HUE: Pretrained Model and Dataset for Understanding Hanja Documents of Ancient Korea

Haneul Yoo[1], **Jiho Jin**[1], **Juhee Son**[1], **JinYeong Bak**[2], **Kyunghyun Cho**[3], **Alice Oh**[1]
[1]KAIST, South Korea, [2]Sungkyunkwan University, South Korea, [3]New York University, USA
{haneul.yoo, jinjh0123, sjh5665}@kaist.ac.kr,
jy.bak@skku.edu, kyunghyun.cho@nyu.edu, alice.oh@kaist.edu

Abstract

Historical records in Korea before the 20th century were primarily written in Hanja, an extinct language based on Chinese characters and not understood by modern Korean or Chinese speakers. Historians with expertise in this time period have been analyzing the documents, but that process is very difficult and time-consuming, and language models would significantly speed up the process. Toward building and evaluating language models for Hanja, we release the Hanja Understanding Evaluation dataset consisting of chronological attribution, topic classification, named entity recognition, and summary retrieval tasks. We also present BERT-based models continued pretraining on the two major corpora from the 14th to the 19th centuries: the Annals of the Joseon Dynasty and Diaries of the Royal Secretariats. [1] We compare the models with several baselines on all tasks and show there are significant improvements gained by training on the two corpora. Additionally, we run zero-shot experiments on the Daily Records of the Royal Court and Important Officials (DRRI). The DRRI dataset has not been studied much by the historians, and not at all by the NLP community.

1 Introduction

Large-scale historical records in Korea were mostly produced during the Joseon dynasty (1392-1897), and the Institute for the Translation of Korean Classics (ITKC) keeps a comprehensive database of Korean classics at a scale of approximately 9 billion characters. This digital archive is a great resource for Korean historians, but the documents remain in the original Hanja language [2]. Hanja is an extinct language, and as Table 1 illustrates with a simple sentence, Hanja is lexically and syntactically different from modern Korean, as well as simplified and traditional Chinese. Understanding the documents in the digital archive is thus difficult and would benefit greatly from a Hanja language model which can also be used to accelerate the expert translation (Vale de Gato, 2015). There are two corpora we can use to train the language model, the Annals of Joseon Dynasty (AJD), first introduced to the NLP community in (Bak and Oh, 2015), and the Diaries of the Royal Secretariats (DRS) (Kang et al., 2021).

Language	Sentence
Hanja	上 御 經筵 。
Modern Korean	임금 이 경연 에 나아갔다 .
Simplified Chinese	国王 参加 了 皇家讲座 。
Traditional Chinese	國王 參加 了 皇家講座 。
English	The King attended the Royal Lecture .

Table 1: Example sentence in AJD

In this paper, we provide the HUE (**H**anja **U**nderstanding **E**valuation) dataset consisting of chronological attribution, topic classification, named entity recognition and summary retrieval, a suite of tasks to help build and evaluate the Hanja language model. In addition to AJD and DRS, we also work with the Daily Records of the Royal Court and Important Officials (DRRI). Unlike AJD and DRS that have been analyzed by historians and contain their annotations, DRRI lacks such systematic analysis, and we use it for zero-shot learning and introduce it to the NLP community.

We also provide pretrained language models (PLMs) for Hanja trained on AJD and DRS, fine-

[1]All codes, models, and dataset are available at https://github.com/haneul-yoo/HUE.git

[2]Hanja is a set of characters (script) used in ancient Korean, while Hanmun is a writing style (language) in the same era. However, we will refer to Hanja as a language following the conventions of the previous works.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1832 - 1844

tuned for each task in HUE. Our pretrained models on the corpora from that era outperform the existing language models built for ancient Chinese, confirming the need for specially-trained Hanja language models. We also run additional experiments based on the analyses of entity- and word-level changes on AJD by controlling input conditions by masking named entity and giving the time period as input. Finally, we demonstrate the effectiveness of our Hanja language model for analyzing unseen documents, running zero-shot experiments for chronological attribution and named entity recognition tasks on DRRI.

Our main contributions are as follows:

- We release the HUE dataset and Hanja PLMs to support historians to understand and analyze a large volume of historical documents written in Hanja. To the best of our knowledge, this is the first work proposing Hanja language models and releasing a NLP benchmark dataset for ancient Hanja documents.
- We demonstrate that providing key information such as named entity and document age as input improves the performance of Hanja language model on the HUE tasks.
- We run zero-shot experiments on several HUE tasks from DRRI which have not been discussed in the NLP community, and demonstrate the performance of our Hanja language models on unseen historical documents.

2 Background

2.1 Hanja

Hanja, the writing system based on ancient Chinese characters, was the main writing system in Korea before the 20th century, while Hangul, the unique Korean alphabet, has been the main writing system in Korea from the last century. Formal records from the Joseon dynasty (1392-1897) are written in Hanja, while spoken language and some written documents were in Hangul, developed in the 15th century. This co-existence of the written and colloquial languages has led Hanja to evolve to have the basic syntax of classical Chinese, mixing with the lexical, semantic, and syntactic characteristics of colloquial Korean.

Hanja is significantly different from both modern Korean and modern Chinese. Modern Korean uses a different alphabet and structure, and traditional Chinese shares some characters with Hanja, while the lexicon has evolved greatly to reflect the temporal, geographical, and cultural differences between the Joseon dynasty and modern-day China. Simplified Chinese, the current written language in China has diverged more because of the simplification of the characters. These differences between Hanja and other related languages would lead to suboptimal performance when adopting the current Chinese language models to NLU tasks for the Korean historical Hanja documents.

2.2 Dataset

We describe the three corpora of records written in Hanja during the Joseon dynasty, whose contents and additional information such as topic and named entities are provided by historians in IKTC [3]. Table 2 shows the list of the Hanja corpora used.

Annals of the Joseon Dynasty (AJD) also called Veritable Records of the Joseon Dynasty, is a corpus of 27 sets of chronological records, and each set covers one ruler's reign. AJD has been translated into Korean from 1968 to 1993 and includes relevant tags such as the named entities and dates of the documents [4]. We use AJD for both training our Hanja language models and building the HUE dataset of NLP tasks.

Diaries of the Royal Secretariat (DRS) is a corpus of detailed records of daily events and official schedules of the court from the first King Taejo to the last (27th) Sunjong. Many of the earlier records were lost, and we use the extant records starting from the 16th King Injo. DRS is known to hold the largest amount of authentic historic records and state secrets of the Joseon Dynasty [5]. We use DRS to continue pretraining the language models.

Daily Records of the Royal Court and Important Officials (DRRI) is a corpus of journals written from the 21st King Yeongjo to the last Emperor Sungjong and presumably initiated from the diaries of the crown prince who became the 22nd King Jeongjo after he was enthroned. DRRI has official daily records from both the central and the local governments, so encompasses all events in the country and reports to the king with summaries. DRRI is known to include details and events of

[3] https://db.itkc.or.kr/
[4] http://esillok.history.go.kr/
[5] http://english.cha.go.kr/

Dataset	Size	Training data	Downstream Tasks	Zero-shot	King
AJD	230K	✔	CA, TC, NER	-	Taejo (1^{st}) - Sunjong (27^{th})
DRS	1,380K	✔	-	-	Injo (16^{th}) - Sunjong (27^{th})
DRRI	426K	-	SR	CA, NER	Yeongjo (21^{st}) - Sunjong (27^{th})

Table 2: Source corpora chosen for building HUE dataset and PLMs

the late Joseon Dynasty not recorded in the AJD or DRS [5], thus making it a good corpus for zero-shot experiments. We use DRRI for the supervised summary retrieval task and the zero-shot experiments for chronological attribution and named entity recognition.

3 HUE Dataset

The HUE (**H**anja **U**nderstanding **E**valuation) dataset is built to assist history scholars to understand Korean historical records written in Hanja. HUE consists of chronological attribution, topic classification, named entity recognition, and summary retrieval, which are tasks that can provide helpful information for studying the documents. We expect that the language models based on HUE will ultimately help historians to interpret unseen historical documents and public to grasp basic concept of those documents. We describe each task in detail below.

3.1 Task Description

Chronological Attribution (CA) is a classification task predicting the ruling king when the document was written. When given a Hanja document from AJD, a classifier outputs one of the 27 kings of the Joseon dynasty. Chronological attribution of the undiscovered document is the first step in anthology to interpret and translate it. Korean historians mostly divide the history of the Joseon Dynasty based on the reigning king, so that we treat chronological attribution as a classification task.

Topic Classification (TC) is a multi-class and multi-label classification task to find the topics of the given document. For TC, we use Hanja document from AJD. We suggest two levels of topics, namely major and minor categories. The major categories consist of 4 broad topics: politics, economy, society, and culture. The minor categories go with 106 sub-topics such as diplomacy, agriculture, and science.

Named Entity Recognition (NER) is a sequence tagging task, identifying the two types of named entities, person and location, from the Hanja document from AJD. We divide train, validation, and test sets such that there are no overlapping entities across the sets.

Summary Retrieval (SR) is a task to find the most relevant summary that matches the content among the summary candidates. For this task, we use DRRI, in which each document is a pair of summary (*gang*) and detailed content (*mok*). Among 426k articles, 265k articles in DRRI dataset contain both *gamg* and *mok*. Also, we exclude those with *gang* longer than *mok*, in which *gang* is not the summary of *mok*. The final dataset contains 213K pairs of content and the corresponding summaries. We describe more details of the preprocessing in the Appendix.

4 Hanja Pretrained Model

As far as we know, there have been no pretrained language models for the Hanja language. One can use related LMs, the pretrained models for ancient Chinese as well as multilingual BERT (Devlin et al., 2019) which includes traditional Chinese in its training corpus. AnchiBERT (Tian et al., 2021) is pretrained in ancient Chinese with the Chinese anthologies written around 1000BC to 200BC. There is some vocabulary overlap between the Hanja documents and traditional Chinese corpora, we can adopt multilingual BERT and AnchiBERT to learn the representations of the Hanja texts.

We propose the pretrained language models suitable for Hanja documents by continuing pretraining those two models on both AJD and DRS. Table 3 shows the ratio of unknown tokens in the test set of AJD by each model. It implies that existing AnchiBERT and multilingual BERT can also be exploited as language models for Hanja documents written in the Joseon dynasty, but the second phase of pretraining on the corpora of that era remarkably decreases the ratio of unknown tokens.

	AnchiBERT (Tian et al., 2021)	mBERT (Devlin et al., 2019)
original	0.88%	0.76%
+AJD/DRS	0.04%	0.04%

Table 3: Unknown token ratio of each model in test set for CA, TC, and NER task in HUE. The first row indicates the results of the original PLMs without any additional pretraining, and the second row indicates those with continued pretraining on AJD and DRS.

5 Experiment

5.1 Experimental Settings

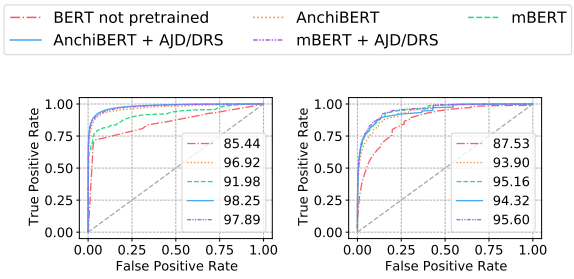
We conduct experiments on HUE with our pretrained model. For the baseline model, we use BERT without pretraining and compare it to various BERT models described in Section 3.1. Specifically, we fine-tune each model to act as a re-ranker in the retrieval task for the summary retrieval. We first retrieve top-k relevant *gang*s (summaries) with BM25 (Robertson and Zaragoza, 2009) among all *gang*s in the dataset. Then, we fine-tune the model as a binary classifier to determine whether the summary matches the content of the *mok* with the cross-entropy loss (Nogueira and Cho, 2019). If the ground truth summary is not included in the top-k relevant summary candidates, we replace the last k^{th} summary with the ground truth. We use $k = 12$ for training and $k = 100$ for inference. As the representative metrics, we present F1 scores for CA, TC, and NER tasks, and Mean Reciprocal Rank (MRR) for SR. The detailed results with other metrics are also available in Appendix.

5.2 Overall Results

Table 4 shows the overall experimental results. The bold and the underlined texts in the table specify the best and the second best result, respectively. BERT without any pretraining shows the poorest results across all the tasks. AnchiBERT and mBERT, which are existing language models on the relevant domains, show better results, and the models continued pretraining on Hanja documents achieve the best performance among all tasks. This tendency indicates that all these tasks on understanding historical documents require pretraining language models on time-specific and domain-specific data.

AnchiBERT pretrained on the Hanja corpora shows slightly better performance than mBERT pretrained on the same corpora. We assume this is because the original training corpora of AnchiBERT are much closer to the Hanja documents, even the era of those two corpora are completely different. The writing style of both Hanja documents in the Joseon dynasty and anthologies in ancient China had come from Classical Chinese and share similarities. On the other hand, the training corpora of mBERT is a contemporary texts whose characters contains Traditional Chinese, but the structure and the format might be considerably changed.

Chronological Attribution (CA) Our models continued pretraining on Hanja corpora outperform other baselines on CA. Detailed analysis on CA result is illustrated in Section 6.1.



(a) Div 1 (4 classes) (b) Div 2 (106 classes)

Figure 1: ROC Curve and AUC for Topic Classification. Each value in the legend indicates the AUC score.

Topic Classification (TC) Figure 1 gives ROC curves and AUC values of each model on each task. Our models show the similar trends to the overall results, outperforming other language models. For the evaluation results including F1 score, we find and set the best threshold to each label by Youden's index (Youden, 1950).

While F1 score goes down as the number of classes increases from 4 to 106, there is no significant difference on AUC value. This might result from consistently high recall achieving around 90% on both tasks. It indicates that the threshold is too low and models tend to predict plausible topics as many as they can, which might be solved by controlling the threshold. AnchiBERT pretrained on AJD and DRS, which shows the best performance, predicts 6.39 labels in average, while the average number of ground truth labels of the minor categories is 1.97. This is probably due to the meaning overlaps in minor categories. For instance, minor categories such as revenue, finance, general price level, and commerce are the sub-categories of economy in the major categories, whose use case cannot be strictly distinguished. It would be more appro-

	CA	TC		NER		SR
		Major	Minor	Person	Location	
	F1	F1	F1	F1	F1	mrr
BERT not pretrained	54.26	68.91	61.52	92.13	87.10	52.85
mBERT (Devlin et al., 2019)	75.29	79.59	76.46	91.63	86.02	67.06
AnchiBERT (Tian et al., 2021)	75.74	85.81	75.22	**93.28**	**88.01**	67.92
mBERT + AJD/DRS	77.77	87.13	77.84	92.83	87.90	73.88
AnchiBERT + AJD/DRS	**79.33**	**88.33**	**78.10**	93.13	87.91	**74.29**

Table 4: Evaluation results of PLMs on HUE dataset

priate in this case to provide all plausible topics roughly rather than suggesting the one only with high certainty.

Named Entity Recognition (NER) NER also indicates similar trends to the overall benchmark tasks, but with a small gap among models including BERT without pretraining. It implies that NER in Hanja documents is a comparably easy task. This might result from certain patterns in named entities in Hanja. Most of the person entities are 3 letters starting with the common characters (family name), and most of the location entities end with the common characters meaning locations or buildings. All models tend to predict person entities better than location entities.

Summary Retrieval (SR) All fine-tuned BERT-based re-rankers outperform BM25 whose MRR is merely 29.87%, mostly retrieving the ground truth answer at the first trial. Likewise, our models shows the best results, while BERT without pretraining show the lowest MRR. It additionally implies that BERT-based re-rankers might be exploited for retrieving relevant documents from different chronicles in terms of written style or contents.

5.3 Effect of Entity and Document Age on Language Model

We investigate whether providing additional information as input can improve the performances of language models. For CA, we mask all named entities in the input and fine-tuned the language model on masked data to examine the impact of entity information. For TC and NER, we concatenate document age information to the input text and run comparative experiments to verify the importance of time period on historical texts.

	CA (Masked)
BERT not pretrained	75.07 (Δ 20.81)
mBERT (Devlin et al., 2019)	83.44 (Δ 8.15)
AnchiBERT (Tian et al., 2021)	82.45 (Δ 6.71)
mBERT + AJD/DRS	83.57 (Δ 5.80)
AnchiBERT + AJD/DRS	**83.58** (Δ 4.25)

Table 5: F1 scores on Chronological Attribution given named entities masked. The value inside parenthesis indicates the increase in performance after masking the named entities.

Entity-Masked Chronological Attribution Table 5 shows the difference on experimental results in CA when the named entities in the given input texts are masked. Compared to the default settings which does not mask named entities, all models show significant improvements. This is probably because models can truly focus on the content and changes in writing style without any disturbance of location entities consistently used for the whole era. Fine-tuned models with entity masked inputs also achieve nearly the same level of performance in the inference with plain inputs including named entities. It suggest to fine-tune models masking named entity in CA, considering the real scenario whose inference texts lack time period information.

Topic Classifciation and Named Entity Recognition with the Age of Document It is for granted to regard that historical texts written over several eras reveal the time changes with respect to lexical choices and contents. Section 6.2 confirms the hypothesis above in terms of n-grams changes over time. Table 6 shows gaps between experimental results of TC and NER given which king reigned when the document was written. Providing document age definitely increases the performance of classifying topics and tagging named

	TC		NER	
	Major	**Minor**	**Person**	**Location**
BERT not pretrained	86.99 (Δ 18.08)	70.58 (Δ 9.29)	91.90 (Δ -0.23)	86.43 (Δ -0.68)
mBERT (Devlin et al., 2019)	**93.57 (Δ 13.98)**	73.64 (Δ 3.63)	93.71 (Δ 2.08)	87.84 (Δ 1.83)
AnchiBERT (Tian et al., 2021)	89.32 (Δ 3.51)	73.57 (Δ 4.27)	94.82 (Δ 1.54)	89.84 (Δ 1.83)
mBERT + AJD/DRS	90.02 (Δ 2.89)	**74.84 (Δ 3.56)**	**94.88 (Δ 2.05)**	**89.88 (Δ 1.98)**
AnchiBERT + AJD/DRS	90.02 (Δ 1.69)	74.54 (Δ 2.47)	94.70 (Δ 1.57)	89.46 (Δ 1.54)

Table 6: F1 scores on topic classification and named entity recognition given document age. The value inside parenthesis indicates the difference on performance after providing document age.

entities. There was a big gap on difference with non-pretrained BERT in TC, which is probably due to the poor performance of itself in the original setting. All models show similar trends on both tasks with improved performances compared to the original settings without document age as input. It is an obvious result considering that the first step for ancient manuscript is assuming the written era. It implies the significance of chronological attribution task in HUE, conveying that chronological attribution task might improve the performance of other HUE tasks.

5.4 Zero-shot Experiment

Countless number of Hanja documents still remain without any analysis and new documents continue to be unearthed. Therefore, we run zero-shot experiments to verify the effectiveness of our language models on extracting information from the historical documents irrelevant to the training corpora. We use DRRI dataset which is not included in both pretraining and fine-tuning data of our Hanja language models and execute CA and NER.

Table 7 shows experimental results with CA and NER on DRRI. All models perform comparably well on the both tasks, regarding that random model will achieve approximately 3.70% performances with 27 classes in CA. Also, all models in CA commonly show high precision which might be due to the monotonous and redundant phrases in the veritable records. It shows similar trends compared to the Table 4, but the gap among models was notably emphasized in the zero-shot settings. This results imply that our CA models might be exploited for the time period prediction of unseen documents in anthology with a reliable level.

Our models outperform others on NER achieving absolutely high performances, though entity maps between AJD and DRRI do not match strictly. Interestingly, all models tend to predict location entities better than person entities which is the opposite result compared to the original NER on AJD. It is probably due to the characteristics of each entity, where location entities are all commonly used in nationwide while person entity might differ by situation. Further analysis on person and location entities in the view of time changes is described in Section 6.2. We present that our models trained on the corpora of the Joseon dynasty provide reliable results on unseen records, implying that our model can be exploited for the low-resourced documents.

6 Further Analysis

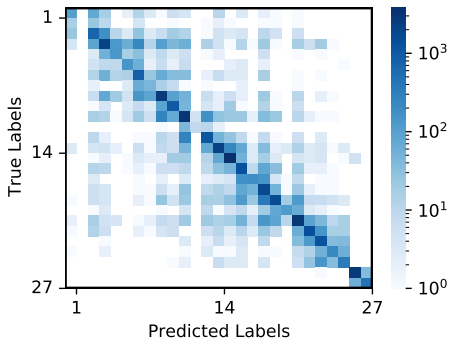
6.1 Do Historical Events Affect Language Models?

To figure out the effect of historical events on models' prediction, we analyzed the output of language models on CA. Figure 2 (a) shows a log-scale confusion matrix of AnchiBERT continued pretraining on AJD and DRS, and Figure 2 (b) indicates the mean absolute error between the predicted king order and the ground truth per each King. The x-axis in the Figure 2 (b) means the changes of time by the king reign period. Each bar in Figure 2 (b) indicates the mean absolute error between the order of ground truth king and the one of predicted, and the line graph means the number of samples in the test set.

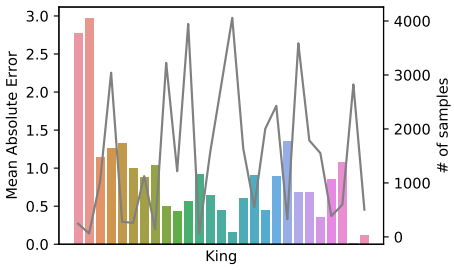
The results of the last two Emperors, Gojong and Sunjong, are remarkable in that the model rarely gets confused with those two labels to others and tends not to fail, showing notable difference on Figure 2 (a) and significantly low mean absolute error on Figure 2 (b). We believe that this is because our model learns the difference between those two records and the others in the historical view and get cues to distinguish them. The last two records are not treated as AJD in general, since those records were inspected and produced by the

	CA	NER	
		Person	Location
BERT not pretrained	26.94	68.50	39.86
mBERT (Devlin et al., 2019)	32.19	72.08	**83.36**
AnchiBERT (Tian et al., 2021)	30.65	71.85	77.23
mBERT + AJD/DRS	35.28	**88.48**	72.08
AnchiBERT + AJD/DRS	**35.85**	82.86	76.53

Table 7: F1 scores of chronological attribution and named entity recognition task on DRRI in zero-shot settings



(a) Log-scale Confusion Matrix



(b) Mean Absolute Error per King

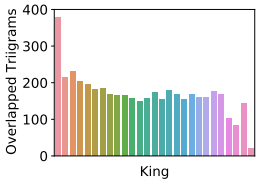
Figure 2: Chronological attribution results with AnchiBERT + AJD/DRS. Figure (a) shows the confusion matrix. In Figure (b), each bar indicates the mean absolute error and the line indicates the number of samples for each king.

Office of Governer-General during the Japanese colonial period with the view of Empire of Japan who ruled Korea [6].

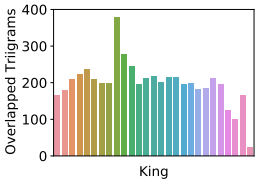
The mean absolute error of the predicted order of each king achieved the difference around one, except for the first King Taejo and the second King Jeongjong, whose errors are almost doubled. We hypothesize that this is mainly because there are too small number of examples in those classes. A similar tendency where the more samples are, the less mean absolute error be has been observed in other classes. Also, the writing style of AJD had settled down from the third King Taejong, and those noisy records might confuse models to predict the exact dates.

6.2 Do Time Changes Affect Written Texts?

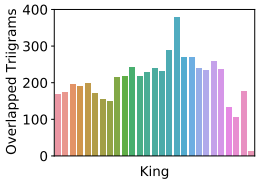
It stands to reason that AJD written in five centuries reveal the features of the language changes. In this section, we investigate the hypothesis above with respect to the named entities and n-grams.



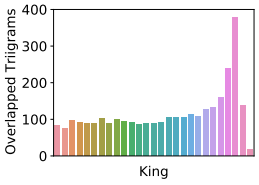
(a) Taejo (1st King)



(b) Seongjong (9th King)



(c) Hyojong (17th King)

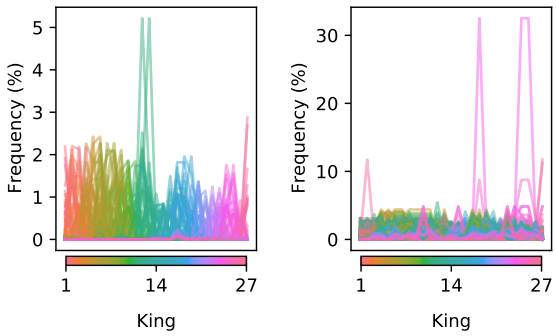


(d) Cheoljong (25th King)

Figure 3: Number of overlapped trigrams per king era. It shows the changes of trigrams over kings whose x-axis shows the kings and the y-axis shows the number of overlapped trigrams.

Words Change over Time We analyze how frequently words change over time. For each king, we plot how many trigrams overlap by each king era in the order. Figure 3 shows overlapped trigrams in the 1st, 9th, 17th, and 25th king and the detailed results with all kings are described in Appendix. It is consistently observed that the closer the king era is, the more trigrams are overlapped. These changes result from not named entity but lexical choices, considering that person and location enti-

[6] http://esillok.history.go.kr/



(a) Person (b) Location

Figure 4: Relative frequency change of top-10 named entities per king. Each line indicates the change of the relative frequency of one entity over time, and the color of the line indicates the king era in which the entity is contained in the top-10 entities. The x-axis represents the time (the kings in the order) and the y-axis represents the relative frequency in each king era.

ties account for 6.38% and 2.05% in the characters of AJD, respectively. It verifies that words used in Joseon dynasty had changed over time gradually, and it enables the language models to capture those features.

Named Entity Changes over Time We investigate how the named entities had been used over time. In particular, we show frequency rates of top-10 frequently-used named entities by each king era and how they change over time in Figure 4. It implies a strong correlation between person entity and the passage of time, while there is no explicit correlation to location entity. Most person entities include officials of the time or the previous kings, relevant to the time. In contrast, most location entities include neighboring countries or place names in the Joseon, which are less dependent on the time. The examples of frequently appeared named entities are described in Appendix.

7 Related Work

ML based NLP techniques have been recently applied to anthology to discover historical documents such as authorship attribution (Ouamour and Sayoud, 2012; Sayoud and Ouamour, 2017; Reisi and Mahboob Farimani, 2020; Hossain et al., 2020), NER (Won et al., 2018; Palladino et al., 2020), and manuscript age detection (Adam et al., 2018). Assael et al. (2022) proposed Ithaca to restore ancient Greek inscriptions and perform geographical attribution and chronological attribution of them.

Along with these works, several works provide language models suited for historical texts in ancient languages and evaluate those models on existing NLU tasks, which aims to support understanding those documents considering that the target languages are mostly extinct. Bamman and Burns (2020) propose Latin BERT for part-of-speech tagging in ancient Latin script. Tian et al. (2021) suggest AnchiBERT and evaluate their model on some NLP tasks including poem topic classification.

However, there has been no research attempting to propose language models in Hanja, which is a dead language in Korea but absolutely necessary to explore Korean history. Most of the studies with Hanja only shed lights on translating historical Hanja documents and use AJD as their corpus (Park et al., 2020; Jin et al., 2020; Kang et al., 2021).

8 Conclusion

We present HUE (**H**anja **U**nderstanding **E**valuation) dataset and BERT-based pretrained language models for classical Hanja documents. HUE dataset includes diverse tasks that can support analyzing historical documents written in Hanja which is an extinct language in Korea: Chronological Attribution (CA), Topic Classification (TC), Named Entity Recognition (NER), and Summary Retrieval (SR). Our models pretrained on Hanja corpora outperform other language models and we observe their performance on zero-shot settings with DRRI which is the dataset never been introduced in NLP community. The experimental results in king prediction imply that our models capture the historical events or facts disclosed in the texts. We also explore several methods to support Hanja language models such as masking named entities and giving document age as input sources, based on the analyses on textual features in AJD.

Help of adequate resources in Hanja documents might could fill some caveats in our work which lacks additional experiments and analyses on the records of different genre such as poetry, novel, and humanities resulting from the low resources that we can exploit. However, we expect that our dataset and accompanying language models might facilitate future works on historical documents written in Hanja by providing fundamental resources to leverage unknown Hanja corpora.

Acknowledgements

We would like to thank Yoonman Heo (Institute for the Translation of Korean Classics) providing expertise on hanja and Korean Classical Chinese. This research was supported by the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921). This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00421, Artificial Intelligence Graduate School Program (Sungkyunkwan University)). Kyunghyun Cho was supported by the NYU Center for Data Science National Science Foundation (Award 1922658) and Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI).

References

Kalthoum Adam, Asim Baig, Somaya Al-Maadeed, Ahmed Bouridane, and Sherine El-Menshawy. 2018. Kertas: dataset for automatic dating of ancient arabic manuscripts. *International journal on Document Analysis and Recognition (IJDAR)*.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

JinYeong Bak and Alice Oh. 2015. Five centuries of monarchy in Korea: Mining the text of the annals of the Joseon dynasty. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*.

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL*.

Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining*.

Anika Samiha Hossain, Nazia Akter, and Md. Saiful Islam. 2020. A stylometric approach for author attribution system using neural network and machine learning classifiers. In *Proceedings of the International Conference on Computing Advancements*.

KyoHoon Jin, JeongA Wi, KyeongPil Kang, and YoungBin Kim. 2020. Korean historical documents analysis with improved dynamic word embedding. *Applied Sciences*.

Kyeongpil Kang, Kyohoon Jin, Soyoung Yang, Soojin Jang, Jaegul Choo, and Youngbin Kim. 2021. Restoring and mining the records of the Joseon dynasty via neural language modeling and machine translation. In *Proceedings of the NAACL: Human Language Technologies*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*.

Siham Ouamour and Halim Sayoud. 2012. authorship attribution of ancient texts written by ten arabic travelers using a smo-svm classifier. In *International Conference on Communications and Information Technology (ICCIT)*.

Chiara Palladino, Farimah Karimi, and Brigitte Mathiak. 2020. Ner on ancient greek with minimal annotation. In *https://dh2020. adho. org/*.

Chanjun Park, Chanhee Lee, Yeongwook Yang, and Heuiseok Lim. 2020. Ancient korean neural machine translation. *IEEE Access*.

Ehsan Reisi and Hassan Mahboob Farimani. 2020. Authorship attribution in historical and literary texts by a deep learning classifier. *journal of Applied Intelligent Systems and Information Sciences*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *The Probabilistic Relevance Framework (PRF)*.

Halim Sayoud and Siham Ouamour. 2017. Score fusion based authorship attribution of ancient arabic texts. In *Florida Artificial Intelligence Research Society Conference*.

Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. Anchibert: A pre-trained model for ancient chinese language understanding and generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*.

Margarida Vale de Gato. 2015. The collaborative anthology in the literary translation course. *The Interpreter and Translator Trainer*.

Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. 2018. Ensemble named entity recognition (ner): Evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*.

William J Youden. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Appendix

A Model

Table 8 shows hyperparameter settings of our models. We used Intel(R) Xeon(R) Silver 4114 (40 CPUs) and GeForce RTX 2080 Ti 10GB (4 GPUs) for all experiments including training, fine-tuning, and inference.

Hyperparameter	Value
Batch Size	32
Early Stopping Patience	3
Hidden Size	768
Learning Rate	2e-5
Learning Rate Scheduler	Linear
Max Sequence Length	512
Number of Hidden Layers	12
Optimizer	AdamW
Vocab Size	11270

Table 8: Model configuration

B HUE Dataset

B.1 Dataset Size

	Train	Dev	Test
CA	330,469	41,309	41,309
TC	330,424	41,303	41,304
NER	385,915	13,417	13,418
SR	169,840	21,570	21,296

Table 9: Data split in HUE dataset

B.2 Source Corpora

B.2.1 Data Collection Process

We crawl AJD [7], DRS [8], and DRRI [9] from the comprehensive database for Korean classics which are publicly available published by IKTC. All source corpora are fully tagged with the written ages and named entities, while their entity maps differ to each other. AJD also provides topics which is tagged by the experts in the translation process.

B.2.2 Dataset Preprocessing

Table 10 shows good and bad example in DRRI to use as summary retrieval dataset. Bad examples mostly written from the 21st King Yeongjo to early in the 22nd King Jungjo describe daily lifes of the crown prince who is King Jeongjo. The bad example in Table 10 is depicting his study. These examples tend to present extremely short *mok*s which cannot be treated as summary and content, while the offical records on administrative has much longer *mok*s.

C Discussion

C.1 Trigram Changes Over Time

Figure 5 shows the changes of trigrams over all kings. It clearly delivers the changes of trigrams as time goes by. We can observe the same trend in either unigram or bigram.

C.2 Top-5 Named Entities by Kings

Table 11 shows top-5 person and location named entities in three King reigns. All person entities except King Sejong, the most frequent entity in Munjong, are officials in the reign periods, which is different by time changes. In contary, all location entities are the name of place, palace or site, showing some entities overlap among kings.

D Experimental Results

For CA, we measure the Quadratic Weighted Kappa score (QWK score) as metrics that treat each king label hierarchically.

Since TC is a multi-label classification task whose example might have multiple labels as the answer, we measure Hamming score along with accuracy. In this case, accuracy is the exact match score, and the hamming score is the accuracy of subset matched, $\mid T \cap P \mid / \mid T \cup P \mid$, where T is set of true labels and P is set of predicted labels (Godbole and Sarawagi, 2004). For the evaluation results in Table 13, we find and set the best threshold to each label by Youden's index. All pretrained models outperform BERT without pretraining, and two LMs re-trained on hanja documents show the best performances.

[7] http://sillok.history.go.kr/main/main.do

[8] http://sjw.history.go.kr/main.do

[9] http://kyudb.snu.ac.kr/series/main.do?item_cd=ILS

	Good Examples	Bad Examples
Date	King Jeongjo 17 (1793) Feb 06	King Yeongjo 50 (1774) May 15
Gang	遞承旨徐榮輔以沈晉賢代之前望也	行召對于尊賢閣。兼弼善洪景顔。說書李駿。翊贊李應重。
Mok	榮輔不仕進政院請牌招以許遞前望單子入之待下批牌招察任	講續綱目
Gang (En)	Royal secretary Yeong-bo Seo was replaced and Jin-hyeon Shim was appointed.	Crown Prince held a royal lecture at the Office of Crown Prince. Kyung-ahn Hong, a lecturer for the Crown Prince, Sangjoon Lee, the second tutor of the Office of Lectures for the Crown Prince, and Eungjoon Lee, a Guard of Crown Prince attended.
Mok (En)	When Yeong-bo Seo did not resign, the king summoned his servants through his royal secretary and ordered him to do so, saying, “wait for appointing the royal secretary among candidates and let him check his job”.	They delivered the lecture on 《Sokgangmok (Comprehensive Mirror for Aid in Government)》.

Table 10: Good and bad examples in DRRI

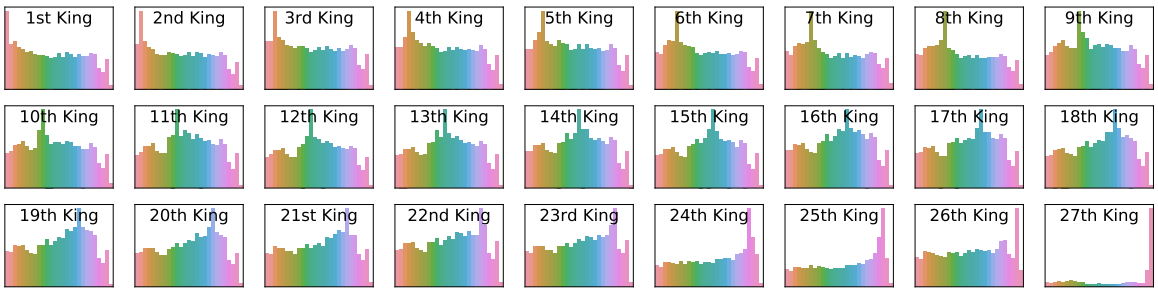


Figure 5: Trigram changes over time

King (Reigning period)	Munjong (5th King) (1450-1452)		Seonjo (14th King) (1567-1608)		Sunjo (23rd King) (1790-1834)	
Person	世宗	2.42%	柳成龍	1.08%	南公轍	1.21%
	金宗瑞	2.26%	李德馨	0.81%	金載瓚	1.09%
	李季甸	1.97%	尹斗壽	0.72%	李時秀	0.85%
	皇甫仁	1.82%	李恒福	0.65%	沈象奎	0.83%
	鄭苯	1.78%	李元翼	0.57%	李相璜	0.77%
Person (En)	King Sejong	(1392-1397)	Seongryong Ryu	(1542-1607)	Gongcheol Nam	(1760-1840)
	Jongseo Kim	(1383-1453)	Deokhyeong Lee	(1561-1613)	Jaechan Kim	(1746-1827)
	Kyejeon Lee	(1404-1459)	Dushu Yun	(1533-1601)	Sisu Lee	(1745-1821)
	Boin Hwang	(1387-1453)	Hangbok Lee	(1556-1618)	Sangkyu Shim	(1766-1838)
	Bun Jeong	(1394-1454)	Weonik Lee	(1547-1634)	Sanghwang Lee	(1763-1841)
Location	平安	4.40%	平壤	2.10%	春塘臺	2.02%
	咸吉	3.24%	朝鮮	1.93%	漢城府	1.19%
	黃海	2.28%	京畿	1.17%	仁政殿	1.71%
	輝德殿	2.21%	全羅	1.76%	平安	1.51%
	京畿	2.10%	慶	1.64%	景慕宮	1.49%
Location (En)	**Pyong-an**	**(Province)**	Pyongyang	(Province)	Chundangdae	(Site)
	Hamgyong	(Province)	Joseon	(Country)	Hanseong Magistracy	(Province)
	Hwanghae	(Province)	**Gyeonggi**	**(Province)**	Injeongjeon	(Hall)
	Hwideokjeon	(Hall)	Jeolla	(Province)	**Pyong-an**	**(Province)**
	Gyeonggi	**(Province)**	Gyeongsang	(Province)	Gyeongmogung	(Palace)

Table 11: Top-5 named entities in 5th, 14th, and 23rd kings

	Acc	F1	Pre	Rec	QWK
BERT not pretrained	56.59	54.26	55.08	56.59	87.09
AnchiBERT (Tian et al., 2021)	76.31	75.74	76.45	76.31	93.99
mBERT (Devlin et al., 2019)	76.02	75.29	75.80	76.02	93.76
AnchiBERT + AJD/DRS	**79.50**	**79.33**	**80.06**	**79.50**	**95.46**
mBERT + AJD/DRS	77.99	77.78	78.92	77.99	95.04

Table 12: Evaluation results for our PLMs on chronological attribution

	Major (4 classes)					**Minor (106 classes)**				
	Acc	F1	Pre	Rec	Ham	Acc	F1	Pre	Rec	Ham
BERT not pretrained	68.52	68.91	74.82	70.65	79.04	26.48	61.29	54.86	84.59	42.45
AnchiBERT (Tian et al., 2021)	68.99	85.81	85.24	88.44	83.00	31.47	69.30	61.63	90.26	51.63
mBERT (Devlin et al., 2019)	56.78	79.59	77.60	87.78	73.64	32.85	70.01	62.17	90.84	54.28
AnchiBERT + AJD/DRS	**70.48**	**88.33**	**86.61**	**92.71**	**84.80**	31.77	**72.07**	**64.87**	91.24	52.25
mBERT + AJD/DRS	69.15	87.13	85.48	91.36	83.81	**33.96**	71.28	63.50	**91.56**	**55.91**

Table 13: Evaluation results for our PLMs on topic classification

	Overall				**Person**			**Location**		
	Acc	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec
BERT not pretrained	98.67	89.40	90.58	88.25	92.13	93.98	90.36	87.10	87.76	86.46
AnchiBERT	98.72	90.30	90.98	89.62	93.13	94.47	91.83	87.91	88.08	87.75
mBERT	98.52	88.57	89.52	87.64	91.63	93.60	89.74	86.02	86.18	85.85
AnchiBERT + AJD/DRS	**98.76**	**90.42**	**91.31**	89.55	**93.28**	**94.77**	91.84	**88.01**	**88.43**	87.59
mBERT + AJD/DRS	98.69	90.16	90.36	**89.95**	92.83	93.51	**92.17**	87.90	87.74	**88.06**

Table 14: Evaluation results for our PLMs on named entity recognition

	MRR	Top-1	Top-10
BM25	29.87	25.58	33.98
BERT not pretrained	52.85	99.20	99.64
AnchiBERT (Tian et al., 2021)	67.92	99.20	99.85
mBERT (Devlin et al., 2019)	67.06	99.32	99.50
AnchiBERT + AJD/DRS	**74.29**	**99.64**	**99.91**
mBERT + AJD/DRS	73.88	99.44	99.59

Table 15: Evaluation results for our PLMs on summary retrieval