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. 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 the 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.

In the realm of Chinese language processing, prevailing datasets like ACE2005 mainly focus 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 1for instance, 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, there is a critical necessity to advance event extraction from individual sentences to entire documents.

However, 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, the event arguments in these datasets are mostly generic (ChFinAnn with 60%, DuEE-fin with 51%). Overall, existing datasets suffer from limited domain coverage and insufficient granularity in event argument types. Therefore, there is an urgent need to construct a dataset with fine-grained event argument types and wide domain coverage to accelerate research in Chinese document-level event extraction.

In our paper, we introduce DocEE-zh, a fine-grained Chinese dataset for document-level event extraction. 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

2.1 Sentence-level Event Extraction Dataset

Automatic Content Extraction (ACE2005-zh) consists of 633 documents, covering 8 event types and 33 subtypes. LEVEN (Yao et al., 2022) is

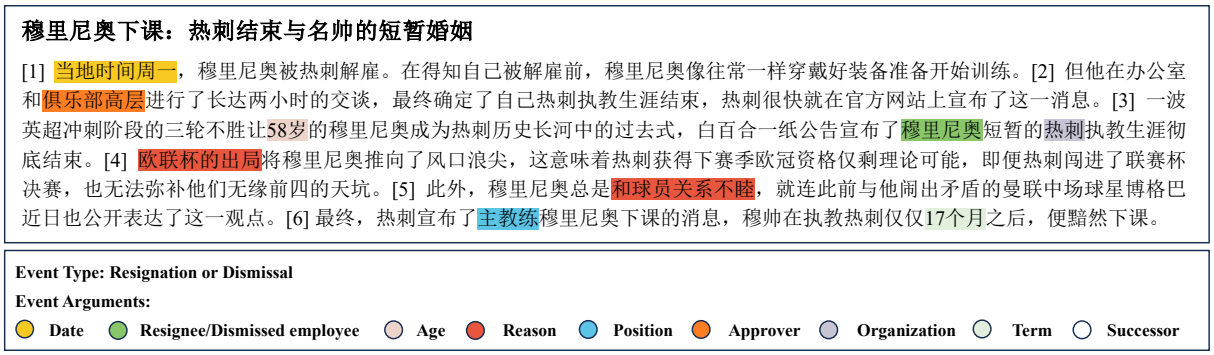

Figure 1: An example from DocEE-zh. Each document in DocEE-zh is annotated with event type and involved event arguments. In the example, the document mainly describes a *Resignation or Dismissal* event which contains the following arguments: *Date, Age, Reason and Term* and etc. We use different colors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

a Chinese legal event detection dataset containing 108 event types. Chinese Emergency Corpus (CEC) focuses on sudden events in Chinese, comprising 5 categories and 332 articles. DuEE (Li et al., 2020) consists of 19,640 events, divided into 65 event types and 121 argument roles. 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 constructed using distant supervision. ChFinAnn comprises 5 event types and 35 event arguments. DuEE-fin includes 13 event types and 92 event arguments. However, the majority of their event arguments are generic, with 21 out of 35 in ChFinAnn and 47 out of 92 in DuEE-fin, which deviates from real-world scenarios. In summary, existing datasets are confined to limited domains and lack refined event argument schema.

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. Detailed information is provided in Appendix. This schema covers influential events of human concern such as earthquakes, floods, diplomatic summits, etc., which cannot be extracted at the sentence level and require multi-sentence descriptions. The setting of event arguments is also consistent with DocEE. The 59 event types collectively define 344 event argument types, averaging 5.8 arguments per event.

3.2 Candidate Data Collection

We utilized Wikipedia as our annotation source, which comprises two event types: historical events and timeline events (Hienert and Luciano, 2012). Historical events are those with dedicated Wikipedia pages, while timeline events are chronologically organized news events. We chose these event types to ensure a balanced data distribution; relying solely on historical events could result in uneven representation, whereas timeline events provide complementary data. To maintain article length consistency, we excluded articles with fewer than 5 sentences and truncated excessively lengthy articles (over 50 sentences). Ultimately, we curated 44,000 candidate events from Wikipedia.

3.3 Crowdsourced Annotation

The crowdsourced annotation process comprises two stages.

Stage 1: Event Classification Annotators classify candidate events into predefined event types. This is a single-label classification task focusing on primary events, primarily depicted in the title and article body. Each candidate event $e =< t, a >$,

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)

where t denotes the title and a denotes the article, is assigned a label y from the 59 defined event types.

Stage 2: Event Argument Extraction Annotators extract event arguments from the entire article for each candidate event $e =< t, a >$. This involves identifying all arguments related to the event type y from the article a.

Annotation Quality Cohen's kappa coefficient is used to assess inter-annotator agreement (IAA), following (Artstein and Poesio, 2008; Hsi, 2022). The IAA scores for event classification and event argument extraction are 93% and 82%, respectively, indicating substantial agreement.

4 Data Analysis of DocEE-zh

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

Overall Statistic DocEE-zh has annotated 36,729 document-level events and 216,496 event arguments, with an average of 5.9 event arguments annotated per document. Among these, the event *Awards ceremony* has the highest average number of event arguments per document (11.6), while the event *Financial Crisis* has the lowest average number of event arguments per document (3.3). We compare DocEE-zh to various representative event extraction datasets in Table 1, including sentence-level EE datasets ACE2005, KBP2017 and Chinese document-level EE dataset ChFinAnn, DuEE-fin.

Event Type Statistic Figure 2 depicts the distribution of the top 15 most common event types in DocEE-zh, representing the highest number of occurrences. These event types encompass various categories such as *sports competitions* (9.8%), *organization fines* (9.4%), *fires* (6.9%), *appointments/inaugurations* (6.1%), *resignations/dismissals* (5.3%), among others. Our annotated data exhibits a long-tail distribution typical of real-world data, where class distributions are often uneven. Notably, classes with over 500 instances constitute 36.2%, while those with over 200 instances represent 79.3%. For further details, please refer to the Appendix.

Event Arguments Statistic We initially analyzed event argument types in DocEE-zh, finding 86% arguments specific to particular events, highlighting the fine-grained of our design. Then, from 1000 randomly selected DocEE-zh documents, we examined 4072 event arguments. Their mention frequency analysis revealed 84.6% arguments mentioned only once, challenging the recall capability of models. Arguments were further categorized by mention length, with 76.9% under 10 characters, mainly named entities. 16.5% had fewer than 20 characters, and 6.6% exceeded 20 words, often involving accident causes or investigation results.

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. Subsequently, we delve into potential future research directions for 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 involved randomly sampling 10

samples for each event type, totaling 590 events, to form the test set.

5.1 Event Classification

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

Baselines We introduce the following mainstream baselines for evaluation: 1) **BERT_Seq** (one of the baseline 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 EE. 5) **ReDEE** (Liang et al., 2022) introduces a customized transformer for capturing multi-scale, multi-quantity parameter relationships. 6) **GPT-4** is a large language model with great contextual understanding and reasoning capabilities. We conducted experiments using a zero-shot learning approach.

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
ReDEE	53.23	34.38	41.78
GPT-4	58.54	83.60	68.86

Table 3: Overall Performance on Event Argument Extraction.

Overall Performance As shown in Table 3, supervised baseline models perform poorly, potentially due to catastrophic forgetting in neural networks. Document-level event extraction underscores the model's capability to handle lengthy texts, necessitating a comprehensive understanding of the entire document before determining argument spans.

GPT-4 demonstrates outstanding performance owing to its robust contextual comprehension and reasoning abilities. It can leverage common-sense knowledge to infer answers and even rectify errors in news reports. This is a problem that our extractive annotation approach cannot correct (as the original information is erroneous). However, GPT-4 tends to make more predictions, necessitating precision improvement while maintaining recall.

We believe that utilizing large language model for document-level EE is a promising research direction. Additionally, our future work may focus on elevating event extraction to higher levels, such as cross-document-level analysis.

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.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: 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.

Daniel Hienert and Francesco Luciano. 2012. Extraction of historical events from wikipedia. In *KNOW@LOD*.

Andrew Hsi. 2022. Event Extraction for Document-Level Structured Summarization.

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.

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.

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.

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.

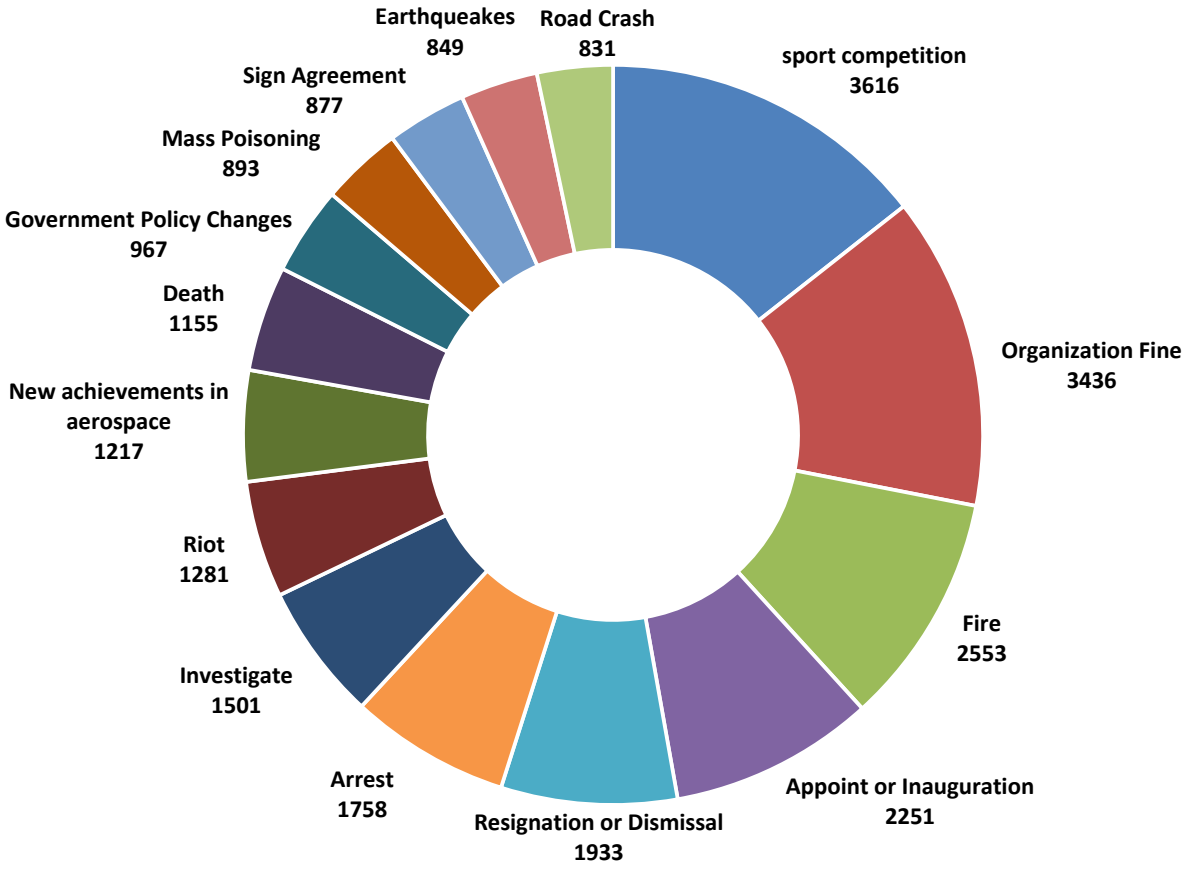

Figure 2: Top 15 event types in DocEE-zh.