Journal of the Korea Convergence Society
Vol. 12. No. 5, pp. 1-7, 2021

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2021.12.5.001

병렬 말뭉치 필터링을 적용한 Filter-mBART기반 기계번역 연구

문현석[1], 박찬준[1], 어수경[1], 박정배[2], 임희석[3*]
[1]고려대학교 컴퓨터학과 석·박사통합과정, [2]고려대학교 Human Inspired AI연구소 교수, [3]고려대학교 컴퓨터학과 교수

Filter-mBART Based Neural Machine Translation Using Parallel Corpus Filtering

Hyeonseok Moon[1], Chanjun Park[1], Sugyeong Eo[1], JeongBae Park[2], Heuiseok Lim[3*]
[1]Master&Ph.D Combined Student, Department of Computer Science and Engineering, Korea University
[2]Research Professor, Department of Human Inspired AI Research, Korea University
[3]Professor, Department of Computer Science and Engineering, Korea University

요 약 최신 기계번역 연구 동향을 살펴보면 대용량의 단일말뭉치를 통해 모델의 사전학습을 거친 후 병렬 말뭉치로 미세조정을 진행한다. 많은 연구에서 사전학습 단계에 이용되는 데이터의 양을 늘리는 추세이나, 기계번역 성능 향상을 위해 반드시 데이터의 양을 늘려야 한다고는 보기 어렵다. 본 연구에서는 병렬 말뭉치 필터링을 활용한 mBART 모델 기반의 실험을 통해, 더 적은 양의 데이터라도 고품질의 데이터라면 더 좋은 기계번역 성능을 낼 수 있음을 보인다. 실험결과 병렬 말뭉치 필터링을 거친 사전학습모델이 그렇지 않은 모델보다 더 좋은 성능을 보였다. 본 실험결과를 통해 데이터의 양보다 데이터의 질을 고려하는 것이 중요함을 보이고, 해당 프로세스를 통해 추후 말뭉치 구축에 있어 하나의 가이드라인으로 활용될 수 있음을 보였다.

주제어 : 딥러닝, 자연어처리, 기계번역, 병렬 말뭉치 필터링, 사전학습 모델

Abstract In the latest trend of machine translation research, the model is pretrained through a large mono lingual corpus and then finetuned with a parallel corpus. Although many studies tend to increase the amount of data used in the pretraining stage, it is hard to say that the amount of data must be increased to improve machine translation performance. In this study, through an experiment based on the mBART model using parallel corpus filtering, we propose that high quality data can yield better machine translation performance, even utilizing smaller amount of data. We propose that it is important to consider the quality of data rather than the amount of data, and it can be used as a guideline for building a training corpus.

Key Words : Deep Learning, Natural Language Process, Machine Translation, Parallel Corpus Filtering, Pretrained model

*This work was supported by Institute for 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) and 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).
*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received February 22, 2021 Revised March 22, 2021
Accepted May 20, 2021 Published May 28, 2021

1. 서론

트랜스포머(Transformer)는 자연어처리의 많은 하위 분야에서 활용되고 있다[1]. 특히 최근 연구에서는 해당 구조를 활용한 사전학습-미세조정 학습 방법론(Pretraining-Finetuning Approach, PFA)이 주로 적용된다. PFA기반의 학습이란, 대용량 말뭉치를 이용하여 모델을 사전학습(Pretraining)한 후 목표로 하는 작업을 위한 말뭉치로 모델을 미세조정(Finetuning)하는 방법론을 의미한다. 대표적으로 BERT[2]에서는 대용량의 단일 언어 말뭉치를 통해 언어 모델을 사전학습하는 과정을 거침으로써 감정분석, 개체명 인식, 형태소분석 등의 세부 작업(Sub-Task) 분야에서 가장 좋은 성능을 얻었다. 특히 BERT에서 제안한 사전학습 방법론인 마스크 모델링(Masked Language Modeling, MLM)과 다음 문장 예측(Next Sentence Prediction, NSP) 방법은 단일 언어 말뭉치를 활용하는 자기 지도 학습(Self Supervised Learning)방법으로, 현재 XLM[3], MASS[4], BART[5]등 여러 자연어처리 모델에 벤치마킹 되어 활용되고 있다.

PFA기반의 학습 방법론은 기계번역에서도 활발하게 적용되고 있다. 특히, 연구가 진행됨에 따라 사전학습과 미세조정에 더 많은 양의 말뭉치가 활용되고 있는데, 이는 기계번역에서 훈련 말뭉치가 성능 향상에 기여하는 바가 매우 크기 때문으로 볼 수 있다[6].

그러나, 단순히 학습에 이용되는 말뭉치의 양을 늘리는 것만이 기계번역 성능 향상을 위한 유일한 해답은 아니다. 말뭉치 내의 저품질 문장(Noisy data)은 품질 저하를 일으키는 매우 큰 요인이라는 연구가 발표되었으며[7], 더 고품질의 훈련 말뭉치를 구축하기 위한 연구가 진행되고 있다. 현재 기계번역에서 가장 영향력 있는 컨퍼런스인 WMT에서도 말뭉치 내의 저품질 문장을 정제하려는 병렬 말뭉치 필터링(Parallel Corpus Filtering) 연구를 공통 과제(Shared Task)로 진행하고 있다[8].

본 연구에서는 훈련 말뭉치의 양보다 질에 집중하고 있는 최신 연구들을 바탕으로 mBART[9] 기반의 기계번역 모델을 설계한다. 사전학습 단계에서 병렬 말뭉치 필터링을 적용한 후 미세조정 단계에서도 이를 활용한 새로운 모델 훈련 방법론인 Filter-mBART를 제안한다.

즉 양질의 말뭉치와 대량의 말뭉치를 통해 각각 mBART 모델을 사전학습한 후 병렬 말뭉치 필터링을 거친 말뭉치와 거치지 않은 말뭉치로 각각 번역을 학습시킴으로써 사전학습 단계와 미세조정 단계에서 모두 양질의 데이터가 미치는 영향을 확인한다. 이를 통해 단순하게 학습데이터의 양을 늘리는 것보다, 데이터의 품질을 고려하면서 학습데이터를 확보하는 것이 기계번역 성능 향상에 더 중요한 요인임을 보인다.

2. 최신 Cross Lingual Language Model 연구

대부분의 자연어처리 세부 작업(Sub-Task)에서는 목표로 하는 작업을 학습하기 위해 레이블(Label)이 부착된 데이터를 활용한다. 하지만 레이블 부착을 위하여 사람의 추가적인 작업이 요구되기 때문에 훈련 말뭉치 구축에 비교적 많은 시간과 비용이 소요된다. 특히 기계번역 분야에서는 학습을 위한 병렬 말뭉치를 생성하기 위해 전문 인력이 요구되기 때문에, 많은 양의 말뭉치를 확보하는 데에 어려움이 발생한다. 이에 따라 웹 크롤링(Crwaling)등을 통해 쉽게 구할 수 있는 언레이블(Unlabeled) 말뭉치를 활용하여 모델을 사전학습 하는 방법이 많이 연구되고 있다. 이러한 방법론은 자기 지도 학습(Self Supervised Learning)이라 불리며, 일반적으로 언레이블 말뭉치 내의 문장 일부를 변형한 후(Noising), 이를 원본 문장으로 복원하는 작업(Denoising)을 학습시킨다.

BERT[2]와 같은 다양한 언어모델(LM)기반 사전학습 모델들이 나왔지만, 대부분 하나의 언어를 기반으로 한 이루어진 연구들이 진행되었다. 최근 이러한 영어 중심으로 편향된 문제를 완화하고자 Cross-Lingual Language Model 연구가 진행되고 있다. 이는 하나 이상의 언어를 기반으로 사전학습 모델을 만드는 방법론을 의미한다. 이러한 방법론은 특히 기계번역에서 많이 활용되고 있다.

대표적으로 하나의 언어에 대한 단일 언어 말뭉치가 아닌, 여러 언어에 대한 단일 언어 말뭉치를 활용하여 MLM을 학습시킨 mBERT[2]가 연구된 바 있다. mBERT는 이를 통해 병렬 말뭉치를 통한 학습 없이도 다중 언어에 대한 이해를 가능하게 하였고, 현재까지 기계번역의 여러 세부 분야에서 활용되고 있다.

이후 XLM[3]에서는 단일 언어 말뭉치뿐 아니라 병렬 말뭉치도 사전학습에 활용하는 새로운 사전학습 방법으로 번역 언어 모델(Translation Language Modeling, TLM)을 제안하였다. TLM에서는 입력을 구성하기 위해

서 소스 문장과 타겟 문장을 하나의 문장으로 연결하고, 이렇게 연결된 문장의 일부를 [MASK]토큰으로 치환한 뒤 원래 문장으로 복원하는 작업을 진행한다. 이렇게 일부가 가려진 문장을 원래 문장으로 복원하는 작업에서 소스 문장과 타겟 문장의 정보를 모두 반영하게 된다. 이를 통해 이종 언어 간의 관계를 더 잘 파악할 수 있다는 점에서, TLM은 다중 언어 학습에 매우 효과적인 방법론으로 알려져 있다.

최근에는 TLM이나 MLM과 같이 문장 내의 단어들을 임의대로 선택하여 [MASK]토큰으로 치환하는 것보다, 연속된 단어들을 치환하는 것이 언어 이해에 있어 더 큰 도움이 된다는 연구가 진행되었다[10]. 기계번역에서는 대표적으로 MASS[4]가 해당 방법론을 적용하였고, 이를 통해 비지도 학습(Unsupervised Learning) 기반 영어-프랑스어 번역에서 가장 좋은 성능을 내었다.

현재 기계번역에서 가장 좋은 성능을 보이는 BART의 경우, 문장 일부를 [MASK]로 치환하는 것 이외로 Token Deletion, Sentence Permutation, Document Rotation, Text Infilling과 같은 여러 Denoising Scheme을 통해 모델을 사전학습하였다.

최근 기계번역 연구는 이렇게 자기 지도 학습 기반의 사전학습에 이용되는 단일 언어 말뭉치의 양을 늘리는 방향으로 이루어지고 있다. 하지만 대부분의 연구에서는 사전학습 말뭉치의 양에 집중할 뿐, 사전학습 말뭉치의 품질에 집중한 연구는 많이 이루어지지 않았다. 이에 본 논문에서는 병렬 말뭉치 필터링을 적용하여 말뭉치의 품질을 높이고, 양질의 말뭉치를 통해 사전학습을 거치는 PFA 기반 모델에 대한 실험을 진행하였다.

3. Filter-mBART

3.1 데이터 구축 및 병렬 말뭉치 필터링

본 연구에서는 병렬 말뭉치 필터링을 통해 정제된 말뭉치를 기반으로 mBART 모델을 훈련하는 Filter-mBART를 제안한다. 기존의 사전학습 모델에서는 대용량의 단일 언어 말뭉치를 활용하여 사전학습을 진행했지만, 학습 말뭉치에 포함된 Noisy data가 기계번역의 품질을 낮춘다는 연구 결과를 고려했을 때[7], 사전학습에 단순히 대용량의 말뭉치를 활용하는 것이 필수적이라고 보기는 어렵다. 이에 본 연구에서는 병렬 말뭉치 필터링을 통해 말뭉치의 품질을 높이고, 이렇게 품질이 검증된 데이터만을 사전학습 단계에 활용함으로써, 양질의 말뭉치를 활용한 PFA기반 기계번역 모델을 설계하였다.

본 실험에서 적용한 말뭉치 필터링은 [11]에서 진행한 방법론과 동일하며, 각 사항은 다음과 같다. 먼저 말뭉치 내의 평균적인 문장 길이를 벗어나 과도하게 길이가 긴 문장을 제거하였다. 이를 위해 500어절 이상의 문장이 포함된 문장 쌍과 1,000음절 이상의 문장이 포함된 문장 쌍을 제거하였다. 그리고 한국어와 영어간의 번역을 원활하게 학습하기 위하여 한국어 단어나 영어 단어가 제대로 포함되어 있지 않은 문장 쌍을 제거하였다. 본 실험에서는 9개 이상의 특수기호가 포함된 문장 쌍, 영어 문장에서 알파벳이 아닌 기호가 50% 이상으로 구성된 문장 쌍, 빈칸, 혹은 탭 기호가 전체 문장의 30%이상을 차지하는 문장 쌍, 그리고 영어 문장과 한국어 문장 위치에 동일한 문장이 들어가 있는 문장 쌍을 제거하였다

마지막으로 Gale&Church 방법론[12] 을 적용하여 추가적인 정제 작업을 진행하였다. Gale&Church 방법론이란, 긴 문장은 긴 문장과 쌍을 이루어야 하는 등, 병

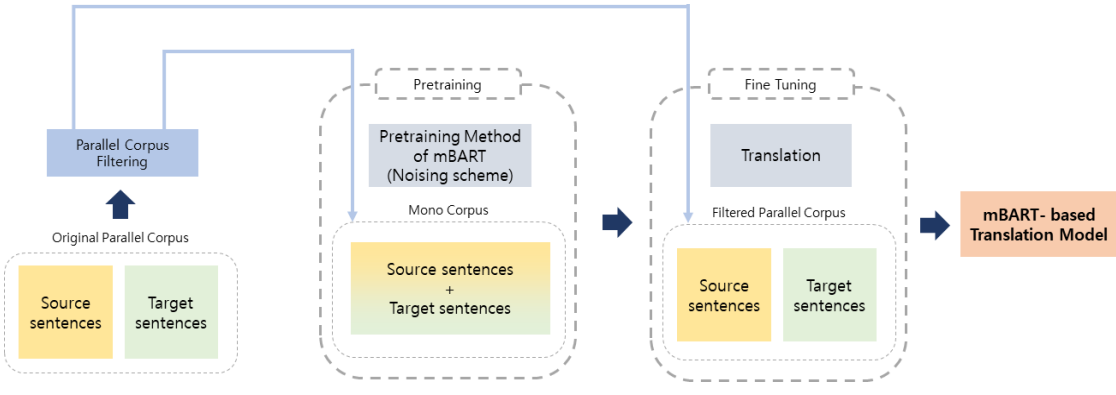

Fig. 1. Overall Training Process

렬 말뭉치 내의 문장 쌍 간의 문장 길이가 유사해야 한다는 이론 아래 두 문장 쌍의 정렬(Align)을 맞추는 작업을 의미한다.

병렬 말뭉치 필터링 과정을 통해 전체적인 말뭉치의 양은 줄어들었지만, 질적으로는 우수한 말뭉치를 확보할 수 있다. Filter-mBART에서는 이 과정을 통해 좋은 품질임이 보증된 말뭉치를 사전학습과 미세조정 모든 단계에 이용한다.

3.2 사전학습 및 미세조정

Filter-mBART는 mBART[9]에서 제안된 사전학습 방법론을 활용하여 모델을 학습한다. mBART에서는 BART[5]와 유사하게 사전학습 단계에서 원본 문장을 고의로 훼손하여 입력 데이터를 만든 후 이를 원본 문장으로 복원하는 작업을 학습한다. 이때, 원본 문장을 훼손하는 방법(Noising Scheme)으로는 연속된 단어를 하나의 [MASK]토큰으로 치환하는 Text Infilling 방법과 입력 내에서 문장의 순서를 바꾸는 Sentence Permutation 방법 두 가지를 활용한다. 본 연구에서는 번역 미세조정에 적용되는 두 언어에 대한 사전학습만을 거친 mBART02[9] 모델과 동일하게, 한국어와 영어 데이터를 통한 사전학습을 진행한다.

mBART에 적용된 사전학습 방법론 중, 특히 Text Infilling은 기계번역과 같은 문장 생성 작업을 미세조정하는 경우 매우 효과적으로 작용한다고 알려져 있으며, 다음과 같은 절차를 통해 진행된다.

가장 먼저 문장을 서브워드(subword)단위로 분절한다. 이후, 분절된 문장에서 연속된 서브워드들을 하나의 [MASK] 토큰으로 치환하는 Text Masking 과정을 거친다. [MASK]로 치환할 연속된 토큰의 개수는 포아송 분포(Poisson Distribution)를 따른다. 포아송 분포는 식 (1)과 같이 표현된다.

$$f(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!} \quad (1)$$

본 실험에서 사전학습 단계에서 가려질 토큰의 개수 n이 선택될 확률은 포아송 분포 함수에 의해 $f(n;3.5)$으로 결정된다. 확률값에 따라 가려질 토큰의 개수가 결정되기 때문에, 가려질 토큰이 없는 경우도 발생하게 된다. 이에 따라 모델은 사전학습 단계에서 [MASK]된 주변 단어들을 통해서 가려진 토큰을 원래 문장의 토큰대로 복원해야 할뿐 아니라, 복원해야 할 토큰의 개수까지 결정하는 작업을 학습하게 된다. 이 과정을 통해 하나의 [MASK]가 반드시 하나의 토큰만을 의미했던 MLM보다 더 깊은 문맥적 이해를 얻을 수 있게 된다. 본 실험에서는 이 과정을 통해 치환되는 토큰의 총 개수를, 입력 문장 총 토큰 개수의 35%가 되도록 설정한다.

Filter-mBART의 전체적인 훈련 프로세스는 Fig. 1.과 같다. 먼저 병렬 말뭉치 필터링으로 학습 말뭉치의 질을 높인 후, 이를 사전학습 단계와 미세조정 단계에 모두 적용하여 모델을 학습시킨다. 사전학습 데이터로 병렬 말뭉치가 활용되기 때문에 Filter-mBART는 한국어와 영어에 대한 이해를 모두 학습하게 된다. 이종언어에 대한 사전학습을 마친 이후 최종적으로 번역 학습을 미세조정시킴으로써 한-영 기계번역에 특화된 모델을 도출하게 된다.

4. 실험 및 실험결과

4.1 데이터 및 실험환경

먼저 모델의 학습을 위하여 웹상에서 병렬 말뭉치를 수집한다. 본 실험에서는 OpenSubtitle[1], AIhub[2], Iwslt 17[13] 한-영 병렬 말뭉치 총 278만 문장 쌍을 수집하였다. 이후 병렬 말뭉치 필터링을 통해 62만 문장쌍을 제거하여 216만쌍의 정제된 병렬 말뭉치를 확보하였다.

모델의 훈련 및 평가는 Facebook AI Research의 Fairseq[14]기반으로 이루어졌다. 모델 구조는 Bart-base[5]와 동일하게 인코더층과 디코더 층의 개수를 6개로 설정하였고, 임베딩과 은닉층의 크기는 768로 설정하였다. 단어사전의 개수는 50,000으로 설정하였다. 서브워드 분리는 센텐스피스 (Sentencepiece)를 이용하였다[15]. 학습은 GeForce GTX 1080Ti 3대로 구성된 환경에서 진행하였고, 성능 평가는 BLEU score[16] 기준으로 삼는다.

모델의 성능 평가를 위하여 네이버 어학사전에 존재하는 예문을 크롤링(crawling)하여 3000개의 한-영 문장쌍으로 구성된 테스트셋을 구성하였다.

4.2 실험 결과

본 실험에서는 필터링을 거친 말뭉치와 거치지 않은 말뭉치를 활용한 번역 모델의 성능을 비교실험 함으로써 Filter-mBART의 성능을 확인한다. 이때, 미세조정에 이용되는 말뭉치뿐 아니라 사전학습에 이용되는 말뭉치에

1) https://opus.nlpl.eu/OpenSubtitles-v2018.php

2) https://aihub.or.kr/

서도 필터링 과정을 적용함으로써, 본 연구에서 제안하는 방법론이 번역에서 어느 정도의 성능 향상을 만들어 낼 수 있는지 정량적으로 분석한다. 실험결과는 Table 1과 같으며, 표에서 Filter란 정제된 말뭉치로 학습한 모델을, No-Filter란 정제되지 않은 원본 말뭉치로 학습한 모델을 의미한다.

Table 1. Experimental Results of Filter-mBART

Model	Pretrain	Finetuning	BLEU
mBART based NMT model	Filter	Filter	**25.62**
		No-Filter	25.49
	No-Filter	Filter	24.50
		No-Filter	24.49

실험결과, 사전학습과 미세조정 단계에서 모두, 병렬 말뭉치 필터링을 거친 Filter-mBART 모델이 가장 좋은 성능을 보였다. 더불어 사전학습 단계(Pretrain-Stage)에서 필터링을 적용한 모델이 그렇지 않은 모델보다 미세조정 단계(Finetuning-Stage)에서의 필터링 적용 여부와 상관없이 모두 좋은 성능을 보였다. 이는 사전학습 단계에서 병렬 말뭉치 필터링을 적용하는 것이 미세조정 단계에서 필터링을 적용한 것보다 더 효과적임을 알 수 있다.

사전학습 단계에서 필터링을 적용하지 않으면서 미세조정에서 필터링을 적용한 모델은 24.50 BLEU score, 적용하지 않은 모델은 24.49 BLEU score를 보였다. 두 모델의 성능 차이가 거의 없는 점은 더욱 사전학습 단계에서 필터링 적용 여부가 중요함을 보여준다.

본 실험결과는 크게 두 가지로 해석할 수 있다. 첫 번째로, 사전학습과 미세조정에 적용한 말뭉치 필터링은 모두 기계번역의 성능을 향상시킨다. 그러나 사전학습에서 필터링을 적용하는 것이 더 큰 성능향상 효과를 볼 수 있다. 기존 연구들은 대부분 미세조정 단계에서만 말뭉치 필터링을 활용하지만 본 논문은 사전학습을 위한 말뭉치에도 필터링을 적용함으로써 추가적인 성능 향상을 얻을 수 있었다. 즉 미세조정 말뭉치의 품질뿐 아니라, 사전학습 말뭉치의 품질도 기계번역에서 매우 중요하게 고려되어야 하는 요인임을 확인할 수 있다.

두 번째로 단순히 말뭉치의 양을 늘리는 것 보다, 학습에 악영향을 끼치는 데이터가 포함되지 않은 양질의 말뭉치를 활용하는 것이 번역 성능 향상에 더 큰 도움이 될 수 있다는 것을 확인할 수 있다. 이는 새로운 모델 구조나 학습 구조의 도입 없이, 학습 말뭉치의 전처리만을 통해서도 기계번역의 성능을 향상시킬 수 있음을 의미한다[17]. 또한, 최적의 번역 성능을 위해서는 데이터 수집 단계에서 데이터의 양만을 고려한 수집이 아닌, 데이터의 품질을 고려한 수집이 이루어져야 한다고 해석할 수도 있다.

본 실험을 통해 적은 양이라도 품질이 좋은 말뭉치를 구축한다면, 이를 통해 충분히 우수한 성능을 보이는 번역 모델을 설계할 수 있음을 볼 수 있다. 이에 따라 말뭉치 구축에 있어 시간과 비용적 측면의 절감 뿐만 아니라 학습속도에서도 이점을 얻을 수 있다. 이에 더해 필터링을 통해 번역 성능이 오른 것으로 보아, 필터링을 통해 정제된 말뭉치들은 번역 성능에 악영향을 끼치는 요인으로 해석될 수 있으며, 이는 추후 말뭉치 구축에 있어 하나의 가이드라인으로 활용될 수 있다.

5. 결론

본 연구에서는 병렬 말뭉치 필터링과 mBART에서 제안한 사전학습방법을 결합한 Filter-mBART를 제안하였다. 말뭉치 필터링을 거친 데이터와 거치지 않은 데이터로 각각 사전학습, 미세조정한 모델들의 성능을 비교 실험함으로써 두 단계에서 모두 필터링을 거친 데이터를 활용한 Filter-mBART가 가장 좋은 성능을 낸다는 것을 보였다. 더 나아가 미세조정 단계뿐 아니라 사전학습 단계에서도 필터링을 거치는 것이 번역 성능 향상에 매우 큰 영향을 끼치는 요인임을 확인하였다.

이를 통해 단순히 말뭉치의 양, 혹은 모델의 크기를 늘리는 것이 번역 성능 향상의 유일한 해답이 아님을 보였다. 이는 좋은 품질의 학습 말뭉치를 확보한다면 비록 말뭉치의 양이 적더라도 번역 성능이 향상될 수 있음을 의미하고, 이를 통해 이후 학습속도나 말뭉치 구축에 요구되는 시간을 줄일 수 있을 것으로 기대한다.

REFERENCES

[1] A. Vaswani et al. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762.*

[2] J. Devlin, M. W. Chang, K. Lee & K. Toutanova. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[3] G. Lample & A. Conneau. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291.*

[4] K. Song, X. Tan, T. Qin, J. Lu & T. Y. Liu. (2019). Mass: Masked sequence to sequence pre-training for

language generation. *arXiv preprint arXiv:1905.02450.*

[5] M. Lewis et al. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461.*

[6] C. Park & H. Lim. (2020). A Study on the Performance Improvement of Machine Translation Using Public Korean-English Parallel Corpus. *Journal of Digital Convergence, 18(6), 271-277.*
DOI : 10.14400/JDC.2020.18.6.271

[7] H. Khayrallah & P. Koehn. (2018). On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282.*
DOI : 10.18653/v1/w18-2709

[8] P. Koehn, V. Chaudhary, A. El-Kishky, N. Goyal, P. J. Chen & F. Guzmán. (2020, November). Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. *In Proceedings of the Fifth Conference on Machine Translation* (pp. 726-742).

[9] Y. Liu et al. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8, 726-742.
DOI : 10.1162/tacl_a_00343

[10] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer & O. Levy. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics, 8, 64-77.*
DOI : 10.1162/tacl_a_00300

[11] C. Park, Y. Lee, C. Lee & H Lim. (2020). "Quality, not Quantity? : Effect of parallel corpus quantity and quality on Neural Machine Translation," *The 32st Annual Conference on Human Cog-nitive Language Technology.*

[12] W. A. Gale & K. Church. (1993). A program for aligning sentences *in bilingual corpora. Computational linguistics,* 19(1), 75-102.

[13] M. Cettolo et al. (2017). Overview of the iwslt 2017 evaluation campaign. *In International Workshop on Spoken Language Translation* (pp. 2-14).

[14] M. Ott et al. (2019). fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038.*
DOI : 10.18653/v1/n19-4009

[15] T. Kudo & J. Richardson. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226.*
DOI : 10.18653/v1/P18-1007

[16] K. Papineni, S. Roukos, T. Ward & W. J. Zhu. (2002, July). Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).*

[17] C. Park, Y. Yang, K. Park & H. Lim. (2020). Decoding strategies for improving low-resource machine translation. *Electronics, 9(10), 1562.*

문 현 석(Hyeonseok Moon) [학생회원]

· 2021년 2월 : 고려대학교 수학과 (이학사)
· 2021년 3월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Neural Machine Translation, Natural Language Processing
· E-Mail : glee889@korea.ac.kr

박 찬 준(Chanjun Park) [학생회원]

· 2019년 2월 : 부산외국어대학교 언어처리창의융합전공 (공학사)
· 2018년 6월 ~ 2019년 7월 : SYSTRAN Research Engineer
· 2019년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Machine Translation, Grammar Error Correction, Deep Learning
· E-Mail : bcj1210@naver.com

어 수 경(Sugyeong Eo) [학생회원]

· 2020년 8월 : 한국외국어대학교 언어인지과학과, 언어와공학전공 (문학사, 언어공학사)
· 2020년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Neural Machine Translation, Quality Estimation, Deep Learning
· E-Mail : djtnrud@korea.ac.kr

박 정 배(Jeongbae Park) [정회원]

· 2002년 2월 : 백석대학교 컴퓨터 학과 (공학사)
· 2014년 8월 : 고려대학교 컴퓨터교육학과 (이학석사)
· 2020년 2월 : 고려대학교 컴퓨터학과 (공학석사)
· 2020년 7월 ~ 현재 : 고려대학교 Human Inspired AI연구소 교수
· 관심분야 : Natural Language Processing, Educational Data Mining, Social Network Analysis
· E-Mail : insmile@korea.ac.kru

임 희 석(Heuiseok Lim) [종신회원]

· 1992년 : 고려대학교 컴퓨터학과(이학사)
· 1994년 : 고려대학교 컴퓨터학과 (이학석사)
· 1997년 : 고려대학교 컴퓨터학과 (이학박사)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 기계학습, 인공지능
· E-Mail : limhseok@korea.ac.kr