Automatic Discovery of Parallel Textual Variants in Classical Chinese Texts: A Post-print

Authors: Liang Yuan, Wang Dongbo, Huang Shuiqing

Date: 2023-04-01T16:02:50+00:00

Abstract

[目的/意义] Textual variants are a common phenomenon in ancient texts and constitute an important research subject. Traditional collation of ancient texts involves manually searching through vast quantities of ancient documents for collation materials, including textual variants. This approach is not only time-consuming, labor-intensive, and burdensome, but also may not yield precise and comprehensive data. Computer-assisted automatic discovery of textual variants can extract effective information from larger-scale corpora. Furthermore, collation methods incorporating automatic discovery of textual variants enable exhaustive retrieval, which is of significant importance for the external collation methodology of ancient texts, thereby offering novel approaches and methodologies for ancient text collation research in the new era.

[方法/过程] This study employs the Spring and Autumn Annals and the "Three Commentaries on the Spring and Autumn Annals" as experimental corpora, introduces the concept of parallel corpora commonly utilized in the field of text translation, integrates deep learning algorithms, conducts comparative experiments on LSTM and BERT models against the classical SVM model, and further explores and discusses the content related to colleague textual variants—wherein different expressions are used across two ancient texts to describe the same event.

[结果/结论] The experiments yield a deep learning model for automatic discovery of colleague textual variants applicable to the "Three Commentaries on the Spring and Autumn Annals", thereby demonstrating the feasibility of integrating emerging technologies such as deep learning into research on ancient text knowledge base construction. Simultaneously, the combination of deep learning technology and the parallel corpus approach can exert substantial influence on textual variant research, offering practical support for the application of digital humanities in Chinese language and literature studies.

Full Text

Preamble

Research on Automatic Mining of Variant Texts Expressing the Same Event in Ancient Books

Liang Yuan[1],[2], Wang Dongbo[1],[2], Huang Shuiqing[1],[2]
[1]College of Information Management, Nanjing Agricultural University, Nanjing 210095
[2]Research Center for Humanities and Social Computing, Nanjing Agricultural University, Nanjing 210095

Abstract:
[Purpose/Significance] Variant texts are a common phenomenon and an important research object in ancient Chinese books. Traditional collation of ancient texts involves manually searching for collation materials, including variant texts, from large collections of ancient literature—a process that is not only time-consuming, labor-intensive, and demanding, but also may yield data that is neither accurate nor comprehensive. Automatic mining of variant texts through computers can extract effective information from larger-scale corpora. Moreover, collation methods combined with automatic variant mining can achieve exhaustive retrieval, which is of great significance for the "ta jiao" method (using other texts as collateral evidence) in ancient book collation, providing new ideas and methods for collation research in the new era. [Method/Process] This study takes the *Spring and Autumn Annals* and the "Three Biographies of the Spring and Autumn Period" as experimental corpora, introduces the parallel corpus concept commonly used in text translation, combines deep learning algorithms, and conducts comparative experiments on LSTM and BERT models against the classic SVM model. It further explores and discusses content related to variants expressing the same event with different formulations in two ancient books. [Result/Conclusion] The experiments yield a deep learning model suitable for automatic mining of same-event variants in the "Three Biographies," proving the feasibility of integrating emerging technologies such as deep learning into the construction of ancient book knowledge bases. Meanwhile, the combination of deep learning technology and parallel corpus concepts can play a significant role in variant text research, providing practical support for the application of digital humanities in Chinese language and literature studies.

Keywords: Three Biographies of the Spring and Autumn Period; variant texts; BERT; automatic mining; digital humanities
Classification Number: G255.1
DOI: 10.13266/j.issn.0252-3116.2021.09.011

China is rich in ancient book resources with countless works, where variant texts are extremely common. Broadly speaking, variant texts include not only character differences between different versions of the same work, but also distinctions in wording when the same event is quoted or narrated [1]. From the

perspective of philology, variant texts can illuminate interchangeable characters, ancient vs. modern characters, and variant character forms. In lexicology, they can reveal synonyms, cognate words, and disyllabic words. In grammar, they can clarify word order and certain special syntactic relationships. Whether in philology, lexicology, or grammar, research findings on variant texts may be applied [2-3].

As the mainstream ideology in ancient China, Confucianism has profoundly influenced China and even the world. The *Spring and Autumn Annals* is one of the most classic Confucian works and China's earliest annalistic historical text, recording traditional culture and ancient wisdom. "The Spring and Autumn style records profound meanings in subtle words." The "Three Biographies of the Spring and Autumn Period" that comment on the *Annals* supplement these obscure contents comprehensively and vividly, presenting a colorful Spring and Autumn period from perspectives of historical background, social customs, and political rituals. This study begins with the *Spring and Autumn Annals* and its three biographies, focusing primarily on different expressions of the same event between two ancient books—i.e., same-event variants. It introduces the parallel corpus concept from text translation and combines deep learning algorithms to further explore automatic mining of same-event variants in ancient books, aiming to integrate new technologies into ancient book knowledge base construction and provide new ideas and methods.

2 Related Research

The study of variant texts can be traced back to ancient scholars' commentaries and sub-commentaries on classical texts. In modern times, many disciplines involve variant text research. As the saying goes, "know what it is and know why it is so," the causes of variant texts have always been a focus for researchers. Luo Jiyong [4] summarized four causes of variant texts in ancient books, proposing that variants are not only valuable for collation and exegetical studies, but also important references for interpreting ancient texts. Subsequently, Wang Yankun [5] and Shi Yunsun [6] respectively studied variant texts in ancient books and modern discourse from different angles, analyzing their causes.

Ancient books are numerous and diverse, divided into four categories (Classics, History, Masters, and Collections) with forty-four subcategories. Studying variants from different categories may yield different results. Many scholars have researched variant phenomena in classical poetry and Buddhist scriptures. Deng Yawen [7] and Wang Xuejun [8] respectively analyzed variants in Tang poetry and Song lyrics, summarizing their types and causes. Zeng Liang and Jiang Kexin [9] focused on Buddhist scripture variants, analyzing their causes based on knowledge of popular characters, and argued that variants could be used to study cognate words and correct Buddhist texts. Some scholars specialized in variants within specific ancient books: Jiang Linchang [10], Zhou Fuyun [11], and Chen Weiling [12] respectively examined *Chu Ci*, *Li Sao*, and *Huai Sha*, analyzing causes of variants and their research significance. Other scholars [13-14]

compared variants between the *Sui Ren Shu Chu Shi Song* and *Wen Xuan*'s *Chu Shi Song*, as well as between the *Daoist Canon* and Dunhuang versions of *Taishang Dongyuan Shenzhou Jing*, summarizing variant phenomena and discussing their practical research significance.

One significant research direction is the role of variants in exegetical studies and textual criticism. Liu He [15] distinguished variants in different versions of the same book, different chapters of the same book, different books recording similar events, and quotations between books from the perspective of exegetical and collation studies, summarizing their importance for semantic interpretation and textual collation. Bian Xingcan [16] proposed the "variant-guided" exegetical method to compensate for the unscientific nature of variant verification, discussing variants' important role in exegetical studies. Wang Yankun [17], Wu Xinchou [18], and Yu Ting [19] emphasized variants' important academic value in characters, phonology, exegetics, vocabulary, grammar, and version collation based on different ancient texts. Recent research has focused on specific works: Di Biyun et al. [20] stated that variant materials are important references for ancient book collation, and that research on variants in *Lingshu Jing* would benefit traditional Chinese medicine and historical linguistics. Bo Yingying [21] studied different types of variants in *Chu Ci Shu· Jiu Zhang*, proposing their significance for Chinese language history, *Chu Ci* exegetical history, and medieval language studies.

Variants also play an important role in ancient book organization and source identification. Feng Qing [22] and Chen Lihua [23] respectively analyzed variant vocabulary in *Shishuo Xinyu* and *Jin Shu*, and variant phenomena in the translated sutra *Sheng Jing*, discussing variants' role in compiling the *Great Chinese Dictionary*. Chen Renren [24] interpreted hexagram variants in four versions of the *Book of Changes* (received text, silk manuscript, Fuyang text, and bamboo slips), providing clues for *Zhou Yi* version evolution research. Ren Lu [25] analyzed version variants and different translations in the Buddhist text *Shuo Wu Gou Cheng Jing*, discussing variants' importance for compiling large-scale Chinese dictionaries. Zhang Qi [26] examined two different versions of the Tang poem *Guan Qi*, judging its original and revised authors.

Some researchers have taken alternative approaches, analyzing variants from educational perspectives. Cui Dasong et al. [30] studied the relationship between variants and ancient Chinese teaching, arguing that variant teaching helps students learn Chinese and expand their knowledge. Guo Dianchen and Guo Zhiyuan [29] compared Li Bai's poetry variants based on *He Yue Ying Ling Ji* with other collections, discussing artistic conception, sound, and the poet's experiences. Zhou Fuyun [28] listed over a dozen variants in Wang Wei's poems from *Complete Tang Poems*, analyzing the relationship between variant semantics and poetic sentiment. Wang Xiang [27] analyzed a variant in Li Bai's *Shu Dao Nan* from perspectives of physical logic, poetic language origins, version differences, and linguistic expression.

With information technology development and researchers' continuous promo-

tion of ancient book digitization, automatic mining of variants—an indispensable part of ancient book research—has gradually become a research hotspot. Automatic variant mining can extract text information for collation from large ancient text collections and achieve exhaustive retrieval [32]. This method helps solve the disadvantage of traditional collation requiring massive manual labor while obtaining more systematic and comprehensive information, providing ideas for improving collation techniques. Jiang Huimin et al. [33] took *Fangzhi Wuchan* as research object, proposing an automatic merging algorithm based on local chronicle style patterns that compares paragraph content, seeks common ground while preserving differences, and annotates variant versions to achieve automatic merging of chronicle versions from different eras in the same region. Xiao Lei and Chen Xiaohe [34] took *Three Biographies of the Spring and Autumn Annals* as example, using bigram to calculate sentence similarity matching and remove identical texts to automatically discover ancient book version variants. Li Yue [35] implemented automatic discovery of same-event variants in *Zuo Zhuan* and *Shiji* from a computational linguistics perspective based on sentence similarity algorithms, improving edit distance algorithms to provide ideas for large-scale ancient book variant processing. Zhao Hong [36] emphasized the importance of variant characters in Turpan documents for Chinese historical corpus construction, arguing that preserving variants in ancient Chinese corpora would facilitate automatic retrieval and discovery. Xie Jing [37] conducted automatic variant discovery research on *Huangdi Neijing* based on sentence matching algorithms, providing important references for traditional Chinese medicine ancient book research.

Most of the above studies analyze variants only from the perspective of Chinese language and literature, with relatively few focusing on automatic variant mining. This study refines the data source to the classic annalistic historical text *Spring and Autumn Annals* and its three biographies, introduces parallel corpus concepts, and attempts to apply deep learning algorithms to facilitate automatic variant mining in large-scale corpora.

3 Model Application

Automatic variant mining can be understood to some extent as automatic matching of two sentences with semantic similarity. Based on this characteristic, this study introduces parallel corpus concepts while attempting to apply the classic Support Vector Machine (SVM), Long Short-Term Memory (LSTM) networks, and classification tasks in the BERT (Bidirectional Encoder Representations from Transformers) model to achieve automatic variant mining.

3.1 SVM

Support Vector Machine is recognized as one of the most effective machine learning algorithms before the emergence of deep learning. It is a supervised learning method that can solve classification and prediction problems. Its basic idea is to find a dividing hyperplane in the sample space based on the training

set to separate samples of different categories [38]. In variant identification tasks, two sentences from candidate pairs (e.g., "车马曰醵" and "乘马曰醵") are represented and concatenated as features to be classified. Through training on a large number of labeled candidate sentence pairs, a hyperplane dividing "0" and "1" is established to achieve the automatic variant mining model. Therefore, this study selects SVM as a baseline for exploring automatic variant mining models, attempting to obtain more effective new models through comparison.

3.2 LSTM

Long Short-Term Memory networks [39] are a variant of Recurrent Neural Networks (RNN). This study adopts Siamese LSTM, primarily used to solve binary classification problems. It is called "Siamese" because the model shares the same weights for sentences on both left and right sides. In this study, the classification process of Siamese LSTM involves inputting two candidate sentences into the model: model a inputs "克者何" while model b inputs "克之者何", then calculating the Manhattan distance between their hidden layer vectors to evaluate sentence similarity. The formula is as follows (similarity range 0-1) [40]:

$$D = \exp(-\|h(\text{left}) - h(\text{right})\|_1)$$

Based on Formula (1), we can determine whether two sentences constitute a variant pair. As shown in Figure 1 [Figure 1: see original paper], given a sample [a, b, y], x and y represent input and output respectively, where y results in [0, 1].

3.3 BERT

BERT is a deep learning model proposed by Google [41], i.e., bidirectional encoder representations based on Transformer. It can simultaneously utilize contextual information of the target vocabulary to solve problems and learn inter-sentence relationships to determine whether two sentences are related. Therefore, BERT typically performs excellently in text classification. Taking a variant sentence pair from this experimental data as an example, inputting "公何以不言即位" and "不书即位" into the model, then through pre-training methods such as random masking, replacement, or next sentence prediction to establish a language model, subsequently predicts the relationship between candidate sentence pairs—that is, predicts sentence categories, as shown in Figure 2 [Figure 2: see original paper].

4 Construction of Variant Text Annotation System

For digital collation research of Chinese historical literature, the "Chinese Text Project" is second to none. This is an online open e-book platform that breaks the barriers of print publishing, containing over 30,000 works and representing the largest collection of historical Chinese literature currently available. This

study obtained original text data of the "Three Biographies"—*Gongyang Zhuan*, *Guliang Zhuan*, and *Zuo Zhuan*—from this platform, then performed cleaning, deduplication, and proofreading as preprocessing steps.

4.1 Data Preprocessing

As this study is an initial exploration of automatic same-event variant mining, it prioritizes short-sentence level variant matching and classification. The author segmented sentences using punctuation marks (commas, periods, exclamation marks, question marks, semicolons, colons), obtaining 67,693 short sentences. Since subsequent research requires utilizing each short sentence's positional information and contextual context, each short sentence was assigned an identifier (see Table 1), where the serial number is sequential. The processed results are shown in Figure 3 [Figure 3: see original paper].

4.2 Variant Annotation Standards

Based on the definition of variants, the preprocessed "Three Biographies" and *Spring and Autumn Annals* texts were annotated for variants with one-to-one correspondence between variant sentences. Specific standards are as follows:

Complete semantic correspondence with high text similarity. For example, "成公意也" and "成公志也" express the same meaning and are annotated as variant sentences.

Partial semantic and textual correspondence. Matching uses short sentences as units, such as "公将平国而反之桓" and "将以让桓也" forming a variant pair.

High semantic similarity but low text similarity, containing synonyms, etc. Also treated as variant pairs, e.g., "盟纳季子也" and "请复季友也" both express that the alliance's purpose was to invite Ji You back to the state, though formulations differ significantly.

Almost no textual similarity but expressing the same event. For example, "则齐国尽子之有也" and "举国而授" are annotated as a variant pair.

Partial components omitted or abbreviated. For instance, "孙良夫率师侵宋" and "伯宗、夏阳、孙良夫、宁相、郑人、伊雒之戎、陆浑、蛮氏侵宋" both express the event of invading Song, but the latter provides more detailed description of the invading parties.

Time and seasonal expressions are not annotated as variant pairs, such as "元年", "春", "三月", etc.

High text similarity but different core semantic expression are not treated as variant pairs. For example, "宋师及齐师战于鞫" expresses that Song and Qi fought at 鞫, while "宋败齐师于鞫" indicates the defeated party and outcome, so these are not variant pairs.

Cases not covered by the above seven standards are analyzed individually by consulting relevant materials to ensure annotation consistency.

Following these standards, manual annotation yielded 1,692 variant sentence pairs.

4.3 Construction of Variant Parallel Corpus

The candidate sentence pair set consists of variant and non-variant sentence pairs in a 1:1 ratio, totaling 3,384 candidate pairs. Variant pairs were manually generated according to Section 4.2 standards. Non-variant pairs consist of sentences from Section 4.1 data, where sentence b is automatically generated from the paragraph corresponding to sentence a's paragraph in the other two biographies, typically the 5th sentence after sentence b's corresponding sentence in variant pairs. Subsequently, using the parallel corpus format, sentence pairs are aligned with "0-1" classification, where 0 indicates non-variant pairs and 1 indicates variant pairs. Partial data from the variant parallel corpus is shown in Table 2 .

5 Experimental Results and Evaluation of Automatic Variant Mining Models

5.1 Experimental Environment and Parameter Settings

The experimental environment uses Ubuntu 16.04 OS, 16GB DDR4 RAM, 4GB GDDR5 GPU memory, Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz, and NVIDIA Quadro K1200 GPU. To ensure comparability of results across the three models, input data consists of uniformly processed identical corpora under the same experimental environment. Parameter settings for the three models are shown in Table 3 .

5.2 Evaluation Metrics

For model performance evaluation, this study adopts Precision (P), Recall (R), and F-measure (F). Specific calculation formulas are:

$$\text{Precision } P = \frac{\text{Correctly identified pairs}}{\text{Correctly identified pairs} + \text{Incorrectly identified pairs}} \times 100\%$$

$$\text{Recall } R = \frac{\text{Correctly identified pairs}}{\text{Correctly identified pairs} + \text{Unidentified pairs}} \times 100\%$$

$$\text{Harmonic mean } F = \frac{2 \times P \times R}{P + R}$$

5.3 Experimental Results

This experiment applied SVM, LSTM, and BERT models on the same dataset using ten-fold cross-validation. Evaluation metric values are shown in Table 4 .

As shown in Table 4, overall LSTM (F=50.14%) and BERT (F=56.90%) perform significantly better than SVM (F=29.61%), with SVM's F-value even below

30%. Although the F-value difference between LSTM and BERT is relatively small, the latter shows considerable improvement. BERT's three evaluation metrics are almost all higher than LSTM's corresponding experimental groups, with BERT's best model achieving F=61.34%. While these evaluation metrics may not seem very high, they are relatively ideal compared with previous results. Notably, prior research using deep learning or even machine learning for automatic variant mining is scarce. Li Yue [35] achieved best results of P=46.02%, R=90.15%, F=60.93% using improved edit distance algorithms for same-event variant discovery. This experiment's optimal model exceeds previous algorithmic results with significantly improved precision.

This study applies deep learning models to automatic variant mining in the "Three Biographies" and compares results with the classic SVM machine learning algorithm on the same dataset, proving deep learning's clear advantages in this research domain. Therefore, for automatic variant mining model selection, BERT performs excellently, LSTM is also relatively good, and both are significantly superior to the classic SVM model.

6 Analysis Based on Automatic Variant Mining Results

This study further analyzed the 1,692 annotated variant sentence pairs. The distribution of variant pairs is shown in Table 5 : 21 pairs between *Spring and Autumn Annals* and *Gongyang Zhuan*, 19 pairs between *Annals* and *Guliang Zhuan*, 970 pairs between *Annals* and *Zuo Zhuan*, 513 pairs between *Gongyang Zhuan* and *Guliang Zhuan*, 89 pairs between *Gongyang Zhuan* and *Zuo Zhuan*, and 80 pairs between *Guliang Zhuan* and *Zuo Zhuan*.

The three biographies each have distinct linguistic styles. *Zuo Zhuan* mostly features concise language, often using the same or similar expressions as the *Annals* to briefly mention events, but provides exhaustive detail for events requiring special expansion. It clearly narrates event backgrounds while vividly portraying protagonists. *Zuo Zhuan*'s vivid character dialogues are also a linguistic feature, which simultaneously increases difficulty for automatic variant identification—yet this is also part of the significance of such research. Additionally, *Zuo Zhuan* devotes considerable space to historical backgrounds of events, often appearing as "Zuo appendices." This study excludes "Zuo appendix" content from the annotation system primarily because it generally introduces causes, consequences, and contexts of periods or events, rarely forming variants with texts in the other three classics.

Relatively speaking, *Guliang Zhuan* and *Gongyang Zhuan* have more similar narrative styles, mostly using rhetorical questions in a question-and-answer format. Answers to previous questions raise subsequent ones, with clear logic and systematic analysis. Content focuses on recording reasons for events, their special characteristics, or connections with other events. These variant pairs are relatively easier to identify.

However, these two classics sometimes provide different or even opposite inter-

pretations of the same event. For example, regarding *Annals* entry "齐人伐山戎" (Qi people attacked Shan Rong), *Gongyang Zhuan · Duke Zhuang 30th Year* explains: "This refers to the Qi Marquis. Why is he called 'people'? To denigrate." *Guliang Zhuan · Duke Zhuang 30th Year* explains: " 'Qi people' refers to the Qi Marquis. Why is it phrased this way? Out of love for the Qi Marquis, not wanting him mentioned together with Shan Rong." Since these candidate sentence pairs express completely opposite meanings, they are not treated as variant pairs in this study.

Digital humanities represents an excellent approach for organic integration of new technology and humanities research. In recent years, as ancient book digitization progresses, ancient book-related research has gradually occupied a place in digital humanities, with variants being an indispensable and important component of ancient book research, especially digital ancient book studies. Automatic variant mining can not only significantly reduce manual labor costs in traditional collation but also verify the feasibility of integrating new technologies with classical studies. Therefore, this study takes the *Spring and Autumn Annals* and its three biographies as experimental corpora, introduces parallel corpus concepts from text translation, constructs a variant parallel corpus and annotation standards, and through comparative analysis of SVM, LSTM, and BERT model experiments, finds that deep learning models perform well in automatic variant mining. This demonstrates that deep learning and parallel corpora can play significant roles in variant research, providing feasible solutions for automatic variant mining that are worthy of expansion and reuse. The authors will continue exploring new technologies in future research to improve model performance while extending application corpora to more classical texts.

Furthermore, analysis of identified variant pairs suggests several directions for improvement and exploration:

For recognizing low-frequency variant types, attempt to improve models or algorithms, as non-variant pairs with high text similarity and variant pairs with low text similarity are difficult to identify correctly.

This study only attempted one-to-one short sentence matching, converting many-to-one sentences into one-to-one form by ignoring some short sentences and only annotating core semantic sentences as variant pairs, which may affect mining effectiveness. Future research will attempt more annotation forms.

The "Three Biographies" are all annalistic historical texts. This study begins with this genre, but future research will expand to biographical histories like *Shiji*, hoping to apply to more ancient book types.

Current models do not incorporate contextual features or synonym dictionaries. Subsequent research will automatically obtain context for variant pairs using serial numbers to improve identification, while introducing dictionaries of personal names, official titles, and part-of-speech synonyms to obtain more advantageous automatic variant mining methods.

References

[1] Huang Peirong. Analysis of Variant Texts in Ancient Books [J]. *Chinese Studies*, 1991, 9(2): 395.
[2] Li Juan. Research on Cognate Words in Variant Texts between *Shiji* and *Hanshu* [J]. *Journal of Hubei Normal University (Philosophy and Social Sciences Edition)*, 2011, 31(4): 60-63.
[3] Li Juan. Research on the Exegetical Value of Variant Texts between *Shiji* and *Hanshu* [D]. Huangshi: Hubei Normal University, 2012.
[4] Luo Jiyong. Variant Texts and Interpretation [J]. *Journal of Ancient Books Collation and Studies*, 1986(2): 58-62.
[5] Wang Yankun. On the Causes of Variant Texts in Ancient Books [J]. *Jinan Journal: Philosophy and Social Sciences Edition*, 1989(4): 78-85.
[6] Shi Yunsun. Variant Texts in Discourse [J]. *Journal of Anqing Teachers College (Social Science Edition)*, 1996(2): 2-8.
[7] Deng Yawen. On Tang Poetry Variants [J]. *Journal of Hubei University of Science and Technology*, 2002, 22(5): 68-70.
[8] Wang Xuejun. Exploration of Song Lyric Variants [J]. *Data of Culture and Education*, 2010(18): 32-36.
[9] Zeng Liang, Jiang Kexin. Buddhist Scripture Variants and Word Studies [J]. *Research in Ancient Chinese Language*, 2013(2): 43-48.
[10] Jiang Linchang. Examples of *Chu Ci* Variants [J]. *Literature*, 1991(3): 3-14.
[11] Zhou Fuyun. Examples of *Li Sao* Variants [J]. *Journal of Huaiyin Teachers College (Philosophy and Social Sciences Edition)*, 1993(2): 20-24.
[12] Chen Weiling. Textual Research on *Huai Sha* Variants [J]. *Journal of Vocational University*, 2007(1): 46-47.
[13] Yi Min. Variants between *Sui Ren Shu Chu Shi Song* and *Wen Xuan* [J]. *Journal of Jinggangshan Teachers College*, 2005(1): 5-8.
[14] Niu Shangpeng. Textual Research on Variants in *Taishang Dongyuan Shenzhou Jing* [J]. *Journal of Yangtze Normal University*, 2016, 32(1): 73-78.
[15] Liu He. Variant Texts and Collation [J]. *Journal of Northeast Normal University (Philosophy)*, 1986(2): 62-67.
[16] Bian Xingcan. On the Role of Variant Texts in Exegetical Studies [J]. *Journal of Zhejiang University (Humanities and Social Sciences Edition)*, 1998(3): 135-140.
[17] Wang Yankun. On the Application of Ancient Book Variants [J]. *Jinan Journal: Philosophy and Social Sciences Edition*, 1987(1): 75-81.
[18] Wu Xinchou. Types and Values of Bamboo and Silk Variants [J]. *Journal of South China Normal University (Social Science Edition)*, 2000(4): 37-42.
[19] Yu Ting. Historical Perspective on Using Variants in Exegetical Practice [J]. *Yangtze River Academic*, 2009(3): 131-138.
[20] Di Biyun, Sun Zhaojie, Fan Dengmai. On Variant Research in *Lingshu Jing* [J]. *Journal of Traditional Chinese Medical Literature*, 2013, 31(3): 12-14.
[21] Bo Yingying. Research on Variants in *Chu Ci Shu · Jiu Zhang* [J]. *Journal of Language*, 2016(24): 59, 107.

[22] Feng Qing. Variant Vocabulary and Lexical History Research [J]. *Journal of Harbin Normal University (Social Sciences Edition)*, 2010, 1(1): 52-55.
[23] Chen Lihua. Research on Variants in *Sheng Jing* [D]. Changsha: Hunan Normal University, 2011.
[24] Chen Renren. Interpretation of Bi Hexagram Variants [J]. *History of Chinese Philosophy*, 2010(3): 54-62.
[25] Ren Lu. Research on Variants in *Shuo Wu Gou Cheng Jing* [D]. Guiyang: Guizhou Normal University, 2015.
[26] Zhang Qi. Research on Author and Variants of *Guan Qi* [J]. *Beijing Social Sciences*, 2016(2): 83-88.
[27] Wang Xiang. A Variant in Li Bai's *Shu Dao Nan* [J]. *Higher Education Management*, 1990(2): 16-17.
[28] Zhou Fuyun. Exploration of Wang Wei Poetry Variants [J]. *Journal of Taizhou University*, 1998(1): 75-79.
[29] Guo Dianchen, Guo Zhiyuan. Research on Li Bai Poetry Variants—Centered on *He Yue Ying Ling Ji* [J]. *Journal of Mianyang Teachers College*, 2014, 33(1): 12-18.
[30] Cui Dasong, Zhan Xuzuo, Chu Taisong, et al. Variant Comparison and Ancient Chinese Teaching [J]. *Journal of Chuzhou University*, 2008, 10(1): 22-25.
[31] Guo Yuchen. Textual Research on Variants in *Su Xinshi Xugongdian* [J]. *Jiannan Literature (Second Half)*, 2016(5): 33-34.
[32] Ju Mingku. Ancient Book Digitization and Traditional Philology [J]. *Journal of Tsinghua University (Philosophy and Social Sciences Edition)*, 2011, 26(5): 154-158, 161.
[33] Jiang Huimin, Bai Zhentian, Zhou Jinshui, et al. Research and Implementation of Automatic Merging of Chronicle Versions from Different Eras in the Same Region [J]. *Guangxi Local Chronicles*, 2008(5): 29-34.
[34] Xiao Lei, Chen Xiaohe. Automatic Discovery of Ancient Book Version Variants [J]. *Journal of Chinese Information Processing*, 2010, 24(5): 50-55.
[35] Li Yue. Automatic Discovery and Analysis of Same-Event Variants between *Zuo Zhuan* and *Shiji* [D]. Nanjing: Nanjing Normal University, 2014.
[36] Zhao Hong. Thoughts on Turpan Documents and Chinese Corpus Construction [J]. *Journal of Nanjing Normal University (School of Chinese Language and Literature)*, 2014(3): 155-158.
[37] Xie Jing. Research on Automatic Variant Discovery in *Huangdi Neijing* Based on Sentence Matching [J]. *Science and Technology Vision*, 2015(35): 53-54.
[38] Hearst MA, Dumais ST, Osuna E, et al. Support Vector Machines [J]. *IEEE Intelligent Systems & Their Applications*, 1998, 13(4): 18-28.
[39] Graves A, Schmidhuber J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures [J]. *Neural Networks*, 2005, 18(5/6): 602-610.
[40] Mueller J, Thyagarajan A. Siamese Recurrent Architectures for Learning Sentence Similarity [C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona: AAAI, 2016: 2786-2792.

[41] Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]//Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis: ACL, 2019: 4171-4186.

Author Contributions

Liang Yuan: Data processing, drafting the manuscript;
Wang Dongbo: Research conceptualization, methodology design;
Huang Shuiqing: Conceptualization, overall research design, manuscript revision.

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