Journal of the Korea Convergence Society
Vol. 12. No. 7, pp. 37-44, 2021

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2021.12.7.037

최신 기계번역 품질 예측 연구

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

Research on Recent Quality Estimation

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

요 약 기계번역 품질 예측(Quality Estimation, QE)은 정답 문장(Reference sentence) 없이도 기계번역 결과의 질을 평가할 수 있으며, 활용도가 높다는 점에서 그 필요성이 대두되고 있다. Conference on machine translation(WMT)에서 매년 이와 관련한 shared task가 열리고 있고 최근에는 대용량 데이터 기반 Pretrained language model(PLM)을 적용한 연구들이 주로 진행되고 있다. 본 논문에서는 기계번역 품질 예측 task에 대한 설명 및 연구 동향에 대한 전반적인 survey를 진행했고, 최근 자주 활용되는 PLM의 특징들에 대해 정리하였다. 더불어 아직 활용된 바가 없는 multilingual BART 모델을 이용하여 기존 연구들인 XLM, multilingual BERT, XLM-RoBERTa와의 비교 실험 및 분석을 진행하였다. 실험 결과 어떤 사전 학습된 다중언어 모델이 QE에 적용했을 때 가장 효과적인지 확인하였을 뿐 아니라 multilingual BART 모델의 QE 태스크 적용 가능성을 확인했다.

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

Abstract Quality estimation (QE) can evaluate the quality of machine translation output even for those who do not know the target language, and its high utilization highlights the need for QE. QE shared task is held every year at Conference on Machine Translation (WMT), and recently, researches applying Pretrained Language Model (PLM) are mainly being conducted. In this paper, we conduct a survey on the QE task and research trends, and we summarize the features of PLM. In addition, we used a multilingual BART model that has not yet been utilized and performed comparative analysis with the existing studies such as XLM, multilingual BERT, and XLM-RoBERTa. As a result of the experiment, we confirmed which PLM was most effective when applied to QE, and saw the possibility of applying the multilingual BART model to the QE task.

Key Words : Quality Estimation, Neural Machine Translation, Deep Learning, 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-2020-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and this research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience program(IITP-2021-0-01819) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)
*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received April 7, 2021 Revised April 22, 2021
Accepted July 20, 2021 Published July 28, 2021

1. 서론

기계번역(Machine translation, MT) 연구가 활발해짐에 따라 병렬 코퍼스 필터링(Parallel corpus filtering, PCF), 기계번역 품질 예측(Quality estimation, QE), 기계번역 자동 사후 교정(Automatic post editing, APE), 성능평가(Metrics)와 같은 MT의 하위분야에 대한 관심이 증대되고 있다. 특히 QE에 대한 지속적인 연구가 이루어지고 있다[1,2].

QE란 정답 문장(reference sentence)을 참고하지 않고 소스 문장(source sentence)과 기계번역 문장(MT sentence)만을 활용하여 기계번역 결과에 대한 품질을 예측하는 연구이다[3,4]. QE에서는 기계번역 문장에 대한 품질을 수치 또는 오류 태그와 같은 품질 주석을 (quality annotations) 통해 나타낸다. 이를 활용하여 여러 기계번역 시스템 중 어떤 시스템의 번역 결과가 가장 좋은지를 선택하거나, 결과에 대한 순위(ranking)를 매길 수 있다[5]. 또한 품질이 낮은 기계번역 문장의 경우, 어절 단위로 부착된 품질 주석을 활용하여 품질이 낮은 어절만을 수정함으로써 사후 교정 시 효율을 높일 수 있다. 이와 같이 기계번역에서 QE는 폭넓은 적용이 가능하다는 점에서 그 중요성이 부각되고 있다.

기계번역 분야에서 가장 영향력 있는 컨퍼런스 중 하나인 WMT(Conference on Machine Translation)에서는 2012년을 기점으로 QE task가 매년 개최되고 있다. 최근 연구동향을 살펴보면 사전 학습된 다중언어 모델(pre-trained cross-lingual language model)에 기반하여 QE task를 진행하는 경우들이 다수를 이루고 있다[6,7]. 기존 연구들의 경우 XLM[8], multilingual BERT(mBERT)[9], XLM-RoBERTa(XLM-R)[10] 등을 활용하는 경향을 보인다. 그러나 사전 학습된 다중언어 모델들에 대한 비교가 선행되지 않은 채 활용되고 있으며, 특정 기준 없이 활용되고 있다는 점에서 최적화된 모델을 특정하기가 어렵다는 한계점이 있다.

본 논문에서는 WMT를 기준으로 전반적인 QE 연구에 대한 survey를 진행했으며, 최근 연구동향에 따라 사전 학습된 다중언어 모델들을 선택적으로 활용할 수 있도록 각 모델들에 대한 학습 방법 및 특징들에 대해 요약했다. 더불어 최근 연구에 대한 한계를 완화하기 위해 기존에 활용된 바 없는 사전 학습된 다중언어 모델인 multilingual BART (mBART)[11]를 새롭게 활용하여 QE에 적용해보았다. mBART는 기계번역에서 가장 좋은 성능을 보이는 모델이나 QE에서는 이에 대해 기존에 활용된 바가 없다. 본 논문에서는 mBART에 대한 실험과 함께 기존에 활용된 모델들에 대해 비교 실험을 진행해 봄으로써 QE 태스크에서 어떤 모델들이 좋은 성능을 내는지를 확인하고 결과에 대한 분석을 진행했다.

본 논문은 다음과 같이 구성된다. 2장에서는 연도별 QE 연구의 흐름에 대해 설명하고 3장에서는 QE의 sub-task들에 대해 소개하며, 4장에서는 최신 QE task에서 활용되는 사전 학습된 다중언어 모델 및 mBART에 대해 설명한다. 5장에서는 4장에서 언급한 다중언어 모델들에 대해 각각 QE task 실험을 진행하고 실험 결과를 비교 분석한다. 이후 6장의 결론으로 마무리한다.

2. 기계번역 품질 예측 역사

2.1 전통적인 기계번역 품질 예측 연구

전통적인 QE 연구들은 대부분 품질 예측 학습을 위한 자질을 추출하거나(feature extraction) 자질을 선정하는(feature selection) 방식으로 진행되었다. 자질 선정 시에는 주로 Gaussian Process, SVM(Support vector machine), Regression Trees 등 기계학습 알고리즘들을 주로 활용하였다[12,13]. 자질 추출의 경우 데이터들에 대해 파서(parser), 태거(tagger), 개체명 인식기(named entity recognizer) 등 외부 자원(external resources)을 활용하여 언어학적 자질(linguistic features)이나 pseudo-reference 자질 등을 추출했다[14,15].

그러나 이들은 자질들과 정답 간의 복잡한 관계성을 찾아내는 데 초점이 맞추어져 있으며, 최적화된 자질을 선정 및 추출하는 과정은 휴리스틱한 과정을 요구하기에 한계가 존재한다.

2.2 WMT16부터 WMT19까지의 연구

WMT16의 QE task에서는 통계기반 모델들을 활용하는 연구들과 더불어 딥러닝(deep learning)을 적용한 연구들이 등장하기 시작했다. Cdacm[16]과 포항공과대학교(Postech)[17]에서는 순환신경망(Recurrent neural network, RNN)[18] 또는 장단기 메모리(Long short-term memory, LSTM)[19]를 활용하였으며, 이들은 각각 phrase 레벨과 sentence 레벨에서 1위를 차지했다.

WMT17에서부터는 대다수의 연구가 딥러닝을 기반으로 진행되었다. 특히 Postech에서 Predictor-Estimator

구조를 새롭게 제안하여 모든 sub-task에서 압도적인 성능 향상을 보이며 1위를 차지했다[20]. Predictor-Estimator 구조는 각각 순환신경망 모델로 구성된다. Predictor는 이중언어 및 양방향 기반 순환신경망 단어 추정 모델(word prediction model)로, 대용량의 병렬 말뭉치 중 타겟 문장에 무작위로 단어를 선택해 마스킹하고 이를 예측한다. 예측 시에는 소스 문장과 타겟 문맥을 참고하며, 예측한 타겟 단어에 대한 지식 정보(Quality estimation feature vectors, QEFVs)는 Estimator 구조로 전달(knowledge transfer)된다. Estimator에서는 QE 데이터를 활용하여 문장(sentence), 단어(word) 및 구(phrase) 레벨의 sub-task를 수행한다. 이 구조는 제한된 양의 기계번역 품질 데이터에 대해 추가 대용량 병렬 말뭉치를 활용할 수 있도록 하면서 데이터 부족 문제를 완화할 수 있었으며, 동시에 비약적인 성능 향상을 이끌어냈다. 최근까지도 이 구조를 변형 및 활용하는 연구들이 계속해서 이루어지고 있다.

WMT18에서는 Predictor-Estimator 구조와 유사한 QE brain[21]이 1위를 차지했다. QE brain은 두 가지 프로세스로 진행되는데, Transformer[22] 모델을 활용하여 자질을 추출하는 과정과 양방향 장단기메모리(Bi-LSTM)를 이용하여 QE를 진행하는 과정이 있다. 자질 추출 과정에서는 소스와 타겟 문장 간 고수준의 공동 잠재 의미 표현(high-level joint latent semantic representations)들을 뽑아내며 이를 인간이 만든 자질(human-craft features)과 결합하여 Estimator에 넣어주게 된다.

Pretrained language model(PLM)의 등장으로 WMT19부터는 이에 기반한 연구들이 진행되었다. Unbabel[23]에서는 Predictor-Estimator의 구조로 학습을 진행하면서, Predictor 부분을 사전 학습된 BERT 또는 XLM 모델로 대체한 연구를 함께 진행했다. ETRI에서는 mBERT를 기반으로 QE task에 대해 사전학습 및 미세조정을 진행하였다[24].

2.3 WMT20 연구

WMT20에서도 PLM을 활용하는 연구들이 활발하게 수행되었다. 문장 레벨의 sub-task 1에서 1위를 차지한 Transquest는 MonoTransquest, SiameseTransquest의 두 가지 구조를 제안했다[25]. 전자는 사전 학습된 다중언어 모델인 XLM-R에 대해 미세 조정을 진행했고, 후자는 소스 문장과 타겟 문장 각각에 대해 개별 XLM-R 모델을 활용하여 각 결과 값에 대한 코사인 유사도(cosine similarity)를 측정해 문장 레벨에 대한 번역 품질을 예측했다. Sub-task 2의 단어 레벨 중 영어-독일어 QE에서 1위에 등극한 Bering lab 역시 XLM-R을 활용했고 두 가지 학습 절차를 거쳤다[26]. 첫 번째 학습에서는 병렬 말뭉치를 활용하여 pseudo 번역 오류율(Translation error rate, TER) 점수를 만들어내 데이터 증강을 실시했으며, XLM-R 사전 학습 모델을 로드한 후 pseudo 데이터에 대해 학습을 진행했다. 이후 WMT에서 제공하는 데이터를 활용하여 미세조정을 실시했다. Sub-task 2의 영어-독일어 문장 레벨 및 단어 레벨 중 영어-중국어에서 1위를 차지한 Huawei Translation Service Center(HW-TSC)는 사전 학습된 언어모델을 활용하지 않았으며 사전 학습된 Transformer 모델을 Predictor로, task에 구체화된 회귀 및 분류기(task-specific regressors or classifiers)를 Estimator 구조로 간주하여 학습을 진행했다[27]. 학습 과정에서는 전이학습의 효율을 개선하고 과적합(overfitting)을 방지하기 위해 병목 어댑터 층(Bottleneck adapter layer, BAL)을 새롭게 추가했다. 또한 APE shared task에서 제공하는 데이터 및 번역 모델을 활용하여 데이터 증강을 실시해 성능 향상을 도모했다. Sub-task 2의 영어-중국어 단어 레벨에서 1위를 한 Tencent에서도 Predictor-Estimator 구조를 활용했다[28]. 이들은 Transformer 기반과 XLM 기반의 두 가지 Predictor를 활용했으며, Estimator는 LSTM 또는 Transformer를 이용하였다.

사전 학습된 다중언어 모델은 많은 언어들 및 대용량의 데이터에 대해 학습을 진행했기 때문에 자연어처리의 다양한 sub-task들에서 우수한 성적을 보이고 있다. QE task에서도 이를 활용한 연구가 주를 이루고 있으며 WMT21에서도 이러한 접근법이 적극적으로 활용될 것으로 보인다.

3. 기계번역 품질 예측 Sub-Task 소개

WMT의 QE sub-task는 매년 조금씩 변경되나 크게 문장 레벨(sentence-level), 단어 레벨(word-level), 문서 레벨(document-level)로 구성된다. 각 sub-task에 따라 활용 용도를 달리할 수 있으며 이들을 잘 활용하기 위해서는 태스크에 대한 이해가 선행되어야 한다. 문장 및 단어 레벨은 2013년을 시작으로 매년 열리고 있으며, 문서 레벨의 경우 2015년에 처음 task가 열렸다.

2020년에는 QE의 sub-task가 문장 레벨의 직접 평

가(sentence-level direct assessment task), 단어 및 문장 레벨의 사후교정 노력 정도(word and sentence-level post-editing effort), 문서 레벨의 기계번역 품질 예측(document-level QE)의 세 가지 분야가 진행되었다.

3.1 Sentence-level direct assessment task

Sub-task 1은 문장 레벨의 직접 평가이며 2020년에 새롭게 도입되었다. 이 task에서는 기계번역 문장에 대한 품질을 직접 평가 점수(Direct assessment, DA)로 나타내는데, 여기서 DA 점수란 적어도 3명의 번역 전문가들이 제공된 DA 형식에 따라 0점에서 100점 사이로 기계번역 문장의 품질을 평가한 값이다. 데이터는 소스 문장, 기계번역 문장, DA 점수와 이를 z표준화한 값의 평균 등으로 구성되며, 총 7천 개의 학습 데이터와 1천 개의 평가 데이터를 활용하여 z-표준화된 직접 평가 점수의 평균을 예측하게 된다.

3.2 Word and sentence-level post-editing effort task

Sub-task 2는 단어 및 문장 레벨의 사후교정 노력 정도이다. 단어 레벨에서는 기계번역의 품질을 어절마다 옳음(OK) 또는 잘못됨(BAD)으로 예측한다. 데이터는 소스 문장, 번역 문장과 사후 교정을 진행한 문장, 단어 얼라인, 소스 문장 태그, 기계번역 문장 및 갭 토큰에 대한 태그, HTER 점수 등으로 구성되어 있다. 이 데이터를 자유롭게 활용하여 단어 레벨에서는 태그를, 문장 레벨에서는 점수를 예측하게 된다.

단어 레벨의 경우 소스 문장과 타겟 문장의 어절들 각각에 대해 태깅(tagging)을 진행하게 되는데, 타겟 문장의 경우 어절 사이에 빠진 단어들이 있는 경우를 고려하여 어절 사이마다 갭(GAP) 태그를 추가한다. 타겟 문장의 어절 개수가 N개라면, 태그된 어절 개수는 타겟 문장의 어절에 해당하는 N개에 갭 태그 개수 N+1개를 더하여 총 2N+1개의 태그를 예측한다.

문장 레벨의 경우 task 1과 비슷하게 참가자들이 문장 전체에 대한 품질 평가 점수를 예측해야 하는데, 이 때 점수는 휴먼 번역 오류율(Human translation error rate, HTER)[29]을 기준으로 한다. 휴먼 번역 오류율은 TER과 비슷하게 측정되는데, 여기서 TER이란 기계번역 결과를 정답 문장과 비교하여 얼마나 최소한의 변경(substitution), 삭제(deletion), 추가(insertion)를 진행해야 옳은 문장이 되는지를 비율로 나타낸 값이다. 휴먼 번역 오류율은 정답 문장과는 별개로 기계번역 문장에 대해 사람이 더 최소한의 수정을 거쳐 옳은 문장으로 변경한 횟수를 확률로 나타낸다. HTER과 TER은 모두 번역 오류율을 나타내므로 값이 낮을수록 번역 오류가 적다.

3.3 Document-level QE task

Sub-task 3은 문서 레벨의 QE이다. 각 문서의 문장별 어떤 부분이 얼마나 잘못되었는지를 스팬(span)과 스팬 길이(span length)로 나타낸다. 또한 오류가 있는 부분에 대한 심각 정도를 3가지로 분류하는데, 번역 오류는 있지만 의미는 제대로 전달되는 경우 마이너(minor), 의미 오류로 이어질 경우 메이저(major), 의미가 잘못 번역되면서 동시에 화자로 하여금 오해의 소지를 불러일으킬 수 있는 오류의 경우 크리티컬(critical)로 나타낸다. 참가자들은 문서 중 오류의 부분과 오류의 심각 정도를 찾아내게 된다. 데이터로는 문서와 문장들, 스팬, 심각 정도, 오류 유형 등이 제공되며 이를 선택적으로 활용하여 문서 레벨의 질을 예측하게 된다.

4. 최신 기계번역 품질 예측 모델

최근 사전 학습된 다중언어 모델을 활용하는 흐름에 따라 본 논문에서는 QE에서 주로 적용되고 있는 다중언어 모델에 대해 정리한다. 또한 본 논문에서 추가로 적용해볼 mBART에 대해서도 설명한다.

먼저 다중언어 모델이 사전 학습 시에 활용된 데이터는 다음과 같다. XLM과 mBERT에서는 100개 언어에 대한 wikipedia 문장들을 활용하여 학습을 진행했다. XLM-R에서는 CommonCrawl[30] 데이터 중 100개 언어에 해당하는 CC100을, mBART에서 25개 언어에 해당하는 CC25를 사전 학습에 활용하였다. 사전 학습된 다중언어 모델을 활용한 연구들은 대부분 모델을 로드한 후 추가 사전 학습 또는 미세조정 등을 진행하는 방법을 활용하며, 모델의 은닉 상태 값들(hidden states)을 활용하여 최종 예측 값을 뽑아내는 방식으로 진행된다.

4.1 multilingual BERT

multilingual BERT(mBERT)[31]는 Google에서 발표한 BERT의 다국어 버전으로 104개 언어의 wikipedia corpus를 모두 사용하여 pre-training을 진행하였다는 것 외에 기존 BERT와 동일하게 학습이 진행된다. 즉

BERT에서 사전 학습 시 활용한 Masked language model(MLM)과 Next sentence prediction(NSP)를 동일하게 적용하여 단일 말뭉치에 대해 학습을 진행하되 단일 말뭉치는 104개 언어의 말뭉치로 확장했다. 여기서 BERT의 MLM이란, 무작위로 15%의 토큰을 마스킹한 후 이를 예측하도록 하며, NSP는 두 문장에 대해 순서를 변경하여 모델이 순서에 대한 정보를 학습할 수 있도록 한다.

4.2 Cross-lingual language model(XLM)

Cross-lingual language model(XLM)은 다중언어의 representation을 학습하는 목적으로 기존 언어모델 학습 방법론을 확장한 구조이다. XLM에서는 단일 언어(monolingual) 말뭉치에 대해 비지도 학습을 수행하는 Causal language model(CLM)과 MLM 그리고 병렬 말뭉치에 대해 지도학습을 수행하는 Translation language model(TLM)을 제안하였다. CLM은 이전 단어들을 참고하여 다음 단어를 학습하는 모델이다. MLM은 BERT의 MLM과 동일하게 15%의 무작위 마스킹을 거친 후 이를 예측하는데, BERT의 경우는 마스킹된 문장 쌍을 입력으로 넣어주지만 XLM의 MLM에서는 마스킹된 전체 문장들(sentence stream)에 대해 256토큰씩 끊어가며 입력으로 넣어준다. TLM은 다중언어에 대한 지식을 함께 학습하기 위해 고안되었다. 병렬 데이터를 활용하여 무작위로 마스킹된 소스 문장과 얼라인(align)된 타겟 문장을 연결(concatenation)시켜주고 마스킹된 부분을 예측하도록 한다. 마스킹된 토큰을 예측할 시 주변 언어에 대한 문맥을 고려하여 예측을 진행하면서 동시에 함께 입력으로 넣어준 다른 언어의 문장 맥락을 참고할 수 있도록 하여 다중언어에 대한 지식을 습득하게 된다. 또한 다중 언어정보를 학습하기 위하여 각 언어에 대한 언어 임베딩(language embedding)을 사용하며 각 언어의 시작과 끝을 나타내기 위해 position embedding도 언어별로 수행한다.

4.3 multilingual BART

multilingual BART(mBART)는 BART[32]를 다국어로 확장한 시퀀스 투 시퀀스(sequence-to-sequence) 구조이다. BART는 대용량의 영어 단일 말뭉치에 대해 문장의 순서를 섞거나(sentence permutation) 무작위로 토큰을 마스킹(token masking), 토큰 삭제(token deletion), 문장 중 포아송 분포(Poisson distribution)에 따른 스팬 길이(span length)를 하나의 [MASK] 토큰으로 치환(text infilling), 문서 중 무작위로 토큰을 뽑아 그 토큰으로 시작하게끔 문서를 섞는(document rotation) 노이즈를 추가하고 이를 Transformer 구조를 활용하여 완벽한 본래의 문장으로 복원하는 디노이징(denoisiong)을 진행한다. mBART의 경우 다국어의 대용량 단일 말뭉치에 대해 sentence permutation과 text infilling만을 활용하여 텍스트를 복원하는 디노이징 사전 학습을 수행한다. 이렇게 mBART 사전 학습을 진행하는 과정에서 모델은 언어에 대한 보편적인 표현(universal representation)을 학습할 수 있다.

4.4 XLM-RoBERTa

XLM-RoBERTa(XLM-R)는 XLM을 확장한 구조이다. 100개의 언어를 포함하는 2TB 이상의 CommonCrawl 데이터에 대해 BERT의 MLM을 활용하여 학습을 진행했다. 단일 언어 데이터를 활용하여 비지도 학습으로 다중언어 표현(cross-lingual representation)을 학습한다.

XLM-R에서는 다중언어의 저주(curse of multilinguality)를 문제 삼고 모델 수용력(model capacity)을 크게 확장했다는 특징이 있다. 다중언어의 저주란, 모델 수용력이 고정되어 있을 때 언어를 추가할수록 처음에는 고자원 언어들(high-resource languages)에 의해 비슷한 저자원 언어들(low-resource languages)의 성능이 향상되지만 특정 이상으로 넘어가면 성능이 다시 하락하며, 동시에 언어를 추가할수록 고자원 언어 하나가 가지는 모델 수용력의 크기가 줄어들어 고자원 언어에 대한 성능마저 떨어지는 상황을 말한다. 이에 대해 모델 수용력을 크게 확장함으로써 저자원 언어의 향상 및 고자원 언어의 성능을 유지할 수 있도록 했다.

5. 실험

5.1 데이터 및 모델

본 논문에서는 사전 학습된 다중언어 모델을 활용하여 WMT20의 sub-task 2에 대한 실험을 진행하였다. 언어쌍은 영어-독일어에 대해 진행했으며, WMT20에서 제공하는 학습 및 평가 데이터를 활용하였다. 다중언어 모델은 XLM, XLM-R, mBERT와 QE에서는 처음 시도하는 mBART까지 총 5가지 사전 학습 모델에 대해 실험을 진행했고, HuggingFace[33]에서 배포하는 사전 학습

모델을 활용하였다. 본 논문은 다중언어 모델별 미세조정 성능을 평가하고자 하므로 모델 학습 시 추가적인 데이터 증강은 수행하지 않았다.

미세조정 시에는 사전 학습된 모델들을 초기 값으로 활용하였고, 각 모델의 마지막 은닉 상태 값(last hidden state)을 선형 분류기(linear classifier)에 넣어주고 문장에 대한 점수 값을 뽑아낼 수 있도록 했다. 손실 값으로는 예측한 점수와 정답 점수 간의 차이를 측정하기 위해 mean squared error(MSE) 손실 값을 활용하였다.

5.2 실험의 필요성

기존 QE 연구들은 대부분 WMT shared task에서 가장 좋은 성능을 거두기 위한 연구만을 진행하고 있다. 이러한 연구는 학습 및 테스트 데이터에 의존적이며, 데이터 증강이 필수적으로 요구될 뿐만 아니라 해당 기법으로 인해 성능이 좌우되는 양상을 보인다. 더불어 충분한 다중언어 모델간의 비교 없이, 극 대용량 데이터로 사전 학습을 진행한 XLM-R을 기반으로 QE 모델을 대부분 학습한다.

그러나 본 논문은 데이터 증강의 효과를 없애고 다중언어모델 간의 순수한 성능 비교가 필요하다고 생각되어 WMT 20의 task 2를 기반으로 대표적인 다중언어 모델간 비교 연구를 진행하였다.

5.3 실험 결과

Sub-task 2(문장 레벨의 사후교정 노력 정도 task)에서 사전 학습 모델별 미세조정을 진행한 결과는 Table 1과 같다.

Table 1. Fine-tuning results for cross-lingual language models in EN-DE subtask 2

Model	Pearson r	MAE	RMSE
XLM[8]	0.332	0.158	0.200
mBERT[9]	0.449	0.179	0.229
XLM-RoBERTa[10]	**0.501**	**0.144**	**0.185**
mBART[11]	0.477	0.140	0.181

실험 결과 XLM-R이 피어슨 상관계수 0.501으로 가장 좋은 성능을 보였다. mBART와 mBERT가 각각 0.477, 0.449의 비슷한 성능을 보였고, XLM의 성능이 0.332로 가장 낮게 나왔다. Mean absolute error(MAE), Root mean squar error(RMSE)의 경우에도 XLM-R이 가장 좋은 결과가 나왔다.

이를 통해 XLM-R이 다중언어 모델 중 가장 좋은 성능을 냈음을 확인할 수 있다. XLM-R의 경우 100개 언어로 구성된 2TB 이상의 대용량 코퍼스에 대해 사전 학습을 진행했으며 이는 다른 모델들이 학습한 데이터에 비해 훨씬 방대하다는 차별점이 있다. Fig. 1을 보면 전체적으로 CommonCrawl 데이터가 wikipedia 데이터보다 훨씬 더 그 크기가 크며, 저자원 언어에 대해서는 특히나 더 많은 데이터를 가지고 있음을 알 수 있다. 즉 wikipedia 데이터 또는 25개 언어로 구성된 CC25 데이터로 학습한 기존 모델에 비해 XLM-R이 훨씬 더 많은 학습량과 모델 수용력(model capacity)으로 사전 학습했다는 점이 성능 향상에 대한 가장 큰 요인으로 해석될 수 있다.

추가적으로 XLM과의 비교결과 피어슨 상관계수 0.169 라는 큰 성능 향상 폭을 보인 것으로 보아 XLM-R의 사전 학습 시 활용한 코퍼스의 양이 성능 향상에 큰 영향을 미쳤으며, 기존 모델들보다 좋은 성능을 보였음을 알 수 있다.

mBART에서는 이에 더 나아가 문장 순서 변경뿐만 아니라 여러 토큰을 하나로 마스킹하여 예측하도록 하는 추가적인 노이징 전략을 사용함으로써 두 모델보다 높은 성능을 낼 수 있었다고 할 수 있다.

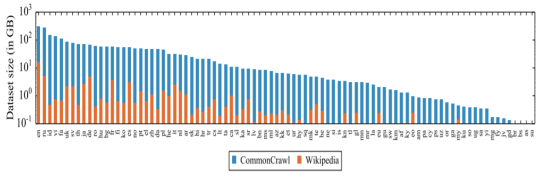

Fig. 1. Comparison of the amount of data for the 88 languages of the CommonCrawl-100 corpus and the Wikipedia-100 corpus[10]

6. 결론

기계번역 품질 예측은 번역 문장에 대한 오류 정도를 수치화하며 구체적으로 어떤 어절에 대한 번역 수정이 필요한지를 알려주는 활용도 높은 기계번역 sub-task이다. 본 논문에서는 QE의 최신 동향 survey와 더불어 QE가 처음 개최된 WMT12부터 현재까지 어떠한 흐름으로 주요 연구들이 진행되어왔는지를 다루었다. 추가로 mBERT, XLM, XLM-R과 더불어 기존 연구에서는 이용되지 않았던 mBART 모델을 QE task에 적용해보았다. 추후 다양한 기계번역의 전처리 기술 중 하나인 parallel corpus filtering을 적용하여 QE의 성능을 향상시킬 계획이다[34,35].

REFERENCES

[1] L. Specia, F. Blain, V. Logacheva, R. Astudillo & A. Martins. (2018). Findings of the wmt 2018 shared task on quality estimation. Association for Computational Linguistics.
DOI : 10.18653/v1/W18-6451

[2] E. Fonseca, L. Yankovskaya, A. F. Martins, M. Fishel & C. Federmann. (2019). Findings of the WMT 2019 shared tasks on quality estimation. I*n Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2),* 1-10.
DOI : 10.18653/v1/W19-5401

[3] 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.

[4] L. Specia, C. Scarton & G. H. Paetzold (2018). Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies, 11(1),* 1-162.
DOI : 10.2200/S00854ED1V01Y201805HLT039

[5] 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

[6] D. Lee. (2020). Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. *In Proceedings of the Fifth Conference on Machine Translation.* (pp. 1024-1028).

[7] Y. Baek, Z. M. Kim, J. Moon, H. Kim & E. Park. (2020). Patquest: Papago translation quality estimation. *In Proceedings of the Fifth Conference on Machine Translation.* (pp. 991-998).

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

[9] 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

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

[12] E. Biçici. & A. Way. (2014). Referential translation machines for predicting translation quality. Association for Computational Linguistics.
DOI : 10.18653/v1/w15-3035

[13] 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. (pp. 145-151).*

[14] N. Q. Luong, B. Lecouteux & L. Besacier. (2013). LIG system for WMT13 QE task: Investigating the usefulness of features in word confidence estimation for MT. *In 8th Workshop on Statistical Machine Translation.* (pp. 386-391).

[15] C. Hardmeier, J. Nivre & J. Tiedemann. (2012). Tree kernels for machine translation quality estimation. *In Seventh Workshop on Statistical Machine Translation, Montréal, Canada, June 7-8, 2012.* (pp. 109-113). Association for Computational Linguistics.

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

[17] H. Kim & J. H. Lee. (2016). Recurrent neural network based translation quality estimation. *In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers.* (pp. 787-792).
DOI : 10.18653/v1/w16-2384

[18] K. Cho et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078.*
DOI : 10.3115/v1/d14-1179

[19] S. Hochreiter & J. Schmidhuber. (1997). Long short-term memory. *Neural computation, 9(8),* 1735-1780.
DOI : 10.1162/neco.1997.9.8.1735

[20] H. Kim, J. H. Lee & S. H. Na. (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).
DOI : 10.18653/v1/w17-4763

[21] J. Wang, K. Fan, B. Li, F. Zhou, B. Chen, Y. Shi & L. Si. (2018). Alibaba submission for WMT18 quality estimation task. *In Proceedings of the Third Conference on Machine Translation: Shared Task Papers.* (pp. 809-815).
DOI : 10.18653/v1/w18-6465

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

[23] F. Kepler et al. (2019). Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. *arXiv preprint arXiv:1907.10352.*
DOI : 10.18653/v1/W19-5406

[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 (Volume 3: Shared Task Papers, Day 2).* (pp. 85-89).
DOI : 10.18653/v1/W19-5407

[25] T. Ranasinghe, C. Orasan & R. Mitkov. (2020). TransQuest at WMT2020: Sentence-Level Direct

Assessment. *arXiv preprint arXiv:2010.05318*.

[26] D. Lee. (2020). Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. I*n Proceedings of the Fifth Conference on Machine Translation.* (pp. 1024-1028).

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

[28] H. Wu et al. (2020, November). Tencent submission for WMT20 Quality Estimation Shared Task. *In Proceedings of the Fifth Conference on Machine Translation.* (pp. 1062-1067).

[29] M. Snover, B. Dorr, R. Schwartz, L. Micciulla & J. Makhoul. (2006, August). A study of translation edit rate with targeted human annotation. *In Proceedings of association for machine translation in the Americas (Vol. 200, No. 6).*

[30] G. Wenzek et al. (2019). Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

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

[32] M. Lewis et al. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461.*
DOI : 10.18653/v1/2020.acl-main.703

[33] T. Wolf et al. (2019). HuggingFace's Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

[34] 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.

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

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 : Machine Translation, Grammar Error Correction, Deep Learning
·E-Mail : bcj1210@naver.com

문 현 석(Hyeonseok Moon) [학생회원]

· 2021년 2월 : 고려대학교 수학과 (이학사)
· 2021년 3월 ~ 현재: 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Neural Machine Translation
· 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) [정회원]

· 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