Can Artificial Intelligence Translate the Annals of the Joseon Dynasty?

Chanjun Park[1], Sungjin Park[1], Kinam Park[2], Jaechoon Jo[3], Heuiseok Lim[4*]
[1] *Master&Ph.D Combined Student, Department of Computer Science and Engineering Korea University, Seoul*
[2] *Professor, Creative Information and Computer Institute, Korea University, Seoul*
[3] *Professor, Division of Computer Engineering, Hanshin University, Osan*
[4*] *Professor, Department of Computer Science and Engineering Korea University, Seoul*

{bcj1210[1], genom1324[1], spknn[2], limhseok[4] }@korea.ac.kr[1], jaechoon@hs.ac.kr*
Corresponding author: Heuiseok Lim*

Abstract Ancient translation is a task that translates the ancient language of the past into the modern language, requiring both historical and linguistic knowledge costly. This work can be a source of information that will become the actual content of various digital media and can help various fields such as natural phenomena, medicine, and science. A variety of attempts to translate classics have been presented worldwide in different approaches according to this necessity. But the task requires professional help, it is difficult to train professionals and, above all, takes a lot of time to translate. Based on these problems, recent researches to restore ancient characters using machine translation technology have been studied, but there are currently no studies in Korean language. To this end, we propose a Ancient Korean Neural Machine Translation using Transformer. As a result, the BLEU score was 24.43 points. In addition, the model was distributed in the form of a platform[1].

Keywords: *Neural Machine Translation, Ancient Korean Translation, Transformer, Platform, Artificial Intelligence*

1. Introduction

고전번역이란 조선왕조실록, 승전원일기와 같은 고어를 번역하는 것을 의미한다. 기계번역이란 소스문장(Source Sentence)을 타겟문장(Target Sentence)으로 컴퓨터가 번역하는 시스템을 의미하며 이를 고전번역에 적용할 경우 소스문장에 고어 타겟문장에 한국어가 적용될 수 있다.

현재 고전에 대한 사회적 수요가 증가하고 있다. 고전을 번역하기 위한 많은 노력들이 진행되고 있으며 대표적으로 조선왕조실록이 완역되었다. 이로 인해 많은 문화 컨텐츠들이 제작되어 사회 문화적으로 파급효과가 발생하고 있다. 또한 고전번역의 학술적 활용은 분명히 증가하고 있는 추세다. 특히, 번역된 자료들의 DB 화가 진척되면서, 그 활용추세가 증가함이 눈에 띄게 나타나고 있다. 과거에는 주로 인문학분야에서 주로 고전을 활용했다면, 현재는 전통적인 인문학의 범위를 넘어서서 사회과학 분야와 자연학, 기술공학 및 예체능 분야의 학문에서도 활용이 되고 있다.

이러한 필요성에도 불구하고 고전번역에는 많은 한계점이 존재한다. 현재 대한민국에 있는 전문 고전번역가 인력을 총 동원해서 현전하는 승정원일기 기록을 수동으로 완역하려면 80 년 가량이 걸린다고 한다. 그런데 이 계산은 이 모든 인력이 다른 일을 아예 하지 않는다는 가정 하에 산출된 것이니 실제로는 80 년 이상이 걸린다. 또한 현재 고전번역 전문가는 200 여명 수준이며 고전번역자 양성 기간은 관련학과 졸업자기준으로 10 년이상 소요된다. 고전번역을 위한 관련 지식 및 실력에서 개인별 편차가 존재하게 되며 이에 따라 자연스레 번역결과물의 품질 편차가 발생하게 된다. 즉 사람의 힘으로만 고전을 번역하는 것은 엄청난 비용과 시간이 소요된다는 단점이 있다.

이러한 한계점을 극복하기 위하여 본 논문은 Neural Machine Translation(NMT)를 이용하여 문제를 해결하고자 한다. NMT 를 이용하여 고전번역기를 제작하게 되면 기존 고전번역사들의 업무 효율성 강화되며 빠른 시간에 번역이 가능하다. 또한 품질 편차를 최소화하고 일관된 번역 품질을 만들어 낼 수 있다. 더 나아가 규장각 도서 등 아직 미번역 상태의 방대한 고전문헌들의 번역에도 도움이 될 수 있다.

그러나 현재 한국 고전번역 관련하여 연구 발표된 논문이 많지 않은 실정이다. 이에 본 논문은 딥러닝 기반 한국어 고전번역 모델을 제안한다. 한국 고전번역 관련하여 현재 학술적으로 진행된 연구는 많지 않다. 그러나 일본의 KuroNET[1], 그리스의 Greek epigraphy[2]에 대한 연구 등 고어에 대한 연구에 움직임이 시작되고 있다. 따라서 본 논문은 NMT 모델인 Transformer[3]를 이용하여 한국 고전번역 모델을 만들었으며 이를 플랫폼 형태로 배포하였다.

[1] http://nlplab.iptime.org:32242/

2. Korean Ancient Literature

한국고전번역원[2]은 우리 조상들의 정신문화를 담고 있는 한국고전의 수집, 정리, 번역을 통해 한국학 연구의 기반을 구축하고 나아가 전통문화를 계승, 발전시키기 위하여 2007 년 11 월 설립된 교육부 산하 기타공공기관으로 고전문헌을 번역함으로써 한국학 연구의 기반을 구축하고 전통문화를 계승 및 발전시키는데 노력을 다하고 있다. 대게 조선왕조실록, 승정원일기와 같은 역사서에 대한 번역 작업을 지속하고 있으며 대부분 수작업으로 번역을 진행하고 있다.

조선왕조실록은 조선시대 역대 임금들의 실록을 합쳐서 부르는 책 이름이다. 실록은 각 왕 별로 만들어졌기 때문에 조선왕조 임금의 숫자만큼 되는데, 태조부터 철종까지 25 대의 왕의 실록에 선조, 현종, 경종의 수정실록 각 1 건씩이 추가되어 모두 28 종이 된다. 승정원일기는 조선시대 승정원에서 처리한 왕명 출납과 제반 행정사무, 의례적 사항 등을 기록한 일기이다.

이러한 고전 문서들은 그 당시 국정 및 각종 생활들의 가장 자세한 기초 사료이며 기본적인 정책 수립의 자료들이다. 따라서 자료적 가치가 매우 높다. 이러한 자료들을 통해 조선시대 자연현상에 대한 기초자료를 알 수 있으며 현대 천문학 등의 자연과학의 연구에도 활용할 수 있으며 국왕의 진료기록 등은 의학사 연구에도 귀중한 자료를 제공할 수 있다.

따라서 본 논문은 조선왕조실록 등 이미 DB 화가 이뤄진 번역물들을 기반으로 NMT 기반 번역기를 만들어 내었으며 아직 미 번역된 고문서에 대해 활용 가능하도록 플랫폼 형태로 해당 모델을 배포하였다. 이를 통해 과거의 일상적인 삶의 모습과, 당대 생활 자체 복원에 대한 도움이 될 것이다.

3. Ancient Korean Neural Machine Translation Platform

인공신경망 기계번역 기술을 고전 문헌 번역에 활용한 'AI 기반 고전문헌 자동번역시스템'을 구축했다. 대표적인 Sequence to Sequence 모델인 Transformer 기반의 모델을 이용하여 고전번역기를 제작하였다. Transformer 란[3] Convolution 과 Recurrence 없이 오직 Attention 만을 이용하여 기계번역 분야의 획기적으로 성능을 올린 Google 에서 개발한 모델이다. Transformer 는 Query, Key, Value 를 기반으로 입력과 출력에 대해 각각 Self Attention 을 학습한 후 입력과 출력 사이의 Attention 을 학습하는 구조이다. Computational parallelize 가 가능하여 다른 모델보다 training time 이 빠르다는 장점이 존재하며 최근 해당 모델을 이용하여 다양한 NLP 응용 시스템이 개발되고 있다.

플랫폼에 활용된 모델의 Hyper Parameter 의 경우 Vocab 사이즈는 각각 32000 개를 설정했으며 Batch Size 는 4096, Optimization 은 Adam, Noam Decay 를 사용하였고 Loss 함수는 Cross Entropy 를 사용하였다.

데이터셋 같은 경우 한국 고전번역원에서 제공하는 조선왕조실록 고전 DB 정보를 크롤링하였다. 길이가 80 글자 미만인 것은 학습에서 제거하였으며 총 52778 개의 고어-한국어 병렬데이터를 구축하였다. 테스트셋 같은 경우 전체 학습셋에서 5000 개를 랜덤하게 추출하였으며 Validation set 도 5000 개를 랜덤하게 추출하였다. 실험결과 24.43 의 BLEU 점수를 보였다.

4. Conclusion

본 논문은 한국 고전번역기에 대한 연구를 진행하였으며 이를 통해 인간 번역자의 수고를 덜어줄 수 있으며 그만큼 번역에 드는 시간을 단축함과 동시에 번역의 질을 올릴 수 있는 토대를 마련하였다. 추후 고전번역 데이터의 Alignment 작업을 추가적으로 진행한 후 번역의 단위에 따라서 어떠한 성능 변화가 있는지에 대한 연구를 진행할 예정이다.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.NRF-2017M3C4A7068189)

References

[1] Clanuwat, Tarin, Alex Lamb, and Asanobu Kitamoto. "KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning." arXiv preprint arXiv:1910.09433 (2019).

[2] Assael, Yannis, Thea Sommerschield, and Jonathan Prag. "Restoring ancient text using deep learning: a case study on Greek epigraphy." arXiv preprint arXiv:1910.06262 (2019).

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[2] http://www.itkc.or.kr/