DocEE-zh: A Fine-grained Benchmark for Chinese Document-level Event Extraction

Minghui Liu[1], **Meihan Tong**[1*], [1], **Yangda Peng**[1], **Lei Hou**[1]
Juanzi Li[1], **Bin Xu**[1]
[1]Knowledge Engineering Laboratory, Tsinghua University, Beijing, China
liumh0121@gmail.com, tongmeihan@gmail.com

Abstract

Event extraction aims to identify events and then extract the arguments involved in those events. In recent years, there has been a gradual shift from sentence-level event extraction to document-level event extraction research. Despite the significant success achieved in English domain event extraction research, event extraction in Chinese still remains largely unexplored. However, a major obstacle to promoting Chinese document-level event extraction is the lack of fine-grained, wide domain coverage datasets for model training and evaluation. In this paper, we propose DocEE-zh, a new Chinese document-level event extraction dataset comprising over 36,000 events and more than 210,000 arguments. DocEE-zh is an extension of the DocEE dataset, utilizing the same event schema, and all data has been meticulously annotated by human experts. We highlight two features: focus on high-interest event types and fine-grained argument types. Experimental results indicate that state-of-the-art models still fail to achieve satisfactory performance, with an F1 score of 45.88% on the event argument extraction task, revealing that Chinese document-level event extraction (DocEE) remains an unresolved challenge. DocEE-zh is now available at https://github.com/tongmeihan1995/DocEE.git.

1 Introduction

Event Extraction (EE) aims to detect events from text, encompassing both event classification and event element extraction. EE is an important task of information retrieval in natural language processing (Xiang and Wang, 2019) with a wide range of applications. For instance, it can automatically detect and analyze major events in news reports, providing timely information for decision-makers (Tanev et al., 2008; Piskorski et al., 2007; Atkinson et al., 2013). In conclusion, advancements in event extraction technologies and systems can benefit numerous domains.

Significant progress has been made in event extraction, particularly in the English domain. Notable datasets such as ACE2005[1] have been extensively used for sentence-level event extraction, laying a foundation for numerous research studies. The TAC KBP[2] Event Nugget dataset extends event extraction to a broader context by including event nuggets and their arguments. The Rich ERE (Entities, Relations, Events) (Song et al., 2015) dataset further advances the field by offering a more detailed annotation schema and expanding the scope to document-level extraction. Recently, the DocEE (Tong et al., 2022) dataset has emerged as a comprehensive resource for document-level event extraction, offering wide coverage of event types and fine-grained annotations, greatly contributing to the advancement of this field.

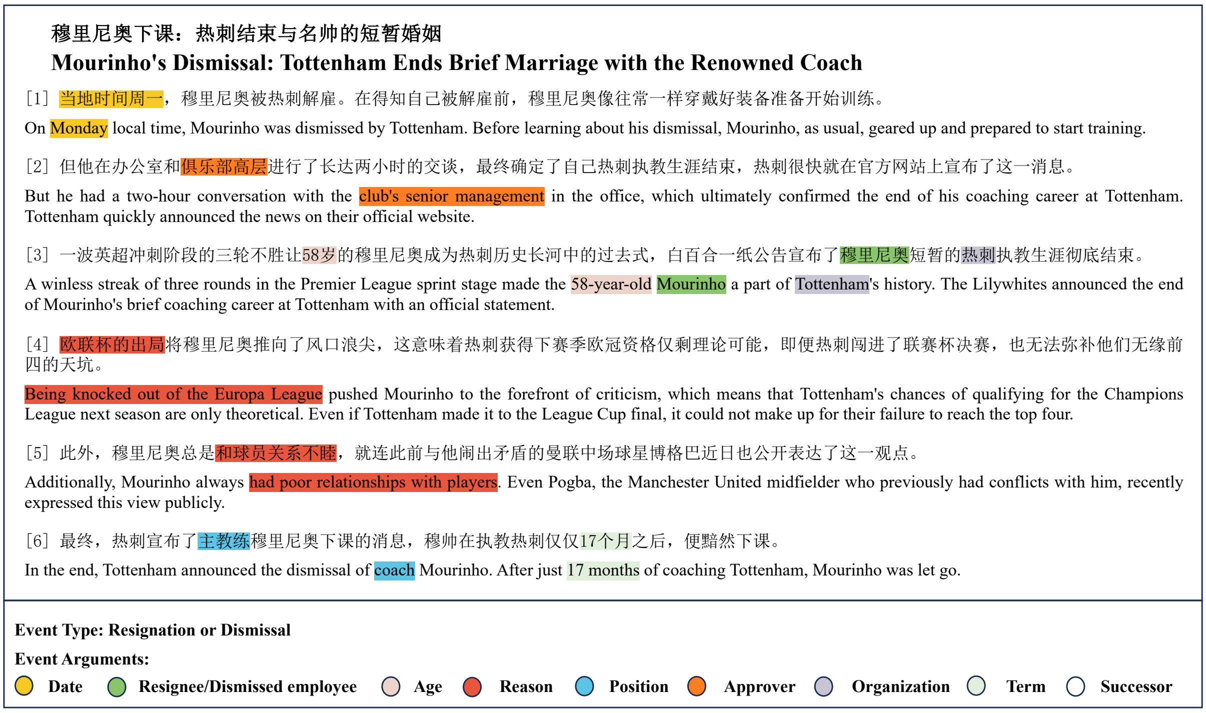
In contrast, Chinese language processing predominantly relies on the Chinese portion of the ACE2005 dataset, which mainly focuses on event extraction at the sentence level. However, events are often spread across entire documents, resulting in event arguments being dispersed across multiple sentences. As depicted in Figure 1, identifying the "Date" argument may require information from sentence [1], while understanding the "Reason" may involve synthesizing data from sentences [4] and [5]. This highlights the need for multi-sentence reasoning and modeling long-range dependencies, which go beyond the scope of sentence-level event extraction. Therefore, advancing event extraction from individual sentences to entire documents is critically necessary.

Currently, there are few Chinese datasets available for document-level event extraction, most of which focus on the financial domain, such as Ch-

[*]Corresponding author.

[1]https://catalog.ldc.upenn.edu/LDC2006T06
[2]https://tac.nist.gov/2017/KBP/

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 637–649
November 12-16, 2024 ©2024 Association for Computational Linguistics



穆里尼奥下课：热刺结束与名帅的短暂婚姻

Mourinho's Dismissal: Tottenham Ends Brief Marriage with the Renowned Coach

[1] 当地时间周一，穆里尼奥被热刺解雇。在得知自己被解雇前，穆里尼奥像往常一样穿戴好装备准备开始训练。

On Monday local time, Mourinho was dismissed by Tottenham. Before learning about his dismissal, Mourinho, as usual, geared up and prepared to start training.

[2] 但他在办公室和俱乐部高层进行了长达两小时的交谈，最终确定了自己热刺执教生涯结束，热刺很快就在官方网站上宣布了这一消息。

But he had a two-hour conversation with the club's senior management in the office, which ultimately confirmed the end of his coaching career at Tottenham. Tottenham quickly announced the news on their official website.

[3] 一波英超冲刺阶段的三轮不胜让58岁的穆里尼奥成为热刺历史长河中的过去式，白百合一纸公告宣布了穆里尼奥短暂的热刺执教生涯彻底结束。

A winless streak of three rounds in the Premier League sprint stage made the 58-year-old Mourinho a part of Tottenham's history. The Lilywhites announced the end of Mourinho's brief coaching career at Tottenham with an official statement.

[4] 欧联杯的出局将穆里尼奥推向了风口浪尖，这意味着热刺获得下赛季欧冠资格仅剩理论可能，即便热刺闯进了联赛杯决赛，也无法弥补他们无缘前四的天坑。

Being knocked out of the Europa League pushed Mourinho to the forefront of criticism, which means that Tottenham's chances of qualifying for the Champions League next season are only theoretical. Even if Tottenham made it to the League Cup final, it could not make up for their failure to reach the top four.

[5] 此外，穆里尼奥总是和球员关系不睦，就连此前与他闹出矛盾的曼联中场球星博格巴近日也公开表达了这一观点。

Additionally, Mourinho always had poor relationships with players. Even Pogba, the Manchester United midfielder who previously had conflicts with him, recently expressed this view publicly.

[6] 最终，热刺宣布了主教练穆里尼奥下课的消息，穆帅在执教热刺仅仅17个月之后，便黯然下课。

In the end, Tottenham announced the dismissal of coach Mourinho. After just 17 months of coaching Tottenham, Mourinho was let go.

Event Type: Resignation or Dismissal

Event Arguments:

🟡 **Date** 🟢 **Resignee/Dismissed employee** ○ **Age** 🔴 **Reason** 🔵 **Position** 🟠 **Approver** 🟣 **Organization** ○ **Term** ○ **Successor**

Figure 1: An example from DocEE-zh. Each document in DocEE-zh is annotated with event type and involved event arguments. In the example, the document mainly describes a *Resignation or Dismissal* event which contains the following arguments: *Date, Age, Reason and Term* and etc. We use different colors to distinguish event arguments.

FinAnn (Zheng et al., 2019) and DuEE-fin (Han et al., 2022). Moreover, a significant portion of the event arguments in these datasets are generic and used across multiple events, with specific arguments tailored to particular event types being relatively scarce. For instance, in ChFinAnn, 60% of the arguments are general, and in DuEE-fin, this figure is 51%. This prevalence of generic arguments limits the ability of models to accurately capture the nuances of specific events, reducing the effectiveness of event extraction systems in identifying and differentiating between unique event types. In summary, existing datasets for Chinese document-level EE fail in the following aspects: limited coverage of domains, and insufficient refinement of argument types.

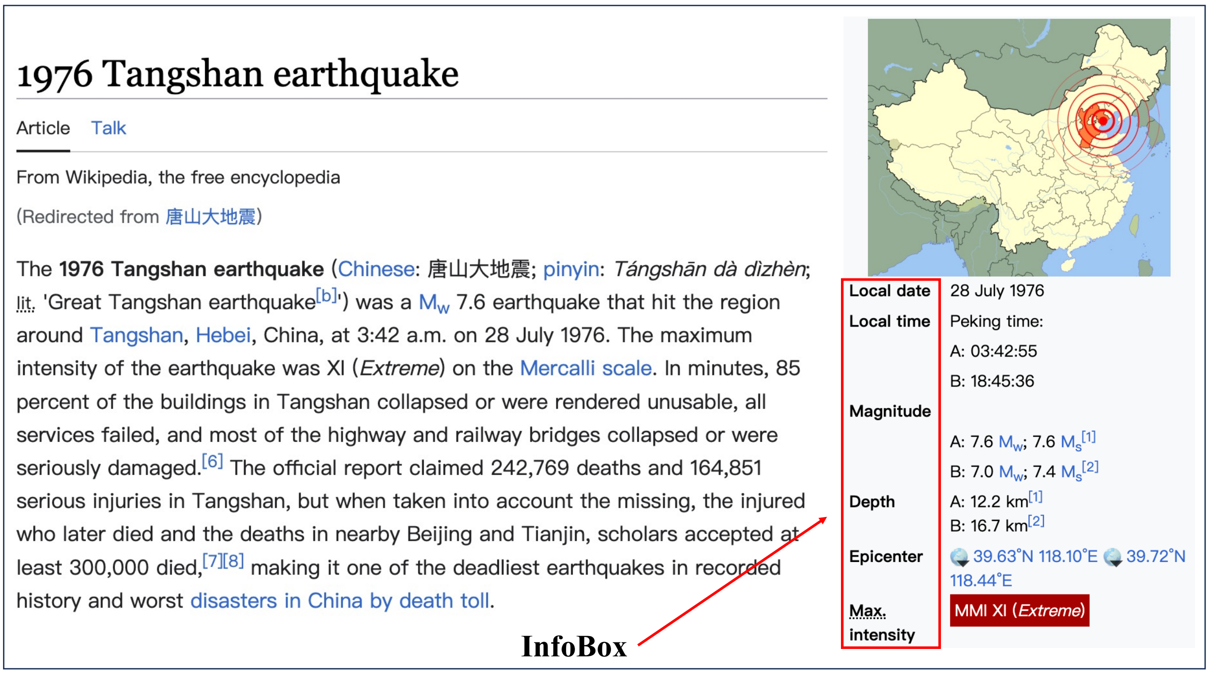
In our paper, we introduce DocEE-zh, a fine-grained Chinese dataset for document-level event extraction. DocEE-zh focuses on the extraction of the main event, following a *one-event-per-document* approach. Figure 1 illustrates an example of DocEE-zh. Our contribution encompasses two key aspects: 1) High-interest event types: DocEE-zh has curated 59 event types derived from various news categories, encompassing domains such as politics, military, entertainment, sports, and others. 2) Fine-grained event argument types: DocEE-zh incorporates a total of 344 argument types, personalized event-specific arguments have been devised for each event type. In DocEE-zh, 86% of the event arguments are specific to individual events.

2 Related Datasets

In recent years, the field of event extraction has made significant advancements, particularly with the development of various datasets tailored for both sentence-level and document-level tasks. These datasets provide valuable resources for research and facilitate the development of more complex models. In this section, we overview the most representative Chinese event extraction datasets and compare them with several important English-language datasets, analyzing their differences and characteristics.

2.1 Sentence-level Event Extraction Dataset

The Automatic Content Extraction (ACE2005-zh) dataset contains 633 documents covering 8 event types and 33 subtypes, serving as a foundational resource for sentence-level event extraction in Chinese. LEVEN (Yao et al., 2022) is a Chinese legal event detection dataset containing 108 event types, providing a comprehensive resource for legal text



1976 Tangshan earthquake

Article Talk

From Wikipedia, the free encyclopedia

(Redirected from 唐山大地震)

The **1976 Tangshan earthquake** (Chinese: 唐山大地震; pinyin: *Tángshān dà dìzhèn*; lit. 'Great Tangshan earthquake[b]') was a M_w 7.6 earthquake that hit the region around Tangshan, Hebei, China, at 3:42 a.m. on 28 July 1976. The maximum intensity of the earthquake was XI (*Extreme*) on the Mercalli scale. In minutes, 85 percent of the buildings in Tangshan collapsed or were rendered unusable, all services failed, and most of the highway and railway bridges collapsed or were seriously damaged.[6] The official report claimed 242,769 deaths and 164,851 serious injuries in Tangshan, but when taken into account the missing, the injured who later died and the deaths in nearby Beijing and Tianjin, scholars accepted at least 300,000 died,[7][8] making it one of the deadliest earthquakes in recorded history and worst disasters in China by death toll.

Local date	28 July 1976
Local time	Peking time: A: 03:42:55 B: 18:45:36
Magnitude	A: 7.6 M_w; 7.6 M_s[1] B: 7.0 M_w; 7.4 M_s[2]
Depth	A: 12.2 km[1] B: 16.7 km[2]
Epicenter	39.63°N 118.10°E 39.72°N 118.44°E
Max. intensity	MMI XI (*Extreme*)

Figure 2: An example of a Wikipedia event page. The infobox in the page is one of the sources for determining event argument types.

analysis. The Chinese Emergency Corpus (CEC)[3] focuses on sudden events, comprising 5 categories and 332 articles, which are essential for studying emergency response. The DuEE dataset (Li et al., 2020) consists of 19,640 events divided into 65 event types and 121 argument roles, offering a rich resource for detailed analysis.

In addition to Chinese datasets, English datasets also play a crucial role in the field of sentence-level event extraction. GENEVA (Parekh et al., 2022) is a large-scale benchmarking dataset designed to evaluate the generalizability of Event Argument Extraction (EAE) models, featuring 115 event types and 220 argument roles. MAVEN (Wang et al., 2020) and MAVEN-Argument (Wang et al., 2023a) represent significant English-language datasets covering 168 event types, with MAVEN-Argument further expanding the argument identification task.

2.2 Document-Level Event Extraction Datasets

Document-level event extraction tasks have gradually become a research hotspot in both Chinese and English contexts, primarily applied in fields such as finance and news. Our dataset, DocEE-zh, has unique characteristics in its annotation strategy compared to existing datasets.

DEIE (Ren et al., 2024) is a unified large-scale document-level event information extraction dataset containing over 56,000 events and 242,000 arguments. The DEIE dataset includes 64 event types and 158 unique argument types derived from 20,000 Chinese news articles. In comparison, our DocEE-zh dataset offers 59 event types and a more extensive set of 344 argument types, annotated from 36,729 Chinese news articles. One of the key strengths of DocEE-zh lies in the fine-grained nature of its argument annotations, with 86% of the argument types being specific to individual events, such as the magnitude parameter in earthquake events. This level of detail supports a more precise understanding of specific events, enhancing the overall accuracy and applicability of event extraction tasks.

Moreover, DocEE-zh adopts a strategy of annotating only one core event per document. This choice is rooted in practical considerations, particularly when dealing with the vast amount of news generated daily. By focusing on annotating significant core events, we avoid being overwhelmed by less informative content, ensuring that we provide the most critical information. This focus enhances the dataset's practical utility, particularly for users seeking concise and relevant event information, while also facilitating efficient event fusion and inference.

In comparison, ChFinAnn (Zheng et al., 2019)

[3] https://github.com/shijiebei2009/CEC-Corpus

and DuEE-fin (Han et al., 2022) are datasets focused on document-level event extraction in the financial domain. ChFinAnn includes 5 event types, such as Company Earnings, Company Financing, and Company Risks, annotated with 35 event arguments. DuEE-fin, on the other hand, covers 13 event types and includes 92 event arguments. While these financial datasets encompass multiple event types and arguments, their focus is relatively narrow, primarily serving financial event analysis, whereas our DocEE-zh dataset offers a broader range of event types and richer arguments to accommodate more complex news scenarios.

3 Constructing DocEE-zh

Our main goal is to construct a fine-grained Chinese dataset to promote the development of event extraction from the sentence level to the document level. In the following sections, we will first introduce how to build event schema, and then discuss how to collect candidate data and label them through crowdsourcing.

3.1 Event Schema Construction

Referring to the construction method of event schema in DocEE (Tong et al., 2022), we have defined 59 event types based on the theory of hard/soft news, comprising 31 hard news event types and 28 soft news event types. Hard news typically includes topics that are timely, important, and serious, such as politics, economics, and disasters. In contrast, soft news covers more human-interest stories and entertainment, such as lifestyle, culture, and personal achievements. The complete list of event types is provided in the Appendix Table 5. This schema encompasses influential events of significant public concern, including but not limited to earthquakes, floods, and diplomatic summits, which cannot be effectively captured at the sentence level and require multi-sentence descriptions. This classification not only covers the primary event types found in news reporting but also accurately reflects the diversity and complexity of news content. Consequently, it allows the model to adapt to a broader range of information extraction scenarios, facilitating users in accessing the event information they seek with greater ease.

Defining event types is just the first step; assigning specific arguments to each type is crucial for constructing an effective event ontology. This involves identifying key characteristics such as date, location, and participants.

We began by using Wikipedia infoboxes to identify initial event arguments. As shown in the figure 2, Wikipedia pages often include structured information in infoboxes, with key-value pairs like "Magnitude," "Date," "Depth," and "Max Intensity." We collected details from 20 Wikipedia event pages per event type and used automated parsing to create a preliminary list of arguments.

Since Wikipedia may not cover all important arguments, we supplemented this with information from authoritative news sources. We analyzed 20 reports per event type from sources like Xinhua News, and invited five journalism students to identify additional arguments. These students suggested critical details, such as "Tsunami Height" for tsunami events, which might not be listed in Wikipedia but are important for understanding the event's impact.

Finally, we consolidated and deduplicated the arguments to ensure accuracy and conciseness. This process resulted in 344 event arguments for the 59 event types, averaging 5.8 arguments per type. These arguments cover basic information as well as specific details like scale, impact, and causes, providing a comprehensive event description. Figure 3 illustrates examples of three event arguments in DocEE-zh.

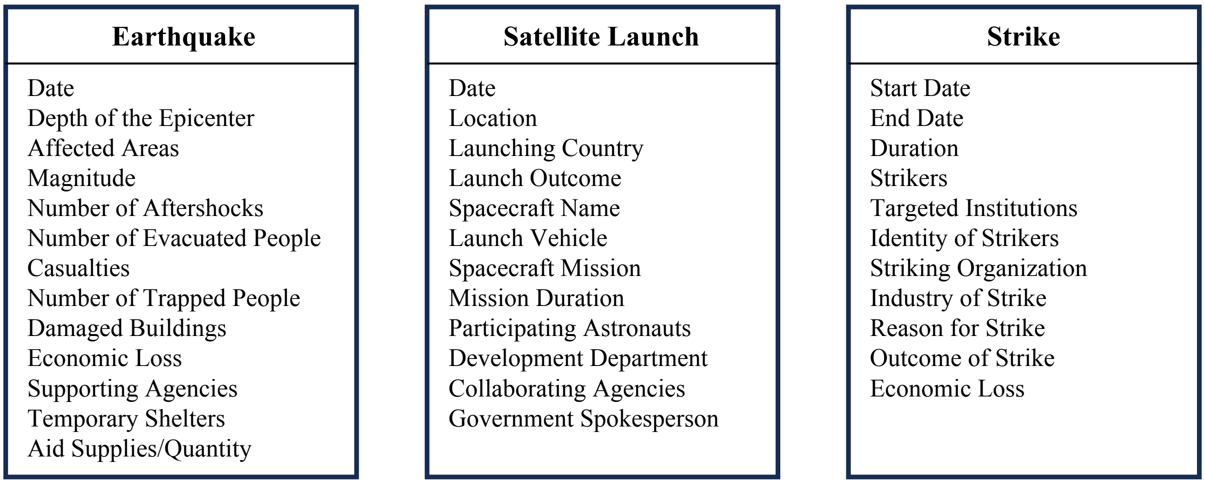
3.2 Candidate Data Collection

In this study, to construct a high-quality Chinese document-level event extraction dataset, we primarily collected data for annotation from two sources: Chinese Wikipedia and the NewsMiner system (Hou et al., 2015).

Specifically, for the Chinese Wikipedia part, we focused on historical events with Chinese entries, such as the "Tangshan Earthquake" shown in Figure 2. These historical events usually have detailed descriptions on Wikipedia, including core information like event time, location, and impact, providing us with a rich corpus of resources.

On the other hand, we selected news reports from the NewsMiner system, spanning from 2019 to 2023, which were published by six major news websites: Tencent News, People's Daily Online, Xinhua News Agency, Sina News, Global Times, and Sohu News. These reports cover a wide range of societal dynamics, including politics, economics, and culture, significantly enhancing the diversity and timeliness of the dataset.

During the screening process, we adopted a high-



Earthquake

Date
Depth of the Epicenter
Affected Areas
Magnitude
Number of Aftershocks
Number of Evacuated People
Casualties
Number of Trapped People
Damaged Buildings
Economic Loss
Supporting Agencies
Temporary Shelters
Aid Supplies/Quantity

Satellite Launch

Date
Location
Launching Country
Launch Outcome
Spacecraft Name
Launch Vehicle
Spacecraft Mission
Mission Duration
Participating Astronauts
Development Department
Collaborating Agencies
Government Spokesperson

Strike

Start Date
End Date
Duration
Strikers
Targeted Institutions
Identity of Strikers
Striking Organization
Industry of Strike
Reason for Strike
Outcome of Strike
Economic Loss

Figure 3: Examples of event arguments in DocEE-zh for three event types: Earthquake, Satellite Launch, and Strike. Each category lists specific arguments to capture comprehensive details about the events, ranging from basic information like date and location to more detailed aspects such as economic loss and participating agencies.

Datasets	#isDocEvent	#EvTyp.	#ArgTyp.	#Doc.	#Tok.	#Sent.	#ArgInst.
ACE2005	✗	33	35	599	290k	15,789	9,590
KBP2017	✗	18	20	167	86k	4,839	10,929
ChFinAnn	✔	5	35	32,040	29,207k	629,338	289,871
DuEE-fin	✔	13	92	7,173	32,959k	684,700	56,806
DEIE	✔	64	158	20,000	11,731k	262,088	243,287
DocEE-zh(ours)	✔	59	344	36,729	36,012k	817,085	216,496

Table 1: Statistics of EE datasets (isDocEvent: whether the event in the corpus at the document-level, EvTyp.: event type, ArgTyp.: event argument type, Doc.: document, Sent.: sentence, ArgInst.: event arguments)

frequency keyword retrieval strategy based on category names and the TF-IDF (Sparck Jones, 1972) algorithm. This method significantly improved the specificity and efficiency of the screening, enabling us to precisely identify reports related to events. Through this series of meticulously designed strategies, we collected approximately 60,000 Chinese articles in total, laying a solid foundation for constructing a comprehensive event extraction dataset.

3.3 Crowdsourced Annotation

The crowdsourced annotation process comprises two stages: event classification and event argument extraction.

3.3.1 Event Classification

In the event classification stage, the focus is on precisely categorizing the core events within news reports. Core events are those prominently highlighted in the news titles and primarily discussed throughout the article. This process aims to identify and annotate the key news events most likely to attract user attention, ensuring that the annotated dataset is directly relevant to the interests of news consumers.

The annotation process is designed to ensure accuracy and consistency, implemented through the following steps:

Pre-annotation Phase In this initial phase, 100 selected news articles are pre-annotated to establish a high-quality annotation standard. This step helps train and calibrate annotators' understanding and application of event classification, providing a reference benchmark for subsequent annotation work.

Annotator Selection Based on the pre-annotation results, annotators with an accuracy rate below 70% are eliminated. This selection mechanism ensures that those participating in the final annotation work possess sufficient quality and capability, thereby enhancing the overall accuracy and reliability of the dataset.

Dual Annotation and Review Mechanism The remaining 48 annotators annotate each news article in pairs. When the classification results of the two annotators differ, a review mechanism is initiated, involving a third annotator to adjudicate and determine the final event classification for that news article. This mechanism effectively reduces the impact of subjective judgment differences, improving the consistency and accuracy of the annotation results.

"Other" Category For news that does not fit into any predefined categories, they are classified as "Other." This approach provides a flexible classification option, ensuring that all news events are appropriately annotated without forcing them into unsuitable categories, maintaining the overall quality and consistency of the dataset.

Through this annotation process, we have effectively achieved precise and consistent classification of core events in the news.

3.3.2 Event Argument Extraction

In the event argument extraction stage, we gathered 90 annotators to accurately extract key event information from complete news articles. To ensure the successful execution of this task, we adopted a strategy combining initial annotation and multiple iterative revisions. Initially, all articles underwent a round of basic annotation. Based on these initial results, common issues were summarized, and a detailed annotation guide was developed, followed by targeted training for the annotators.

Subsequently, the project entered the iterative revision stage, where each article was reviewed in three rounds, with each round handled by different annotators to ensure that each article was reviewed by at least three annotators. After each round, the identified issues were fed back to the annotation team to make corresponding adjustments in the subsequent annotations.

Through this continuous iterative revision process, the annotation accuracy significantly improved from an initial 56.24% to 76.83%, eventually reaching 85.96%. This improvement process demonstrates the effectiveness of the adopted methods in enhancing annotation quality.

During the event argument annotation, to ensure the completeness of the work, if an event argument is mentioned multiple times in the document, all mentions are recorded. For example, as shown in Figure 1, the "Reason" event argument is mentioned through "exit from the Europa League" and "tumultuous rapport with several squad members", both of which are included in the annotation task.

3.3.3 Annotation Quality Analysis

Following the studies of Artstein and Poesio (2008) and McHugh (2012), we used Cohen's kappa coefficient to measure the inter-annotator agreement (IAA) for assessing annotation data consistency. In the event classification stage, the kappa value reached 93%, while in the event argument extraction stage, it was 82%. These high kappa values indicate significant consistency among annotators, ensuring the high reliability of the entire dataset. Additionally, the annotation cost was controlled within 2 RMB per data entry.

4 Data Analysis of DocEE-zh

In this section, we conduct a comprehensive analysis of DocEE-zh to provide a deep understanding of the dataset and the task of document-level event extraction.

Overall Statistics DocEE-zh contains annotations for 36,729 document-level events and 216,496 event arguments, averaging 5.9 arguments per document. Notably, the event type *Awards ceremony* exhibits the highest average number of event arguments per document at 11.6, while the *Financial Crisis* event type shows the lowest at 3.3.

The average document length in our dataset is 1005 characters, encompassing an average of 23.36 sentences per document. This highlights the substantial amount of information available for analysis. Each document is dedicated to a single event, focusing on the core event described in the news. Given the high volume of daily news, our goal is to extract the most impactful events of interest to users. This approach facilitates more focused and efficient subsequent tasks such as event fusion and event reasoning.

Table 1 presents a comparison of DocEE-zh with several representative event extraction datasets, including the sentence-level ACE2005 and KBP2017 datasets, as well as the Chinese document-level ChFinAnn, DuEE-fin, and DEIE datasets. As shown in Table 1, DocEE-zh demonstrates significant strengths across various dimensions. Although DEIE contains a higher number of event types (64), DocEE-zh still offers a highly detailed set of annotations, covering 59 event types and the largest

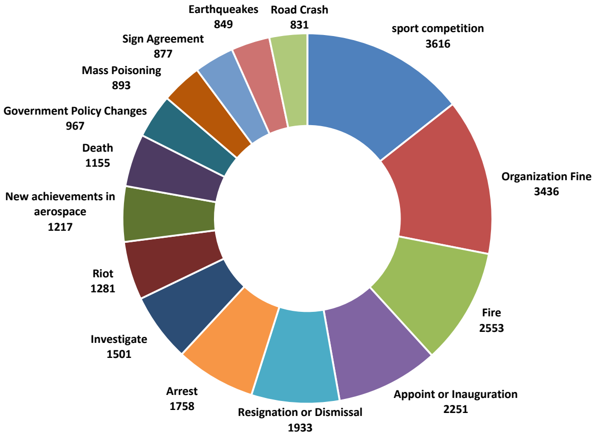


Figure 4: Top 15 event types in DocEE-zh.

number of event argument types (344), highlighting the dataset's diversity and fine-grained annotation.

Additionally, DocEE-zh contains the largest number of documents (36,729) and tokens (36,012k), reflecting its richness in textual resources. The dataset's extensive sentence count (817,085) and event arguments (216,496) further underscore its comprehensive coverage and the fine granularity of its event argument annotations, especially across such a large corpus. Compared to ChFinAnn and DuEE-fin, DocEE-zh not only includes a broader variety of event types, but also features more detailed event argument annotations, capturing nuanced semantic information and supporting more complex event extraction tasks. While DEIE surpasses DocEE-zh in the total number of event arguments, DocEE-zh still excels in the precision and depth of its annotations, making it particularly suitable for handling more complex and diverse task scenarios.

Event Type Statistics Figure 4 illustrates the distribution of the top 15 most common event types in DocEE-zh, representing the highest frequency of occurrences. These event types include categories such as *sports competitions* (9.8%), *organization fines* (9.4%), *fires* (6.9%), *appointments/inaugurations* (6.1%), and *resignations/dismissals* (5.3%), among others. Our annotated data exhibits a long-tail distribution typical of real-world datasets, where class distributions are often uneven. Notably, event types with over 500 instances constitute 36.2%, while those with over 200 instances represent 79.3%. **Further details can be found in the Appendix.**

Event Arguments Statistics We initially analyzed the event argument types in DocEE-zh, finding that 86% of arguments are specific to particular events, highlighting the fine-grained nature of our annotations. From a random sample of 1000 DocEE-zh documents, we examined 4072 event arguments. Frequency analysis revealed that 84.6% of arguments are mentioned only once, posing a challenge for model recall. Arguments were further categorized by mention length: 76.9% are under 10 characters (mainly named entities), 16.5% are under 20 characters, and 6.6% exceed 20 characters, often involving complex information such as accident causes or investigation results.

Overall, we identified 344 unique event argument types, of which 49 are shared across multiple events, accounting for only 14.2%. This low percentage of shared arguments underscores the fine-grained and diverse nature of our dataset. Events typically span an average of 7.1 sentences, presenting a significant challenge for models to extract information accurately across multiple sentences.

5 Experiments on DocEE-zh

In this section, we elucidate the challenges posed by DocEE-zh through comprehensive experimentation employing state-of-the-art models. We commence by delineating the experimental setup, followed by conducting experiments on event classification and event argument extraction tasks. Finally, we discuss the implications of our findings and suggest potential directions for future development in Chinese document-level event extraction.

Experiment Settings We partitioned the data into training (80%), validation (10%), and test (10%) sets. For transformer-based methods, we utilized the base version of pretrained models with a learning rate of 2e-5, batch size of 32, and maximum document length of 512. Additionally, experiments with GPT-4 adopted a zero-shot learning approach, involving randomly sampling 10 samples for each event type, totaling 590 events, to form a separate test set. Appendix Table 6 demonstrates the zero-shot experimental methodology of GPT-4.

In evaluating GPT-4's performance on event argument extraction, we observed that as a generative model, GPT-4 occasionally produced arguments that did not exactly match the given descriptions but were semantically correct. Initially, we employed exact matching to assess performance, but this method failed to fully capture the model's capabilities. As a result, we incorporated manual evaluation for a more comprehensive assessment. However, because manual evaluation was based on semantic correctness, this broader standard may have led to seemingly inflated performance results for GPT-4.

5.1 Event Classification

Task Definition Assign a predefined event type label to a document. The output is a single event type label.

Baselines We employ various baseline methods: 1) **TextCNN** (Kim, 2014) utilizes CNN kernel sizes for text classification. 2) **BERT** (Devlin et al., 2019) utilizes unsupervised objectives like Masked Language Model and Next Sentence Prediction. 3) **RoBERTa** (Liu et al., 2019) extends BERT with larger training batches and learning rates. 4) **ERNIE 3.0** (Sun et al., 2021) is pretrained on a 4TB corpus, focusing on language understanding. 5) **GPT-4** (OpenAI, 2023) is a multimodal model processing both image and text inputs. Evaluation metrics include Precision, Recall, and Macro-F1 score following (Kowsari et al., 2019).

Method	Precision	Recall	F1
TextCNN	88.15	82.32	83.40
BERT	89.60	87.21	87.78
RoBERTa	91.75	87.88	89.16
ERNIE 3.0	91.88	87.68	88.71
GPT-4	67.19	71.07	66.39

Table 2: Overall Performance on Event Classification.

Overall Performance Table 2 shows experimental results for event classification, highlighting: 1) Transformer-based models (BERT, RoBERTa, ERNIE 3.0) outperform TextCNN, benefiting from pretraining on large-scale unlabeled corpora and possessing extensive background semantic knowledge. 2) GPT-4 scores lower than supervised models, possibly due to the presence of many similar event types in the data, demanding strong identification of primary event features, posing a challenge for GPT-4 without specialized fine-tuning.

5.2 Event Argument Extraction

Task Definition Given a document with an identified primary event and its relevant argument types, extract the event arguments such as date, location, and participants. The output is a set of extracted arguments.

Baselines We introduce the following mainstream baselines for evaluation: 1) **BERT_Seq** (one of the baselines in Du and Cardie (2020a)) utilizes the pre-trained BERT model to sequentially label words in the article. 2) **MG-Reader** (Du and Cardie, 2020a) proposes a novel multi-fine-grained reader to dynamically aggregate information at the sentence and paragraph levels. 3) **BERT_QA** (Du and Cardie, 2020b) queries the article for answers using the argument type as a question. 4) **Doc2EDAG** (Zheng et al., 2019) generates an entity-based directed acyclic graph for document-level event extraction. 5) **PTPCG** (Zhu et al., 2021) proposes a pseudo-trigger-aware pruned complete graph approach for efficient document-level event extraction. 6) **ProcNet** (Wang et al., 2023b) utilizes procedural generation techniques to dynamically create event extraction templates by capturing global event information. 7) **ReDEE** (Liang et al., 2022) introduces a customized transformer for capturing multi-scale, multi-quantity parameter relationships. 8) **PAIE** (Ma et al., 2022), a generation-based model, employs prompt-based learning to enhance argument extraction by leveraging pre-trained language models. 9) **GPT-4** (OpenAI, 2023), a large language model, excels in contextual understanding and reasoning capabilities.

Method	Precision	Recall	F1
BERT_Seq	42.32	41.76	42.04
MG-Reader	40.43	46.36	43.19
BERT_QA	41.46	48.47	44.69
Doc2EDAG	49.45	31.06	38.15
PTPCG	46.49	35.93	40.53
ProcNet	53.64	40.08	45.88
ReDEE	53.23	34.38	41.78
PAIE	48.33	39.17	43.27
GPT-4	21.15	30.21	24.88
GPT-4 (Human)	58.54	83.60	68.86

Table 3: Overall performance on Event Argument Extraction. (ME denotes manual evaluation; other methods utilize exact matching for evaluation.)

Overall Performance The experimental results in Table 3 demonstrate significant performance differences among various models on the DocEE-zh dataset. Traditional Transformer-based models, such as BERT_Seq, MG-Reader, and BERT_QA, show relatively stable performance, but their capabilities are limited when it comes to modeling complex events. Although these models achieve F1 scores between 42% and 45%, they encounter difficulties in handling multi-sentence reasoning and long-distance dependencies. For example, MG-Reader and BERT_QA exhibit improved recall rates (46.36% and 48.47%, respectively), but their precision scores are relatively low, indicating that these models struggle to effectively differentiate core events from irrelevant information.

More advanced event extraction models, such as Doc2EDAG, PTPCG, ProcNet, and ReDEE, integrate additional contextual information and structured reasoning, resulting in slightly better overall performance than traditional BERT-based models. Notably, ProcNet achieves an F1 score of 45.88%, suggesting its potential in document-level event extraction tasks. However, these models still face significant challenges. While Doc2EDAG demonstrates high precision (49.45%), its recall rate is only 31.06%, indicating limitations in capturing complex event types.

The generative model PAIE also shows competitive performance in this task, with an F1 score of 43.27%, reflecting its ability to understand and generate event-related information to a certain extent. Nevertheless, generative models still face issues such as blurred event boundaries and heightened sensitivity to noise when compared to traditional matching-based models.

GPT-4's performance is particularly noteworthy. Under automatic evaluation, GPT-4 achieves an F1 score of only 24.88%, significantly lower than most models. However, in manual evaluation (ME), GPT-4's performance improves dramatically, reaching an F1 score of 68.86%. This indicates that large generative models, when relying on more complex reasoning and contextual understanding, can exhibit outstanding performance, although current automatic evaluation metrics may not fully capture their true capabilities. This also highlights the limitations of existing evaluation methods in addressing diverse and complex tasks.

Overall, the performance of these models on the DocEE-zh dataset remains suboptimal, primarily due to the dataset's diverse and complex event types, which require models to possess detailed processing capabilities. The primary challenges for these models include multi-sentence reasoning, modeling long-distance dependencies, and excluding irrelevant information to identify core events in the documents. These limitations represent key obstacles preventing models from fully capturing the complexity of the DocEE-zh dataset.

6 Conclusion

In this paper, we propose DocEE-zh, a document-level event extraction dataset, to foster the development of Chinese document-level event extraction. DocEE-zh contains over 36,000+ events and 210,000+ arguments, and includes more fine-grained event arguments. Experiments demonstrate that Chinese document-level event extraction remains an open problem.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62277033). It is also supported by the project from Tsinghua-SPD Bank Joint-Lab. Thanks the support from National Engineering Laboratory for Cyberlearning and Intelligent Technology, and Beijing Key Lab of Networked Multimedia.

Limitations

Our dataset design ensures that each document contains only a single event, aiming to highlight the core events reported in news articles. Given the large volume of daily news, this approach focuses on extracting the most impactful events, improving the efficiency of tasks like event fusion and reasoning. However, this design may have limitations in cases where multiple events occur in a single document. If the goal is to extract all events, our approach may not fully capture the complexity of such documents, potentially limiting its broader applicability. Future work could explore incorporating a fine-grained event schema into multi-event annotations to enhance model flexibility.

With the rise of large language models, interest in leveraging them for event extraction tasks has grown. However, current extractive annotation methods may affect the evaluation of these models. For instance, in Table 4, although the news article does not explicitly state the event date, the model correctly infers it from the surrounding context. Our extractive method, however, does not capture such inferences, potentially hindering the accurate evaluation of model performance. Future research should develop evaluation methods that better account for inferential reasoning, advancing the field of event extraction.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Martin Atkinson, Mian Du, Jakub Piskorski, Hristo Tanev, Roman Yangarber, and Vanni Zavarella. 2013. Techniques for multilingual security-related event extraction from online news. *Computational Linguistics: Applications*, pages 163–186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Xinya Du and Claire Cardie. 2020a. Document-level event role filler extraction using multi-granularity contextualized encoding. In *Annual Meeting of the Association for Computational Linguistics*.

Xinya Du and Claire Cardie. 2020b. Event extraction by answering (almost) natural questions. In *Conference on Empirical Methods in Natural Language Processing*.

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 *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 172–183. Springer.

Lei Hou, Juanzi Li, Zhichun Wang, Jie Tang, Peng Zhang, Ruibing Yang, and Qian Zheng. 2015. Newsminer: Multifaceted news analysis for event search. *Knowledge-Based Systems*, 76:17–29.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing*.

Kamran Kowsari, K. Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. 2019. Text classification algorithms: A survey. *Inf.*, 10:150.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020. Duee: a large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, pages 534–545. Springer.

Yuan Liang, Zhuoxuan Jiang, Di Yin, and Bo Ren. 2022. RAAT: Relation-augmented attention transformer for relation modeling in 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 4985–4997, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? paie: Prompting argument interaction for event argument extraction. *arXiv preprint arXiv:2202.12109*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

OpenAI. 2023. Gpt-4 technical report.

Tanmay Parekh, I Hsu, Kuan-Hao Huang, Kai-Wei Chang, Nanyun Peng, et al. 2022. Geneva: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. *arXiv preprint arXiv:2205.12505*.

Jakub Piskorski, Hristo Tanev, and Pinar Oezden Wennerberg. 2007. Extracting violent events from on-line news for ontology population. In *Business Information Systems: 10th International Conference, BIS 2007, Poznan, Poland, April 25-27, 2007. Proceedings 10*, pages 287–300. Springer.

Yubing Ren, Yanan Cao, Hao Li, Yingjie Li, Zixuan ZM Ma, Fang Fang, Ping Guo, and Wei Ma. 2024. Deie: Benchmarking document-level event information extraction with a large-scale chinese news dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4592–4604.

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.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Ouyang Xuan, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *ArXiv*, abs/2107.02137.

Hristo Tanev, Jakub Piskorski, and Martin Atkinson. 2008. Real-time news event extraction for global crisis monitoring. pages 207–218.

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.

Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, et al. 2023a. Maven-arg: Completing the puzzle of all-in-one event understanding dataset with event argument annotation. *arXiv preprint arXiv:2311.09105*.

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. *arXiv preprint arXiv:2004.13590*.

Xinyu Wang, Lin Gui, and Yulan He. 2023b. Document-level multi-event extraction with event proxy nodes and hausdorff distance minimization. *arXiv preprint arXiv:2305.18926*.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale Chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 183–201, Dublin, Ireland. 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 Jing Yuan, and Min Zhang. 2021. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph. *arXiv preprint arXiv:2112.06013*.

Inferred Hidden Information
News: According to Overseas Network on April 25, citing the UK "Mirror" on the 24th, a recent protest march in London against the lockdown turned violent. Several police officers were injured during clashes with the protesters, with head injuries and bleeding. ... Three people were arrested for allegedly assaulting the police. **Event Type:** Protest or Online Condemnation **Event Argument:** Date **Annotated Answer:** 24th **GPT-4 Answer:** April 24

Table 4: An example of event extraction by GPT-4, where the LLM correctly infers the event date based on the mentioned date in the text, providing complete event argument information.

Event Type	Event Subtype
Economic Event	Organization merger, economic assistance, organization establishment economic crisis, organization penalty, organization bankruptcy
Diplomatic Event	Joining organization, signing agreement, diplomatic visit withdrawing from organization, tearing up agreement diplomatic negotiations
Political Event	Government policy change, taking office, election, resignation
Natural Disaster	Earthquake, fire, snowstorm, tsunami, famine, drought Flood, pest disaster, volcanic eruption, mudslide
Human-induced Disaster	Bank robbery, air crash, vehicle accident, mass poisoning gas explosion, Train collision, shipwreck, mine collapse
Violent Conflict Event	Military exercise, protest activity, strike, political turmoil armed conflict, riot
Public Health Event	Disease outbreak, environmental pollution
Science and Technology Event	Record-breaking, archaeological discovery, solar eclipse lunar eclipse, satellite launch
Public Figure Event	Death event, lawsuit event, recovery event, marriage event investigation event, Divorce event, speech event, sentencing event trial event, illness event, release event
Sports and Entertainment Event	Award ceremony, sports competition

Table 5: Event type of DocEE-zh

Task	Prompt
Event classification	Known event type list: [’type1’, ’type2’, ’type3’, ...] Given the text: "XXXXX..." Q: What is the core type of event in this text?
Event Argument Extraction	Given the text: "XXXXXX..." This text primarily describes the "XXX" event, and the corresponding list of argument roles for the "XXX" event includes: [’Arg1’, ’Arg2’, ’Arg3’,...]. Based on the provided argument roles, please extract the event arguments and output them in JSON format.

Table 6: Prompt for GPT-4 on Event Extraction