Journal of Digital Convergence
Vol. 19. No. 7, pp. 199-208, 2021

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2021.19.7.199

최신 기계번역 사후 교정 연구

문현석[1], 박찬준[1], 어수경[1], 서재형[1], 임희석[2*]
[1]고려대학교 컴퓨터학과 석·박사통합과정, [2]고려대학교 컴퓨터학과 교수

Recent Automatic Post Editing Research

Hyeonseok Moon[1], Chanjun Park[1], Sugyeong Eo[1], Jaehyung Seo[1], Heuiseok Lim[2*]
[1]Master&Ph.D Combined Student, Department of Computer Science and Engineering, Korea University
[2]Professor, Department of Computer Science and Engineering, Korea University

요 약 기계번역 사후교정이란, 기계번역 문장에 포함된 오류를 자동으로 교정하기 위해 제안된 연구 분야이다. 이는 번역 시스템과 관계없이 번역문의 품질을 높이는 오류 교정 모델을 생성하는 목적을 가진 연구로, 훈련을 위해 소스 문장, 번역문, 그리고 이를 사람이 직접 교정한 문장이 활용된다. 특히, 최신 기계번역 사후교정 연구에서는 사후교정 데이터를 통한 학습을 진행하기 이전에, 사전학습된 다국어 언어모델을 활용하는 방법이 적용되고 있다. 이에 본 논문은 최신 연구들에서 활용되고 있는 다국어 사전학습 언어모델들과 함께, 해당 모델을 도입한 각 연구에서의 구체적인 적용 방법을 소개한다. 나아가 이를 기반으로, 번역 모델과 mBART모델을 활용하는 향후 연구 방향을 제안한다.

주제어 : 딥러닝, 자연어처리, 언어 융합, 기계번역, 기계번역 사후교정, 사전학습 모델

Abstract Automatic Post Editing(APE) is the study that automatically correcting errors included in the machine translated sentences. The goal of APE task is to generate error correcting models that improve translation quality, regardless of the translation system. For training these models, source sentence, machine translation, and post edit, which is manually edited by human translator, are utilized. Especially in the recent APE research, multilingual pretrained language models are being adopted, prior to the training by APE data. This study deals with multilingual pretrained language models adopted to the latest APE researches, and the specific application method for each APE study. Furthermore, based on the current research trend, we propose future research directions utilizing translation model or mBART model.

Key Words : Deep Learning, Natural Language Process, Language Convergence, Machine Translation, Automatic Post Editing, Pretrained model,

1. 서론

기계번역이란 병렬 말뭉치(Parallel Corpus)를 이용하여 소스 문장(Source Sentence)을 이에 대응하는 번역문인 타겟 문장(Target Sentence)으로 번역하는 자연어처리의 하위 분야(Sub Task)이다[1]. 기계번역 연구는 모델의 변화, 학습 데이터의 증강을 통한 성능 향상 연구뿐 아니라 학습 데이터의 전후처리 및 번역문의 사후교정 연구도 활발하게 이루어지고 있다[2]. 대표적으로 기계번역 사후교정(Automatic Post Editing, APE), 병렬

*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 May 3, 2021 Revised July 6, 2021
Accepted July 20, 2021 Published July 28, 2021

말뭉치 필터링(Parallel Corpus Filtering), 기계번역 품질 예측(Quality Estimation)분야가 존재하고, 이 분야들은 특히 World Machine Translation (WMT) 컨퍼런스에서 공동 연구 과제(Shared Task)로도 지정되어 현재 활발하게 연구되고 있다[3-5].

이들 중, 특히 APE 연구는 번역 모델을 수정하지 않고도 번역문의 품질을 향상시킬 수 있다는 점에서 큰 이점을 가지고 있다[6]. APE는 번역문에 포함된 오류를 자동으로 교정하기 위해 제안된 연구 분야로, 번역 시스템을 분석하거나 수정하지 않고(Black Box Machine Translation Engine) 이를 통해 생성한 번역문의 오류를 교정하기 위한 연구가 진행된다.

초기 APE 연구는 번역문에 대한 분석만으로 오류를 수정하는 방법을 활용하였으나, 최근에는 번역문뿐 아니라 번역 이전의 소스 문장까지 교정에 활용하는 방법으로 APE 연구를 진행하고 있다. 이에 따라 APE 연구에는 한-영(Source Sentence)과 번역 시스템을 통해 이를 번역한 문장(Machine Translation), 그리고 번역한 문장을 전문가가 직접 교정한 문장(post edit)이 요구된다. APE는 번역 모델과 관계없이 번역문의 품질을 높일 수 있을 뿐 아니라, 도메인 특화 번역문 생성에도 기여할 수 있다는 점에서 여러 분야에 유용하게 활용되고 있다[3,7].

최근 APE 연구는 다량의 데이터를 통해 학습된 사전학습 언어모델(Pretrained Language Model)을 활용하려는 방향으로 이루어지고 있다[3,8,9]. APE 데이터 생성을 위해서는 번역문의 오류를 수정하는 사람의 직접적인 교정작업이 요구되기 때문에 데이터 구축에 많은 시간과 비용이 소요되고, 이에 따라 데이터 구축에 어려움이 발생하는데, 사전학습 언어모델은 이에 따라 발생하는 APE의 데이터 부족 문제를 해소하는 데 도움을 주는 방법론이라 말할 수 있다[9]. 특히 2019년 WMT컨퍼런스(WMT19)에서는 사전학습된 mBERT모델을 활용한 모델이 가장 좋은 성능을 보였고, 2020년 WMT컨퍼런스(WMT20)에도 대부분의 논문이 사전학습된 언어모델을 활용하는 방법을 활용하였다[10,11].

이에 본 논문에서는 최신 APE 연구 동향을 소개하며, 현재 APE 분야에서 활용되는 다국어 사전학습 언어모델과 향후 연구에서 이용될 수 있는 사전학습 언어모델을 소개한다. 그리고 이를 바탕으로 향후 APE 연구 방향에 대해 고찰한다.

2. 기계번역 사후교정 선행 연구

기계번역 사후교정(APE) 연구는 기계번역 모델을 통해 생성한 번역문에 포함되어있는 오류를 수정하기 위해 등장한 연구 분야이다. 초기 APE는 번역문에서 자주 발생하는 오류들에 대해 사람이 미리 수정 규칙 세우고, 이를 통해 오류를 수정하는 방식이 주로 활용되었고[12], 이후 사람이 직접 오류 수정 규칙을 세우는 것이 아닌, 대용량 말뭉치를 통계적으로 분석하여 오류 수정 규칙을 세우는 방법론이 활용되었다[13]. 이렇게 규칙기반(Rule-Based) 오류 수정 모델, 혹은 통계 분석을 기반으로 한 통계기반(Statistical-Based) 오류 수정 모델은 번역 이전의 소스 문장(Source Sentence)를 고려하지 않고, 번역문에 대한 분석만을 통해 이를 교정문으로 수정하는 작업이다.

최근 APE 연구는 단순히 번역문만을 통한 오류 수정을 넘어, 번역 이전의 소스 문장까지 고려하는 방법이 활용되고 있다. 즉, 소스 문장과 번역문을 통해 교정문(Post Edit)을 생성하는, 일종의 이중 소스 번역(Dual Source Translation) 작업으로 해석할 수 있다. 소스 문장은 교정하려는 문장과 다른 언어로 구성되어있지만, 교정하고자 하는 번역문이 소스 문장으로부터 생성되었다는 점을 고려했을 때, 이는 오류 수정에 긍정적인 영향을 미친다고 볼 수 있다[14]. 2018년도에는 소스 문장과 번역문을 함께 처리하기 위하여 두 개의 개별적인 인코더를 활용하는 방법론이 주로 활용되었으며[14,15], 최근에는 이런 소스 문장과 번역문을 각각 독립적으로 인코딩(encoding) 하지 않고, 소스 문장과 번역문을 함께 연관지어 인코딩하는 방법을 활용하고 있다[8,910].

3. 다국어 사전학습 언어모델

최근 APE 연구는 주로 사전학습된 언어모델을 활용하는 방향으로 이루어지고 있다. 특히 여러 언어 말뭉치에 대한 사전학습을 통해 다중 언어에 대한 교차정보가 습득된 언어모델이 주로 활용된다. 비교적 쉽게 구할 수 있고, 다량으로 존재하는 데이터를 통해 사전학습된 모델을 사용함으로써, APE 작업에 있어 데이터 증강 효과를 기대할 수 있다[9].

사전학습 모델을 활용하는 방법으로는 대표적으로 APE를 학습하기 이전에 모델을 사전학습 모델의 변수값으로 초기화시키는 전이학습(Transfer Learning)방법론이 존재한다. 이러한 방법론은 2019년 WMT에서 mBERT를 활용한 APE 모델이 당해 가장 좋은 성능을

낸 이후, 그 흐름이 가속화되었다. 현재 공개되어 APE를 비롯한 기계번역 분야에서 폭넓게 활용되고 있는 다국어 사전학습 모델은 다음과 같이 정리할 수 있다.

3.1 mBERT

BERT[16]는 트랜스포머(transformer)[17]의 인코더 구조만을 활용한 언어모델로, 대용량의 단일 언어 말뭉치를 통한 자기 지도 학습(Self Supervised Learning)을 통해 훈련하는 과정을 거친다. 이때 사용한 자기 지도 학습 방법론으로는 Masked Language Model(MLM), 그리고 Next Sentence Prediction(NSP) 방법론 두 가지가 존재한다.

MLM이란, 문장을 고의로 훼손시킨 후 이를 소스 문장으로 복원하는 방법론으로, 해당 과정에서는 입력 문장의 15%에 해당하는 토큰을 임의로 선택하고, 선택한 토큰들 중 80%는 [MASK]토큰으로, 10%는 임의의 다른 토큰으로, 그리고 10%는 기존 토큰 그대로 두는 방식으로 문장을 훼손시킨다. 즉, 레이블이 없는(Unlabeled) 단일 언어 말뭉치에서, 기존 문장을 고의로 훼손시키고 소스 문장으로 복원함으로써 언어에 대한 양 방향적 문맥을 파악할 수 있게 된다.

NSP란, 입력 문장을 구성할 때, 50%는 기존의 순서대로 연속된 문장을, 그리고 50%는 임의로 선택된 문장을 연결함으로써, 두 문장 간의 문맥적 의미를 파악하는 작업을 의미한다. 두 문장이 연속된 문장일 때는 1을, 연속된 문장이 아닌 경우에는 0을 도출하게 함으로써 문장 간의 연관성이 타당한지 판별하는 작업을 학습하게 된다.

mBERT에서는 이러한 방법론을 활용하여 104개 언어에 대한 Wikipedia 단일 언어 말뭉치를 통해 모델을 사전학습시킨다. 여러 언어에 대한 단일 언어 말뭉치를 활용함으로써, 해당 모델은 병렬 말뭉치를 통한 훈련 없이 여러 언어에 대한 교차 언어정보를 습득할 수 있었다[18]. 이를 통해 훈련되지 않은 언어에 대한 작업(Zero Shot Learning)에서도 우수한 수행 능력을 보였고, 기계번역과 APE 등, 다중 언어 정보(Multilingual Representation)이 요구되는 여러 작업에서 사전학습 모델로 활용되었다[9].

3.2 XLM

XLM[19]에서는 BERT와 동일하게 트랜스포머 인코더 구조를 활용한다. 이때, BERT에서 활용된 MLM 뿐 아니라, 이전 토큰들을 통해 다음 토큰을 예측하는 Casual Language Modeling(CLM) 학습을 진행한다. CLM이란, 이전 토큰들을 기반으로 다음 토큰을 예측하는 작업을 의미하며, 모델은 훈련 과정에서 식(1)와 같이 확률 p를 모델링한다.

$$p = P(w_t|w_1,\ldots,w_{t-1},\theta) \quad (1)$$

즉, XLM 모델 θ는, CLM을 학습하는 과정에서 이전 토큰 $w_1,\ldots,w_{t-1}$를 통해 다음 토큰 w_t를 예측하는 작업을 학습함으로써 문장 구조에 대한 이해를 얻게 된다.

XLM은 여러 언어데이터를 활용한 MLM과 CLM학습을 통해 여러 언어에 대한 교차언어 정보를 학습한다. 이에 더해, 병렬 말뭉치를 통한 사전학습 방법인 Translation Language Model(TLM)을 훈련함으로써 교차언어 정보를 더 깊게 학습하게 된다. TLM이란, 병렬 말뭉치에 존재하는 소스 문장과 타겟 문장을 하나의 입력으로 연결한 후, MLM과 같이 문장의 일부를 [MASK]로 치환하고, 이를 소스 문장으로 복원하는 작업을 의미한다. TLM 사전학습을 통해 모델은 이종 언어간의 연관성을 더 잘 파악하게 된다. 이때, 입력 문장 내에서 소스 문장과 타겟 문장을 구분하기 위하여 소스 언어와 타겟 언어에 언어 임베딩(Language Embedding)을 각 문장에 더해준다.

해당 논문에서는 여러 언어들에 대한 사전학습을 위해 Wikipedia Data를 사용하였다. 이때, 언어마다 구축한 데이터의 양이 다르기 때문에, 해당 데이터를 아무런 처리 과정 없이 그대로 사용하게 된다면, 다량의 데이터가 수집된 언어에 대해서는 많은 학습이 이루어지지만, 데이터의 양이 상대적으로 적은 언어에 대해서는 학습이 제대로 이루어지지 않게 된다. 이에 따라 해당 논문은 저자원언어에 대한 학습을 촉진하기 위하여, 각 언어에 대해 구축된 데이터의 양에 따라 데이터 샘플링(Sampling) 비율을 다르게 설정하였다. 여기에 사용한 비율은 Multinomial Distribution으로, 각 언어의 데이터가 훈련 단계에서 선택될 확률은 식(2)의 q_i와 같이 결정된다.

$$q_i = \frac{p_i^{\alpha}}{\sum_{j=1}^{N} p_j^{\alpha}} \quad with \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k} \quad (2)$$

여기서 N은 데이터셋 내의 총 언어 개수를, n_k는 k번째 언어데이터의 문장 개수를 의미하며, p_i는 구축된 전체 데이터셋 내의 문장 개수에 대한 i번째 언어데이터의 문장 개수의 비율을 의미한다. 위 식에서

α값을 달리해줌으로써 언어별로 샘플링될 확률을 조절해주게 된다. XLM에서는 이때 활용되는 α값을 0.5로 설정하여 사전학습에 이용하였다.

해당 모델은 발표 당시 영어-독일어 번역에서 가장 뛰어난 성능을 보였을 뿐 아니라, 영어-독일어, 영어-프랑스 등의 언어 쌍에 대한 비지도(Unsupervised) 번역 학습에서도 가장 좋은 성능을 보였다. 이는 언어마다 다른 비율을 주어 MLM과 CLM을 학습시킨 방법론이 매우 뛰어난 교차언어 이해 효과를 주며, 병렬 말뭉치를 통한 사전학습이 여러 언어에 대한 학습을 진행함에 있어 많은 도움이 되는 것으로 해석할 수 있다.

3.3 XLM-R

XLM-R[20]은 사전학습 데이터의 양을 큰 폭으로 늘린 대규모 사전학습 방법을 활용하였다. XLM과 유사하게 트랜스포머 인코더 구조의 모델을 활용했으며, XLM의 사전학습 방법론 중 TLM과 CLM은 활용하지 않고, MLM만을 적용한 사전학습이 진행되었다. 사전학습에는 100개 언어에 대한 Common Crawl[21] 단일 언어 말뭉치가 이용된다. mBERT와 XLM의 사전학습에 이용되었던 Wikipedia data보다 더 방대한 양의 데이터를 사전학습단계에 활용하고, 기존 모델들보다 더 큰 크기의 모델을 설계함으로써, 여러 언어를 학습시킴으로 인해 발생하는 성능저하(Curse of Multilinguality)를 막고, 저자원 언어에서의 성능을 향상시켰다.

XLM-R에서는 사전학습 말뭉치 구성을 위한 sampling 비율을 식(2)와 같이 Multinomial distribution을 통해 결정한다. 단, 이때 사용되는 α값은 0.3으로 조정한다. 이는 XLM보다 더 낮은 α값을 사용함으로써 저자원 데이터가 샘플링될 확률을 상승시켰고, 이에 따라 저자원 데이터에 대한 학습을 강조했다고 볼 수 있다. 이에 더해, XLM과는 다르게 각 언어에 대한 언어 임베딩(Language Embedding)을 추가하지 않음으로써, 언어쌍들에 대한 코드 변환(Code Switching)을 더 용이하게 하였다.

3.4 mBART

기존 mBERT, XLM과 XLM-R의 경우, 트랜스포머의 인코더만을 활용한 모델 구조를 사용하였다. 이를 통해 개체명 인식이나 Cross Natural Language Inference (XNLI)[22]와 같은 Natural Language Understanding (NLU) 작업에서는 좋은 성능을 보일 수 있었으나, 디코더 구조를 포함하지 않기 때문에 번역과 같은 언어 생성 작업에 이용하기에는 한계가 있었다. 이에 mBART[23]는 트랜스포머의 인코더와 디코더를 모두 활용한 모델을 여러 언어에 대한 단일 언어 말뭉치를 통해 사전학습함으로써 번역에 특화된 언어모델을 도출하였다.

사전학습에는 25개 언어에 대한 Common Crawl 단일 언어데이터를 활용하였고, BART[24]에서 제안된 자기 지도 학습 방법론을 차용하였다. BART에서는 5가지 문장 훼손 방법(Noise Scheme)을 활용한 사전학습이 제안되었고, 각 방법은 다음과 같다. 첫 번째로 소스 문장에 임의대로 토큰을 추가하는 Insertion, 두 번째로 소스 문장의 토큰을 임의대로 삭제하는 Deletion, 세 번째로 연속된 문장의 순서를 바꾸는 Sentence Permutation, 네 번째로 입력 내의 토큰을 임의로 선택하여, 그 토큰을 중심으로 입력문 전체를 회전시키는 Document Rotation, 그리고 마지막으로 문장 내의 연속된 토큰을 하나의 [MASK]로 치환하는 Text Infilling 방법론이 있다. 이렇게 고의로 훼손시킨 문장을 원래 문장으로 복원하는 작업을 훈련함으로써, 모델은 양방향적 문맥 및 문장에 대한 이해를 학습하게 된다.

mBART에서는 BART에서 활용된 Noise Scheme을 모두 활용하지 않고, 이들 중 Text Infilling과 Sentence Permutation 방법론만을 활용하여 모델을 사전학습시켰다. Text Infilling을 통한 사전학습 단계에서, [MASK]로 치환할 토큰의 개수는 전체 문장 토큰 개수의 30%로 설정되며 마스킹하는 연속적인 토큰의 길이는 포아송 분포에 따라 결정되는데, 이는 식(3)과 같이 표현된다.

$$f(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!} \tag{3}$$

mBART에서는 마스킹을 위한 연속적인 토큰 개수를 λ값이 3.5인 포아송 분포를 따르게 하는데, 기존 BART에서 진행한 λ값보다 더 크게 설정해줌으로써, 더 깊은 문맥적 이해가 학습되게 하였다.

mBART에서는 저자원 언어의 원활한 학습을 위한 방법으로 Up-Down Sampling 방법론을 적용하였다. 이는 데이터 양이 적은 언어에서는 동일한 데이터를 복사함으로써 그 양을 증강시키고, 데이터 양이 많은 언어에서는 데이터의 일부를 제거함으로써, 저자원언어와 고자원언어간의 데이터 비율을 맞춰주는 작업을 의미한다. 이때, 데이터셋 내의 i번째 언어에 대한 데이터를 증강 비율λ_i는 식(4)와 같이 결정된다.

$$\lambda_i = \frac{1}{p_i} \frac{p_i^\alpha}{\sum_i p_i^\alpha} \quad (4)$$

p_i란, 데이터셋 내에서 i번째 언어의 데이터가 차지하는 비율을 의미한다. 식(2)를 통하여, 작은 p_i값을 가지는 저자원 언어들의 데이터는 증강(Up Sampling)되고, 큰 p_i값을 가지는 고자원 언어들의 데이터는 감소(Down Sampling)된다. mBART에서는 α를 0;7로 설정함으로써 저자원언어의 학습을 강화시켰다.

4. 기계번역 사후교정 연구 동향 분석

최신 APE 연구에서는, APE 데이터만을 통해 모델을 훈련하는 것 이상으로, 대용량의 말뭉치를 통해 사전학습된 언어모델을 활용하는 방법론이 주로 연구되고 있다. 현재 APE 연구에서 사전학습 모델을 활용하는 방법론은 크게 세 가지로 분류할 수 있으며, 각 방법론은 다음과 같다.

4.1 인코더 전이학습

mBERT와 XLM과 같은 다국어 사전학습 언어모델들은, 여러 언어 데이터를 통한 사전학습을 통해 다국어에 대한 교차 언어 정보(Cross Lingual Representation)를 학습하였고, 이에 따라 XNLI[22] 등, 다국어 이해가 요구되는 자연어처리 작업에서 뛰어난 성능을 보여주었다. 최근 APE 연구에서는 해당 모델을 APE모델로 전이학습(Transfer Learning)[25]함으로써, 다국어 언어 이해 모델들의 교차 언어 이해 능력을 활용하는 방법을 적용하고 있다. 전이학습이란, APE 데이터를 통해 모델을 학습하기 이전에 사전학습 모델의 변수(Parameter)값으로 모델을 초기화(Initialize)해주는 방법을 의미한다.

4.1.1 mBERT 기반 전이학습

WMT19에서 가장 좋은 성능의 APE모델을 제안한 Unbabel[9]은 mBERT를 사전학습 모델로 활용하였다. 해당 논문에서 사용된 모델 구조는 Fig.1과 같다.

해당 논문에서는 mBERT를 인코더로 활용한 트랜스포머를 설계하였고, 디코더의 각 변수값도 mBERT와 동일하게 초기화하는 방법을 활용하였다. 이때, 디코더의 Cross Attention 구조는 인코더에 포함되어있지 않기 때문에, 인코더의 Self Attention 구조의 변수값과 동일하게 초기화시켰다.

APE작업은 소스 문장와 번역문을 함께 처리해야 하기 때문에 이중 소스(Dual Source) 접근법으로 문장을 생

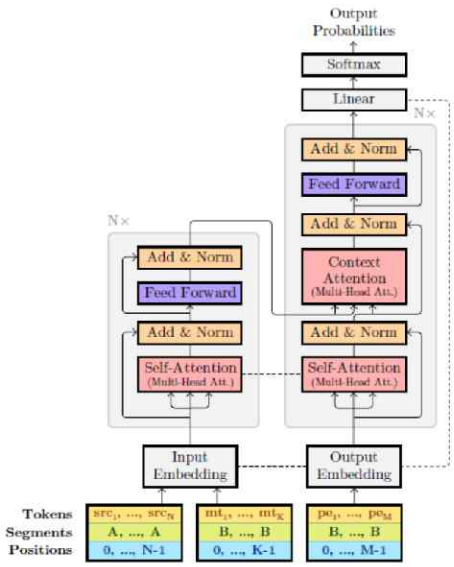

Fig. 1. WMT19 Unbabel's APE Model[9]

성해야 한다. 이에 따라 해당 모델에서는 소스 문장와 번역문을 [SEP]토큰으로 연결하여 하나의 입력을 구성하고, 이를 통해 교정문을 생성하는 구조를 적용하였다. 이때, 소스 문장과 번역문에 각기 다른 Segment Id를 더해줌으로써, 동일한 입력 문장 내에서 소스 문장와 번역문을 구분하였다.

4.1.2 XLM 기반 전이학습

Fig.1과 유사한 구조를 활용하여 POSTECH에서는 WMT20에서 XLM기반의 APE모델을 발표하였다[10]. 해당 모델에서는 트랜스포머 기반 APE모델의 인코더에 mBERT가 아닌 XLM을 전이(Transfer)시킴으로써 기존 모델보다 더 향상된 교차 언어 이해 능력을 얻었다. mBERT를 활용한 전이학습은 APE에서 많은 성과를 보였으나[9,26], mBERT는 병렬 말뭉치를 통해 두 언어 쌍에 대한 관계를 직접 학습하지는 않기 때문에 소스 문장과 번역문을 함께 입력으로 받는 APE 모델 구조에 적용하기에는 한계점이 발생할 수 있다[27]. 이에 해당 모델은 병렬 말뭉치를 통한 TLM 사전학습을 거친 XLM을 활용함으로써, 이종 언어 문장이 연결된 입력을 처리하는데 있어 보다 뛰어난 성능을 기대할 수 있었다[10].

모델의 훈련은 WMT19 Unbabel[9]에서 활용된 방법

과 유사하게 진행된다. 소스문장과 번역 문장을 하나로 합쳐서 Joined Representation[1]을 구성하고 이를 입력으로 받아 교정문장을 생성하는 작업을 학습하며, APE 데이터를 통한 학습 이전에 XLM의 변수 값과 동일하게 모델의 인코더를 초기화시킨다. 단, WMT19 Unbabel의 모델 구조와는 달리, 디코더의 변수들을 인코더와 같은 변수 값으로 초기화하지 않고, 모두 무작위 값으로 초기화하였다. 이후 APE 데이터를 통해 전체 모델 구조를 미세조정(Fine Tuning)함으로써, APE에 특화된 트랜스포머 모델을 생성하였다.

4.2 번역 모델 기반 교정 모델

Unbabel과 Postech등의 APE 연구에서는 MLM과 같은 자기 지도 학습방법으로 훈련된 언어모델을 APE 모델로 전이 학습하는 방법론을 적용하였다. 하지만 WMT20 우승 모델인 Huawei Translation Services Center(HW-TSC)에서는, 이러한 연구의 흐름에서 벗어나 번역 모델을 사전학습 모델로 활용하는 방법을 제안하였다. 해당 논문에서는 WMT19 뉴스 번역 공동 작업(News Translation Shared Task)에서 공개된 병렬 말뭉치를 통해 학습된 트랜스포머 기반의 번역 모델을 활용하였고, 해당 모델을 APE 데이터로 미세조정하였다.

이때, 사전학습된 번역 모델을 APE 특화 모델로 미세조정하는 과정에서 전체 모델을 학습시키지 않고, Bottleneck Adapter Layer(BAL)구조[28]를 도입하는 방법을 활용하였다. 이는 미세조정시, 모델에 BAL구조를 삽입한 후, 기존에 사전학습된 모델 구조는 학습하지 않고(Freeze) BAL 구조만을 학습하는 방법론으로, 이를 통해 전체 모델을 학습시킬 때보다 더 적은 양의 변수를 학습하면서도 뛰어난 미세조정 성능을 얻을 수 있다. BAL은 두 개의 Dense Layer와 그 사이의 Relu 활성화 함수(Activation Function)로 구성되어있으며, 두 Dense Layer 사이의 중간 차원(Dimension)은 기존 트랜스포머 모델의 은닉층(Hidden Layer)크기의 2배로 설정한다. HW-TSC에서 제안한 APE모델 구조는 Fig. 2와 같다.

Fig.2를 통해 확인할 수 있듯, 해당 모델의 학습 및 추론단계에서 입력을 구성할 때, 소스 문장(src)와 번역문(mt)를 연결하고, 외부 번역 모델을 통한 번역문 mt'을 추가적으로 연결해준다. 해당 논문에서는 구글 번역기를 통해 생성한 번역문을 mt'으로 활용하였으며, 이는 추가적인 번역문을 통해 외부 번역모델에서의 정보를 활용할 수 있다는 점에서, 하나의 데이터 증강 기법으로 이해할 수 있다.

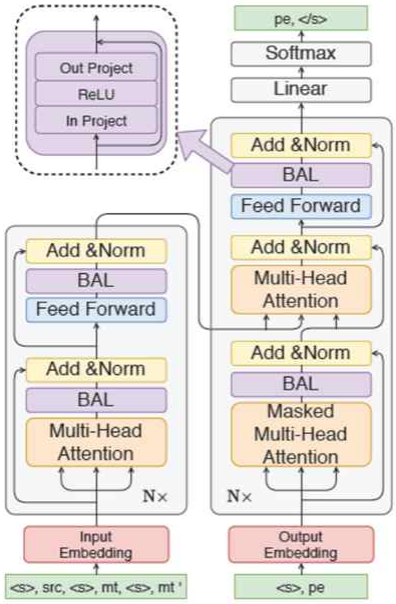

Fig. 2. WMT20 HW-TSC APE Model[8]

4.3 기계번역 품질 예측 기반 사후 교정

Berling Lab에서는 WMT20에서 XLM-R기반의 기계번역 품질 예측(Quality Estimation, QE)모델을 APE에 적용하는 방법을 활용하였다[11]. QE란, 실제 정답 문장과의 비교를 통해 번역문의 품질을 예측하는 것이 아닌, 소스 문장과 번역문을 통해 번역문의 품질을 예측하는 작업을 의미하며[29], 해당 연구에서는 단위(word level) 품질 예측과 문장 단위 품질 예측을 활용하였다.

먼저 단어 단위 예측에서는 소스 문장을 통해 번역문 내의 각 단어에 대한 번역 품질을 예측하고, 품질에 따라 OK/BAD 태그를 부착한다. 해당 논문에서는 단어 단위 예측모델을 통한 태그 부착 이후, BAD 태그가 부착된 단어를 [MASK]로 치환하고, 이렇게 일부가 가려진 문장을 MLM이 학습된 XLM-R모델을 통해 복원함으로써 번역문의 오류를 교정한다.

이후 문장 단위 예측을 통한 문장 선택 과정을 거친다. 문장 단위 예측이란, 번역문 전체에 대한 번역 품질을 예측하는 작업으로, 해당 논문에서는 교정 이전의 번역문과 교정 이후의 번역문의 품질을 소스 문장을 통해 예측하고, 예측한 두 문장의 품질을 비교함으로써 고품질의 문장을 선택하게 된다. 이는 APE의 Over Correction[30,31]문제를 해소하는 방법으로, 번역을 통해 옳은 문장을 더 저품질의 문장으로 바꾸어버리는 문제를 해결하기 위해 도

입된다.

Berling Lab의 논문에서는 이렇게 APE 데이터를 활용하지 않고도, 단어 단위 품질 예측 및 마스크 복원 모델과 문장 단위 품질 예측모델을 통해 번역문의 품질을 향상시킬 수 있음을 보였다. 이는 APE 데이터 없이도 사후교정을 이룰 수 있는 방법론으로 해석할 수 있으며, 특히 APE 데이터가 존재하지 않는 한국어에서도 적용할 수 있는 연구라는 부분에서 의의가 있다.

5. 향후 연구 방향성 논의

현재 APE 연구에서는 mBERT, XLM, XLM-R 등 인코더 구조의 다국어 사전학습 모델이 활용되고 있다. 하지만 인코더 구조로 이루어진 모델을 APE 작업에 적용하기 위해서는 디코더를 추가로 설계해야 하는 문제가 발생한다[9,10]. 기존의 접근 방법에서는 디코더의 변수값을 인코더와 같게 설정해주거나, 디코더를 새롭게 생성하는 방법을 사용했으나, 두 방법론 모두 사전학습 모델을 온전히 이용하는 것으로 보기는 어렵다. 이 한계점은 디코더가 포함된 다국어 사전학습 모델인 mBART를 적용함으로써 해소할 수 있을 것으로 예측된다. 모델 구조에 디코더가 포함되어있기 때문에, APE 미세조정에서 디코더를 새롭게 도입할 필요 없이, 교차언어 정보가 습득된 사전학습 모델을 APE에 온전히 활용할 수 있게 된다.

mBART모델 구조를 활용한 APE모델의 학습은, 선행연구들의 결과를 고려했을 때, 전이학습 기반의 생성 방법을 활용하는 것이 이상적으로 보인다. 현재 APE에서 사전학습 언어모델을 이용하는 방법은 크게 전이학습을 적용하는 방법과 노이즈 복원모델로 활용하는 방법 두 가지로 나눌 수 있다. 전이학습 기반의 연구는 POSTECH에서 XLM을 활용한 연구가 대표적이며[10], 이는 교정문 Z를 생성함에 있어 원문 X와 번역문 Y에 대하여 식(5)와 같은 학습 목적을 가지고 있다.

$$\max \sum_{i=1}^{n} \log(P(z_i|X, Y, z_{j<i}, \theta)) \qquad (5)$$

이는 원문과 교정문을 모두 참고하여 자동회귀(auto regressive)방식으로 교정문의 각 토큰 z_i를 생성하는 방법이다. 사전학습 모델을 노이즈 복원 모델로 활용하는 연구는 대표적으로 XLM-R을 활용한 Berling lab의 연구가 존재하며[11], 식(6)과 같은 학습 목적을 가지고 있다.

$$\max \sum_{i=1}^{n} \log(P(z_i|X, Y', \theta)) \qquad (6)$$

여기서 Y'는 QE모델을 통해 품질이 낮은 단어를 마스킹한 문장을 의미한다. 식(6)을 통해 확인할 수 있듯이, 해당 방법을 활용하는 경우 자동회귀적으로 문장을 생성하지 않고, 각 토큰 위치마다 독립적인 마스크 복원 및 토큰 생성 작업이 진행된다. 본 방법으로 사전학습 언어모델을 활용하는 경우, TER 1.10점의 성능 향상을 이룰 수 있었으나[11], 식(5)과 같은 학습방법으로 모델을 활용했을 때 TER 4.34점의 성능 향상을 이뤘다는 점과 비교한다면[10] 그 수치는 미미한 수준이었다. 이는 식(5)와 같은 학습 목적으로 전이학습 기반의 APE 미세조정을 진행함으로써 더 큰 성능 향상을 이룰 수 있음을 보여주며, mBART를 APE에 적용하는 경우에도, 해당 방법을 적용하는 것이 이상적일 것이라 예상할 수 있다. 해당 방법으로 mBART 기반 APE모델을 생성하는 경우, 모델 구조는 Fig.1과 동일하며, APE 미세조정 시 입력 구조는 [10]에서와 유사하게, 원문과 번역문을 한데 합친 문장을 활용할 수 있을 것으로 보인다.

나아가 본 모델과 XLM간의 성능을 비교함으로써, APE의 성능 향상에 도움이 되는 사전학습 방법에 대한 해답을 얻을 수도 있을 것으로 기대된다. 이는 mBART 모델의 경우 다량의 단일언어 언레이블 데이터로 사전학습을 진행했고, XLM모델의 경우 mBART보다는 상대적으로 적은 양의 단일 언어 데이터를 활용하였으나 병렬 데이터를 통한 TLM을 진행했다는 점을 통해 확인할 수 있다. 각 모델 기반의 APE 성능을 비교함으로써 APE에 효과적인 학습방법에 대해 확인할 수 있으며, 이를 통해 향후 APE 연구를 진행함에 있어 최적의 연구 방향을 제안해줄 수 있을 것으로 보인다.

또한, 현재 WMT20에서 발표된 APE연구들의 성과를 바탕으로, 향후 연구의 방향을 확인할 수 있었다. 먼저 WMT20 영어-독일어 APE 모델들의 성능 표는 Table 1과 같다.

Table 1. WMT20 APE results table[3]

	TER	BLEU
HW-TSC [8]	20.21	66.89
Alibaba [32]	26.99	55.77
Postech (XLM) [10]	27.02	56.37
Berlinglab [11]	27.61	54.71
Postech (Noise) [33]	28.22	54.51
Baseline	31.36	50.21

Table.1에서 확인할 수 있듯, 기계번역을 사전학습 모델로 활용한 HW-TSC의 APE 모델은 다른 연구들보다 월등히 좋은 성능을 보였다. 특히 해당 모델은 XLM 기반의 APE모델을 설계했던 Postech의 모델과 비교했을 때 TER에서 6.79점 향상된 결과를 보였는데, 이를 통해 번역이 학습된 모델을 APE에 적용하는 것이, MLM등의 denoising이 사전학습된 언어모델을 활용하는 것보다 더 효과적으로 작용할 수 있음 확인할 수 있다. 나아가, MLM과 같이 언레이블 데이터를 기반으로 자기 지도 학습방법을 활용하여 문장을 복원하는 작업을 훈련하는 것보다, 번역 학습을 통해 이종 언어간의 관계성을 학습한 모델이 APE에서 더 효과적으로 작용하는 것으로도 해석할 수 있다. WMT20에서 HW-TSC연구가 발표되기 이전 대부분의 APE연구들에서 denoising작업을 학습한 언어모델을 활용했다는 점을 고려했을 때, 앞으로 번역 모델을 기반으로 한 APE연구를 진행한다면 보다 발전된 모델을 얻을 수 있을 것으로 기대한다.

위 고찰을 바탕으로 본 논문에서는 mBART기반 번역 모델을 바탕으로 APE를 미세조정하는 방법을 제안한다. mBART는 다국어 언어이해뿐 아니라 기계번역에도 특화된 모델로, 한-영 언어 쌍 등, 여러 언어 쌍들에 대해 현재까지도 가장 좋은 번역 성능을 보여주고 있다. 이에 따라, 기존 언어모델이나 트랜스포머 기반 번역모델을 사전학습 모델로 채용했던 기존 APE 연구들에 mBART를 적용한다면, 기존 성능보다 뛰어난 성능의 교정 모델을 얻을 수 있을 것으로 전망한다.

6. 결론

본 논문에서는 APE 연구의 최신 동향 및 활용되는 다국어 사전학습 언어모델을 소개하였고, 이를 바탕으로 향후 APE 연구 방향을 제안하였다. 이전 연구에서는 mBERT, XLM등, 인코더 구조로 이루어진 다국어 사전학습 모델을 활용하였으나, 이는 디코더를 새롭게 학습해야 한다는 한계점이 있었고, 이에 따라 본 논문에서는 위 한계를 극복하는 방법으로 디코더구조가 포함된 다국어 사전학습 모델인 mBART를 활용하는 방안을 제안하였다.

또한, 최신 APE 연구들 간의 비교를 통해 자기 지도학습 방법 기반의 사전학습을 활용했을 때보다 번역 모델을 전이 학습하였을 때 더 뛰어난 성능의 APE 모델을 얻을 수 있음을 확인하였고, 이에 따라 번역 모델을 기반으로 한 APE 모델 연구를 제안하였다. 이를 바탕으로 본 논문에서는 mBART기반 번역 모델을 사전학습모델로 활용함으로써, 기존 연구보다 뛰어난 성능의 APE 모델을 얻을 수 있을 것으로 전망하였다.

REFERENCES

[1] Park, C., & Lim, H. (2020). Automatic Post Editing Research. *Journal of the Korea Convergence Society*, 11(5), 1-8.

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

[3] Chatterjee, R., Freitag, M., Negri, M., & Turchi, M. (2020, November). Findings of the WMT 2020 shared task on automatic post-editing. *In Proceedings of the Fifth Conference on Machine Translation*, (pp. 646-659).

[4] Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P. J., & Guzmán, F. (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).

[5] Specia, L. et al. (2020, November). Findings of the WMT 2020 shared task on quality estimation. *In Proceedings of the Fifth Conference on Machine Translation* (pp. 743-764).

[6] Pal, S., Herbig, N., Krüger, A., & van Genabith, J. (2018, October). A transformer-based multi-source automatic post-editing system. *In Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (pp. 827-835).

[7] Ive, J., Specia, L., Szoc, S., Vanallemeersch, T., Van den Bogaert, J., Farah, E., ... & Khalilov, M. (2020, May). A Post-Editing Dataset in the Legal Domain: Do we Underestimate Neural Machine Translation Quality?. *In Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 3692-3697).

[8] Yang, H., Wang, M., Wei, D., Shang, H., Guo, J., Li, Z., ... & Chen, Y. (2020, November). HW-TSC's Participation at WMT 2020 Automatic Post Editing Shared Task. *In Proceedings of the Fifth Conference on Machine Translation* (pp. 797-802).

[9] Lopes, A. V., Farajian, M. A., Correia, G. M., Trénous, J., & Martins, A. F. (2019). Unbabel's Submission to the WMT2019 APE Shared Task: BERT-based Encoder-Decoder for Automatic Post-Editing. *arXiv preprint arXiv:1905.13068.*

[10] Lee, J., Lee, W., Shin, J., Jung, B., Kim, Y. G., & Lee, J. H. (2020, November). POSTECH-ETRI's Submission to the WMT2020 APE Shared Task: Automatic Post-Editing with Cross-lingual Language Model. *In Proceedings of the Fifth Conference on Machine Translation* (pp. 777-782).

[11] Lee, D. (2020, November). Cross-Lingual Transformers for Neural Automatic Post-Editing. *In Proceedings of the Fifth Conference on Machine Translation* (pp. 772-776).

[12] Allen, J., & Hogan, C. (2000, April). Toward the development of a post editing module for raw machine translation output: A controlled language perspective. *In Third International Controlled Language Applications Workshop* (CLAW-00) (pp. 62-71).

[13] Simard, M., Goutte, C., & Isabelle, P. (2007, April). Statistical phrase-based post-editing. *In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference* (pp. 508-515).

[14] Shin, J., & Lee, J. H. (2018, October). Multi-encoder transformer network for automatic post-editing. *In Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (pp. 840-845).

[15] Junczys-Dowmunt, M., & Grundkiewicz, R. (2018). MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. *arXiv preprint arXiv:1809.00188*.

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

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762.*

[18] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual bert?. *arXiv preprint arXiv:1906.01502.*

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

[20] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116.*

[21] Wenzek, G., Lachaux, M. A., Conneau, A., Chaudhary, V., Guzman, F., Joulin, A., & Grave, E. (2019). Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359.*

[22] Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053.*

[23] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8, 726-742.

[24] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461.*

[25] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22(10)*, 1345-1359.

[26] Correia, G. M., & Martins, A. F. (2019). A simple and effective approach to automatic post-editing with transfer learning. arXiv preprint arXiv:1906.06253.

[27] Jihyung L, WonKee L, Young-Gil K, Jonghyeok L. (2020). Transfer Learning of Automatic Post-Editing with Cross-lingual Language Model. *KIISE 2020*, 392-394.

[28] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. *In International Conference on Machine Learning* (pp. 2790-2799). PMLR.

[29] Kim, H., Lee, J. H., & Na, S. H. (2017, September). Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. *In Proceedings of the Second Conference on Machine Translation* (pp. 562-568).

[30] Park, C., Yang, Y., Lee, C., & Lim, H. (2020). Comparison of the Evaluation Metrics for Neural Grammatical Error Correction With Overcorrection. *IEEE Access,* 8, 106264-106272.

[31] Park, C., Kim, K., Yang, Y., Kang, M., & Lim, H. (2020). Neural spelling correction: translating incorrect sentences to correct sentences for multimedia. *Multimedia Tools and Applications,* 1-18.

[32] Wang, J., Wang, K., Fan, K., Zhang, Y., Lu, J., Ge, X., ... & Zhao, Y. (2020, November). Alibaba's Submission for the WMT 2020 APE Shared Task: Improving Automatic Post-Editing with Pre-trained Conditional Cross-Lingual BERT. In Proceedings of the Fifth Conference on Machine Translation (pp. 789-796).

[33] Lee, W., Shin, J., Jung, B., Lee, J., & Lee, J. H. (2020, November). Noising Scheme for Data Augmentation in Automatic Post-Editing. In Proceedings of the Fifth Conference on Machine Translation (pp. 783-788).

문 현 석(Moon, Hyeonseok) [학생회원]

· 2021년 2월 : 고려대학교 수학과 및 인공지능학과(이학사, 공학사)

· 2021년 3월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통

합과정
· 관심분야 : Neural Machine Translation, Natural Language Processing
· E-Mail : glee889@korea.ac.kr

박 찬 준(Park, Chanjun) [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(Eo, Sugyeong) [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

서 재 형(Seo, Jaehyung) [학생회원]

· 2020년 8월 : 고려대학교 영어영문학과 및 경영학과(문학사, 경영학사)
· 2020년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Graph Encoder, Commense Reasoning
· E-Mail : seojae777@korea.ac.kr

임 희 석(Lim, Heuiseok) [정회원]

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