Journal of The Korea Convergence Society
Vol. 12. No. 11, pp. 35-43, 2021

I-SSN 2233-4890 / e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2021.12.11.035

Zero-Shot 기반 기계번역 품질 예측 연구

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

Study on Zero-shot based Quality Estimation

Sugyeong Eo[1], Chanjun Park[1], Jaehyung Seo[1], Hyeonseok Moon[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 최근 다언어모델(Cross-lingual language model)을 활용하여 한 번도 보지 못한 특정 언어의 하위 태스크를 수행하는 제로샷 교차언어 전이(Zero-shot cross-lingual transfer)에 대한 관심이 증가하고 있다. 본 논문은 기계번역 품질 예측(Quality Estimation, QE)을 학습하기 위한 데이터 구축적 측면에서의 한계점을 지적하고, 데이터를 구축하기 어려운 상황에서도 QE를 수행할 수 있도록 제로샷 교차언어 전이를 수행한다. QE에서 제로샷을 다룬 연구는 드물며, 본 논문에서는 교차언어모델을 활용하여 영어-독일어 QE 데이터에 대해 미세조정을 실시한 후 다른 언어쌍으로의 제로샷 전이를 진행했고 이 과정에서 다양한 다언어모델을 활용하여 비교 연구를 수행했다. 또한 다양한 자원 크기로 구성된 언어쌍에 대해 제로샷 실험을 진행하고 실험 결과에 대해 언어별 언어학적 특성 관점으로의 분석을 수행하였다. 실험결과 multilingual BART와 multillingual BERT에서 가장 높은 성능을 보였으며, 특정 언어쌍에 대해 QE 학습을 전혀 진행하지 않은 상황에서도 QE를 수행할 수 있도록 유도하였다.

주제어 : 기계번역 품질 예측, 인공신경망 기계번역, 제로샷, 언어 융합, 자연언어처리

Abstract Recently, there has been a growing interest in zero-shot cross-lingual transfer, which leverages cross-lingual language models (CLLMs) to perform downstream tasks that are not trained in a specific language. In this paper, we point out the limitations of the data-centric aspect of quality estimation (QE), and perform zero-shot cross-lingual transfer even in environments where it is difficult to construct QE data. Few studies have dealt with zero-shots in QE, and after fine-tuning the English-German QE dataset, we perform zero-shot transfer leveraging CLLMs. We conduct comparative analysis between various CLLMs. We also perform zero-shot transfer on language pairs with different sized resources and analyze results based on the linguistic characteristics of each language. Experimental results showed the highest performance in multilingual BART and multillingual BERT, and we induced QE to be performed even when QE learning for a specific language pair was not performed at all.

Key Words : Quality estimation, Neural machine translation, Zero-shot, Language convergence, Natural language processing

*"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)" and this research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).
*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received August 10, 2021 Revised September 2, 2021
Accepted November 20, 2021 Published November 28, 2021

1. 서론

최근 여러 언어로 구성된 말뭉치를 활용하여 전반적인 언어 지식에 대해 사전학습(pre-training)한 다언어 모델(Cross-lingual language models, CLLMs)을 하위 태스크 데이터에 대해 미세조정(fine-tuning)함으로써 여러 언어로 구성된 자연어처리(Natural Language Processing, NLP)의 태스크들을 수행하는 연구가 활발하게 진행되고 있다[1,2]. 동시에 다언어모델을 활용하여 제로샷 교차 언어 전이(Zero-shot cross-lingual transfer)를 진행하는 연구들도 활발히 수행되고 있다[3,4].

제로샷 교차 언어 전이란 한 언어로 된 특정 태스크의 데이터에 대해 미세조정한 후 다른 언어로 구성된 동일한 태스크의 데이터에 대해 평가를 진행하는 것을 말한다. 다언어모델은 학습한 25개, 100개 등의 언어에 대한 지식을 가지고 있으며 한 언어로 구성된 NLP 하위 태스크를 풀어내는 지식도 각각 가지고 있지만, 다른 언어로 구성된 동일 하위 태스크를 풀어내는 지식은 학습하지 않았기 때문에 제로샷이라고 일컫는다. 제로샷 세팅에서는 한 언어 또는 한 언어쌍에 대해서만 학습을 진행했음에도 다른 언어에 대한 태스크들을 수행할 수 있다는 데 큰 장점을 지니고 있다.

기계번역 품질 예측(Quality Estimation, QE)은 정답 번역문 없이 소스 문장과 기계번역 문장만을 활용하여 기계번역 결과에 대한 품질을 예측할 수 있다는 장점을 지닌 태스크이다[5]. BLEU[6], METEOR[7] 등과 같은 성능 측정 지표들과는 달리 정답 번역문을 참조하지 않아도 된다는 장점을 지니지만 [8]에서 언급한 바와 같이 QE를 학습하기 위한 데이터 자체를 구축하는 과정이 오히려 병렬 코퍼스를 구축하는 것보다 더 많은 전문성 및 시간, 비용을 요구한다는 한계점이 있다. 특히나 저자원 언어에서는 이로 인해 QE를 수행하는 것 자체가 제한된다.

본 논문에서는 먼저 QE의 한계점을 보완하면서 동시에 고자원 언어에서 대용량으로 학습한 언어 지식을 교차 언어로 전이함으로써 불특정 언어, 특히나 QE 데이터를 구축하기 어려운 언어에 대해서도 QE를 진행할 수 있도록 교차언어 제로샷 전이를 수행했다. 본 논문에서는 사전학습된 다언어모델에 대해 문장 단위의 QE 태스크를 학습할 수 있도록 고자원 언어인 영어-독일어 데이터셋에 대해 미세조정을 실시했다. 이후 영어-독일어 QE 모델에 대해 고자원 언어인 중국어, 중자원 언어인 루마니아어, 에스토니아어, 저자원 언어인 싱할라어 및 네팔어로의 제로샷 전이를 수행하였다. 다언어모델은 사전학습 전략 및 언어 개수, 모델 수용력 등에 따라 다양하게 나뉘기 때문에 Cross-lingual language model(XLM)[9], XLM-RoBERTa[10], multilingual BERT(mBERT)[11], multilingual BART(mBART)[12]와 같이 다양한 다언어모델을 활용하여 제로샷 연구를 수행하고 각 모델별 성능을 비교 분석한다. 또한, 언어별 제로샷 성능에 대해 어족, 어순 및 알파벳 등 다양한 관점에서 비교함으로써 제로샷 전이에서 특정 언어쌍의 성능이 높거나 낮게 나온 원인에 대해 분석한다.

본 연구의 두 가지 기여는 다음과 같다. 첫째, 본 연구에서는 QE가 가진 데이터 구축적 한계를 지적하고, 제로샷 교차 언어 전이를 수행함으로써 소스 문장과 기계번역 문장을 포함하는 불특정 언어쌍에 대해 기계번역 결과에 대한 품질을 예측할 수 있도록 했다. 둘째, 다언어모델별 및 언어별 비교실험 및 다양한 관점으로의 분석을 통해 최적의 제로샷 교차 언어 전이를 수행하도록 유도했다.

본 논문은 2장에서 제로샷 및 QE에 대한 관련 연구들을, 3장에서 제로샷 전이 방법에 대해 설명한다. 4장에서는 본 연구에서 진행하는 태스크, 활용한 데이터 및 다언어모델에 대한 설명과 함께 실험 결과에 대해 언급하며 5장에서 언어 관점에서 결과를 분석하는 것으로 구성한다.

2. 관련 연구

2.1 제로샷 전이 학습(Zero-shot transfer learning)

2.1.1 태스크 설명

특정 하위 태스크를 수행하도록 미세조정한 학습 내용과는 다르게, 제로샷 전이의 경우 언어모델이 사전학습 및 미세조정 시 학습한 내용들을 토대로 한 번도 보지 못한 다른 태스크를 풀어내거나 다른 언어로 구성된 동일한 태스크를 풀어내게 된다. 제로샷은 모델이 이전에 보지 못했던 것들에 대해 기존에 학습한 내용들을 얼마나 잘 활용하여 풀어내는지를 확인한다는 점에서 근본적인 모델의 능력을 확인할 수 있다는 장점을 지닌다.

2.1.2 연구 동향

최근 다언어모델을 활용하여 하위 태스크에서 한

언어에 대해 미세조정한 후 학습하지 않은 다른 언어로 해당 태스크를 수행하는 방향의 제로샷 연구가 활발히 진행되고 있다[3,4,13]. 언어모델이 한 번도 학습하지 않았던 도메인 또는 태스크를 수행하는 방향으로의 제로샷도 연구되고 있다[14,15]. 특히나 [16]에서는 단어 얼라인 및 생성 스코어와 같은 교차언어 패턴을 학습한 모델에 대해 QE를 진행한다. 또한 [17]과 같이 자연어 이해 태스크가 아닌 자연어 생성 태스크에서도 사전학습된 다언어모델을 활용하여 제로샷을 수행하는 경우도 존재한다. 본 논문에서는 QE에 대해 영어-독일어 쌍으로 미세조정한 후 이 모델을 활용하여 다른 언어의 QE를 수행하는 연구를 진행한다. 이러한 방향의 제로샷 연구는 기존에 진행된 바가 없으며, 본 연구에서 처음으로 시도되었다.

2.2 기계번역 품질 예측 (Quality Estimation)

2.2.1 태스크 설명

QE는 정답 번역문 없이 소스 문장과 기계번역 문장을 참고하여 기계번역 결과에 대한 품질을 예측하는 태스크이다. 기계번역 결과에 대한 품질은 숫자, OK/BAD 태그, minor/major/critical 태그 등으로 나타내기 때문에 번역하는 언어에 대해 생소한 사용자의 경우 기계번역 결과에 대한 신뢰도를 쉽게 파악할 수 있다는 특징을 지닌다. 문서, 문장, 단어 등 다양한 수준에서 품질 주석을 활용한 품질 예측이 가능하며, 다양한 방면에서 활용될 수 있다[18,19]. 기계번역 분야의 워크샵인 Conference on Machine Translation(WMT)에서는 매년 QE 태스크를 개최하고 있으며, 이를 기반으로 QE 태스크에 대한 다수의 연구가 수행 중에 있다.

2.2.2 연구 동향

QE 분야의 연구는 크게 딥러닝 모델 등장 이전의 통계기반 학습[20,21], 딥러닝 모델 등장 이후의 RNN 및 LSTM을 이용한 학습[22,23], 언어모델 등장 이후의 사전학습 및 미세조정 학습[24,25]으로 나눌 수 있다. 최근에는 다언어모델을 활용하는 경우가 대다수를 이루고 있으나, [26]에서는 Transformer 구조를 활용하여 높은 성능을 이루어내기도 하였다. [27]는 QE를 수행하기 위해 가장 적합한 언어모델이 무엇인지에 대한 비교 연구를 수행하였으며, 기존에 활용되지 않았던 인코더-디코더 구조를 지닌 multilingual BART 모델을 추가적으로 활용하였다. [8]에서는 QE의 한계를 지적하고 슈도 데이터 제작 방법론을 활용하여 데이터 문제를 다루어 저자원 언어에서도 QE를 진행할 수 있도록 유도했다.

본 논문은 문장 단위의 QE를 수행하였으며, 다양한 다언어모델의 제로샷 성능을 비교하였다.

3. Zero-Shot기반 기계번역 품질예측 모델

사전학습된 다언어모델을 활용한 제로샷 전이는 Fig. 1과 같이 수행된다. QE 태스크를 학습시키기 위한 미세조정 과정에서는 가장 먼저 사전학습된 다언어모델을 로드하여 파라미터들을 초기화한다. 미세조정을 위한 QE 데이터셋은 모델별 미리 학습된 토크나이저를 활용하여 서브 워드 단위로 분절하며, 분절된 단어들에 대해 모델별 단어사전을 활용하여 숫자화한다. 또한 QE의 경우 소스 문장과 기계번역 결과를 연결하여 동시에 입력으로 넣어주게 되는데, 두 문장 사이에 [SEP] 토큰을 추가하고 문장의 가장 앞머리에 문장 전체에 대한 정보를 담을 [CLS] 토큰도 함께 추가한다. 두 문장에 대한 구분을 위한 분절 임베딩(segment embedding)도 함께 입력으로 넣어주게 되는데, 각 문장은 숫자로 구분된다. XLM-R, mBERT, mBART 모델은 이와 같이 입력을 구성하며, XLM 모델의 경우 추가적으로 각 문장을 구분할 때 언어마다 주어진 인덱스를 활용하여 분절 임베딩을 구성하게 된다. 추가적으로 XLM에서는 단어에 대한 순서 정보도 함께 다언어모델의 입력으로 넣어주는데, 이는 토큰화된 문장에 대해 각 토큰에 해당하는 인덱스로 구성된다.

위와 같이 입력 임베딩을 구성한 후 각 다언어모델에 입력으로 넣어주고 마지막 은닉층의 은닉값 중 문장 전체에 대한 정보를 포함하는 [CLS] 토큰 위치에 해당하는 값을 뽑아낸다. 이후 이 값을 선형 분류기(linear classifier)를 활용해 하나의 값으로 변환하고, 전체 문장에 대한 최종 품질 값을 예측한다. 예측한 값과 정답값 간은 평균 제곱 오차를 활용하여 예측 값이 정답에 가까울 수 있도록 학습을 진행한다. 이렇게 미세조정을 진행한 모델 중 검증 데이터에 대한 손실값이 가장 낮은 체크포인트를 저장하고 제로샷 전이를 위한 미세조정

모델로 활용한다. 제로샷 전이 시에는 위 체크포인트를 활용하여 파라미터를 초기화하며, QE 데이터가 없는 상황을 고려하여 추가적인 학습 과정은 생략한다. 테스트할 언어쌍의 소스 문장과 기계번역 결과를 활용하여 데이터 입력을 구성하는데, 이 과정은 미세조정 시와 동일하게 진행된다. 모델에 입력을 구성하여 넣어주고 출력 값을 예측하는 과정 역시 미세조정 시와 동일하게 진행이 되며, 제로샷 전이 성능을 확인할 수 있도록 모델 예측 값과 평가 데이터의 정답 값 간의 피어슨 상관계수(Pearson correlation coefficient)를 측정한다.

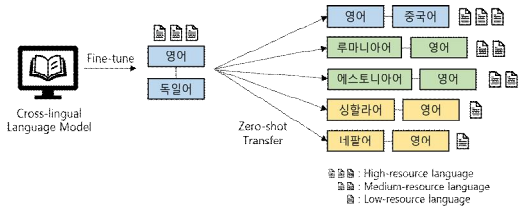

Fig. 1. Overall architecture of zero-shot transfer

4. 실험

4.1 데이터

본 논문은 WMT20의 QE 태스크 중 문장 단위 직접 평가(Sentence-level direct assessment) 하위 태스크에 해당하는 데이터를 활용하여 실험했다. 해당 하위 태스크는 기계번역 문장의 전반적인 품질을 숫자화하여 보여주게 된다.

모든 언어쌍에 대해 7천 개의 학습 데이터, 1천 개의 검증 데이터, 1천 개의 평가 데이터로 구성되어 있다. 본 논문에서는 영어-독일어 언어쌍의 미세조정을 위해 학습, 검증 및 평가 데이터를 모두 활용하였으며, 제로샷 교차언어 전이를 수행하는 언어쌍에 대해서는 1천 개의 평가 데이터만을 활용하였다. 데이터는 소스 문장과 기계번역 문장, 그리고 각 번역 전문가가 평가한 DA 스코어와 그 평균, 그리고 z-표준화된 DA 스코어와 그 평균으로 구성되어 있다. 이 때 모델이 예측해야 하는 값은 평균 z-표준화 DA 스코어이며, 여기서 DA 스코어란 적어도 3명의 번역 전문가가 문장의 품질에 대해 0점에서 100점 사이로 직접 평가한 점수이다. 본 논문은 소스 문장과 기계번역 결과, 평균 z-표준화 DA 스코어만을 학습에 활용하였다. 최종 예측값과 정답 값 간의 평가 매트릭은 피어슨 상관 계수를 활용하였다.

4.2 다언어모델

본 논문에서는 영어 및 독일어를 포함하여 사전학습을 진행한 다언어모델을 활용하였다. 사전학습된 다언어모델은 HuggingFace[28]에서 배포하는 모델을 활용하였으며, 각 모델별 설명은 아래와 같다.

- XLM-R-base: 100개의 언어로 구성된 CommonCrawl 코퍼스에 대해 문장 중 일부 토큰을 마스킹하고 그 토큰을 예측하도록 하는 MLM 노이징 전략을 활용하여 사전학습한 모델이며, 220M 개의 파라미터로 구성되어 있다.
- XLM-R-large: XLM-R-base 모델과 동일하게 학습을 진행했으나 large 모델에서는 수용력(capacity)을 확장하여 학습했으며, 파라미터 수는 550M이다.
- mBERT: 100개의 언어로 구성된 Wikipedia 코퍼스에 대해 MLM 및 문장 순서를 임의로 변경한 후 순서가 옳은지를 예측하도록 하는 NSP 노이징 전략을 활용한 사전학습 모델이며, 110M 개의 파라미터로 구성되어 있다.
- mBART: mBART는 25개의 언어로 구성된 CommonCrawl 데이터에 대해 문장의 여러 단어를 마스킹한 후 이를 예측하는 text infilling, 문장의 순서 일부를 섞는 sentence permutation 노이징 전략을 활용하여 사전학습하였다. 610M 개의 파라미터로 학습했다.
- XLM-CLM: 문장 내에서 이전의 단어들을 보고 다음 단어를 예측하도록 사전학습한 causal language model 이며, 영어 및 독일어의 두가지 언어에 대해 사전학습하였다.
- XLM-MLM: MLM 노이징 전략만을 활용한 모델이며 영어 및 독일어의 두가지 언어에 대해 사전학습을 진행하였다.
- XLM-MLM-17: XLM-MLM과 동일한 방식으로 지행되었으나, 언어를 17개로 확장하였다.
- XLM-MLM-100: XLM-MLM과 동일한 방식으로 지행되었으나, 언어를 100개로 확장하였다.
- XLM-TLM: 언어가 다른 두 문장을 연결하고, 연결된 문장 중 일부를 마스킹한 후 이를 예측하도록 사전학습한 모델이다. 한 언어의 마스킹된 단어를 예측할 때 연결된 다른 언어의 문장도 함께 고려하기 때문에 얼라인 정보를 학습할 수 있도록 유도되었다.

Table 1. Comparison of zero-shot transfer performance of various cross-lingual language models for each language pair. The Pearson correlation coefficient for EN-DE is the result of fine-tuning, not zero-shot

Model	EN-DE [27]	EN-ZH	ET-EN	NE-EN	SI-EN	RO-EN
XLM-R-base [10]	0.328	0.124	0.109	0.210	0.145	0.185
XLM-R-large [10]	0.298	0.060	0.180	0.239	0.196	0.255
mBERT [11]	0.382	0.150	**0.309**	0.239	0.245	**0.414**
mBART [12]	0.351	**0.207**	0.290	**0.300**	**0.384**	0.321
XLM-CLM [9]	0.253	0.023	0.062	0.029	0.004	0.178
XLM-MLM [9]	0.206	0.093	0.046	0.003	0.105	0.121
XLM-MLM17 [9]	0.253	0.146	0.132	0.174	0.136	0.236
XLM-MLM100 [9]	0.232	0.100	0.214	0.179	0.157	0.245
XLM-TLM [9]	**0.394**	0.076	0.192	0.127	0.172	0.137

4.3 실험 결과

4.3.1 다언어모델별 성능 비교

본 논문에서는 제로샷을 위한 미세조정 학습 과정에서 무작위로 초기화된 값들에 의한 성능 변화의 영향을 줄이고자 동일한 모델에 대해 각각 다른 무작위 값을 활용하여 총 다섯 번을 실험하였다. 미세조정 시 각 체크포인트를 저장하여 총 다섯 번의 제로샷 실험을 수행하였으며 각 결과에 대한 최저, 최대, 평균값을 추출하였다. 다언어모델별 각 언어쌍 마다 제로샷 전이를 진행한 결과는 Table 1과 같다.

실험 결과, 제로샷 전이에서 가장 좋은 성능을 낸 모델은 mBART와 mBERT였다. mBART 모델은 제로샷 실험을 진행한 모든 언어에 대해 사전학습을 통해 지식을 가지고 있으며, 다른 모델들과는 차별적으로 text infilling, sentence permutation이라는 추가적인 노이징 전략을 활용하였는데 이러한 요소가 다언어모델의 전반적인 언어 이해를 높였으며, 제로샷 교차언어 전이에서도 좋은 성능을 낼 수 있었던 요인으로 분석할 수 있다.

Table 1의 1열은 영어-독일어 언어쌍에 대한 미세조정 성능인데, [27]에 따르면 두 언어 간의 정보를 학습했다는 점에서 XLM-TLM이 가장 높은 성능을 보였으나 이 모델은 본 실험에서 중국어를 제외한 나머지 언어에 대해 학습하지 않았기 때문에 성능이 낮게 나왔다고 해석할 수 있다. 마찬가지로 XLM-CLM과 XLM-MLM의 경우 역시 사전학습을 영어와 독일어에 대해서만 수행했기 때문에 태스크에 대한 이해는 있지만 영어, 독일어를 제외한 나머지 언어는 완전히 처음 보았기 때문에 성능이 매우 저조함을 알 수 있다.

중국어 외의 다른 언어는 사전학습하지 않은 XLM-MLM17과 모든 언어에 대한 사전학습 지식이 있는 XLM-MLM100의 결과를 비교해보면, 중국어의 경우는 17개의 언어로 사전학습한 모델이 성능이 더 높았다. 각 언어에 대한 지식이 있는 XLM-MLM100 모델의 경우 중국어를 제외한 나머지 언어에서 모두 성능이 더 높았다. 이를 통해 언어에 대한 사전 이해가 있는 경우 제로샷 성능이 높아지지만, [10]에서 언급한 다중언어의 저주(curse of multilinguality) 문제로 인해 언어의 개수가 늘어날수록 고자원 언어인 중국어는 성능이 낮아졌다고 분석할 수 있다.

또한 두 개의 언어로 학습한 XLM-MLM과 비교했을 때 XLM-MLM17의 성능이 뚜렷하게 더 높은 것을 확인할 수 있다. 이는 비록 XLM-MLM17 모델이 중국어를 제외한 나머지 언어들에 대한 지식이 없다 하더라도, 17개의 언어들 중 본 논문에서 실험한 언어들과 어족이나 알파벳이 유사한 경우 제로샷을 진행했을 때 유사 언어에 대한 정보를 활용하여 성능에 긍정적인 영향을 주었다고 분석할 수 있다.

XLM-R의 경우 중국어를 제외하면 base 모델보다 large 모델이 일관적으로 더 높은 성능을 보였다. 따라서 모든 환경은 동일하게 하되 파라미터 수를 변경하였을 때 성능 변화가 있는 것으로 보아, 파라미터 수가 성능에 영향을 미치는 요소라고 할 수 있다.

4.3.2 언어쌍별 성능 비교

모든 다언어모델을 참조하여 각 언어쌍마다 가장 높은 피어슨 상관 계수들을 뽑아낸 결과는 Table 2와 같다. Table 2에서 확인할 수 있듯 동일한 1천 개의 평가 데이터셋에 대한 제로샷 성능은 언어마다 각각 다르게 분포되고 있는데, 본 논문에서는 이러한 결과가 나온 이유

를 언어의 특성 관점에서 분석해보고자 국가별 어족, 어순 및 알파벳에 대해 정리하였고, 표는 Table 3과 같다.

Table 2. Zero-shot transfer results for cross-lingual language models by language. The Pearson correlation coefficient for EN-DE is the result of fine-tuning, as shown in Table 1.

Language	EN-DE	EN-ZH	ET-EN	NE-EN	SI-EN	RO-EN
Maximum	0.442	0.256	0.395	0.347	0.419	0.465
Minimum	0.336	0.163	0.263	0.258	0.323	0.368
Average	0.394	0.207	0.309	0.300	0.384	0.414

Table 3. Comparison of language families, word order, and alphabets by language

	어족	어순	알파벳
독일어(DE)	인도유럽	SVO	라틴
중국어(ZH)	중국티베트	SVO	한자
에스토니아어(ET)	우랄	SVO	라틴
네팔어(NE)	인도유럽	SOV	데바나가리/란자나
싱할라어(SI)	인도유럽	SOV	싱할라
루마니아어(RO)	인도유럽	SVO	라틴

미세조정한 독일어 데이터셋과 비교하여 어족과 어순, 알파벳이 모두 동일한 언어는 루마니아어인데, Table 2의 실험 결과에서도 루마니아어에서 가장 높은 제로샷 전이 성능을 보였다. 이를 통해 국가별 언어적 특성이 동일하거나 비슷한 언어에서 교차언어 제로샷 전이가 더 잘 이루어졌다고 분석해볼 수 있다.

평균 피어슨 상관계수 0.207으로 가장 성능이 낮았던 중국어의 언어적 특성을 보면 독일어와 비교했을 때 어순을 제외한 어족과 알파벳이 모두 달랐다. 추가적으로 중국어에서는 띄어쓰기를 하지 않는데 이러한 특징들로 인해 가장 성능이 낮게 나왔다고 미루어 짐작할 수 있다. 또한 에스토니아어의 경우 루마니아어와 비교했을 때 어순과 알파벳이 동일하지만 어족이 다른 것을 확인할 수 있다. 독일어에서 에스토니아어로의 제로샷 교차언어 전이를 했을 때의 평균 피어슨 상관계수 0.309보다 동일한 어족인 루마니아어로의 전이를 했을 때 0.414로 더욱 좋은 성능을 보인 것으로 보아 [3]에서의 결과와 일관되게 어족이 동일한 경우 제로샷 교차언어 전이가 더 잘 이루어진다고 예측할 수 있으며, 제로샷 교차언어 전이에서 어족은 하나의 고려해야 할 요소라고 할 수 있다.

4.3.3 제로샷 교차언어 전이 및 미세조정 성능 비교

본 논문에서는 제로샷 성능과 미세조정 성능을 비교하기 위해 언어별 미세조정을 추가적으로 실시했다. 각 언어쌍별 미세조정을 위한 학습 데이터로는 4.1에서 언급한 바와 동일하게 WMT20에서 제공하는 데이터 중 학습, 검증 데이터를 추가로 활용하였다. 학습은 영어-독일어 QE 데이터에 대한 미세조정과 동일한 방법으로 수행하였으며, 동일하게 무작위 초기화 값으로 인한 성능 변동의 영향을 최소화하기 위해 총 다섯 번의 실험을 거쳐 평균값을 추출했다. 실험 결과는 Table 4와 같다. 실험 결과 미세조정을 한 경우 뚜렷하게 피어슨 상관계수가 압도적으로 높았다. 제로샷은 각 언어로 구성된 QE 태스크에 대한 지식, 뿐만 아니라 각 언어에 대한 지식마저도 없는 상황에서 QE 태스크를 수행했기 때문에 이러한 결과가 나왔음을 알 수 있다. 비록 QE에 대한 데이터가 존재하는 경우 성능은 훨씬 높아질 수 있으나, 제로샷 교차언어 전이 방법은 상당한 전문성을 요구하는 QE 학습 데이터를 구축할 수 없는 상황에서도 QE 수행을 가능하게 한다는 장점이 있다. QE에서의 제로샷 성능 향상을 위한 방법들은 향후 연구로 남기고자 한다.

Table 4. Comparison of zero-shot cross-lingual transfer results and fine-tuning results

Language	EN-DE	EN-ZH	ET-EN	NE-EN	SI-EN	RO-EN
Average Zero-shot	-	0.207	0.309	0.300	0.384	0.414
Average Fine-tuning	0.394	0.412	0.639	0.734	0.571	0.839

5. 결론

기계번역 품질 예측은 정답 번역문에 대한 참조 없이도 문장 전체에 대한 품질을 예측할 수 있다는 점에서 장점을 지니고 있다. 그러나 QE 학습을 위한 데이터의 구축 시 상당한 전문성이 요구되며, 이로 인해 QE 수행이 제한된다. 이러한 한계점을 다루기 위해 본 논문에서는 영어-독일어 언어쌍 데이터로 QE 모델을 학습한 후 다른 언어쌍으로의 제로샷 교차언어 전이를 수행함으로써 특정 언어에 대한 QE 데이터가 없는 경우에 대해서도 QE를 수행할 수 있도록 유도하였다. 또한 다양한 다언어모델에 대한 비교실험을 통해 어떤 모델이 제로샷 교차 전이에서 가장 강력한 성능을 보이는지를 확인하였다. 중국어, 에스토니아어, 네팔어, 싱할라어 및 루마니아어와 같이

다양한 자원의 언어에 대해서도 제로샷 실험을 진행하였고, 실험 결과 어족이나 어순, 알파벳이 유사하거나 동일한 언어로의 제로샷 전이에서 더 좋은 성능을 보였음을 확인하였다. 향후 연구로는, 데이터 중심의 방법론[29,30,31]에 대한 연구 활성화 움직임을 기반으로 데이터 필터링 기법[32,33] 등을 바탕으로 한 실험을 진행해 볼 예정이다.

REFERENCES

[1] T. Ranasinghe, C. Orasan & R. Mitkov. (2020). TransQuest at WMT2020: Sentence-Level Direct Assessment. *arXiv preprint arXiv:2010.05318.*

[2] Z. Chi, L. Dong, S. Ma, S. H. X. L. Mao, H. Huang & F. Wei. (2021). mT6: Multilingual Pretrained Text-to-Text Transformer with Translation Pairs. *arXiv preprint arXiv:2104.08692.*

[3] T. Pires, E. Schlinger & D. Garrette. (2019). How multilingual is multilingual BERT?. *arXiv preprint arXiv:1906.01502.*
DOI : 10.18653/v1/p19-1493

[4] G. Chen et al. (2021). Zero-shot Cross-lingual Transfer of Neural Machine Translation with Multilingual Pretrained Encoders. *arXiv preprint arXiv:2104.08757.*

[5] L. Specia, K. Shah, J. G. De Souza & T. Cohn (2013). QuEst-A translation quality estimation framework. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* 79-84.

[6] K. Papineni, S. Roukos, T. Ward & W. J. Zhu. (2002). Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics,* 311-318.
DOI : 10.3115/1073083.1073135

[7] S. Banerjee & A. Lavie. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization,* 65-72.
DOI : 10.3115/1626355.1626389

[8] S. Eo, C. Park, H. Moon, J. Seo & H. Lim. (2021). Dealing with the Paradox of Quality Estimation. *In Proceedings of the 4rd Workshop on Technologies for MT of Low Resource Languages,* 1-10.

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

[10] A. Conneau et al. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116.*
DOI : 10.18653/v1/P19-4007

[11] 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.*
DOI : 10.18653/v1/N19-1423

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

[13] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat & M. Johnson. (2020, November). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. *In International Conference on Machine Learning,* 4411-4421.

[14] G. Campagna, A. Foryciarz, M. Moradshahi & M. S. Lam. (2020). Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. *arXiv preprint arXiv:2005.00891.*
DOI : 10.18653/v1/2020.acl-main.12

[15] A. Lauscher, V. Ravishankar, I. Vulić & G. Glavaš, (2020). From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633.*
DOI : 10.18653/v1/2020.emnlp-main.363

[16] L. Zhou, L. Ding & K. Takeda. (2020). Zero-shot translation quality estimation with explicit cross-lingual patterns. *arXiv preprint arXiv:2010.04989.*

[17] Z. Chi, L. Dong, F. Wei, W. Wang, X. L. Mao & H. Huang. (2020). Cross-lingual natural language generation via pre-training. *In Proceedings of the AAAI Conference on Artificial Intelligence,* 7570-7577.
DOI : 10.1609/aaai.v34i05.6256

[18] L. Specia, D. Raj & M. Turchi (2010). Machine translation evaluation versus quality estimation. *Machine translation, 24(1),* 39-50.
DOI : 10.1007/s10590-010-9077-2

[19] D. Lee. (2020). Cross-lingual transformers for neural automatic post-editing. *In Proceedings of the Fifth Conference on Machine Translation,* 772-776.

[20] K. Shah, T. Cohn & L. Specia. (2015). A bayesian non-linear method for feature selection in machine translation quality estimation. *Machine Translation, 29(2),* 101-125.
DOI : 10.1007/s10590-014-9164-x

[21] R. Soricut, N. Bach & Z. Wang. (2012). The SDL language weaver systems in the WMT12 quality estimation shared task. *In Proceedings of the Seventh Workshop on Statistical Machine Translation,* 145-151.

[22] H. Kim, J. H. Lee & S. H. Na. (2017). Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. *In Proceedings of the Second Conference on Machine Translation,* 562-568.
DOI : 10.18653/v1/W17-4763

[23] R. N. Patel. (2016). Translation quality estimation using recurrent neural network. *arXiv preprint arXiv:1610.04841.*
DOI : 10.18653/v1/W16-2389

[24] H. Kim, J. H. Lim, H. K. Kim & S. H. Na. (2019). QE BERT: bilingual BERT using multi-task learning for neural quality estimation. *In Proceedings of the Fourth Conference on Machine Translation (3),* 85-89.
DOI : 10.18653/v1/W19-5407

[25] S. Eo, C. Park, H. Moon, J. Seo & H. Lim. (2021). Research on Recent Quality Estimation. *Journal of the Korea Convergence Society, 12(7),* 37-44.

[26] M. Wang et al. (2020). Hw-tsc's participation at wmt 2020 quality estimation shared task. *In Proceedings of the Fifth Conference on Machine Translation,* 1056-1061.

[27] 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.
DOI : 10.3390/app11146584

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

[29] C. Park, Y. Yang, K. Park & H. Lim. (2020). Decoding strategies for improving low-resource machine translation. *Electronics, 9(10),* 1562.
DOI : 10.3390/electronics9101562

[30] C. Park, S. Eo, H. Moon & H. S. Lim. (2021). 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,* 97-104.
DOI : 10.18653/v1/2021.naacl-industry.13

[31] 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,* 106-116.
DOI : 10.18653/v1/2021.wat-1.10

[32] 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

[33] C. Park, Y. Lee, C. Lee & H. Lim. (2020). Quality, not quantity?: Effect of parallel corpus quantity and quality on neural machine translation. *In The 32st Annual Conference on Human Cognitive Language Technology,* 363-368.

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 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

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

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)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 기계학습, 인공지능
· E-Mail : limhseok@korea.ac.kr