Challenges in Processing Chinese Texts Across Genres and Eras

Minghao Zheng and **Sarah Moeller**
University of Florida, FL, U.S.A.
{minghao.zheng, smoeller} @ufl.edu

Abstract

Pre-trained Chinese Natural Language Processing (NLP) tools show reduced performance when analyzing poetry compared to prose. This study investigates the discrepancies between tools trained on either Classical or Modern Chinese prose when handling Classical Chinese prose and Classical Chinese poetry. Three experiments reveal error patterns that indicate the weaker performance on Classical Chinese poems is due to challenges identifying word boundaries. Specifically, tools trained on Classical prose struggle recognizing word boundaries within Classical poetic structures and tools trained on Modern prose have difficulty with word segmentation in both Classical Chinese genres. These findings provide valuable insights into the limitations of current NLP tools for studying Classical Chinese literature.

1 Introduction

The creation of Classical Chinese treebanks for prose and poetry has enabled training NLP systems across different eras and genres. This study conducts a comparative analysis, examining the performance of NLP tools trained on either Classical and Modern Chinese and either poems or prose. Especially, we analyze how NLP systems trained on prose perform when applied to poems, finding that high performance on Classical Chinese prose (Yasuoka, 2019) is reduced on poetry and providing a preliminary explanation why this reduction happens on Classical Chinese poems.

Classical Chinese, the language of ancient Chinese literature, differs significantly from Modern Chinese in style, vocabulary, and grammar. Classical Chinese texts are characterized by the absence of spaces and punctuation, appearing as continuous character strings (Yasuoka, 2019), presenting a challenge for word segmentation. For example, a Classical Chinese sentence with 11 characters translates to Modern Chinese as 29 characters. Classical Chinese prose is usually expressed in longer, variable-length sentences, whereas classical Chinese poetry is usually tightly constrained to five- or seven-character lines. Prose also tends to employ more elaborate syntactic structures, in contrast to poetry’ s concise, rhythmically disciplined forms that often create a more striking aesthetic (Li, 2020).

The application of NLP to Classical Chinese poetry and prose is divided into two categories. First is the archiving and generation of classical poetry and prose. The second is the theme and emotion classification of classical poetry (Liu, 2024). Since the first step in emotion analysis is text pre-processing, including word segmentation (Liu, 2024), the present study can help improve emotion analysis.

Our findings confirm that the reduced performance in poetry is due to difficulties in identifying word boundaries. These insights can be used to enhance the use of existing NLP tools in the study of Classical Chinese literature by highlighting opportunities for adapting NLP technologies from other Chinese eras and genres.

2 Data and Tools

This study analyzed parser performance on two syntactic treebanks of Classical Chinese: one made up of prose and one for poems. We compared the performance of Stanza’ s Modern Chinese (traditional) pipeline (Qi et al., 2020) on both datasets. Stanza is an open-source Python NLP toolkit supporting 66 human languages. We used Stanza instead of the Kyoto processor for both datasets for consistency. Stanza’ s Modern Chinese (traditional) pipeline is able to align the traditional characters used in both Modern and Classical Chinese, avoiding the need to convert characters. This pipeline was originally trained on Modern Chinese prose.

Dataset 1: Classical Chinese Poems: Classical Chinese poetry is connected with particular historical periods, such as the poetry of the Tang dynasty. The first dataset used in this study is sourced from the CityU Treebank of Classical Chinese Poems (Lee and Kong, 2012). Dataset 1 contains 60 poems written by the esteemed poet Du Fu (712-770AD), totaling 300 sentences. Each poem follows a five- or seven-character fixed-length format. Table 1 presents the size and characteristics of Dataset 1. Furthermore, Table 5 (see Appendix A) presents the proportions of various POS tag categories. Because CityU Treebank is not open-source, access to its tokenized words, sentences, POS tags, and dependency relations is limited and Dataset 1 contains the manually extracted content from the portal. Due to the non-open-source nature of the CityU treebank, annotated data was not available. No NLP tool has been specifically trained on the CityU Treebank. Also, it does not provide its own NLP tool so we used Stanza's Modern Chinese (traditional) pipeline (Qi et al., 2020).

Description	**Num**	**Pct**
Total Words	3069	100%
Total Characters	3383	-
Single-char Words	2833	92.31%
Two-char Words	227	7.40%
Three-char+ Words	9	0.29%

Table 1: Number and Distribution of Dataset 1

Dataset 2: Classical Chinese Prose: This dataset comprises the first 500 sentences from the Kyoto Treebank (Yasuoka, 2019). The Kyoto Treebank features complete texts of the Four Books, which are written in prose. Sentences generally longer than seven characters, in contrast to the poetry in Dataset 1. The Kyoto Python NLP tool (Yasuoka, 2019) which is trained and tested on the Kyoto treebank of Classical Chinese prose for tokenization (99.5% accuracy) and POS tagging (90.8% accuracy) remains the first and only Python-based NLP tool for Classical Chinese. Since the Kyoto tool was trained on the Kyoto treebank that makes up Dataset 2, we used Stanza's Modern Chinese (traditional) pipeline in our experiments with Dataset 2 rather than the the Kyoto parser.

3 Experiment Setup

To better understand the performance of a tool trained on prose when asked to parser Classical poems, we conducted three experiments and then analyzed error patterns in two prose parsers applied to Classical Chinese poems. We evaluate two NLP tools, one for Classical Chinese and one trained on Modern Chinese on two datasets: Classical Chinese poems and Classical Chinese prose. We analyze the tools' ability to handle word segmentation and POS tagging in the two Classical Chinese genres. Dependency parsing was excluded due to the complexity of manually extracting relations from the CityU Treebank of Classical Chinese Poems. Our genre-specific analysis compares the performance of both tools on Classical poetry and assesses the Modern Chinese tool on Classical prose. We compute accuracy, recall, precision, and F1-score for segmenting one- and two-character words and then identify frequently misclassified POS categories and analyze specific misclassification pairs.

Experiment 1 examines how the Classical Chinese tool handles poems in Classical Chinese (Dataset 1).

Experiment 2 evaluates how a Modern Chinese tool handles Classical Chinese prose (Dataset 2), serving as a comparative benchmark for Experiment 3.

Experiment 3 evaluates the effectiveness of a Modern Chinese tool on Dataset 1.

4 Results

In this section, we present the experiment results and the main findings. Table 2 and Table 3 summarize the results.

4.1 Word Segmentation

4.1.1 Overall Segmentation Accuracy

Experiment 1 Given the Classical Chinese tool's 99.5% accuracy in tokenizing Classical Chinese prose, we anticipated its performance in Classical poems would be equally high. The word segmentation accuracy of 84.37% is high but not as high as expected from a tool that was trained on texts from the same era.

Experiment 2 achieved a word segmentation accuracy of 74.30%.

Experiment 3 achieved a word segmentation accuracy of 56.64%, significantly lower than the 74.30% in Experiment 2.

	One-char. Words				Two- char. Words			
	Proportion	Recall	Precision	F1-score	Proportion	Recall	Precision	F1-score
Exp 1	98.81%	0.97	0.85	0.90	1.19%	0.09	0.51	0.15
Exp 3	61.02 %	0.43	0.85	0.57	36.61%	0.45	0.12	0.19

Table 2: Recall, precision, and F1 score for word segmentation shows the relative performance of the tools on Classical Chinese poems. Proportions of single and two-character words relative to all segmented words are also shown.

	Overall		Correctly Segmented
	Segmentation	POS	POS
Exp 1	84.37%	55.10%	65.30%
Exp 2	74.30%	36.10%	48.59%
Exp 3	56.64%	25.44%	44.92%

Table 3: Overall accuracy on word segmentation and POS tagging shows the relative performance of the tools on Classical Chinese poems. Accuracy on POS tagging on the correctly segmented words only (last column) shows the impact of word segmentation for downstream processing.

4.1.2 Number and Distribution of segmented words in various word lengths

We take a closer look at segmented words from the prose processors.

Experiment 1 The Classical prose processor segmented 6.74% more words in the poems than expected. As shown in Table 1, Dataset 1 contains 3,069 words, but the Kyoto processor generated 3,276. Second, while it can detect single-character words, it struggles with multi-character words, identifying fewer two-character words than exist and failing to recognize any three-character words. For example, the proper noun 小有天 'The name of the cave passed down by Taoists' in the sentence 萬古仇池穴，潛通小有天 'The Qi-uchi cave, which has been passed down through the ages, is secretly connected to Xiao Youtian' was incorrectly segmented into three words: 小 'small', 有 'have', and 天 'heaven'.

Experiment 3 Table 2 and Table 4 show that the Modern Chinese processor had difficulty segmenting each character as a separate word in Classical Chinese poems compared to its performance with Classical Chinese prose. Specifically, only approximately 61.02% of the words segmented by the tool were single-character words, whereas around 92.31% of the original poems were single-character.

4.1.3 Recall, Precision, and F1-score for segmenting one- and two-character words

Experiment 1 As shown in the confusion matrix in Table 6 (see Appendix B), only 20 two-character words in the poems were correctly segmented by the Kyoto processor. Given this very low number of true positives, it is not surprising that we found a low recall of 0.09% and an F1-score of 0.15% for segmenting two-character words in Table 2. In contrast, the F1-score for segmenting single-character words is high.

Experiment 3 In the poems, only 1,212 out of 2,833 one-character words (42.78%) were accurately segmented by the processor. As shown in Table 2, the Modern Chinese processor performed moderately when segmenting one-character words (F1-score 0.57), but when segmenting multi-character words, a notably low precision (0.12) and F1-score (0.19) were observed.

4.2 Overall Segmentation Results

These findings indicate that both prose processors' weaker performance in analyzing poems stems from difficulties identifying 'words' within poetic structures, which impacts tokenization accuracy between prose and poetry.

In Experiment 3, unlike Experiment 1, the Classical Chinese prose processor effectively identified single-character words but struggled with multi-character words. In contrast, the Modern Chinese processor had difficulty segmenting characters as individual words and produced a greater number of multi-character words in Classical Chinese poems. Alongside overall segmentation accuracy of 84.37% and 56.64% for both prose processors, additional recall, precision, and F1-score metrics offer a more nuanced view of both tools' diminished performance on poems, particularly in segmenting multi-character words.

4.3 POS tagging

To better understand how segmentation performance affects POS tagging accuracy, we investigate both prose tools' POS tagging accuracy across all words and among correctly segmented words. The Kyoto processor achieved 55.10% accuracy when tagging all words, which increased to 65.30% after controlling for segmentation. In contrast, the Modern Chinese processor started with only 25.44% accuracy across all words, rising to 44.92% after segmentation control. This figure is very similar to the 48.59% POS accuracy achieved by the Modern Chinese processor on Classical Chinese prose (Experiment 2) after segmentation control. This comparison suggests that genre differences within the same era of Chinese do not significantly impact POS tagging performance when segmentation is taken into account.

In Experiment 1, verbs were the most frequently misclassified, followed by nouns and proper nouns. Among the incorrectly tagged POS category pairs, the pair ('ADJ', 'VERB') was the most frequent, constituting 24.71% of the total misclassified pairs, indicating that adjectives are often misclassified as verbs. This was followed by the pair ('NOUN', 'VERB'), where nouns are misclassified as verbs in 14.39% of the cases. Lastly, the pair ('ADV', 'VERB') occurred in 11.16% of the cases. These common misclassified pairs suggest that the tool frequently labels other categories as verbs.

In Experiment 3, particles were the most frequently misclassified，followed by proper nouns, nouns, and verbs. In contrast to the findings in Experiment 1, the Modern Chinese processor performed better when tagging verbs but worse when tagging particles. In both cases, nouns and proper nouns were tricky for both tools. Furthermore, among the incorrectly tagged POS category pairs, the pair ('NOUN, 'PART’) was the most frequent, indicating that actual nouns are most often misclassified as particles, constituting 22.87% of the total misclassified pairs. This was followed by the pair ('NOUN’ ,'PROPN’), where nouns are misclassified as proper nouns in 18.60% of the cases. Lastly, the pair ('NOUN’ ,'VERB') accounts for 5.10% of the total incorrect cases. These common misclassified pairs suggest that the Modern Chinese processor frequently mislabels nouns.

	Num	**Proportion**
Total Words	2327	-
Three-char	40	1.72%
Four-char	6	0.26%
Five-char	8	0.35%
Seven-char	1	0.04%

Table 4: Proportions of words over two characters relative to all segmented words in Experiment 3.

5 Conclusion

This study explores differences in NLP parsers trained on Classical or Modern Chinese prose in handling prose and poetry from the different eras and analyzes the error patterns to better understand the differences in performance. We aim to contribute to developing a robust NLP tool that accurately distinguishes between these eras and genres. The findings suggest that reduced performance in Classical poetry analysis is due to difficulties in identifying word boundaries by a tool trained on a different genre or era of the language. Tools trained on Classical prose struggle to segment words in poetic structures, while Modern parsers struggle with word segmentation in both Classical Chinese genres.

We show how this difficulty affects downstream POS tagging in Classical Chinese poems. In Classical Chinese poems, verbs are commonly misclassified by the Classical Chinese processor, whereas the Modern Chinese parser commonly misclassifies particles and nouns. Future research should look at those common misclassified categories more closely to evaluate whether there are more detailed patterns that may help improve performance on Classical Chinese poetry.

Future work should apply Large Language Models (LLMs) to full-scale datasets and compare their outputs against standard parsers. For example, a Classical Chinese-specific LLM, TongGu (Cao et al., 2024), was recently developed. We prompted the GPT OSS 20B language model to perform word segmentation and POS tagging on two Classical Chinese poems (86 characters in total, 10 sentences). The model’ s performance fell short of both the Kyoto Processor and Stanza with segmentation accuracy at approximately 42% and POS tagging accuracy achieving only about 1%.

Our analysis informs the challenges of adapting NLP technologies to various eras and genres in the same language, highlighting the limitations of current tools in studying Classical Chinese literature, especially poems. These insights could be used to enhance text preprocessing, thereby improving emotion classification and other NLP tasks for Classical Chinese poetry. Moreover, these insights can support many languages that lack sufficient early texts for training parsers, and so using a modern language version is a practical bootstrap solution.

Limitations

The study is limited by the relatively small size of the poem dataset, which consists of only 300 sentences. A larger corpus of poems could yield more accurate comparisons of tool performance across genres. The current size was chosen to facilitate manual extraction of content and POS tags from a designated website, as a publicly accessible poem treebank was unavailable. Additionally, while this study focused on tokenization and POS tagging, incorporating error analysis for other NLP tasks, such as dependency parsing, lemmatization, and sentiment analysis, could provide a more comprehensive evaluation of tool performance across different genres and eras.

References

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024. Tonggu: Mastering classical chinese understanding with knowledge-grounded large language models. *arXiv preprint arXiv:2407.03937*.

J. S. Lee and Y. H. Kong. 2012. A dependency treebank of classical chinese poems. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 191–199.

Dingguang 李定广 Li. 2020. *中国诗词名篇名句赏析: 上*. Sino-Culture Press, Beijing.

Jinghan Liu. 2024. Research on the application of natural language processing in the analysis of ancient poems and texts. In *AIP Conference Proceedings*, volume 3194. AIP Publishing.

P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 101–108.

Masaki Yasuoka. 2019. Universal dependencies treebank of the four books in classical chinese. In *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pages 20–28. Digital Archives and Digital Humanities.

A POS tag proportions for Dataset 1

POS Tag	Proportion
NOUN	45.50%
VERB	20.95%
ADJ	14.04%
ADV	7.92%
NUM	6.22%
PRON	2.39%
ADP	2.03%
DET	0.69%
PART	0.13%
CONJ	0.10%
SCONJ	0.03%

Table 5: Proportions of POS Tags in Dataset 1

B Confusion Matrix of Experiment 1: Two-character-word Segmentation

	Actual 2-char words	Non-2-char words
Segmented as 2-char words	20	19
Not segmented as 2-char words	207	3030

Table 6: Confusion Matrix for Two-character-word Segmentation of Classical Chinese Processor in Classical Chinese Poems