IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 March 2024, accepted 8 April 2024, date of publication 17 April 2024, date of current version 3 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3390181

RESEARCH ARTICLE

Exploiting Hanja-Based Resources in Processing Korean Historic Documents Written by Common Literati

HYEONSEOK MOON[1], MYUNGHOON KANG[1], JAEHYUNG SEO[1], SUGYEONG EO[1], CHANJUN PARK[2], YEONGWOOK YANG[3], AND HEUISEOK LIM[1]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea
[2]Upstage, Yongin-si, Gyeonggi-do 17006, Republic of Korea
[3]Department of Computer Science and Engineering, Gangneung-Wonju National University, Wonju 26403, Republic of Korea

Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)

This work was supported in part by ICT Creative Consilience Program through the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government (MSIT) under Grant IITP-2024-2020-0-01819, in part by the National Research Foundation of Korea (NRF) grant funded by Korean Government (MSIT) under Grant 2022R1A5A7026673, and in part by the Basic Science Research Program through NRF funded by the Ministry of Education under Grant NRF-2021R1A6A1A03045425.

ABSTRACT This research aims to explore the comprehension of historical Korean archives authored by common literati. Numerous endeavors have been made to study Korean historical documents; however, the majority of these endeavors focus solely on royal documents. By comparing the distinct linguistic characteristics between royal and commoner languages, this study challenges the applicability of the royal language-centric approach to commoner documents. In particular, we investigate the feasibility and limitations of existing resources that share the same writing system (Hanja) as historical Korean documents for processing Korean common literati documents. Through our investigation, we propose a simple yet effective methodology that enables the utilization of Hanja-based language resources in processing Korean common literati documents: the removal of special characters. We demonstrate that aligning characteristics of Hanja-based resources allows considerable performance improvements. To the best of our knowledge, our study represents the first research endeavor to concentrate on the comprehension of common literati documents.

INDEX TERMS Natural language processing, deep learning, named entity recognition, sentence segmentation, ancient language processing.

I. INTRODUCTION

Hangul, the present Korean script, was invented by King Sejong in AD 1443. Prior to its creation, Korea predominantly used the character system known as Hanja, which shares the same system with ancient Chinese [1], [2], for written communication [3]. As a result, a majority of Korea's historical records are documented in Hanja. With the evolution of the language system, speakers who are solely familiar with Hangul face difficulties in understanding historical Korean records [3], [4]. Considering the socio-political insights these kinds of historical literature can provide, specialized efforts are crucial for deciphering and understanding traditional documents [5].

Several historical Korean records were released as open-source by the *Institute for the Translation of Korean Classics (ITKC)*.[1] Among them, documents from the Korean royal court, particularly from the Joseon dynasty (such as the Annals of the Joseon Dynasty (AJD) and "Seungjeongwon Ilgi". We denote these as **KrR** in this study), stand out as structured and data-rich. Such prominence has steered a majority of Korean historical research towards exclusively focusing on these royal manuscripts [5], [6], [7], and [8].

However, within this research trend, historical documents written by the common literati (**KrC**) are being overlooked

The associate editor coordinating the review of this manuscript and approving it for publication was Daniel Augusto Ribeiro Chaves.

[1]https://db.itkc.or.kr/

(a) KrR (written by Royal Court)

(b) KrC (written by Common Literati)

FIGURE 1. Sample images for the (a) KrR and (b) KrC. Red square denotes special characters that indicate puncutation or white spaces. (a) takes several sentence spliters while (b) does not. This distinction suggests the disparity of the characteristics between KrR and KrC.

TABLE 1. Data samples for KrR and KrC. KrR comprises whitespaces or special characters for identifying semantic units (such as word), while KrC does not.

KrR
先生曰: 鹽在北, 而移司於南, 卽金不至, 亦剽奪盡矣
KrC
仍堤宅李士瞻鄭述甫亦在座聞權棨訃七日朝

and only **KrR** are being studied as representatives of the historical records. In this study, we first emphasize the need to differentiate between these two types of documents, based on their distinct data characteristics. Despite sharing the same writing system, **KrR** are characterized by structured sentence formations with scarce omissions, while **KrC** freely employ several abbreviations (such as omission of prepositions or conjunctions) [5] and [9]. Notably, as shown in Figure 1 and Table 1, **KrC** seldom adopt special characters, including white spaces and punctuation, while **KrR** attaches clear sentence markers [5], [7].

These distinction pose limitations to the utilization of existing royal-centric studies and resources for **KrC**. We point out that **KrR** has challenge to cover **KrC**. To validate the previously mentioned problems, we collected 1,860 documents from the Joseon dynasty, written by contemporary scholars (36,000 sentences approximately) from the 17C to the 19C. Specifically, we prioritize two tasks with direct relevance to the document curation: sentence splitting (SenS) and Named Entity Recognition (NER) [8], [10].

In this study, based on the characteristics of **KrC**, we analyze the applicability of existing Hanja-based resources for **KrC**, and explore the strategies that need to be applied to make existing resources more suitable for **KrC**. A simplistic yet effective solution we propose is standardizing sentence structures by removing special characters. Unlike the other languages, **KrC** employs minimal punctuation, as evidenced in Figure 2. Our experiments reveal that merely removing special characters significantly enhances the transferability of Hanja-based resources to **KrC**. By training models on punctuation-stripped **KrR**, we achieved up to a 4-point F1 score improvement in NER. Even for linguistically distinct languages like Chinese, removing punctuation led to over a 40-point F1 score improvement in sentence splitting. This methodology facilitates the effective transfer and adaptation of existing Hanja-based resources to **KrC**.

To the best of our knowledge, this research marks the first attempt to endeavor to understand Korean commoner literati documents. Through this exploration, we aim to highlight the challenges and potential areas of enhancement in existing research methodologies, fostering a more holistic understanding of historical Korean manuscripts.

II. RELATED WORK

Differences between modern and ancient languages exist for virtually all languages [11], [12], [13], [14], and the attempt to understand them and the exploration of appropriate methods is considered an essential area of research [15], [16], [17], [18]. Such studies encompass Chinese [19], [20] and several other alphabetic languages [15], [21], [22], [23].

In the case of Korean, a totally different writing system has been developed for the modern language (Hangul, which is a phonogram) [24], [25]. Ancient Korean is rather similar to the ancient Chinese that shares Hanja writing system [1], [26]. Hanja are ideograms, meaning that each character has a separate meaning, and while the meaning of each character is shared between Chinese and Korean, the usage of each character in a sentence and the way the word is combined are different [26] and [27].

Based on this shared characteristic of the writing system, several attempts have been made to understand traditional Korean writings through the ancient Chinese [2], [5], and [6]. This is because, although the word order and specific meanings may differ, the shared writing system made it effective to transfer the knowledge between them. In particular, [1] created a language model for ancient Korean, by post-training historical Korean records written by the royal court on AnchiBERT [28] and mBERT [29], and found that it is a more effective way to understand ancient Korean using AnchiBERT, an ancient Chinese language model.

However, we find out that the majority of research for the ancient Korean focused solely on the **KrR** [1], [2], [5], and [6]. Through our subsequent analyses, we discern that **KrR** cannot sufficiently cover **KrC**, and propose a strategy that elevating the utility of existing Hanja-based resources for exploiting them to **KrC**.

III. DATA ANALYSIS

A. DATA COLLECTION

We have curated the **KrC** dataset in collaboration with the *Advanced Center for Korean Studies*.[2] Specifically, our dataset draws from two primary sources: diary data (Ilok) and letter data (Ganchal). In this experiment, we use two

[2]http://www.ugyo.net/

TABLE 2. Detailed data statistics for all of the datasets.

Data Statistics		NER			SenS		
		# of Data	Average chr Length	% of Special chr	# of Data	Average chr Length	% of Special chr
KrC	Train	1,860	193.93	1.44	1,860	193.93	1.44
	Validation	252	191.78	1.75	252	191.78	1.75
	Test	231	195.69	1.02	231	195.69	1.02
KrR	Train	374,191	68.00	22.85	417,602	123.74	15.91
	Validation	12,839	44.43	24.28	14,987	124.50	15.90
	Test	12,856	45.29	24.18	14,991	124.71	15.92
ChC	Train	12,919	25.67	15.04	731	219.02	27.44
	Validation	4,308	25.23	15.05	155	215.41	28.01
	Test	4,304	25.00	14.70	148	217.80	28.29

documentary data, Gyeam Ilok[3] and Ganchal of Andong Hansan family, as the main data. Gyeam Ilok is a meticulously maintained diary from the Joseon dynasty, authored by the literati Kim-Yeong. This diary meticulously chronicles his official and hermit life, from July 1603 to March 1641. The Ganchal data we employ is sourced from the 18th century. These datasets were provided through the support of the National Research Foundation of Korea and a government-funded project.

We evaluate the suitability of several existing Hanja-based resources for the **KrC** understanding task. Firstly, we employ the **KrR** dataset, a widely used resource for Korean ancient literature research. We assess its transferability and adaptability to enhance practical understanding of **KrC**. The **KrR** consists of records documenting events in the royal court of Joseon from the 14C to the 19C and is referred to as the Annals of the Joseon Dynasty (AJD) in previous studies [1] and [2]. For NER data, we utilized previously released data [1], and for SenS, we directly preprocessed the released data.[4]

To investigate the viability of data that shares the same language system (Hanja), we also adopt the ancient Chinese datasets written by the Chinese common literati, which we denote as a **ChC**. We have utilized publicly available data that fit our purpose. The NER data[5] are written records from the Song Dynasty, a period that shares Chinese characters with Korea. Historical Thought in Song and Yuan Dynasty (宋元學案) are mainly about various Confucian academic ideas throughout the Song Dynasty, which started a pattern of writing in the "Xuean" style. With 100 chapters and approximately 2,000 eminent thinkers, this influential book is recognized as a well-known masterwork of Chinese philosophy. For SenS,[6] we use officially released data from LT4HALA workshop [30].

The statistical information of the data used in this experiment can be found in Table 2. For **KrC**, the documents were partitioned into chunks so that the character length of each data was less than 200. This resulted in a total of 1,860 documents, which were identically adopted for SenS and NER. **KrR** represents the most abundant data source, with all available data being utilized in this study. Further elaboration on the implications of this statistical information will be provided in a subsequent sections.

B. SPECIAL CHARACTER PROCESSING

Comparing with existing resources, we find that the most distinguishing feature of **KrC** is the frequency of special characters. Here, the term "special character" encompasses all symbols, including punctuation that marks sentence units and quotation marks and brackets (denoted as markers in this paper) that indicate the function of a specific phrase within a sentence. Utilizing the priorly established rule,[7] we define sentence splitting characters and simple special characters as shown in Table 3. In particular, sentence splitters serve as crucial indicators for sentence splitting, and we consider them as additional sentence distinguishing factors in addition to the already established sentence units. This is motivated by the varying definition of split point among **KrC**, **KrR**, and Chinese documents from common literati (**ChC**) datasets. We aim to establish a standardized definition of a sentence unit. By applying the same preprocessing rules to these different language data, we can objectively examine the role of language symbols and existing resources' adaptability to **KrC**.

The necessity of these segmentation rules can be seen in Table 2. **KrC** has less than 1.5% of the total number of characters as special characters, while **KrR** has 15% and Ch_Literati has 25%. This can be attributed to the number of punctuation marks present in the document, as well as differences in the tendencies of the transcriber when building the data in the first phase. Since Hanja is essentially ideograms with each character carrying its meaning [3], we argue that the role of special characters and non-character punctuation plays an important role in language understanding.

To mitigate the unintended influence and bias caused by special characters, we generate a new dataset by removing all

[3]Please refer https://encykorea.aks.ac.kr/Article/E0076329 for more details.

[4]https://sillok.history.go.kr/

[5]NER: https://github.com/MescoCoder/AncientChineseProject

[6]SenS: https://circse.github.io/LT4HALA/2022/EvaHan

[7]https://sillok.history.go.kr/intro/rulePopup.do?type=03

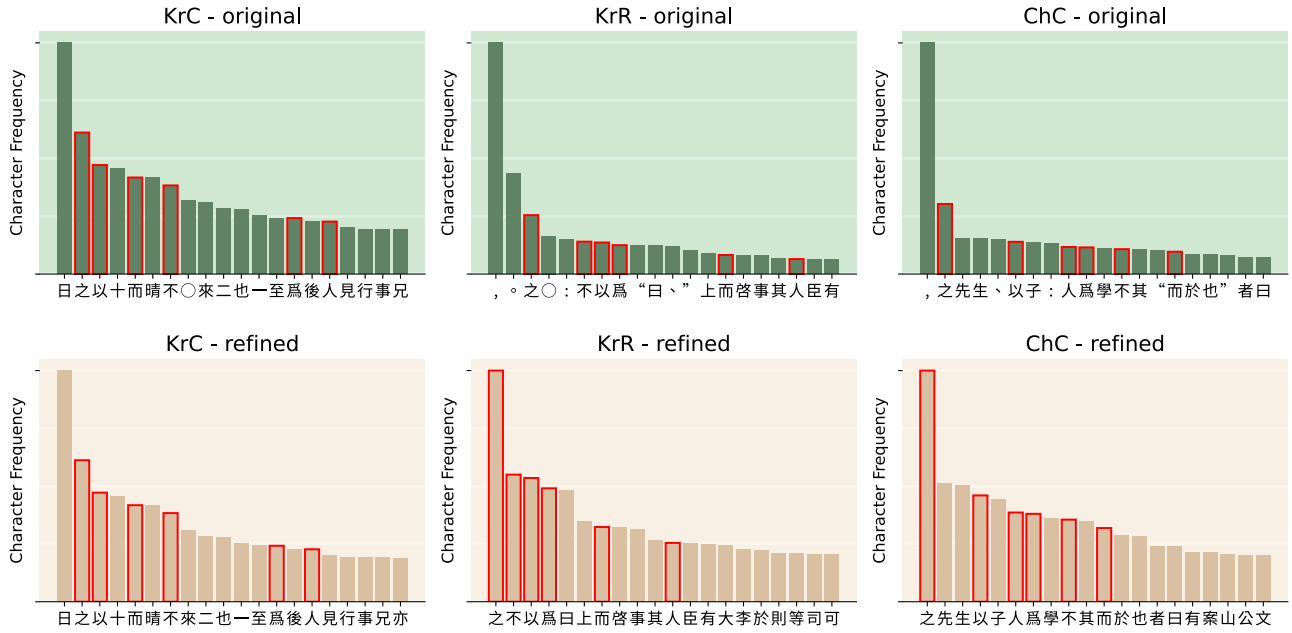

FIGURE 2. Token distributions of NER datasets. We find that `KrC` seldomly adopt special characters while others frequently utilize them. Red squared characters denote sharing characters across three datasets.

TABLE 3. Special character list defined in this paper. In processing each sentence, if we encounter spliter, we split the prior and posterior parts into the separate sentence. Marker and others are removed universally across all the datasets.

Spliter	。 ! ? . `
Marker	“” ‘’ [] 〔〕 () 《》 【】
Etc.	○ □ : ; , /

special characters. This enables more objective evaluation on the impact of special characters and facilitates an assessment of inter-dataset transfer performance with minimal symbol bias. The token distributions of the dataset with and without special characters are shown in Figure 2. We extracted the twenty-five most frequent characters in the dataset. Among the most frequent characters, those shared by all three datasets are indicated by red edges. Specifically, ‘以’, ‘而’ and ‘之’ serve as common ‘‘postposition’’ characters, and 不 functions as a universally common word expressing denial (the same as ‘‘not’’) in both Chinese and Korean.

C. WHY SPECIAL CHARACTERS MATTER?

The handling of special characters, which may seem like a simple strategy, is considered a crucial task in dealing with Korean historical documents. This is due to ambiguity in sentences written in Hanja. As logographic characters, each Hanja character holds its unique meaning, and hence, compared to phonographic languages, the interpretations can significantly fluctuate depending on sentence delimiters or word delimiters.

Particularly in **KrC**, such ambiguity is exaggerated by the absence of whitespaces and function words. For instance, ‘‘金公克邦來’’ can be interpreted as ‘‘**Dr. KimKeukBang**(金公克邦)’’ has **arrived**(來)’’ by taking word delimiters as ‘‘金公克邦□來’’. However, if we view this sentence as ‘‘金公□克邦□來’’, the meaning changes to ‘**Dr Kim**(金公) and **KeukBang**(克邦) have **arrived**(來)’, which can also be a valid interpretation.

In this sense, the presence of special characters greatly determines the difficulty in comprehending a sentence Although a simple whitespace can resolve ambiguity, their scarcity and the frequent omission of conjunctions such as ‘and’ worse the ambiguity of the sentences in **KrC**. Considering the differences in characteristics between well-structured resources(i.e. **KrR**) and unstructured ones like **KrC**, standardising such characteristics can greatly enhance the utility of existing resources.

D. DATA STATISTICS FOR NER

Table 4 presents the statistical data pertaining to NER. Considering the entity length and document length, it can be inferred that the named entity occurrence frequency is similar among the datasets we experimented. However, we noted that the document length of **KrC** is substantially larger compared to **KrR** or **ChC**. To mitigate the potential risk arising from this disparity, we adjusted the test data of **KrC** to a length similar to the training data when testing **KrC** data from model trained with **KrR** or **ChC**. By experimenting with language transfer in settings where the statistics across datasets were similar, we were able to obtain relatively objective experimental results.

IV. TASK DESCRIPTION

This paper adopts two main evaluation objectives: named entity recognition (NER) and sentence splitting (SenS). NER is a fundamental task in comprehensing sentence structure [10]. SenS is a crucial task, particularly in the documentation of ancient Korean records, where punctuation or splitting markers are rarely attached.

TABLE 4. Data statistics of the NER dataset.

KrC	Avg # of character per doc	Avg # of "Person" entity per doc	Avg # of "Location" entity per doc	Avg Length of "Person" entity	Avg Length of "Location" entity
Train	193.93	9.975	2.727	2.500	1.996
Valid	191.78	10.079	2.833	2.463	2.021
Test	195.69	9.333	2.628	2.505	1.993
KrR	Avg # of character per doc	Avg # of "Person" entity per doc	Avg # of "Location" entity per doc	Avg Length of "Person" entity	Avg Length of "Location" entity
Train	68.00	2.161	0.764	2.561	2.366
Valid	44.43	0.881	0.570	2.835	2.492
Test	45.29	0.853	0.569	2.840	2.500
ChC	Avg # of character per doc	Avg # of "Person" entity per doc	Avg # of "Location" entity per doc	Avg Length of "Person" entity	Avg Length of "Location" entity
Train	25.67	1.056	0.400	1.991	2.023
Valid	25.23	1.057	0.403	1.984	2.020
Test	25.00	1.056	0.384	1.997	2.018

A. NAMED ENTITY RECOGNITION (NER)

NER is the task of identifying mentions of rigid designators from text belonging to predefined semantic types [31]. We gauge the basic capability of language understanding for each language resource [32]. In implementing NER, we trained each PLM to return the probability to be classified to each entity type. We choose this approach for the NER since our focal point is on comparing each language model's capability. Our methodology is centered on evaluating the inherent ability of different language models or training datasets to comprehend the given language effectively. By keeping the approach straightforward, we aim to gauge the raw capacity of each resource without being swayed by auxiliary techniques or methodologies.

1) ENTITY TYPE UNIFORMIZATION

The performance of NER varies depending on the number of entity types being classified [33], [34]. In our experimental setup, we utilize datasets with distinct entity types. **KrC** encompasses six entity types (person, location, official position, time, event), while **KrR** only includes two (person, location), and **ChC** contains three (person, location, official position) entity types. To ensure a fair evaluation and comparison across these settings, we restrict our analysis to the two entity types (person, location) common across all datasets, serving as labels.

B. SENTENCE SPLIT (SENS)

SenS refers to the task of finding segmenting points for a sentence in the multi-sentence document. Considering the lack of punctuation and irregular segmentation in the ancient language (especially **KrC**), it highly relies on the manual labeling of human annotators. Taking this into account, we leverage SenS to validate the language model's capability of understanding the ancient language's syntactic structure.

C. LANGUAGE MODEL POST-TRAINING

To our best knowledge, there is only one language model for the ancient Korean [1], which is post-trained to the ancient Chinese language model trained on BC data initially (AnchiBERT [28]), with **KrR**. To further analyze the effectiveness of constructing a language model using **KrR**, we build our language model by post-training the Chinese language model SikuBERT [35], which was trained on ancient Chinese data from a similar period as **KrR**.

We follow the Masked Language Modeling (MLM) method, which utilizes unlabeled data to train the language model in an unsupervised manner [29]. By comparing the performance of using the Chinese language model directly and the language model post-trained with **KrR**, we aim to analyze the impact of Royal court language resources on understanding **KrC**.

V. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETTING

In this experiment, we employ pre-trained language models (PLMs) trained on Hanja character systems. Specifically, we utilize two representative examples, namely SikuBERT [35][8] and AnchiBERT [28]. SikuBERT is trained on the data from 18C while AnchiBERT is trained on BC data. In order to construct Korean-specific language models, we further train each language model using **KrR** data. For AnchiBERT, we used a Korean-specific language model released by [1], and for SikuBERT, we trained our own model in this experiment. As such, we denote the language models trained with the **KrR** as SikuBERT_Kr and AnchiBERT_Kr, respectively. In addition, we adopt XLM-Roberta-large model [36], which is a widely used multilingual pre-trained language model, to further validate

[8]https://github.com/hsc748NLP/SikuBERT-for-digital-humanities-and-classical-Chinese-information-processing

TABLE 5. Experiments on the PLM adaptability. We mainly report F1-score and auxiliarily denote precision and recall.

Training / Test	KrC	
Task	NER	SenS
AnchiBERT	75.37 (76.25 / 76.26)	93.99 (94.23 / 93.99)
AnchiBERT_Kr	77.59 (78.95 / 77.79)	93.07 (93.05 / 93.55)
SikuBERT	**79.95 (81.35 / 80.11)**	**95.60 (95.73 / 95.65)**
SikuBERT_Kr	76.96 (78.26 / 77.25)	95.06 (95.10 / 95.24)
XLM-R	74.62 (76.07 / 75.21)	93.66 (94.42 / 93.21)

the effectiveness and adaptability of existing resources to the **KrC**.

Post-training of the model and fine-tuning of NER and SenS were all performed using the huggingface library [37].[9] For training, we select 1e-04 as our learning rate through the pilot study among {1e-05, 5e-05, 1e-04, 2e-04}, with the fixed batch size 64 for all experiments. We trained each model with AdamW optimizer [38] upto 50 epochs with early stopping at the highest validation performance. Each training is performed with a single NVIDIA RTX A100 GPU, and the training time was less than 2 hours for **KrC** and **ChC**, and 20 hours for **KrR**.

B. EXPERIMENTS ON THE PLM ADAPTABILITY

First, we train each PLM on the **KrC** data to evaluate the suitability of each language model. The experimental results are presented in Table 5.

As shown in the results, SikuBERT attains the best performance. This can be attributed to the similarity of the time period of the pre-training data to the **KrC**. It demonstrates that even the Chinese language model can derive the best effectiveness in understanding **KrC**. XLM-R yields the lowest performance among the five PLMs, yet it still achieves decent performance with F1-scores of 74.62 for NER and 93.66 for SenS. This is even encouraging as XLM-R is not directly trained with the ancient Korean corpus, which indicates the potential applicability of multilingual models in understanding historical documents.

In particular, by comparing the performance of SikuBERT and SikuBERT_Kr, we can observe the lower performance of SikuBERT_Kr. This results indicate that auxiliary training with **KrR** can lead to a performance degradation in understanding **KrC**. Unlike SikuBERT, AnchiBERT_Kr outperforms AnchiBERT. These results suggest that the training data of AnchiBERT, which is the BC data, facilitates the effectiveness of **KrR** in comprehending relatively modern language (**KrC**). Such results imply that **KrR** may not be the main contributing factor. We further refine this discussion in the following experiments.

C. CROSS-LANGUAGE EXPERIMENTS

Then, we validate the suitability of the existing data for **KrC**. For this purpose, we trained models to perform NER and SenS on **KrR** and **ChC**, respectively, and checked their performance on **KrC**. In doing so, we aim to investigate the efficacy of the Hanja-based language data in comprehending **KrC** data. We utilize the original data without special character processing during the training and testing phases. The experimental results are reported in Table 6. We analyze the results of this experiment as follows.

Both **KrR** and **ChC** exhibit performance above 95 when tested with the same language as the trained dataset. Specifically, **KrR** models achieve F1 scores above 96 for NER and up to 99.85 for SenS. However, such performance is not maintained in applying to **KrC**. **ChC** exhibits even greater performance degradation, with maximum NER and SenS scores of 39.25 and 46.70, respectively. While some performance degradation can be expected in zero-shot language transfer, such a substantial difference in performance suggests limitations in cross-lingual transfer despite the shared language systems in these datasets.

The decrease in performance can be attributed, in part, to the inherent difficulty of the **KrC** test data. Upon comparing the results of the previous baseline experiment, it can be observed that the model trained with **KrR** gives decent performance for NER, with similarly high performance as the model trained with **KrC**. This indicates that the model trained on **KrR** can handle the documents in **KrC** to some extent, even without being explicitly trained on them. However, it should be noted that there is a significant difference in the amounts of **KrR** data and **KrC** training data. This observation suggests that the decent performance of **KrR** may be attributed to the vast amount of the data size, and it still remains a necessity to construct and investigate training data specifically for **KrC** documents. Note that the **KrR**-trained model achieves a maximum SenS performance of 87.08 on **KrC**, representing a significant decrease considering that the data trained on **KrC** attained performance above 95.

D. SPECIAL CHARACTERS ON CROSS-LANGUAGE TRANSFERABILITY

Our previous results demonstrate that models trained on **KrR** and **ChC** exhibit a decrease in performance when applied to **KrC**. We hypothesize that this lack of transferability can be attributed to differences in characteristics between the datasets. Among several factors, we propose ''special character'' as one of the most explicit and unambiguous characteristics that distinguish **KrC** from **KrR** and **ChC**. In this section, we demonstrate the benefits of removing special characters in utilizing existing Hanja-based resources to **KrC**. The experimental results are shown in Table 7.

Our findings indicate that, in the majority of cases, detaching special characters leads to significant improvements in transfer performance. Specifically, for the **ChC**, we observed a maximum enhancement of 40.19 for the SenS task. This observation suggests that standardizing special characters alone can have a substantial impact on the robustness of Hanja-based language systems. Thus, even a simplistic feature unification approach, such as the removal of special characters, can highly aid for handling **KrC**.

[9] https://github.com/huggingface/transformers

TABLE 6. Results on the cross-language experiments. We trained model with KrR and ChC, and tested their performance on KrC. For the comparison, we report its original performance on the same language of the trained dataset.

Training	KrR			
Test	KrC		KrR	
Task	NER	SenS	NER	SenS
AnchiBERT	67.37 (74.99 / 65.40)	78.80 (84.48 / 76.10)	96.30 (96.40 / 96.54)	99.74 (99.75 / 99.74)
AnchiBERT_Kr	65.00 (71.27 / 65.32)	80.46 (88.18 / 76.82)	96.57 (96.61 / 96.88)	99.77 (99.79 / 99.77)
SikuBERT	**76.18 (79.38 / 76.70)**	**87.08 (93.46 / 83.35)**	96.57 (96.63 / 96.86)	99.77 (99.80 / 99.76)
SikuBERT_Kr	69.35 (75.32 / 68.69)	83.78 (91.87 / 79.73)	**96.69 (96.76 / 96.91)**	**99.85 (99.86 / 99.85)**
XLM-R	59.91 (69.59 / 58.20)	82.58 (87.09 / 80.46)	96.25 (96.33 / 96.49)	97.69 (96.60 / 99.13)
Training	**ChC**			
Test	**KrC**		**ChC**	
Task	**NER**	**SenS**	**NER**	**SenS**
AnchiBERT	34.24 (49.93 / 32.50)	45.84 (42.53 / 50.01)	95.05 (95.36 / 95.47)	99.28 (99.24 / 99.32)
AnchiBERT_Kr	37.38 (52.95 / 35.52)	45.84 (42.53 / 50.00)	94.07 (94.44 / 94.55)	97.03 (98.96 / 95.41)
SikuBERT	36.85 (53.09 / 34.34)	**46.70 (46.06 / 50.48)**	**95.48 (95.79 / 95.83)**	**99.26 (99.21 / 99.32)**
SikuBERT_Kr	**39.25 (53.97 / 37.10)**	45.85 (42.75 / 50.00)	94.24 (94.65 / 94.68)	99.26 (99.20 / 99.32)
XLM-R	34.18 (46.27 / 33.15)	45.87 (44.82 / 49.56)	94.07 (94.28 / 94.77)	98.10 (97.51 / 98.80)

TABLE 7. Results on the cross-language transferability derived by the special characters. We report F1 score on this table, and its derivative with the results on the Table 6.

Training / Test	KrR / KrC	
Task	NER	SenS
AnchiBERT	69.48 (△ **2.11**)	84.21 (△ **5.41**)
AnchiBERT_Kr	68.00 (△ **3.00**)	84.82 (△ **4.36**)
SikuBERT	75.40 (▽ **-0.78**)	85.86 (▽ **-1.22**)
SikuBERT_Kr	73.84 (△ **4.49**)	84.93 (△ **1.15**)
XLM-R	70.98 (△ **11.07**)	86.43 (△ **3.85**)
Training / Test	**ChC / KrC**	
Task	**NER**	**SenS**
AnchiBERT	31.94 (▽ **-2.30**)	77.90 (△ **32.06**)
AnchiBERT_Kr	38.11 (△ **0.73**)	80.88 (△ **35.04**)
SikuBERT	37.39 (△ **0.54**)	85.76 (△ **39.06**)
SikuBERT_Kr	39.77 (△ **0.52**)	86.04 (△ **40.19**)
XLM-R	35.13 (△ **0.95**)	74.58 (△ **28.71**)

TABLE 8. Experiments on the performance difference by the special characters.

Training / Test	KrC / KrC	
Task	NER	SenS
AnchiBERT	74.62 (▽ **-0.75**)	93.87 (▽ **-0.12**)
AnchiBERT_Kr	76.73 (▽ **-0.86**)	94.04 (△ **0.97**)
SikuBERT	78.00 (▽ **-1.95**)	95.85 (△ **0.25**)
SikuBERT_Kr	76.29 (▽ **-0.67**)	95.67 (△ **0.61**)
XLM-R	76.95 (△ **2.33**)	94.02 (△ **0.36**)
Training / Test	**KrR / KrC**	
Task	**NER**	**SenS**
AnchiBERT	95.77 (▽ **-0.53**)	97.87 (▽ **-1.87**)
AnchiBERT_Kr	96.31 (▽ **-0.26**)	97.96 (▽ **-1.81**)
SikuBERT	96.24 (▽ **-0.33**)	98.01 (▽ **-1.76**)
SikuBERT_Kr	96.49 (▽ **-0.20**)	98.02 (▽ **-1.83**)
XLM-R	95.83 (▽ **-0.42**)	98.01 (△ **0.32**)
Training / Test	**ChC / KrC**	
Task	**NER**	**SenS**
AnchiBERT	94.44 (▽ **-0.61**)	92.06 (▽ **-7.22**)
AnchiBERT_Kr	93.67 (▽ **-0.40**)	90.06 (▽ **-6.97**)
SikuBERT	95.16 (▽ **-0.32**)	94.18 (▽ **-5.08**)
SikuBERT_Kr	94.59 (△ **0.35**)	93.85 (▽ **-5.41**)
XLM-R	93.69 (▽ **-0.38**)	90.32 (▽ **-7.78**)

E. SPECIAL CHARACTERS ON THE TASK PERFORMANCE

To delve deeper into the influence of special characters, we conduct an extensive analysis of the performance on the identical language of the trained dataset. Specifically, we examine the disparity in performance between models trained on datasets devoid of any special characters and models trained on the original dataset. The experimental results are detailed in Table 8.

First, we observed that the removal of special characters in **KrC** resulted in a decrease in performance for NER. This can be interpreted that symbols such as markers aid comprehension of the phrases within a sentence by indicating their respective roles. However, in SenS, we observed a performance improvement. We find that such results are related to the irregular attachment of punctuation marks in **KrC**, where seldomly adopted punctuation marks hinder understanding of the splitting point for a sentence.

In contrast, our empirical results demonstrate that in processing **KrR** and **ChC** datasets, removal of special characters generally leads to a considerable decline in performance. This decline is particularly pronounced in the case of SenS. More specifically, the removal of special characters resulted in a degradation of 1.87 F1 score for **KrR** and 7.22 score for **ChC**. It implies special characters play a significant role in sentence comprehension and segmental parsing, especially for highly structured data where punctuation rules are uniformly applied across datasets.

F. QUALITATIVE ANALYSIS

To further highlight the limitations of directly exploiting **KrR** to **KrC**, we conducted a qualitative analysis of error cases arising from the trained model. Specifically, we focus

on the error cases of AnchiBERT_Kr, which showed the lowest performance when utilizing **KrR** as a training data to process **KrC** (Table 7). Notably, we have verified the practical effectiveness of our methodology of removing special characters in the training data, by comparing two cases: one where **KrR** was used as training data directly (Royal_original), and one where the special characters were removed before using it as training data (Royal_refined).

The experimental results are shown in Table 9. These results align with our quantitative analysis results, showing higher accuracy in the model where special characters were removed. Specifically, in models trained with data retaining special characters, we observed difficulties in identifying sentence splitting points within **KrC** or pinpointing the location of named entities.

This can be perceived as the significant influence of special characters on word interpretation within sentences. In the case of Royal_original, it tends to focus more on the meaning signified by the special characters present in the training data, making it difficult to comprehend sentences like **KrC** where special characters are rare. On the contrary, when special characters are removed in the training data, the trained model can focus more on the role of each Hanja character within the sentence. This enables acquisition of robust knowledge applicable to **KrC**, even with training on **KrR**.

VI. DISCUSSION

In response to potential concerns related to this study, we wish to provide the following explanations.

A. OTHER FEATURES BESIDE SPECIAL CHARACTERS

One of the most significant characteristics of **KrC** documents, suggested in this paper, is the use of special characters such as word or sentence delimiters. **KrC** uses special characters markedly less frequently compared to **KrR** or **ChC**. This characteristic is not exclusively relevant to the data used in our experiment; rather, it is a low-level feature commonly shared among nearly all **KrC** documents.[10] Since our method does not consider the unique characteristics of the data domain, we believe it can robustly apply to any existing or future corpus.

There could be multiple differences in properties between **KrC** and existing Hanja-based language sources (i.e. **KrR** and **ChC**) beyond the attachment of special characters. However, analyzing these differences could lead to research that is narrowly valid within the data we collected, and raise questions about its generalizability. As a first step of the **KrC** research, we have focused primarily on its fundamental and universal features, thereby allowing our proposed directions to be easily re-implemented in other studies.

We have also confirmed that even with very simple low-level actions such as removing special characters, the usefulness of existing resources greatly increases in the context of work on **KrC**. By experimentally validating that mere alignment of characteristics can considerably enhance cross-lingual transferability, we open avenues for potential future studies related to the investigation of finer data characteristics.

[10]https://www.krpia.co.kr/knowledge/ugyo/main

B. RELATIVELY SMALL DATA SIZE

While the dataset used in our experiment may seem relatively small, when this is restructured into sentence-level data, an average sentence length of 12.54 yields a total of 36,640 sentences. Although this amount may not be able to fully represent the characteristics of all existing **KrC** documents, we believe it is sufficient to depict its general properties. Especially in historical Korean documents, due to the semantic richness of each character (ideograph) significantly differing from English, a sentence can be formed by very few characters. For instance, a 4-character length sentence such as "試令遣諭" from classical Korean can be translated to "As an experimental measure, we dispatched individuals to notify the participants" in English. From this perspective, we believe that although the amount of data used for the experiment is not large, it is sufficient for demonstrating our argument.

C. TASK SELECTION AND ITS IMPLICATION

We begin by acknowledging that the majority of Korean historical documents written by common literati are not yet digitized. While government-led efforts are underway to accomplish this task, the scale of the work necessitates a large workforce and progresses at a relatively slow pace due to its heavy reliance on human resources.

The data we used for our experiments originated from non-digitized, real-world documents. The process of transcribing these offline documents into usable data surfaced two key tasks that the workers needed most: NER and Sens. These tasks are immensely important in classical Korean texts recorded in Hanja writing system, where neither word separators (including whitespace) nor sentence separators (including punctuation) are used. In documents where there is no distinction between words or even sentences (as shown on the right side of Figure 1), it is challenging to decipher the meaning of a sentence, identify nouns, or determine the subject within the sentence. In this context, the two tasks of distinguishing sentence units (SenS) and the role of words within sentences (NER) are deemed very significant for constructing a database from scratch. Considering many **KrC** documents have yet to be digitized, the importance of the task we experimented cannot be overstated.

In this context, exploring existing sources and analyzing their applicability can be regarded as practical efforts. Notably, our experiments presented in Table 5 reveal that utilizing Chinese language data might be a more effective strategy than using Korean royal language data for sentence splitting tasks for commoner's documents. This insight, which cannot be gleaned from simply applying methodologies without taking into account data specifics, illustrates the

TABLE 9. Qualitative analyses. Above samples are inference results obtained by training AnchiBERT_Kr model with KrR, and test it with KrC.

Text	洪君受過瞥見路上盖子**瞻君**受向**蘆川**也
Meaning	**Hong Gunsu** stopped by and watched them for a while on the road. Usually, **Zachum** and Gunsu were on their way to the **Nocheon.**
NER Label	{**洪君受**: Person} {**瞻君**: Person} {**蘆川**: Location}
Royal_refined	{**瞻君**: Person} {**蘆川**: Location}
Royal_original	{**瞻**: Person} {**川**: Location}
Text	**以志**仍祭于**清叔父**墓
Meaning	Then **Yiji** held a memorial service for his **uncle Ubcheong** in the graveyard.
NER Label	{**以志**: Person} {**清叔父**: Person}
Royal_refined	{**以志**: Person} {**清**: Person}
Royal_original	No Named Entity
Text	**孟厚**服史**雲洞**慰柬封
Meaning	Mourner from **Undong** sent a letter of sympathy to a **Manghu** who is in funeral
NER Label	{**孟厚**: Person} {**雲洞**: Location}
Royal_refined	{**孟厚**: Person} {**雲洞**: Location}
Royal_original	{**史雲洞**: Location}
Text	客榻問狀柯亭金謹封固慮有行李之勞
Meaning	To ask how one is doing in a foreign country. Sent by Kim from Gajeong. Best regards. I was originally worried that you would have travel fatigue.
SenS Label	客榻問狀 // 柯亭金謹封 // 固慮有行李之勞
Royal_refined	客榻問狀 // 柯亭金謹封 // 固慮有行李之勞
Royal_original	客榻問狀柯亭金謹封固慮有行李之勞
Text	吾亦欲同往子無可得無可奈何
Meaning	I wanted to go with you. However, I couldn't find a horse to get down there. There was nothing I could do.
SenS Label	吾亦欲同往 // 子無可得 // 無可奈何
Royal_refined	客榻問狀 // 柯亭金謹封 // 固慮有行李之勞
Royal_original	客榻問狀柯亭金謹封 // 固慮有行李之勞
Text	聰明水在臺東步致遠庵又在其南山路甚狹絶壁萬
Meaning	Chongmingshu was a few steps east of Nudae. Qi Yuan Temple was just south of it. The mountain path was very narrow. The cliffs were ten thousand feet high.
SenS Label	聰明水在臺東步 // 致遠庵又在其南 // 山路甚狹 // 絶壁萬
Royal_refined	聰明水在臺東步 // 致遠庵 // 又在其南 // 山路甚狹 // 絶壁萬
Royal_original	聰明水在臺東步 // 致遠庵又在其南 // 山路甚狹絶壁 //萬

need for thoughtful consideration of data when leveraging existing resources in working with **KrC**.

Specifically, to alleviate ambiguity and clarify meanings in **KrC**, experts traditionally decipher the sentences by considering their context, which is not intuitive and requires intensive human labor. Accordingly, the application of existing resources to assist in this deciphering process can be deemed a highly practical approach.

VII. CONCLUSION

This study prompted initial attempts to processing historical Korean documents written by the Korean common literati (**KrC**). Specifically, considering that the **KrC** shares the same writing systems with ancient Chinese (**ChC**) and Korean documents written by the royal court (**KrR**), we evaluate the suitability of Hanja-based resources. Through this evaluation, we have demonstrated that the direct application of **KrR** for **KrC** may bear several limitations, and **ChC** also exhibits low applicability. We hypothesize that this low suitability is due to differences in the characteristics of the data, and we have proposed a simple yet effective method of removing special characters to address this issue. As a result, we have significantly increased the potential applicability of existing Hanja-based resources to **KrC**. To the best of our knowledge, this study is the first attempt to understand historical Korean common literati documents, and it provides a foundation for strategies to utilize existing resources for understanding. As suggested by our findings, we aim to further enhance this transferability through more advanced methods that unify the characteristics of the language, taking into consideration the specific nature of the language.

REFERENCES

[1] H. Yoo, J. Jin, J. Son, J. Bak, K. Cho, and A. Oh, "HUE: Pretrained model and dataset for understanding hanja documents of ancient Korea," in *Proc. Findings Assoc. Comput. Linguistics*, 2022, pp. 1832–1844.

[2] J. Bak and A. Oh, "Five centuries of monarchy in korea: Mining the text of the annals of the joseon dynasty," in *Proc. 9th SIGHUM Workshop Lang. Technol. Cultural Heritage, Social Sci., Humanities (LaTeCH)*, 2015, pp. 10–14.

[3] J.-R. Cho and H.-C. Chen, "Orthographic and phonological activation in the semantic processing of Korean hanja and hangul," *Lang. Cognit. Processes*, vol. 14, nos. 5–6, pp. 481–502, Oct. 1999.

[4] J.-R. Cho and H.-C. Chen, "Semantic and phonological processing in reading Korean hangul and hanja words," *J. Psycholinguistic Res.*, vol. 34, no. 4, pp. 401–414, Jul. 2005.

[5] K. Kang, K. Jin, S. Yang, S. Jang, J. Choo, and Y. Kim, "Restoring and mining the records of the joseon dynasty via neural language modeling and machine translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2021, pp. 4031–4042.
[6] S. Yang, M. Choi, Y. Cho, and J. Choo, "HistRED: A historical document-level relation extraction dataset," 2023, *arXiv:2307.04285*.
[7] J. Son, J. Jin, H. Yoo, J. Bak, K. Cho, and A. Oh, "Translating hanja historical documents to contemporary Korean and English," 2022, *arXiv:2205.10019*.
[8] S. L. Kim, T. Jang, J. Ahn, H. Lee, and J. Lee, "Transfer learning across several centuries: Machine and historian integrated method to decipher royal secretary's diary," 2023, *arXiv:2306.14592*.
[9] S. Kim, "The new aspect of the narrative literature written in Chinese characters in the later period of joseon dynasty," *J. Korean Classical Chin. Literature*, vol. 18, no. 1, pp. 7–37, Jun. 2009.
[10] E. Boros, E. L. Pontes, L. A. Cabrera-Diego, A. Hamdi, J. G. Moreno, N. Sidère, and A. Doucet, "Robust named entity recognition and linking on historical multilingual documents," in *Proc. Conf. Labs Eval. Forum (CLEF)*, vol. 2696, 2020, pp. 1–17.
[11] Y. Assael, T. Sommerschield, B. Shillingford, M. Bordbar, J. Pavlopoulos, M. Chatzipanagiotou, I. Androutsopoulos, J. Prag, and N. de Freitas, "Restoring and attributing ancient texts using deep neural networks," *Nature*, vol. 603, no. 7900, pp. 280–283, Mar. 2022.
[12] M. Corazza, F. Tamburini, M. Valério, and S. Ferrara, "Unsupervised deep learning supports reclassification of bronze age cypriot writing system," *PLoS One*, vol. 17, no. 7, Jul. 2022, Art. no. e0269544.
[13] T. Clanuwat, A. Lamb, and A. Kitamoto, "KuroNet: Pre-modern Japanese kuzushiji character recognition with deep learning," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 607–614.
[14] F. Lombardi and S. Marinai, "Deep learning for historical document analysis and recognition—A survey," *J. Imag.*, vol. 6, no. 10, p. 110, Oct. 2020.
[15] R. Quirk and C. Wrenn, *An Old English Grammar*. Evanston, IL, USA: Routledge, 2002.
[16] X. Chang, F. Chao, C. Shang, and Q. Shen, "Sundial-GAN: A cascade generative adversarial networks framework for deciphering Oracle bone inscriptions," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 1195–1203.
[17] Y. Assael, T. Sommerschield, and J. Prag, "Restoring ancient text using deep learning: A case study on Greek epigraphy," in *Proc. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 6368–6375.
[18] A. Hamid, M. Bibi, M. Moetesum, and I. Siddiqi, "Deep learning based approach for historical manuscript dating," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 967–972.
[19] Y. Tian and Y. Guo, "Ancient Chinese word segmentation and part-of-speech tagging using data augmentation," in *Proc. 2nd Workshop Lang. Technol. Historical Ancient Lang.*, 2022, pp. 146–149.
[20] A. Peyraube and R. Woodard, "Ancient Chinese," in *The Routledge Encyclopedia of the Chinese Language*. New York, NY, USA: Routledge, 2016, pp. 39–55.
[21] D. Bamman and P. J. Burns, "Latin BERT: A contextual language model for classical philology," 2020, *arXiv:2009.10053*.
[22] T. Yousef, C. Palladino, D. J. Wright, and M. Berti, "Automatic translation alignment for ancient Greek and Latin," in *Proc. 2nd Workshop Lang. Technol. Historical Ancient Lang.*, 2022, pp. 101–107.
[23] K. P. Johnson, P. J. Burns, J. Stewart, T. Cook, C. Besnier, and W. J. B. Mattingly, "The classical language toolkit: An NLP framework for pre-modern languages," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., Syst. Demonstrations*, 2021, pp. 20–29.
[24] J. Kim, "Investigating phonological processing in visual word recognition: The use of Korean hangul (alphabetic) and hanja (logographic) scripts," Ph.D. dissertation, UNSW Sydney, Sydney, NSW, Australia, 1998.
[25] C. Park, C. Lee, Y. Yang, and H. Lim, "Ancient Korean neural machine translation," *IEEE Access*, vol. 8, pp. 116617–116625, 2020.
[26] M. Cartwright, "Ancient Korean & Chinese relations," *Ancient Hist. Encyclopedia*, vol. 30, Nov. 2016.
[27] C. Park, S. Lee, J. Seo, H. Moon, S. Eo, and H.-S. Lim, "Priming ancient Korean neural machine translation," in *Proc. 13th Lang. Resour. Eval. Conf.*, 2022, pp. 22–28.
[28] H. Tian, K. Yang, D. Liu, and J. Lv, "AnchiBERT: A pre-trained model for ancient Chinese language understanding and generation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.
[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.
[30] R. Sprugnoli and M. Passarotti, "Proceedings of the second workshop on language technologies for historical and ancient languages," in *Proc. 2nd Workshop Lang. Technol. Historical Ancient Lang.*, 2022.
[31] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Mar. 2020.
[32] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[33] M. Peng, R. Ma, Q. Zhang, L. Zhao, M. Wei, C. Sun, and X. Huang, "Toward recognizing more entity types in NER: An efficient implementation using only entity lexicons," in *Proc. Findings Assoc. Comput. Linguistics*, 2020, pp. 678–688.
[34] A. Roy, "Recent trends in named entity recognition (NER)," 2021, *arXiv:2101.11420*.
[35] D. Wang, C. Liu, Z. Zhao, S. Shen, L. Liu, B. Li, H. Hu, M. Wu, L. Lin, X. Zhao, and X. Wang, "GujiBERT and GujiGPT: Construction of intelligent information processing foundation language models for ancient texts," 2023, *arXiv:2307.05354*.
[36] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8440–8451.
[37] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 38–45.
[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2018.

HYEONSEOK MOON received the B.S. degree from the Department of Science in Mathematics and Engineering in Artificial Intelligence, Korea University, Seoul, South Korea, in 2021. He is currently pursuing the Ph.D. degree in computer science and engineering with Korea University. His research interests include machine translation, natural language generation, and efficient tuning for NLP engineering.

MYUNGHOON KANG received the B.S. degree from the Department of Urban Sociology and Data Science, University of Seoul, Seoul, South Korea, in 2022. He is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul. Currently, under an integrated master and Ph.D. course. He is a part of the Natural Language Processing and Artificial Intelligence Laboratory. His research interests include natural language understanding and fake news detection.

JAEHYUNG SEO received the B.S. degree from the Department of English Language and Literature, Korea University, Seoul, South Korea, in 2020. He is currently pursuing the Ph.D. degree in computer science and engineering with Korea University. Currently, under an integrated master's and Ph.D. course. He is a part of the Natural Language Processing and Artificial Intelligence Laboratory Team. His research interests include language generation and decoding strategies, where he attempts to find inspiration from how humans do so and build a generative model based on common-sense reasoning.

SUGYEONG EO received the B.A. degree in linguistics and cognitive science, language, and technology from Hankuk University of Foreign Studies, Yongin, South Korea, in 2020. She is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul, South Korea. She is a Faculty Member with the Natural Language Processing and Artificial Intelligence Laboratory. Her research interests include neural machine translation, quality estimation, and question generation.

CHANJUN PARK received the Ph.D. degree from Korea University, under the supervision of Prof. Heuiseok Lim, with a focus on "data-centric neural machine translation," in 2023. From 2018 to 2019, he was with SYSTRAN, as a Research Engineer. He is currently a Researcher of natural language processing (NLP), with a focus on data-centric AI, machine translation, and large language models (LLM). He is also a Principal Research Engineer and the Technical Leader of Upstage LLM Team. He is the Founder and the Chief Scientist of the KU-NMT Group. He has published more than 180 articles in the field of NLP. He has initiated projects, such as SOLAR, Open Ko-LLM Leaderboard, Dataverse, Evalverse, and up one trillion token club.His current research interest includes forming an LLM-based ecosystem. He received the Naver Ph.D. Fellowship in 2021. He served as the Virtual Social Chair for COLING 2022. He is also serving as the Program Chair for the WiNLP Workshop and the Publication Chair for the DMLR Workshop. He is selected for Forbes 30 Under 30 Korea in the SCIENCE/SW field.

YEONGWOOK YANG received the master's degree in computer science education and the Ph.D. degree from the Department of Computer Science and Engineering, Korea University, Seoul, South Korea. He was a Research Professor with the Department of Computer Science and Engineering, Korea University, for one year. He was a Senior Researcher with the University of Tartu, Tartu, Estonia. He was an Assistant Professor with the Division of Computer Engineering, Hanshin University. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Gangneung-Wonju National University. His research interests include information filtering, recommendation systems, educational data mining, and deep learning.

HEUISEOK LIM received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1992, 1994, and 1997, respectively. He is currently a Professor with the Department of Computer Science and Engineering, Korea University. His research interests include natural language processing, machine learning, and artificial intelligence.