Postprint: War Event Extraction Techniques for Zuo Zhuan

Authors: Zhangchao Li, Li Zhongkai, He Lin

Date: 2023-04-01T16:15:51+00:00

Abstract

[Purpose/Significance] This study investigates war events in Zuo Zhuan, which holds significant reference value for research on pre-Qin history and Chinese national culture. [Methods/Process] Based on frame theory, we construct a fundamental framework system for war events in Zuo Zhuan. Pattern matching is employed for war sentence identification. A conditional random field model combined with feature templates is utilized to recognize and extract seven named entities, including war time and warring parties. Finally, war events are analyzed and visualized based on the obtained structured data. [Results/Conclusion] The results demonstrate that the conditional random field model can be effectively applied to war event extraction from Zuo Zhuan. Feature selection influences entity recognition outcomes. Regarding specific content, states including Jin, Chu, Qi, and Zheng exhibited high participation frequencies during the Spring and Autumn period, with Jin serving as the primary aggressor and Zheng as the primary defender.

Full Text

Preamble

ChinaXiv Collaborative Journal
Volume 64, Issue 7, April 2020

Research on War Event Extraction Technology in *Zuo Zhuan*

Li Zhangchao, Li Zhongkai, He Lin
Department of Information Management, Nanjing Agricultural University, Nanjing 210095

Abstract: [Purpose/Significance] This research focuses on war events in *Zuo Zhuan*, offering significant reference value for studies of pre-Qin history and Chinese national culture. [Method/Process] Based on frame theory, we construct a basic framework system for war events in *Zuo Zhuan*, utilize pattern

matching methods for war sentence identification, and employ a Conditional Random Field (CRF) model combined with feature templates to recognize and extract seven named entities including war time and warring parties. Finally, we analyze and visualize the war events based on the structured data obtained. [Result/Conclusion] The results demonstrate that the CRF model can be effectively applied to war event extraction from *Zuo Zhuan*; feature selection affects entity recognition performance. In terms of specific content, the states of Jin, Chu, Qi, and Zheng participated most frequently in wars during the Spring and Autumn period, with Jin as the primary aggressor and Zheng as the main defender.

Keywords: *Zuo Zhuan*; war events; event extraction
Classification Number: G255
DOI: 10.13266/j.issn.0252-3116.2020.07.003

Classics refer to ancient Chinese texts of exceptional value that serve as crucial carriers of Chinese culture and symbols of five millennia of Chinese civilization. In the digital era, numerous classics have been digitized and made publicly available online, primarily through two forms: digital preservation of ancient books (original scanning and replication) and digital collation of ancient books. Digital preservation represents the most common approach and constitutes foundational work in digital humanities. However, this method merely stores classics in digital format, which does not facilitate retrieval, acquisition, information processing, or in-depth research of ancient materials. Meanwhile, existing research predominantly focuses on modern Chinese, with relatively few studies on classical Chinese. Moreover, due to significant differences in vocabulary, syntax, and linguistic structures between modern and classical Chinese, targeted research on classical Chinese is essential.

The emergence of the fourth paradigm of scientific research also poses new questions for digital humanities: Can we effectively organize the natural language in classics using new technologies such as entity knowledge mining to provide comprehensive and accurate classic information for historical research? Reviewing existing studies and relevant literature, we find that research on ancient classics exhibits the following characteristics: (1) In terms of research content, numerous studies in classical Chinese and historical fields have explored usage patterns and construction rules for entities in classics, covering various aspects of ancient politics, economy, society, and military affairs. For instance, Huang Shuiqing et al. applied the Conditional Random Field (CRF) model to automatically identify ancient Chinese place names in *Zuo Zhuan* and *Guoyu* [3]. (2) Regarding research methods, the approach has gradually transitioned from manual processing to computer-based natural language processing of classics, transforming paper resources into digital resources and constructing large-scale corpora for research use. For example, C.L. Liu et al. mined biographical information from over 220 Chinese local gazetteers using language models and CRF models [4]; Qian Zhiyong et al. conducted automatic word segmentation and annotation experiments on *Chu Ci* using Hidden Markov Models [5].

Zuo Zhuan, China's first narrative chronicle, possesses exceptional historical and literary value. Meanwhile, war represents a typical manifestation of intensified internal and external contradictions in specific historical stages, reflecting the political, economic, and cultural elements of the time. Therefore, this paper focuses on war event extraction from *Zuo Zhuan*, specifically including: (1) war event knowledge representation based on frame theory, (2) event sentence extraction using rule-based methods, (3) entity annotation research, (4) automatic entity recognition using sequence labeling methods, and (5) war event visualization demonstration.

2 Literature Review

2.1 Overview of Event Extraction Research

Event extraction, a crucial component of information extraction, employs computer technology to extract parameters related to specific events, event elements, or relationships from natural language texts, including named entity recognition and relation extraction [7]. Currently, event extraction primarily utilizes pattern matching and machine learning methods. Pattern matching identifies and extracts specific event types based on patterns, matching sentences and templates according to corresponding algorithms, offering high accuracy and domain specificity. For example, M. Surdeanu et al. developed the open-domain event extraction system FSA. In machine learning approaches, the principle involves selecting and constructing classifiers for classification [8], mainly including event type identification and event element recognition (slots in event templates and event participants). In existing research, L.C. Chieu et al. introduced maximum entropy classifiers into event extraction, facilitating event element identification [9]; D. Ahn applied MegaM and Timbl machine learning methods to study event type and element recognition, effectively processing ACE English corpora [10]; Zhao Yanyan et al. extracted candidate events from texts based on trigger word sets, employed binary classifiers to select appropriate candidate events, and introduced maximum entropy classifiers for event recognition [8]; Yu Jiangde et al. studied Chinese text event extraction using Hidden Markov Models, extracting candidate event statements through trigger word detection and constructing HMMs based on features of each event element type [11], then extracting event elements from statements using these models.

Additionally, scholars from various fields have conducted event extraction research combined with their backgrounds. For instance, Wu Pingbo et al. focused on extracting key information from network events, formulated information extraction rules using sentence patterns, identified events to be extracted based on rules, and extracted high-quality event information through time phrase recognition and basic phrase identification, enabling segmentation of different events [12]; Jiang Jifa studied role information extraction for casualty personnel in disaster events based on "HowNet," proposing a cross-sentence Chinese event information extraction method that achieved high recall and accuracy rates [13]; Zheng Jiaheng et al. researched acquisition methods for crop variety description

patterns, discovering an inverse relationship between research object scale and result accuracy [14]; Yang Erhong studied information acquisition from breaking news reports, constructing emergency event information extraction models based on contextual, part-of-speech, tagging, and indicator word features of event texts to facilitate extraction of specific information and structures [15].

2.2 Progress in Ancient Chinese Text Processing

Against the backdrop of rapid digitization of classics, the maturation of natural language processing technology has advanced ancient Chinese text processing [16]. Ancient Chinese text processing involves using information technology to process the phonology, morphology, and semantics of classical Chinese, enabling deep mining and knowledge discovery [5]. In terms of content, ancient Chinese information resources are primarily obtained through digitization, recording and storing classics in computer-readable media. Methodologically, with continuous development of computer technology, scholars have begun introducing machine learning methods into digital humanities to process and analyze ancient Chinese texts, mainly including word segmentation and named entity recognition for ancient Chinese.

Word segmentation is fundamental and critical for ancient Chinese information processing using technical methods. Text segmentation approaches primarily include rule-based methods and statistical machine learning methods. Rule-based methods suit structured texts with known sentence pattern features but yield unsatisfactory results for unstructured texts. Consequently, more scholars opt for statistical machine learning methods for text segmentation, referencing modern text segmentation approaches while incorporating lexicons such as place name tables, personal name tables, and commentary word tables to assist computers [17]. Automatic segmentation using machine learning has achieved good results and been applied to many ancient texts, such as *Chu Ci* [4] and *Mencius* [18].

Research on named entity recognition for ancient Chinese has also attracted increasing scholarly attention, laying a solid foundation for knowledge mining from ancient texts. Ancient Chinese named entity recognition primarily includes personal names and place names, with the CRF model being most widely used and achieving satisfactory results in classics such as *Romance of the Three Kingdoms* [19] and *Spring and Autumn Annals* [2].

2.3 Review

In recent years, event extraction and ancient Chinese text processing have garnered widespread attention from scholars, progressing from digitization of classics to intelligent processing of ancient Chinese, achieving promising results in automatic segmentation and named entity recognition. Currently, some studies have utilized pattern matching and machine learning methods to process and analyze ancient Chinese texts, but subsequent research suffers from insufficient specificity and applicability. Therefore, this paper constructs a basic framework

system for war events in *Zuo Zhuan*. Based on pattern matching methods, we first construct a trigger word table to filter and obtain candidate war sentence sets, then establish a series of rules to extract war sentences from the candidate set, thereby constructing a *Zuo Zhuan* war sentence corpus. Simultaneously, according to the previously constructed war event framework and based on the CRF model, we conduct multiple entity recognition experiments by combining features of context window length, part-of-speech, and indicator words in *Zuo Zhuan* war sentences, selecting optimal solutions to extract seven entities: war time, attacker, defender, war location, war trigger cause, war result, and reinforcements. Finally, we analyze and visualize the data using statistical methods and E-Charts tools.

3 Research Framework

3.1 Overall Research Approach

Based on the aforementioned analysis and summary, our research framework is shown in Figure 1 [Figure 1: see original paper]. First, we construct the *Zuo Zhuan* war event framework based on *Zuo Zhuan* corpora for structured description and knowledge representation of war events. In the event extraction stage, we use the more controllable pattern matching method for war sentence identification. By constructing a trigger word table and matching it to obtain preliminary event sentence sets, we achieve event sentence extraction through rule base matching. Second, we conduct named entity recognition and extraction. Through observation and analysis of the extracted war event sentence set and considering features such as context window length, part-of-speech, and indicator words in *Zuo Zhuan* texts, we manually annotate the war event sentence set based on a five-position tagging system and use the CRF model to recognize and extract named entities.

3.2 Construction of *Zuo Zhuan* War Event Framework

Historical research identifies the main components of war events as war time, warring parties (attacker and defender), war location, war trigger cause, and war result. Through reading *Zuo Zhuan*, we find that war descriptions in the text also encompass these elements. Additionally, *Zuo Zhuan* descriptions include rescue events. Therefore, we categorize wars into two major types: expeditionary and rescue, adding the "reinforcements" entity to rescue-type war events for more specific and complete description of war events in *Zuo Zhuan*.

Accordingly, we construct the basic information framework for *Zuo Zhuan* war events, as shown in Figure 2 [Figure 2: see original paper].

Example 1: "In the spring of the tenth year, the Qi army attacked us. We fought at Changshao, and the Qi army was defeated."

Based on this framework, we extract information for the expeditionary-type war event in Example 1, with the resulting information framework shown in Figure

3 [Figure 3: see original paper].

Example 2: "In autumn, the Viscount of Chu besieged Xu to rescue Zheng. The feudal lords rescued Xu, and then returned."

Based on this framework, we extract information for the rescue-type war event in Example 2, with the resulting information framework shown in Figure 4 [Figure 4: see original paper].

4 Key Technologies

4.1 War Sentence Identification

Currently, common event sentence extraction methods include pattern matching and machine learning approaches. Pattern matching suits texts with short event sentences and small total corpora. This method extracts event sentences through feature matching, performing syntactic analysis on target texts based on linguistics to identify patterns of target thematic sentences and their differences from other thematic sentences. Trigger words represent important features expressing differences between event sentences. By constructing a trigger word table to locate sentences containing these words and, while prioritizing recall rate, eliminating non-compliant sentences through rule formulation, we ultimately obtain the *Zuo Zhuan* war event thematic sentence set.

Given the complexity of syntactic features in *Zuo Zhuan* corpora, we conduct pattern matching based on principles of completeness, specificity, and feasibility through these steps: (1) Construct the *Zuo Zhuan* war event trigger word table. War event descriptions in *Zuo Zhuan* follow certain patterns, allowing rapid location of target sentences through special war-related words such as "attack" and "invade." We compile and organize these special words to create the trigger word table. (2) Locate sentences containing trigger words and extract candidate sentences. Using the trigger word table, we identify *Zuo Zhuan* sentences containing specific trigger words and extract them as candidate war sentences. (3) Eliminate non-war sentences. Finally, based on established patterns and principles, we filter out quasi-war sentences from the war and quasi-war sentence set obtained in step two.

4.2 Trigger Word Table Construction

Trigger verbs are special verbs pre-summarized before pattern extraction, serving as the key and foundation for extracting war-related elements. Trigger words can narrow text scope and improve efficiency and specificity of rule formulation. Based on Zhang Qiuxia's research on expeditionary verbs in *Zuo Zhuan*, we categorize them into nine types: mobilization, engagement, attack, leadership, harassment, killing, defense, capture, and outcome, as shown in Table 1 .

Based on reading *Zuo Zhuan* and summarizing trigger words in war sentences, we simplify Zhang Qiuxia's nine categories by removing words not used independently in war descriptions or used concurrently with other expeditionary

words (e.g., “mobilize,” “lead,” “kill,” “slay”). Additionally, referencing Deng Yong et al.’s definitions of war events in *Zuo Zhuan*, we add two categories: betrayal and rescue [21], obtaining the following trigger word table: “raid/v, attack/v, camp/v, battle/v, rebel/v, capture/v, meet/v, assault/v, punish/v, surrender/v, pursue/v, gate/v, enter/v, rescue/v, follow/v, invade/v, defeat/v, encroach/v, conquer/v, assist/v, besiege/v, besiege/v, destroy/v, array/v, chase/v, pillage/v.” After establishing the trigger word table, we preliminarily identify the entire *Zuo Zhuan* corpus, extracting sentences containing trigger words with verb part-of-speech, ultimately obtaining a set containing both war and quasi-war sentences.

4.3 Named Entity Recognition

Before applying the CRF model, we need to design a series of experimental algorithms, mainly including sequence labeling, feature selection, and feature template formulation.

4.3.1 Sequence Labeling Both character-based and word-based sequence methods have applications, each with advantages and disadvantages: character-based sequences provide richer features for machine learning but pose difficulties in entity boundary determination; word-based sequences, while unable to utilize character-level features, offer advantages in entity boundary judgment. However, small corpora using word-level sequences suffer from data sparsity issues, leading to insufficient training and affecting entity recognition results. Given the small scale of *Zuo Zhuan* corpora and the predominant use of character-based units in ancient Chinese entity recognition, we select single characters as the unit for sequence labeling.

Character-based sequence labeling for *Zuo Zhuan* involves classifying each Chinese character. In named entity recognition, classification is typically performed based on existing entity categories to indicate each character’s position within an entity, generally using W (single-character entity), B (entity beginning), M (entity middle), and E (entity end). Accordingly, we define a set Q containing 25 categories for sequence labeling-based entity recognition: {B-ATT, M-ATT, E-ATT, W-ATT, B-DEF, M-DEF, E-DEF, W-DEF, B-TIME, M-TIME, E-TIME, W-TIME, B-HEL, M-HEL, E-HEL, W-HEL, B-RES, M-RES, E-RES, W-RES, B-REA, M-REA, E-REA, W-REA, O}.

To intuitively demonstrate entity sequence labeling, we use the war sentence “Zheng Bo defeated Duan at Yan” from *Zuo Zhuan* as an example. The statement after correct entity recognition is shown as: B-ATT E-ATT O W-DEF O W-LOC. From the information embedded in subcategories B, E, and S, we determine that the statement contains entities such as Zheng Bo (person name), Duan (person name), and Yan (place name). Thus, we identify entities in the example war sentence through sequence labeling.

4.3.2 Feature Selection Feature selection is critical to named entity recognition, directly affecting model performance. Features are generally understood as elements representing categories in classification models. In sequence labeling models, characters or words can be considered features, with additional features such as surname, place name, and part-of-speech incorporated during named entity recognition. For ancient Chinese classics named entity recognition tasks, we add features including context window length, tags, word part-of-speech, and entity indicator words beyond character or word features.

4.3.3 Feature Template Formulation Feature templates in CRF models define methods for extracting features from training sets by combining information from surrounding characters and features according to required sequence unit lengths during training. In the renowned CRF++ open-source toolkit, feature templates extract features from training and test texts through template files, which are then used for CRF model calculation with feature parameters from the training set. Therefore, appropriate feature templates must be selected based on training corpora characteristics before CRF training. Through continuous experimentation and adjustment, we set template window sizes of [-1,1], [-2,2], and [-3,3], observing final effects through simplified feature templates.

5 Experiments and Results Analysis

5.1 Experimental Design

5.1.1 Data Source The Spring and Autumn period witnessed continuous warfare and hegemonic struggles among feudal lords. *Zuo Zhuan* provides comprehensive war records with detailed narratives, complete causality, and intact structures, offering significant advantages. Meanwhile, *Zuo Zhuan*'s syntactic structures and forms exhibit regularity, facilitating processing. The full text contains approximately 180,000 characters, recording 9,671 vocabulary items with rich and research-applicable lexical resources. Additionally, *Zuo Zhuan* texts feature distinct marker words, enabling rapid annotation, localization, and key sentence acquisition, thus shortening data processing time. To further ensure research accuracy, we adopt the *Zuo Zhuan* corpus constructed by Chen Xiaohe's team at Nanjing Normal University, which has already undergone text proofreading and word segmentation. Based on this, we use pattern matching methods to identify and construct the required war sentence corpus for this study.

5.1.2 Evaluation Metrics Each experimental result requires corresponding evaluation metrics. In this experiment, we adopt accuracy, recall, and F-score as evaluation metrics, defined as follows:

Accuracy = (Number of correctly annotated entity words by system / Number of entity words annotated by system) × 100%
Recall = (Number of correctly annotated entity words by system / Number of

entity words appearing in test set) × 100%
F-score = 2 × Accuracy × Recall / (Accuracy + Recall) × 100%

5.1.3 Experimental Environment (1) CRF++ Toolkit Selection. Currently, major open-source CRF-based tools include pocketcrf, flexcrf, and CRF++. Based on previous research experience, CRF++ is the most popular toolkit among developers, indicating its good performance. Therefore, we select the CRF++ toolkit (version 0.58) for this study.

(2) CRF++ Toolkit Usage. The CRF++ toolkit requires six files: crf_{learn}.exe: CRF++ training program; crf_{test}.exe: CRF++ prediction program; libcrfpp.dll: static link library required by training and prediction programs; template.data: file storing feature templates; train.data: file storing training corpora; test.data: file storing test corpora. The complete CRF++ toolkit file composition is shown in Figure 5 [Figure 5: see original paper].

CRF++ imposes strict format requirements on corpora. Generally, training and test texts in CRF models contain multiple tokens (meaning markers in lexical analysis), with each line representing one token. Each line includes two or more columns: the first column represents the character, the last column represents the annotation for that character, and middle columns are optional (zero or more) representing linguistic features related to that character. In other words, CRF model training texts generally include observed values, corresponding features, and state values. Table 2 shows a training text example with one added feature.

The test corpora format for CRF models is largely identical to training corpora, with the only difference being that test corpora may omit the last annotation column. The result data format generated by CRF model training is the same as training corpora but with an additional result column obtained after training, as shown in Table 3 .

CRF++ execution involves four steps: Place crf_{learn}.exe, crf_{test}.exe, libcrfpp.dll, and template.data from the CRF++ toolkit in the same folder, modifying the feature template file as needed. Simultaneously, place training and test corpora texts that have undergone format conversion into this folder. Train the corpora using the command: "crf_{learn} template train.data model," where crf_{learn} is the CRF learning algorithm, template is the feature template filename, train.data is the training corpus filename, and model is the model file generated during training. Test using the command: "crf_{test} -m model test.data > output.txt," where crf_{test} is the CRF testing algorithm, model is the trained model file, test.data is the test corpus, and output.txt is the test result file. Evaluate the generated test result file using the command: "conlleval.pl < output.txt."

5.2 Evaluation Results

5.2.1 War Sentence Identification Effectiveness Evaluation Based on part-of-speech features and annotation system compliance features, we conduct experiments using three different feature templates with context window lengths of [-1,1], [-2,2], and [-3,3] respectively. We perform 33 experiments under three different feature templates, selecting the best results from each template as shown in Table 4 .

Table 4 shows that optimal entity recognition performance is achieved with a context window length of [-1,1], reaching an F-score of 82.6999%, followed by [-2,2] and [-3,3]. Additionally, we observe that using the CRF model for named entity recognition achieves relatively high accuracy and recall rates around 80% regardless of feature template selection. Furthermore, both accuracy and recall rates decrease as context window length increases.

In summary, we conclude that window length impacts entity recognition results, with longer context windows yielding lower accuracy. CRF-based ancient Chinese named entity recognition is feasible and demonstrates notable effectiveness.

5.2.2 Named Entity Recognition with Different Feature Templates (1) CRF Entity Recognition Experiment with Part-of-Speech Features. In this experiment, we select context window lengths of [-1,1] for Template 1, [-2,2] for Template 2, and [-3,3] for Template 3 based on previous experiments. The results are shown in Table 5 .

Table 5 data reveals that adding part-of-speech features yields slight improvements in accuracy and recall (fluctuating within 0.5%), with minimal impact from different window length feature templates, not exhibiting the inverse relationship between context window length and accuracy observed in previous experiments. This may occur because part-of-speech features provide limited assistance to entities themselves or because they exert counteracting effects. However, we observe that adding part-of-speech features makes experimental metrics very close across trials, indicating that part-of-speech features enhance experimental stability.

In summary, we conclude that incorporating part-of-speech features into entity recognition has minimal impact on experimental results. We will add various entity indicator words in subsequent experiments to improve named entity recognition performance.

(2) CRF Entity Recognition Experiment with Entity Indicator Word Features. In this experiment, we only use feature templates with window lengths of 1 and 2. To verify feature template effects, we additionally create a template with the same window length but different formulation (Template 3 is slightly simplified compared to Template 2) for comparative experiments. Template 1 has a window length of 1; Templates 2 and 3 have window lengths of 2. This experiment incorporates various entity indicator words as features,

including warring party indicators, time indicators, cause indicators, result indicators, location indicators, and reinforcement indicators. Each indicator word occupies one column, marked as Y when present and N otherwise. Annotated examples are shown in Table 6 .

Selecting optimal results from cross-experiments with three templates, we obtain experimental results shown in Table 7 .

Table 7 data shows that war entity recognition performance significantly improves after adding entity indicator word features, with accuracy rates reaching 87.54% on average. Comparing Template 1 and Template 2 reveals that window length affects entity recognition effectiveness, with window length 2 achieving higher accuracy than window length 1 but lower recall. Comparing Template 2 and Template 3 shows that slight modifications to template formulation also impact results, indicating that feature template writing requires continuous experimentation and improvement. Meanwhile, we observe that F-scores from all three feature templates are very close.

5.3 Data Application

5.3.1 War Event Statistics *Zuo Zhuan* employs multiple expressions for attackers and defenders, including personal names, place names, and surname plus official position. To achieve data uniformity, we establish correspondences between warring party entities and states by mapping personal names, place names, and states, ultimately obtaining the *Zuo Zhuan* Warring Parties Table (excerpt) shown in Table 8 . Using the attacker and defender columns in this table, we 统计各国参与战争的频次 to intuitively understand war participation during the Spring and Autumn period. We also generate a word cloud of participating states using Tableau software, shown in Figure 6 [Figure 6: see original paper]. Additionally, our statistical analysis of war events in *Zuo Zhuan* reveals 69 rescue-type events and 1,020 expeditionary-type events.

Figure 6 shows that Jin, Chu, Qi, Zheng, Lu, Wei, Song, Wu, Qin, Chen, Cai, and other states participated frequently in wars during the Spring and Autumn period, representing major war participants. The war frequency word cloud generally describes relative participation counts but cannot differentiate between attackers and defenders. Therefore, we separately 统计各国进攻和防守次数 (exceeding 20 instances individually), with results shown in Figure 7 [Figure 7: see original paper] and Figure 8 [Figure 8: see original paper]. Combined analysis reveals that Jin primarily participated as an attacker, launching 152 attacks and defending 41 times, while Zheng mainly participated as a defender, though with relatively numerous attacks.

Simultaneously, we 统计交战地点 (with over 10 wars), shown in Figure 9 [Figure 9: see original paper]. Figure 9 indicates that Zheng, Song, and Wei, as frequent defenders, experienced the most battles on their territories. Remaining war locations were mostly states situated between warring parties, such as Chen and Xu between Zheng and Song.

5.3.2 Spring and Autumn Period Map We designed a dynamic *Zuo Zhuan* war visualization using HTML, CSS, and E-Charts 3 technologies to provide more intuitive understanding of wars during the Spring and Autumn period, as shown in Figure 10 [Figure 10: see original paper].

The dynamic map generation process involves: First, converting online Spring and Autumn map image resources into vector maps usable in HTML and E-Charts, using ArcGIS for vectorization through steps including loading base maps, creating SHP files, setting polygon attributes, and obtaining an SHP-format Spring and Autumn vector map. Second, converting the SHP-format map into JSON-format map data parseable by E-Charts using Mapshaper. Third, using the pure JavaScript chart library E-Charts to convert war data into intuitive, interactive, and personalized visual charts through: initializing E-Charts instances via echarts.init() and placing them in div containers; asynchronously loading map and war entity data using jQuery's JSON retrieval statement $.getjson(); configuring frameworks and populating data via setOption() to dynamically generate war maps.

Conclusion

Pre-Qin classics contain rich content and active ideas, embodying the thoughts and wisdom of pre-Qin masters. Among them, *Zuo Zhuan* stands as one of the most representative historical works of the pre-Qin period. Research on *Zuo Zhuan* can assist studies in classical Chinese linguistics, archaeology, and other historical and literary fields. Meanwhile, using *Zuo Zhuan* as experimental corpora aims to explore effective ancient Chinese information extraction methods while providing references for natural language processing. This paper constructs a basic framework system for *Zuo Zhuan* war events based on frame theory, uses pattern matching for war sentence identification, employs the CRF model combined with feature templates to recognize and extract seven named entities (war time, attacker, defender, war location, war trigger cause, war result, and reinforcements), and analyzes and visualizes war events based on structured data. This approach features: (1) Combining CRF models with frame theory, feature templates, and pattern matching methods to improve completeness, specificity, and feasibility of event extraction; (2) Designing and selecting appropriate annotation systems and feature templates based on *Zuo Zhuan* text characteristics to achieve good experimental results; (3) Conducting multiple experiments to verify effects of different window length feature templates and different features to obtain optimal results.

Our research yields the following conclusions: (1) The CRF model can be effectively applied to war event extraction from *Zuo Zhuan*; (2) Feature selection affects entity recognition results; (3) Regarding specific content, Jin, Chu, Qi, Zheng, and other states participated frequently in wars during the Spring and Autumn period, with Jin as the primary attacker and Zheng as the main defender.

References

[1] Huang Shuiqing, Wang Dongbo. Current status and trends of ancient Chinese text processing research[J]. Library and Information Service, 2017, 61(12): 43-49.
[2] Shi Chenlu. What restrains the digitization of ancient books[EB/OL]. [2019-05-15]. https://www.jfdaily.com/news/detail?id=53981#
[3] Huang Shuiqing, Wang Dongbo, He Lin. Research on automatic ancient Chinese place name recognition model construction based on pre-Qin corpora[J]. Library and Information Service, 2015, 59(12): 135-140.
[4] Liu C L, Huang C K, Wang H S, et al. Mining local gazetteers of literary Chinese with CRF and pattern based methods for biographical information[C]//Proceedings of the IEEE international conference on big data. Santa Clara: IEEE, 2015: 1629-1638.
[5] Qian Zhiyong, Zhou Jianzhong, Tong Guoping, et al. Research on automatic word segmentation and annotation of Chu Ci based on HMM[J]. Library and Information Service, 2014, 58(4): 105-110.
[6] Zhu Xiaohong. Research on pre-Qin military law thought[D]. Xi'an: Northwest University, 2010.
[7] Liu Min. Research and application of information extraction and new knowledge discovery system based on professional domain literature[D]. Jinan: Shandong University, 2018.
[8] Zhao Yanyan, Qin Bing, Che Wanxiang. Research on Chinese event extraction technology[J]. Journal of Chinese Information Processing, 2008, 22(1): 3-8.
[9] HAI L C, NG H T. A maximum entropy approach to information extraction from semi-structured and free text[C]//Eighteenth national conference on artificial intelligence. San Jose: American Association for Artificial Intelligence, 2002.
[10] AHN D. The stages of event extraction[C]//Workshop on annotating & reasoning about time & events. Sydney: Association for Computational Linguistics, 2006.
[11] Yu Jiangde, Xiao Xinfeng, Fan Xiaozhong. Chinese text event information extraction based on Hidden Markov Model[C]//Proceedings of the national open distributed and parallel computer academic conference (Volume 2). Nanning, 2007.
[12] Wu Pingbo, Chen Qunxiu, Ma Liang. Research on clue event extraction and integration system based on spatiotemporal analysis[J]. Journal of Chinese Information Processing, 2006, 20(1): 21-28.
[13] Jiang Jifa. A cross-sentence Chinese event information extraction method[J]. Computer Engineering, 2005, 31(2): 27-29.
[14] Zheng Jiaheng, Wang Xingyi, Li Fei. Research on automatic information extraction pattern generation methods[J]. Journal of Chinese Information Processing, 2004(1): 48-54.
[15] Yang Erhong. Research on emergency event information extraction[D]. Beijing: Beijing Language and Culture University, 2005.

[16] Gao Juan, Liu Jiazhen. Problems and countermeasures of ancient book digitization in mainland China[J]. Journal of Library Science in China, 2013(4): 110-119.
[17] Wang Jialing. Automatic word segmentation of medieval Chinese: A case study of Han Shu[D]. Nanjing: Nanjing Normal University, 2014.
[18] Liang Shehui, Chen Xiaohe. Research on automatic word segmentation methods for pre-Qin literature Mencius[J]. Journal of School of Chinese Language and Literature, Nanjing Normal University, 2013(3): 175-182.
[19] Wang Zheng. Research on automatic ancient book place name recognition based on CRF[D]. Nanning: Guangxi University for Nationalities, 2008.
[20] Zhang Qiuxia. Research on expeditionary verbs in Zuo Zhuan[D]. Changchun: Jilin University, 2009.
[21] Deng Yong. Hegemony: Justice and order[D]. Wuhan: Wuhan University, 2007.

Author Contributions

Li Zhangchao: Paper writing and revision
Li Zhongkai: Algorithm implementation and data analysis
He Lin: Topic selection, research framework design, and paper revision suggestions

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