Journal of The Korea Convergence Society
Vol. 12. No. 10, pp. 27-35, 2021
I-SSN 2233-4890 / e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2021.12.10.027

전이학습 기반 기계번역 사후교정 모델 검증

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

The Verification of the Transfer Learning-based Automatic Post Editing Model

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 사후교정 (Automatic Post Editing, APE)이란 번역 시스템을 통해 생성한 번역문을 교정하는 연구 분야로, 영어-독일어와 같이 학습데이터가 풍부한 언어쌍을 중심으로 연구가 진행되고 있다. 최근 APE 연구는 전이학습 기반 연구가 주로 이루어지는데, 일반적으로 self supervised learning을 통해 생성된 사전학습 언어모델 혹은 번역모델이 주로 활용된다. 기존 연구에서는 번역모델에 전이학습 시킨 APE모델이 뛰어난 성과를 보였으나, 대용량 언어쌍에 대해서만 이루어진 해당 연구를 저 자원 언어쌍에 곧바로 적용하기는 어렵다. 이에 본 연구에서는 언어 혹은 번역모델의 두 가지 전이학습 전략을 대표적인 저 자원 언어쌍인 한국어-영어 APE 연구에 적용하여 심층적인 모델 검증을 진행하였다. 실험결과 저 자원 언어쌍에서도 APE 학습 이전에 번역을 한차례 학습시키는 것이 유의미하게 APE 성능을 향상시킨다는 것을 확인할 수 있었다.

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

Abstract Automatic post editing is a research field that aims to automatically correct errors in machine translation results. This research is mainly being focus on high resource language pairs, such as English-German. Recent APE studies are mainly adopting transfer learning based research, where pre-training language models, or translation models generated through self-supervised learning methodologies are utilized. While translation based APE model shows superior performance in recent researches, as such researches are conducted on the high resource languages, the same perspective cannot be directly applied to the low resource languages. In this work, we apply two transfer learning strategies to Korean-English APE studies and show that transfer learning with translation model can significantly improves APE performance.

Key Words : Deep learning, Natural language process, Language convergence, Machine translation, Automatic post editing, Pretrained model

*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 August 3, 2021
Revised August 18, 2021
Accepted October 20, 2021
Published October 28, 2021

1. 서론

기계번역 사후교정(Automatic Post Editing, APE)이란 기계번역 연구의 하위 분야로, 번역 시스템을 통해 생성된 번역문을 교정해주기 위한 연구이다[1]. APE는 번역문 내에 포함된 오류를 교정해주기 위한 사람의 노력을 경감시킨다는 점과[2] 도메인 특화번역에 성과를 낼 수 있다는 점에서[3,4] 주목받고 있는 연구 분야이며, 현재 Conference on Machine Translation (WMT)에서 공유 과제로 지정되어 활발하게 연구되고 있다[5].

최근 APE 연구는 전이학습(Transfer Learning) 기반으로 연구가 이루어지고 있다[6]. 대용량의 언레이블(Unlabeled) 데이터를 통한 자기 지도학습 방법론(Self Supervised Learning)을 통해 학습한 언어모델을 사전학습 모델로 활용하거나[7], 번역모델을 사전학습 모델로 활용하는 방법[8] 두 가지로 나눌 수 있다. 특히, 번역모델에 APE 작업을 미세조정(fine-tuning)하는 것은 APE 모델 생성에 있어 매우 효과적인 접근이라는 것이 최근 연구를 통해 밝혀졌다[4].

하지만 이들 연구 결과를 통해, APE에 있어 번역모델을 전이 학습시키는 것이 항상 효과적인 접근법이 된다고 속단하기는 어렵다. 이는 지금까지의 APE 연구가 영어-독일어와 같이 데이터의 양이 매우 풍부한 언어쌍에만 집중하고 있을 뿐, 한국어-영어와 같이 비교적 저 자원 언어쌍에 대한 연구는 진행되지 않았기 때문이다. 특히, 학습데이터의 양은 번역모델의 성능에 매우 큰 영향을 주기 때문에[9], 번역모델을 활용하여 높은 APE 성능을 얻었던 고 자원 언어쌍에서의 학습 전략(Training Strategy)이, 저 자원 언어쌍에서도 동일하게 효과적으로 적용할 수 있을 것으로 보기는 어렵다. 이는 한국어-영어와 같은 저 자원 언어쌍에 대해서, 추가적인 논의 없이 영어-독일어의 기존 연구와 동일한 관점을 적용할 수는 없음을 보여준다.

이에 본 연구에서는 저 자원 언어쌍에 대해서도 번역모델 기반 전이학습 방법론이 APE에서 효과적으로 작용할 수 있는지 검증한다. 이를 위하여 본 연구에서는 번역모델 기반의 APE 모델과, 언레이블(Unlabeled) 언어 데이터를 활용해 자기 지도학습 방법론으로 훈련된 다국어 사전학습 언어모델 기반 APE 모델을 각각 생성하였다. 다국어 사전학습 언어모델로는 현재 많은 기계번역 연구들에서 널리 활용되어 뛰어난 성능을 보여주고 있는 mBART[10]를 활용하였으며, 번역모델로는 트랜스포머(Transformer) [11] 모델을 활용하였다. 본 연구에서는 이들 간의 성능을 비교함으로써, 저 자원 언어쌍에 대해서도 번역모델을 기반으로 한 APE 학습 전략이 효과적으로 작용할 수 있는지 검증하였다. 모든 실험은 대표적인 저 자원 언어쌍인 한국어-영어 언어쌍에 대하여 진행하였다. 특히, 한국어-영어와 같은 저 자원 언어쌍에 대해서는 사람의 교정작업을 통해 생성한 APE 데이터가 존재하지 않기 때문에, 병렬 말뭉치로부터 노이즈 생성방법(Noising Scheme)을 활용해 직접 생성한 APE 데이터를 훈련에 활용하였다.

최종적으로, 본 연구에서는 다국어 사전학습 모델과 번역모델에 각각 전이학습시킨 APE 모델의 성능에 대한 정량적 및 정성적 분석을 통해 APE에서 최적의 연구 방향을 제안한다.

2. 관련 연구

APE 데이터 생성을 위해서는 번역문의 오류를 수정하는 사람의 직접적인 교정작업이 요구되기 때문에 데이터 구축에 많은 시간과 비용이 소요되고, 이에 따라 다량의 데이터 구축에 어려움이 발생한다. 이에 따라 일반적으로 APE 연구에서는 데이터 부족 문제가 존재하며, 최근 APE 연구는 데이터 부족 문제를 해결하기 위한 연구 중심으로 이루어지고 있다[5]. 대표적으로는 번역 시스템을 통해 병렬 말뭉치로부터 유사 APE 데이터(pseudo-APE triplet)을 생성하는 방법을 제안한 연구가 존재한다[12]. 해당 연구에서는 병렬 말뭉치 내에 존재하는 원문(source sentence)과 타겟 문장(target sentence)을 각각 APE 데이터의 원문, 교정문으로 간주하고, 직접 설계한 번역 시스템을 통해 원문을 번역하여 APE 데이터를 생성하는 방법론을 제안하였다. 해당 방법론을 통해 생성한 데이터를 활용하는 것은 APE 모델의 성능을 유의미하게 상승시키며, 최근 많은 연구에서도 해당 방법론을 차용한 데이터 증강 기법을 적용하고 있다[7]. 하지만 해당 방법론에서는 번역문과 교정문이 각각 독립적으로 생성되기 때문에, 생성한 번역문이 실제 APE 모델을 통해 교정되어야 하는 정보를 담고 있지 않을 수 있다는 점이 지적된다[13].

위 방법 이외로도, 병렬 말뭉치를 통한 데이터 증강 기법은 최근에도 활발하게 연구되고 있다. 포스텍에서는 WMT20에서 노이즈 생성방법론을 기반으로 유사 APE 데이터(pseudo-APE triplet)를 생성하는 방법론을 제

안하였고, 단순하게 번역 기반의 데이터 증강만을 적용했을 때보다 더 우수한 결과를 얻을 수 있음을 실험을 통해 입증하였다[14]. 또한, 번역 시스템을 통해 병렬 말뭉치로부터 APE 데이터를 증강하는 경우, 교정문 사이의 관계성이 충분히 고려되지 않는다는 점을 지적하면서, 역번역(Back-Translation)방법[15]과 유사하게 데이터를 증강하는 방법론 또한 제안되었다[13].

데이터 부족 문제를 극복하기 위한 데이터 증강 기법으로, 사전학습 언어모델을 활용하는 방법도 활용되고 있다. 이는 APE 모델을 생성할 때, 단순히 APE 훈련데이터로만 학습을 진행하는 것이 아닌, 다량의 데이터를 통해 한차례 학습이 끝난 언어모델에 APE를 미세조정해주는 방법론을 의미한다. 이는 다량의 언어데이터를 통해 학습한 정보가 APE 작업을 원활하게 진행하는 데 도움을 준다고 볼 수 있다[5]. 이러한 관점은, 자연어처리 기반 분류 작업을 비롯한[16] 여러 자연어처리 연구 분야에서도 활발하게 적용되고 있으며[17], 주로 BERT[18]나 mBART[10]와 같은 사전학습 언어모델들이 활용되고 있다.

단, APE에서는 사전학습 모델을 활용하는 경우, 고려할 수 있는 두 가지 선택사항이 존재한다. 이는 마스크 모델링(Masked Language Modeling, MLM)[18], 번역 모델링(Translation Language Modeling, TLM)[19]과 같이 기존 문장에 noise를 주고, 이를 원본 문장으로 복원하는 사전학습 방법을 통해 훈련된 모델을 활용하는 방법과[6,7] 사전학습된 번역모델을 활용하는 방법[8] 두 가지로 분류할 수 있다. 최근 WMT20에서의 결과를 분석해봤을 때, 번역모델에 APE task를 미세조정하는 것은, 다국어 사전학습 언어모델을 활용한 APE 모델들의 성능보다 월등히 뛰어난 성능을 보인다는 것을 확인할 수 있다[5].

3. 제안하는 방법

3.1 사전학습 언어모델 전이학습

기존 연구들을 통해, 번역모델을 사전학습 언어모델로 활용하여 전이학습 방법론으로 APE 모델을 생성하는 것은 매우 효과적이라는 것이 밝혀졌다[5]. 하지만 이들 연구는 영어-독일어와 같이 비교적 병렬 말뭉치의 양이 풍부한 언어 쌍에 대해서 실험이 이루어졌기 때문에, 비교적 저 자원 언어쌍에 속하는 한국어-영어와 같은 연구에 곧바로 적용하기에는 무리가 있다. 즉, 기존 연구들의 실험결과만을 통해, 한국어-영어 APE에서 최적의 전이학습 접근방법을 단정 지을 수는 없다.

이에 본 연구에서는 다국어 사전학습 언어모델을 기반으로 전이 학습한 APE 모델과 번역모델을 기반으로 전이 학습하여 생성한 APE 모델 간의 성능 비교가 이루어져야 함을 제안한다. 이때 다국어 사전학습 언어모델로는 현재 기계번역 여러 분야에서 활발하게 활용되고 있는 mBART 모델[10]을 활용했으며, 번역모델로는 자체적으로 학습시킨 Transformer[11] 기반 번역모델을 활용하였다.

본 연구에서 전이학습방법론을 통해 각 언어모델에 APE 작업을 미세조정하는 방법은 다음과 같이 설명된다. 기본적으로 원문 X, 번역문 Y, 교정문 $Z = \{z_i\}_{i=1}^{n}$로 구성된 APE 훈련 데이터셋 D에 대하여, 학습하고자 하는 APE 모델 θ의 훈련 목표(Training Objective)는 식(1)과 같다.

$$\max_{\theta} \sum_{(X,Y,Z) \in D} \left(\sum_{i=1}^{n} \log P(z_i | X, Y, z_{k<i}, \theta) \right) \quad (1)$$

이는 APE 데이터에 존재하는 원문과 번역문을 연결하여 하나의 입력구조를 생성하고, 이를 입력으로 받아 교정문을 생성하는 시퀀스 투 시퀀스(Sequence to Sequence)[20] 기법을 적용한 것이라고 말할 수 있다. 본 연구에서 활용한 입력구조는 식(2)와 식(3)과 같다.

$$[bos]\, X\, [bos]\, Y \quad (2)$$

$$[bos]\, X\, [bos]\, Y\, [eos]\, [lid] \quad (3)$$

여기서 $[bos]$와 $[eos]$는 각각 문장의 시작과 끝을 알리는 특수 토큰을 의미하며, $[lid]$는 번역문의 언어를 알리는 특수 토큰으로, mBART의 사전학습에 활용되었던 언어 토큰과 동일하게 활용되었다. 식(2)는 Transformer 기반 APE 모델을 훈련하는 데 이용한 입력구조이며, 이는 기존 연구에서 번역모델 기반 APE 모델을 생성할 때 활용한 전이학습 방법과 동일하다[8]. 식(3)은 mBART 기반 APE 모델을 훈련에 이용한 입력구조로, mBART의 사전학습 단계에서 활용했던 입력구조와 일관성 있는 구조를 생성하였다. 원문과 번역문을 하나의 문장으로 연결하여 입력구조를 생성하는 방법은 현재 많은 연구에서 활용되고 있는 방법으로[6,7], 본 연구에서도 이들과 유사한 방법론을 적용하였다.

3.2 병렬 말뭉치를 활용한 APE 데이터 생성

현재 한국어-영어 APE 모델 생성을 위한 데이터는 공개된 바가 없다. 이에 본 연구에서는 전문인력이 교정한 APE 데이터가 없는 상황에서 모델을 생성하기 위하여, 병렬 말뭉치를 기반으로 한 유사 APE 데이터를 생성 방법을 활용한다.

먼저 병렬 말뭉치 내의 원문과 타겟 문장을 APE 데이터의 원문과 교정문으로 각각 간주한다. 그리고 타겟 문장을 고의로 훼손시킴으로써(Noising Scheme) APE 데이터의 번역문 역할을 할 문장을 자동으로 생성한다. 해당 방법론은 WMT20에서 포스텍이 제안한 방법과 유사하다[14]. 단, 이렇게 기존 문장을 훼손시킴으로써 생성한 데이터를, 번역 시스템을 기반으로 증강한 데이터와 힘께 활용했던 이전 연구와는 다르게, 본 연구에서는 모든 데이터를 noising scheme을 기반으로 생성하였다.

기존 대부분의 연구에서 활용하고 있는 APE 데이터 생성 방법인 번역모델을 기반 데이터 증강 기법[12]에서는, 번역모델의 성능이 학습데이터의 양에 크게 의존하기 때문에[9], 병렬 데이터가 충분히 존재하지 않는 언어쌍에 대해서는 낮은 품질의 훈련데이터가 생성될 수 있다는 우려가 존재한다. 이는 저 자원 언어쌍에 대해서는 원하는 품질의 번역문을 생성할 수 없음을 의미하고[21], 실제 모델 활용 단계에서 입력으로 주어질 번역문의 품질과 관계없이, 낮은 품질의 번역문 데이터를 통해 APE의 학습을 진행해야 한다는 문제점이 발생한다. Noising scheme을 통한 데이터 증강은, 이러한 한계를 해소할 수 있으면서, 훈련데이터 내의 번역문 품질을 사용자의 임의대로 조절하여, 일관적인 품질의 훈련데이터를 만들 수 있다는 강점이 있다.

기존 연구를 바탕으로 본 논문에서 사용한 noise scheme은 insertion, deletion, substitution, shifting의 4가지가 존재한다. Insertion은 원본 문장에 추가적인 token을 집어넣음으로써 noise를 생성하는 방법이고, deletion은 원본 문장의 기존 토큰을 일부 삭제하는 방법, substitution은 원본 문장의 토큰을 다른 토큰으로 치환하는 방법, 그리고 shifting은 원본 문장 내의 토큰의 위치를 변경하는 방법이다. 이런 noise scheme들을 원본 문장에 무작위로 적용시킴으로써 noised sentence를 생성하고, 해당 과정을 통해 생성된 문장은 삼중항(triplet)으로 이루어진 APE 데이터 (원문, 번역문, 교정문) 중 번역문으로써 활용된다. 본 논문은 영어-독일어 번역에 대해서, 추가적인 데이터로써만 활용되었던 noising scheme 기반 APE 데이터를 한국어-영어 APE에 적용시켰다. 또한, 병렬 데이터에서 noise를 기반으로 합성한 데이터만을 훈련데이터로 활용하였고 그 이외의 데이터는 배제하여 APE 모델을 훈련시켰다.

4. 실험 및 실험결과

4.1 데이터셋 및 평가 지표

본 실험에서는 AIhub[1]에서 제공된 전문 한영 병렬 말뭉치 160만 문장이 활용된다. 해당 말뭉치는 번역 전문가들의 검수 과정을 통해 구축되어 품질에 대한 신뢰성이 보증된 말뭉치로, 구어체, 대화체, 뉴스데이터, 한국문화, 조례, 지자체 6개 도메인 데이터들로 구성되어 있다. 본 연구에서는 더욱 일관성 있는 훈련데이터의 생성을 위하여 데이터의 개수가 2개보다 작거나 200개보다 많은 데이터를 필터링하였다[22,23]. APE 모델의 훈련 및 검증을 위하여, 본 실험에서는 위 데이터의 각 도메인에서 2만 문장쌍씩 추출하였다. 그리고 이들 데이터 중, 16,000개씩은 훈련데이터로, 그리고 2,000개씩은 각각 검증과 테스트 데이터로 활용하였다. AIhub말뭉치 내의 데이터 중, APE를 위해 추출된 데이터를 제외한 나머지 데이터는 Transformer 기반 번역모델의 학습에 활용되었다.

현재 전문인력을 통해 생성된 APE 데이터가 존재하지 않기 때문에, 생성한 APE 모델의 성능을 평가하기 위해서, 우리는 여러 외부 번역 시스템을 통해 테스트 데이터를 생성하였다. 이는 테스트 데이터로 정한 병렬 말뭉치의 원문과 타겟 문장을 각각 APE 데이터의 원문, 교정문으로 간주하고, 번역 시스템을 통해 원문을 번역하여 APE 데이터의 번역문을 생성하는 방법이다.

이를 통해 외부 번역 시스템을 통해 생성한 문장이, 본 연구에서 제안한 APE 모델을 거침으로써 어떤 교정 효과를 볼 수 있는지 확인할 수 있다. 해당 실험은 실제 서비스 관점에서, 이 연구가 제안하는 모델이 실제적인 성과를 낼 수 있는지 확인하는 지표로 활용될 수 있으며, 본 연구에서는 정량적 분석과 더불어 정성적 분석으로 해당 실험결과를 분석한다. 정량적 분석은 Translation Edit Rate(TER)[24]과 BLEU[25]를 통해 확인한다. TER과 BLEU는 현재 WMT를 비롯한 대부분의 APE

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

Table 1. Quantitative Analysis of each APE model

APE Model	Before Editing		mBART		mBART Translation		Transformer Translation	
	TER ↓	BLEU ↑	TER ↓	BLEU ↑	TER ↓	BLEU ↑	TER ↓	BLEU ↑
Google	52.983	34.49	48.379	39.68	46.636	41.86	45.402	43.81
Microsoft	59.478	26.49	53.993	33.46	51.828	36.21	45.677	43.52
Amazon	60.421	23.32	52.958	33.02	49.977	37.14	45.580	43.62

연구들에서 모델 성능 평가를 위해 활용하는 대표적인 평가 지표이다[4]. 특히 TER은 번역문 내의 오류를 수정하기 위해서 총 몇 번의 교정이 더 필요한지를 정량화한 값으로, APE 모델 성능 평가의 중심 지표이다. 본 실험에서는 해당 지표를 확인하기 위한 소프트웨어로 각각 Tercom에서 제공한 TER측정 도구[2]와, Moses에서 제공하는 BLEU측정 도구[3]를 활용한다.

4.2 사전학습 언어모델

본 연구에서는 mBART와 transformer를 기반으로 한 APE 모델을 설계하였다. mBART는 Huggingface[26]에서 제공하는 모델 구조를 활용하였다. 이는 25만 개의 vocab size를 가지고 있고, 은닉층의 차원이 1,024인 인코더, 디코더 층 각각 12개로 구성되어있다. Transformer 기반 번역모델의 경우 은닉층의 차원이 512인 인코더, 디코더층 6개로 이루어진 기본 모델 구조를 활용하였다. Transformer 번역모델에서 토크나이저(Tokenizer)는 병렬 말뭉치 내의 한국어 문장들과 영어 문장들을 한데 합친 말뭉치를 생성한 이후, 해당 말뭉치를 통해 센텐스피스(Sentencepiece)[27]모델을 학습하는 방법을 활용하였다[28]. 해당 토크나이저의 단어 개수는 5만 개로 설정하였다.

효율적인 학습을 위하여 본 연구에서는 사전학습 언어모델로부터 APE를 미세조정 할 때, 모델 구조 안에 적응층(Adapter Layer)를 추가하여 학습하는 방법론을 적용하였다[8,29]. 본 연구에서의 모든 실험은 RTX 8000 4대를 통해 진행되었다.

4.3 정량적 분석

본 실험에서는, 한국어-영어 APE에서 최적의 성능을 내는 접근방법을 확인하기 위하여 여러 APE 모델들의 성능을 비교 분석한다. 먼저 다국어 사전학습 언어모델인 mBART을 기반으로 전이학습 하여 APE 모델을 생성하고, Transformer기반 번역모델에 전이 학습하여 APE 모델을 생성한다. 그리고 이에 더하여, Transformer 기반 번역모델 학습 시 활용했던 데이터와 동일한 데이터로 mBART에 번역을 학습을 진행하고, 이를 기반으로 한 APE 모델을 설계한다. 이들 모델 간의 비교 실험을 통해 검증하고자 하는 사항은 다음 두 가지로 나눌 수 있다. 첫 번째로, 저 자원 언어에 속하는 한국어-영어 언어쌍에 대해서도, 번역모델을 사전학습 모델로 활용하여 APE 모델을 전이 학습시키는 것이 유의미한 성과를 내는지 확인한다. 그리고 두 번째로, 전이학습 기반으로 APE 모델을 생성하는 경우, 최적의 성능을 내기 위해 선택해야 하는 모델 구조를 확인한다. 실험결과는 Table 1과 같다.

실험결과를 통해 확인할 수 있듯, 세 가지 테스트 데이터에서 모두, Transformer 번역모델을 기반으로 한 APE 모델이 가장 좋은 성능을 보여주었으며, mBART에 번역을 한차례 학습한 이후 APE 미세조정을 진행한 모델이, mBART에 APE를 곧바로 미세조정한 모델에서보다 더 뛰어난 성능을 보였다. 해당 실험결과는 다음 두 가지 관점에서 해석할 수 있다. 첫 번째로, 전문인력의 교정을 통해 생성된 APE 데이터 없이도 준수한 성능의 APE 모델을 생성할 수 있다는 것을 확인할 수 있다. 이는 저 자원 언어에 대해서도, 병렬 말뭉치로부터 APE 데이터를 합성하는 방식의 noising scheme 기반 데이터 생성방법이 매우 효과적이라는 것을 보여주며, 공식적인 APE 데이터가 없는 상황에서도 해당 방법론을 통해 유의미한 성능의 APE 모델을 생성할 수 있음을 보여준다. 두 번째로, 저 자원 언어에 대해서도 번역 시스템을 사전학습 모델로 활용하는 것은 APE에서 매우 효과적이라는 것을 확인할 수 있다. 이는 Transformer 기반 번역모델 기반의 APE 모델이, mBART 기반 모델보다 더 뛰어난 성능을 보였으며, mBART에 APE를 곧바로 미세조정하는 것보다, 번역을 한차례 학습한 경우 더 뛰어난 성능을 보인다는 것을 통해 유추할 수 있다.

2) http://www.cs.umd.edu/~snover/tercom/

3) https://github.com/moses-smt/mosesdecoder

Table 2. Representative examples of each APE model

Source	아우내 전통 고추장은 1990년대 후반 죽계리 부녀회원들이 만들기 시작했다.			
Reference	Aunae Traditional Gochujang was started to be produced from the late 1990s by the members of the Jukgye-ri Women's Association			
	Before Editing	mBART	mBART translation	Transformer Translation
Google	Aouna traditional gocujang began to be made by young women in the late 1990s.	Aunae traditional gochujang began making by women members of Jukgye-ri in the late 1990s.	Aunae traditional gochujang began to be made by women members of Jukgye-ri in the late 1990s.	Aunae Traditional Gochujang was started by the members of Jukgye-ri Women's Association in the late 1990s.
Microsoft	The traditional red pepper field was started by members of the bamboo family in the late 1990s.	The traditional red pepper field was started by members of the bamboo family in the late 1990s.	Aunae's traditional red pepper paste was started to be made by women members of Jukgye-ri in the late 1990s.	The traditional Aunae red pepper paste was produced by the members of Jukgye-ri Women's Association in the late 1990s.
Amazon	Aouna traditional gocujang began to be made by young women in the late 1990s.	Aouna Traditional Gochujang began to be made by young women members in the late	Aouna traditional gochujang began to be made by female members of Jukgye-ri in the late 1990s.	Aunae Traditional Gochujang was started by members of Jukgye-ri Women's Association in the late 1990s.

번역모델을 기반으로 한 APE 모델의 전이학습 방법이 mBART와 같은 다국어 사전학습 언어모델을 활용한 방법에서보다 더 뛰어난 성과를 얻을 수 있다는 것은 훈련 효율성 측면에서도 유의미한 결과이다. 본 실험에서 활용한 transformer기반 번역모델의 총 파라미터 수는 약 9,800만개로, mBART의 총 파라미터 수인 6.1억개와 비교했을 때 1/6수준에 그친다. 즉, 더 작은 크기의 모델을 활용하더라도 그보다 뛰어난 APE 성능을 낼 수 있음을 보여주며, 이는 더 적은 GPU 자원으로도 훈련을 진행시킬 수 있음을 보여준다. 즉, 병렬 말뭉치의 양이 상대적으로 적은 저 자원 언어쌍에서도, mBART와 같은 대용량 사전학습 모델을 도입할 필요 없이, 상대적으로 작은 크기의 transformer기반 번역모델에 APE를 전이 학습함으로써 뛰어난 고성능의 APE 모델을 생성할 수 있음을 보여준다. 이는 향후 전이학습 기반으로 APE 모델을 생성하는 경우, 사전학습 언어모델 선택 전략에 대한 가이드라인으로 활용될 수 있다.

4.4 정성적 분석

Transformer 번역모델을 기반으로 한 APE의 경우, 기존 번역문보다 더 뛰어난 품질의 교정 결과를 얻을 수 있었으나, 세 가지 번역문에 대한 교정에서 모두 정량적으로 유사한 결과를 도출하였다. 예를 들어, 구글 번역기를 통한 번역 결과와 아마존 번역기를 통한 번역 결과는, 교정 이전의 품질을 비교했을 때, TER 7.438점, BLEU 11.17점의 차이가 존재하였으나, 이들을 교정한 결과물의 품질 차이는 TER 0.178점, BLEU 0.19점 차이로 줄어들었다. 이는 번역 시스템을 기반으로 생성한 APE 모델이, 번역문의 품질을 고려하지 않고 원문만을 고려하게 되는 것이 아닌지에 대한 우려를 남긴다. 즉, 정량적 분석 결과만을 확인한다면, APE 모델이 원문과 번역문을 모두 고려하여 번역문을 생성하는지, 혹은 번역문과는 관계없이 원문만을 고려하여 교정문을 생성하는지 확신하기 어렵다. 이에 본 실험에서는 각 번역문에 대한 교정 결과를 정성적으로 분석하여 APE 모델의 실제적인 성능을 확인하였다. 실제 교정 결과는 Table 2와 같다. Table 2에서 적색으로 표시된 단어들은, 교정 이전 번역문에서 청색으로 표시된 단어들이 각 APE 모델들을 통해 교정된 결과를 의미한다.

실제 교정 결과를 통해 확인할 수 있듯, Transformer 번역모델 기반으로 생성한 APE 모델에서의 교정 결과물들은, 정량적 분석에서는 서로 유사한 모습을 보여주었으나, 실제 교정 결과는 기존 번역문의 특성을 반영하는 형태로 출력되는 것을 확인할 수 있다. 예를 들어 실제 교정 결과를 통해 확인할 수 있듯, Google 번역 시스템이나 Amazon 번역 시스템에서는 “고추장”이라는 단어를 고유 명사 “gocujang”으로 번역하는 스펠링 오류를 보여줬고, Microsoft 번역 시스템에서는 “red pepper field”로 번역하는 동의어 해석 오류를 보여주었다. 그리고 이러한 오류들은 Transformer 번역모델 기반 APE 모델을 통해 각각 “Gochujang”과 “red pepper paste”로 교정되었다. 이는 번역모델을 기반으로 생성된 APE 모델이, 입력으로 주어지는 원문만을 활용하여 교정문을 생성하는 것이 아니라 번역문 내의 오류를 포

착하여 수정하는 방향으로 교정을 진행하는 것이라 유추할 수 있다. 즉, 번역모델을 기반으로 생성한 APE 모델은 정량적으로 우수한 성능을 낼뿐 아니라, 실제적인 교정 품질도 우수하다는 것을 보여준다.

5. 결론

본 연구에서는 전문가의 교정작업을 통해 생성한 APE 데이터가 없는 상황에서, 병렬 말뭉치만으로 APE 모델을 훈련할 수 있는 노이즈 기반 데이터 생성방법을 적용하였고, 이를 통해 준수한 성능의 APE 모델을 생성할 수 있음을 확인하였다. 또한, APE 모델 생성을 위하여 선택할 수 있는 두 가지 전이학습 접근방법에 대해 실험하고, 한국어-영어와 같은 저 자원 언어쌍에서 최적의 모델 훈련 방법을 확인하였다. 특히, APE 모델 생성에 있어, 사전에 번역 작업을 학습시키는 것이 APE 성능 향상에 매우 효과적이라는 것을 실험을 통해 확인했으며, 이러한 방법으로 생성한 APE 모델은 정량적으로나 정성적으로 매우 우수한 교정 성능을 얻을 수 있음을 보였다. 우리는 향후 APE 모델 성능을 향상시키기 위해 적용할 수 있는, 다양한 APE 데이터 자동 생성방법을 연구할 예정이다[30].

REFERENCES

[1] C. Park & H. Lim. (2020). Automatic Post Editing Research. *Journal of the Korea Convergence Society, 11(5),* 1-8.
DOI : 10.15207/JKCS.2020.11.5.001

[2] S. Pal, N. Herbig, A. Krüger & J. van Genabith. (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).*

[3] P. Isabelle, C. Goutte & M. Simard. (2007). Domain adaptation of MT systems through automatic post-editing. *MT Summit XI,* 102.

[4] S. Chollampatt, R. H. Susanto, L. Tan & E. Szymanska. (2020). Can Automatic Post-Editing Improve NMT?. *arXiv preprint arXiv:2009.14395.*

[5] R Chatterjee, M. Freitag, M. Negri & M. Turchi. (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).*

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

[7] J. Lee, W. Lee, J. Shin, B. Jung, Y. G. Kim & J. H. Lee. (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).*

[8] H. Yang et al. (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] B. Zoph, D. Yuret, J. May & K. Knight. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201.*

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

[11] A. Vaswani et al. (2017). Attention is all you need. *In Advances in neural information processing systems (pp. 5998-6008).*

[12] M. Negri, M. Turchi, R. Chatterjee & N. Bertoldi. (2018). ESCAPE: a large-scale synthetic corpus for automatic post-editing. *arXiv preprint arXiv:1803.07274.*

[13] W. Lee, B. Jung, J. Shin & J. H. Lee. (2021, April). Adaptation of Back-translation to Automatic Post-Editing for Synthetic Data Generation. *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 3685-3691).*

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

[15] R. Sennrich, B. Haddow & A. Birch. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709.*

[16] J. Lim, H. Moon, C. Lee, C. Woo & H. Lim. (2021). An Automated Industry and Occupation Coding

System using Deep Learning. *Journal of the Korea Convergence Society, 12(4), 23-30.*

[17] S. Eo, C. Park, H. Moon, J. Seo & H. Lim. (2021). Comparative Analysis of Current Approaches to Quality Estimation for Neural Machine Translation. *Applied Sciences, 11(14), 6584.*

[18] 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.*

[19] A. Conneau & G. Lample. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems, 32, 7059-7069.*

[20] I. Sutskever, O. Vinyals & Q. V. Le. (2014). Sequence to sequence learning with neural networks. *In Advances in neural information processing systems (pp. 3104-3112).*

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

[22] H. Moon, C. Park, S. Eo, J. Park & H. Lim. (2021). Filter-mBART Based Neural Machine Translation Using Parallel Corpus Filtering. *Journal of the Korea Convergence Society, 12(5),* 1-7. DOI : 10.15207/JKCS.2021.12.5.001

[23] 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

[24] M. Snover, B. Dorr, R. Schwartz, L. Micciulla & J. Makhoul. (2006). A study of translation edit rate with targeted human annotation. *In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers (pp. 223-231).*

[25] 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).*

[26] T. Wolf et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771.*

[27] T. Kudo & J Richardson. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226.*

[28] C. Park, S. Eo, H. Moon & H. S. Lim. (2021, June). Should we find another model?: Improving Neural Machine Translation Performance with ONE-Piece Tokenization Method without Model Modification. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers (pp. 97-104).*

[29] N. Houlsby et al. (2019, May). Parameter-efficient transfer learning for NLP. *In International Conference on Machine Learning (pp. 2790-2799). PMLR.*

[30] C. Park, J. Seo, S. Lee, C. Lee, H. Moon, S. Eo. & H. Lim. (2021). BTS: Back TranScription for Speech-to-Text Post-Processor using Text-to-Speech-to-Text *In Proceedings of the 8th Workshop on Asian Translation (pp.106-116)*

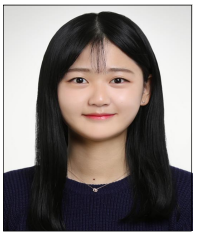
문 현 석(Hyeonseok Moon) [학생회원]

· 2021년 2월 : 고려대학교 수학과 및 인공지능학과(이학사, 공학사)
· 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 : Data-Centric AI, 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

서 재 형(Jaehyung Seo) [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(Heuiseok Lim) [종신회원]

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