Joint Models for Sentence Segmentation and Named Entity Recognition in Literary Sinitic Text

DongNyeong Heo[1], Yunhee Kang[2], Chul Heo[3], Heeyoul Choi[1] and Kyounghun Jung[4,*]

[1]*Handong Global University, Korea*
[2]*Baekseok University, Korea*
[3]*Pusan University, South Korea*
[4]*Wonkwang University, Korea*
E-mail: sjglsks@gmail.com; yhkang@bu.ac.kr; heochul@gmail.com; hchoi@handong.edu; juilam@hanmail.net
[*]*Corresponding Author*

Received 30 September 2025; Accepted 03 December 2025

Abstract

It is challenging to understand Literary Sinitic text from the Joseon dynasty, since there is a lack of explicit word separators, which creates significant semantic ambiguity. To address this, both sentence segmentation and named entity recognition (NER) are essential. We propose a Transformer-based analyzer that performs these two tasks simultaneously. Trained on a labeled corpus from the Seungjeongwon Ilgi, our model effectively segments sentences and identifies named entities, thereby significantly improving the understanding of sentence structure and overall context.

Keywords: Literary sinitic, sentence segmentation, NER, transformer.

Journal of Web Engineering, Vol. 25_1, 19–32.
doi: 10.13052/jwe1540-9589.2512
© 2026 River Publishers

1 Introduction

Literary Sinitic called *Hanmun* 漢文 in Korean is a common written language used in Korea, China, Japan, Vietnam, and other countries that are part of the Han-character cultural sphere during the medieval period. Literary Sinitic is used as the standard written language in these countries at the time, and it was used to record all documents. In Korea, it was used from around the 3rd or 4th century AD until the early 20th century, not only for historical records, but also in literature, philosophy, and all kind of documents. It is often referred to as Old Chinese, and it has various linguistic characteristics different from modern Chinese. In order to analyze Literary Sinitic sentence, which has a different linguistic structure and grammatical characteristics from modern Chinese, different types of language analysis tools are required.

Currently, main morphological analyzers developed in China include Wu Yu Dian developed by the Digital Humanities Research Institute of Peking University, Gu Lian Shuzi's Automatic Punctuation by Zhonghua Shuguo, Gushiwen Duanju by Beijing Normal University, and Zhino Biaodian by Rushi Guji. These analyzers are known to use modern language analysis algorithms such as BERT, Bi-LSTM, and CRF [3, 6]. However, even though the same Literary Sinitic sentences are used across different countries, differences in linguistic habits create certain limitations for applying morphological analyzers developed in China to the analysis of specific Korean classical texts. Particularly, the Korean documents include numerous proper nouns, like personal names, place names, and official titles. To analyze such texts, a suitable analyzer is required. Furthermore, the writing and printing methods for Literary Sinitic sentences differ greatly from Western styles, as the basic writing method of Literary Sinitic sentences involves a string of characters without punctuation, delimiters or other marks. Therefore, to understand Literary Sinitic sentences effectively, the reader must possess the ability to segment text and have knowledge of its specific grammar. To properly interpret a Chinese sentence, sentence segmentation and named entity recognition (NER) are essential for reducing ambiguity in meaning.

NER plays an important role in the knoweldge extraction and utilization of Literary Sinitic [11]. Compared with NER in modern Chinese, NER in Literary Sinitic texts from the *Joseon dynasty* is more challenging due to semantic ambiguity. The hurdles of Literary Sinitic NER originate from linguistic characteristics like the lack of word separators in Chinese, since Literary Sinitic has its own unique characteristics. It may require significantly

diverse language processing due to differences in grammar and vocabulary compared to modern Chinese. To address this issue, sentence segmentation is necessary, as most Literary Sinitic texts were originally written without punctuation.

In this paper, we build a Transformer based analyzer which performs sentence segmentation and NER in Literary Sinitic texts from the Joseon Dynasty. Our model consists of a transformer encoder, delimeter classifier for sentence segmentation, and NER classifier. That is, the model for sentence segmentation includes the transformer encoder and the delimeter classifier, and the model for NER includes the transformer encoder and the NER classifier, where the transformer encoder is shared by the two tasks. We train the model with *Seungjeongwon Ilgi* (承政院日記) as labeled input. In the experiment, our model effectively segments sentences and identifies named entities, thereby significantly improving the understanding of sentence structure and overall context.

2 Background

2.1 Seungjeongwon Ilgi

The *Daily Records of the Royal Secretariat* 承政院日記 (Seungjeongwon Ilgi) is a diary that recorded royal commands, administrative affairs, ceremonial matters, and other activities handled by the Royal Secretariat (承政院) during the Joseon dynasty. Records have been preserved from 1623, the first year of King *Injo*, to July 1894, the 31st year of King *Gojong*.

The data digitalizaiton began in 2001 and was out in 2015. It included the application of modern punctuation, markup Entity such as names, places, and titles, and the addition of article titles for volumes 1–34. For volumes 35–141, only modern punctuation and markup for keywords terms were added. The *Seungjeongwon Ilgi* is one of the most influential texts in the Joseon dynasty.

2.2 Morphological analysis in Literary Sinitic text

Nanjing Normal University morphological analyzer provides MA for each word in a given sentence, which is based on Bidirectional LSTM-CRF [6]. This network can efficiently use past input features via the LSTM layer and sentence level tag information via the CRF layer. It is used for sequence tagging including part-of-speech (POS) tagging, chunking, and NER. However, Literary Sinitic employs complex and distinctive grammatical structures,

including syntax, word order, and rhetoric, among other aspects, which differ significantly from modern Chinese. Machine-based punctuation and NER analyzers developed in China exhibit performance variations due to the inclusion of Korean-specific reign titles, place names, personal names, and official titles.

We construct a corpus which is used to annotate a text as an input of a model building. To build the corpus, we use *Seungjeongwon Ilgi* as the sample Chinese texts. Consider the constraints of Processing Literary Sinitic: the writing and printing method for Literary Sinitic sentences differs greatly from Western styles. The basic writing method of Literary Sinitic sentences involves a string of characters without punctuation, delimiters or other marks. Thus, given a sentence, sentence segmentation and NER recognition processes are necessary.

3 Proposed Methods

In order to analyze Literary Sinitic sentences, which have different linguistic structures and grammatical characteristics from modern Chinese, different types of language analysis tools are required. In this section, we describe the sentence segmentation and NER tasks in Literary Sinitic sentences, then propose Transformer-based models for the two tasks.

3.1 Sentence Segmentation

Sentence segmentation (SS) is a natural language processing task that predicts the positions of delimiters to separate a long string into individual sentences. The SS task predicts the positions of delimiters, ',' or '。', given the Hanmun string without delimiters. Previous SS algorithms include rule-based and learning-based methods, and learning based methods include deep learning approaches like LSTM and BERT (or Transformer) [4, 14].

In this work, we conduct the SS task on the Seungjengwon Ilgi dataset. Specifically, we trained a Transformer-based text encoder and classifier that predicts the position of delimiters. By doing this, we expect this work can contribute to the automated SS process which has been done by human experts. Beyond this, we expect the automatically segmented string of sentences can enhance subsequent natural language processing tasks, such as named entity recognition.

3.2 Named Entity Recognition

Named entity recognition (NER) is an NLP component that identifies defined categories of objects in a body of text. These features can include part-of-speech tagging (POS tagging), word embeddings and contextual information, among others [2,5,12]. Through NER, we can accurately identify and extract specific categories of information, such as personal names, location names, and organization names, which plays a crucial role in the extraction and utilization of knowledge of Literary Sinitic books. Chinese NER is a very challenging task due to the lack of word separators in Chinese. Compared with NER in modern Chinese, NER in Literary Sinitic is more difficult. One of the reasons is the concise nature of Literary Sinitic texts, often characterized by single character and short sentences, which frequently results in semantic ambiguity.

Many researches on NER in Literary Sinitic have progressively shifted towards leveraging deep neural network models. *SikuBERT* is a pre-trained language model proposed to enhance the representation of Literary Sinitic texts using ancient text corpora [10], to integrate both named entity recognition and sentence segmentation tasks [4].

3.3 Transformer-based Seungjengwon Ilgi Analyzer

In this section, we describe the details of our proposed Transformer-based analyzer model that aims to conduct the SS and NER tasks on the Seungjengwon Ilgi data, which contains delimiters and named entities annotated by human experts. We consider this annotated outcome as ground-truth prediction result. To obtain a training dataset for SS and NER tasks, we performed the preprocessing procedure in the reverse direction of the (human-based) annotation process. Specifically, we eliminated named entities (e.g., 'person_name') and delimiters (e.g., ',' and '。') from the annotated result to obtain the original text. Then, we labeled the positions where delimiters were located and characters that were annotated as named entity before the elimination. Finally, we provided the original text as input to the model, and the labeled positions and characters as targets, as presented in Figure 1.

Our proposed model consists of one main encoder and two subsequent classifiers for SS and NER tasks, respectively. The main encoder follows the architecture of Transformer [9], which is the fundamental building block of modern large language models, including BERT [3] and GPT [1] (with causal masking). The subsequent classifiers are designed as a multi-layer perceptron

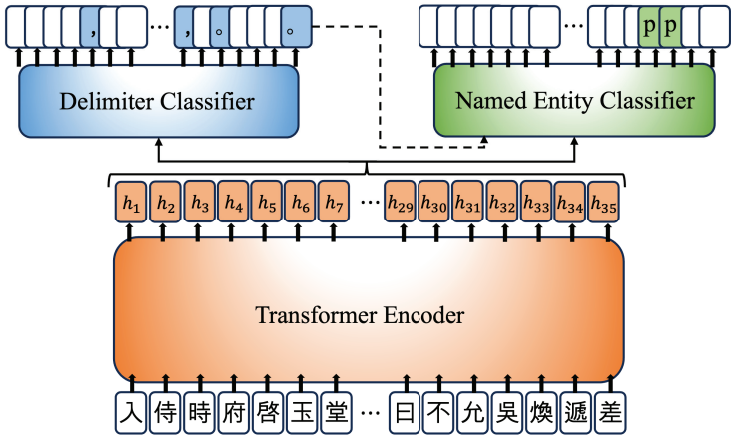


Figure 1 The proposed Transformer-based analyzer: The input letters and targets are pre-processed results of the example in Section 2.2. White target tokens mean 'None' class, while other colorful target tokens mean specific classes other than 'None'. For example, 'p' target token on top of named entity classifier means 'person_name' class.

(MLP) with the same hidden dimension of the main encoder and the number of classes for the output. Our main encoder is basically similar to the backbone network of SikuBERT, so it looks natural to fine-tune SikuBERT for our main tasks. However, a vocabulary mismatch issue poses challenges for directly fine-tuning SikuBERT. We found that SikuBERT was pre-trained using subword tokenization techniques [7,8], whereas our proposed approach needs character-level tokenization for the model's fine-grained SS and NER predictability at the character-level.

Unlike to the conventional work, such as SikuBERT, we share the encoder model for both SS and NER tasks and train whole model with a multi-task learning fashion. Therefore, our model is efficient in the number of parameters and can be regularized by both tasks' objectives. In addition, as we conjectured, the result of SS could be helpful for the NER task (or vice versa). Therefore, we additionally propose to input SS result to the NER classifier. We provide ground-truth and predicted sequence of delimiters for training and testing, respectively. Note that this additional proposal could be applied in the opposite direction, that is, inputting named entities to the delimiter classifier. There is a class imbalance problem as shown in Table 1, and to mitigate the problem during training, we differently set weights for each class in loss computation. Specifically, we set 0.2 for 'None' class and 1.0 for other classes following the ratio.

Table 1 Resulting statistics of preprocessed Seungjengwon Ilgi dataset. 'Ratio of Delimiters' are percentages of 3 classes [None / ',' / '。']. 'Ratio of Named Entities' are percentages of 6 classes [None/person_name/location_name/additional_explanation/signature/working_status]

# of Tokens/Strings	# of Vocab.	Ratios of Delimiters (%)	Ratios of Named Entities (%)
272M / 3M	13,789	79.2/15.0/5.8	78.9/19.2/1.3/0.3/0.2/0.1

4 Results

4.1 Datasets

In this section, we describe details of preprocessing for the Seungjengwon Ilgi dataset. Mainly, we processed to eliminate the delimiters and named entities from the completely annotated string. During this elimination process, we labeled the position and characters to indicate the ground-truth prediction results as we explained in Section 3.3. After this process, we tokenized the strings in character-level. In result, the total number of preprocessed tokens/strings, the number of unique vocabularies, types of classes, and ratios of classes are entirely summarized in Table 1. We split the preprocessed strings into training/validation/testing sets with ratios of 98%/1%/1%, which are 2.98M/10K/10K strings, respectively.

4.2 Model Training

As we explained in Section 3.3, we used Transformer-based encoder, especially with PreLN architecture [13]. The total number of parameters is around 27M. We had our models trained on the preprocessed Seungjengwon Ilgi training dataset. At every 1K iteration, we evaluated model on the validation dataset and saved the checkpoint if it achieves its best performance. We early-stopped training if the model did not renew its previous best performance 30 times. We used GTX1080Ti GPU for whole experiments, and they took 134 hours on average.

4.3 Evaluations

Table 2 demonstrates the experiment results of our proposed models. We ran three experiments with the original Seungjengwon Ilgi analyzer model, as described in Section 3.3. We also tried two variants: feeding delimiters to the named entity classifier, 'S2N', and feeding entities to the delimiter classifier, 'N2S'. We evaluated all our models on the testset with respect to precision (Prec.), recall (Rec.), F1 score, and accuracy (Acc.).

Table 2 Results of Seungjengwon Ilgi analyzer models. 'Original' model means our proposed model introduced in Section 3.3. '+S2N' is the model inputting delimiters (output of SS) to named entity classifier, and '+N2S' with the opposite direction

Model	Sentence Segmentation				Named Entity Recognition			
	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
Original	0.9287	0.9372	0.9322	0.9729	0.9670	0.9717	0.9689	0.9947
Original+S2N	**0.9309**	0.9400	0.9347	0.9749	**0.9735**	**0.9767**	**0.9747**	**0.9958**
Original+N2S	0.9304	**0.9421**	**0.9355**	**0.9751**	0.9702	0.9743	0.9719	0.9953

Table 3 Two prediction examples of our proposed model

Example 1 - SS (F1: 1.0, Acc.: 1.0) / NER (F1: 1.0, Acc.: 1.0)	
Input	都承旨鄭廣敬左承旨崔右承旨李行遠左副承旨韓興一右副承旨韓亨吉 同副承旨李景曾注書李道長一員未差假注書柳事變假注書金廈樑金振
SS Pred.	都承旨鄭廣敬。左承旨崔。右承旨李行遠。左副承旨韓興一。右副承旨韓亨吉。 同副承旨李景曾。注書李道長一員未差。假注書柳。事變假注書金廈樑金振。
SS GT	都承旨鄭廣敬。左承旨崔。右承旨李行遠。左副承旨韓興一。右副承旨韓亨吉。 同副承旨李景曾。注書李道長一員未差。假注書柳。事變假注書金廈樑金振。
NER Pred.	都承旨 鄭廣敬 。左承旨 崔 。右承旨 李行遠 。左副 承旨 韓興一 。右副承旨 韓亨吉 。同副承旨 李景曾 。 注書 李道長 一員未差。假注書 柳 。事變假注書 金廈樑金振 。
NER GT	都承旨 鄭廣敬 。左承旨 崔 。右承旨 李行遠 。左副 承旨 韓興一 。右副承旨 韓亨吉 。同副承旨 李景曾 。 注書 李道長 一員未差。假注書 柳 。事變假注書 金廈樑金振 。
Example 2 - SS (F1: 0.8151, Acc.: 0.8095) / NER (F1: 1.0, Acc.: 1.0)	
Input	開元寺崇恩殿親祭三嚴卯初一刻日出時還宮
SS Pred.	開元寺, 崇恩殿親祭, 三嚴, 卯初一刻, 日出時, 還宮。
SS GT	開元寺崇恩殿親祭, 三嚴卯初一刻。日出時還宮。
NER Pred.	<l>開元寺, 崇恩殿</l>親祭, 三嚴, 卯初一刻, 日出時, 還宮。
NER GT	<l>開元寺崇恩殿</l>親祭, 三嚴卯初一刻。日出時還宮。

We found that our original model achieves quite high performances on both SS and NER tasks, such as 0.9322/0.9729 F1/accuracy performances on SS and 0.9689/0.9947 on NER, respectively. Interestingly, when we additionally input one classifier's outcome to the other classifier, we found that it helped to improve performance further. For example, 'S2N' model achieved the best NER performances and 'N2S' model achieved the best SS performances. We believe that natural language analysis methods for Literary Sinitic text can enhance by each other as we expected.

Table 3 presents two examples of model predictions. We demonstrated the predictions of 'Original+S2N' model. The second example's SS prediction contains multiple incorrect delimiters. We found that our model tends to

insert more delimiters than the ground-truth. Furthermore, due to the class imbalance problem, we found that our model has tendency that outputs ',' more than '。'. However, even though our model contains errors somehow, it successfully conducts the SS and NER tasks as demonstrated in the first example.

To evaluate the performance of sentence segmentation in the perspective of morphological analysis, we measured the morphological analysis processing failure rate with and without sentence segmentation. With sentence segmentation, the morphological analysis processing failure rate is 4.61% of the whole dataset, while the failure rate is 94.95% without sentence segmentation.

5 Conclusions

Our work primarily focuses on analyzing the Literary Sinitic texts. It is a minority field due to the lack of annotated data and relatively few application scenarios. In this paper, we proposed a transformer model for sentence segmentation and named entity recognition in the Literary Sinitic dataset, *Seungjengwon Ilgi*. In the performance evaluation, the proposed model achieved high accuracy in the sentence segmentation task, which reduced the ambiguity of Literary Sinitic sentence. Also, we showed that the proposed model achieved high accuracy in the NER task, which can be used to identify categories of information that can improve the understanding of Literary Sinitic text.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2021S1A5C2A02089896).

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

[2] Chang, Y., Kong, L., Jia, K., and Meng, Q. (2021). Chinese named entity recognition method based on bert. In *ICDSCA 2021*, pages 294–299.

[3] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, Minneapolis, Minnesota.

[4] Ge, S. (2022). Integration of named entity recognition and sentence segmentation on Ancient Chinese based on siku-BERT. In Hämäläinen, M., Alnajjar, K., Partanen, N., and Rueter, J., editors, *International Workshop on Natural Language Processing for Digital Humanities*, pages 167–173, Taipei, Taiwan.

[5] Guo, W., Lu, J., and Han, F. (2022). Named entity recognition for chinese electronic medical records based on multitask and transfer learning. *IEEE Access*, 10:77375–77382.

[6] Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

[7] Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

[8] Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

[9] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

[10] Wang, D., Liu, C., Zhu, Z., Jiang, Feng, Hu, H., Shen, S., and Li, B. (2021). Construction and application of pre-training model of "Siku Quanshu" oriented to digital humanities.

[11] Wang, S., Li, X., Meng, Y., Zhang, T., Ouyang, R., Li, J., and Wang, G. (2022). $k$nn-ner: Named entity recognition with nearest neighbor search. *arXiv preprint arXiv:2203.17103*.

[12] Wu, H., Ji, J., Tian, H., Chen, Y., Ge, W., Zhang, H., Yu, F., Zou, J., Nakamura, M., and Liao, J. (2021). Chinese-named entity recognition from adverse drug event records: Radical embedding-combined dynamic embedding–based bert in a bidirectional long short-term conditional random field (Bi-LSTM-CRF) model. *JMIR Med Inform*, 9(12):e26407.

[13] Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. (2020). On layer normalization in the transformer architecture. In *ICML*, pages 10524–10533. PMLR.

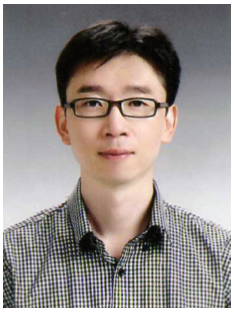
[14] Yu, J. S., Wei, Y., and Zhang, Y. W. (2019). Automatic ancient chinese texts segmentation based on BERT. *Journal of Chinese Information Processing*, 33:57–63.

Biographies

DongNyeong Heo received his B.S. and M.S. from Handong Global University, Korea, in 2019 and 2021, respectively. He is expected to receive his Ph.D. from Handong Global University, Korea, in February 2026. His research interests cover machine learning-based natural language processing, and generative models.

Yunhee Kang earned a BS in Computer Engineering (1989) and an MS in Computer Engineering (1993), both from Dongguk University in Seoul, Korea. He received a PhD in Computer Science (2002) from Korea University in Seoul, Korea. He has been working as a Full Professor at Baekseok University in Cheonan, Korea since March 2002. His research interests include Trusted Computing, Cloud computing, Applied AI, Blockchain and Web3.

Chul Heo earned a BS(1996) and MS(2000) in Hanmun (Literary Sinitic) Education, both from SungKyunKwan University in Seoul, Korea. He received a PhD in Chinese Linguistic and Character(2010) from Beijing Normal University, China. He currently serves as a Researcher at the Jeom Pil Jae Research Institute at Pusan National University in South Korea, while also holding appointments as Distinguished Professor at Sichuan Tourism University China and Yangzhou University, China, and as a Distinguished Research Fellow at the Nishan World Center for Confucian Studies and Mengzi Research institute in China. His research interests focus on Global Han-characters and Hanmun(Literary Sinitic) Education, Digital Humanities for East Asian Ancient Texts, and Cultural Exchange within the East Asian Sinosphere.



Heeyoul Choi received his B.S. and M.S. from Pohang University of Science and Technology, Korea, in 2002 and 2005, respectively, and the Ph.D. from Texas A&M University, Texas, in 2010. He is a professor at Handong Global University. His research interests cover machine learning (deep learning), and natural language processing.

Kyounghun Jung earned a bachelor's degree in Chinese literature (1997) and a master's degree in Korean literature (1999) from Chungnam National University. He received a doctorate in Korean literature (2005) from Sungkyunkwan University. He has been an assistant professor at Wonkwang University in Iksan, Korea since March 2021. His research interest is in building and utilizing knowledge base data for Chinese literature records.