CHED: A Cross-Historical Dataset with a Logical Event Schema for Classical Chinese Event Detection

Congcong Wei [*], **Zhenbing Feng** [*], **Shutan Huang, Wei Li, Yanqiu Shao** [†]
Information Science School, Beijing Language and Culture University
Language Resources Monitoring and Research Center
15 Xueyuan Road, HaiDian District, Beijing, 100083
weicongcong0214@163.com, zbfengblcu@163.com
liweitj47@blcu.edu.cn, yqshao163@163.com

Abstract

Event detection (ED) is a crucial area of natural language processing that automates the extraction of specific event types from large-scale text, and studying historical ED in classical Chinese texts helps preserve and inherit historical and cultural heritage by extracting valuable information. However, classical Chinese language characteristics, such as ambiguous word classes and complex semantics, have posed challenges and led to a lack of datasets and limited research on event schema construction. In addition, large-scale datasets in English and modern Chinese are not directly applicable to historical ED in classical Chinese. To address these issues, we constructed a logical event schema for classical Chinese historical texts and annotated the resulting dataset, which is called classical Chinese Historical Event Dataset (CHED). The main challenges in our work on classical Chinese historical ED are accurately identifying and classifying events within cultural and linguistic contexts and addressing ambiguity resulting from multiple meanings of words in historical texts. Therefore, we have developed a set of annotation guidelines and provided annotators with an objective reference translation. The average Kappa coefficient after multiple cross-validation is 68.49%, indicating high quality and consistency. We conducted various tasks and comparative experiments on established baseline models for historical ED in classical Chinese. The results showed that BERT+CRF had the best performance on sequence labeling task, with an f1-score of 76.10%, indicating potential for further improvement. [1]

1 Introduction

Event detection (ED) is a significant research area in natural language processing (NLP). The ED task mainly includes two steps. Firstly, recognizing and labeling triggers (words that best represent the occurrence of events) in the text, and secondly, determining the event types to which triggers belongs. For example, 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), and we label this sentence as a "职位-官位-免职" (*Position-Official_position-Dismiss_from_a_position*) event triggered by the word "罢" (ba).

Constructing high-quality datasets for specific domains is critical for ED tasks. Several high-quality ED datasets exist for English and Chinese, such as 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). However, classical Chinese lacks such datasets due to complex semantics and special era. Large-scale datasets in English and modern Chinese are not directly applicable to classical Chinese ED. The current research on ED in classical Chinese is limited by the lack of high-quality datasets that are specific, systematic, and scalable.

To address these crucial issues and enhance the accuracy and efficiency of classical Chinese ED, we have constructed the classical Chinese Historical Event Dataset (CHED). This dataset has the potential to serve as a benchmark for developing and evaluating ED algorithms for classical Chinese historical

[*] Equal contribution
[†] Corresponding Author
[1] The CHED data is released on https://github.com/lcclab-blcu/CHED

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

texts. The hierarchical and logical event schema of the CHED can be extended and adapted to other NLP domains, making it a valuable resource not only for NLP researchers but also for scholars in other humanities fields. Moreover, CHED offers a unique historical perspective for exploring ancient societies, enhancing our comprehension of their cultures and interconnections. It also supports the digital humanities research and helps preserve cultural heritage through the study of classical Chinese texts.

During the construction of our dataset, we encountered three primary challenges: **1)** developing an event schema that could encompass the majority of events described in classical Chinese literature; **2)** accurately identifying and classifying events within cultural and linguistic contexts while accounting for the ambiguity resulting from multiple meanings of words in historical texts; **3)** ensuring consistent annotation results, which was essential throughout the entire dataset construction process.

To address these challenges, we proposed several approaches. One such approach involved subjecting the processed data and preliminary event schema to trial annotation and expert review. Through several revisions and validations, we constructed a hierarchical and logical event schema with fine granularity, consisting of 9 major event categories and 67 subcategories that cover significant events in ancient Chinese history. The 9 major categories of events include ***Life***, ***Position***, ***Communication***, ***Movement***, ***Ritual***, ***Military***, ***Law***, ***Economy***, and ***Nature***. The complete event schema has been placed in the appendix A, as shown in Figures 11 and 12. In addition, we have annotated a total of 8,122 valid sentences.

To ensure further accuracy, our annotators possessed extensive knowledge of classical Chinese and actively sought expert opinions while constructing the dataset. Multiple cross-validation were also conducted, yielding an average Kappa coefficient of **68.49%**, which denotes a high level of consistency and quality. Additionally, we conducted various tasks and comparative experiments on established baseline models for historical ED in classical Chinese. The outcomes indicated that BERT+CRF exhibited the highest performance on sequence labeling task, achieving an f1-score of **76.10%**.

We conclude three main contributions as follows: **1)** We constructed the CHED, which provides a rich cross-historical data foundation for classical Chinese ED, making it a valuable resource for scholars and researchers. The dataset contains 8,122 valid sentences; **2)** We proposed a hierarchical and logical event schema, which has a fine-grained structure that can be adapted more effectively to other NLP domains; **3)** We excavated a unique and profound historical perspective from the CHED, promoting the advancement of digital humanities research.

2 Related work

In the realm of event detection (ED) tasks in deep learning, sparse and imbalanced training data, complex text, and semantic ambiguity still pose problems, highlighting the importance of dataset construction and feature extraction through text refinement processing.

A high-quality dataset is essential for ED tasks. It should be large enough to support various learning algorithms, has high accuracy and consistency in labeled data, and contains diverse event types. Several high-quality annotated ED datasets have been constructed, including the widely used English dataset ACE 2005 (Walker et al., 2006), the legal ED dataset LEVEN (Yao et al., 2022), the large-scale cross-domain ED dataset MAVEN (Wang et al., 2020), the electrical power ED dataset PoE (Li et al., 2022) and the Chinese event dataset DuEE (Li et al., 2020) based on real-world scenarios. While many studies have summarized the primary methods of Chinese ED based on literature, classical Chinese field ED faces challenges due to differences in context and expression of historical texts.

There have been studies using deep learning methods to investigate historical ED in classical Chinese texts (Jiuming Ji, 2015), such as researching the war events in the ZuoZhuan (左传). For example, the RoBERTa-CRF model was established (Xuehan Yu, 2021), and pattern matching and CRF models were used to extract events from the ZuoZhuan (左传) (Zhongbao Liu, 2020). Additionally, mixed techniques using information extraction have been applied to classical Chinese texts, including entity recognition and event extraction, with the extracted information being visualized using electronic charts (Li, 2019). Furthermore, studies have been conducted on extracting historical events and event elements from Shiji (史记) and ZuoZhuan (左传) (Dang, 2021). However, these studies have only produced coarse-grained event type constructions, mostly focused on a single text and based on relatively small dataset sizes.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

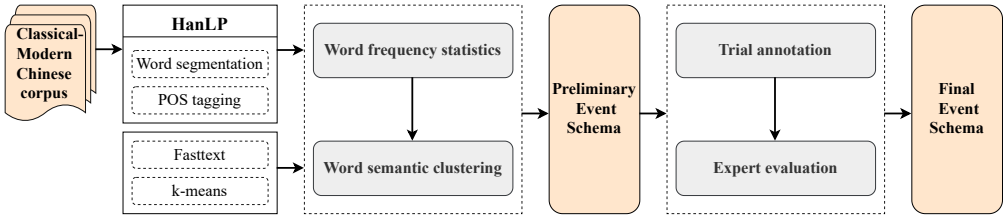

Figure 1: This is the complete process for constructing the event schema. The preliminary construction was based on word frequency statistics and semantic clustering of the translations corpus, and it was finalized through trial annotation and expert evaluation.

3 Event Schema Construction

The construction of event types in a given context should fulfill the criteria of comprehensive coverage, precise granularity, and high accuracy. To achieve these goals, we mainly carried out work in four aspects, as shown in Figure 1: **1) Word frequency statistics**; **2) Word semantic clustering**; **3) Trial annotation**; **4) Expert evaluation**. Eventually, we constructed an event schema that includes 9 major categories and 67 subcategories. Figure 2 depicts the structure of one of the major categories, ***Position***.

We assume that the words with higher frequency in the text reflect the main content and central theme of the text, which is closely related to the event types. Therefore, it is necessary to conduct comprehensive word frequency statistics on the text to ensure the coverage of event types. We selected the translated works of the Twenty-Four Histories from NiuTrans [1]and used HanLP [2] for basic word segmentation and part-of-speech tagging on the corpus, and conducted word frequency statistics based on the results. After removing stop words and irrelevant part-of-speech tags, we analyzed the word frequency statistics results of nouns, verbs, and gerunds. We discovered that certain high-frequency words, such as "进攻" (attack), could serve as event types for historical events in classical Chinese.

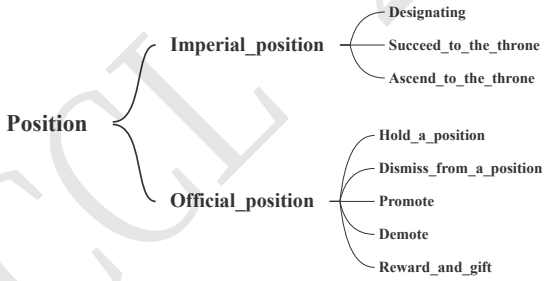

Figure 2: ***Position*** is one of the 9 major event categories in the CHED event schema, and this diagram shows the complete hierarchical structure of ***Position***.

Semantic clustering analysis was further conducted on words to automatically classify similar semantic words, aiming to provide more refined classification references for the construction of classical Chinese event types. We used Fasttext[3] to generate vector representations for each word and the k-means clustering algorithm to cluster words with high semantic similarity. Based on the analysis, and inspired by the ACE (Walker et al., 2006), MEVEN (Wang et al., 2020), LEVEN (Yao et al., 2022) and other datasets, we preliminarily summarized the classical Chinese historical event types, including 15 major categories and 73 subcategories.

To evaluate the actual event coverage in real-world texts, we randomly selected 15 volumes from the Benji (本纪) and Liezhuan (列传) sections of each book in the Corpus of the Twenty-Four Histories pro-

[1]https://github.com/NiuTrans/Classical-Modern
[2]https://github.com//hankcs/pyhanlp
[3]https://github.com/facebookresearch/fastText

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

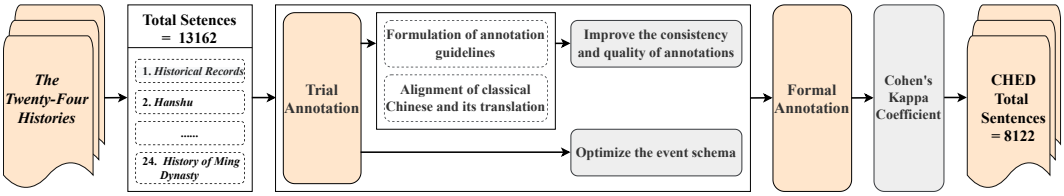

Figure 3: The entire annotation process from raw corpus to dataset is presented, including two main stages, as well as the measures taken to ensure the quality of annotation.

vided by the Hancheng website[4], which included a total of 8,304 sentences,for trial annotation. Finally, we obtained 2,913 annotated sentences and 4,047 event labels. Based on the trial annotation results and the actual situation during the annotation process,we modified and merged some event types.

In addition, to ensure the accuracy of classical Chinese historical event types and avoid personal subjective bias, we invited experts and students with linguistic and computer science backgrounds to evaluate our event types. After these efforts, we constructed the final event schema for CHED.

4 Annotation Process

We used Figure 3 to illustrate our process.

4.1 Document Selection

In order to ensure the completeness and high quality of the corpus, we chose the published book *The Twenty-Four Histories (12 volumes of annotated editions with comparison of classical Chinese and modern Chinese)* published by Xianzhuang Shuju (线装书局) as our main source of annotated corpus.

There are three main reasons for choosing published books: **1) High-quality corpus**: the corpus in published books has been carefully selected and strictly reviewed multiple times; **2) Reduced workload**: the standardized typesetting of books eliminates the need for additional data prepossessing; **3) Provide reference translations**: the books provide high-quality aligned classical Chinese and modern Chinese corpus, facilitating reference for annotation personnel.

We mainly focused our annotations on the Benji(本纪) and Liezhuan (列传) (the main body of the histories), selecting 2-3 complete volumes at random from each of the Twenty-Four Histories to ensure complete historical figure records. In total, we selected 61 volumes, comprising 61 historical figures, 13,159 sentences, and 236,842 characters. Our main objective is to identify and label triggers in classical Chinese texts, and determine the event categories to which these triggers belong.

4.2 Annotation stage

The annotation process mainly consisted of two stages: **1)Trial annotation** was to preliminary test and refine the types of historical events in classical Chinese, as well as to unify the annotation discrepancies between the two annotators. This was helpful for improving the consistency and accuracy of the formal annotation stage; **2) Formal annotation**: the two annotators were assigned different tasks. Annotator 1 was responsible for annotating the first 12 books of the Twenty-Four Histories, while annotator 2 for the latter 12 books. Specifically, as shown in the figure 4, we created a sequence annotation project on the Doccanno platform [5] and split the documents into units of sentences delimited by periods for ease of annotation.

4.3 Annotation quality

In this section, we introduce our three main measures taken to ensure the accuracy and consistency of the annotated corpus.

[4]https://guoxue.httpcn.com/zt/24shi/
[5]https://github.com/doccano

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

Figure 4: The Doccano annotation interface contains three events in this example, with triggers “见” (jian), “取” (qu), and “生” (sheng). We can select the corresponding event type below each trigger for annotation.

4.3.1 Annotation Guidelines

To ensure dataset quality and improve manual annotation consistency, rules and standards have been established for selecting triggers.

Contextual and semantic priority. We should focus on the semantics of the translation and its original context because the problem of polysemy is particularly prominent in classical Chinese, and the process of annotation is prone to errors in understanding. In example (1) and (2), “胜之” (sheng zhi) and “败之” (bai zhi) have different usages, but both semantically denote victory. We annotated both of them as “ Military-Ceasefire-Vanquish ” based on the semantic meaning of the translated text.

(1) 军事-停战-战胜：四月，友宁引兵西，至兴平，及李茂贞战于武功，大败之。

(*Military-Ceasefire-Vanquish: In April, Youning led his army westward to Xingping and fought against Li Maozhen in Wugong, where he achieved a resounding victory.*)

(2) 军事-停战-战胜：与晋战河阳，胜之。

(*Military-Ceasefire-Vanquish: In the battle against Jin at Heyang, they emerged victorious.*)

Simplest trigger. It’s best to use simple triggers that are easy to understand and annotate, as this reduces the time and cost of annotation, minimizes subjective differences among annotators, simplifies subsequent processing and analysis, and ultimately improves the accuracy and reliability of the annotated data. For example (3), we only label the noun “水” (flood), while “大” (massive) is not labeled.

(3) 自然-灾害-水灾：秋七月乙酉，三郡大水。

(*Nature-Disaster-Flood/Drought: In the second month of autumn, there was a severe flood in three counties.*)

Event property. It is difficult to immediately determine event attributes such as tense and polarity in classical Chinese because crucial information is often omitted. We have adopted LEVEN’s event annotation guidelines (Yao et al., 2022) and annotate any events that are mentioned. In example (4), even if the attack has not yet taken place, we still annotate it.

(4) 军事-攻击-征伐：引兵欲攻燕，屯中山。

(*Military-Attack-Conquest: The army is preparing to attack Yan kingdom and stationed at Zhongshan.*)

Incorporation of ancient cultural knowledge. Classical Chinese contains a wealth of historical and cultural background knowledge that must be taken into consideration when constructing event schema and annotating them. For example, Classical Chinese has specific vocabulary expressions for the change of official positions, such as “去” (qu) and “罢” (ba), which means “dismiss”.

4.3.2 Alignment of classical Chinese and its translation

It was necessary to provide annotators with an objective reference translation standard during the annotation process to ensure consistency, given the difficulty of understanding the semantics of classical Chinese. Our aligned classical Chinese and modern Chinese data mainly came from the *Twenty-Four Histories (12 volumes of annotated editions with comparison of classical Chinese and modern Chinese)* published by Xianzhuang Shuju (线装书局).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

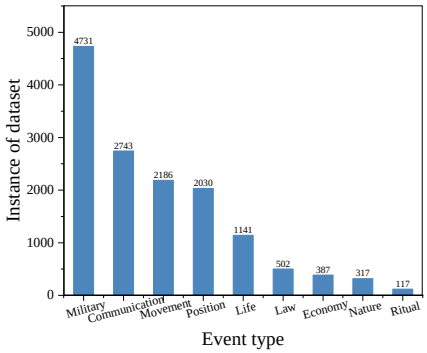

Figure 5: Distribution of event types in CHED.

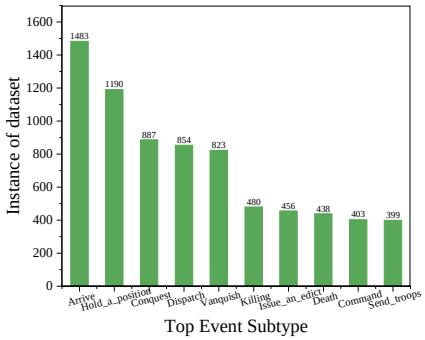

Figure 6: Top event sub-types in CHED.

4.3.3 Cohen's Kappa Coefficient

To verify the consistency of the annotations and ensure the validity and reliability of the dataset, we conducted cross-validation using Cohen's Kappa coefficient. Specifically, the labeled sentences were divided into two datasets, A and B, with annotator 1 and annotator 2 each annotating a portion of the sentences. A random sample of 10% of the sentences was taken from each dataset, and the annotators swapped datasets to annotate the sampled sentences.

Regarding the calculation standard for Cohen's kappa coefficient, we considered the annotation to be consistent if both annotators labeled the same event labels for the same sentence, and considered it to be inconsistent if they labeled different event labels. After conducting 4 rounds of cross-validation, the average kappa coefficient was **68.49%**, indicating a relatively high level of consistency between the two annotators and a high level of reliability for the annotation results.

Inconsistent annotations often stem from ambiguity resulting from multiple meanings of words in historical texts. Such as example (5), the character "屯" may have been incorrectly labeled as the trigger for the "Military-Garrisoning" event. However, it is actually a noun that means "military camp". Therefore, the sentence should be annotated with "还" as the trigger word for the "Movement-Arrive" event type.

(5) 坚还屯。 (*Sun Jian returned to the military camp.*)

*Annotator 1: *Military-Garrisoning*: 坚还屯。 Annotator 2: *Movement-Arrival*: 坚还屯。

5 Data Analysis

In this section, we mainly introduce the scale and distribution of the dataset, as well as the phenomenon of data sparsity that has been observed, and provide possible explanations for it.

5.1 Data Size

The dataset consists of 61 volumes and 61 historical figures from the Twenty-Four Histories, comprising a total of 13,159 sentences and 236,842 characters. Among them, there are 8,122 sentences with event labels, totaling 145,973 characters, and a total of 14,154 labels.

The scale of the dataset we finally constructed is moderate due to the difficulty and high cost of cross-historical annotation. However, it contains rich information on classical Chinese history texts from different dynasties and historical figures, and has certain representativeness. It can be used in the future to train and evaluate algorithms and models for classical Chinese historical event detection.

5.2 Data Distribution

An imbalanced distribution of event types is indicated by Figure 5 in the CHED dataset. The major event types-including *Military*, *Communication*, *Movement* and *Position*, account for the vast majority of the dataset. Among the event sub-types depicted in Figure 6, including *Arrive*, *Hold_a_position*, *Conquest*, *Dispatch* and others, the proportions are higher. This imbalance may result in insufficient recognition of minority events by models, posing a challenge for future classical Chinese ED tasks.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

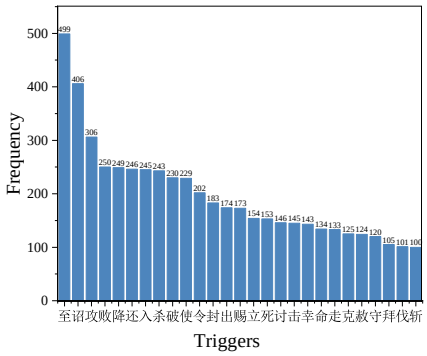

Figure 7: The triggers that appear with a frequency greater than 100 in the sentences of the CHED.

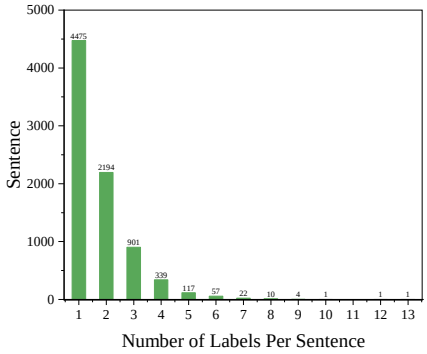

Figure 8: The number of event labels that appear per sentence in the CHED.

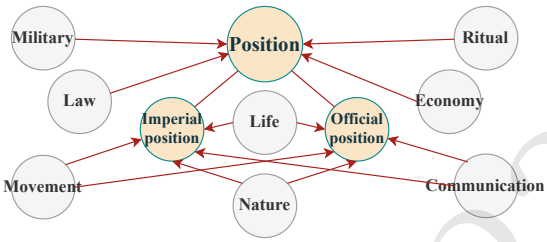

Figure 9: The figure shows the ***Position*** event divided into ***Imperial position*** and ***Official position*** connected through ***Movement*** and ***Communication***. ***Military***, ***Law***, ***Ritual***, and ***Economy*** events serve ***Position*** while ***Nature*** events affect people represented by ***Imperial position*** and ***Official position***.

Following our previous annotation standards, Figure 7 displays the triggers that appear at a frequency greater than 100 in the sentences, which primarily consist of monosyllabic words. The frequency distribution of triggers corresponds to the proportion of event types. For instance, "至" (zhi) corresponds to the *Arrival* event. This indicates that identifying high-frequency triggers in a sentence to predict the corresponding event type is a vital aspect in classical Chinese historical ED.

Displaying the number of event labels that appear in a single sentence, Figure 8 reveals that a single sentence typically contains one or multiple event types, with 1-3 event types being the most common. This poses a challenge for accurately detecting multiple event types in classical Chinese.

Several possible explanations for the imbalanced distribution observed in the CHED dataset have been identified based on historical facts from ancient China. The frequency of certain events in historical texts reveals significant aspects of ancient Chinese political, social, and cultural life. The emphasis on posthumous honor is shown by the disparity in the frequency of birth and death events. The prevalence of imperial edict events indicates a society governed by men rather than laws, while the high proportion of position events is a result of the imperial examination system. The frequent occurrence of military events reflects the challenges to the legitimacy and orthodoxy of feudal monarchy. Overall, these findings align with historical reality and demonstrate the potential for effective digitization of ancient literature.

5.3 Event Logical System Construction

We have constructed a complete and logically consistent ontology of classical Chinese historical event types that exhibit a hierarchical relationship and entail connections between the major categories of event types, as shown in Figure 9. It is our belief that the central theme of records in Twenty-Four Histories continues to revolve around political power struggles and the pursuit of authority. Therefore, we focus on the ***Position*** events as the core, which are further divided into ***imperial position*** and ***official position***, reflecting the two major relationships between emperors and officials in ancient China.

The transition of political power is generally reflected in ***Military*** events, which seize power through

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

Table 1: The detailed statistics of subsets of CHED

Dataset	Sentences	Event_labels	Characters
Training	5,685	9,979	102,636
Validation	1,218	2,056	21,618
Test	1,219	2,119	21,719
Total	8,122	14,154	145,973

warfare and maintain power through ***Law*** events, supported by the ***Economy*** events that are centered around the taxation system. In order to strengthen the legitimacy of their political power, emperors often hold ***Ritual*** events, including the worship of heavenly deities to emphasize the divine right of emperors, the worship of ancestral spirits to emphasize the continuity of their bloodline-based inheritance system centered around the eldest son, and the worship of sages (e.g. Confucius) to provide a source of legitimate political ideology for their regimes.

In a political system that centers around imperial power, there exists a relationship between emperors and officials, where ***Movement*** and ***Communication*** events are utilized to facilitate the transmission of political orders and the implementation or abolition of measures from top to bottom. The ***Life*** events mainly refer to the lives of the emperor and the officials, which are the main records of figures in the Benji (本纪), Liezhuan (列传) and Shijia (世家) sections of the Twenty-Four Histories.

At the same time, the records of ***Nature*** events in the Twenty-Four Histories mainly focus on how natural events affected the behavior of the emperor and the officials. For example, in Volume One of Song Shi (宋史), in the Benji (本纪) of Taizu (太祖), the sentence following contains *famine* event, which affected the emperor's subsequent actions, namely, ordering the opening of granaries to provide relief for the people due to the occurrence of famine in eight provinces.

(6) 辛亥，澶、滑、卫、魏、晋、绛、蒲、孟八州饥，命发廪振之。

(*In Xinhai year, there was a famine in eight provinces, including Chanzhou, Huazhou, Weizhou, Jinzhou, Jingzhou, Puqizhou, and Mengzhou. The emperor commanded the opening of granaries to provide relief for the people.*)

6 Experiments

6.1 Setting

We randomly shuffled the dataset and divided it into training set, validation set, and test set in a ratio of 0.7:0.15:0.15. The sizes of each part of the dataset are shown in the Table 1.

Regarding the hyper parameters of the model, including BERT, BiLSTM, IDCNN, CRF, we set the seed number of the random number generator to 123 to ensure the reproducibility and stability of the model. We set the maximum input sequence length to 150 to ensure model performance. The train batch size is set to 32, and the eval batch size is set to 12 for training and validation batches, respectively. Due to the specificity of the corpus and the imbalance of the labels, we set the number of training epochs to 30, the learning rate to 3e-05, dropout to 0.3, and adam epsilon to 1e-08 to prevent the model from over-fitting.

Inspired by Leven (Yao et al., 2022), we used two perspectives of micro and macro for the evaluation metrics of the model, including precision, recall, and f1-score. This was because we noticed the imbalance of the labels for classical Chinese event types. The micro perspective focuses on categories with a large number of samples, considering the frequency of each category's occurrence in the samples. The macro perspective treats each category equally, enabling us to evaluate the model from multiple aspects.

6.2 Baseline

We approached the ED task by dividing it into two tasks: **1) Sequence labeling task**: We labeled the event type corresponding to the triggers to detect events in a sentence, using BERT, BiLSTM, IDCNN, and CRF as baseline models. BERT from chinese-bert-wwm-ext (Devlin et al., 2019) was used as the

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

Table 2: The experimental results by modeling ED as a sequence labeling task on the CHED.

Model	Micro			Macro		
	Precision	Recall	F1-score	Precision	Recall	F1-score
BERT	74.58	77.11	75.82	**67.95**	65.05	65.19
BERT+CRF	**75.15**	77.06	**76.10**	67.69	65.22	64.98
BiLSTM	70.40	64.98	67.58	58.40	51.36	53.15
BiLSTM+CRF	70.24	66.73	68.44	60.77	52.91	54.76
IDCNN	71.70	60.97	65.90	57.98	44.40	49.05
IDCNN+CRF	71.04	63.66	67.15	55.50	46.88	49.44
BERT+BiLSTM+CRF	72.93	**77.68**	75.23	66.17	**66.64**	**65.23**

Table 3: The experimental results by modeling ED as a multi-class classification task on the CHED.

Model	Micro			Macro		
	Precision	Recall	F1-score	Precision	Recall	F1-score
BERT+Prompt	87.36	87.36	87.36	86.88	74.26	76.27
T5+Prompt	87.93	87.93	87.93	83.39	74.70	75.69

input vector representation for BiLSTM and IDCNN models, and the project code was based on tianshan1994 [6] (Shi et al., 2011); **2) Multi-class classification task**: We utilized BERT (Devlin et al., 2019) and T5 (Raffel et al., 2019) with human-crafted prompts to predict the upcoming sentence given the known context, and transformed the multi-label classification problem into a binary classification problem to detect events in a sentence. We designed a prompt template: (***["placeholder":"text_a"] Does the sentence contain ["placeholder":"text_b"]? [MASK]***), and "text_a" represents the sentence text and "text_b" represents the event type. The project code was based on Openprompt (Ding et al., 2022) [7].

The baseline models used in each task were: BERT (Devlin et al., 2019) and T5 (Raffel et al., 2019) are pre-trained language models that have demonstrated state-of-the-art performance on a range of NLP tasks. BiLSTM is a widely used sequence modeling method that captures bidirectional context (Hochreiter and Schmidhuber, 1997). IDCNN is a convolutional neural network that uses different dilation kernel sizes to capture contextual information at different ranges (Cao and Yusup, 2022). CRF is a commonly used sequence labeling model that improves labeling accuracy by considering the dependencies between labels (Lafferty et al., 2001). Prompt is a novel technique for zero-shot learning tasks that allows the model to perform new tasks without any training examples by adding special prompts (Ding et al., 2022).

6.3 Result and Analysis

In the sequence labeling task, overall, the micro-average results outperformed the macro-average results, due to the imbalanced distribution of event labels where some labels had fewer instances in the dataset, resulting in insufficient learning by the model. The results showed that the BERT+CRF model performed the best,while the performance of the BiLSTM and IDCNN models was inferior, respectively. Additionally, the BERT+BiLSTM+CRF model had the highest macro-average f1-score while the IDCNN model had the lowest macro-average f1-score.

These results indicate that the BERT+CRF model is better suited for this task than other models, as it can capture richer contextual information and the use of CRF can address label dependencies and enhance algorithm performance. However, in the historical ED task in classical Chinese texts, triggers are often monosyllabic, and label dependencies may not be as strong, hence the influence of the CRF model may not be as significant.

In multi-class classification tasks, the difference between the results of micro-average and macro-

[6]https://github.com/taishan1994/pytorch_bert_bilstm_crf_ner
[7]https://github.com/thunlp/OpenPrompt

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

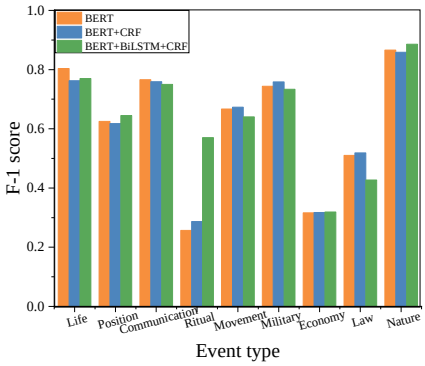

Figure 10: The comparison results of f1-scores for different models across different event types in CHED.

average is not significant compared to sequence labeling tasks. This may be because Prompt is more suitable for handling datasets with few samples, and it provides additional information to the pre-trained language model through manually designed prompts, enabling the model to better utilize existing knowledge for classification tasks. Moreover, the Prompt method performed well. Unlike sequence labeling tasks, multi-class classification tasks focus more on the classification of historical events in classical Chinese texts, and therefore, the BERT / T5 + Prompt model may have an advantage in classification.

There may be several reasons for such results: **1) Model structure**: The superior performance of BERT in historical ED tasks in classical Chinese texts may be attributed to its pre-trained Transformer-based architecture that effectively captures contextual information, compared to traditional neural network models like BiLSTM and IDCNN that may be affected by sequence length limitations and gradient vanishing. However, combining BERT with BiLSTM and CRF in the BERT+BiLSTM+CRF model did not yield the expected performance level, possibly due to increased noise or conflicts resulting from the introduction of more complexity and parameters. **2) Annotation errors**: Despite our efforts to ensure the quality and consistency of the annotations, the complexity of the context and cultural context of classical Chinese, as well as the ambiguity of word meanings, may lead to some annotation errors in the dataset, especially when the annotator's knowledge level is limited. These errors may have an impact on the performance of the models.

3) Sparse samples: As shown in the figure 10, the f1-scores of different event types on different models are displayed. We can see that the performance of the Ritual, Economy, and Law events is poorer compared to other events, and the number of samples for these three event types in the dataset is also the smallest. With an imbalanced distribution, the presence of some noise or mislabeling may lead to poor recognition ability of the model for certain event types and stronger recognition ability for other types.

Overall, the BERT+CRF model performed the best in the task of historical ED in classical Chinese texts. The Prompt method also performed well. However, there is still significant room for improvement and challenges in future research.

7 Conclusion and Future Work

In conclusion, we have constructed a hierarchical and logical schema for classical Chinese events and used it to create the CHED based on the Twenty-four Histories corpus. The CHED can effectively facilitate the advancement of digital humanities research by providing a unique and profound historical perspective. Despite encountering various challenges during the construction of the dataset, we ensured the consistency and quality of the annotations. We assessed the effectiveness and quality of the dataset by testing it against several baselines and calculating kappa scores, and we obtained satisfactory results. Nevertheless, there is scope for further enhancement, and our future work will concentrate on expanding and optimizing the dataset to meet a wider range of application needs. Our dataset is a valuable resource not only for natural language processing but also for classical literature and cultural studies. Furthermore, it makes a significant contribution to the field of event detection in classical Chinese, and we anticipate that it will inspire further research and exploration.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

Acknowledgements

This research project is supported by the National Natural Science Foundation of China (61872402), Science Foundation of Beijing Language and Culture University (supported by " the Fundamental Research Funds for the Central Universities ") (18ZDJ03) .

References

Yingjie Cao and Azragul Yusup. 2022. Chinese electronic medical record named entity recognition based on BERT-WWM-IDCNN-CRF. In *9th International Conference on Dependable Systems and Their Applications, DSA 2022, Wulumuqi, China, August 4-5, 2022*, pages 582–589. IEEE.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. Openprompt: An open-source framework for prompt-learning. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 105–113. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

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).

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

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.

Zhongkai Li. 2019. The study on the extraction of war events in zuo zhuan based on mixed approaches. Master's thesis, Nanjing Agricultural University.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Xiaodong Shi, Yidong Chen, and Xiuping Huang. 2011. Key problems in conversion from simplified to traditional chinese characters. In *International Conference on Asian Language Processing*.

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

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.

Lin He Jian Xu Xuehan Yu. 2021. Extracting events from ancient books based on roberta-crf. *Data Analysis and Knowledge Discovery*, 5(26–35).

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.

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

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

A Event schema of the CHED

Figure 11: Event schema of the CHED in English

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

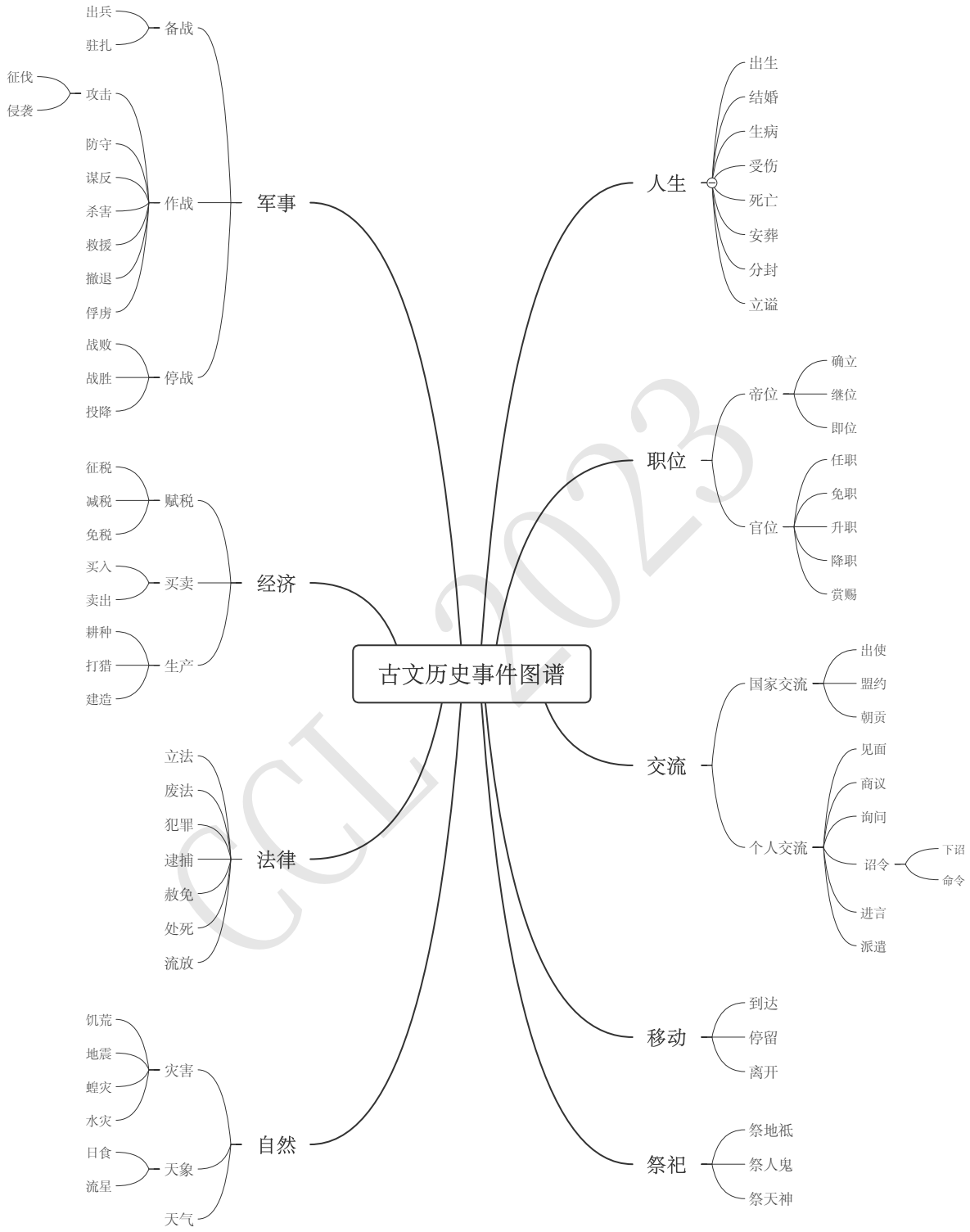

Figure 12: Event schema of the CHED in Chinese

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.