A Joint Model of Automatic Word Segmentation and Part-Of-Speech Tagging for Ancient Classical Texts Based on Radicals

Bolin Chang[1,2], **Yiguo Yuan**[2], **Bin Li**[1,2]✉, **Zhixing Xu**[1,2], **Minxuan Feng**[1,2], **Dongbo Wang**[3,2]
[1]School of Chinese Language and Literature, Nanjing Normal University, Nanjing, China
[2]Center of Language Big Data and Computational Humanities, Nanjing Normal University, Nanjing, China
[3]College of Information Management, Nanjing Agricultural University, Nanjing, China
✉libin.njnu@gmail.com

Abstract

The technique of word segmentation and part-of-speech tagging in ancient Chinese plays a crucial role in the field of information processing in ancient Chinese. The current state of ancient Chinese word segmentation and part-of-speech tagging technology presents pressing issues that require immediate attention, such as low accuracy and efficiency. This study employs a methodology that combines word segmentation and part-of-speech tagging. It establishes a correlation between fonts and radicals, trains the Radical2Vector radical vector representation model, and integrates it with the SikuRoBERTa word vector representation model. Finally, it connects the BiLSTM-CRF neural network.The study investigates the combination of word segmentation and part-of-speech tagging through an experimental approach using a specific data set. In the evaluation dataset, the F1 score for word segmentation is 95.75%, indicating a high level of accuracy. Similarly, the F1 score for part-of-speech tagging is 91.65%, suggesting a satisfactory performance in this task. This model enhances the efficiency and precision of the processing of ancient books, thereby facilitating the advancement of digitization efforts for ancient books and ensuring the preservation and advancement of ancient book heritage.

1 Introduction

The challenge of automatically segmenting words and assigning part-of-speech tags to ancient Chinese text is a crucial area of study within the discipline of natural language processing. The primary objective of this project is to employ computer technology for the precise identification of word boundaries in ancient Chinese writings, as well as the exact assignment of appropriate part-of-speech labels to these words, including nouns, verbs, conjunctions, and others. By implementing this procedure, the conventional task of manual labelling can be efficiently alleviated, leading to notable enhancements in both labelling efficiency and accuracy. The progress of this technology not only facilitates the processing of ancient Chinese texts, but also exerts a significant influence on interconnected disciplines, including literature, history, philology, and digital humanities.

The maturation of ancient Chinese automatic word segmentation and part-of-speech tagging technologies is occurring with the ongoing advancement of computer technology. Nevertheless, the current utilisation of these methodologies is confronted with two distinct obstacles as a result of the numerous distinctive characteristics of ancient Chinese. Firstly, it is worth noting that there remains potential for enhancing the precision of word segmentation and part-of-speech labelling in the context of ancient Chinese. While several technologies currently available have the capacity to partially substitute manual labelling, their level of accuracy falls short of totally replacing manual labelling. Consequently, a significant amount of proofreading labour is necessary during the later stages. Secondly, there is a need for additional enhancement in the effectiveness of word segmentation and part-of-speech tagging in the context of ancient Chinese. The prevailing approach involves conducting word segmentation as the initial step, followed by part-of-speech tagging. Nevertheless, this sequencing will result in diminished processing efficiency and has the potential to propagate errors in word segmentation to the subsequent part-of-speech tagging phase, so exacerbating the influence of these errors and subsequently diminishing overall accuracy.

This paper utilizes the Word2Vec model to incorporate the radical information of Chinese characters. It proceeds to train the Radical2Vector model and combines it with SikuRoBERTa to form the Embedding layer. Subsequently, the BiLSTM-CRF neural network is connected to conduct an experiment on the integration of word segmentation and part-of-speech tagging in ancient Chinese. The

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 122–132,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_015

utilization of ancient Chinese word segmentation and part-of-speech tagging facilitates the exploration of profound insights within ancient texts, thereby advancing the digital advancement and utilization of these texts. Furthermore, it contributes to the preservation and progression of ancient literary works.

2 Related Work

The co-examination of automatic word segmentation and part-of-speech tagging in the context of ancient Chinese is a common area of research. Huang (2002) conducted a study on part-of-speech tagging in ancient Chinese using the hidden Markov model. They applied this model to analyze "The Analects of Confucius" and "Tao Te Ching". Although the study employed a set of 22 part-of-speech tags, it made significant contributions to the field. In their study, Fang (2009) developed a text segmentation program called Yu Segmentation Program. The researchers focused on ancient books such as "The Classic of Tea" and employed a model algorithm that utilized tree pruning to achieve efficient text segmentation of these classical texts. The F1 score for word segmentation has been reported to be approximately 86% by Min Shi (2010). A comparative experiment was conducted to evaluate the performance of the Conditional Random Fields (CRF) model in the tasks of automatic word segmentation, part-of-speech tagging, and integration of ancient Chinese. Both features and integrated processing contribute to the enhancement of the F1 value. Runhua Xu (2012) proposed a method that utilizes structured annotations to enhance the word segmentation process. In their study, Shuiqing Huang (2015) employed the CRF model to analyze word categories, phonetics, and probability features. Notably, their analysis yielded a remarkable F1 value of 97.47%. According to the study conducted by Xiaoyu Wang (2017), This paper examines the issue of automatic word segmentation in Middle Ancient Chinese by employing a combination of the CRF model and a dictionary. It also investigates the impact of inconsistent word segmentation on the results of artificial word segmentation in Middle Ancient Chinese through experimental analysis. Additionally, the paper introduces character classification as part of the research methodology. The dictionary information exhibits two notable features. Firstly, the word segmentation F1 value achieved a remarkable accuracy rate of over 99% in the closed test. Secondly, in the open test, the word segmentation F1 value ranged between 89% and 95%, further highlighting the effectiveness of the dictionary information. Ning Cheng (2020) employed the Word2Vec-BiLSTM-CRF model to investigate the amalgamation of part-of-speech tagging for sentence segmentation and part-of-speech analysis in ancient Chinese texts.

Numerous studies have been conducted on the utilization of vector representations of strokes, parts, components, and radicals to facilitate Chinese information processing, both in contemporary and ancient contexts. In their study, Tao (2019) introduces a new model called Dual-channel Word Embedding (DWE) that aims to effectively capture both sequential and spatial information of characters. The author argues that this model demonstrates a logical and advantageous approach in representing the morphology of Chinese language. In their study, Zhang (2021) presents a novel model called the Feature Subsequence based Probability Representation Model (FSPRM) for the purpose of acquiring Chinese word embeddings. The model incorporates both morphological and phonetic features, specifically stroke, structure, and pinyin, of Chinese characters. By designing a feature subsequence, the model captures a wide range of semantic information pertaining to Chinese words. The efficacy of the proposed method is substantiated through a series of comprehensive experiments conducted on various tasks including word analogy, word similarity, text classification, and named entity recognition. The results of these experiments consistently indicate that the proposed method surpasses the performance of the majority of existing state-of-the-art approaches. In the study conducted by Shi (2015), a novel deep learning technique referred to as "radical embedding" is introduced. The author provides a rationale for this approach by drawing upon principles derived from Chinese linguistics. Furthermore, the feasibility and usefulness of this technique are assessed through a series of three experiments. In their study, Yu (2017) presents a method for simultaneously embedding Chinese words, characters, and subcharacter components at a detailed level. The performance of our model is shown to be superior through evaluation on both word similarity and word analogy tasks. In their study, Han (2018) utilized a shared radical level embedding approach to address the task of Simplified and Traditional Chinese Word Segmen-

tation. Notably, their method does not require any additional conversion from Traditional to Simplified Chinese. The integration of radical and character embeddings results in a reduction in parameter count, while facilitating the sharing and transfer of semantic knowledge between the two levels. This integration significantly enhances performance. In their recent publication, Tang (2021) introduces a pioneering model named Moto, which aims to enhance embedding through the incorporation of multiple joint factors. The empirical findings indicate that our Moto model attains state-of-the-art performance with an F1-score of 0.8316, representing a 2.11% improvement, when applied to Chinese news titles. Furthermore, it achieves an accuracy of 96.38 (a 1.24% improvement) on the Fudan Corpus dataset and 0.9633 (a 3.26% improvement) on the THUCNews dataset. Among the various research endeavors, the investigation into the utilization of radical vectors stands out as the most prominent. On one hand, this phenomenon can be attributed to the relatively straightforward acquisition of the corresponding relationship data between Chinese characters and their radicals. On the other hand, the inclusion of radical vectors has been found to enhance the efficacy of Chinese information processing tasks.

It is evident that among the aforementioned studies, only one specifically addresses the topic of ancient Chinese classical Chinese, with a specific focus on automating sentence segmentation tasks. The absence of vector representations for strokes, parts, components, and radicals in ancient Chinese information processing has the potential to enhance the morphology of ancient books. This article endeavors to analyze the impacts of research. This study exclusively focuses on the radical vector representation and application of ancient Chinese characters, primarily due to limited resources.

3 Model Architecture

3.1 Embedding

The embedding layer, also known as the input layer, is a fundamental component in neural network architectures. It is responsible for transforming input data into enhancing the caliber of vectorized representation of historical Chinese text within the coding layer of the model constitutes a pivotal aspect in advancing the automated processes of sentence segmentation and word segmentation in ancient Chinese. In order to utilize natural language as input for the neural network model, it is necessary to convert it into a vector representation. The BERT model, constructed by Transformer's bidirectional encoder, is currently one of the most advanced technologies for language vector representation. Therefore, this research has opted to utilize the BERT model. The SikuRoBERTa model serves as the foundational approach for generating vector representations of Chinese characters. SikuRoBERTa is a vector representation model developed by Wang Dongbo et al. that is specifically designed for ancient Chinese. This model is built upon the BERT architecture. The training corpus utilized in this study is the renowned Wenyuange "Siku Quanshu" collection, which consists of approximately 500 million word instances. The word list encompasses a total of 21,128 characters.

The exclusive reliance on Chinese character vectors is insufficient in fully capturing the interrelationships among Chinese characters. It is imperative to delve into a comprehensive characterization of the intrinsic information embedded within Chinese characters. Chinese characters are a form of semantic and phonetic characters, wherein the radicals, components, and even strokes of these characters possess a certain capacity to convey meaning. Hence, in the domain of character-based sequence labeling, the inclusion of semantic information from these entities is frequently employed to enhance the precision of lexical analysis. precision. Firstly, it is imperative to differentiate between the four concepts of strokes, parts, component, and radicals of Chinese characters. This article aligns with the principles outlined in "A General Theory of Modern Chinese" edited by Jingmin Shao (2017).

(1) The stroke represents the fundamental building block of regular script glyphs.

(2) The part refers to a unit of character construction in the Chinese writing system. It is comprised of strokes, can be utilized autonomously, and serves the purpose of constructing Chinese characters. Components can also be considered as units of word formation that are derived through one or more segmentations of the complete word.

(3) The component refers to the structural component obtained through a single segmentation of the combined character using the dichotomy method.

(4) The radical is component or subcomponent that can combine to create characters in groups. The characters that share a common component

are grouped together in the "character set", with this component being positioned at the forefront as the leading unit. This arrangement serves as the foundation for character retrieval.

Using the character "時" as a case study, Table 1 reveals the presence of a shared denotative symbol "日" in the parts, components, and radicals of "時". This symbol serves as a pictograph, also known as a meaning, for "時" characters. However, it is important to note that there is no direct and exclusive correspondence between the pictographs and radicals found in Chinese characters. In certain instances, the complete representation of a Chinese character necessitates the inclusion of all its constituent components or radicals. For instance, the pictograph for the character "闖" is denoted by the combination of "門" and "馬", which collectively convey the meaning of door. This character "膽", in turn, signifies "肉".

Word building unit	Composition of "時"
strokes	一一一一一一一一一一
parts	日士寸
components	日寺
radicals	日

Table 1: The strokes, parts, components and radicals of "時"

This paper establishes a mapping between fonts and radicals based on a dataset comprising over 70,000 Chinese characters and their corresponding radicals. Subsequently, the ancient Chinese traditional corpus known as "Siku Quanshu" is converted into radicals using the established mapping relationship between fonts and radicals. Please refer to Figure 1. This study utilizes the radical corpus and employs the Word2Vec training methodology to train the Radical2Vector model, which represents radical vectors. The Word2Vec algorithm is widely recognized as a prominent method for training word vectors. It effectively maps words or radicals onto a continuous vector space, enabling the identification and representation of semantic and morphological similarities among them. By utilizing the Radical2Vector model that has undergone rigorous training, it is possible to acquire the vectorized representation of individual radicals.

While the radical vector does contain internal information pertaining to Chinese characters, its informational capacity is restricted. Consequently, it cannot serve as a standalone vector representation for Chinese characters, necessitating its utilization in conjunction with word vectors. There exist two methods for integrating character vectors and radical vectors. The first approach involves concatenating the radical vectors and character vectors to form extended vectors, which are subsequently fed into a Bidirectional Long Short-Term Memory (BiLSTM) feature extractor. The second method entails combining the radical vectors and character vectors without further elaboration. The vectors are inputted into two distinct BiLSTM feature extractors, with the exception of the hidden size, the hyperparameters of these two feature extractors remain consistent. In conclusion, it is imperative to meticulously adjust the hyperparameters in order to achieve the most optimal radical vector representation model and input methodology.

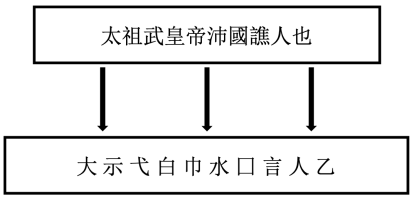

Figure 1: Transformation from traditional Chinese corpus to radical corpus

3.2 Neural Networks

The neural network layer is connected subsequent to the Embedding layer. The neural network architecture comprises two distinct layers, BiLSTM layer and CRF layer.

The BiLSTM is a type of neural network that incorporates bidirectional long short-term memory units. The recurrent neural network under consideration possesses the capability to effectively model sequential data. The BiLSTM model encompasses both forward and backward directions, enabling the simultaneous consideration of contextual information. This characteristic renders it highly effective for tasks involving sequence labeling. By utilizing BiLSTM, the model is able to acquire a greater amount of global semantic information.

The CRF model, also known as the conditional random field, is a statistical model used in machine learning and pattern recognition. The proposed approach is a statistical model designed for sequence labeling tasks, with the capability to optimize the labeling results on a global scale. Given that the output of the BiLSTM model is a probability matrix, it can be observed that the outcomes at each

time step are mutually independent. Consequently, the impact of the preceding label on the current label cannot be taken into account. To address this issue, the current innovation opts for CRF model and integrates it following BiLSTM model. The CRF is a graph model that can be used to represent the joint probability distribution of a label sequence given an observation sequence. It is commonly employed to enforce constraints on the labeling results produced by BiLSTM model, ensuring that the output labels adhere to the rules of a valid sequence. Furthermore, the CRF can also be utilized to compute the optimal solution of the BiLSTM output sequence, thereby enhancing the effectiveness of sequence labeling.

The Embedding layer incorporates both word vectors and radical vectors, resulting in the formation of two distinct model structures when the neural network is spliced. These structures are illustrated in Figure 2 and Figure 3. Based on the analysis of Figure 2 and Figure 3, it is evident that the two input methods for radical vectors exhibit distinct characteristics. The former approach involves the concatenation of word vectors and radical vectors within the embedding layer, requiring the construction of a set of hidden layers using BiLSTM. Conversely, the latter method necessitates the integration of radical vectors with other components. The word vector and radical vector are separately fed into two distinct BiLSTM hidden layers in order to generate two sets of BiLSTM feature vectors. These LSTM feature vectors are subsequently concatenated.

4 Integrated Labeling Strategy

The tasks of Chinese automatic word segmentation and part-of-speech tagging are typically performed independently, with the outcome of automatic word segmentation serving as the foundation for part-of-speech tagging. Hence, the general approach in Chinese lexical analysis involves the sequential implementation of automatic word segmentation followed by part-of-speech tagging. The concept of integrated tagging can be attributed to Shuanhu Bai (1996), who proposed a combined approach for word segmentation and part-of-speech tagging to address the issue of ambiguous domains in contemporary Chinese automatic word segmentation. However, Shuanhu Bai did not conduct a comprehensive assessment of the practicality of integrated tagging. Ng (2004) provide a comprehensive analysis of the viability of integrated tagging in their scholarly work. The authors conducted a comparative analysis of two strategies for Chinese word segmentation, namely part-of-speech tagging and integrated tagging, using the maximum entropy model. The findings indicate that the integrated method, which relies on word annotation, demonstrates superior performance. The initial utilization of the integrated tagging method in the domain of ancient Chinese can be attributed to the research conducted by Min Shi (2010). The CRF model was employed to carry out experiments on word segmentation and part-of-speech tagging for Pre-Qin Chinese. The findings of the study indicated that the integrated strategy was effective. In comparison to the two-step strategy, it demonstrates a notable enhancement in the efficacy of word segmentation and part-of-speech tagging. Hence, this study also employs an integrated labeling approach. To achieve integrated labeling, the output label of each word is determined by combining the word's position and its corresponding part of speech. The lexical tagging system for word position information consists of a set of four lexemes: B for begin, I for middle, E for end, and S for a single word. The "Basic Collection of Pre-Qin Chinese Parts of Speech Tags" prescribes the use of part-of-speech tags. For instance, the tag "v" is employed to indicate verbs, while the tag "n" is used to indicate nouns, among others. The hyphen (-) serves as a connector between the lexeme marker and the part-of-speech marker. An illustration of this can be seen in the token B-v, which represents a word that initiates a verb.

5 Experiment

5.1 Dataset

The selection of "Zuo Zhuan" as the experimental corpus is based on the following rationales: The "Zuo Zhuan" holds the distinction of being the inaugural chronicle history book in our nation's history, encompassing a comprehensive narrative. Furthermore, it boasts the highest word count among all pre-Qin literature publications. The extensive body of literature, consisting of over 200,000 characters, is well-suited for conducting automated word segmentation and part-of-speech tagging experiments on ancient Chinese through the application of deep learning techniques. Furthermore, the reliability of the electronic corpus of "Zuo Zhuan" utilized in this study is reasonably assured. In ad-

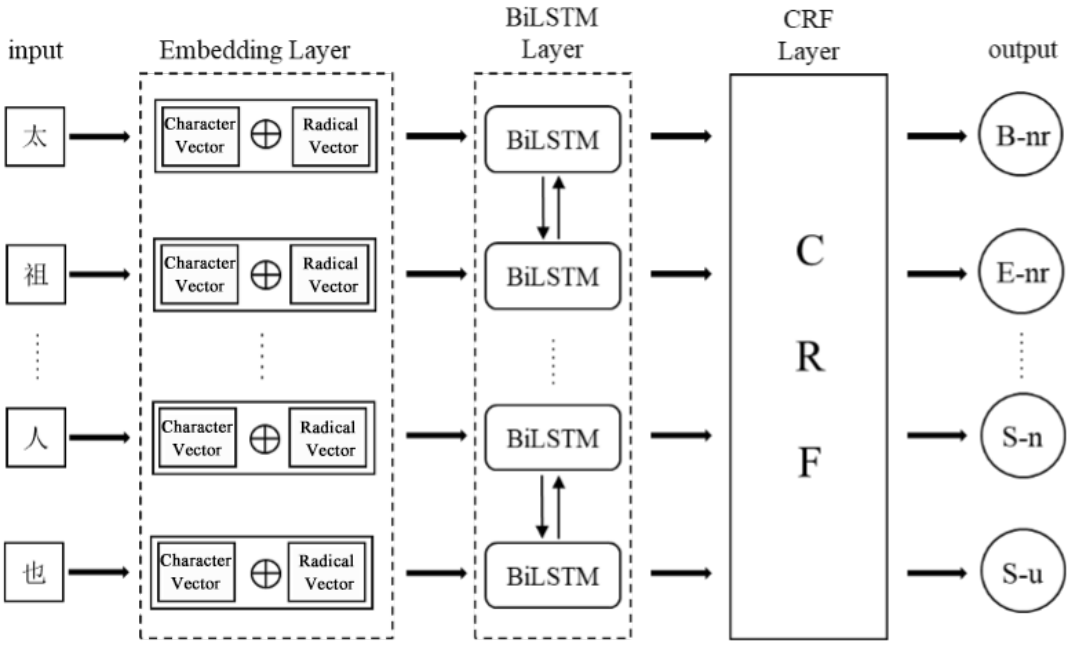

Figure 2: The first input method of radical vector

dition to addressing punctuation and collation, the research group also examined the matter of variant texts in relation to Yang Bojun's (1990) work titled "Spring and Autumn Zuozhuan Zhuan". Furthermore, our research team has conducted artificial segmentation and tagging of the electronic corpus of "Zuo Zhuan". The aforementioned tagged corpus exhibits a commendable level of quality and is deemed appropriate for utilization as an experimental corpus. In their respective studies, Min Shi (2010), Chengming Li (2018), and Ning Cheng (2020) employed the "Zuo Zhuan" as the corpus for conducting automated lexical analysis of ancient Chinese. In order to facilitate a meaningful comparison with their experimental findings, it is imperative for this study to employ the identical "Zuo Zhuan" annotated corpus during the experimentation process.

Hence, the partitioning of the "Zuo Zhuan" dataset in this study aligns with the experimental design of the baseline model. Specifically, the initial ten volumes of "Zuo Zhuan" serve as the training corpus, while the final two volumes are utilized as the test corpus. Table 2 displays the precise scale of the experimental set.

Dataset	Tokens	Types
Training set	194,995	166,141
Test set	33,298	28,131

Table 2: "Zuo Zhuan" training set and test set size

Among the datasets, the ratio of word case occurrences in the training set to the test set is approximately 5.86, while the ratio of overall word case occurrences is approximately 5.91. In general, the training set is approximately six times larger than the test set in terms of size ratio.

This study employs the conventional word tagging technique to accomplish the task of automated lexical analysis. To do so, we must develop a tag set that is suitable for both word segmentation and part-of-speech tagging tasks.

5.2 Equipment and Environment

The model employed in this study was constructed using the PyTorch 1.7.1 framework, with the programming language of choice being Python 3.8. Regarding the system configuration, the central processing unit (CPU) employed is the Intel i7-13700F operating at a clock speed of 2.90GHz. The memory capacity of the system amounts to 64GB, while the graphics processing unit (GPU) utilized is the NVIDIA GeForce RTX 4090. Furthermore, the memory size associated with the GPU is 24GB. This particular system configuration has the capability to guarantee both the efficiency and speed of model training.

5.3 Hyper-parameters

Radical2Vector can be described as a vector representation model that captures the essence of ancient

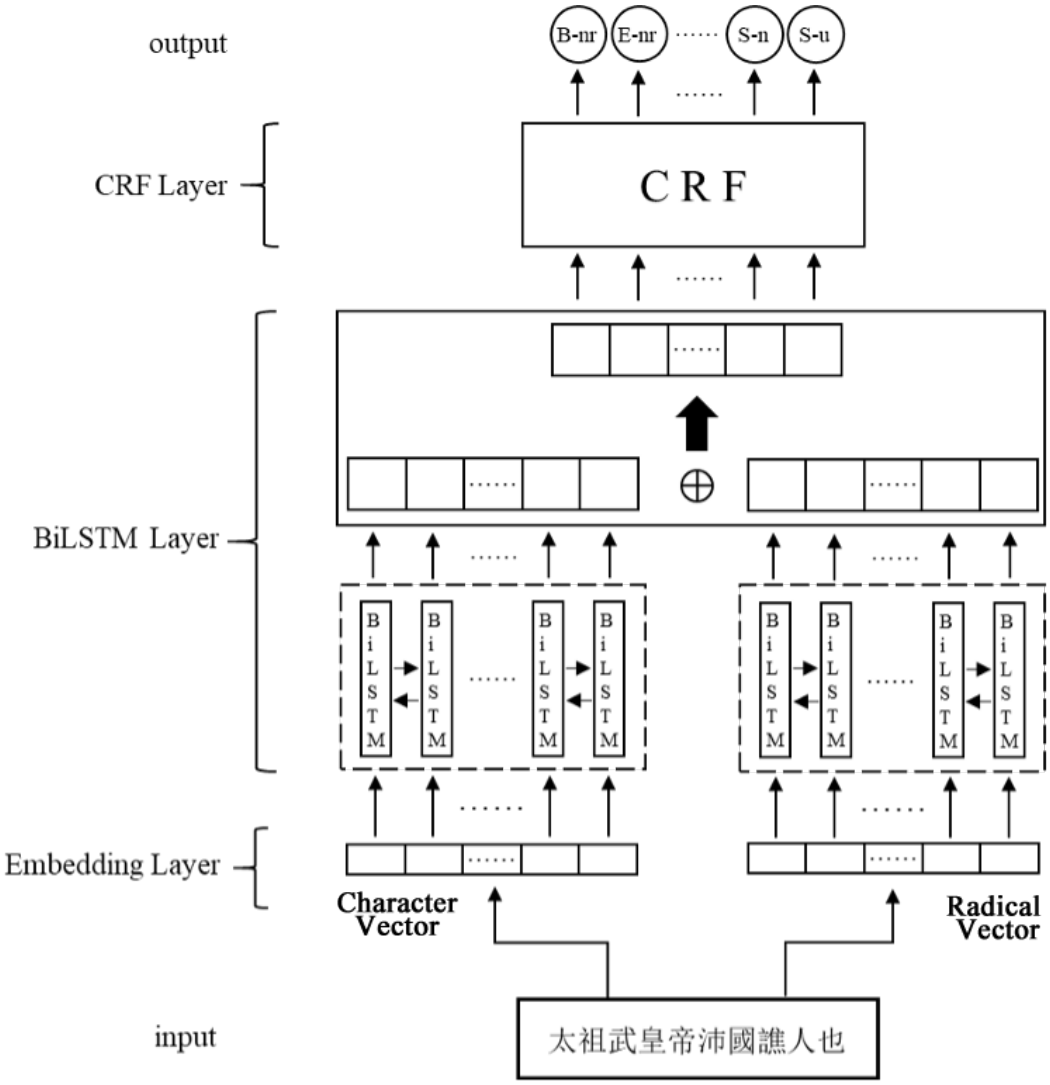

Figure 3: The second input method of radical vector

Chinese radicals. This model is constructed by applying the Word2Vec training method to a radical corpus sourced from "Siku Quanshu," which contains over 700 million word examples. During the training process of Word2Vec models, it is common to encounter the need for adjusting four key hyperparameters. These hyperparameters include the choice of training algorithm, which encompasses both Continuous Bag of Words (CBOW) and Skip-Gram methods, as well as the feature vector dimension, the number of iterations, and the window size. The CBOW model is a technique that utilizes contextual information to predict the current word or words as part of a training task. On the other hand, the Skip-Gram model is a method that employs the current word or words to predict the surrounding context as part of a training task. The training tasks. The dimension of the feature vector is a crucial parameter in the Word2Vec model as it dictates the size of the vector representation for words or radicals in the continuous space. The term "number of iterations" pertains to the frequency at which the corpus is traversed during the training process. In each iteration, the parameters of the model will be updated in order to optimize the vector representation of words or radicals. The term "window size" pertains to the maximum distance separating the context and the present word (or words), thereby determining the extent of the context. This paper combines various hyperparameter selections as outlined in Table 3 and conducts an initial experiment for parameter tuning.

This study initially selects the initial splicing technique of word vector and radical vector, and proceeds to conduct a comparative experiment on the training algorithm. Initially, the CBOW and Skip-Gram models underwent training with vector dimensions of 128, 256, and 512, respectively. This

Hyperparameters	Value
training algorithm	CBOW/Skip-Gram
vector dimension	128/256/512/768
iterations	5/10/15/20/25
window size	3/4/5/6/7/8

Table 3: Hyperparameters for Radical2Vector Model

Model name	Word segmentation	POS tagging
CBOW-128d	95.73	91.54
CBOW-256d	95.56	91.45
CBOW-512d	**95.75**	**91.65**
Skip-128d	95.58	91.38
Skip-256d	95.73	91.35
Skip-512d	95.63	91.35

Table 4: F1 value (%) of CBOW and Skip-Gram models with different radical vector dimensions

training was conducted with iteration number 10 and window size 5. The resulting models were labeled as CBOW-128d, CBOW-256d, CBOW-512d, Skip-128d, Skip-256d, and Skip-512d. Next, employ the initial radical vector input approach to carry out a comprehensive experiment involving word segmentation and part-of-speech tagging on the "Zuo Zhuan" dataset. The empirical findings are presented in Table 4.

Figures 4 and 5 display the performance of CBOW-512d across various iterations with a window size of 5, as well as the performance of CBOW-512d across different window sizes with a fixed number of iterations at 10. These figures aim to investigate the impact of the number of iterations and window size on the model's influence.

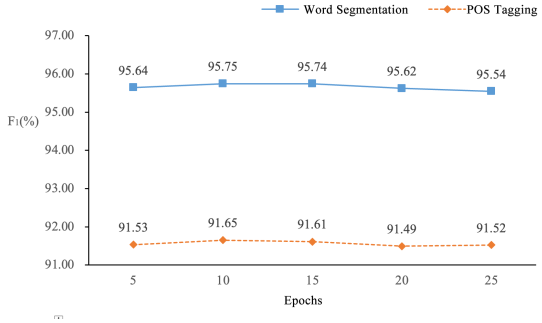

Figure 4: The performance of CBOW-512d at different iterations when the window size is 5

The chart illustrates that the CBOW-512d model demonstrates the most favorable outcome. Additionally, the Word2Vec method's radical vector

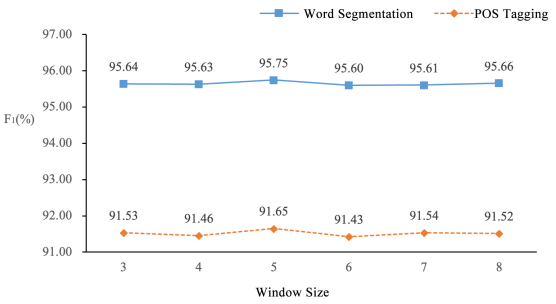

Figure 5: The performance of CBOW-512d at different iterations when the window size is 5

training does not significantly contribute to word segmentation; however, it does enhance the efficacy of part-of-speech tagging. The Skip-Gram approach is not deemed appropriate for the training of radical vectors. To ascertain the potential enhancement of the model's performance with a larger radical vector dimension, this study employs the Continuous Bag-of-Words (CBOW) approach to train a model with a vector dimension of 768, referred to as CBOW-768d. In comparison to the CBOW-512d model, the integrated tagging of this model exhibits a decrease of 0.1 in the F1 value for word segmentation and a decrease of 0.17 in the F1 value for part-of-speech tagging.

In general, the selection of the CBOW training method and the configuration of a vector dimension of 512 are deemed more suitable. This study made adjustments to the number of iterations and window size of CBOW-512d, based on the given rationale. Based on the findings presented in Figure 4 and Figure 5, this study ultimately determines that the optimal number of iterations for CBOW-512d is 10, while the most effective window size is 5. The model is referred to as Radical2Vector.

5.4 Vector Composition

The preceding data represents the performance of Radical2Vector in the initial approach of radical vector input, wherein the word vector and radical vector are combined and fed into a series of BiLSTM hidden layers to produce LSTM feature vectors. In this particular instance, there is a slight enhancement observed in the impact of part-of-speech tagging. This study employs the Radical2Vector methodology to carry out experiments pertaining to the second input modality. The experimental results of the two input methods are presented in Table **??**. The second input method of

the radical vector, Radical2Vector, exhibits minimal improvement efficacy.

6 Evaluation

In this study, the Radical2Vector model was selected as the representation model for ancient Chinese radicals. The radical vector was combined with the word vector and fed into the same set of BiLSTM hidden layers. This approach, referred to as the first radical vector input method, was employed. The incorporation of radical vectors enhances the efficacy of part-of-speech tagging; however, its impact remains somewhat constrained.

To assess the impact of the model proposed in this research paper, the evaluation metrics employed include accuracy rate (P), recall rate (R), and harmonic mean (F1). The model presented in this study is then compared to the outcomes achieved by participating teams in the open test TestA of the first international ancient chinese word segmentation and pos tagging bakeoff(Li et al., 2022). The findings are juxtaposed, as illustrated in Table 6. The training and test sets utilised in this study align with the evaluation dataset. Furthermore, the training approach and outcome statistics presented in this article adhere to the criteria outlined for open evaluation. Consequently, the model's computational findings can be compared to the evaluation results to assess its impact on word segmentation and part-of-speech tagging.

7 Discussion

Innovation can improve ancient Chinese word segmentation and part-of-speech labeling. Reasons include these. Ancient Chinese radicals are related with form and meaning. The word segmentation and part-of-speech tagging model captures ancient Chinese character structural similarities better by integrating radical information. The resemblance helps the model reliably identify and categorize ancient Chinese characters, improving word segmentation precision. Radicals also relate to ancient Chinese character semantics. Radical information helps the model learn radical semantics and apply them to part-of-speech labeling. Some radicals associate with nouns, whereas others with verbs or adjectives. Semantic information can improve part-of-speech labeling. Ancient Chinese word segmentation and part-of-speech tagging require knowledge of ancient literature and culture. Radicals are a vital part of ancient Chinese characters. Innovative information improves the word segmentation and part-of-speech tagging model's understanding of historical manuscripts' lexicon and expressions, improving ancient Chinese language processing computational capabilities.

Lexical analysis is better with integrated tagging. The integrated labeling technique reduces category labels during multi-classification tasks like lexical analysis. This improves lexical analysis. This work uses the four-lexeme tag set for automatic word segmentation and 21 part-of-speech tags for tagging. Integrated tagging reduced the training set of "Zuo Zhuan" to 59 integrated tags. Strategy has 84 category labels. This is because ancient Chinese auxiliary words (u), quantifiers (q), and concurrent words (j) were single-character terms. These linguistic elements are only combined with the single-word marker (S), not with beginning (B), medial (I), or final (E) markers. The "Zuo Zhuan" dataset contains terms without three-character words. This applies to prepositions, adverbs, modal particles, and onomatopoeia. This method reduces class labels further by adding in-word (I) tagging. Certain characters vary and limit the part-of-speech scope of their words on different lexemes. This limits character consequences. Thus, the integrated tagging technique integrates external knowledge and automated processing by leveraging the interrelated and complimentary nature of word segmentation and part-of-speech information. Thus, this study labels everything.

8 Conclusion

This study employs deep learning techniques to extract the radical information of Chinese characters, thereby achieving the integration of automatic word segmentation and part-of-speech tagging in ancient texts. This study utilizes a dataset comprising over 70,000 Chinese characters and their corresponding radicals to establish a correlation between fonts and radicals. Additionally, it employs the Radical2Vector model to train a radical vector representation. An experiment was conducted on the "Zuo Zhuan" dataset to examine the integration of word segmentation and part-of-speech tagging, utilizing the SikuRoBERTa-Radical2Vector-BiLSTM-CRF model in conjunction with the original SikuRoBERTa. The model's automatic word segmentation achieved an F1 value of 95.75% on the test set, while the automatic part-of-speech tagging achieved an F1 value of 91.65%. The present

Input Method	Task	P	R	F1
First	Word Segmentation	95.52	95.97	**95.75**
	POS Tagging	91.44	91.86	**91.65**
Second	Word Segmentation	95.38	95.85	95.61
	POS Tagging	91.08	91.53	91.31

Table 5: The integrated labeling effect of Radical2Vector on the two input methods (%)

Evaluation	Word Segmentation			POS Tagging		
	P	R	F1	P	R	F1
FDU	95.81	96.88	**96.34**	92.05	93.07	**92.56**
	95.73	96.84	96.28	91.88	92.94	92.41
ZNNU	92.78	90.18	91.46	88.97	86.48	87.71
HIT	91.2	93.49	92.33	85.41	87.56	86.47
	91.09	93.41	92.24	85.27	87.45	86.35
BLCU	90.91	92.4	91.65	83.55	84.92	84.23
	90.56	92.29	91.41	83.13	84.72	83.92
NJUPT	78.14	86.31	82.02	57.35	63.35	60.2
This article	95.52	95.97	95.75	91.44	91.86	91.65

Table 6: Comparison between the model in this paper and the results of the evaluation teams (%)

study introduces an integrated model that utilizes radicals for word segmentation and part-of-speech tagging in ancient Chinese. This model demonstrates a high level of performance, significantly enhancing the efficiency and accuracy of tagging ancient book corpora. Consequently, it facilitates the digitization process of ancient books and actively contributes to the advancement of research in this field. The topic of discussion pertains to the concepts of inheritance and development.

9 Acknowledgments

This research was supported by National Language Commission Project (YB145-41), Key Project of Ancient Books Work (22GJK006) and National Social Science Foundation of China major project (21&ZD331, 22&ZD262). We are grateful to the reviewers for comments which helped us to improve the paper.

References

Shuanhu Bai. 1996. An integrated method of chinese word segmentation and part-of-speech automatic tagging. *Chinese Information*, 2:46–48.

Ning Cheng, Bin Li, Sijia Ge, Xingyue Hao, and Minxuan Feng. 2020. A joint model of automatic sentence segmentation and lexical analysis for ancient chinese based on bilstm-crf model. *Journal of Chinese Information Processing*, 34(4):1–9.

Miao Fang, Yi Jiang, Qi Zhao, and Xin Jiang. 2009. Automatic word segmentation for chinese classics of tea based on tree-pruning. In *2009 Second International Symposium on Knowledge Acquisition and Modeling*, volume 1, pages 438–441. IEEE.

Han He, Lei Wu, Xiaokun Yang, Hua Yan, Zhimin Gao, Yi Feng, and George Townsend. 2018. Dual long short-term memory networks for sub-character representation learning. In *Information Technology-New Generations: 15th International Conference on Information Technology*, pages 421–426. Springer.

Liang Huang, Yinan Peng, Huan Wang, and Zhenyu Wu. 2002. Pcfg parsing for restricted classical chinese texts. In *COLING-02: The First SIGHAN Workshop on Chinese Language Processing*.

Shuiqing Huang, Dongbo Wang, and Lin He. 2015. Exploring of word segmentation for fore-qin literature based on the domain glossary of sinological index series. *Library and Information Service*, 59(11):127.

Bin Li, Yiguo Yuan, Jingya Lu, Minxuan Feng, Chao Xu, Weiguang Qu, and Dongbo Wang. 2022. The first international ancient chinese word segmentation and pos tagging bakeoff: Overview of the evahan 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 135–140.

Chengming Li. 2018. *Research on Lexical Analysis of Ancient Books Based on Deep Learning*. Ph.D. thesis, Nanjing, China: Nanjing Normal University.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *Proceedings of the*

2004 Conference on Empirical Methods in Natural Language Processing, pages 277–284.

Jingmin Shao. 2017. *General Theory of Modern Chinese*. Shanghai Educational Publishing House.

Min Shi, Bin Li, and Xiaohe Chen. 2010. Crf based research on a unified approach to word segmentation and pos tagging for pre-qin chinese. *Journal of Chinese Information Processing*, 24(2):39–45.

Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao Liu. 2015. Radical embedding: Delving deeper to chinese radicals. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 594–598.

Xunzhu Tang, Rujie Zhu, Tiezhu Sun, and Shi Wang. 2021. Moto: Enhancing embedding with multiple joint factors for chinese text classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2882–2888. IEEE.

Hanqing Tao, Shiwei Tong, Tong Xu, Qi Liu, and Enhong Chen. 2019. Chinese embedding via stroke and glyph information: A dual-channel view. *arXiv preprint arXiv:1906.04287*.

Xiaoyu Wang and Bin Li. 2017. Automatically segmenting middle ancient chinese words with crfs. *Data Analysis and Knowledge Discovery*, 1(5):62–70.

Runhua Xu and Xiaohe Chen. 2012. A method of segmentation on zuo zhuan by using commentaries. *Journal of Chinese Information Processing*, 26(2):13r17.

Bojun Yang. 1990. *Annotations to the Spring and Autumn Annals*. Zhonghua Book Company.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint embeddings of chinese words, characters, and fine-grained subcharacter components. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 286–291.

Yun Zhang, Yongguo Liu, Jiajing Zhu, and Xindong Wu. 2021. Fsprm: A feature subsequence based probability representation model for chinese word embedding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1702–1716.