Information Research - Vol. 31 No. iConf (2026)

Towards automated genre conversion: aggregating thematic events in classical Chinese chronological histories

Litao Lin and Shiyan Ou

DOI: https://doi.org/10.47989/ir31iConf64279

Abstract

Introduction. This study explores the automatic restructuring of chronological historical texts into historical narratives. It aims to identify and aggregate dispersed event records to reconstruct macro-narrative structures.

Method. We propose a framework combining event extraction and unsupervised clustering. First, an event detection model tailored for classical Chinese is developed. Next, we employ contrastive learning to train a semantic representation model using the thematic text *Tongjian Jishi Benmo*. Finally, unsupervised clustering aggregates vectorised paragraphs into event-specific groups. A mapping dataset linking the chronological *Zizhi Tongjian* to thematic chapters was created for quantitative evaluation.

Results. Experiments indicate that the contrastive learning model combined with the DBSCAN algorithm yields the best performance, with an adjusted rand index (ARI) of 0.43 and normalised mutual information (NMI) of 0.78. The model successfully aggregates semantically related paragraphs, demonstrating an initial capability to transform chronological annals into event-centered accounts.

Conclusions. While precision in event boundaries needs improvement, this research validates the feasibility of automated narrative reconstruction, offering methodological insights for digital historical knowledge organisation.

Introduction

In the millennia-long tradition of Chinese historical writing, the biographical style □ □ □ □ □ and the chronological style □ □ □ □ □ are the two principal forms, the former centres on individuals, presenting history through biographical narratives, while the latter organises records according to the sequence of years and months, providing a clear temporal framework. While the chronological style, exemplified by *Zizhi Tongjian*(*Comprehensive Mirror to Aid in Governance*, □ □ □ □), offers precise timelines, it often fragments complex events across long periods, obscuring causal logic. To address this limitation, the historical-narrative style (□ □ □ □ □), represented by works such as *Tongjian Jishi Benmo*(*Comprehensive Mirror to Aid in Governance with Events in Their Entirety*□ □ □ □ □ □ □ □), gradually emerged in later periods. This style structures around major historical events as its narrative backbone and reorganises the scattered records from chronicles into thematic narratives, thereby helping readers grasp the full context of historical events.

In the digital age, the automated reconstruction of narrative logic has become a key focus in knowledge organisation of histories. Recent studies, such as those by (Zhang et al., 2022), have successfully extracted fine-grained knowledge units (e.g., time, people, places) from the biographical-style histories and reorganised them into chronological style. While these structured representations support multi-dimensional retrieval, they often sacrifice the contextual integrity of original text, making it difficult to reconstruct the complete background of historical events.

Current studies on the knowledge organisation of historical texts predominantly focuses on fine-grained knowledge extraction (such as named entity recognition and relationship extraction) rather than information aggregation at the macro-historical event level. This gap limits the ability of digital systems to present the developmental arc of history. To address this, this study proposes a general framework to automate the extraction of thematic events from chronological style histories. Simulating the transformation logic exemplified by *Tongjian Jishi Benmo*, the proposed method identifies, and aggregates scattered paragraphs related to the same major historical events into coherent thematic clusters.

The specific objectives of this study include: (1) Semantic Representation: To construct a model capable of representing paragraphs describing related events closely in the vector space. space. (2) Dataset Construction: To create a mapping dataset between *Zizhi Tongjian* and *Tongjian Jishi Benmo* for quantitative evaluation. (3) Unsupervised Clustering: To validate the effectiveness of clustering algorithms in automatically discovering thematic structures.

This study shifts the focus of historical knowledge organisation from fine-grained knowledge extraction and knowledge reorganisation to macro-event context restoration, offering a novel pathway for the automated generation of historical-narrative style, which centres on thematic events.

Method and implementation

Overview of the conversion process

To transform linear chronological records into historical narratives, we designed a computational framework shown in Figure 1. The method proceeds in three steps: firstly, extracting valid paragraphs that describing a specific event (hereafter abbreviated as '*event paragraphs*') from source texts; secondly, mapping these paragraphs to a vector space via semantic representation; and thirdly, applying unsupervised clustering to reconstruct event boundaries.

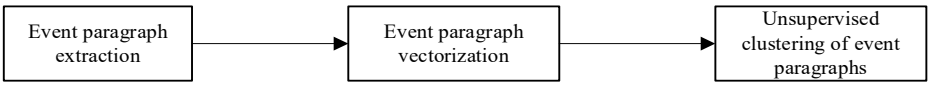

Figure 1. The process of extracting thematic events from chronological style histories

Event paragraph extraction

The goal of event paragraph extraction is to refine historical information by extracting factual descriptions and removing the author's commentary.

Currently, research on event paragraph extraction from ancient Chinese texts primarily focuses on improving and evaluating algorithmic models. These studies are often based on small-scale corpora and use custom sets of event trigger words and event classification systems (Y. Wang et al., 2023; Yu Xuehan et al., 2023; Zhangchao Li et al., 2020). A broad consensus has yet to be reached regarding the definition of an event or its classification system.

To advance research in this area, the 23rd China National Conference on Computational Linguistics (CCL24-Eval) organised an evaluation on historical event type extraction from classical Chinese. This task used the historical event classification system developed by the Computational Linguistics and Culture Research Lab at Beijing Language and Culture University(Wei et al., 2023) as the basis for event detection and indexing.

This team selected The Twenty-Four Histories as their research object. Through comprehensive word frequency statistics and semantic clustering, they refined and built a hierarchical classification system comprising 9 major categories and 67 subcategories. This system covers most event types described in ancient Chinese historical documents and is both highly generalised and finely granular. The system's construction process involved multiple rounds of experimental annotation and evaluation by experts with linguistic and computer science backgrounds, ensuring its accuracy and scientific validity(Wei et al., 2023).

The team also released a corresponding event detection dataset, meticulously annotated by humans, which can be used for training and evaluating event detection models. Since the event classification system used in this dataset was built on all The Twenty-Four Histories, it covers all historical periods of ancient China and has good universality and standardisation. Therefore, we propose using this dataset and event classification system to perform event information extraction from ancient Chinese texts and apply it to this research.

Event paragraph vectorisation

The vectorisation of event paragraphs aims to represent natural language event paragraphs in a vector format that can be understood and computed by a computer. The core goal is to make sentences '*belonging to the same event*' closer in the vector space, while making sentences from different events farther apart. This provides the foundation for the subsequent unsupervised clustering of similar event paragraphs. To achieve this, we propose using a supervised contrastive learning paradigm to train the event paragraph representation. Contrastive learning is a method that has made significant progress in representation learning in recent years. Its basic idea is to pull similar samples closer and push different samples farther apart in the vector space to obtain a more discriminative semantic representation model. Specifically, we borrows from the supervised training mechanism of SimCSE and combines it with the Supervised Contrastive Learning framework proposed by (Khosla et al., 2020). The chapter structure of historical materials like The *Tongjian Jishi Benmo*, which are indexed by topic events, is used as a supervised label. '*Paragraph pairs belonging to the same topic event*' are constructed as positive samples, and '*paragraph pairs from different chapters*' are constructed as negative samples. The training objective is to maximise the cosine similarity between event paragraph vectors from the same event and minimise the similarity between those from different events.

Unsupervised clustering of event paragraphs

The unsupervised clustering of event paragraphs is designed to automatically group semantically related event paragraphs into the same text cluster, which mimics the thematic organisational structure of chronological-style historical texts. The key reason for choosing an unsupervised clustering approach is that the number of independent thematic events in chronological historical

materials is unknown in real-world applications. This makes supervised methods, which require a predefined number of categories, difficult to apply. Unsupervised methods can automatically discover the underlying thematic structure based purely on the semantic features of the text. This provides an objective and flexible technical foundation for converting chronological historical texts into the chronological-style format.

Implementation of the method

Construction of event detection model

Our study selected the Chinese Historical Event Detection dataset (CHED)(lcclab-blcu, 2023/2024) as the training data for the event detection model and the historical event classification system constructed by the Computational Linguistics and Culture Research Lab at Beijing Language and Culture University (Wei et al., 2023) as the basis for event information detection and indexing.

Regarding model selection, the experimental results of (Wei et al., 2023) based on the CHED dataset showed that BERT-BiLSTM-CRF performed excellently. Further experiments by (Litao et al., 2024) on the same dataset indicated that a coarser-grained event annotation system is more conducive to improving the performance of various models. Furthermore, when the GujiBERT (D. Wang et al., 2023) pre-trained model was chosen as the semantic encoder, its event detection performance was superior to the fine-tuned large-scale classical Chinese model, Xunzi(*Xunzi-LLM-of-Chinese-Classics/XunziALLM*, n.d.).

Based on this, this study combines the GujiBERT (D. Wang et al., 2023) pre-trained model to construct a GujiBERT-BiLSTM-CRF event detection model. The model is fine-tuned based on the CHED dataset and the coarse-grained event annotation system. The F1 score on the official CHED validation set is used as the criterion for selecting the optimal model weights. This process ultimately yielded an event detection model with an F1 score of 0.9873, demonstrating a practical level of performance.

Construction of the vectorisation model for the same-event paragraphs

(1) Corpus acquisition and preprocessing

This study selected The *Tongjian Jishi Benmo* as its training data. Compiled by the Southern Song dynasty historian Yuan Shu, this work is a significant historical achievement following Sima Guang's *Zizhi Tongjian*. Its ground-breaking thematic-chronicle style (Jishi Benmo) reorganises the scattered historical facts from *Zizhi Tongjian*, which were originally arranged chronologically—by topic, allowing readers to follow a single historical event from its cause to its course and conclusion. The book meticulously selects and categorises more than a hundred of the most representative major events from *Zizhi Tongjian*, dedicating a separate chapter to each. This rigorous and well-organised structure greatly enhances the readability and comprehension of historical narratives.

The full text of The *Tongjian Jishi Benmo* was acquired from Wikisource (□□□□□□ - □□□□□□□□□, n.d.) using a web crawler. The data was saved while preserving the original paragraph divisions, and each paragraph was annotated with its number and chapter title. The content was then converted to simplified Chinese to ensure compatibility with mainstream Chinese pre-trained models. Following this, the pre-trained event detection model was applied to the text to accurately filter out paragraphs containing explicit event information.

This series of processing steps resulted in 13,580 event paragraphs from 235 thematic event chapters, with an average paragraph length of 153.5 characters. This provided a rich foundational corpus for subsequent model training. As shown in Figure 2, the character count for the vast majority of paragraphs is under 500, indicating that a lightweight model is sufficient for handling

the sentence length requirements of the semantic representation and downstream tasks in this study.

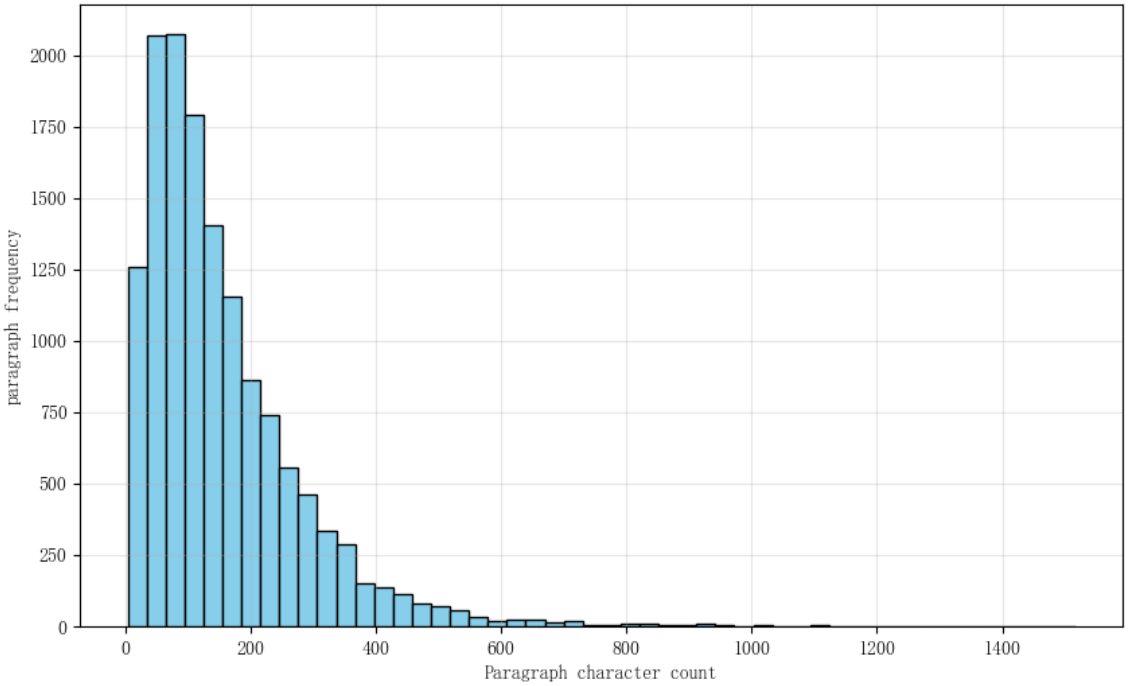

Figure 2. Character distribution of paragraphs in *Tongjian Jishi Benmo*

(2) Model architecture and paragraph vectorisation

To cluster semantically related paragraphs, we employ a supervised contrastive learning paradigm adapted from the SimCSE framework. We utilise the chapter organisation of *Tongjian Jishi Benmo* as weak supervision: paragraphs belonging to the same thematic chapter form positive pairs, while those from different chapters serve as negative samples.

The model is implemented as a Siamese Network using the SentenceTransformer library(Thakur et al., 2021). This dual-tower architecture enables independent sentence encoding, which is essential for efficient large-scale similarity retrieval.

For the underlying encoder, we selected GujiRoBERTa (D. Wang et al., 2023), a model pre-trained on 1.7 billion characters of classical Chinese. This choice is grounded in prior research by (Ye et al., 2024), whose experiments on ancient text intertextuality demonstrated that GujiRoBERTa consistently outperforms other pre-trained models (e.g., GujiBERT, Qwen-7B) in semantic similarity tasks. By adopting this validated backbone, we focus our contribution on the aggregation framework rather than repeating base model comparisons.

(3) Training dataset construction and model fine-tuning

Directly pairing all paragraphs creates severe class imbalance. To address this, we designed a hierarchical sampling strategy to curate a balanced dataset: (1) Positive Sampling: We adopted a 'neighbor-first, distance-weighted' principle. This prioritises adjacent paragraphs (high narrative continuity) and samples non-adjacent pairs based on proximity. (2) Negative Sampling: We combined '*simple*' random pairs with '*hard*' negatives. Hard negatives were selected from different chapters using TF-IDF cosine similarity (threshold >0.2) to challenge the model. (3) Dynamic Balancing: We implemented a curriculum learning approach, linearly increasing the ratio of hard negatives from 0.3 to 0.7 during training.

This process resulted in a final refined dataset of 48,042 positive pairs and 12,233 negative pairs.

Clustering algorithm

To automatically aggregate content-related paragraphs into '*thematic event clusters*'—simulating the *thematic-chronicle* style—we employ unsupervised clustering methods. We conducted a comparative experiment between K-Means and DBSCAN to determine the optimal algorithm and parameter configuration for this task.

Evaluation and results

This section evaluates the framework's ability to identify and aggregate semantically related event paragraphs from dispersed chronological records. The overall evaluation workflow is illustrated in Figure 3. We first detail the construction of the alignment dataset, followed by a quantitative analysis of the clustering results.

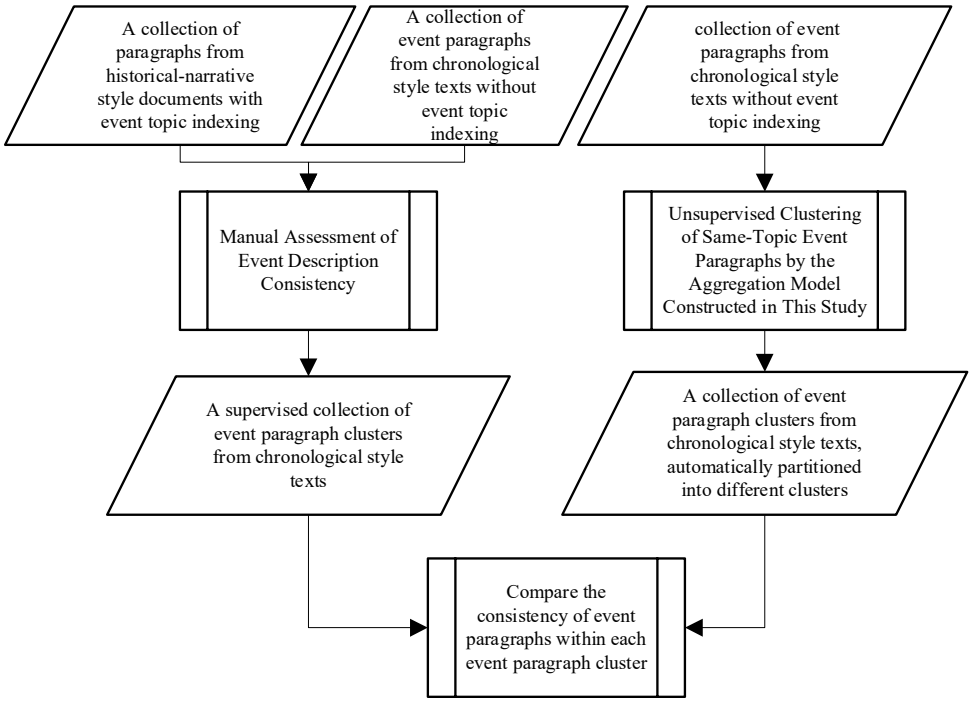

Figure 3. Evaluation process of event paragraph clustering.

Construction of the evaluation dataset

We constructed a labeled dataset to map chronological *Zizhi Tongjian* paragraphs to thematic *Tongjian Jishi Benmo* events. The *Zizhi Tongjian* corpus was obtained from Gushiwen.cn, cleaned of commentary, and annotated with time indices. We first extracted 27,480 event paragraphs from *Zizhi Tongjian* using our detection model. Subsequently, we aligned these with *Tongjian Jishi Benmo* content using the Longest Common Subsequence (LCS) algorithm. By applying a strict similarity threshold of 0.75 and conducting manual verification, we obtained 9,925 high-quality matched pairs. This dataset serves as the ground truth for evaluating the accuracy of unsupervised event clustering.

Evaluation metrics

To comprehensively assess the consistency between unsupervised clusters and manual annotations, we employ three complementary metrics: Adjusted Rand Index (ARI)(Hubert & Arabie, 1985), Normalised Mutual Information (NMI)(Strehl & Ghosh, 2003), and Fowlkes-Mallows Index

(FMI)(Fowlkes & Mallows, 1983). These indices measure clustering performance from diverse perspectives, providing a robust evaluation of the thematic event extraction capability.

Evaluation process

We first encode the *Zizhi Tongjian* evaluation corpus using our *Tongjian Jishi Benmo*-trained SentenceTransformer model. The dataset is split 1:1 into development and test sets. For clustering, we compare K-Means (using the Elbow method for K) and DBSCAN. DBSCAN parameters are optimised via grid search: min_cluster_size in [3, 15], min_samples in [2, 10], and merge_threshold in [0.0, 0.5]. Additionally, we evaluate the impact of UMAP dimensionality reduction10. The optimal configuration identified on the development set is applied to the test set to ensure rigorous comparative evaluation.

Evaluation results

Quantitative scores

To evaluate the performance of different representation models and clustering algorithms in the same-event paragraph clustering task, this study constructed multiple comparative experiments using three semantic modeling approaches: the LDA model, the original GujiRoberta representation, and the GujiRoberta-contrastive-model, a semantic representation model optimised with contrastive learning. These were combined with common clustering methods such as K-Means, DBSCAN, and UMAP dimensionality reduction.

The experimental results from the development set showed that the K-Means algorithm performed best with K=17, while the DBSCAN algorithm was optimal with a minimum cluster size of 15, a number of density estimation samples of 2, and a cluster merge threshold of 0. Table 1 displays the performance of each combination across four metrics: ARI, NMI, FMI, and the number of valid clusters (excluding noise points).

Model	ARI	NMI	FMI	Number of valid clusters
LDA	0.064156	0.485	0.075	198
GujiRoberta-DBSCAN	0.000151	0.021	0.091	4
GujiRoberta-UMAP-DBSCAN	0.01158	0.2	0.082	8
GujiRoberta-DBSCAN	0.146241	0.644	0.27	32
GujiRoberta-contrastive-model-DBSCAN	0.432043	0.782	0.444	145
GujiRoberta-KMEANS	0.000151	0.021	0.091	4
GujiRoberta-UMAP-DBSCAN	0.050584	0.375	0.069	49
GujiRoberta-contrastive-model-KMeans	0.380861	0.764	0.458	49
GujiRoberta-contrastive-model-UMAP-KMeans	0.385319	0.759	0.443	49

Table 1. Evaluation of event clustering performance

Case study and error analysis

Based on a manual review of misclustered samples, we summarised typical failure scenarios. While not statistically exhaustive, these cases (see Table 2) highlight the primary challenges, such as ambiguous temporal markers or complex multi-entity narratives.

Clustering Result	Paragraph in *Tongjian Jishi Benmo*	Paragraph in *Zizhi Tongjian*	Qualitative Analysis
Correct	十一年，秦败韩师于西山。(In the 11th year, Qin defeated the Han army at Xishan.)	显王十一年，秦败韩师于西山。(In the 11th year of King Xian, Qin defeated the Han army at Xishan.)	The time, person, and outcome of the battle are clear, and the expression is concise.
Correct	或谓裕曰：张纲有巧思，若得纲使为攻具，广固必可拔也。(Someone said to Yu, *'Zhang Gang possesses mechanical ingenuity. If we can obtain him and employ him to construct siege engines, Guanggu will surely be captured.')*	安皇帝庚义熙五年，或谓裕曰：'张纲有巧思……'(In the 5th year of the Yixi reign of Emperor An, someone said to Yu, *'Zhang Gang possesses mechanical ingenuity...')*	A dialogue-based paragraph with a clear character and unique context.
Incorrect	文公从之，资苏秦车马，以说赵肃侯曰……(Duke Wen heeded his advice and provided Su Qin with carriages and horses to lobby Marquis Su of Zhao, saying...)	显王三十六年，初，洛阳人苏秦说秦王以兼天下之术……(In the 36th year of King Xian, earlier, Su Qin, a native of Luoyang, lobbied the King of Qin with a strategy for unifying the realm...)	Involves multiple diplomatic figures and arguments, with a large narrative span.
Incorrect	冬十月，南凉王傉檀攻吕隆于姑臧。(In winter, in the tenth month, Rutan, King of Southern Liang, attacked Lü Long at Gusang.)	安皇帝丁元兴元年，南凉王傉檀攻吕隆于姑臧。(In the first year of the Yuanxing reign of Emperor An, Rutan, King of Southern Liang, attacked Lü Long at Gusang.)	Similar military events are described in a similar way, but without a time marker, they are easily confused with other battles.

Table 2. Typical case studies and qualitative analysis of event paragraph clustering.

Clustering success correlates strongly with clear entity identifiers (time, person, place). Misclassification occurs primarily in passages with (1) multi-actor complexity or (2) high textual similarity across chronologically distinct events. Consequently, future work must incorporate external entity knowledge as constraints to enhance the model's discriminative power and precision in boundary detection.

Conclusion and future work

This study proposes a framework for the automated reconstruction of historical narratives from chronological-style historical texts. By facilitating the automatic aggregation of dispersed events in classics such as *Zizhi Tongjian* and Shiji, this approach significantly enhances the efficiency of traditional manual compilation and supports the restoration of historical event sequences. While our experiments confirm the feasibility of unsupervised semantic aggregation, we acknowledge current limitations regarding low cluster purity and imprecise event boundary control. Furthermore, the present study has not yet leveraged state-of-the-art large language models (LLMs), lacks a quantitative analysis of error patterns, and requires further validation across broader historical periods and genres. To address these gaps, future research will aim to integrate advanced LLMs to enhance the understanding of classical Chinese contexts and implicit relationships. Additionally, we plan to explore structured semantic representation methods to capture core event elements, laying a robust foundation for more precise clustering.

Acknowledgement

This work is supported by the project of National Social Science Fund of China, titled '*Research on Multi-source Knowledge Fusion and Multi-dimensional Knowledge Reorganisation of Historical Books and Classics Based on Spatiotemporal Knowledge Graphs*' (Grant No. 24BTQ035) and the project of Postgraduate Research & Practice Innovation Program of Jiangsu Province '*Construction and Influence Analysis of Place Networks Based on Multi-source Historical Documents*' (Grant No. KYCX25_0124).

About the authors

Litao Lin is a PhD student in the School of Information Management at Nanjing University, China, and is currently a Research Fellow at Harvard University, USA. His research interests focus on knowledge organisation and knowledge discovery based on Chinese historical literature. He can be contacted at litaolin@smail.nju.edu.cn.

Shiyan Ou is a Professor in the School of Information Management at Nanjing University, China. She received her Ph.D. from Nanyang Technological University, Singapore. Her research interests include knowledge organisation and scientometrics. She can be contacted at oushiyan@nju.edu.cn.

References

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, 78(383), 553–569. https://doi.org/10.2307/2288117

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218. https://doi.org/10.1007/BF01908075

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, 18661–18673.

lcclab-blcu. (2024). Lcclab-blcu/CHED [Computer software]. https://github.com/lcclab-blcu/CHED (Original work published 2023)

Litao, L., Mengcheng, W., Xueying, S., Jiaxin, Z., & Shiyan, O. (2024). Multi-model classical chinese event trigger word recognition driven by incremental pre-training. In H. Lin, H. Tan, & B. Li (Eds.), Proceedings of the 23rd chinese national conference on computational linguistics (volume 3: Evaluations) (pp. 178–190). Chinese information processing society of China. https://aclanthology.org/2024.ccl-3.20/

Strehl, A., & Ghosh, J. (2003). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res., 3(null), 583–617. https://doi.org/10.1162/153244303321897735

Thakur, N., Reimers, N., Daxenberger, J., & Gurevych, I. (2021). Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks (arXiv:2010.08240). arXiv. https://doi.org/10.48550/arXiv.2010.08240

Wang, D., Liu, C., Zhao, Z., Shen, S., Liu, L., Li, B., Hu, H., Wu, M., Lin, L., Zhao, X., & Wang, X. (2023, July 11). Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. arXiv.Org. https://arxiv.org/abs/2307.05354v1

Wang, Y., Wang, H., Zhu, H., & Li, X. (2023). Research on the Construction of an Event Recognition Model for Historical Antique Books Based on Text Generation Technology. Library and Information Service, 67(3), 119–130. https://doi.org/10.13266/j.issn.0252-3116.2023.03.011

Wei, C., Feng, Z., Huang, S., Li, W., & Shao, Y. (2023). CHED: A Cross-Historical Dataset with a Logical Event Schema for Classical Chinese Event Detection. In M. Sun, B. Qin, X. Qiu, J. Jing, X. Han, G. Rao, & Y. Chen (Eds.), Chinese Computational Linguistics (pp. 289–305). Springer Nature. https://doi.org/10.1007/978-981-99-6207-5_18

Xunzi-LLM-of-Chinese-classics/XunziALLM. (n.d.). Retrieved September 12, 2025, from https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM

Ye, W., Hu, D., Wang, D., Zhou, H., & Liu, L. (2024). Research on Unsupervised Automatic Intertextual Discovery Based on Large Models of Ancient Books. Library and Information Service, 68(23), 41–51. https://doi.org/10.13266/j.issn.0252-3116.2024.23.004

Yu Xuehan, He Lin, & Wang Xianqi. (2023). Research on Event Extraction from Ancient Books Based on Machine Reading Comprehension. Journal of The China Society for Scientific and Technical Information, 42(3), 316–326.

Zhang, Q., Wang, D., Huang, S., & Deng, S. (2022). Multi-Dimensional Knowledge Reorganisation and Visualisation of History Books: Based on Records of the Grand Historian. Journal of the China Society for Scientific and Technical Information, 41(2), 130–141.

Zhangchao Li, Zhongkai Li, & Lin He. (2020). Study on the Extraction Method of War Events in Zuo Zhuan. Library and Information Service, 64(07), 20–29. https://doi.org/10.13266/j.issn.0252-3116.2020.07.003

通鉴纪事本末—维基文库，自由的图书馆. (n.d.). Retrieved May 20, 2025, from https://zh.wikisource.org/zh-hans/通鑑紀事本末