applied sciences

MDPI

Article

Korean Historical Documents Analysis with Improved Dynamic Word Embedding

KyoHoon Jin [1], JeongA Wi [1], KyeongPil Kang [2] and YoungBin Kim [1,*]

[1] Department of Image Science and Arts, Chung-Ang University, Dongjak, Seoul 06974, Korea; fhzh123@cau.ac.kr (K.J.); placeja@cau.ac.kr (J.W.)
[2] Scatterlab, Seongdong-gu, Seoul 04766, Korea; kyeongpil@scatterlab.co.kr
* Correspondence: ybkim85@cau.ac.kr

Received: 29 September 2020; Accepted: 6 November 2020; Published: 9 November 2020

check for updates

Abstract: Historical documents refer to records or books that provide textual information about the thoughts and consciousness of past civilisations, and therefore, they have historical significance. These documents are used as key sources for historical studies as they provide information over several historical periods. Many studies have analysed various historical documents using deep learning; however, studies that employ changes in information over time are lacking. In this study, we propose a deep-learning approach using improved dynamic word embedding to determine the characteristics of 27 kings mentioned in the Annals of the Joseon Dynasty, which contains a record of 500 years. The characteristics of words for each king were quantitated based on dynamic word embedding; further, this information was applied to named entity recognition and neural machine translation.In experiments, we confirmed that the method we proposed showed better performance than other methods. In the named entity recognition task, the F1-score was 0.68; in the neural machine translation task, the BLEU4 score was 0.34. We demonstrated that this approach can be used to extract information about diplomatic relationships with neighbouring countries and the economic conditions of the Joseon Dynasty.

Keywords: historical documents; deep-learning; dynamic word embedding; named entity recognition; neural machine translation; transformer

1. Introduction

Historical documents—besides being old texts—carry considerable information, including observations of ideology and phenomena; this information can be used for reconstructing the past. Most research on such documents is performed via a close reading of a small number of documents [1–4]. These attempts have allowed us to understand the meaning of a large corpus of historical documents and identify their patterns, thereby helping us discover new information or reconfirm known facts [5,6]. Developments in these related technologies have improved the possibility of analysing larger historical documents.

Historical documents generally maintain an account of long-term records; for example, the Journal of the Royal Secretariat contains approximately 300 years of records from 1623 to 1910; similarly, the Ming Shilu provides us with nearly 300 years of records from 1368 to 1644. These historical documents were analysed to determine information related to specific periods or to long periods of time as a longitudinal study. For such analyses, it is necessary to identify the characteristics considering changes over time because the meaning and usage of words can vary over time. For example, the word 'apple' was initially used to refer to a fruit, but it is now frequently used to refer to electronics or other products related to the company 'Apple' or the 'iPhone'. Therefore, knowledge about the changing meaning of words is an important factor for deciphering historical documents written over long periods of time. Research has been performed to understand various languages which have embedded words, such as Portuguese [7] or Lithuanian [8], but the studies are limited in their analysis of the passage of time. Several researchers are focused on studying the changes over time [9,10], however, they are concentrating only on decades, so the work is not suitable for understanding and analysing the changes in word meanings over a long period of time. Therefore, we aim to capture semantic changes over time in historical documents.

We utilise a representative Korean historical record, i.e., the Annals of the Joseon Dynasty (AJD). The AJD—a UNESCO World Record Heritage (http://www.unesco.org/new/en/communication-and-information/memory-of-the-world/register/full-list-of-registered-heritage/registered-heritage-page-8/the-annals-of-the-choson-dynasty/)—is an extensive historical document that contains a considerable amount of information related to politics, economy, culture, society, and weather during the Joseon Dynasty, and it has information that spans over 500 years, ruled by 27 kings. Th AJD comprises 50 million characters in the 1893 volumes of the 888 books; it is a detailed and comprehensive historical document. The original text has now been digitised; in addition, a version translated by experts is available.

The AJD has been used in research in the fields of politics, culture, metrology, and medicine [11–17]. However, these analyses were performed over the entire period of 500 years, thereby making it difficult to differentiate between the specific characteristics or ideologies of each king. The AJD covers the longest continual period of a single dynasty compared to other historical documents. We used improved dynamic word embedding (DWE) to capture the semantic changes in the historical document. The semantic changes were quantified based on the embedding vector obtained from DWE. This information was then used to improve the performance of named entity recognition (NER) and neural machine translation (NMT) of historical documents. The entire process is illustrated in Figure 1.

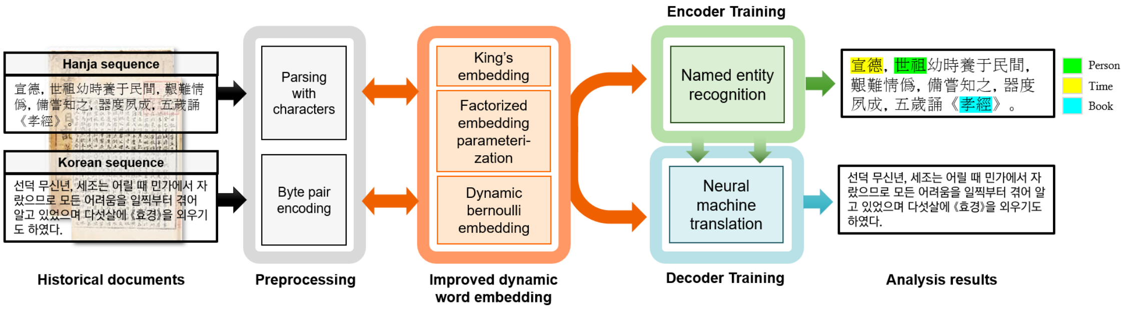

Hanja sequence
宣德，世祖幼時養于民間，艱難情僞，備嘗知之，器度夙成，五歲誦《孝經》。
Korean sequence
선덕 무신년, 세조는 어릴 때 민가에서 자랐으므로 모든 어려움을 일찍부터 겪어 알고 있었으며 다섯살에 《효경》을 외우기도 하였다.
Historical documents
Parsing with characters
Byte pair encoding
Preprocessing
King's embedding
Factorized embedding parameterization
Dynamic bernoulli embedding
Improved dynamic word embedding
Encoder Training
Named entity recognition
Neural machine translation
Decoder Training
宣德，世祖幼時養于民間，艱難情僞，備嘗知之，器度夙成，五歲誦《孝經》。
Person
Time
Book
선덕 무신년, 세조는 어릴 때 민가에서 자랐으므로 모든 어려움을 일찍부터 겪어 알고 있었으며 다섯살에 《효경》을 외우기도 하였다.
Analysis results

Figure 1. Overview of the proposed workflow of dynamic word embedding (DWE), named entity recognition (NER), neural machine translation (NMT) in historical documents.

The contributions of this study are listed below:

- We proposed an improved DWE using factorised embedding parameterization to identify the temporal change in the meaning of words in historical documents. We analysed the words after converting them to dense vectors using the improved DWE; we identified the change in the relationship between countries over time and the taxation structure that varied based on the king. Through this, it was able to better reflect the times than when DWE only was used.
- We confirmed the effectiveness of the improved DWE by incorporating it with tasks such as NER. Through the method we proposed, we were able to improve the F1-score by 3% to 7%. The improved DWE helps us identify the change in object name information or the usage of words for each king; further, the integration of this information and the NER model helped enhance the performance.
- We found that the application of parameters obtained from the NER model integrated with the improved DWE enhanced the effectiveness of historical document translations. Through the method we proposed, we were able to improve the BLEU score by 2% to 8%.

2. Related Work

The analysis of historical documents requires considerable resources because most of them are considerably large. Recent developments in machine learning, text mining and natural language processing methods have also developed rapidly, and these methods are now being used to analyse and understand the meaning of large-scale texts data such as online community [18–21] and social media [22]. Likewise, many studies focusing on historical documents use deep learning or machine learning at various stages, for example during digitisation of historical documents based on deep learning [23–25], automatic extraction of information using methods such as topic modelling [1,26], and during analysis of extracted information [2].

The AJD has been investigated in many studies. The data on the decisions of the king were collected and used to analyse the decision-making pattern of the kings [3,4]; the weather records were used to analyse the weather conditions for that specific period [11,12,17]. Further, the AJD has been used in many other cases, including the analysis of celestial motion and infections [13,16,27]. However, most of these studies perform the analysis by extracting information from a specific period or by applying the same definition to the entire period. Although these methods are effective when analysing the entire document, they have a limitation when identifying variability depending on time.

Many researchers have attempted to incorporate the temporal information in words. Commonly used word embedding methods, such as Word2Vec [28] or GloVe [29] are not effective in using word information over time as they are trained for the entire document. To overcome this problem, DWE techniques have been and are still considered the first distributional semantics, where learning is performed by comparing the frequency of the simultaneous occurrences of words over time and by using an existing word embedding technique [30–32]. However, many of these techniques require a large amount of data for each period, which prevents the application of DWE for analysing historical documents as the amount of accumulated information for each time period in historical documents is irregular [33].

To address this problem, we presented an improved DWE, which is based on dynamic Bernoulli embedding [33]; the performance was improved by incorporating it with factorised embedding parameterization. Thus, the improved DWE allowed us to capture semantic changes. In addition, the improved DWE enhanced the performance of NER and NMT tasks used for analysing historical documents.

3. Methodology

This study used the improved DWE to numerically analyse the semantic changes for each period; then, based on the results, the NER model effectively classified objects such as persons and organisations found in the text. The effectiveness of translating historical documents was further improved by using parameters trained through the NER model in the NMT.

3.1. Dynamic Word Embedding

The word set of the entire document is set as $(x_1, \dots, x_N)$ and the size of the vocabulary is denoted by V. Assuming that the occurrence frequency of each word follows a Bernoulli distribution, data points x_{iv} that have a vocabulary with size V at time point i are defined as $x_{iv} \in 0, 1$. It is assumed that c_i is a set of positions in the neighbourhood of position i and x_{c_i} is a collection of data points indexed by these positions. Further, the embedding vector $\rho_v \in \mathbb{R}^K$ and context vector $\alpha_v \in \mathbb{R}^K$ are assigned for each index (i, v), and the conditional distribution of x_{iv} is given as

$$p(x_{iv}|x_{c_i}) \sim Bern(p_{iv}) \tag{1}$$

where η_{iv} denotes a log odds value assigned through dynamic Bernoulli embedding; the formula is given by

$$\eta_{iv} = logit(\rho^{(t_i)\top}(\sum_{j \in c_j} \sum_{v'} \alpha_{v'} x_{jv'})) \tag{2}$$

A zero-centred Gaussian random walk was used as a prior probability of embedding, similar to that for dynamic Bernoulli embedding [33].

$$\alpha_v, \rho_v^{(0)} \sim \mathcal{N}(0, \lambda_0^{-1} I) \tag{3}$$

$$\rho_v^{(t)} \sim \mathcal{N}(\rho_v^{(t-1)}, \lambda^{-1} I) \tag{4}$$

We adopted the factorised embedding parameterization used in ALBERT [34]. This method has two benefits: (i) The operation amount is reduced from $\mathcal{O}(V \times H)$ in the existing method to $\mathcal{O}(V \times D + D \times H)$ in the new method (here, H denotes the set embedding dimension and D is the dimension less than H). (ii) The comprehension of context is improved as the conversion of words into a dense vector through a single layer increases the training for each word, and the vectors passing through one layer allow comprehending the interaction among words. If factorised embedding parameterization is applied to equation (x),

$$\eta_{iv} = logit(W^{d_{t_i}}(\rho^{(t_i)\top}(\sum_{j \in c_j} \sum_{v'} \alpha_{v'} x_{jv'})) + b^{d_{t_i}}) \tag{5}$$

where $W^{d_{t_i}} \in \mathbb{R}^{d_{emb} \times d_{model}}$ and $b^{d_{t_i}} \in \mathbb{R}^{d_{model}}$.

3.2. Named Entity Recognition & Neural Machine Translation

In addition to inputting each word as an embedding vector, we input additional information into the model. For a more effective reflection of information for each period, information about the 27 kings appearing in the AJD was added. After producing the embedding vectors for each king and making the dimensions of the vectors identical to those of the DWE, we use a bilinear function and transform them into vectors with identical dimensions. This is formulated as

$$\eta_{iv}^{\top} \times A \times K_{emb} \tag{6}$$

where K_{emb} means kings' embedding which $(W^k k_i + b^k)$ where k_i is the information of a king with the ith context, $W^k \in \mathbb{R}^{27 \times d_e}$ and $b^k \in \mathbb{R}^{d_e}$. And $A \in \mathbb{R}^{d_e \times d_e}$ is bilinear parameter and d_e is the pre-designated embedding dimension. Non-linearity is added more effectively when these bilinear functions are used.

We applied the transformer architecture [35] for performing the NER task. Following BERT [36], we performed the NER task using only the encoder part from the transformer architecture. The encoder of the transformer and the fully connected were sequentially subjected to NER.

The decoder of the transformer architecture is used in the NMT. In this process of learning, the bilinear function trained for NER and the values processed by the encoder of the transformer and the Korean embedding vector were used as inputs. This helps determine the effect of performing contextualised word embedding using NER as pre-training. All procedure is visualize in Figure 2.

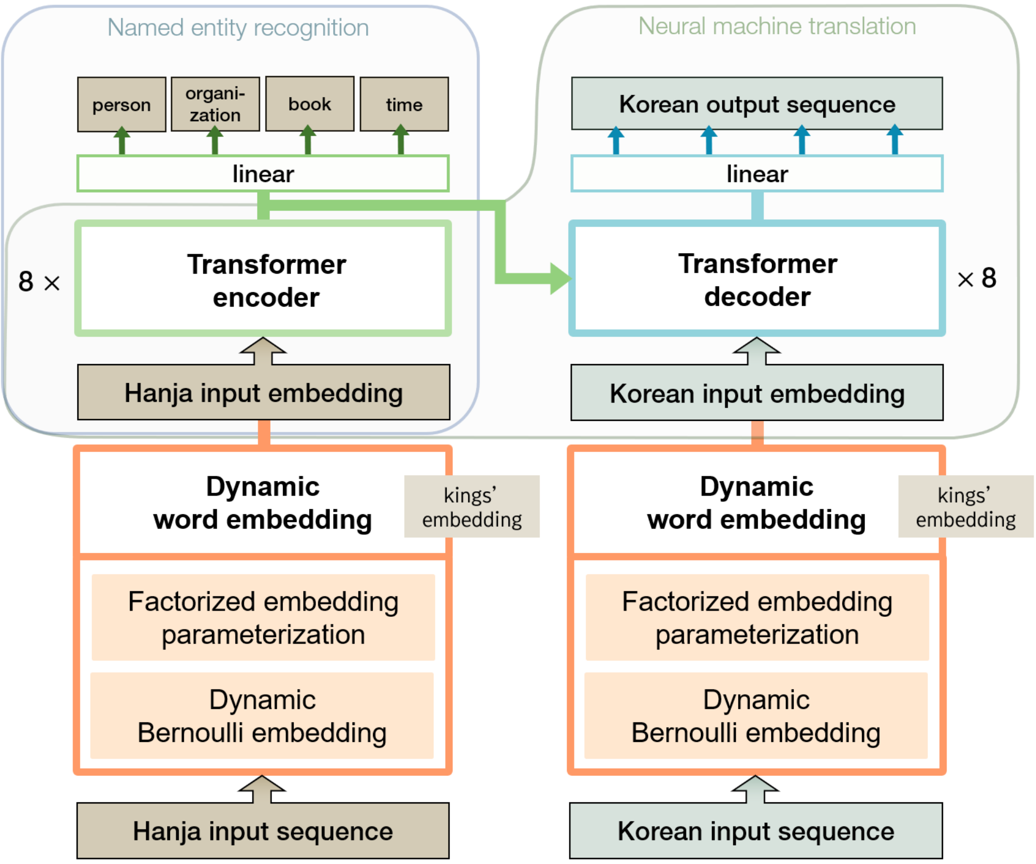


Figure 2. Overview of the proposed model for the NER and NMT tasks.

4. Experiment

4.1. Dataset

To train the deep learning model to analyse historical documents, we used the AJD data. We crawled the original text of the AJD, the object name data labelled by experts, and the data translated by experts. We collected a total of 310,000 paired sentences that included an average of 72 and 149 original (Hanja) characters and translated (Korean) sentences, respectively. Four types of object names were collected for the NER: person, organization, book, and time. The Hanja data were parsed with the character as a unit, similar to that in previous studies [37,38]. In addition, we tokenise each Korean sentence based on byte pair encoding (BPE) [39] provided by Google's SentencePiece [40] library (https://github.com/google/sentencepiece). For stable training, sentences with lengths under 300 characters were used for learning both Hanja and Korean, which accounted for 95% of the entire data. These preprocessed data were divided into training, validation, and test data with a ratio of 8:1:1 before they were used. We conducted validation and the test data consists of 30,000 paired sentences each. In experiments, we used hyperparameters for test data that showed the best performance in validation.

4.2. Experimental Setup

We used four RTX-2080 graphics processing units to train the models. We used eight layers in the encoder of the transformer for the NER task, and another eight layers in the decoder of the transformer for the NMT task. Further, we used the AdamW optimizer for optimization [41]. The learning rate was initially set to 5×10^{-5} and 1×10^{-6} for the NER and NMT tasks, respectively. We used the Warmup learning rate scheduler [42] such that the warmup step is 12000 iterations and then using a linear decay scheduler. In deep neural network learning, multiplication by large weights can lead to an excessive update step, which can cause the algorithm to diverge inappropriately. Therefore, to avoid such a divergence, we used gradient clipping [43] and set the maximum norm to 5. To prevent overfitting, we used dropout [44] at a rate of 0.3. All embedding dimensions were 256, and the other dimensions were 512. For the NMT, we used BPE for sub-word segmentation, and the vocabulary size for the BPE was set to 24,000.

4.3. Analysis of Dynamic Word Embedding

The performance of the proposed improved DWE was tested using quantitative and qualitative tests. For the quantitative test, the loss values were learned using pseudo log likelihood [45], similar to that in a previous study [33]. The loss values decreased in each iteration until it converged to a proper value, as shown in Figure 3a. Separating this process, the pseudo log likelihood of the data x representing the loss values is given by

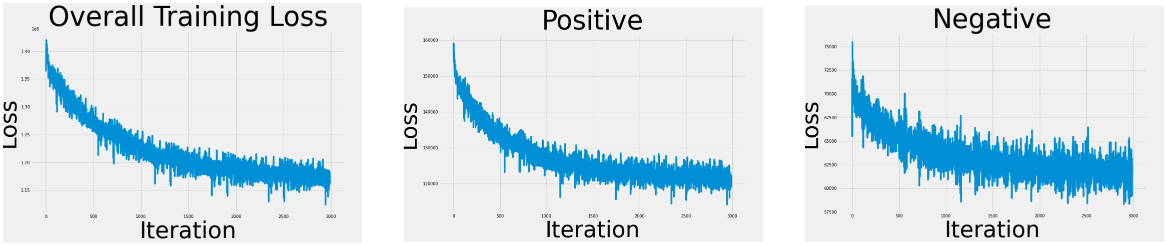
$$\mathcal{L}_{pos} = \sum_{i=1}^{N}\sum_{v=1}^{V} x_{iv} log\sigma(\eta_{iv}) \tag{7}$$

$$\mathcal{L}_{neg} = \sum_{i=1}^{N}\sum_{v=1}^{V} (1 - x_{iv}) log(1 - \sigma(\eta_{iv})) \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function and η_{iv} is same as (5).

$\mathcal{L}_{pos}$ indicates how good the model is at positively predicting the target word from the context, and $\mathcal{L}_{neg}$ indicates how good the model is at negatively predicting the negative samples from the context. In Figure 3b,c, both values were found to converge roughly, although $\mathcal{L}_{neg}$ was more unstable than $\mathcal{L}_{pos}$. This demonstrated that the proposed method performed satisfactory learning in numerical terms.

Qualitative tests were performed to avoid restricting the performance of word embedding to only numerical aspects. We determined the nearest neighbourhood vector of an embedding vector and investigated the change in the 'neighbourhood vector' over time.



(**a**) Overall training loss of proposed improved DWE method.

(**b**) Likelihood of positive samples

(**c**) Likelihood of negative samples

Figure 3. Losses of proposed improved DWE.

The words nearest to the given words for the 1st, 9th, 18th, and 27th kings were compared as summarised in Table 1. For a more effective and convenient analysis, we applied the proposed improved DWE to both Hanja and Korean texts. As mentioned above, the Hanja and Korean texts were segmented based on characters and sub-words, respectively, and the outputs of the two languages were also segmented on the same basis. All distances between words were calculated using Euclidean distance.

For the words relating to Japan, most of them in the earlier Joseon Dynasty were associated with fishery, such as 'fishnet', 'sea salt', and 'salt'. However, it was found that the distances between words such as '虜', which means slave or prisoner of war, decreased after the Imjin War and the Eulsaneukyak. Historically, the Joseon Dynasty was influenced more by the Ming Dynasty than by Japan [46], and toward the later period, the influence of war against Japan was the strongest. Thus, it can be concluded that the performance of the proposed method is excellent.

Through the Meiji restoration, Japan actively accepted Western cultures from other countries such as the England and the USA. Further, the results of the analysis showed that the word 'Japan' became closer to the word 'England' and the word 'USA' in the later Joseon Dynasty. The strongest power in China changed from the Ming Dynasty in the early Joseon Dynasty to the Qing Dynasty during the later Joseon Dynasty. In addition, most words near 'mine' represent place names, which are useful to infer the location of mines for each period. In terms of the 'tax', most taxes in the early Joseon Dynasty were collected for lands, while such taxes were often replaced by labour in the middle period and those collected for fisheries occupied most of them in the later periods.

Table 1. Results of proposed improved DWE method. The words closest to the target words in order by the Joseon Dynasty. Notation * means location and ‡ means job position. 'lang' means language and (o) means original which Hanja or Korean, and (e) means English.

King	Lang	Target Word	
		倭	**일본**
태조 (1st)	(o)	罒, 儵, 猬, 㛐, 嬿	대내전, 유구국, 대마도, 소이전, 귀국
	(e)	fishnet, accident, Wi *, sister-in-law, gorgeous	Daenaejeon ‡, Ryukyu Kingdom *, Tsushima Island *, Soijeon ‡, remigrate
성종 (9th)	(o)	罒, 塩, 猬, 儵, 㛐	유구국, 중국, 대마도, 명, 요동
	(e)	fishnet, salt, Wi *, accident, sister-in-law	Ryukyu Kingdom *, China, Tsushima Island *, Ming Dynasty, Liaodong *
선조 (18th)	(o)	儵, 虜, 煑, 扮, 嗎	중국, 요동, 대마도, 귀국, 명
	(e)	accident, capture, sea salt, grasp, scold	China, Liaodong *, Tsushima Island *, remigrate, Ming Dynasty
숙종 (27th)	(o)	儵, 虜, 券, 煑, 攎	영국, 여진, 총병관, 미국, 몽고
	(e)	accident, capture, weary, cook, oblige	England, Jurchen, admiral, USA, Mongolia
King	**Lang**	**Target Word**	
		중국	**백성**
태조 (1st)	(o)	명, 일본, 본국, 청, 귀국	흉년 ,왜적, 서울, 변장, 오랑캐
	(e)	Ming Dynasty, Japan, home country, Qing Dynasty, remigrate	lean year, Japanese burglar, Seoul *, military attache, barbarian
성종 (9th)	(o)	명, 일본, 청, 귀국, 본국	흉년, 민중, 도민, 기전, 굶주려
	(e)	Ming Dynasty, Japan, Qing Dynasty, remigrate, home country	lean year, people, residents, metropolitan area surrounding, starve
선조 (18th)	(o)	일본, 명, 귀국, 조선, 사신	민중, 가난한, 기전, 굶주려, 도민
	(e)	Japan, Qing Dynasty, remigrate, Joseon, envoy	people, poor, metropolitan area surrounding, starve, residents
숙종 (27th)	(o)	청, 일본, 본국, 몽고, 조선	흉년, 민생, 토병, 소민, 민간
	(e)	Qing Dynasty, Japan, home country, Mongolia, Joseon	lean year, public welfare, native troops, plebeian, civil
King	**Lang**	**Target Word**	
		광산	**세금**
태조 (1st)	(o)	남양, 배천, 단양, 여산, 풍덕	공물, 전결, 부세, 공전, 요역
	(e)	Nam-yang *, Bae-cheon *, Dan-yang *, Yeo-san *, Pung-duck *	tribute, field tax, duty, national land, corvee
성종 (9th)	(o)	남양, 단양, 무안, 이천, 배천	전결, 공물, 부세, 요역, 잡역
	(e)	Nam-yang *, Dan-yang *, Mu-an *, I-cheon *, Bae-cheon *	field tax, tribute, duty, corvee, chores
선조 (18th)	(o)	인천, 배천, 무안, 의령, 용인	전결, 잡역, 소출, 공납, 부세
	(e)	In-cheon *, Bae-cheon *, Mu-an *, Ui-ryeong *, Yong-in *	field tax, chores, crops, local products payment, duty
숙종 (27th)	(o)	인천, 경성, 용인, 평양, 철산	잡역, 잡물, 미곡, 어염, 신역
	(e)	In-cheon *, Kyung-sung *, Yong-in *, Pyongyang *, Cheolsan *	chores, sundries, rice, fishery tax, physical labor

4.4. Results of Named Entity Recognition and Neural Machine Translation

To evaluate the performance of the proposed method, various word embeddings were performed before the NER task; Table 2 summarises these results. In the table, W2V represent the Word2Vec [28] method; DW2V is dynamic word embedding from Yao et al. (2018) [32]; and DBE is dynamic Bernoulli embedding from Rudolph et al. (2017) [33]. The * mark indicates that the method adopted a bilinear function to contain information about the kings.

When word embedding was pre-trained using the skip-gram method in Word2Vec (generally used without DWE), the F1-score was 0.61, thereby suggesting that it achieved the worst performance compared to the others that used DWE. The proposed method that combined DWE with factorised embedding parameterization showed a higher F1-score than other existing word embedding methods. This implies that word embedding has limitations in reflecting temporal information; further, it demonstrated that the addition of temporal information through factorised embedding parameterization and the bilinear function enhances the performance in various tasks.

Further, NMT was compared with GRU [47] and a transformer-based model to test the performance of the proposed method. Learning in GRU also used the seq2seq [48] with attention mechanism [49] method supported by the encoder and the decoder, and the hyperparameters were set to the same values as those in the proposed model. Three metrics (BLEU4 [50], METEOR [51], and ROUGE-L [52]) were used in the test; the results are summarised in Table 3.

Table 2. Results of proposed NER method. Test results was evaluated on parameter set with best validation F1-score. W2V represent the Word2Vec [28] method; DW2V is dynamic word embedding from Yao et al. (2018) [32]; and DBE is dynamic Bernoulli embedding from Rudolph et al. (2017) [33]. The * mark indicates that the method adopted a bilinear function to contain information about the kings.

Method	Precision	Recall	F1-Score
W2V [28]	0.581	0.637	0.607
DW2V [32]	0.627	0.640	0.633
DW2V * [32]	0.582	0.665	0.620
DBE [33]	0.629	0.639	0.633
Ours	0.650	0.646	0.648
Ours *	0.679	0.690	0.684

Table 3. Results of the performance of the translation task. "Transformer (from scratch)" and "Ours" represent the model trained only using the machine translation task and the model trained using NER and bilinear function to encoder and trained decoder only using the NMT task, respectively.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L
RNN Seq2seq (GRU) [47]	0.450228	0.350884	0.281607	0.250351	0.298210	0.428376
Transformer (from scratch) [35]	0.547702	0.438370	0.359203	0.314636	0.372762	0.585471
Ours	0.542257	0.447176	0.379800	0.337939	0.394580	0.656273

Table 3 indicates that the proposed method achieved higher performance compared to the other methods for historical documents. Results shows our model performs better than other NMT task-specifically learned model. It demonstrated that our model enhance NMT performance. Table 4 summarises the results of the translation test for the test dataset; the first sentences were translated accurately, although most of them contain place-names. This is attributed to the fact that the application of the NER improved the translation of texts related to names of places, and therefore, vectors pre-trained through NER are effective for learning the representation of historical documents.

Table 4. Examples of original Hanja sentences, predicted sentences, and translated predicted sentences. For readability, we appended English sentences corresponding to the predicted sentences in each row.

Original	平安道 江界 渭原 寧邊 龜城等地雨雹.
Predicted	평안도의 강계 · 위원 · 영변 · 귀성 등 지방에 우박이 내렸다.
Predicted (Eng.)	Hail fell in the provinces of Gang-gye, Wi-won, Yong-byon, and Gui-seong of Pyeong-an.
Original	太白晝見.
Predicted	태백성이 낮에 나타났다.
Predicted (Eng.)	Venus appeared during the day.
Original	臺諫啓李希雍等事, 不允.
Predicted	대간이 이희옹 등의 일을 아뢰었으나 윤허하지 않았다.
Predicted (Eng.)	Daegan referred to Lee Hee-ong's work, but he was not allowed.
Original	壬辰/詣宗廟景慕宮展拜, 王世子隨詣行禮.
Predicted	종묘와 경모궁에 나아가 전배하였는데, 왕세자가 따라가 예를 거행하였다.
Predicted (Eng.)	King went to Jongmyo and Gyeongmogung and worshiped them, and the Prince followed them to celebrate.

4.5. Discussion

Historical documents can derive insight from any country as well as from the country where they are produced. However, the method we proposed has only translated Hanja into Korean and has not yet been used to translate other languages such as English, French or Chinese. Our method adopted the best-performing method of translating Hanja into Korean, but this may not be as effective in other languages where the Poisson distribution or multinomial distribution may be more appropriate than the Bernoulli distribution.

Since historical documents are records written in the past, there is a limit to further data collection. We used dynamic word embedding to reflect the historical background of historical documents, but to use such dynamic word embedding, we must have accurate information on when the document was written and more than a certain number of data at that time.

5. Conclusions & Future Work

This paper proposed an improved DWE technique to quantitate semantic changes in historical documents; the performance of the proposed technique was evaluated via application to the AJD. The proposed technique revealed the semantic changes, and it was demonstrated that such information can be used for NER and NMT tasks, which facilitated an enhancement in performance for various tasks related to historical documents.

The NER achieved an F1-score of 0.68 via a combination of the improved DWE and information about the king using a bilinear function. The NMT achieved a BLEU4 score higher than that of previous models (by 0.02) by adding the information of the NER obtained from operation based on improved DWE.

The proposed method can be applied to other historical documents such as the Journal of the Royal Secretariat, for which the translation remains incomplete owing to the vast amount of information it contains (over four times that of AJD), and the Ming Shilu.

In future work, we plan to make multi-lingual models with diverse word distributions. We also plan to conduct further studies on a general model for historical documents that incorporates regularisation techniques such as augmentation [53–57] which will enable the exploration of results based on interactions. The proposed methods are expected to reduce the cost of analysing and comprehending historical documents.

Author Contributions: Conceptualization, K.J., J.W., K.K. and Y.K.; methodology, K.J., K.K. and Y.K.; software, K.J.; validation, K.J.; formal analysis, K.J., J.W., and Y.K.; investigation, J.W.; resources, K.J. and K.K.; data curation, K.J. and K.K.; writing—original draft preparation, K.J. and J.W.; writing—review and editing, Y.K.; visualization, J.W.; supervision, Y.K.; project administration, Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Chung-Ang University Excellent Student Scholarship in 2019 and National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2018R1C1B5046461).

Conflicts of Interest: KyeongPil Kang is an employee of Scatterlab. KyoHoon Jin, JeongA Wi, and YoungBin Kim declare no potential conflict of interest.

References

1. Yang, T.I.; Torget, A.; Mihalcea, R. Topic modeling on historical newspapers. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Portland, OR, USA, 24 June 2011; pp. 96–104.
2. Zhao, H.; Wu, B.; Wang, H.; Shi, C. Sentiment analysis based on transfer learning for Chinese ancient literature. In Proceedings of the 2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014), Shanghai, China, 30 October–1 November 2014; pp. 1–7.
3. Bak, J.; Oh, A. Five centuries of monarchy in Korea: Mining the text of the annals of the Joseon dynasty. In Proceedings of the SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Beijing, China, 26–31 July 2015.
4. Bak, J.; Oh, A. Conversational Decision-Making Model for Predicting the King's Decision in the Annals of the Joseon Dynasty. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018.
5. Storey, G.; Mimno, D. Like Two Pis in a Pod: Author Similarity Across Time in the Ancient Greek Corpus. *J. Cult. Anal.* **2020**, *2371*, 4549.
6. Vellingiriraj, E.; Balamurugan, M.; Balasubramanie, P. Information extraction and text mining of Ancient Vattezhuthu characters in historical documents using image zoning. In Proceedings of the 2016 International Conference on Asian Language Processing (IALP), Tainan, Taiwan, 21–23 November 2016; pp. 37–40.
7. Sousa, T.; Gonçalo Oliveira, H.; Alves, A. Exploring Different Methods for Solving Analogies with Portuguese Word Embeddings. In Proceedings of the 9th Symposium on Languages, Applications and Technologies (SLATE 2020), Barcelos, Portugal, 13–14 July 2020.
8. Kapočiūtė-Dzikienė, J.; Damaševičius, R. Intrinsic evaluation of Lithuanian word embeddings using WordNet. In *Computer Science On-Line Conference*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 394–404.
9. Barzokas, V.; Papagiannopoulou, E.; Tsoumakas, G. Studying the Evolution of Greek Words via Word Embeddings. In Proceedings of the 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, 2–4 September 2020; pp. 118–124.
10. Jiang, Y.; Liu, Z.; Yang, L. The Dynamic Evolution of Common Address Terms in Chinese Based on Word Embedding. In Proceedings of the Workshop on Chinese Lexical Semantics, Chiayi, Taiwan, 26–28 May 2018; pp. 478–485.
11. Yoo, C.; Park, M.; Kim, H.J.; Choi, J.; Sin, J.; Jun, C. Classification and evaluation of the documentary-recorded storm events in the Annals of the Choson Dynasty (1392–1910), Korea. *J. Hydrol.* **2015**, *520*, 387–396. [CrossRef]
12. Hayakawa, H.; Iwahashi, K.; Ebihara, Y.; Tamazawa, H.; Shibata, K.; Knipp, D.J.; Kawamura, A.D.; Hattori, K.; Mase, K.; Nakanishi, I.; et al. Long-lasting Extreme Magnetic Storm Activities in 1770 Found in Historical Documents. *Astrophys. J.* **2017**, *850*, L31. [CrossRef]
13. Lee, K.W.; Yang, H.J.; Park, M.G. Orbital elements of comet C/1490 Y1 and the Quadrantid shower. *Mon. Not. R. Astron. Soc.* **2009**, *400*, 1389–1393. [CrossRef]
14. Jeong, H.Y.; Choi, K.H.; Lee, K.S.; Jo, B.M. Studies on conservation of the beeswax-treated Annals of Joseon Dynasty. *J. Korea Tech. Assoc. Pulp Pap. Ind.* **2012**, *44*, 70–78. [CrossRef]
15. Ki, H.C.; Shin, E.K.; Woo, E.J.; Lee, E.; Hong, J.H.; Shin, D.H. Horse-riding accidents and injuries in historical records of Joseon Dynasty, Korea. *Int. J. Paleopathol.* **2018**, *20*, 20–25. [CrossRef]

16. Kang, D.H.; Ko, D.W.; Gavart, M.; Song, J.M.; Cha, W.S. King Hyojong's diseases and death records-through the Daily Records of Royal Secretariat of Joseon Dynasty Seungjeongwonilgi (承政院日記). *J. Korean Med. Class.* **2014**, *27*, 55–72. [CrossRef]
17. Park, M.; Yoo, C.; Jun, C. Consideration of documentary records in the Annals of the Choson Dynasty for the frequency analysis of rainfall in Seoul, Korea. *Meteorol. Appl.* **2017**, *24*, 31–42. [CrossRef]
18. Kang, K.; Choo, J.; Kim, Y. Whose opinion matters? analyzing relationships between bitcoin prices and user groups in online community. *Soc. Sci. Comput. Rev.* **2020**, *38*, 686–702. [CrossRef]
19. Kim, Y.B.; Kang, K.; Choo, J.; Kang, S.J.; Kim, T.; Im, J.; Kim, J.H.; Kim, C.H. Predicting the currency market in online gaming via lexicon-based analysis on its online forum. *Complexity* **2017**, *2017*, 4152705. [CrossRef]
20. Kim, Y.B.; Lee, J.; Park, N.; Choo, J.; Kim, J.H.; Kim, C.H. When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. *PLoS ONE* **2017**, *12*, e0177630. [CrossRef]
21. Christensen, K.; Nørskov, S.; Frederiksen, L.; Scholderer, J. In search of new product ideas: Identifying ideas in online communities by machine learning and text mining. *Creat. Innov. Manag.* **2017**, *26*, 17–30. [CrossRef]
22. Chen, W.F.; Ku, L.W. Utcnn: A deep learning model of stance classificationon on social media text. *arXiv* **2016**, arXiv:1611.03599.
23. Poncelas, A.; Aboomar, M.; Buts, J.; Hadley, J.; Way, A. A Tool for Facilitating OCR Postediting in Historical Documents. *arXiv* **2020**, arXiv:2004.11471.
24. Can, Y.S.; Kabadayı, M.E. Automatic CNN-Based Arabic Numeral Spotting and Handwritten Digit Recognition by Using Deep Transfer Learning in Ottoman Population Registers. *Appl. Sci.* **2020**, *10*, 5430. [CrossRef]
25. Chen, K.; Seuret, M.; Liwicki, M.; Hennebert, J.; Ingold, R. Page segmentation of historical document images with convolutional autoencoders. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1011–1015.
26. Riddell, A.B. How to read 22,198 journal articles: Studying the history of German studies with topic models. In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*; Boydell & Brewer: London, UK, 2014; pp. 91–114.
27. Jeon, J.; Noh, S.J.; Lee, D.H. Relationship between lightning and solar activity for recorded between CE 1392–1877 in Korea. *J. Atmos. Sol. Terr. Phys.* **2018**, *172*, 63–68. [CrossRef]
28. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.
29. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
30. Hamilton, W.L.; Leskovec, J.; Jurafsky, D. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv* **2016**, arXiv:1605.09096.
31. Bamler, R.; Mandt, S. Dynamic word embeddings. *arXiv* **2017**, arXiv:1702.08359.
32. Yao, Z.; Sun, Y.; Ding, W.; Rao, N.; Xiong, H. Dynamic word embeddings for evolving semantic discovery. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Los Angeles, CA, USA, 5–9 February 2018; pp. 673–681.
33. Rudolph, M.; Blei, D. Dynamic Bernoulli embeddings for language evolution. *arXiv* **2017**, arXiv:1703.08052.
34. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
37. Zhang, M.; Zhang, Y.; Che, W.; Liu, T. Character-level chinese dependency parsing. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 1326–1336.

38. Li, H.; Zhang, Z.; Ju, Y.; Zhao, H. Neural character-level dependency parsing for Chinese. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, Orleans, LA, USA, 2–7 February 2018.
39. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
40. Kudo, T.; Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv* **2018**, arXiv:1808.06226.
41. Loshchilov, I.; Hutter, F. Fixing Weight Decay Regularization in Adam. 2018. Available online: https://www.semanticscholar.org/paper/Fixing-Weight-Decay-Regularization-in-Adam-Loshchilov-Hutter/45dfef0cc1ed96558c1c650432ce39d6a1050b6a#featured-content (accessed on 9 November 2018).
42. Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; He, K. Accurate, large minibatch sgd: Training imagenet in 1 h. *arXiv* **2017**, arXiv:1706.02677.
43. Pascanu, R.; Mikolov, T.; Bengio, Y. On the difficulty of training recurrent neural networks. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; pp. 1310–1318.
44. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
45. Arnold, B.C.; Castillo, E.; Sarabia, J.M. Conditionally specified distributions: An introduction (with comments and a rejoinder by the authors). *Stat. Sci.* **2001**, *16*, 249–274.
46. Jungshin, L. KoreansPerception of the Liaodong Region During the Chosŏn Dynasty: Focus on Sejong sillok chiriji (Geographical Treatise in the Annals of King Sejong) and Tongguk yŏji sŭnglam (Augmented survey of the geography of Korea). *Int. J. Korean Hist.* **2016**, *21*, 47–85.
47. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
48. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
49. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
50. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
51. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, 29 June 2005; pp. 65–72.
52. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, 25–26 July 2004.
53. DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* **2017**, arXiv:1708.04552.
54. Jang, S.; Jin, K.; An, J.; Kim, Y. Regional Patch-Based Feature Interpolation Method for Effective Regularization. *IEEE Access* **2020**, *8*, 33658–33665. [CrossRef]
55. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
56. Guo, H.; Mao, Y.; Zhang, R. Augmenting data with mixup for sentence classification: An empirical study. *arXiv* **2019**, arXiv:1905.08941.
57. Marivate, V.; Sefara, T. Improving short text classification through global augmentation methods. In Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Dublin, Ireland, 25–28 August 2020; pp. 385–399.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.