Journal of the Korea Convergence Society
Vol. 12. No. 5, pp. 23-29, 2021

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2021.12.5.023

인공신경망 기계번역에서 말뭉치 간의 균형성을 고려한 성능 향상 연구

박찬준[1], 박기남[2], 문현석[1], 어수경[1], 임희석[3*]
[1]고려대학교 컴퓨터학과 석·박사통합과정, [2]고려대학교 정보창의교육연구소 연구교수, [3]고려대학교 컴퓨터학과 교수

A study on performance improvement considering the balance between corpus in Neural Machine Translation

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

요 약 최근 딥러닝 기반 자연언어처리 연구들은 다양한 출처의 대용량 데이터들을 함께 학습하여 성능을 올리고자 하는 연구들을 진행하고 있다. 그러나 다양한 출처의 데이터를 하나로 합쳐서 학습시키는 방법론은 성능 향상을 막게 될 가능성이 존재한다. 기계번역의 경우 병렬말뭉치 간의 번역투(의역, 직역), 어체(구어체, 문어체, 격식체 등), 도메인 등의 차이로 인하여 데이터 편차가 발생하게 되는데 이러한 말뭉치들을 하나로 합쳐서 학습을 시키게 되면 성능의 악영향을 미칠 수 있다. 이에 본 논문은 기계번역에서 병렬말뭉치 간의 균형성을 고려한 Corpus Weight Balance (CWB) 학습 방법론을 제안한다. 실험결과 말뭉치 간의 균형성을 고려한 모델이 그렇지 않은 모델보다 더 좋은 성능을 보였다. 더불어 단일 말뭉치로도 고품질의 병렬 말뭉치를 구축할 수 있는 휴먼번역 시장과의 상생이 가능한 말뭉치 구축 프로세스를 추가로 제안한다.

주제어 : 기계번역, 병렬말뭉치, 휴먼번역, 고품질 데이터, 딥러닝, 언어융합

Abstract Recent deep learning-based natural language processing studies are conducting research to improve performance by training large amounts of data from various sources together. However, there is a possibility that the methodology of learning by combining data from various sources into one may prevent performance improvement. In the case of machine translation, data deviation occurs due to differences in translation(liberal, literal), style(colloquial, written, formal, etc.), domains, etc. Combining these corpora into one for learning can adversely affect performance. In this paper, we propose a new Corpus Weight Balance(CWB) method that considers the balance between parallel corpora in machine translation. As a result of the experiment, the model trained with balanced corpus showed better performance than the existing model. In addition, we propose an additional corpus construction process that enables coexistence with the human translation market, which can build high-quality parallel corpus even with a monolingual corpus.

Key Words : Machine Translation, Parallel Corpus, Human Translation, High Quality Data, Deep Learning, Language Conversion

*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 the MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience program(IITP-2021-2020-0-01819) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation).
*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received February 22, 2021 Revised March 18, 2021
Accepted May 20, 2021 Published May 28, 2021

1. 서론

최근 글로벌 교역이 지속적으로 증가하여 다국어 번역에 대한 필요가 증가하고 있으며 기계번역 시장은 날로 성장하고 있다. 대한민국에 대표적으로 네이버의 Papago, 카카오 번역, ETRI, 한컴인터프리의 지니톡, LLsoLLu, 플리토, SYSTRAN, 현대자동차 등 많은 기업에서 기계번역에 관련된 연구를 진행하고 있다. 해외의 경우 구글, 마이크로소프트, 아마존, Unbabel 등 많은 기업들이 기계번역과 관련된 연구 및 사업화를 진행하고 있다.

과거 기계번역 관련하여 많은 프로젝트들이 1950년대 말부터 시작되었으나 딥러닝의 등장 전까지는 대부분 성공보다는 실패한 사례가 더 많았다. 그러나 딥러닝의 등장으로 Neural Machine Translation(NMT)가 개발되면서 사람들이 이전보다 만족할 만한 성능의 기계번역기들이 개발되고 있다[1-5]. 과거 기계번역 연구는 규칙기반 및 통계기반 방식을 이용했으나 최근에는 딥러닝 기반 방식으로 많은 기술적인 성과를 이루어냈다. GPU의 등장으로 인한 행렬 연산의 병렬처리를 통한 컴퓨팅 파워의 개선, Tensorflow 및 Pytorch 등의 오픈소스 프레임워크의 등장으로 인한 개발환경의 개선, 웹을 통한 빅데이터 확보 가능, 획기적인 딥러닝 모델 개발 등을 바탕으로 딥러닝을 이용한 다양한 분야에서 엄청난 성과를 보이고 있으며 기계번역도 마찬가지이다. 그러나 여전히 사람들에게 실질적인 만족감을 부여하기 위해서는 아직 많이 개선되어야 할 사항들이 많다.

대표적으로 데이터의 품질이 개선되어야 한다. 고품질의 학습데이터를 구축하는 일은 딥러닝의 전 분야에 걸쳐서 공통된 중요사항이다. 그러나 양질의 데이터를 구하는 일은 저작권 확보의 문제, 정렬 작업의 어려움, 상당한 비용과 시간의 투자 등을 이유로 쉽지 않은 상황이다. 기계번역의 경우에도 이종 언어로 된 병렬말뭉치를 구하기 어려우며 정제 및 정렬에는 고도의 기술이 필요하고 단일 말뭉치를 원하는 이종 언어로 번역하는 데는 많은 비용과 시간이 많이 필요하다. 즉 기본적인 NMT 학습을 위해서는 최소 200만 이상의 병렬말뭉치가 필요한데 이런 대용량의 인공지능 학습용 말뭉치를 준비하기는 쉽지 않은 실정이다.

이에 질 좋은 학습데이터를 구축하기 위한 연구들이 많이 진행되고 있다. 정제 및 필터링 작업을 거친 말뭉치로 학습을 한 모델이 그렇지 않은 모델보다 BLEU 점수가 더 높게 나오는 Parallel Corpus Filtering에 대한 연구가 진행 중에 있다[6-8]. 이는 통계기반 방식에서는 데이터의 양이 많으면 많을수록 좋았으나 딥러닝 방식에서는 데이터의 양보다는 데이터의 질이 더 중요함을 알 수 있다.

그러나 해당 연구들은 말뭉치 간의 편향성에 대한 것을 고려하지 않고 있다. Pretrain-Finetuning Approach(PFA)란 대용량의 데이터로 모델을 사전학습시킨 후 세부 작업을 위한 데이터로 미세조정을 진행하는 작업을 의미하며 최신 자연언어처리 연구의 핵심 트렌드이다. PFA를 활용하기 위해서는 사전학습을 위한 많은 양의 학습 데이터를 갖추는 것이 중요하다. PFA를 사용하는 많은 논문들에서 다양한 출처의 데이터를 하나의 데이터로 합쳐서 사전학습을 진행하고 있다. 그러나 다양한 출처의 데이터를 하나의 데이터로 합쳐서 사용하게 되면 데이터간의 불균형성으로 인하여 성능 하락이 발생할 여지가 존재한다. 대표적으로 기계번역의 경우 구어체, 문어체, 격식체가 혼용되어 있는 말뭉치로 학습을 시킬 경우 모델이 종결어미를 선택함에 있어서 혼란을 초래할 가능성이 존재한다.

본 논문은 다양한 출처의 말뭉치를 학습데이터로 사용할 때 말뭉치 간 균형성을 부여하여 성능을 향상 시키는 Corpus Weight Balance (CWB) 학습 방법론을 제안한다. 더불어 이러한 균형성과 고품질의 학습데이터를 구축할 수 있는 병렬 말뭉치 구축 가이드라인 또한 제안한다. 즉 고품질의 기계번역 데이터의 구축방향성을 제시하며 휴먼번역시장과 상생 가능한 데이터 구축 프로세스를 새롭게 제안한 [9]를 확장한 논문이다. [9]의 논문의 경우 대략적인 데이터 구축 프로세스에 대한 제안만 진행하였으며 본 논문은 해당 프로세스에 대한 자세한 설명과 더불어 데이터의 편향성을 고려한 Corpus Weight Balance (CWB) 학습 방법론을 새롭게 제안하였다.

2. Corpus Weight Balance 학습 방법론

최근 자연언어처리 연구들은 데이터의 양을 늘리고 모델의 파라미터를 대폭 상승시키고 있다. 또한 다양한 출처의 말뭉치들을 함께 합쳐 학습에 사용하고 있다. 대표적으로 RoBERTa[10], GPT3[11] 등이 존재한다. RoBERTa의 경우 학습데이터로 Book Corpus, English Wikipedia, CC-News, OpenWebText, Stories의 총 5개의 데이터를 합쳐 160GB의 텍스트 데이터로 언어모델을 학습시킨다.

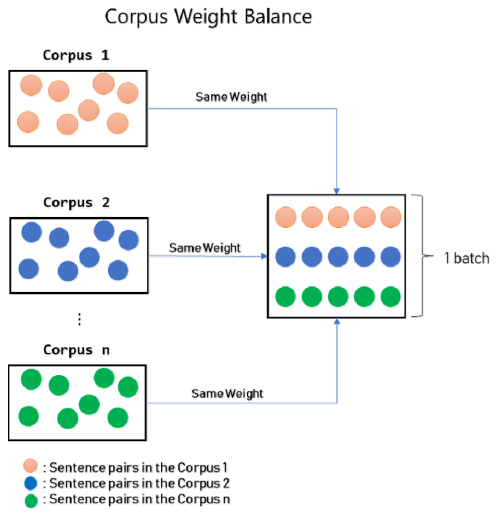

Fig. 1. Concept of Corpus Weight Balance

GPT3도 Common Crawl, WebText2, Books1, Books2, Wikipedia 등의 데이터를 합쳐 모델을 훈련하게 된다. 그러나 말뭉치 간의 특성 및 특징(어투, 문체, 도메인 등)이 다름에도 하나의 데이터로 합쳐서 훈련하는 것은 성능 하락과 직결될 수 있다. 이에 본 논문은 다양한 출처의 데이터들을 하나의 데이터로 합쳐서 훈련시키는 것이 아닌 각각의 말뭉치들에게 균등한 가중치를 부여하여 학습을 진행하는 Corpus Weight Balance (CWB) 학습 방법론을 제안한다.

Corpus Weight Balance란 각기 다른 출처의 말뭉치 간의 여러 차이들(어투, 문체, 도메인 등)로 인해 모델의 성능 하락이 발생할 수 있음을 완화시키는 학습 방법론이다. 예를 들어 한-영 기계번역을 학습함에 있어서 AIhub, OpenSubtitles, Iwslt 총 3가지 데이터를 함께 학습시킨 연구가 존재하는데 이는 말뭉치간의 균형성을 고려하지 않고 하나의 데이터로 합쳐서 훈련을 진행하였다[6]. 그러나 OpenSubtitles는 영화 자막 도메인, Iwslt는 음성인식 번역 도메인, AI Hub는 일반 도메인이나 문어체, 구어체 등이 혼용되어 있다. 이렇게 다양한 코퍼스를 하나의 데이터로 합쳐서 학습시키는 것보다 각각의 특성을 고려하여 균등한 비율로 학습시키는 것이 성능향상에 더 도움이 될 가능성이 존재한다.

세상에 존재하는 문서에는 여러 전문 도메인이 존재한다. 따라서 동일한 단어라도 도메인에 따라 다양하게 번역이 되는 경우가 존재한다. 예를 들어 “trans” 의 경우 경제 용어로는 “주식거래”, 일반용어로 “수송”, “번역” 등 10가지 이상으로 다양하게 쓰이게 된다. 이에 “trans”가 “주식거래”로 번역된 문장과 “수송”이라고 번역된 문장이 균형성 없이 함께 학습이 된다면 성능의 악영향을 미칠 수 있다. 따라서 문장 간의 비중을 동일하게 학습을 진행하는 것이 더 좋은 성능을 낼 수 있다.

이에 Vocabulary를 추출함에 있어 각각의 말뭉치에서 균등한 비율로 추출하는 방법론을 제안한다. 이로 인해 말뭉치 간의 편향성을 완화하여 성능 향상으로 이어질 가능성이 존재한다. 또한 Batch를 구성할 때도 말뭉치 간 동일한 비율로 구성하는 방법론을 제안한다.

Corpus Weight Balance의 구성도는 Fig. 1과 같다. 해당 구조도를 살펴보면 Corpus1부터 Corpus n까지 있다고 가정했을 때 각각의 말뭉치별로 동일한 비중을 두어 Batch를 구성함을 볼 수 있다. 또한 각 말뭉치에서 동일한 비율로 Vocabulary를 뽑아내어 어휘 단계에서도 말뭉치 간 균형성을 이룰 수 있다. 즉 딥러닝 학습 시 Batch 구성과 Vocabulary 구성에 있어서 말뭉치 간의 동일한 비율로 학습을 진행하여 균형성을 바탕으로 안정적인 모델을 제작할 수 있다.

3. 기계번역 학습용 데이터 구축 방안 제안

4차 산업혁명에 힘입어 정보의 교류와 소통이 급격히 늘어나 번역에 대한 수요가 급증하고 있다. 정보기술(IT) 산업의 성장에 따른 자동번역 기술의 발전에도 불구하고, 자동번역기의 한국어-영어 번역 성공률이 70~80%에 머물러 사용자들이 만족할 만한 수준에 미치지 못하고 있다. 아울러, 기업에서 기계번역을 서비스함에 있어 애로사항은 다음과 같다.

초기 번역 솔루션 구축의 비용과 시간의 장벽이 존재하고 양질의 데이터를 확보하기가 어렵다. 또한 NMT 성능품질 유지의 어려움이 있고 도메인에 특화된 언어쌍별 데이터 확보의 어려움 및 도메인에 특화된 NMT 솔루션 확보의 어려움이 존재한다. 즉 대부분의 애로사항은 번역 데이터 즉 병렬말뭉치의 부족으로 인한 문제이다. 또한 저작권이 해결된 데이터를 확보하기가 매우 어려우며, 확보에는 수많은 비용이 발생하여 이는 곧 인공지능 기반산업의 스타트업 기업이나 관련 기술혁신을 준비하는 기업에게는 매우 큰 애로사항으로 작용하고 있다.

일반적으로 단일 말뭉치의 경우 구하기 쉽고 충분한 양이 확보 가능하나, 병렬말뭉치의 경우 구하기도 만들기도 어려운 것이 현실이다. 또한 병렬말뭉치를 구축한

다고 하여도 실질적인 원본말뭉치의 정제 및 가공에 수많은 고도의 기술이 필요하며, 단일 말뭉치를 원하는 이종 언어로 번역하는 데도 많은 비용과 시간이 필요하다. 기본적으로 NMT 학습을 위해서는 최소 100만 ~ 200만 이상의 병렬말뭉치가 필요한데 이런 대량의 고품질 말뭉치를 준비하고 학습하기란 쉽지 않다.

온라인 무료번역은 단시간 내 빠른 번역 데이터 확보가 용이하여 구글 등 대기업이 대용량의 데이터를 중심으로 사업화가 용이하나, 번역 품질의 한계, 데이터 보안이 취약하여 고품질의 휴먼 번역과 상호 보완이 되어야 하는 문제점을 가지고 있다. 이에 본 논문은 고품질이 병렬말뭉치를 효율적으로 구축할 수 있는 프로세스를 제안한다.

병렬말뭉치를 구축하는 것은 많은 시간과 돈이 들며 대부분의 사람들은 Web Crawling을 통하여 Mono Corpus 데이터만 가지고 있다. 이러한 사람들을 위하여 병렬말뭉치를 만드는 Process를 제시하고자 한다. 전체적인 프로세스는 Fig. 2와 같다.

1단계로 먼저 원문의 전처리 작업을 위하여 Mono Corpus Cleaning작업과 Grammar correction 즉 문법교정기를 거쳐서 Mono Corpus의 질을 높인다. Web Crawling을 통하여 얻은 데이터들은 문법이 잘못된 경우가 많으며 검증되지 않은 데이터들이기 때문에 이 과정을 거쳐야 한다. 이를 통해 기존 Mono Corpus의 질을 한단계 상승시켜 줄 수 있다.

2단계로 이 원문을 NMT를 통해 번역을 진행한다. 기업 혹은 학교가 보유하고 있는 In-House NMT를 이용할 수도 있고 상용화 시스템을 이용할 수도 있다. 이를 통해 1차적으로 병렬 말뭉치가 구축이 된다.

3단계로 A.P.E(Automatic Post Editing)[12]를 거쳐 후처리를 진행하여 병렬말뭉치를 구축한다. 이를 통해 기계번역 된 문장의 질을 상승시킬 수 있다.

4단계로 원문과 A.P.E를 거친 번역 결과를 가지고 기계번역 품질 예측 즉 Quality Estimation[13]을 진행한다. 문장 수준의 성능 평가 척도로는 Pearson's correlation, Mean Average Error(MAE), Root Mean Squared Error(RMSE)이 사용되며 어절 및 구 수준의 성능 평가 척도로는 multiplication of F1-OK,F1-BAD가 사용된다. 해당 점수 척도를 이용하여 감수 Level을 정하게 된다. 총 3단계(High, Middle, Low)의 감수 Level로 분류하여 High인 부분만 선별하게 된다. High인 부분은 고품질의 병렬 말뭉치로 간주하여 데이터로 즉시 사용하며 Middle과 Low 부분은 가격을 다르게 측정하여 휴먼번역을 맡겨 데이터의 질을 높인다.

5단계는 선택사항으로 4단계 직후 Back Translation[14]을 적용하여 2차 감수를 진행하게 된다.

마지막으로 2차 감수 후 사용자는 이 문장을 말뭉치 데이터로 사용할지 아니면 저렴한 가격으로 번역 감수를

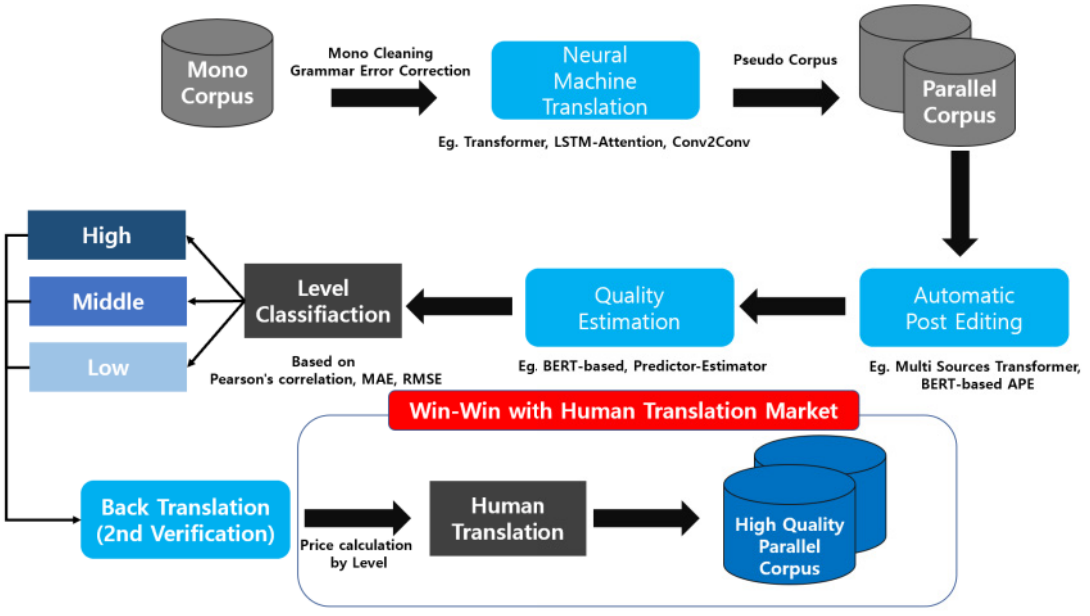

Fig. 2. Schematic diagram of the proposed parallel corpus constructing process for Neural Machine Translation

맡길지 결정하게 된다. 이를 통해 전문 번역가 2차 감수 과정을 통하여 양질의 인공지능 학습용 데이터를 구축하게 될 것으로 기대된다.

데이터의 품질 향상을 위하여 궁극적으로 사람의 손을 거치는 것이 가장 신뢰성 있고 고품질의 데이터가 구축될 수 있다. 그러나 모든 데이터에 대하여 사람의 손을 거쳐서 데이터를 구축하게 된다면 막대한 비용과 시간이 투자되어야 할 것이다. 따라서 어느정도 컴퓨터가 자동으로 품질을 판단하여 어느 일정 수준 이상의 품질이면 사람의 손을 거치지 않고 일정 수준 미만의 데이터만 사람의 손을 거쳐 검증 및 후처리 작업을 진행하면 좋을 것이다. 따라서 기계번역의 휴먼번역 시장을 함께 활성화 시킬 수 있는 데이터 구축 방법론을 제안하였다.

이를 위하여 많은 양의 문서를 번역하고 관리하기 위한 전문 크라우드 소싱 플랫폼은 필수 시스템이다번역 작업을 프로젝트와 테스크로 나누고 해당분야 고급 전문 번역가를 선정하여 일을 나누어 주고 번역 진행 과정을 관리해야하며 번역가는 크라우드 소싱 플랫폼을 통해 NMT 번역 결과와 TM(번역메모리), 용례 사전을 참조하여 모든 번역가들이 일관된 번역을 할 수 있도록 지원해야한다.

본 프로세스의 장점으로 먼저 휴먼 번역 2차 감수 가격을 다르게 책정하여 비용을 줄일 수 있다. High level 같은 경우 이미 양질의 데이터이기에 감수 작업을 진행하지 않거나 시간을 조금만 들여도 쉽게 감수 작업을 진행 할 수 있다. 반면 Low level 같은 경우 집중적으로 심도 있게 감수 작업을 진행해야 한다. 이를 통해 시간을 단축할 수 있으며 감수 작업의 효율을 향상시킬 수 있다.

결론적으로 시간과 비용의 절약 및 감수 작업의 효율을 향상시켜 고품질 병렬말뭉치를 확보 할 수 있다. 이를 통해 고품질의 인공지능 학습용 데이터가 구축이 가능해져 비전문가도 필요로 하는 해외 데이터를 손쉽게 얻는 것이 가능해짐으로 '정보화 격차'의 해소에 도움이 될 뿐 아니라 국가 경쟁력 제고에 기여할 수 있다. 또한 이기종 산업 같은 성공사례를 전파시키고, 유사 산업군 광학인식(OCR), 사물인터넷(IoT), 로봇(Robot)간의 융합 가능성을 제시함으로 인해 콜라보를 통한 상생 전략이식이 가능하며 다국어기반 신규 번역가 양성 및 관련 산업 분야 고용창출 효과를 기대할 수 있다. 번역 생산성 및 번역 수요 증가로 인한 시니어 일자리 창출 및 사회고용 안정화에도 기여 가능하다.

4. 실험 및 실험결과

4.1 데이터 및 모델

Corpus Weight Balance 방법론의 성능 검증을 위한 학습데이터로 AI Hub, OpenSubtitles, Iwslt에서 공개한 한-영 병렬 말뭉치를 이용하였다. 총 2,781,758 문장 쌍의 병렬 말뭉치가 구축되었다.

모델은 Vanilla Transformer를 이용하였으며 모든 하이퍼파라미터는 [1]의 논문과 동일하게 설정하였다. Subword Tokenization 같은 경우 google의 sentencepiece를 적용하였으며 vocab size 같은 경우 32,000개로 설정하였다. Vocab의 경우 AI Hub, OpenSubtitles, Iwslt 데이터에서 1대1대1 비율로 동일하게 추출된다. 이것이 Corpus Weight Balnce 학습 방법론의 차별화된 특성으로 정의할 수 있다.

검증 데이터는 학습데이터 에서 5,000개를 랜덤하게 추출하였으며 성능평가는 BLEU 점수[15]를 기준으로 진행하며 Moses의 multi-bleu.perl script를 이용한다.

테스트 셋의 경우 기존의 연구들은 Iwslt 16, Iwslt 17을 사용하였으나[2,4,9] 해당 데이터는 TED 도메인을 기반으로만 구축되었기에 다양한 도메인을 포괄하면서 데이터의 중립성을 보증할 수 있는 네이버 사전 예문을 크롤링하였다. 즉 모델의 성능 평가를 위한 테스트셋으로 네이버 어학사전에 존재하는 예문을 크롤링(crawling)하여 3000개의 한-영 문장 쌍으로 설정하였다.

4.2 실험결과

제안하는 Corpus Weight Balance 방법론에 대한 검증을 위하여 말뭉치의 비율을 고려한 모델과 그렇지 않은 모델 간의 비교실험을 진행하였다. Decoding 시 두 모델 모두 Beam size는 5로 통일하였다. 실험결과는 Table 1과 같다.

Table 1. Experimental Results

Model	BLEU
Vanilla Transformer	22.75
Corpus Weight Balance	**23.55**

실험결과 Corpus Weight Balance를 적용한 학습 방법론이 그렇지 않은 모델보다 더 좋음을 알 수 있었다. 이는 학습 데이터간의 불균형성이 성능의 악역향을 미침을 알 수 있었으며 다른 최신 자연언어처리 연구에도 각

기 다른 출처의 데이터 간에는 균등한 비율로 학습시키는 것이 더 성능이 좋을 수 있음을 나타낸다.

5. Conclusion

본 논문은 기계번역에서 말뭉치 간의 불균형성을 완화할 수 있는 Corpus Weight Balance 학습 방법론을 제안하였다. 이를 통해 사전학습 시 출처가 다른 데이터 간의 불균형성을 완화하고 성능을 더 향상시킬 수 있음을 실험을 통해 증명하였다. 해당 방법론은 최신 대용량 언어모델 연구에도 충분히 활용될 수 있는 방법론이다. 더불어 NMT에서 중요한 요소 중 하나인 고품질의 병렬말뭉치를 구축하는 정책과 프로세스를 제시하였다. 추후 해당 방법론을 이용하여 PFA와 도메인특화 기계번역에도 적용해볼 예정이며 제안하는 프로세스를 기반으로 실제 데이터 구축을 진행할 예정이다. 본 논문을 기반으로 국가적 차원에서 고품질의 학습 데이터가 구축되기를 기대해본다.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser & I. Polosukhin. (2017). Attention is all you need. *In Advances in neural information processing systems,* 5998-6008.

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

[3] C. Park, C. Lee, Y. Yang & H. Lim. (2020). Ancient Korean Neural Machine Translation. *IEEE Access, 8,* 116617-116625.

[4] 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

[5] K. Song, X. Tan, T. Qin, J. Lu & T. Y. Liu. (2019). Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450.*

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

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

[8] Sen, Sukanta, Asif Ekbal, and Pushpak Bhattacharyya. "Parallel Corpus Filtering based on Fuzzy String Matching." *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2).* 2019.

[9] C. J. Park, Y. D. Oh, J. K. Choi, D. P. Kim & H. Lim. (2020). Toward High Quality Parallel Corpus Using Monolingual Corpus. *The 10th International Conference on Convergence Technology (ICCT 2020), Volume 10,* 146-147.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen & V. Stoyanov. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal ... & D. Amodei. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

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

[13] E. Fonseca et al. (2019). "Findings of the WMT 2019 Shared Tasks on Quality Estimation." *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2).* 2019.

[14] S. Edunov et al. (2018). "Understanding back-translation at scale." *arXiv preprint arXiv:1808.09381.*

[15] 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* 311-318.

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

박 기 남(Kinam Park) [정회원]

- 2004년 2월 : 백석대학교 컴퓨터학과 (이학학사)
- 2006년 2월 : 한신대학교 컴퓨터정보학과(이학석사)
- 2011년 8월 : 고려대학교 컴퓨터교육학과(이학박사)
- 2011년 9월 ~ 현재 : 고려대학교 연구교수
- 관심분야 : 인공지능, 인지과학, 스마트교육
- E-Mail : spknn@korea.ac.kr

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

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

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