Event Recognition Model Construction for Historical Classical Texts Based on Text Generation Technology

Authors: Wang Yanying, Wang Hao, Zhu Hui, Li Xiaomin, Wang Hao

Date: 2022-08-31T00:00:00+00:00

Abstract

Objective To compare the performance of sequence labeling and text generation methods for event recognition in historical ancient texts, and to construct an event recognition model for such texts. Methods This study selects *Records of the Three Kingdoms* as the raw corpus. For sequence labeling experiments, the event dataset of *Records of the Three Kingdoms* is annotated with BMES tags, and a BBCN-SG model is constructed. For text generation experiments, a T5-SG model is constructed to compare the performance of the two approaches. Additionally, RoBERTa-SG and NEZHA-SG models are built for comparative experiments among generative models. By integrating the three text generation models and incorporating the Stacking ensemble learning paradigm, a Stacking-TRN-SG model is constructed. Results For event recognition modeling in historical ancient texts, text generation methods significantly outperform sequence labeling methods. Among the text generation methods, the three models rank as RoBERTa-SG > T5-SG > NEZHA-SG in performance. Stacking ensemble learning substantially improves the recognition effectiveness of the generative models. Limitations This study is constrained by limited computational resources, and the Stacking-TRN-SG model lacks application research on other historical ancient text corpora. Conclusion The Stacking-TRN-SG model constructed in this paper preliminarily achieves automatic event recognition for historical ancient texts.

Full Text

Research on the Construction of Event Recognition Models for Historical Ancient Texts Based on Text Generation Technology

Yanying Wang[1],[2], Hao Wang[1],[2],*, Hui Zhu[1],[2], Xiaomin Li[1],[2]
[1] School of Information Management, Nanjing University, Nanjing 210023, China
[2] Jiangsu Province Key Laboratory of Data Engineering and Knowledge Services (Nanjing University), Nanjing 210093, China

Abstract

[**Objective**] This study compares the performance of sequence labeling and text generation methods for event recognition in historical ancient texts to construct an effective event recognition model. [**Methods**] Using *Records of the Three Kingdoms* as the corpus, we conducted sequence labeling experiments with BMES annotation to build the BBCN-SG model, and text generation experiments with the T5-SG model. We further compared three text generation models by constructing RoBERTa-SG and NEZHA-SG models, and finally integrated them using Stacking ensemble learning to create the Stacking-TRN-SG model. [**Results**] The text generation approach significantly outperformed sequence labeling for historical ancient text event recognition. Among the text generation models, performance ranked as RoBERTa-SG > T5-SG > NEZHA-SG, with Stacking ensemble learning substantially improving recognition effectiveness. [**Limitations**] Computational resources were limited, and the Stacking-TRN-SG model lacks application studies on other historical corpora. [**Conclusions**] The proposed Stacking-TRN-SG model achieves preliminary automatic event recognition for historical ancient texts.

Keywords: Historical ancient texts; Event recognition; Text generation; Sequence labeling; Ensemble learning
Classification Number: G254

This work was supported by the National Natural Science Foundation of China General Program "Semantic Parsing and Humanities Computing Research on Chinese Intangible Cultural Heritage Texts Driven by Linked Data" (72074108), the Central Universities Basic Research Project "Semantic Analysis and Knowledge Graph Research of Local Gazetteer Texts for Humanities Computing" (010814370113), and talent development programs including Jiangsu Young Social Science Talents and Nanjing University Zhongying Young Scholars.

Corresponding Author: Hao Wang, E-mail: ywhaowang@nju.edu.cn

---

1 Introduction

Historical ancient texts represent a crucial marker of the profound and extensive nature of Chinese civilization. Digitizing and applying these texts enables better cultural heritage preservation, while constructing knowledge graphs from historical ancient texts can visually present history to the public. Event recognition constitutes a vital component of information extraction in knowledge graph construction. However, event recognition in historical ancient texts faces unique challenges: classical Chinese semantics are difficult to comprehend, and the predominance of single characters makes event summarization challenging. Manual event extraction is time-consuming, labor-intensive, and highly subjective, constrained by researchers' knowledge levels and prone to omissions or errors. Therefore, automatic event recognition for historical ancient texts is essential.

This study addresses several key questions: Can machines accurately recognize events in classical Chinese texts like *Records of the Three Kingdoms* that even researchers find challenging? Sequence labeling and text generation represent mainstream event recognition methods, yet comparative studies are lacking. How do these methods perform on classical Chinese, and which is superior? Most sequence labeling methods target named entity recognition with short-distance constraints, whereas event recognition involves summarizing portions of long sentences with numerous long-distance dependencies. Can machines effectively learn these long-distance constraints? For text generation methods, can machines generate events in classical Chinese style? To answer these questions, we investigate event recognition in historical ancient texts using both sequence labeling and text generation techniques to construct an automatic event recognition model.

We selected *Records of the Three Kingdoms* as our corpus and conducted experiments from both perspectives. For sequence labeling, we performed BMES annotation on the event dataset and applied the BERT-BiLSTM-CRF-NER model for training and prediction. For text generation, we utilized the T5 pre-trained model. Since text generation substantially outperformed sequence labeling, we further experimented with RoBERTa and NEZHA models. Finally, we integrated these three text generation models using Stacking ensemble learning to construct the Stacking-TRN-SG model.

2 Related Research

Event recognition [1][2][3] is a critical component of information extraction with established research outcomes. Current domestic and international research primarily categorizes event recognition into rule-based and statistical methods. Rule-based approaches rely on pattern matching, constructing dictionaries beforehand and matching sentences against them according to predefined rules and patterns. These methods achieve high accuracy, as demonstrated by Surdeanu et al. [4][5] who built the FSA system for open-domain event extraction.

However, they heavily depend on dictionaries and suffer from poor portability.

Statistical methods frame event recognition as a classification problem, focusing on classifier selection, construction, and feature engineering. Common techniques include Hidden Markov Models (HMM) [6], Maximum Entropy Models (MEM) [7], Support Vector Machines (SVM) [8], and Conditional Random Fields (CRFs) [9]. These methods require less manual effort, offer greater flexibility, and demonstrate higher portability. For instance, Ahn D. [10] combined MegaM and Timbl machine learning methods to study event type and element identification, achieving favorable results on ACE English corpora. Li Zhangchao et al. applied pattern matching to identify war sentences in *Zuo Zhuan* [19][20][21], but this approach depends on trigger word tables and rule construction, hindering generalizability. Moreover, historical ancient text event recognition faces the characteristic of predominantly single-character words in classical Chinese, with syntax differing from modern Chinese, making rule-based methods unsuitable. Among statistical methods, CRF models offer more flexible feature design and are widely applied in named entity recognition. Classical Chinese named entity recognition [11][12][13][14] primarily focuses on person and location names with short-distance constraints, lacking research on long-distance dependencies. Therefore, we reference named entity recognition methods to investigate the effectiveness of sequence labeling for event recognition with long-distance constraints.

This study transforms event recognition into an abstractive summarization task through text generation methods. In recent years, deep learning has advanced rapidly, with Sequence-to-Sequence (Seq2seq) models [18] achieving significant progress in natural language generation. Cho et al. [15] and Sutskever et al. [16] proposed the Seq2seq model with encoder-decoder architecture, inferring output sequences from input sequence global information. Rush et al. [17] first applied Seq2seq to abstractive summarization, achieving superior results closer to human-generated summaries. Subsequent Seq2seq-based summarization research has flourished, contributing substantially to machine learning.

3.1 Research Framework

Building upon these methods, we selected *Records of the Three Kingdoms* as our historical corpus for event recognition research. The overall experimental framework is illustrated in Figure 1 [Figure 1: see original paper]. Specifically:

First, we compared sequence labeling and text generation methods on the *Records of the Three Kingdoms* event dataset. For the sequence model, we performed BMES annotation on the event dataset, employed the BERT-BiLSTM-CRF-NER base model, retrained it to obtain the BBCN-SG model, and predicted on the test set. For the generation model, we applied the T5 pre-trained model to train on the event dataset, obtaining the T5-SG model for test set prediction. Results demonstrated that text generation substantially outperformed sequence labeling.

Based on this finding, we added RoBERTa and NEZHA pre-trained models for generation experiments, constructing RoBERTa-SG and NEZHA-SG models to compare their performance. Finally, we integrated the three text generation models using Stacking ensemble learning to build the Stacking-TRN-SG model.

3.2 Data Collection and Preprocessing

(1) Data Collection

Experimental data were sourced from chapters 1-30 of *Records of the Three Kingdoms* (the "Book of Wei"). The original text was in traditional Chinese, punctuated with "。", and contained parenthetical annotations. We converted the text to simplified Chinese, replaced "。" punctuation with spaces, and removed parenthetical annotations. Additionally, some characters were unrecognizable due to historical evolution; we replaced them based on comprehensive consideration of websites such as "Gushiwen.cn" and "Shidaquan.com". For example, "士卒无 Z00050 志" was replaced with "士卒无斗志". Table 1 demonstrates the text processing using the first paragraph of Chapter 1, "Biography of Emperor Wu", with the Gushiwen.cn version provided as reference.

(2) Data Preprocessing

We manually extracted original sentences containing events and generated abstractive summaries. Notably, for sequence model experiments, we ensured summary characters were entirely sourced from the original text. Processed data are shown in Table 2, where "source" represents the extracted original text and "target" represents its summary version (a subset of source). The final dataset comprised 671 instances, split 500:171 into training and test sets.

Generation models use source-target pairs directly, while sequence models require conversion to annotated data. We annotated source data according to target data using BMES tagging, with examples shown in Table 3. Annotation rules are: First character in target → "B" in source; Last character in target → "E" in source; Other characters → "M"; Characters absent from target → "S"; If first/last characters appear multiple times, all instances are tagged "B" or "E"; Sentence start/end positions are determined by target rather than source occurrence. For model training, each character and its tag occupy one line, with blank lines separating instances.

3.3 Experimental Methods

(1) Model Selection

The primary challenge is the small size of the *Records of the Three Kingdoms* event dataset, making direct training difficult. Therefore, we selected BERT-BiLSTM-CRF-NER[3] as the base sequence labeling model and T5, RoBERTa, and NEZHA pre-trained models as text generation bases.

BERT (Bidirectional Encoder Representations from Transformers) [22] is a bidirectional encoder based on Transformer architecture, essentially the encoder portion of Transformer. Its pretraining tasks include Masked Language Model (MLM) and Next Sequence Prediction (NSP), trained on 800M words from BooksCorpus and 2500M words from English Wikipedia. This enables easy fine-tuning for downstream tasks with minimal sample data and computational power. BERT has achieved excellent results across NLP tasks, establishing the pretrain+fine-tune paradigm that sparked numerous subsequent pre-trained models. The three text generation models used in this study are optimized improvements upon BERT, fine-tuned on the *Records of the Three Kingdoms* event dataset.

T5 (Text-To-Text Transfer Transformer) [23][24][25][26][27][28] is a Google pretrained model from October 2019 that innovatively frames all NLP tasks as "Text-To-Text" problems. By adding different prefixes to inputs, it handles machine translation, classification, similarity, and summarization through a unified generative approach. Unlike BERT' s encoder-only architecture, T5 employs both encoder and decoder from the original Transformer and uses relative position encoding. It trains on 750GB C4 corpus (Colossal Clean Crawled Corpus), adopts span masking like SpanBERT, increases training steps to 1M, and uses multi-task training with some supervised data in unsupervised corpora.

RoBERTa (A Robustly Optimized BERT Pretraining Approach) [29] is a Facebook-University of Washington model from July 2019. Improvements include dynamic masking (similar to cross-validation) versus BERT' s static masking, removal of NSP objective with FULL-SENTENCES/DOC-SENTENCES construction, expanded 160GB corpus, byte-level BPE encoding versus character-level, and hyperparameter optimization with larger batch sizes and iterations.

NEZHA (Neural Contextualized Representation for Chinese Language Understanding) [30] is a Huawei Noah' s Ark Lab pre-trained model from September 2019 for Chinese NLP tasks, featuring four key improvements: functional relative position encoding using sinusoidal functions of relative positions, whole-word masking instead of random masking, mixed-precision training with single and half precision for acceleration, and LAMB optimizer usage.

(2) Ensemble Learning

Ensemble learning [31][32] was first proposed by Dasarathy and Sheela in 1979 and has become a vital machine learning branch. Three classic algorithms dominate: Bagging [33], Boosting [34], and Stacking [35]. Bagging generates multiple training subsets via bootstrap sampling, trains different classifiers, and combines results through voting. Boosting converts weak learners to strong ones through iterative training. Stacking trains multiple heterogeneous base classifiers on the same dataset, then trains a meta-classifier using their outputs. We combined Stacking and Bagging concepts, integrating three generation model results through voting.

3.4 Model Construction

Experiments comprised three components: sequence model construction, generation model construction, and generation model ensemble learning.

The sequence model employed BERT-BiLSTM-CRF-NER as the base, fine-tuned on the *Records of the Three Kingdoms* event dataset to obtain BBCN-SG for test set prediction. Model performance was evaluated by directly comparing predicted and target tags. For application, predicted tags were reverse-converted to summaries using BMES rules. Key parameters: max_{length}=128, learning_{rate}=1e-5, batch_{size}=16, epochs=10.

Generation models included T5, RoBERTa, and NEZHA pre-trained models, similarly fine-tuned to build T5-SG, RoBERTa-SG, and NEZHA-SG models. Parameters were unified across models: max_{length}=256, learning_{rate}=1e-5, batch_{size}=16, epochs=20. After training, we obtained 171$×$3 predictions on the test set, compared them with target summaries using ROUGE and BLEU metrics, and performed Stacking ensemble learning.

3.5 Evaluation Metrics

Manual evaluation of event recognition is subjective and time-consuming. Therefore, we adopted mainstream quantitative metrics. Sequence models were evaluated using seqeval[4] by directly comparing target and predicted tags. Generation models were assessed using ROUGE and BLEU metrics, with final scores averaged across all predictions.

(1) Sequence Model Metrics

Using BMES annotation, we calculated accuracy, precision, recall, and F1-score by comparing predicted and target tags. Based on the confusion matrix:

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Where N represents total samples, TP are true positives, TN true negatives, FP false positives, and FN false negatives.

(2) BLEU Metric

BLEU (Bilingual Evaluation Understudy), commonly used for machine translation, assesses prediction quality by measuring N-gram overlap between generated and reference texts. Higher overlap indicates better predictions. The formula is:

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \times \log P_n\right)$$

where P_n is N-gram precision (matching N-grams divided by total N-grams) and w_n is weight (typically $1/N$). The BP penalty factor addresses length issues:

$$BP = \begin{cases} \exp\left(1 - \frac{l_r}{l_c}\right) & \text{if } l_c \leq l_r \\ 1 & \text{otherwise} \end{cases}$$

where l_c is generated text length and l_r is reference length. Given classical Chinese' s single/double-character nature and short event lengths, we used N=1,2.

(3) ROUGE Metric

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [36] emphasizes recall over BLEU' s precision focus. It comprises ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. We employed ROUGE-1, ROUGE-2, and ROUGE-L.

ROUGE-N calculates N-gram recall:

$$\text{ROUGE-N} = \frac{p}{q}$$

where p is the count of overlapping N-grams and q is total N-grams in the reference. We used N=1,2.

ROUGE-L uses Longest Common Subsequence (LCS):

$$R_{lcs} = \frac{\text{LCS}(C, S)}{\text{len}(S)}$$

$$P_{lcs} = \frac{\text{LCS}(C, S)}{\text{len}(C)}$$

$$F_{lcs} = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

where C is generated text, S is reference text, and β is typically large (making F_{lcs} approximate R_{lcs} when $\beta \to \infty$).

4.1 Comparison of Sequence and Generation Models

Sequence labeling and text generation are mainstream event recognition methods lacking detailed comparative studies. We conducted comparative experiments on the *Records of the Three Kingdoms* event dataset from a classical Chinese perspective. The sequence model used BERT-BiLSTM-CRF-NER (BBCN-SG) and the generation model used T5 (T5-SG), both fine-tuned and evaluated. Results appear in Table 4 .

The sequence model outputs BMES tags, enabling direct comparison with target tags. The generation model outputs text, evaluated via BLEU and ROUGE. Table 4 shows recall, precision, and F1, with ROUGE-1 for the generation model. The sequence model exhibited lower recall but higher precision and F1. However, BMES tagging is essentially a binary classification problem (only 'B' , 'E' , 'M' , and 'S' tags), so high accuracy doesn' t necessarily indicate good recognition quality, and the accuracy itself wasn' t particularly high for binary classification. Therefore, metric analysis alone cannot determine superiority; semantic accuracy and coherence must be considered.

Sequence model predictions require tag-to-text conversion, which presents challenges. According to conversion rules, 'B' marks the first character and 'M' marks middle characters. When 'M' appears before 'B' , the order becomes ambiguous. As shown in Table 5 , 'B' marks "吴" and the first 'M' marks "年" . Following conversion rules, "年" should appear mid-sentence, but its exact position is uncertain, and semantically, "年" disrupts coherence.

Table 6 compares model predictions. Sequence models perform poorly on long-distance constraints, producing incoherent, incomprehensible outputs that fail to effectively recognize historical events. Generation models maintain semantic coherence even for lengthy classical Chinese. For medium-length texts (instances 4-7), both models produce roughly relevant events, but sequence model outputs lack coherence and are difficult to understand independently. Generation models occasionally produce inaccurate meanings (e.g., instance 5) but maintain understandable statements. For short texts, both models perform reasonably, but generation models are clearly superior. Additionally, generation models recognize multi-character terms like "长史" (Chief Clerk), "合肥" (Hefei), and "袁绍" (Yuan Shao), while sequence models remain character-based. Thus, generation models excel in semantic accuracy and coherence.

In summary, sequence labeling performs worse than text generation on *Records of the Three Kingdoms* event recognition.

4.2 Generation Model Comparison

Comparing generation models revealed superior performance, prompting us to add RoBERTa and NEZHA models fine-tuned on the dataset. Table 7 presents BLEU and ROUGE scores.

BLEU emphasizes precision: RoBERTa-SG scored highest, followed by T5-SG,

with NEZHA-SG slightly lower. ROUGE precision followed the same pattern. However, ROUGE emphasizes recall, where T5-SG achieved the highest scores across ROUGE-1, ROUGE-2, and ROUGE-L, followed by RoBERTa-SG, then NEZHA-SG. As shown in Figure 2 [Figure 2: see original paper], T5-SG excels in recall, RoBERTa-SG in precision, and NEZHA-SG lags behind. Summing all metrics, RoBERTa-SG ranked highest overall: RoBERTa-SG > T5-SG > NEZHA-SG.

4.3 Ensemble Learning for Generation Models

Although RoBERTa-SG performed best overall, T5-SG achieved the highest recall. We therefore examined individual instances across the three models.

As Table 8 shows, no single model consistently outperformed others; they exhibited complementarity. Using BLEU-2 as the criterion, instance 1 favored NEZHA-SG, instance 2 RoBERTa-SG, instance 3 T5-SG, and instance 4 showed ties. This suggested Stacking ensemble learning could significantly improve performance. Stacking learns multiple heterogeneous weak learners in parallel and combines them via a meta-model. Since the three generation models have different architectures, we adopted a voting approach inspired by Bagging. Using BLEU metrics as the voting standard, we ranked predictions and selected the highest-scoring result as the final Stacking output. Given BLEU-2' s stricter requirements than BLEU-1, we applied a three-tier sorting: BLEU-2 first; BLEU-1 if tied; ROUGE if both tied, following T5-SG > RoBERTa-SG > NEZHA-SG order due to T5-SG' s highest recall.

Comparing integrated predictions with target texts, we calculated BLEU and ROUGE scores for each instance and averaged them as the Stacking-TRN-SG final score. Table 9 shows all metrics exceeded those of the three base models, demonstrating substantial improvement.

Table 10 presents ensemble examples. Stacking-TRN-SG selects the optimal output among the three, yielding more accurate and coherent results. Whether for long or short texts, Stacking-TRN-SG produces understandable, appropriately length events that preserve classical Chinese characteristics. However, some semantic inaccuracies remain, requiring further research.

5 Conclusion

Historical ancient texts are vital carriers of Chinese culture. Constructing knowledge graphs from them can visually present history, making automatic event recognition essential. Classical Chinese characteristics—predominantly single characters and difficult semantics—have limited event recognition research. Additionally, while sequence labeling and text generation are mainstream methods, detailed comparative studies are scarce.

This study selected *Records of the Three Kingdoms* for experiments in both sequence labeling and text generation. Sequence labeling used BMES annotation

with BERT-BiLSTM-CRF-NER, while text generation employed T5. BLEU and ROUGE metrics quantified results, supplemented by semantic accuracy and coherence assessment. Text generation substantially outperformed sequence labeling. RoBERTa and NEZHA models were added, with overall performance ranking RoBERTa-SG > T5-SG > NEZHA-SG but showing complementarity. Integrating the three models via Stacking ensemble learning created the Stacking-TRN-SG model, which significantly improved recognition. The model produces semantically coherent, accurate events that reflect classical Chinese characteristics, achieving 70.35% recall and preliminary automatic event recognition for historical ancient texts.

Limitations include constrained computational resources and lack of application studies on other historical corpora.

References

[1] E. Haihong, Z. Wenjing, X. Siqi, C. Rui, H. Yingxi, Z. Xiaosong, N. Peiqing. A Survey on Deep Learning for Entity Relation Extraction [J]. *Journal of Software*, 2019, 30(06): 1793-1818. DOI:10.13328/j.cnki.jos.005817.

[2] Z. Yanyan, Q. Bing, C. Wanxiang, L. Ting. Research on Chinese Event Extraction Technology [J]. *Journal of Chinese Information Processing*, 2008(01): 3-8.

[3] J. Jifa. Research on Information Extraction Pattern Acquisition from Free Text [D]. Doctoral Dissertation, Chinese Academy of Sciences, 2004: 1-18.

[4] M. Surdeanu, S. Harabagiu, J. Williams, et al. Using Predicate-Argument Structures for Information Extraction [A]. In: *Proceedings of ACL* [C]. 2003. 8-15.

[5] M. Surdeanu, S. Harabagiu. Infrastructure for Open-Domain Information Extraction [A]. In: *Proceedings of the Human Language Technology Conference* [C]. 2002. 325-330.

[6] H. Zhou, J. Chen, G. Dong, et al. Detection and Diagnosis of Bearing Faults Using Shift-invariant Dictionary Learning and Hidden Markov Model [J]. *Mechanical Systems & Signal Processing*, 2015(72-73).

[7] D. Luwei, R. Song. Preliminary Study on Automatic Recognition of Missing Topics in Chinese Punctuated Sentences Based on Maximum Entropy Model [J]. *Computer Engineering & Science*, 2015, 37(12).

[8] L. Shuang, H. Degen, M. Tingting, et al. Automatic Recognition of Chinese Names Based on Support Vector Machines [J]. *Computer Engineering*, 2006, 32(19): 188-190.

[9] L. Wu, L. Liu, H. Li, et al. Chinese Place Name Recognition Method Based on Conditional Random Fields [J]. *Geomatics and Information Science of Wuhan University*, 2017.

[10] D. Ahn. The Stages of Event Extraction [C] // *Proceedings of the COLING-ACL 2006 Workshop on Annotating and Reasoning About Time and Events*, 2006: 1-8.

[11] Z. Niuniu, J. Meng, G. Jianling, C. Yaxian. Chinese Named Entity Recognition Method Based on BERT [J]. *Computer Science*, 2019, 46(S2): 138-142.

[12] R. Zhihui, X. Haoyu, F. Songlin, Z. Han, S. Jun. Chinese Word Segmentation Based on LSTM Network for Sequence Labeling [J]. *Application Research of Computers*, 2017, 34(05): 1321-1324+1341.

[13] C. Wei, W. Youzheng, C. Wenliang, Z. Min. Automatic Keyword Extraction Based on BiLSTM-CRF [J]. *Computer Science*, 2018, 45(S1): 91-96+113.

[14] T. Huihui, W. Hao, Z. Zixuan, W. Xueying. Research on Chinese Historical Event Name Extraction Based on Character Annotation [J]. *Data Analysis and Knowledge Discovery*, 2018, 2(07): 89-96.

[15] K. Cho, B. van Merriënboer, C. Gulcehre, et al. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation [C] // *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 2014: 1724-1734.

[16] I. Sutskever, O. Vinyals, Q. V. Le. Sequence to Sequence Learning with Neural Networks [C] // *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Cambridge: MIT Press, 2014: 3104-3112.

[17] A. M. Rush, S. Chopra, J. Weston. A Neural Attention Model for Abstractive Sentence Summarization [C] // *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 2015: 379-389.

[18] L. Shi, R. Ruanxuan, W. Ruibin, C. Ying. A Survey on Abstractive Text Summarization Based on Sequence-to-Sequence Models [J]. *Journal of the China Society for Scientific and Technical Information*, 2019, 38(10): 1102-1116.

[19] Z. Liu, J. Dangfei, Z. Zhijian. Research on Automatic Extraction of Historical Events and Construction of Event Logic Graphs from *Records of the Grand Historian* [J]. *Library and Information Service*, 2020, 64(11): 116-124. DOI:10.13266/j.issn.0252-3116.2020.11.013.

[20] Z. Li, Z. Zhongkai, H. Lin. Research on War Event Extraction Technology for *Zuo Zhuan* [J]. *Library and Information Service*, 2020, 64(07): 20-29. DOI:10.13266/j.issn.0252-3116.2020.07.003.

[21] Y. Xuehan, H. Lin, X. Jian. Research on Ancient Chinese Historical Event Extraction Method Based on RoBERTa-CRF [J]. *Data Analysis and Knowledge Discovery*, 2021, 5(07): 26-35.

[22] J. Devlin, M. W. Chang, K. Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C] // 2018.

[23] C. Raffel, N. Shazeer, A. Roberts, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [J]. *J. Mach. Learn. Res.*, 2020, 21(140): 1-67.

[24] L. Xue, N. Constant, A. Roberts, et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer [J]. arXiv preprint arXiv:2010.11934, 2020.

[25] N. Shazeer. Glu Variants Improve Transformer [J]. arXiv preprint arXiv:2002.05202, 2020.

[26] Y. N. Dauphin, A. Fan, M. Auli, et al. Language Modeling with Gated Convolutional Networks [C] // *International Conference on Machine Learning*. PMLR, 2017: 933-941.

[27] H. W. Chung, T. Fevry, H. Tsai, et al. Rethinking Embedding Coupling in Pre-trained Language Models [J]. arXiv preprint arXiv:2010.12821, 2020.

[28] M. Joshi, D. Chen, Y. Liu, et al. SpanBERT: Improving Pre-training by Representing and Predicting Spans [J]. *Transactions of the Association for Computational Linguistics*, 2020, 8: 64-77.

[29] Y. Liu, M. Ott, N. Goyal, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [J]. arXiv preprint arXiv:1907.11692, 2019.

[30] J. Wei, X. Ren, X. Li, et al. NEZHA: Neural Contextualized Representation for Chinese Language Understanding [J]. arXiv preprint arXiv:1909.00204, 2019.

[31] J. Xu, Y. Yang. A Survey on Ensemble Learning Methods [J]. *Journal of Yunnan University (Natural Sciences Edition)*, 2018, 40(06): 1082-1092.

[32] X. Zhou, D. Lixin, W. Runze, G. Qiang. Research on Classifier Ensemble Algorithms [J]. *Wuhan University Journal (Natural Science Edition)*, 2015, 61(06): 503-508. DOI:10.14188/j.1671-8836.2015.06.001.

[33] L. Breiman. Bagging Predictors [J]. *Machine Learning*, 1996, 24(2): 123-140.

[34] Y. Freund, R. E. Schapire. A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting [J]. *Journal of Computer and System Sciences*, 1997, 55(1): 119-139.

[35] D. H. Wolpert. Stacked Generalization [J]. *Neural Networks*, 1992: 241-259.

[36] C. Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries [C] // *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*. 2004.

Author Contributions

Yanying Wang: Conducted experiments, drafted and revised the manuscript.
Hao Wang: Conceived research idea, supervised experiments, proposed framework, guided revisions.
Hui Zhu: Guided manuscript revisions.
Xiaomin Li: Guided manuscript revisions.

Note: Figure translations are in progress. See original paper for figures.