Priming Ancient Korean Neural Machine Translation

Chanjun Park[1,2]*, Seolhwa Lee[3]*, Jaehyung Seo[1],
Hyeonseok Moon[1], Sugyeong Eo[1], Heuiseok Lim[1]†
[1]Korea University
[2]Upstage
[3]University of Copenhagen
{bcj1210, seojae777, glee889, djtnrud, limhseok}@korea.ac.kr
chanjun.park@upstage.ai
sele@di.ku.dk

Abstract

In recent years, there has been an increasing need for the restoration and translation of historical languages. In this study, we attempt to translate historical records in ancient Korean language based on neural machine translation (NMT). Inspired by priming, a cognitive science theory that two different stimuli influence each other, we propose novel priming ancient-Korean NMT (AKNMT) using bilingual subword embedding initialization with structural property awareness in the ancient documents. Finally, we obtain state-of-the-art results in the AKNMT task. To the best of our knowledge, we confirm the possibility of developing a human-centric model that incorporates the concepts of cognitive science and analyzes the result from the perspective of interference and cognitive dissonance theory for the first time.

Keywords: Ancient-Korean Neural Machine Translation, Priming, Neural Machine Translation

1. Introduction

Any language changes in appearance over time by gaining new meanings of words or through the disappearance of old words. This characteristic is called the history of the language. There are words that have different meanings in the past and present. For example, in the Korean language, the original definition of the word "어리다" was "foolish" but has changed in the present day to "little old." Several studies have been conducted on historical languages that reflect this characteristic (Clanuwat et al., 2019; Assael et al., 2019; Park et al., 2020).

Artificial intelligence (AI) research is being conducted based on a cognitive scientific perspective that imitates the structure of the human brain or the brain function analysis, such as reinforcement learning (Kaelbling et al., 1996), neuroplasticity (Lee et al., 2021), cross-language speech perception (Sirois, 2004; Kuhl, 2000), and priming (Pham et al., 2020). With this trend, research is being conducted to incorporate the cognitive science theory into the neural machine translation (NMT) (Pham et al., 2020).

Based on the examples in previous studies, we reinterpret the ancient-Korean NMT (AKNMT) with priming (Tulving et al., 1982), a representative cognitive science theory. Priming is a process in which exposure to a single stimulus affects the response to subsequent stimuli without conscious guidance or intention. For example, when a person thinks of a category of items, similar items are stimulated by the brain. In this study, we propose priming AKNMT influenced by the fact that two stimuli work effectively when they are in the same pattern. The method we present here is more human-centric than previous methodologies because it reflects information processing in the human brain in the NMT task.

To enable learning of the ancient language and Korean language in the same pattern during NMT embedding, we apply bilingual word embedding (BWE) to improve the performance of the existing AKNMTs. BWE ensures that words that have similar meanings in two different languages are mapped in similar spaces by embedding words from the two languages into a single space, which is consistent with priming.

Furthermore, BWE is learned in terms of restricted subwords rather than in terms of words considering the structural properties of the ancient language, and we use it as the embedding layer of the model.

2. Related Works

Recently, there has been a movement in the research field to restore ancient languages using deep learning. Several studies, such as KuroNET (Clanuwat et al., 2018; Clanuwat et al., 2019) in Japan and Greek epigraphy (Assael et al., 2019) in Greece, have been conducted to restore the respective ancient languages using optical character recognition. Provatorova et al. (2020) attempted to analyze the ancient language based on the named entity recognition and linking techniques. However, very few studies have attempted to translate ancient languages.

Park et al. (2020) is the first group to conduct research on AKNMT. An important factor in ancient language translation is how well an entity is translated, and data from it show that a person's name, place, or institu-

*These authors contributed equally.
†This author is the corresponding author.

tion accounts for most of the sentences. In other words, recording was an important issue in ancient times, and thus, information about the entity is considerably important. Based on the features of these ancient language translations, Park et al. (2020) have proposed the SVER BPE method, which specializes in the ancient-Korean language and enforces restrictions on the task of subword tokenization.

Pham et al. (2020) is the first to apply priming to NMT. They conducted experiments on similar translations based on priming and obtained results that were better than the existing methodologies.

3. Ancient Korean Translation

Ancient Korean translation (AKT) refers to the translation of historical Korean books such as Veritable Records of the Joseon Dynasty [1] and The Daily Records of Royal Secretariat of Joseon Dynasty[2]. At present, Veritable Records of the Joseon Dynasty have been fully translated, and many cultural contents (historical dramas, movies, and webcomics) have been published. There is also an increasing trend of utilizing previous research findings. In the past, ancient literature was mainly used in humanities, but it is now being used beyond the scope of traditional humanities in various disciplines, including social sciences, physics, and art (Lee, 1981; Kim, 1997; Lee, 2003). Despite its usability and ripple effects, AKT has three major limitations.

The first is the time and cost limitation. It will take approximately 80 years to manually translate The Daily Records of Royal Secretariat of Joseon Dynasty by mobilizing all of the professional ancient language translators in South Korea. Thus, translating ancient literature using only human resource incurs nontrivial costs. The next is the threshold of manpower shortages. At present, there are only approximately 200 ancient language translation experts in Korea, and it takes more than 10 years to produce experts. Quality differences can also be one of the limitations. It is not easy to present an accurate translation as the translation results vary depending on each individual. This is due to personal deviations from the relevant knowledge and skills they possess for ancient language translation.

A recent study of AKNMT has been conducted to mitigate these limitations (Park et al., 2020). AKNMT has the advantage of enabling faster translation of ancient literature compared with human translation, thus minimizing the quality deviations and maintaining a consistent translation quality. It is also possible to proceed with draft translations of untranslated documents in a short time. AKNMT will be able to contribute significantly to the translation of other ancient literature works such as the books in Gyujanggak.

4. Proposed Method

In this section, we describe our proposed priming AKNMT method.

4.1. Why Priming?

The process of priming involves the activation of a representation or association in the memory just before another stimulus or task is introduced. For example, exposing someone to the word “red” will evoke a faster response to the word “apple” than it would to an unrelated word such as “banana.”

BWE is a useful approach that is analogous to priming because BWE and priming can infer a target stimulus (i.e., target/Korean) from another stimulus (i.e., source/ancient). Inspired by the human-centric nature of the priming processing, we employ BWE to enhance the model performance and mimic the priming process in our NMT embedding method.

Positive and negative priming Positive priming is a process through which providing stimuli *accelerates processing* to the subsequent presentation of the same stimulus (McLennan et al., 2019). For example, if someone is asked to name a fast-food restaurant starting with the letter M, they will answer “McDonald’s” 50% of the time because M and McDonald’s are closely associated in people’s brains. In contrast, negative priming is an implicit memory effect in which previous exposure to a stimulus *adversely affects* the response to the same stimulus (Mielke and Hume, 2001). Thus, positive and negative priming affect the speed of the priming process, with a faster response achieved due to positive priming and slower response due to negative priming.

4.2. Ancient-Korean Subword Embedding Initialization

In this subsection, we describe the methods used to initialize the training of priming AKNMT using ancient-Korean subword embedding. One of the most broadly utilized approaches for monolingual embeddings, such as fastText (Bojanowski et al., 2017), widens the continuous representation of words using the skip-gram negative sampling method (Mikolov et al., 2013a) to learn the subword information.

Subword embedding models (Bojanowski et al., 2017; Heinzerling and Strube, 2017) have been proposed to solve the problem of failure of the word unit embedding models to capture the internal structure of words. Therefore, subword representations provide better task performance than those trained using only whole words (Chaudhary et al., 2018).

Ancient language structure aware byte pair encoding (BPE) Subword tokenization is a general preprocessing method in NMT, and it represents a word by separating a single word into several subwords because a single word is composed of a combination of several meaningful subwords (Avraham and Goldberg, 2017; Sennrich et al., 2015). It is worthwhile to apply subword

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

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

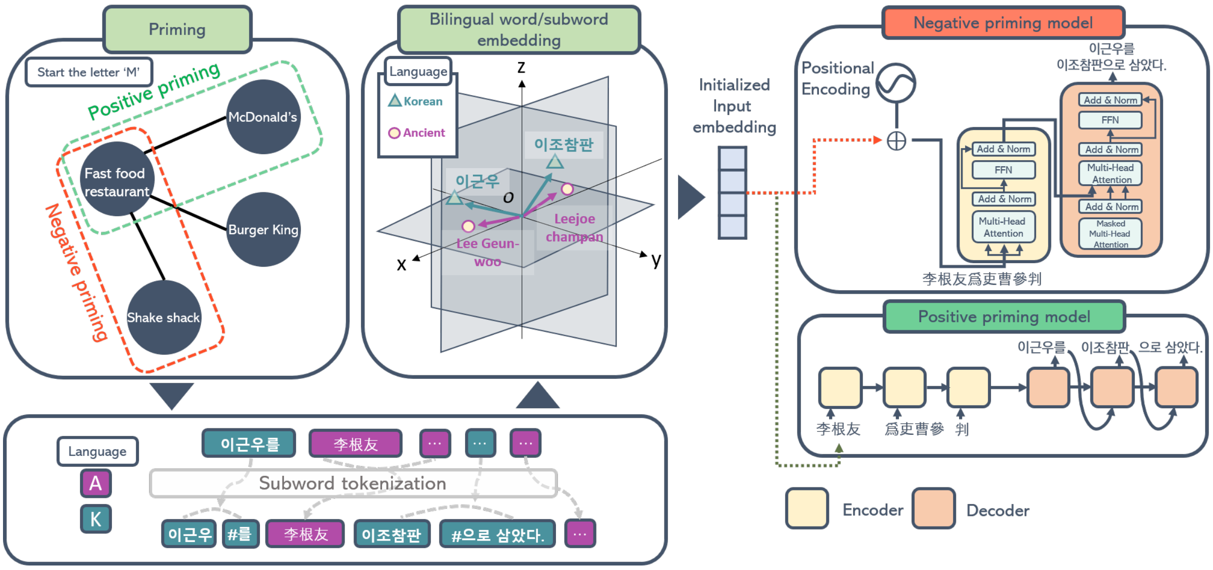

Figure 1: Overall architecture of priming AKNMT. For example, there is fast food restaurant name matching game which consists of two-round, given the correct answers are "Mcdonald's" and "Shake shack" in each round, respectively. It firstly shows the letter M when starting the game at each round. In round 1, the participant response fast answering to "Mcdonald's" due to recall alphabet M. But in round 2, the participant responded (it will say "Shake shack" later.) slower than round 1 because being affected to alphabet M. The former we called positive priming, and the latter is negative priming. We applied this example to a positive priming model for accelerating the processing and negative priming model adversely because of positional encoding.

tokenization to recognize ancient entities because ancient literature includes many entities such as a king's name, location, and the name of the social class.

Therefore, we leverage the method of share vocabulary and entity restriction BPE (SVER BPE) (Park et al., 2020), which is a subword tokenization method specified in the field of ancient language translation for arguing this issue. We shuffled ancient-Korean sentence pairs by each token and then applied its output to the SVER BPE model.

Bilingual subword embedding (BSE) Typically, the NMT model initializes the embedding of the vocabulary for the encoder and the decoder. These embeddings change independently during the NMT training process. The BWE training procedure involves finding the mapping between two languages from a bilingual signal (document alignment). The vector spaces of two languages have a linear relationship (Mikolov et al., 2013b). The fundamental learning process of BWE comprises resolving the following optimization problem, given the first language, X, and the second language, Y. W is a linear map, $W \in \mathbb{R}^{d_X \times d_Y}$, which is learned by solving the optimization problem.

$$\min_{W} \|XW - Y\|_F, \tag{1}$$

where W is the transformation matrix and F is the Frobenius norm. $\mathbb{R}^{d_X}$ indicates the embedding space of the X language. A linear map is used as the training bilingual signal for the NMT model training.

In this study, we employed the fastText library (Bojanowski et al., 2017) to train the BSE. fastText allows us to use a large amount of subword information for initializing the NMT model. Assume that we are given a dictionary of n-grams of size G. We connect a vector representation, z_g, to each n-gram, g, and represent a word as the sum of the vector representations of its n-grams. We calculate a scoring function, S, between a word, w, and a context word, c (i.e., surrounding word to the word w):

$$\mathrm{S}(w, c_{\{A,K\}\in\mathbb{L}}) = \sum_{g\in\mathcal{G}_w} z_g^T v_c, \tag{2}$$

where $\mathcal{G}_w \subset \{1, ..., \mathrm{G}\}$ denotes the set of n-grams present in w, and v_c denotes the context vectors. Given a language, $\mathbb{L}$, c_A and c_K are identified as the ancient context word and the Korean context word, respectively. To summarize, we tokenized the ancient-Korean pairs into subwords using the SVER BPE model, which enables the recognition of the ancient entities efficiently from the ancient language structures, and its output trained BSE using fastText. Consequently, BSE is used as a bilingual signal for initializing the NMT model.

4.3. Priming AKNMT

Hypothesis We conceivably hypothesized that *negative priming* and *positive priming* can be implemented along with the existing computational models. One could infer that *negative priming* would be the transformer (Vaswani et al., 2017), which corrupts the positional information of BWE as a transformer consists of positional encoding (Wang et al., 2019; Ke et al., 2020; Wang et al., 2021; Chen et al., 2021). Hence, this could lead to a decrease in the model performance because positional encoding captures the position of the individual words, not the ordered relationship between single word positions. For *positive priming*, we considered the long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) rather than transformer owing to the remaining positional information available in LSTM.

Overall architecture An ancient-Korean shuffled sentence is given as an input, and it is vectorized using BSE to initialize the NMT model. The model learns the latent expressions of the ancient-Korean sentence in the encoder and translates its output into modern Korean in the decoder. In other words, the NMT model consists of an encoder and a decoder that maximize the conditional log-likelihood as joint training.

We shows the overall architecture of priming AKNMT in Figure 1. We adopted priming process to our proposed method using Ancient language structure aware subword tokenization and BSE/BWE and then applied its output as to initialize input embedding of negative/posotive priming models.

5. Experiments

5.1. Dataset Details

To perform the experiments, we utilized the same training and test data as those used by Park et al. (2020). By leveraging the same data, we performed a fair comparison with the previous study. Additional detailed information on our dataset and specific examples of the training data can be found in Table 1 and 2, respectively.

5.2. Model Details

- **Positive priming model**: For the positive priming model, we used a 2-layer LSTM and Bahdanau attention mechanism (Bahdanau et al., 2014). The hyperparameter, dropout ratio, and mini-batch size were set to 500 hidden units, 0.2, and 128, respectively. We used the Adam optimizer and applied an input feed.

- **Negative priming model**: This is a self-attention-based model developed by Vaswani et al. (2017). A batch size of 4096 and the Adam and Noam decay were used for optimization. Six attention blocks and eight attention heads were used, and the embedding size was set to 512.

We used a vocabulary size of 32,000 words, the cross-entropy loss function for each model. The beam size for decoding was fixed to 5, and all translation results were evaluated using BLEU (Papineni et al., 2002), calculated with the Moses script. [3]

5.3. Experimental Design

We conducted experiments by itemizing the environment into four cases with respect to positive and negative priming: (1) monolingual word embedding (2) monolingual subword embedding (3) BWE, and (4) BSE. Here, (1) and (2) can be regarded as non-priming models, whereas (3) and (4) can be viewed as priming models.

5.4. Experimental Results

Positive versus negative priming model As shown in Table 3, by utilizing BSE in the positive priming model, we achieve the highest model performance. The BLEU score obtained for the corresponding model is 30.45, which is higher than that obtained by Park et al. (2020) by 1.05. This indicates that the architecture of the attention-based LSTM model, which does not contain the structural interference when applying the pretrained embedding, leads to the performance improvement. It is also observed that the application of subword embedding to the positive priming model can enhance its performance, whereas the application of word-based embedding leads to a slight degradation compared to the baseline. This shows that entity-restricted subword segmentation, which reflects the characteristics of AKNMT, plays a significant role in improving the model performance.

In the case of the negative priming model, the application of the pretrained embedding consistently degrades the model performance compared to the model developed by Park et al. (2020). In the human sense, negative priming generally occurs by neglecting the experienced prior stimulation. From this perspective, we can infer that the positional embedding, which is additionally leveraged during the training of the transformer models, affects the model performance. These results support our hypothesis; the LSTM and transformer can be viewed as the positive and negative priming models, respectively.

Interference Theory We can interpret these results using the interference theory (Postman, 1961; Postman and Underwood, 1973). According to this theory, the main causes of the forgetting effect can be subdivided into two possible ways: (i) newly obtained information interfering with the retrieval of the preobtained information and (ii) acquisition of new information. Thus, we can infer that positional embedding, which is added to the pretrained embedding in the transformer, can deteriorate the model performance by hindering the retrieval of the preobtained BWE.

Cognitive Dissonance Theory Cognitive dissonance theory (Brehm and Cohen, 1962; Cooper, 2011) states that human beings tend to change their cognition when they encounter an imbalance between their attitude and behavior, to maintain a balanced state of their cognition. Inspired by this theory, we constructed an ensemble model between positive and negative priming models. In other words, as the pretrained embedding deteriorates

[3] `https://github.com/moses-smt/mosesdecoder`

Information	Ancient-Training	Korean-Training	Ancient-Test	Korean-Test
#Sents	52,778	52,778	3,000	3,000
#Average syllable	39.12	92.78	38.83	91.79
#Max syllable	167	350	141	301
#Min syllable	3	5	4	7

Table 1: Data statistics of training and test set.

Ancient Sentence	賜故上護軍朴淳妻任氏米豆十石 。
Korean sentence	**고**(故) **상호군**(上護軍) **박순**(朴淳)의 처 **임씨**(任氏)에게 쌀•콩 10석을 내려 주었다.
English (Translated)	10 rice and beans were handed down **Mr. Im**, the **wife of deceased Sanghogun Park Soon**
Ancient Sentence	以李根友爲吏曹參判 。
Korean sentence	**이근우**(李根友)를 **이조 참판**으로 삼았다.
English (Translated)	**Lee Geun-woo** was taken as the **leejoe champan**
Ancient Sentence	中批, 以趙秉龜爲戶曹參判 。
Korean sentence	**중비**(中批)로 **조병귀**(趙秉龜)를 **호조 참판**(戶曹參判)으로 삼았다.
English (Translated)	**Cho Byeong-gwi** was taken as the **jungbi** as the **hojo champan**.

Table 2: Example of training data. Most of the sentences are composed of an entity (king's name, person's name, location, name of the social class etc., (blue)), and it can be seen that the Korean sentence is overwhelmingly long.

Model		Pretrained Embedding	BLEU
LSTM	Park et al. (2020)	-	29.40
	None priming model	Mono	28.71 (-0.69)
		Mono Subword	30.00 (+0.60)
	Positive priming model	Bi	29.03 (-0.37)
		Bi Subword	**30.45 (+1.05)**
Trans-former	Park et al. (2020)	-	**29.68**
	None priming model	Mono	28.53 (-1.15)
		Mono Subword	26.01 (-3.67)
	Negative priming model	Bi	28.71 (-0.97)
		Bi Subword	26.36 (-3.32)

Table 3: Experimental results of the comparison between the positive and negative priming models. In the table, Mono denotes monolingual and Bi denotes bilingual. The numbers in the parentheses represent a comparative performance with the model developed by Park et al. (2020).

the performance of the negative priming model but enhances the performance of the positive priming model, we aimed to determine the complementary effect derived by the ensemble method and interpret it using the human sense. For the ensemble model, decoding was performed by using multiple models simultaneously and combining their prediction distributions by averaging (Wu et al., 2016; Park et al., 2021). The results are shown in Table 4.

As observed in the results of the experiment, the ensembled Bi subword model shows an imbalanced state (i.e., PP is "+" and NP is "−") and a better performance than the ensembled Bi model that shows a balanced state (i.e., PP and NP are both "−").

According to the theory of cognitive dissonance, we could infer that the performance is further improved based on the assertion that when an imbalance occurs, it tries to maintain a harmonious state by changing one's cognition to resolve it. In addition, we verified whether the same pattern is present in the non-priming model, and the experimental results are presented in Table 5.

Pretrained embedding	PP	NP	CD	BLEU
Bi	−	−	X	31.50
Bi Subword	+	−	O	33.49

Table 4: Experimental result of cognitive dissonance theory based performance comparison with priming models. PP, NP, and CD indicate positive and negative priming models and cognitive dissonance state, respectively. "−" and "+" indicate the negative and positive degree of improvement, respectively, followed by the parentheses values from Table3. "X" and "O" represent the imbalanced and balanced state, respectively.

Pretrained embedding	LSTM based NP	Transformer based NP	CD	BLEU
Mono	−	−	X	31.11
Mono Subword	+	−	O	31.77

Table 5: Experimental result of cognitive dissonance theory based performance comparison with non-priming models.

To the best of our knowledge, our study has interpreted this from the perspective of the cognitive dissonance theory for the first time.

6. Conclusion

In this study, we reinterpreted the AKNMT task based on the concept of priming and presented the priming AKNMT model, which exhibited state-of-the-art performance. Furthermore, we adopted the method of entity-restricted subword segmentation, which considers the characteristics of AKNMT and improves the model performance. We conducted a quantitative analysis for all the cases of embedding initialization methods and interpreted the results from the perspective of human sense using interference theory and cognitive dissonance theory.

7. Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques).

8. Bibliographical References

Assael, Y., Sommerschield, T., and Prag, J. (2019). Restoring ancient text using deep learning: a case study on greek epigraphy. *arXiv preprint arXiv:1910.06262*.

Avraham, O. and Goldberg, Y. (2017). The interplay of semantics and morphology in word embeddings. *arXiv preprint arXiv:1704.01938*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brehm, J. W. and Cohen, A. R. (1962). Explorations in cognitive dissonance.

Chaudhary, A., Zhou, C., Levin, L., Neubig, G., Mortensen, D. R., and Carbonell, J. G. (2018). Adapting word embeddings to new languages with morphological and phonological subword representations. *arXiv preprint arXiv:1808.09500*.

Chen, P.-C., Tsai, H., Bhojanapalli, S., Chung, H. W., Chang, Y.-W., and Ferng, C.-S. (2021). Demystifying the better performance of position encoding variants for transformer. *arXiv preprint arXiv:2104.08698*.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. (2018). Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*.

Clanuwat, T., Lamb, A., and Kitamoto, A. (2019). Kuronet: Pre-modern japanese kuzushiji character recognition with deep learning. *arXiv preprint arXiv:1910.09433*.

Cooper, J. (2011). Cognitive dissonance theory. *Handbook of theories of social psychology*, 1:377–398.

Heinzerling, B. and Strube, M. (2017). Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. *arXiv preprint arXiv:1710.02187*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Ke, G., He, D., and Liu, T.-Y. (2020). Rethinking the positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*.

Kim, H.-g. (1997). *Understanding Korean Literature*. ME Sharpe.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97(22):11850–11857.

Lee, C., Kim, Y.-B., Ji, H., Lee, Y., Hur, Y., and Lim, H. (2021). On the redundancy in the rank of neural network parameters and its controllability. *Applied Sciences*, 11(2):725.

Lee, P. H. (1981). *Anthology of Korean Literature: From the Earliest Era to the Nineteenth Century*. University of Hawaii Press.

Lee, P. H. (2003). *A history of Korean literature*. Cambridge University Press.

McLennan, K. S., Neumann, E., and Russell, P. N. (2019). Positive and negative priming differences between short-term and long-term identity coding of word-specific attentional priorities. *Attention, Perception, & Psychophysics*, 81(5):1426–1441.

Mielke, S. and Hume, A. (2001). Negative priming as a memory phenomenon: A review of 20 years of negative priming research. *Journal of Psychology*, 215(1):35–5.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Park, C., Lee, C., Yang, Y., and Lim, H. (2020). Ancient korean neural machine translation. *IEEE Access*, 8:116617–116625.

Park, C., Park, S., Lee, S., Whang, T., and Lim, H.-S. (2021). Two heads are better than one? verification of ensemble effect in neural machine translation. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 23–28.

Pham, M. Q., Xu, J., Crego, J. M., Yvon, F., and Senellart, J. (2020). Priming neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 516–527.

Postman, L. and Underwood, B. J. (1973). Critical issues in interference theory. *Memory & Cognition*, 1(1):19–40.

Postman, L. (1961). The present status of interference theory. In *Conference on Verbal Learning and Verbal Behavior, 1959, US*. McGraw-Hill Book Company.

Provatorova, V., Vakulenko, S., Kanoulas, E., Dercksen, K., and van Hulst, J. M. (2020). Named entity recognition and linking on historical newspapers: Uva. ilps & rel at clef hipe 2020. In *CLEF*.

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

Sirois, S. (2004). Autoassociator networks: insights into infant cognition. *Developmental Science*, 7(2):133–140.

Tulving, E., Schacter, D. L., and Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of experimental psychology: learning, memory, and cognition*, 8(4):336.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, B., Zhao, D., Lioma, C., Li, Q., Zhang, P., and Simonsen, J. G. (2019). Encoding word order in complex embeddings. *arXiv preprint arXiv:1912.12333*.

Wang, B., Shang, L., Lioma, C., Jiang, X., Yang, H., Liu, Q., and Simonsen, J. G. (2021). On position embeddings in bert. In *International Conference on Learning Representations*, volume 2, pages 12–13.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.