LREC 2022 Marseille

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

$:Equal Contribution
*: Corresponding Author

Introduction

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

INTRODUCTION

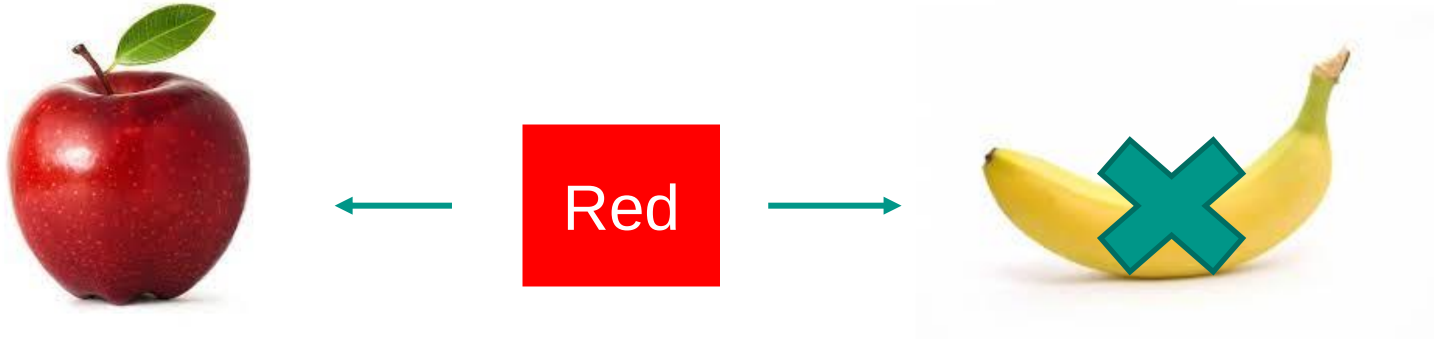
- We attempt to translate historical records in ancient Korean language based on neural machine translation. Inspired by priming, a cognitive science theory that two different stimuli influence each other:
  - We propose novel priming ancient-Korean NMT (AKNMT) using bilingual subword embedding initialization with structural property awareness in the ancient documents

Ancient Korean

Hangul (Korean)

INTRODUCTION

- What is Priming?
  - The process of priming involves the activation of a representation or association in the memory just before another stimulus or task is introduced



INTRODUCTION

- Based on the examples in previous studies (Pham et al., 2020; Lee et al., 2021) in cognitive scientific perspective, we reinterpret the ancient-Korean NMT (AKNMT) with priming a representative cognitive science theory
  - Human-centric approach -> information processing

- We apply bilingual word embedding (BWE) to improve the performance
  - Ancient Korean language – Korean language

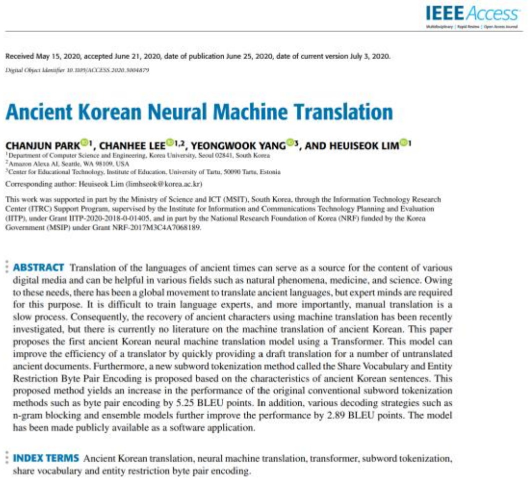
Why AKNMT?

Why AKNMT?

- Ancient Korean translation (AKT) refers to the translation of historical Korean books such as Veritable Records of the Joseon Dynasty

- AKT has three major limitations
  - The first is the time and cost limitation
  - Threshold of manpower shortages
  - Quality differences

- A recent study of AKNMT has been conducted to mitigate these limitations



IEEE Access
Received May 15, 2020, accepted June 21, 2020, date of publication June 25, 2020, date of current version July 3, 2020.
Digital Object Identifier 10.1109/ACCESS.2020.3004879

Ancient Korean Neural Machine Translation

CHANJUN PARK[1], CHANHEE LEE[1,2], YEONGWOOK YANG[3], AND HEUISEOK LIM[1]
[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea
[2]Amazon Alexa AI, Seattle, WA 98109, USA
[3]Center for Educational Technology, Institute of Education, University of Tartu, 50090 Tartu, Estonia
Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program, supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2020-2018-0-01405, and in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIP) under Grant NRF-2017M3C4A7068189.

ABSTRACT Translation of the languages of ancient times can serve as a source for the content of various digital media and can be helpful in various fields such as natural phenomena, medicine, and science. Owing to these needs, there has been a global movement to translate ancient languages, but expert minds are required for this purpose. It is difficult to train language experts, and more importantly, manual translation is a slow process. Consequently, the recovery of ancient characters using machine translation has been recently investigated, but there is currently no literature on the machine translation of ancient Korean. This paper proposes the first ancient Korean neural machine translation model using a Transformer. This model can improve the efficiency of a translator by quickly providing a draft translation for a number of untranslated ancient documents. Furthermore, a new subword tokenization method called the Share Vocabulary and Entity Restriction Byte Pair Encoding is proposed based on the characteristics of ancient Korean sentences. This proposed method yields an increase in the performance of the original conventional subword tokenization methods such as byte pair encoding by 5.25 BLEU points. In addition, various decoding strategies such as n-gram blocking and ensemble models further improve the performance by 2.89 BLEU points. The model has been made publicly available as a software application.

INDEX TERMS Ancient Korean translation, neural machine translation, transformer, subword tokenization, share vocabulary and entity restriction byte pair encoding.

Proposed Method

PROPOSED METHOD

- Priming theory to BWE
  - BWE and priming can infer a target stimulus (i.e., target/Korean) from another stimulus (i.e., source/ancient)
  - Positive priming - providing stimuli accelerates processing to the subsequent presentation of the same stimulus
  - Negative priming - previous exposure to a stimulus adversely affects the response to the same stimulus

PROPOSED METHOD

1) Ancient-Korean Subword Embedding Initialization

- Goal - Initialize the training of priming AKNMT using ancient-Korean subword embedding
- Subword tokenization - Ancient language structure aware byte pair encoding (BPE); tokenization to recognize ancient entities (such as a king's name, location, and the name of the social class.)
- Bilingual word/subword embedding - BWE training to find the mapping between two languages from a bilingual signal

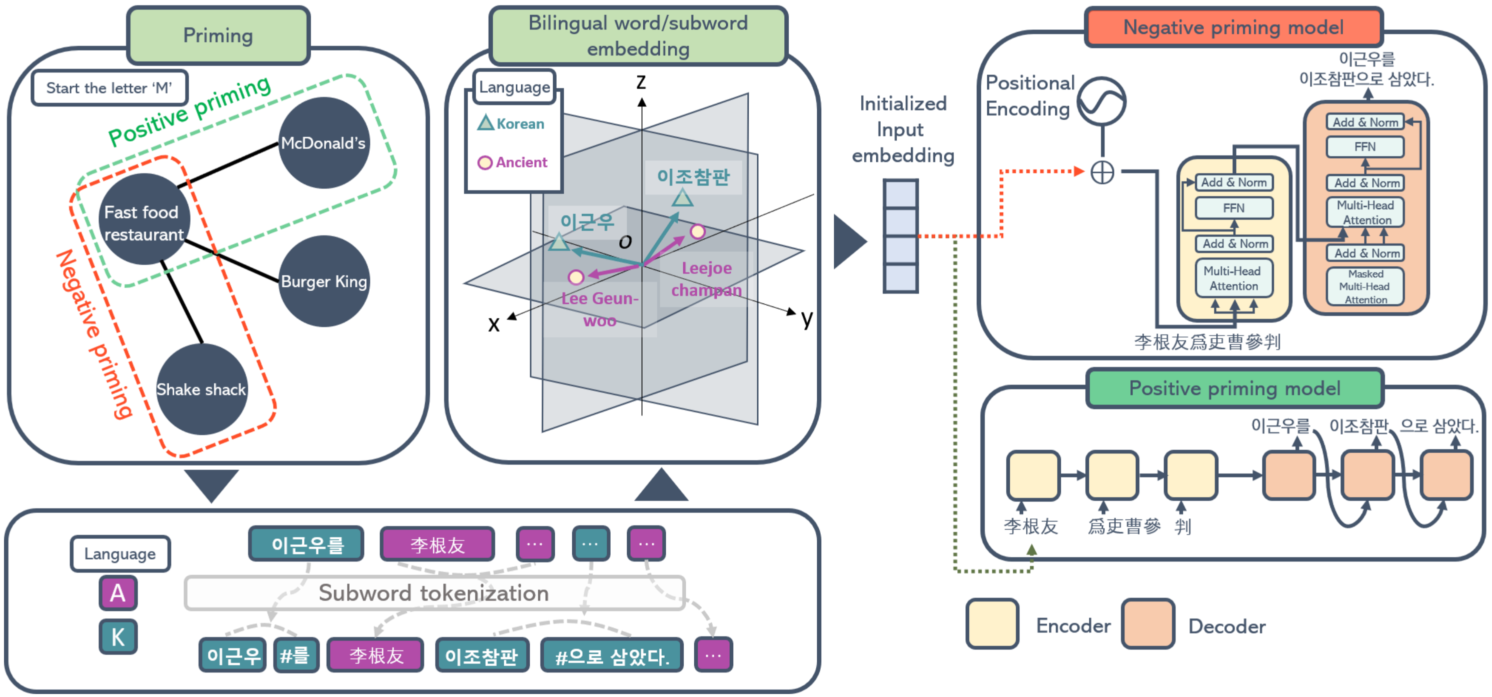
$$\min_{W} \|XW - Y\|_F, \qquad \mathrm{S}(w, c_{\{A,K\}\in\mathbb{L}}) = \sum_{g\in\mathcal{G}_w} z_g^T v_c,$$

PROPOSED METHOD

2) Priming AKNMT

- Hypothesis – negative priming (Transformer) ; positive priming (LSTM)
- Transformer - consists of positional encoding, which lead to a decrease in the model performance
- LSTM - rather than transformer owing to the remaining positional information available in LSTM

PROPOSED METHOD



<Overall architecture>

Experiments

DATASET DETAILS

- We utilized the same training and test data as those used by Park et al. (2020)

IEEE Access
Received May 15, 2020, accepted June 21, 2020, date of publication June 25, 2020, date of current version July 3, 2020.
Digital Object Identifier 10.1109/ACCESS.2020.3004879

Ancient Korean Neural Machine Translation

CHANJUN PARK[1], CHANHEE LEE[1,2], YEONGWOOK YANG[3], AND HEUISEOK LIM[1]
[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea
[2]Amazon Alexa AI, Seattle, WA 98109, USA
[3]Center for Educational Technology, Institute of Education, University of Tartu, 50090 Tartu, Estonia
Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program, supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2020-2018-0-01405, and in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIP) under Grant NRF-2017M3C4A7068189.

ABSTRACT Translation of the languages of ancient times can serve as a source for the content of various digital media and can be helpful in various fields such as natural phenomena, medicine, and science. Owing to these needs, there has been a global movement to translate ancient languages, but expert minds are required for this purpose. It is difficult to train language experts, and more importantly, manual translation is a slow process. Consequently, the recovery of ancient characters using machine translation has been recently investigated, but there is currently no literature on the machine translation of ancient Korean. This paper proposes the first ancient Korean neural machine translation model using a Transformer. This model can improve the efficiency of a translator by quickly providing a draft translation for a number of untranslated ancient documents. Furthermore, a new subword tokenization method called the Share Vocabulary and Entity Restriction Byte Pair Encoding is proposed based on the characteristics of ancient Korean sentences. This proposed method yields an increase in the performance of the original conventional subword tokenization methods such as byte pair encoding by 5.25 BLEU points. In addition, various decoding strategies such as n-gram blocking and ensemble models further improve the performance by 2.89 BLEU points. The model has been made publicly available as a software application.

INDEX TERMS Ancient Korean translation, neural machine translation, transformer, subword tokenization, share vocabulary and entity restriction byte pair encoding.

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

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 참판으로 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.

MODEL DETAILS

- Positive Priming model - two-layer LSTM and Bahdanau attention mechanism (Bahdanau et al., 2014)
- Negative Priming model – Transformer (Vaswani et al. 2017)
- vocabulary size of 32,000 words and 5 beam size
- Evaluation metric - BLEU

EXPERIMENTAL DESIGN

- Four cases

1) Monolingual word embedding
2) Monolingual subword embedding
3) Bilingual Word Embedding (BWE)
4) Bilingual Word Embedding (BSE)

EXPERIMENTAL RESULTS

	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).

- Positive (PP) VS negative priming (NP) model
  - positive priming model, we achieve the **highest** model performance; negative priming model, **degrades** the model performance
- Interference Theory – forgetting effects by (1) newly obtained information interfering 2) acquisition of new information
- Interpretation of results – positional embedding by pretrained embedding can deteriorate the model performance by hindering the retrieval of the pre-obtained BWE

EXPERIMENTAL RESULTS

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 Table 3. "X" and "O" represent the imbalanced and balanced state, respectively.

- Cognitive Dissonance (CD) Theory – 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
- Interpretation of results - The ensembled Bi subword model shows an imbalanced state and a better performance than the ensembled Bi model that shows a balanced state

Conclusion

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

Thank You

http://parkchanjun.github.io/
https://kunmt.org/