Overview of CCL24-Eval Task 5: Classical Chinese Historical Event Detection Evaluation

Zhenbing Feng[1,2] , Wei Li[1], Yanqiu Shao[1,2,*]
Information Science School, Beijing Language and Culture University [1]
Language Resources Monitoring and Research Center [2]
15 Xueyuan Road, HaiDian District, Beijing, 100083
zbfengblcu@163.com, liweitj47@blcu.edu.cn, yqshao163@163.com

Abstract

Event detection involves identifying and extracting event information from natural language texts. The complex syntax and semantics of Classical Chinese, coupled with its limited usage, pose significant challenges for information extraction tasks on classical Chinese texts. At the 23rd China National Conference on Computational Linguistics (CCL 2024), we launched an evaluation task focused on the extraction of historical events from Classical Chinese. We used our constructed Classical Chinese Historical Event Logical Schema to identify event triggers and classify event types. The evaluation utilized the Classical Chinese Historical Event Detection Dataset (CHED), annotated from *The Twenty-Four Histories* corpus, with the aim of enhancing event extraction technologies and advancing the digital study of classical Chinese historical texts. The evaluation included two subtasks and attracted 28 teams, with 15 teams submitting valid results. In the subtask of trigger identification, the best-performing system achieved an Exact match score of 63.6%. In the subtasks of coarse-grained and fine-grained event type classification, the top systems achieved F1-scores of 84.5% and 81.4%, respectively.

1 Introduction

Event detection is the process of identifying and extracting relevant event information from natural language texts. Given the complex syntax and semantics of classical Chinese, coupled with its limited usage scope, the task of information extraction for classical Chinese texts remains a significant challenge.

Constructing high-quality datasets tailored to specific domains is essential for event detection tasks. While several high-quality Event Detection (ED) datasets exist for English and modern Chinese, including ACE 2005 (Walker et al., 2006), LEVEN (Yao et al., 2022), MAVEN (Wang et al., 2020), PoE (Li et al., 2022), and DuEE (Li et al., 2020), classical Chinese lacks such datasets due to its semantic complexity and unique historical context. Large-scale datasets in English and modern Chinese are not directly applicable to classical Chinese ED.

Research in deep learning for event detection has explored historical event detection in classical Chinese texts, such as the study of war events in Shiji and ZuoZhuan (Dang, 2021) (Jiuming Ji, 2015) (Zhongbao et al., 2020). However, there still remain challenges such as sparse training data, complex text structures, semantic ambiguity. These issues highlight the need for constructing specific datasets and refining text features. To address these critical challenges and enhance the accuracy and efficiency of classical Chinese event detection , we have developed the classical Chinese Historical Event Dataset (CHED) (Congcong et al., 2023). This dataset is intended to serve as a benchmark for advancing the development of event detection algorithms in classical Chinese historical texts.

The Language and Culture Computing Laboratory (**LCC-Lab**) of Beijing Language and Culture University conducted the historical event type extraction evaluation as part of the CCL 2024 conference. The task aimed to assess the performance of algorithms in detecting event types in classical texts and to promote research and development in classical Chinese event extraction technology. The competition

* Corresponding Author

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 201-209, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

attracted 28 teams, with 15 submitting valid results. The best-performing systems achieved a match score of 76.3% in Trigger Identification and F1-score of 84.5% and 81.4% in fine and coarse Event Type Classification tasks, respectively.

The evaluation included two subtasks: Trigger Identification and Event Type Classification, corresponding to the two steps of event type detection. We summarize the two main features as follows:

(1) Event Type Schema and Dataset [1]. The evaluation used the classical Chinese Historical Event Logical Schema with 9 major categories and 67 subcategories. The CHED dataset, based on *The Twenty-Four Histories* corpus, includes 8,122 valid event instances, providing a robust data benchmark for event detection in classical Chinese texts.

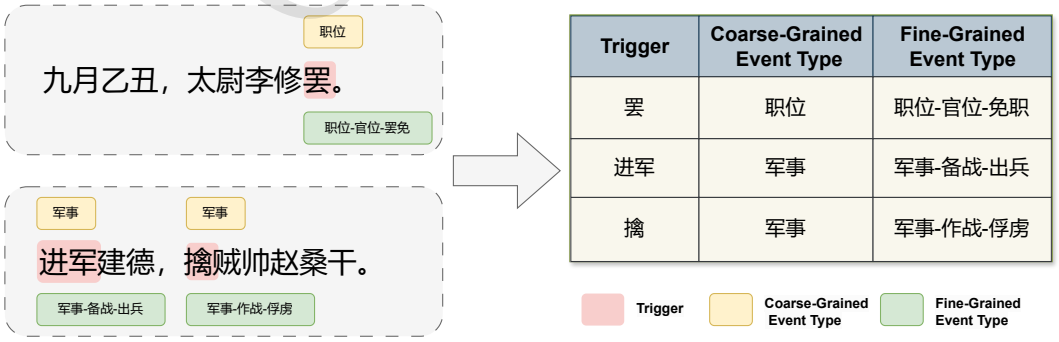
(2) Task Setup. Trigger Identification focused on marking event triggers, while Event Type Classification assigned event types at coarse and fine granularities. This setup ensures a comprehensive assessment of the system's capabilities in accurately identifying event triggers and tests the precision of event type detection algorithms at different granularities, making the evaluation criteria more rational.

2 Task Description

In this section, we will provide a detailed description of the two subtasks.

Subtask 1: Trigger Identification. This subtask involves identifying event triggers within the text and marking their locations. Our dataset is constructed based on the principle of minimal triggers, primarily using single-syllable words that best represent the occurrence of events within the text. The goal is to ensure that the identified triggers are concise yet accurately indicative of the events described. As shown in Figure 1, in the sentence "九月乙丑，太尉李修罢。" (*In September of Yi Chou, General Li Xiu was dismissed.*), the word "罢" (ba) means *dismiss*. Therefore, the trigger in this sentence is "罢" (ba).

Subtask 2: Event Type Classification. In this subtask, each identified trigger word is classified into an event type based on a predefined event type schema. The classification is conducted at both coarse-grained and fine-grained levels, allowing for a detailed understanding of the event's nature. Coarse-grained event types include 9 major categories, and fine-grained event types include 67 subcategories. As shown in Figure 1, in the sentence "进军建德，擒贼帅赵桑干。" (*Advancing to Jiande, capturing the enemy leader Zhao Sanggan*), the word "进军" (advance) could represent the event of dispatching troops to Jiande, and the word "擒" (capture) could represent the event of capturing the enemy leader Zhao Sanggan. Therefore, the trigger "进军" (advance) corresponds to the coarse-grained event type "军事" (*Military*) and the fine-grained event type "军事-备战-出兵" (*Military - Prepare_for_war - Send_troops*). The trigger "擒" (capture) corresponds to the coarse-grained event type "军事" (*Military*) and the fine-grained event type "军事-作战-俘虏" (*Military - Combat - Capture*).



Trigger	Coarse-Grained Event Type	Fine-Grained Event Type
罢	职位	职位-官位-免职
进军	军事	军事-备战-出兵
擒	军事	军事-作战-俘虏

Figure 1: Diagram of classical Chinese Historical Event Detection

[1]Event Type Schema and Dataset is released on https://github.com/NLPInBLCU/CHED2024

3 Data Description

3.1 Event Logical Schema Construction

The construction of an event type schema in a given context should meet the criteria of comprehensive coverage, precise granularity, and high accuracy. As shown in Figure 2 , the initial construction was based on word frequency statistics and semantic clustering of the translated corpus, and it was finalized through trial annotation and expert evaluation.

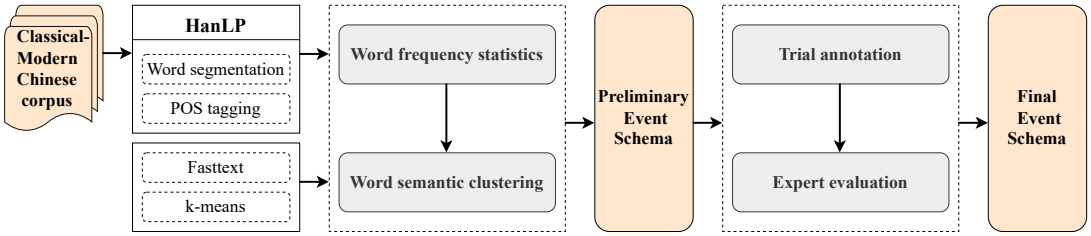


Figure 2: The diagram illustrates the complete process for constructing the event schema.

Word Frequency Statistics. High-frequency words in a text often reflect its main content and central themes, closely related to event types. We selected translated works of *The Twenty-Four Histories* from NiuTrans and used HanLP for word segmentation and part-of-speech tagging, followed by word frequency analysis. High-frequency words such as "进攻" (attack) were identified as potential event types for historical events in classical Chinese.

Semantic Clustering. Further analysis was conducted to classify words with similar semantics automatically. We used Fasttext to generate vector representations for each word and applied the k-means clustering algorithm to group words with high semantic similarity.

Trial Annotation. To evaluate event coverage in texts, we randomly selected 15 documents from the Benji and Liezhuan sections of each book in *The Twenty-Four Histories*. Based on trial annotation results, we modified and merged some event types.

Expert Evaluation. To ensure accuracy and avoid subjective bias, experts and students with backgrounds in linguistics and computer science evaluated our event types. This process led to the final event schema for CHED, including 9 major categories and 67 subcategories. Figure 3 illustrates the complete structure of the military category, which is one of the 9 major categories. The 9 major categories of events include ***Life***, ***Position***, ***Communication***, ***Movement***, ***Ritual***, ***Military***, ***Law***, ***Economy*** and ***Nature***.

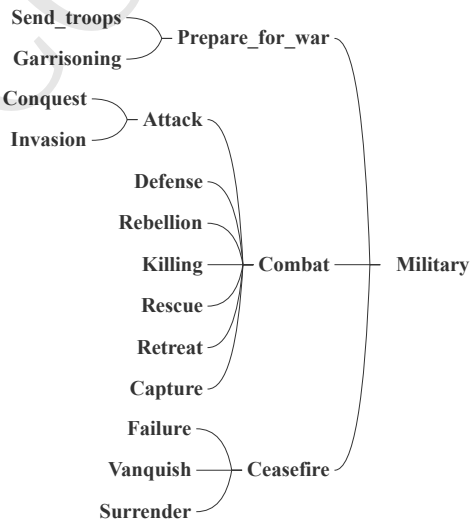


Figure 3: The diagram shows the complete hierarchical structure of the ***Military*** category, including 13 fine-grained event types.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 201-209, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

3.2 Dataset

Through two rounds of annotation, we refined the event type schema, resolved annotation discrepancies, and improved annotation consistency and accuracy. Both annotators were graduate students with backgrounds in linguistics and had received coursework or training in classical Chinese.

We randomly selected 2 to 3 complete volumes from the Benji and Liezhuan sections of each book in *The Twenty-Four Histories*. In total, we selected 61 volumes, covering 61 historical figures and 13,159 sentences for annotation. This effort resulted in the creation of the CHED dataset, which contains 8,122 valid event instances. The distribution of data across different historical books as well as train set, dev set and test set is shown in the table 1.

Number	History Book	Train Set	Dev Set	Test Set	Total
1	Records of the Grand Historian	521	96	98	715
2	Records of the Three Kingdoms	343	86	66	495
3	Book of Zhou	483	118	113	714
4	Book of Liang	38	7	14	59
5	Book of Wei	174	32	24	230
6	Book of Later Han	281	58	61	400
7	Book of Song	75	10	24	109
8	Book of Southern Qi	81	21	30	132
9	Book of Han	417	100	81	598
10	Book of Jin	236	44	44	324
11	Book of Chen	101	21	27	149
12	Book of Northern Qi	141	27	21	189
13	Ming History	249	42	47	338
14	History of the Southern Dynasties	238	52	68	358
15	Book of Sui	205	43	37	285
16	New History of the Five Dynasties	528	117	123	768
17	History of Song	302	49	67	418
18	Book of Northern History	227	60	49	336
19	New Book of Tang	70	25	33	128
20	Old Book of Tang	175	51	49	275
21	History of Liao	98	18	14	130
22	History of Yuan	271	68	78	417
23	History of Jin	175	32	36	243
24	Old History of the Five Dynasties	221	41	50	312
-	**Total**	**5650**	**1218**	**1254**	**8122**

Table 1: Chinese Historical Texts Dataset: Distribution Across Train Set, Dev Set, and Test Set.

4 Evaluation Metrics

This section introduces the evaluation metrics for the two subtasks and the comprehensive assessment.

4.1 Subtask 1: Trigger Identification

Participants will be evaluated on the test set. They must submit a result file containing the position information of triggers within each sentence. This subtask accounts for 40% of the overall score.

The Subset Match Score is calculated as follows:

$$\text{Subset Match Score} = \frac{\sum_{i=1}^{N} \text{SubsetMatch}(S_i, T_i)}{N} \tag{1}$$

where N is the total number of entries, S_i is the set of standard triggers for entry i, T_i is the set of submitted triggers for entry i, and SubsetMatch$(S_i, T_i) = 1$ if all elements of S_i are present in T_i, otherwise SubsetMatch$(S_i, T_i) = 0$.

The Exact Match Score is calculated as follows:

$$\text{Exact Match Score} = \frac{\sum_{i=1}^{N} \text{ExactMatch}(S_i, T_i)}{N} \tag{2}$$

where N is the total number of entries, S_i is the set of standard triggers for entry i, T_i is the set of submitted triggers for entry i, and ExactMatch$(S_i, T_i) = 1$ if S_i is equal to T_i, otherwise ExactMatch$(S_i, T_i) = 0$.

The total score is calculated as:

$$\text{Total Score} = \frac{\text{Subset Match Score} + \text{Exact Match Score}}{2} \tag{3}$$

4.2 Subtask 2: Event Type Classification

Participants must submit a result file containing each trigger and its predicted event type. Subtask 2 is evaluated at both a coarse-grained level (9 major categories) and a fine-grained level (67 subcategories). The evaluation will use both macro-average and micro-average metrics, and this subtask accounts for 60% of the overall score. For **coarse-grained evaluation** , the metrics are calculated based on 9 major categories. For **fine-grained evaluation**, the metrics are calculated based on 67 subcategories, using the same formulas as coarse-grained metrics.

Macro-Average Metrics:

$$\text{Macro Precision} = \frac{1}{L}\sum_{l=1}^{L}\frac{TP_l}{TP_l + FP_l} \tag{4}$$

$$\text{Macro Recall} = \frac{1}{L}\sum_{l=1}^{L}\frac{TP_l}{TP_l + FN_l} \tag{5}$$

$$\text{Macro F1} = \frac{1}{L}\sum_{l=1}^{L}\frac{2 \times TP_l}{2 \times TP_l + FP_l + FN_l} \tag{6}$$

Micro-Average Metrics:

$$\text{Micro Precision} = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L}(TP_l + FP_l)} \tag{7}$$

$$\text{Micro Recall} = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L}(TP_l + FN_l)} \tag{8}$$

$$\text{Micro F1} = \frac{2 \times \text{Micro Precision} \times \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}} \tag{9}$$

Total Score Calculation for Subtask 2:

$$\text{Total Score for Subtask 2} = \frac{\text{Macro F1} + \text{Micro F1}}{2} \tag{10}$$

4.3 Comprehensive Assessment

$$\text{Overall Score} = 0.4 \times \text{Task 1} + 0.3 \times \text{Coarse-grained of Task 2} + 0.3 \times \text{Fine-grained of Task 2} \tag{11}$$

5 Evaluation Results

5.1 Participants and Team Scores

The evaluation attracted a diverse range of participants. A total of 28 teams registered for this evaluation, including 17 teams from academic institutions, 2 teams from commercial organizations, and 9 independent teams or individuals. In the end, 15 teams submitted their results. We released the training and validation datasets for each task on March 5th, and published the answerless test set for the evaluation task on April 5th. All participating teams submitted their result sets before April 25th. The final scores and rankings were announced on May 10th, and some teams' scores were subject to appeal. The scores of the participating teams are shown in the table 2.

Rank	Team	Task 1		Task 2 - coarse		Task 2 - grain		Total Score
		Exact	Subset	Ma-F1	Mi-F1	Ma-F1	Mi-F1	
1	TeleAI	**63.60**	88.90	83.90	**84.50**	74.30	**81.40**	79.12
2	JXNU	63.10	77.50	82.20	82.80	75.50	79.90	76.18
3	SXU	62.00	78.00	82.10	82.60	74.80	79.30	75.82
4	NJU	60.40	77.50	80.90	81.00	75.10	78.60	74.22
5	SCAU	56.90	79.40	80.90	80.70	72.30	78.10	74.11
6	CAS	58.50	77.00	78.70	81.10	70.10	77.60	73.36
7	CPIC	62.80	72.00	77.20	80.30	70.60	74.50	72.53
8	BUPT	60.80	73.60	78.10	78.60	67.70	77.30	72.48
9	XZK	55.50	77.00	77.50	77.10	68.00	72.70	70.02
10	BLCU-1	57.90	70.70	75.60	78.00	60.10	73.50	68.58
11	SEU	54.80	70.00	70.10	78.20	58.10	74.10	68.04
12	XJNU	50.20	60.30	69.90	73.90	54.90	71.20	62.59
13	ZZU	45.40	52.20	66.90	65.60	38.90	48.80	49.10
14	BIT	26.00	47.60	37.80	73.40	18.90	57.70	42.49
15	BLCU-2	0.00	0.00	46.10	51.30	18.80	45.90	28.17

Table 2: **Final Rankings and Scores of Participating Teams in the Evaluation** (Unit: %). Scores for Subtask 1: Trigger Identification, Subtask 2: Coarse-Grained Event Type Classification, and Subtask 2: Fine-Grained Event Type Classification. Please refer to Section 4 for detailed score calculation. Total Score is calculated with a weighting of 4:3:3.

5.2 Methodology and Analysis

We ultimately received system reports from six participating teams. This subsection will summarize and analyze the methods used by the top five ranked teams based on final scores in the evaluation.

TeleAI used a semi-supervised self-training method that combined large and small models. The large language model (Wang et al., 2024) improved accuracy on scarce labeled data, while the small language model (Liu et al., 2023) (Ren et al., 2021) provided flexibility and precision for specific tasks. They generated high-quality pseudo-labels from unlabeled data, initially training the model on labeled data and then creating and filtering pseudo-labels through consistency checks. This method effectively increased training data quantity and diversity, enhancing the model's performance in recognizing event triggers and classifying event types in classical Chinese texts.

JXNU used the EIGC model to jointly extract information by combining external knowledge and a global correspondence matrix. They incrementally pre-trained the BERT-Ancient-Chinese model (Wang and Ren, 2023) on approximately 970,000 classical Chinese texts to enhance its understanding of classical literature. During event extraction, the model incorporated both textual information and external semantic knowledge to improve the accuracy of event trigger and type identification. The EIGC model

used a global correspondence matrix for joint extraction of triggers and event types, employing BERT encoding to incorporate contextual information and integrating part-of-speech data(Yu et al., 2023). This approach effectively increased the accuracy of event extraction in classical Chinese texts and demonstrated broad applicability in handling complex texts .

SXU combined the pre-trained ANCIENT-BERT model (Lample et al., 2016) , optimized specifically for classical Chinese, with Conditional Random Fields (CRF) to enhance the automatic recognition of event triggers and their types in classical texts. ANCIENT-BERT effectively captures the semantic information of traditional and rare characters, while the CRF layer helps find the optimal annotation path globally. They conducted experiments comparing the performance of using ANCIENT-BERT alone versus the combination with CRF. Additionally, they explored the introduction of pointer networks and various loss functions, such as Focal Loss, for further optimization.

NJU used both small-scale and large-scale language models to automatically identify and classify historical event triggers in classical Chinese texts. They selected the small-scale model GujiBERT (Wang et al., 2023), optimized for classical Chinese, and the large-scale model Xunzi-Qwen-14b (Hu et al., 2021). The models employed sequence labeling and sequence-to-sequence evaluation paradigms, respectively, and were pre-trained and fine-tuned for the task. Experimental results showed that GujiBERT, combined with BiLSTM and CRF, outperformed the large-scale model in terms of accuracy and standardized output, especially in high-precision information extraction tasks.

SCAU employed a method based on multi-granularity contrastive learning and Gaussian distribution embeddings. By combining coarse-grained entity-level contrastive learning with fine-grained token-level contrastive learning, they enhanced both the accuracy and efficiency of event detection. They employed the BERT-base-chinese model and applied a CRF schema to refine the learning process. A Gaussian transformation network was employed to create Gaussian embeddings for tokens, assuming that the token semantic representations are distributed according to a Gaussian distribution. Non-linear activation functions were used to produce embeddings for the mean and variance of the Gaussian distribution, and contrastive loss was applied to optimize the relationships between positive and negative samples, further enhancing the model's performance.

Rank	Team Name	Model Used
1	TeleAI	Guwen-NER-RoBERTa-base
2	JXNU	EIGC
3	SXU	ANCIENT-BERT
4	NJU	GujiBERT, Xunzi-Qwen-14b
5	SCAU	BERT-base-chinese

Table 3: Models Used by Top 5 Teams

The participating teams in the CCL24-Eval Task 5 employed various methods to improve trigger identification and event type classification. The table 3 shows the models used by each team. Common approaches included the use of BERT-based models, such as ANCIENT-BERT and BERT-Ancient-Chinese, tailored for understanding classical Chinese texts, and CRF for sequence labeling tasks to identify event triggers and classify event types.

Their optimization was achieved through techniques like semi-supervised learning, incremental pretraining, and the integration of external knowledge. For instance, TeleAI employed pseudo-label generation, JXNU utilized a global correspondence matrix, and SXU combined ANCIENT-BERT with CRF to enhance data diversity, contextual understanding, and optimal annotation paths. Furthermore, NJU demonstrated the effectiveness of using both small and large language models and SCAU applied multi-granularity contrastive learning with Gaussian embeddings. These approaches leverage diverse model architectures and sophisticated learning techniques, aiming to improve both accuracy and efficiency.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 201-209, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

6 Conclusion and Future Work

This paper presents an overview of the evaluation of historical event type extraction from classical Chinese texts. Utilizing the classical Chinese Historical Event Detection dataset (CHED) and the constructed event schema, this evaluation offers a solid data foundation. The evaluation is divided into two subtasks: trigger word identification and event type classification, with the latter further divided into coarse and fine granularity levels. This detailed approach provides a comprehensive reflection of model performance.

A total of 28 teams registered for the competition, with 15 teams submitting valid results and 6 teams submitting system reports. Many participating teams used pre-trained language models (such as BERT) and optimized them through semi-supervised learning, incremental pre-training, and integration of external knowledge to adapt to the event extraction tasks in the field of classical Chinese literature.In the trigger word identification task, the best performing system achieved an Exact match score of 63.6% , while in the fine-grained and coarse-grained event type classification tasks, the best performances reached F1-scores of 84.5% and 81.4%, respectively. Overall, the challenges in event extraction from classical Chinese texts remain rooted in the syntactic and semantic complexity of the language. Enhancing models' ability to understand the deep semantics and subtle contextual nuances of classical Chinese is crucial.

In the future, we will continue to expand the resources for event analysis to provide classical Chinese data support, and continue to conduct evaluation tasks related to classical Chinese event research . We hope that the iterations of future evaluation tasks can continue to promote technological progress in the field of classical Chinese event detection, thereby facilitating the inheritance of classical cultural classics and the advancement of digital humanities.

Acknowledgements

This research project is supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (24YCX167), the National Natural Science Foundation of China (62306045, 61872402), Beijing Language and Culture University's School-level Project (Special Fund for the Basic Scientific Research Business Fee of Central Universities) (18ZDJ03), Beijing Language and Culture University's Phoenix Tree Innovation Platform Project (21PT04).

References

Wei Congcong, Feng Zhenbing, Huang Shutan, Li Wei, and Shao Yanqiu. 2023. CHED: A cross-historical dataset with a logical event schema for classical Chinese event detection. In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 875–888, Harbin, China, August. Chinese Information Processing Society of China.

Jianfei Dang. 2021. Research on knowledge extraction method of chinese classics based on deep learning. Master's thesis, North University of China.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Nan Li Jiqing Sun Jiuming Ji, Jinhui Chen. 2015. Effect analysis of chinese event extraction method based on literatures. *Journal of Modern Information*, 35(12)(3-10).

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

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.

Qian Li, Jianxin Li, Lihong Wang, Cheng Ji, Yiming Hei, Jiawei Sheng, Qingyun Sun, Shan Xue, and Pengtao Xie. 2022. Type information utilized event detection via multi-channel gnns in electrical power systems. *CoRR*, abs/2211.08168.

Shixuan Liu, Chen Peng, Chao Wang, Xiangyan Chen, and Shuangyong Song. 2023. icsberts: Optimizing pre-trained language models in intelligent customer service. In Chrisina Jayne, Danilo P. Mandic, and Richard J. Duro, editors, *International Neural Network Society Workshop on Deep Learning Innovations and Applications, INNS DLIA@IJCNN 2023, Gold Coast, Australia, 23 June 2023*, volume 222 of *Procedia Computer Science*, pages 127–136. Elsevier.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2646–2656. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Pengyu Wang and Zhichen Ren. 2023. The uncertainty-based retrieval framework for ancient chinese CWS and POS. *CoRR*, abs/2310.08496.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A massive general domain event detection dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1652–1671. Association for Computational Linguistics.

Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, and Xiyu Wang. 2023. Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. *CoRR*, abs/2307.05354.

Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Zhongjiang He, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, Yan Wang, Xin Wang, Luwen Pu, Huihan Xu, Ruiyu Fang, Yu Zhao, Jie Zhang, Xiaomeng Huang, Zhilong Lu, Jiaxin Peng, Wenjun Zheng, Shiquan Wang, Bingkai Yang, Xuewei He, Zhuoru Jiang, Qiyi Xie, Yanhan Zhang, Zhongqiu Li, Lingling Shi, Weiwei Fu, Yin Zhang, Zilu Huang, Sishi Xiong, Yuxiang Zhang, Chao Wang, and Shuangyong Song. 2024. Telechat technical report. *CoRR*, abs/2401.03804.

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 Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 183–201. Association for Computational Linguistics.

Zijian Yu, Tong Zhu, and Wenliang Chen. 2023. (syntax-aware event argument extraction). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 196–207.

Liu Zhongbao, Dang Jianfei, and Zhang Zhijian. 2020. Research on automatic extraction of historical events and construction of event graph based on historical records. *Library and Information Service*, 64:116–124.