed in this section were intended to be groundwork for the introduced document-level event summary generation task and we leave developing more tailored methods for future work.

5.3. Document-level Event Relation Extraction

Baselines. We reproduce four representative neural models and report their performances on DEIE, including (1) **BERT$_{base}$** (Ethayarajh, 2019), a widely-used PLM, we adopt it as the backbone and build classification models on top of it. (2)**MAVEN-ERE** (Wang et al., 2022), which provides simple but strong baselines for 4 event relation extraction tasks. (3) **Relative-Time** (Wen and Ji, 2021), which targets to event temporal relation classification. (4) **CF-ECI** (Mu and Li, 2023), which proposes counterfactual reasoning for event causality identification.

Evaluation Metrics. We adopt Precision (P), Recall (R), and F1 score (F1) as evaluation metrics.

Overall Performance. Experimental results for temporal/casual relation extraction are shown in Table 7. We can observe that: (1) The joint model (BERT$_{base}$, MAVEN-ERE) outperforms the relation-specific model (Relative-Time, CF-ECI), demonstrating improvements of 0.5%-5.9% for Temporal F1 and 1.4%-13.2% for Casual F1. The performance indicates that considering the rich interactions between event relations is promising for handling the complex event relation extraction tasks, and demonstrates the benefits of our unified annotation. (2) The achieved performances for the temporal relation extraction task are far from practically usable, which better reflects the inherent challenge of temporal understanding. Specifically, the performances of MAVEN-ERE are higher than Relative-Time (2.6 F1 gap). This is because straightforwardly joint training on relation extraction tasks can bring certain improvements, especially on the tasks with fewer data, i.e., temporal relation extraction. (3) For casual relation extraction, this performance is far from perfect, thus suggesting the challenges for event casual identification and presenting ample research opportunities to improve the performance in the future. (4) DEIE annotates global event pairs within documents, and the lower performance better demonstrates that understanding the diverse and complex event relations is a huge challenge for NLP models and needs more research efforts.

[2]https://tartarus.org/martin/PorterStemmer/

6. Applications

Every event entry in DEIE encompasses both intra-event information (i.e., event trigger, type, argument, summary) and inter-event information (i.e., event-event relation). The comprehensive event information within DEIE has profound implications for the following tasks and applications: (1) **Document-level event extraction.** With such rich event element information, DEIE is extremely meaningful for training an event detection and argument extraction joint model for typical event extraction at the document level. (2) **Event summary generation.** DEIE serves as an invaluable training foundation for creating event summaries from detailed text descriptions or generating the text description of (role, argument) pair of an event. (3) **Event-event relation extraction and inference**. DEIE can plug the gap of unified event-event relation extraction and further make it possible to infer event-event relations. For instance, by having relations like Temporal (A, B) and Temporal (B, C), we can infer Temporal (A, C), which can be invaluable for constructing event taxonomies and enhancing our understanding of event relations. (4) **Event-centric knowledge base construction**. Upon completing the aforementioned steps, the event profile is established, serving as an event entry for an event-centric knowledge base. In other words, the above three steps can be seen as the subtasks of event knowledge base construction.

7. Conclusion

In this paper, we introduce DEIE, a unified large-scale document-level event information extraction dataset designed to advance event-related re-

search spanning from sentence to document level. Compared to existing datasets, DEIE significantly enhances the data scale, featuring over 56,031 events and 243,287 arguments. Moreover, it provides other detailed event information, including event summary and event-event relation. Through our experimental evaluations, we demonstrate that DEIE stands as a challenging and yet untapped benchmark, promising to catalyze further breakthroughs in event-related research.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (NO.2022YFB3102200).

8. Bibliographical References

Collin F. Baker. 2014. FrameNet: A knowledge base for natural language processing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 1–5, Baltimore, MD, USA. Association for Computational Linguistics.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

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, Vancouver, Canada. Association for Computational Linguistics.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. 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.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

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*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Jian-feng Fu, Wei Liu, Zong-tian Liu, and Sha-sha Zhu. 2010. A study of chinese event taggability. In *2010 Second International Conference on Communication Software and Networks*, pages 400–404. IEEE.

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

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. Duee-fin: A large-scale dataset for document-level event extraction. In *Natural Language Processing and Chinese Computing*, pages 172–183, Cham. Springer International Publishing.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Yusheng Huang and Weijia Jia. 2021. Exploring sentence community for document-level event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 340–351, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Heng Ji, Joel Nothman, H Trang Dang, and Sydney Informatics Hub. 2016. Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end cold-start kbp. *Proceedings of TAC*.

Sam N. Lehman-Wilzig and Michal Seletzky. 2010. Hard news, soft news, 'general' news: The necessity and utility of an intermediate classification. *Journalism*, 11:37 – 56.

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

Hao Li, Yanan Cao, Yubing Ren, Fang Fang, Lanxue Zhang, Yingjie Li, and Shi Wang. 2023. Intra-event and inter-event dependency-aware graph network for event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6362–6372, Singapore. Association for Computational Linguistics.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jian Liu, Yubo Chen, and Jun Zhao. 2021a. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3608–3614.

Jian Liu, Yufeng Chen, and Jinan Xu. 2021b. 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. 2022. 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.

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.

Virginia McLean. 1992. Fourth message understanding conference (muc-4).

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, Gothenburg, Sweden. Association for Computational Linguistics.

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

on Events, pages 51–61, San Diego, California. Association for Computational Linguistics.

Feiteng Mu and Wenjie Li. 2023. Enhancing event causality identification with counterfactual reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 967–975, Toronto, Canada. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011. Overview of the epigenetics and post-translational modifications (EPI) task of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 16–25, Portland, Oregon, USA. Association for Computational Linguistics.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, and Marcia Lazo. 2003b. The timebank corpus. *proceedings of corpus linguistics*.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *International Conference on Language Resources and Evaluation*.

Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013. Overview of the cancer genetics (CG) task of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 58–66, Sofia, Bulgaria. Association for Computational Linguistics.

Yubing Ren, Yanan Cao, Fang Fang, Ping Guo, Zheng Lin, Wei Ma, and Yi Liu. 2022. CLIO: Role-interactive multi-event head attention network for document-level event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2504–2514, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. 2023. Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–306, Toronto, Canada. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *ArXiv*, abs/2109.05729.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie. 2020. Improving event detection via open-domain trigger knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5887–5897, Online. Association for Computational Linguistics.

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 the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United

States. Association for Computational Linguistics.

Gaye Tuchman. 1972. Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of Sociology*, 77:660 – 679.

Gaye Tuchman. 2018. *Objectivity as Strategic Ritual: An Examination of Newsmen's Notions of Objectivity 1*, pages 127–146.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

S Van Landeghem, J Björne, C- H Wei, K Hakala, S Pyysalo, S Ananiadou, H -Y Kao, Z Lu, T Salakoski, Y Van de Peer, and F Ginter. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PLoS ONE*, 8(4).

Marc Verhagen, Robert J. Gaizauskas, Frank Schilder, Mark Hepple, Jessica L. Moszkowicz, and James Pustejovsky. 2009. The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43:161–179.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

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.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. Ace 2005 multilingual training corpus-linguistic data consortium. *URL: https://catalog. ldc. upenn. edu/LDC2006T06*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019a. Adversarial training for weakly supervised event detection. 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 998–1008, Minneapolis, Minnesota. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019b. 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.

Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Guo Zhi, and Li Jin. 2021. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4672–4682, Online. Association for Computational Linguistics.

Haoyang Wen and Heng Ji. 2021. Utilizing relative event time to enhance event-event temporal relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10431–10437, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ludwig Wittgenstein. 1994. *Tractatus logico-philosophicus*. Edusp.

Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. 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 3533–3546, Online. Association for Computational Linguistics.

Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2022. A two-stream AMR-enhanced model for document-level event argument extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5025–5036, Seattle, United States. Association for Computational Linguistics.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy. 2020. A two-step approach for implicit event argument detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7479–7485, Online. Association for Computational Linguistics.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event 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 337–346, Hong Kong, China. Association for Computational Linguistics.

Tong Zhu, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, Nicholas Yuan, and Min Zhang. 2022. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4552–4558. International Joint Conferences on Artificial Intelligence Organization. Main Track.

9. Language Resource References

Christopher Walker and others. 2005. *ACE 2005 Multilingual Training Corpus*. Linguistic Data Consortium, ISLRN 458-031-085-383-4.

James Pustejovsky and others. 2006. *TimeBank 1.2*. Linguistic Data Consortium, ISLRN 717-712-373-266-4.

Stephanie Strassel and others. 2014. *TAC KBP English Event Argument - Training and Evaluation Data 2014-2015*. Linguistic Data Consortium, ISLRN 505-557-960-269-8.

Stephanie Strassel and others. 2016. *TAC KBP Event Argument - Comprehensive Training and Evaluation Data 2016-2017*. Linguistic Data Consortium, ISLRN 982-513-576-529-0.