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

Anonymous ACL submission

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 (F1 score of 68%), revealing that Chinese DocEE remains an unresolved challenge.

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 ChFinAnn (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 ar-

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

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.

guments 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 seen significant advancements, particularly with the development of various datasets tailored for both sentence-level and document-level tasks. These datasets have been crucial in driving forward research and enabling the development of more sophisticated models. In this section, we provide an overview of the most prominent Chinese event extraction datasets, highlighting their characteristics and contributions to the field.

2.1 Sentence-level Event Extraction Dataset

Automatic Content Extraction (ACE2005-zh) consists of 633 documents, covering 8 event types and 33 subtypes. This dataset has been a foundational resource for sentence-level event extraction in Chinese, enabling the development of various high-performance models. LEVEN (Yao et al., 2022) is a Chinese legal event detection dataset containing 108 event types, providing a comprehensive resource for legal text analysis. Chinese Emergency Corpus (CEC)[3] focuses on sudden events in Chinese, comprising 5 categories and 332 articles, which are essential for studying emergency

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

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

response and management. DuEE (Li et al., 2020) consists of 19,640 events, divided into 65 event types and 121 argument roles, offering a wide range of event types and roles for detailed analysis. Based on these datasets, various superior models have been proposed to enhance sentence-level sentiment expression, achieving significant success (Orr et al., 2018; Tong et al., 2020; Lu et al., 2021).

2.2 Document-level Event Extraction Dataset

Most Chinese document-level event extraction tasks primarily focus on the financial domain, exemplified by ChFinAnn and DuEE-fin.

ChFinAnn (Zheng et al., 2019) is a Chinese Financial Announcement dataset designed for document-level event extraction. It comprises 5 event types: Company Earnings, Company Financing, Company Changes, Company Investments, and Company Risks, annotated with 35 event arguments such as Company Name, Date, Amount, and Stakeholder. Constructed using distant supervision, ChFinAnn provides a substantial amount of training data, but the reliance on automated annotation techniques may introduce noisier annotations.

DuEE-fin (Han et al., 2022), curated by Baidu, is designed for document-level financial event extraction and includes 13 event types such as IPO, Mergers and Acquisitions, and Financial Reports. It is annotated with 92 event arguments, including Company Name, Date, Amount, and Legal Entity. DuEE-fin is characterized by its high-quality manual annotations. However, the dataset includes a substantial number of general arguments (47 out of 92), which can limit the model's ability to extract fine-grained arguments specific to individual events.

In summary, while ChFinAnn and DuEE-fin provide valuable resources for financial event extraction, their limitations lie in their narrow domain focus and coarse-grained arguments. This highlights the need for datasets with broader domain coverage and more fine-grained arguments to better reflect real-world 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

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.

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

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
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)

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-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.

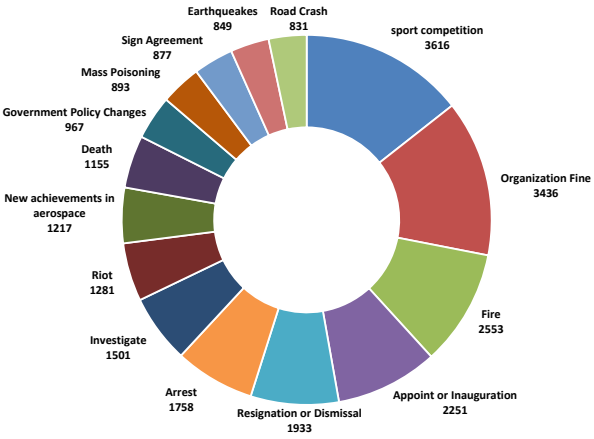

Figure 4: Top 15 event types in DocEE-zh.

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 and DuEE-fin datasets. As shown in Table 1, DocEE-zh outperforms other datasets in several aspects. It includes the highest number of event types (59) and event argument types (344), offering more detailed and diverse event annotations. The dataset also contains the largest number of documents (36,729) and tokens (36,012k), indicating a rich source of textual information. The number of sentences (817,085) and event arguments (216,496) further highlight the extensive coverage and granularity of our dataset. Compared to ChFinAnn and DuEE-fin, which are focused on the financial domain, DocEE-zh provides a broader domain coverage, making it more versatile for various event extraction tasks.

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 in-

clude 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.

The results from GPT-4, being a generative model, mean that the generated event arguments might not always match the descriptions in the text exactly, yet they can be semantically correct. We attempted to validate GPT-4's performance using exact matching, but this did not fully reflect the model's capabilities. Therefore, given the limitations of exact matching, we resorted to manual evaluation to more accurately and objectively assess the capabilities of the GPT-4 model. However, because the manual evaluation was based on semantic correctness, this somewhat broad standard might have led to seemingly inflated 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., 2023) 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.

Overall Performance As shown in Table 3, a variety of models, including traditional transformer-based ones like BERT_Seq, MG-Reader, and BERT_QA, as well as advanced models such as Doc2EDAG, PTPCG, ProcNet, and ReDEE, and generative models like PAIE, demonstrate diverse levels of effectiveness on the DocEE-zh dataset. However, the overall performance of these models still falls short of expectations. This shortfall is largely due to the complex and diverse event types in DocEE-zh, which demand nuanced processing capabilities. Two primary challenges impede model performance: catastrophic forgetting, where models lose previously learned information upon acquiring new data, and a lack of deep semantic understanding necessary for accurately parsing and classifying intricate event arguments. These limitations are critical barriers that prevent the models from fully capturing the complexities of the DocEE-zh dataset.

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	58.54	83.60	68.86

Table 3: Overall Performance on Event Argument Extraction.

GPT-4 significantly outperforms all other models with an F1 score of 68.86%. This remarkable performance is attributed to its exceptional contextual understanding and reasoning capabilities. However, it's important to note that the high F1 score, resulting from manual evaluations emphasizing semantic correctness, may inflate the results.

In conclusion, while GPT-4 demonstrates relatively superior performance on DocEE-zh, the experimental results indicate that Chinese document-level event extraction remains an unresolved challenge. Future research should focus on refining evaluation techniques, enhancing the accuracy of large language models in extracting exact matches, and improving model robustness to address the diversity and complexity of event arguments in datasets like DocEE-zh.

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.

Limitations

Our dataset design focuses on having each document contain only a single event, with the goal of highlighting the core events reported in news articles. Given the high volume of daily news, our objective is to extract the most impactful events that are of significant interest to users. This approach facilitates more focused and efficient subsequent tasks, such as event fusion and event reasoning. However, this design choice has certain limitations. It may not fully align with scenarios where multiple events occur within a single document. If the goal is to extract all events from a document, our current approach may not capture the full complexity and richness of such documents. Consequently, this could limit the dataset's applicability for training models that need to extract multiple events from a single document. In summary, while our single-event-per-document design enhances the clarity and precision of event annotations and supports efficient event-focused tasks, it may also introduce limitations in handling documents with multiple events. Future work could explore incorporating our fine-grained event schema into multi-event annotation tasks to address these limitations and further improve the versatility and robustness of event extraction models.

With the emergence of large language models, there is growing interest in leveraging them for event extraction tasks. However, extractive annotation methods may impact the evaluation of these models. As illustrated in Table 4, the news article does not explicitly state the event date, yet the large language model can correctly infer the event date from other contextual dates. Although this inference is accurate, our extractive annotation method fails to capture answers that are non-contiguous or require inferential reasoning from the original text. Consequently, this limitation may hinder the objective assessment of large language models' performance. Future research should address this limitation by developing more robust evaluation methods that account for inferential capabilities, thereby 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.

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. *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.

Walker Orr, Prasad Tadepalli, and Xiaoli Fern. 2018. Event detection with neural networks: A rigorous empirical evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 999–1004, Brussels, Belgium. Association for Computational Linguistics.

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.

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, 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.

Xinyu Wang, Lin Gui, and Yulan He. 2023. 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