The Paradigm Shift in Event Extraction: An In-Depth Overview with Large Language Models

Dung-Cam Quang[1,2][0000−0002−9324−4199], Vinh Q. Vo[1,2][0009−0003−0014−6317], Xuan-Bach Le[1,2][0009−0003−6848−5403], and Tho Quan[1,2(✉)][0000−0003−0467−6254]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
{quang.dungcse,vinh.vo1205,lexuanbach,qttho}@hcmut.edu.vn

Abstract. The advent of Large Language Models (LLMs) marks a significant breakthrough that has substantially advanced performance across a wide range of Natural Language Processing (NLP) tasks. Event Extraction (EE) is a notable beneficiary of this paradigm shift. As a long-standing challenge in Information Extraction, EE has historically been addressed through diverse methodologies, from early syntactic and semantic rule-based systems to modern Machine Learning and Deep Learning models. Recently, LLMs have emerged as a prominent and effective approach, primarily leveraged through two distinct strategies to boost EE performance: data augmentation and task-specific fine-tuning, capitalizing on their vast pre-trained knowledge. While prior surveys have explored generative or prompt-based models for EE, the rapid evolution of LLMs necessitates a more comprehensive and up-to-date overview. In this survey, we provide a detailed analysis of these LLM-based approaches. Furthermore, unlike previous works, we place a unique emphasis on the analysis of EE datasets, which has been insufficiently covered in prior literature. We detail available datasets across multiple languages, including English, Chinese, and Vietnamese. Finally, we conclude with critical insights, highlight key challenges, and outline promising future research directions for the field.

Keywords: Event extraction · Large language models · Event extraction datasets.

1 Introduction

Event Extraction (EE) identifies event occurrences to support Question Answering and Knowledge Base construction [25]. An example sentence from a news article is provided as follows: “Trước đó, ngày 03/12/2019, Vingroup đã kí thỏa thuận nguyên tắc về việc **sát nhập** công ty *VinCommerce* và công ty con *VinEco* vào *Masan Consumer Holdings*.” (Translation: “Previously, on December 3, 2019, Vingroup signed a principle agreement regarding the merger of VinCommerce and its subsidiary VinEco into Masan Consumer Holdings.”).

For a "Merger" event triggered by "sát nhập" (merger), EE is typically decomposed into four sub-tasks: **(1) Trigger Identification (TI)** to locate the trigger (*sát nhập*); **(2) Type Classification (TC)** to categorize the event (*Merger and Acquisition*); **(3) Argument Identification (AI)** to find entities (*Masan Consumer Holdings, VinCommerce, VinEco*); and **(4) Argument Classification (AC)** to assign roles (*Acquirer* or *Target*).

Previous surveys have analyzed EE from various perspectives, ranging from Machine Learning/Deep Learning methods [10] and neural taxonomies [31] to generative approaches [23]. While broader studies like LLM4IE [32] and LLM-for-DA [3] provide high-level panoramas of generative AI, they lack specific taxonomies for EE's structural complexity. To our knowledge, no existing survey specifically focuses on the recent advancements of LLMs in EE. Our work addresses this gap by reviewing diverse datasets and LLM-based methodologies to identify open challenges and future research directions. Our primary contributions are:

- We present a comprehensive overview of EE approaches that incorporate LLMs. These methods are categorized into two main perspectives: (i) using LLMs for data augmentation, and (ii) leveraging LLMs through prompting or fine-tuning to directly address the EE task.
- We aggregate and present EE-related datasets spanning multiple languages and granularity levels (e.g., sentence-level, document-level). A key focus of our survey is domain-specific datasets, an area that has received limited attention in previous surveys despite its importance for real-world applications.
- By integrating insights from both data resources and methodological developments, we provide a critical analysis of LLM-based approaches, highlight key challenges, and outline promising future research directions for advancing the EE field.

2 Event Extraction Datasets

EE resources are categorized by (i) annotation scope (sentence vs. document-level), (ii) domain (general vs. specific), and (iii) linguistic coverage. While early research focused on sentence-level datasets, since 2020, a shift toward document-level resources has emerged to address cross-sentence argument dispersion and event coreference. Additionally, beyond broad general-domain datasets, domain-specific resources have gained importance in specialized fields like finance, pharmacovigilance, and cybersecurity.

Fig. 1 shows the total of 15 datasets in our survey. These categorizes datasets by analysis level and scope. **Sentence-level** and **document-level** analysis are shown via color blocks. Solid *red* borders denote **general-purpose** datasets, while solid *blue* borders indicate **domain-specific** ones. Dashed lines signify language: *orange* for **Vietnamese**, *pink* for **Chinese**, and others are **English**. In this study, we focus on the aspect of domain applicability, and based on this criterion, we introduce the datasets further categorized by their annotation scope and linguistic coverage.

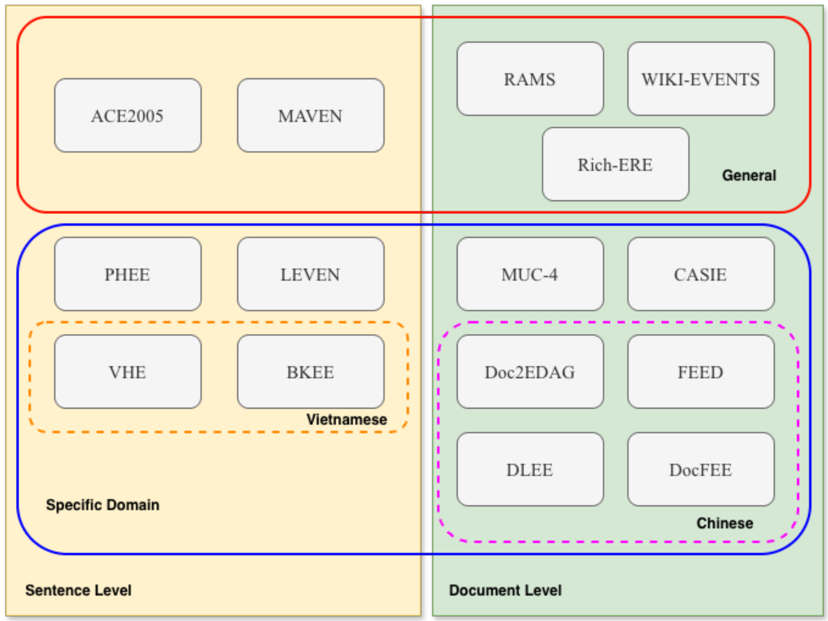

Fig. 1. Overview of Datasets Included in Our Survey

2.1 General Events Datasets

While **ACE 2005** [4] remains the *de facto* sentence-level benchmark despite its limited scale, the field has evolved toward large-scale, distantly supervised corpora like **MAVEN** [29]. To address discourse-level complexities, document-level datasets such as **Rich-ERE** [24], **RAMS** [6], and **WikiEvents** [12] have been introduced to benchmark cross-sentence argument retrieval and coreference chains. In the low-resource landscape, **BKEE** [20] adapts these schemas for Vietnamese, marking a critical step for non-English event extraction (see Table 1).

2.2 Specific-Domain Events Datasets

Domain-specific benchmarks have evolved from pioneering templates like **MUC-4** [7] to modern corpora such as **PHEE** [27] (medical) and **CASIE** [21] (cybersecurity). Notably, Chinese resources like **LEVEN** [36] and **Doc2EDAG** [38] now dominate in scale and document-level complexity. Conversely, Vietnamese resources remain nascent; aside from the historical **VHE** [8], there is a critical lack of modern domain-specific data (see Table 2).

2.3 Summary

Although ACE 2005, Rich-ERE, and RAMS remain cornerstone benchmarks, the field has decisively moved toward large-scale, document-level, and multilingual

Table 1. Comparison of General-Domain EE Datasets

Dataset	Year	Lang	Lvl	Domain	#Typ	#Evt	Key Features / Limitations
ACE 2005* [4]	2005	EN	Sent	News	33	5,349	Golden benchmark. Small scale & imbalanced classes.
Rich-ERE* [24]	2015	EN	Doc	News/Forums	38	8,696	Introduced Event Hoppers. Rich entity mention levels (NAM/NOM/PRO).
MAVEN [29]	2020	EN	Sent	Wiki	168	118,000	Distant Supervision from FrameNet. Massive scale but noisy labels, long-tail.
RAMS [6]	2020	EN	Doc	News	139	9,124	Focus on cross-sentence args. Uses OntoNotes entities.
WikiEvents [12]	2021	EN	Doc	Wiki	50	3,241	Complete event coreference chains (KAIROS ontology). Cross-sentence, small scale.
BKEE [20]	2024	VI	Sent	News	33	9,000	First general Vietnamese EE dataset. Adapted from ACE schema.
*Note: * Licensed by LDC*							

Table 2. Comparison of Specific-Domain EE Datasets

Dataset	Year	Lang	Lvl	Domain	#Typ	#Evt	Key Features / Limitations
English Resources							
MUC-4* [7]	1996	EN	Doc	Terrorism	5	1,700	Pioneering template-filling dataset. Small scale/outdated.
CASIE [21]	2020	EN	Doc	Cyber	5	3,903	High quality ($\kappa = 0.81$). Specialized roles (Attacker, Victim, CVE).
PHEE [27]	2022	EN	Sent	Medical	2	5,019	Pharmacovigilance focus. Nested/sub-events, subject attributes.
Chinese Resources							
LEVEN [36]	2022	ZH	Sent	Legal	108	151,000	Largest legal EE dataset. Hierarchical schema, long-tail, high quality.
DLEE [30]	2024	ZH	Doc	Legal	13	10,000	First document-level legal EE in Chinese. Semi-automated annotation, more than 10k judgment docs.
Doc2EDAG [38]	2019	ZH	Doc	Finance	5	47,824	Distant supervision on ChFinAnn. Addresses multi-event & argument scattering.
FEED [11]	2021	ZH	Doc	Finance	5	46,960	Distant supervision from financial announcements.
DocFEE [2]	2025	ZH	Doc	Finance	9	>35,000	Human-AI collaborative, longest docs (avg 2,277 chars), 19,044 announcements (2020-2024), high accuracy.
Vietnamese Resources							
VHE [8]	2024	VI	Sent	History	35	5,213	Ancient Vietnamese historical texts (Hong Bang dynasty). Limited modern applicability.
*Note: * Licensed by LDC*							

paradigms. Chinese domain-specific datasets now dominate in scale (LEVEN: 150,977 events, Doc2EDAG/FEED/DocFEE: >35,000 events each), while English general-domain document-level resources still hover below 10,000 events. Vietnamese event extraction is particularly under-resourced, with only BKEE and VHE combined offering approximately 14,000 events—both sentence-level and lacking document-level annotations. This substantial gap presents a clear opportunity for future work in low-resource and cross-lingual event extraction.

3 LLM-Based Approaches

Numerous distinct approaches have been proposed for the EE task over time. Many previous survey works have covered these methodologies in detail. In this study, we will only provide a quick overview of these approaches, as illustrated in Fig. 2.

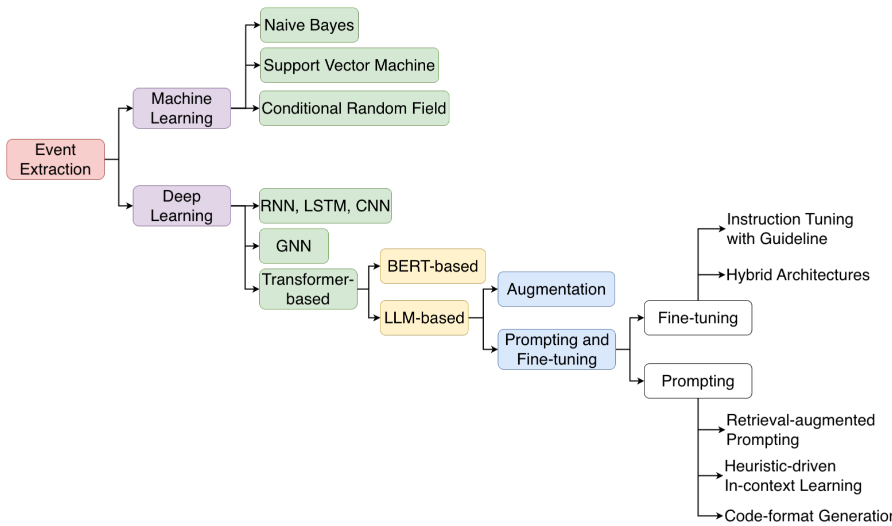

Fig. 2. Overview of Event Extraction Approaches. This figure illustrates the various methodologies proposed for the EE task, with a specific focus on the LLM-based approaches explored in our survey.

Our main focus will be to delve into the **LLM-based approach**, categorized primarily into two classes: leveraging LLMs for **data augmentation** and utilizing LLMs for **model fine-tuning** for the EE task. Finally, we will compile the relevant experimental data to facilitate a comparison between LLM-based approaches and several established baseline models for this task.

3.1 LLMs for Data Augmentation

Existing LLM-based data augmentation methods generally fall into three paradigms: diversity-oriented generation, large-scale dataset construction, and task-aware strategies for long-tail distributions.

Diversity and Scale. To address data scarcity, Meng et al. [17] proposed CEAN, which synthesizes diverse training samples via schema-based composition and rephrasing. Crucially, it incorporates a *contrastive learning objective* to enforce semantic consistency, mitigating the noise often introduced by generative models. Scaling this approach, Liu et al. [15] developed a collaborative framework where multiple LLMs perform fine-grained annotation and filtering. By employing consensus-based voting, this method effectively converts noisy distant supervision into high-quality training data for massive event schemas.

Task-Aware and Long-Tail Strategies. Addressing the long-tail distribution of argument roles, FineCSDA [1] introduces a dual-level augmentation strategy. It generates synthetic samples for *rare event types* (sample-level) and infills missing arguments within specifically constructed contexts for *rare roles* (context-level), regulated by an iterative pruning mechanism to ensure data quality. Alternatively, IDEAE [28] frames extraction as a conditional generation task. It combines rule-based augmentation for label comprehension with *Chain-of-Thought (CoT)* distillation, effectively transferring reasoning capabilities from LLMs to smaller models to handle complex argument generation steps.

Self-Correction Mechanisms. Recently, Yang et al. [34] proposed a self-data augmentation strategy where a single LLM performs both generation and extraction to reduce resource overhead. A key contribution is the *Logical Thoughts for Self-Data Augmentation (LoTSA)* mechanism, which dynamically evaluates and revises generated samples to ensure reliability without external supervision. Specifically, LoTSA dynamically evaluates generated samples through three logical checks (consistency, completeness, contradiction) and revises low-confidence outputs automatically.

3.2 Prompting and Fine-tuning with LLMs

This paradigm aligns general-purpose models with complex event schemas through three primary strategies: context optimization (Prompting), parameter refinement (Instruction Tuning), and hybrid system design.

Retrieval-Augmented In-Context Learning. To overcome context window limits and hallucinations in zero-shot settings, Shiri et al. [22] optimized In-Context Learning (ICL) using *Retrieval-Augmented Examples (RAE)*. By dynamically populating prompts with semantically similar demonstrations and decomposing the task into sequential detection steps, the framework guides the model effectively without parameter updates. To address unseen roles and long contexts, recent studies utilize cognitive analogies, such as HD-LoA [39], to transfer knowledge across schemas. Alternatively, some approaches integrate retrieval modules, like R-GQA [5], to locate scattered arguments prior to generation. HD-LoA employs hierarchical decomposition of argument roles to transfer knowledge across different schemas. R-GQA integrates a retrieval module to locate scattered arguments before generation, significantly reducing hallucination in long documents.

Reasoning-Enhanced Instruction Tuning. For fine-tuning, recent works inject expert knowledge to improve stability. EE-LCE [37] distills the "logic" of extraction by supervising smaller models with CoT explanations generated by GPT-4. Similarly, Srivastava et al. [26] treat extraction as guideline-constrained code generation, directly injecting detailed *Annotation Guidelines* into the instruction prompt to enhance cross-schema generalization. Furthermore, Guideline-Tuning [26] treats extraction as guideline-constrained code generation. While promising for generalization, it requires robust handling of parsing errors and negative sampling to prevent hallucinations in real-world deployment.

Efficiency and Hybrid Architectures. Addressing computational costs, LFDe [9] reformulates extraction as prompt-based sequence labeling, utilizing weak labels (POS/AMR) for efficient pre-training with minimal data. Finally, to balance precision and reasoning, LC4EE [40] proposes a hybrid "detect-then-correct" paradigm. A specialized Small Language Model (SLM) produces initial predictions, which are subsequently verified and refined by an LLM using retrieved error-feedback rules.

3.3 Comparison with baseline models

Table 3. Performance comparison on ACE2005 (Micro-F1 %, Pred. Triggers).

Category	Model	Year	ACE2005	
			Trig-C	**Arg-C**
Baselines	OneIE[13]	2020	74.7	56.8
	FourIE[18]	2021	72.8	58
	DepIE[19]	2022	74.6	61.2
	QGA-EE[16]	2023	-	57.9
LLM-based	CEAN [17]	2024	**82.5**	61.5
	LLM-PEE [15]	2025	72.6	<u>69.2</u>
	EE-LCE [37]	2024	76.3	61.9
	DEE [22]	2024	<u>81.1</u>	58.2
	LC4EE [40]	2024	77.2	54.9
	Guideline-Tuning [26]	2025	80.8	60.3
	R-GQA [5]	2022	-	**72.8**
	LFDe [9]	2024	70.1	53.6

Note: Comparison is approximate due to heterogeneous protocols. Best results are **bolded**, second-best are <u>underlined</u>.

Quantitative analysis highlights distinct performance patterns. On sentence-level ACE2005 (Table 3), despite the generative power of recent LLMs (e.g., CEAN), retrieval-augmented models like R-GQA ($\approx 72.8\%$) remain dominant in argument classification. This suggests that structural priors are still essential for precise role assignment in short contexts.

Quantitative analysis in Table 4 reveals that Prompting (HD-LoA) and Augmentation (FineCSDA) methods merely match traditional baselines, hovering around 50% Span F1. In contrast, IDEAE achieves a superior 59.6%, demonstrating the distinct advantage of Instruction Tuning. This confirms that document-level span matching necessitates deep parameter updates rather than relying solely on inference-based prompting.

Table 4. Performance comparison on RAMS (Gold Triggers).

Category	Model	Backbone	RAMS Metrics	
			Span F1†	**Head F1**‡
Baselines	TSAR [33]	RoBERTa-Large	51.18	58.53
	SCPRG [14]	RoBERTa-Large	52.32	<u>59.66</u>
	TARA [35]	RoBERTa-Large	<u>52.51</u>	**60.86**
LLM-based	HD-LoA [39]	GPT-4	50.4	42.8
	FineCSDA [1]	BERT/GPT	49.72	56.23
	IDEAE [28]	Flan-T5	**59.6**	54.1

† Argument Identification (Span match).
‡ Argument Classification (Head match).
Note: Comparison is approximate due to heterogeneous protocols.

4 Research Challenges and Opportunities

Table 5 maps challenges to opportunities, assessing their difficulty and impact.

Table 5. Research Challenges and Opportunities

Challenge	**Challenges from Literature**	**Research Opportunities**	**Difficulty**	**Impact**
Data Scarcity	Severe language imbalance: Vietnamese EE is restricted to just two datasets (BKEE [20], VHE [8]) and lacks domain-specific resources	Build finance & law datasets using cross-lingual transfer from LEVEN/Doc2EDAG + semi-automatic annotation	High	High
Hallucination	Prompting-based models (e.g., HD-LoA [39]) show errors in nested events	Consensus voting / Constrained decoding / Multi-stage Reasoning	Medium	High
Computational Cost	Prohibitive LLM fine-tuning costs limit widespread adoption	SLM + LoRA + hybrid detect-then-correct (LC4EE [40])	Medium	Medium
Privacy & Ethics	RAG calling external APIs with sensitive data (e.g., legal, medical)	Parameter-efficient tuning (LoRA) + local deployment	Low	High

5 Conclusion

Our survey reviews EE development in the LLM era. Our analysis shows a shift from general tasks to specialized domains like finance and law, while noting limited resources for languages like Vietnamese. We categorize LLM-based methods into data augmentation and direct application through prompting or fine-tuning. Despite remaining challenges in cost, hallucination and low-resource languages, integrating LLMs with domain-specific requirements represents the most promising future direction. Future research should prioritize the integration of LLMs with structured reasoning and domain-specific knowledge to overcome current limitations in scalability, reliability, and multilingual applicability.

Acknowledgments. We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

References

1. Chen, J., Long, K., Li, S., Tang, J., Wang, T.: Finecsda: Boosting document-level event argument extraction with fine-grained data augmentation. In: Wong, D.F., Wei, Z., Yang, M. (eds.) Natural Language Processing and Chinese Computing. pp. 3–15. Springer Nature Singapore, Singapore (2025)
2. Chen, Y., Zhou, T., Li, S., Zhao, J.: A dataset for document level Chinese financial event extraction. Scientific Data **12**(1), 772 (May 2025). https://doi.org/10.1038/s41597-025-05083-9
3. Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A.T., Joty, S.: Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 1679–1705. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting (2024). https://doi.org/10.18653/v1/2024.findings-acl.97
4. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ACE) program – tasks, data, and evaluation. In: Lino, M.T., Xavier, M.F., Ferreira, F., Costa, R., Silva, R. (eds.) Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). European Language Resources Association (ELRA), Lisbon, Portugal (May 2004), https://aclanthology.org/L04-1011/
5. Du, X., Ji, H.: Retrieval-Augmented Generative Question Answering for Event Argument Extraction. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 4649–4666. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/v1/2022.emnlp-main.307
6. Ebner, S., Xia, P., Culkin, R., Rawlins, K., Van Durme, B.: Multi-Sentence Argument Linking. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8057–8077. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.718
7. Grishman, R., Sundheim, B.: Message Understanding Conference- 6: A Brief History. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics (1996)
8. Hoang, T., Nguyen, L., Dinh, D.: VHE: A New Dataset for Event Extraction from Vietnamese Historical Texts. In: Oco, N., Dita, S.N., Borlongan, A.M., Kim, J.B. (eds.) Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation. pp. 619–634. Tokyo University of Foreign Studies, Tokyo, Japan (Dec 2024)
9. Kan, Z., Peng, L., Gao, Y., Liu, N., Qiao, L., Li, D.: Lfde: A lighter, faster and more data-efficient pre-training framework for event extraction. In: Proceedings of the ACM Web Conference 2024. p. 3964–3975. WWW '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589334.3645318
10. Kulkarni, A., Dogra, V.: Comprehensive survey of event extraction methods in natural language processing. In: 2024 International Conference on Sustainable Communication Networks and Application (ICSCNA). pp. 925–929. https://doi.org/10.1109/ICSCNA63714.2024.10864060

11. Li, G., Wang, P., Xie, J., Cui, R., Deng, Z.: FEED: A Chinese Financial Event Extraction Dataset Constructed by Distant Supervision. In: Proceedings of the 10th International Joint Conference on Knowledge Graphs. pp. 45–53. ACM, Virtual Event Thailand (Dec 2021). https://doi.org/10.1145/3502223.3502229
12. Li, S., Ji, H., Han, J.: Document-Level Event Argument Extraction by Conditional Generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 894–908. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.69
13. Lin, Y., Ji, H., Huang, F., Wu, L.: A joint neural model for information extraction with global features. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7999–8009. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.713
14. Liu, W., Cheng, S., Zeng, D., Hong, Q.: Enhancing document-level event argument extraction with contextual clues and role relevance. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 12908–12922. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.817
15. Liu, W., Li, Z., Bai, L., Zuo, Y., Xu, D., Jin, X., Guo, J., Cheng, X.: Towards event extraction with massive types: LLM-based collaborative annotation and partitioning extraction. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 34365–34387. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1743
16. Lu, D., Ran, S., Tetreault, J., Jaimes, A.: Event extraction as question generation and answering. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 1666–1688. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-short.143
17. Meng, Z., Liu, T., Zhang, H., Feng, K., Zhao, P.: CEAN: Contrastive event aggregation network with LLM-based augmentation for event extraction. In: Graham, Y., Purver, M. (eds.) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 321–333. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024). https://doi.org/10.18653/v1/2024.eacl-long.19
18. Nguyen, M.V., Lai, V.D., Nguyen, T.H.: Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 27–38. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.3
19. Nguyen, M.V., Min, B., Dernoncourt, F., Nguyen, T.: Learning cross-task dependencies for joint extraction of entities, events, event arguments, and relations. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 9349–9360. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.634

20. Nguyen, T.N., Tran, B.T., Luu, T.N., Nguyen, T.H., Nguyen, K.H.: BKEE: Pioneering Event Extraction in the Vietnamese Language. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 2421–2427. ELRA and ICCL, Torino, Italia (May 2024)
21. Satyapanich, T., Ferraro, F., Finin, T.: CASIE: Extracting Cybersecurity Event Information from Text. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 8749–8757 (Apr 2020). https://doi.org/10.1609/aaai.v34i05.6401
22. Shiri, F., Moghimifar, F., Haffari, R., Li, Y.F., Nguyen, V., Yoo, J.: Decompose, enrich, and extract! schema-aware event extraction using LLMs. In: 2024 27th International Conference on Information Fusion (FUSION). pp. 1–8. https://doi.org/10.23919/FUSION59988.2024.10706385
23. Simon, , Olsen, H., You, H., Touileb, S., Øvrelid, L., Velldal, E.: Generative approaches to event extraction: Survey and outlook. In: Tetreault, J., Nguyen, T.H., Lamba, H., Hughes, A. (eds.) Proceedings of the Workshop on the Future of Event Detection (FuturED). pp. 73–86. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.futured-1.7
24. Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., Ma, X.: 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. pp. 89–98. Association for Computational Linguistics, Denver, Colorado (2015). https://doi.org/10.3115/v1/W15-0812
25. Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana: Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media, 1st edn., ISBN: 9781492054047
26. Srivastava, S., Pati, S., Yao, Z.: Instruction-tuning LLMs for event extraction with annotation guidelines. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 13055–13071. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.findings-acl.677
27. Sun, Z., Li, J., Pergola, G., Wallace, B., John, B., Greene, N., Kim, J., He, Y.: PHEE: A Dataset for Pharmacovigilance Event Extraction from Text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 5571–5587. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/v1/2022.emnlp-main.376
28. Wang, C., Wang, X., Chen, H.: Instruction tuning with data augmentation for event argument extraction. In: Huang, D.S., Li, B., Chen, H., Zhang, C. (eds.) Advanced Intelligent Computing Technology and Applications. pp. 257–268. Springer Nature Singapore, Singapore (2025)
29. Wang, X., Wang, Z., Han, X., Jiang, W., Han, R., Liu, Z., Li, J., Li, P., Lin, Y., Zhou, J.: MAVEN: A Massive General Domain Event Detection Dataset. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1652–1671. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.129
30. Xian, G., Du, S., Tang, X., Shi, Y., Jia, B., Tang, B., Leng, Z., Li, L.: DLEE: a dataset for Chinese document-level legal event extraction. Neural Computing and Applications **36**(25), 15581–15597 (Sep 2024). https://doi.org/10.1007/s00521-024-09907-4

31. Xie, J., Zhang, Y., Kou, H., Zhao, X., Feng, Z., Song, L., Zhong, W.: A survey of the application of neural networks to event extraction **30**(2), 748–768. https://doi.org/10.26599/TST.2023.9010139
32. Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., Chen, E.: Large language models for generative information extraction: a survey. Frontiers of Computer Science **18**(6), 186357 (Dec 2024). https://doi.org/10.1007/s11704-024-40555-y
33. Xu, R., Wang, P., Liu, T., Zeng, S., Chang, B., Sui, Z.: A two-stream AMR-enhanced model for document-level event argument extraction. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5025–5036. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.370
34. Yang, L., Fan, X., Wang, X., Wang, X., Chen, Q.: Event extraction based on self-data augmentation with large language models. Memetic Computing **17**(1), 3 (Mar 2025). https://doi.org/10.1007/s12293-025-00436-8
35. Yang, Y., Guo, Q., Hu, X., Zhang, Y., Qiu, X., Zhang, Z.: An AMR-based link prediction approach for document-level event argument extraction. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12876–12889. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.720
36. Yao, F., Xiao, C., Wang, X., Liu, Z., Hou, L., Tu, C., Li, J., Liu, Y., Shen, W., Sun, M.: LEVEN: A Large-Scale Chinese Legal Event Detection Dataset. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 183–201. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.findings-acl.17
37. Yu, Y., Wang, Y., Ma, Y., Li, J., Lu, K., Huang, Z., Chua, T.S.: EE-LCE: An event extraction framework based on LLM-generated CoT explanation. In: Cao, C., Chen, H., Zhao, L., Arshad, J., Asyhari, T., Wang, Y. (eds.) Knowledge Science, Engineering and Management. pp. 28–40. Springer Nature. https://doi.org/10.1007/978-981-97-5492-2_3
38. Zheng, S., Cao, W., Xu, W., Bian, J.: 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). pp. 337–346. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1032
39. Zhou, H., Qian, J., Feng, Z., Hui, L., Zhu, Z., Mao, K.: LLMs Learn Task Heuristics from Demonstrations: A Heuristic-Driven Prompting Strategy for Document-Level Event Argument Extraction. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11972–11990. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.647
40. Zhu, M., Zeng, K., JibingWu, J., Liu, L., Huang, H., Hou, L., Li, J.: LC4EE: LLMs as Good Corrector for Event Extraction. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 12028–12038. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting (2024). https://doi.org/10.18653/v1/2024.findings-acl.715