Vector-Based Stylistic Analysis on Ancient Chinese Books: Take the *Three Commentaries on the Spring and Autumn Annals* as an Example

Yue Qi [1],Liu Liu [1]*, Bin Li [2], Dongbo Wang [1]
[1]College of Information Management, Nanjing Agricultural University,Nanjing,China
[2]School of Chinese Language and Literature, Nanjing Normal University,Nanjing,China
liuliu@njau.edu.cn

Abstract

Commentary of Gongyang, *Commentary of Guliang*, and *Commentary of Zuo* are collectively called the *Three Commentaries on the Spring and Autumn Annals*, which are the supplement and interpretation of the content of *Spring and Autumn Annals* with value in historical and literary research. In traditional research paradigms, scholars often explored the differences between the *Three Commentaries* within the details in contexts. Starting from the view of Stylistic Analysis, this paper examines the differences in the language style of the *Three Commentaries* through the representation of language, which takes the methods of deep learning. Specifically, this study vectorizes the context at word and sentence levels. It maps them into the same plane to find the differences between the use of words and sentences in the *Three Commentaries*. The results show that the *Commentary of Gongyang* and the *Commentary of Guliang* are relatively similar, while the *Commentary of Zuo* is significantly different. This paper verifies the feasibility of deep learning methods in stylistics study under computational humanities. It provides a valuable perspective for studying the *Three Commentaries on the Spring and Autumn Annals*.

1 Introduction

Style is an additional component in the process of language expression and expression. It changes due to the social era and environment in which language is used and in various forms due to the user's expression habits and intentions. This characteristic has received longstanding attention from stylistics. Among the study of ancient Chinese classics, *the Spring and Autumn Annals* was known as "having profound meaning in simple words." and the *Historical Records* were called "*Li Sao* without rhyme." These are classic summaries of ancient Chinese books. The language style can also be used to compare and analyze authors, such as Li Bai and Du Fu honored as "Poetic Immortal" and "Poetic Sage". For the study of stylistics, the traditional paradigm generally starts from vocabulary, rhetoric, sentence patterns, etc., with the help of examples, and forms an interpretive logic that is now called "close reading". Corresponding to this is the "distance reading" after the rise of digital humanities. With the help of many computational methods such as lexical statistics, quantitative linguistics, and natural language processing, the study of textual style has increasingly inclined towards results with precision value. This research paradigm, or computational humanities, provides new exploration perspectives for studying stylistics.

This study focuses on the style of ancient books in computational humanities. Compared with traditional methods, the advantage of the computational humanities lies in quantification, which is based on data and computation to obtain objective and verifiable conclusions. The study of stylistics under this paradigm also presents a variety of technical and theoretical frameworks due to the intersection of fields, forming a developing trend of mutual integration. This study of the style of ancient books depends on multi-level observations from Chinese characters to vocabulary and sentences. Representation learning can also provide more comprehensive quantification for the style analysis of ancient books.

This research focuses on the *Three Commentaries on the Spring and Autumn Annals* and related ancient books. The *Three Commentaries* are the most important classics among ancient Chinese books and have also received much attention in computational humanities. On the other hand, stylistic studies on the difference between the *Three Commentaries* have also gained much attention. Specifically, this study will take Hong Ye's *Index on Spring and Autumn Annals and the Three Commentaries* as the data source and use text representation learning in deep learning to examine the style differences between the *Three Commentaries*. As

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 117–121,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_014

an essential content of computational humanities, this study will provide a compelling computational research idea and reference for studying style in ancient books.

2 Related Research

The *Three Commentaries on the Spring and Autumn Annals* revolve around the history of the Lu State recorded in the Spring and Autumn Period in terms of content and ideological system. Still, there are apparent differences in the writing and language style focus. Scholars often draw relevant conclusions based on a careful reading of the *Three Commentaries*(Chen, 2021), which have significantly contributed to the development of historiography but need more accurate, verifiable, and reproducible digital indicators to prove it. Moreover, the entry point for investigation is single, often only starting from a specific problem, needing a macroscopic inspection from a global perspective.

As one of the research directions of ancient Chinese text mining, the metrological research of old books has the characteristics of mature technology and diverse perspectives. According to the different properties of the research objects, it can be divided into different research levels, such as vocabulary, sentences, and text. The measurement research of ancient books based on vocabulary includes word segmentation(Huang et al., 2015), part-of-speech tagging(Zhang et al., 2021), named entity recognition(Liu and Wang, 2018), keyword extraction(Qin and Wang, 2020), etc. In the quantitative research conducted around the sentence level, taking sentence segmentation (Zhao et al., 2022), sentence classification(Liu et al., 2013), and sentence extraction(Zhou et al., 2021) as examples, it is possible to explore the implicit features and inter-sentence relationships of sentences in ancient books. Research at the chapter level includes research on automatic summarization(Xu et al., 2022), bibliographic information measurement(Tong et al., 2021), etc. In summary, studying computational humanities in ancient books has extended research in various directions at different levels and has gradually formed a mature research paradigm. However, there is still a gap in the study of the style and style of ancient books, and there needs to be more research that takes ancient books as the main body and uses quantitative analysis to observe the differences in digital indicators of ancient books. This study takes the *Three Commentaries on the Spring and Autumn Annals* as the research object. It aims at texts of different levels to explore the language style differences formed in writing the three biographies. This is significant for exploring ancient books in the Spring and Autumn Periods.

3 Style Comparison With Text Representation

Words and sentences are important objects in stylistics research; different from the measurement of word frequency in traditional diagrams, this study uses representation learning models in deep learning to automatically obtain the vectorized representation of words and sentences to acquire knowledge about vocabulary and sentence style. Specifically, Word2vec and Sentence-BERT were chosen respectively for vectorizing words and sentences in the *Three Commentaries*, and mapping scatter points with dimensionality reduction was used for the style comparison.

3.1 Model Introduction

Word2vec is a neural network language model that can capture semantic information between contexts, map each word into a word vector, and mine connections between words(Mikolov et al., 2013). The Word2vec model contains two models for training word vectors: the CBOW (Continuous Bag-of-Words Model) model and the Skip-gram model. The former uses N words before and after the feature word to predict the word, and the latter uses the word's context to predict N words before and after. The Word2vec model adds contextual analysis to the context, which makes the semantic analysis more abundant.

Sentence BERT (SBERT for short) is a sentence vector computing model proposed (Reimers and Gurevych, 2019), which maps text into Vector space in sentence units. One vector can represent the semantics expressed by a sentence in the text. Compared with the BERT model, SBERT can better generate sentences. The Embedding vector enables the vectorized expression to carry more semantic information.

3.2 Word Vectorization Mapping

As was written in the name, the *Three Commentaries* were all commentaries on *Spring and Autumn Annals*, which provided detailed descriptions of historical events of the State of Lu in the Spring and Autumn period. Their themes and contents

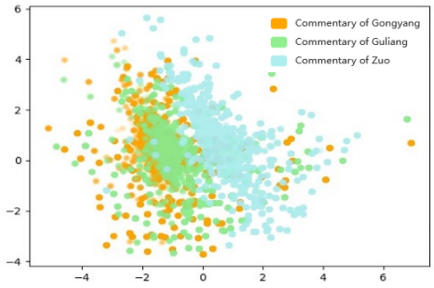

Figure 1: The mapping of words in the *Three Commentaries*

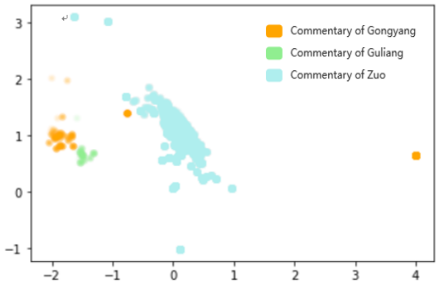

Figure 2: The mapping of single-occurrence words in the *Three Commentaries*

are similar to a certain extent, but their language styles are different. Based on this, it can be considered that the differences shown in the mapping on the scatterplot are more due to the differences in the language styles of the *Three Commentaries*, rather than the "fixed collocation" between words, that is, the differences caused by different recorded content. We use the word-segmented text to train the Word2vec model and generate words into multi-dimensional word vectors. To map in two-dimensional space, PCA is used to reduce the dimensionality of word vectors and map them on the graph through different colors.The distribution of words and single-occurrence words of the three biographies is shown in Figure 1 and Figure 2.

Each dot represents a word in the scatter plot, and the three colors correspond to the *Three Commentaries*. For example, a blue dot represents a word in the Commentary of Zuo. The number of points determines the depth of the color at the coordinate position. Since a point represents a word, the distance between points represents the degree of similarity between the two words. It can be observed that the three biographies have a slight overlap near the point (0, 0) in Figure 1. In addition, the blue word points representing the *Commentary of Zuo* are mainly concentrated in the upper right corner of Figure 1, with a relatively clear and intuitive boundary between the *Commentary of Gongyang* and the *Commentary of Guliang*. Based on this, from the perspective of words, even though the content is similar, the three biographies still have differences in language style. The *Commentary of Gongyang* and the *Commentary of Guliang* are identical in language style and preferred word definition. At the same time, the *Commentary of Zuo* has a unique narrative style that prefers supplementary historical events.

Single-occurrence words refer to the words that occurred only in one of the *Three Commentaries*.Compared with some more general words, these words that only appear in certain commentaries can better reflect the language habits in the process of writing the book. On the map of single-occurring words, there is almost no overlapping part, which conforms to the definition concept of single-occurring words, which can explain that based on similar content, the language styles of the *Three Commentaries* are different in terms of words. And the distribution of each commentaries point is consistent with that shown in Figure 2. The above image also reflects the orange dots near points (4,1). The distribution of single-occurrence words deepens the accuracy and credibility of the above picture from the side.The distribution of single-occurrence words deepens the accuracy and credibility of the above picture from the side, which confirms that the *Three Commentaries* not only show differences in the overall language style but also have different habits in the use of single-occurrence words.

3.3 Sentence Vectorization Mapping

The process of generating sentence vectors is to use the text after the sentence to train the SBERT model, and each generated vector represents a sentence. Similar to the word vector dimensionality reduction method, PCA is used to reduce the dimensionality of the sentence vector so that it can be presented on a two-dimensional graph. The distribution of sentence vectors is shown in Figure 3.

Each point in Figure 3 represents a sentence, and the distance between points represents the sentence's similarity. In the sentence vector, it is observed that the dispersion of the sentence vector is slightly smaller than that of the word vector, and the overlapping area is larger around the point (0,0). Most of Figure 3's color blocks are composed of

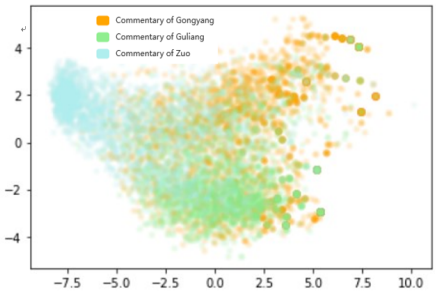

Figure 3: The distribution of sentence vectors in the *Three Commentaries*

mixed and interlaced colors. But the *Commentary of Zuo* still shows differences, converging into a single color block around the point (-7.5, 2). This phenomenon is consistent with what the word vectors offer, and it reflects that the *Commentary of Zuo* is significantly different from the other two biographies in style.

Based on the mapping results of word vectors and sentence vectors, the style of the *Commentary of Gongyang* and the *Commentary of Guliang* are relatively similar, and the *Commentary of Zuo* shows distinct style differences. This result aligns with the views of ancient and modern scholars who have carefully read *Three Commentaries on the Spring and Autumn Annals* and can deepen the conclusion that the *Commentary of Zuo* is different in language style.

4 Conclusion

From the perspective of Natural language processing, uses a deep learning model with good versatility to calculate the language style differences between different levels of the *Three Commentaries on the Spring and Autumn Annals*. It concludes that the *Commentary of Zuo* differs from the *Commentary of Gongyang*, and the *Commentary of Guliang*, realizing the mining research on the language characteristics of ancient books.

In the follow-up research, we will use other methods to examine the differences between the *Three Commentaries*. From the perspective of natural language processing, this study has verified the feasibility of the general language model in discovering the differences in the *Three Commentaries*, and subsequent language models suitable for ancient Chinese, such as GuwenBERT, SikuBERT, and other pre-training based on ancient Chinese domain data enhancement model to further observe differences in language styles. In addition, the difference between the *Three Commentaries* can be observed from multiple perspectives, such as automatically mining different types of entities, or observing the usage habits of words from the part of speech.

From the perspective of quantitative linguistics, the language style differences between the *Three Commentaries* will be observed through different levels of language measurement indicators. At the word level, the index selects the average word length to measure the difference in word length, selects the word density, standard type ratio, and single word ratio to measure the difference in the richness of the *Three Commentaries* vocabulary, and observes the information carrying capacity of the richness of the *Three Commentaries* by calculating the entropy of text information. At the sentence level, the writing characteristics of the *Three Commentaries* were examined through average sentence length, sentence dispersion, sentence fragmentation, and other indicators. From the above two perspectives, the linguistic characteristics and stylistic differences of the *Three Commentaries on the Spring and Autumn Annals* can be examined from a new perspective, which provides new verification ideas for the research related to *Three Commentaries on the Spring and Autumn Annals*.

5 Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant number 72004095], and the National Social Science Fund of China [grant number 21&ZD331].

References

Wangheng Chen. 2021. Political Aesthetic Thoughts of the Three Commentaries on the Spring and Autumn Annals. *Wuhan University Journal (Philosophy & Social Science)*, 74(6):80–94.

Shuiqing Huang, Dongbo Wang, and Lin He. 2015. Discussion on Automatic Word Segmentation of Pre-Qin Classics Using Sinological Index Series as the domain vocabulary. *Library and Information Service*, 59(11):127–133.

Liu Liu, Bin Li, Weiguang Qu, and Xiaohe Chen. 2013. Automatic Acquisition of Age Characteristics of Pre-Qin Vocabulary and Automatic Judgment of Document Age. *Journal of Chinese Information Processing*, 27(5):107–113.

Liu Liu and Dongbo Wang. 2018. A Review of Named Entity Recognition. *Journal of the China Society for Scientific and Technical Information*, 37(3):329–340.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Heran Qin and Dongbo Wang. 2020. Application of Keyword Extraction in Pre-Qin Ancient Chinese under the Digital Humanities——Taking Chun Qiu Jing Zhuan as an Example. *Library Journal*, 39(11):97–105.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv:1908.10084 [cs].

Lin Tong, Bing Liu, Jingpeng Deng, Dajun Liu, Yongsheng Yang, Xueyuan Hu, Zefeng Kang, and Ruili Huo. 2021. Quantitative analysis and thinking on bibliographic information of existing ancient Chinese ophthalmology books. *Journal of Traditional Chinese Ophthalmology*, 31(6):449–452.

Runhua Xu, Dongbo Wang, Huan Liu, Yuan Liang, and Kang Chen. 2022. Research on Automatic Summarization of History as a Mirror for the Digital Humanities of Ancient Books——Taking the SikuBERT Pre-training Model as an Example. *Library Tribune*, 42(12):129–137.

Qi Zhang, Chuan Jiang, Youshu Ji, Minxuan Feng, Bin Li, Chao Xu, and Liu Liu. 2021. Construction of an automatic tagging model for part-of-speech integration of word segmentation for multi-domain pre-Qin classics. *Data Analysis and Knowledge Discovery*, 5(3):2–11.

Lianzhen Zhao, Yiqin Zhang, Jiangfeng Liu, Dongbo Wang, Minxuan Feng, and Bin Li. 2022. Research on Automatic Punctuation of Pre-Qin and Han Classics for Digital Humanities—Taking SikuBERT Pre-training Model as an Example. *Library Tribune*, 42(12):120–128+137.

Hao Zhou, Dongbo Wang, and Shuiqing Huang. 2021. Research on Automatic Recognition of Context of Citations in Ancient Books——Taking Commentary Documents as an Example. *Information studies: theory & application*, 44(9):169–175.