Journal of The Korea Convergence Society
Vol. 13. No. 2, pp. 21-29, 2022
e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2022.13.02.021

언어적 특성과 서비스를 고려한 딥러닝 기반 한국어 방언 기계번역 연구

임상범[1], 박찬준[2], 양영욱[3*]
[1]강남대학교 소프트웨어응용학부 학생, [2]고려대학교 컴퓨터학과 석박사통합과정, [3]한신대학교 컴퓨터학과 교수

Deep Learning-based Korean Dialect Machine Translation Research Considering Linguistics Features and Service

Sangbeom Lim[1], Chanjun Park[2], Yeongwook Yang[3*]
[1]Student, Department of Software Application, Kangnam University
[2]Master & Ph.D. Combined Student, Department of Computer Science and Engineering, Korea University
[3]Professor, Department of Computer Science and Engineering, Hanshin University

요 약 본 논문은 방언 연구, 보존, 의사소통의 중요성을 바탕으로 소외될 수 있는 방언 사용자들을 위한 한국어 방언 기계번역 연구를 진행하였다. 사용한 방언 데이터는 최상위 행정구역을 기반으로 배포된 AIHUB 방언 데이터를 사용하였다. 방언 데이터를 바탕으로 Transformer 기반의 copy mechanism을 적용하여 방언 기계번역기의 성능 향상을 도모하는 모델링 연구와 모델 배포의 효율성을 도모하는 Many-to-one 기반의 방언 기계 번역기를 제안한다. 본 논문은 one-to-one 모델과 many-to-one 모델의 성능을 비교 분석하고 이를 다양한 언어학적 시각으로 분석하였다. 실험 결과 BLEU점수를 기준으로 본 논문이 제안하는 방법론을 적용한 one-to-one 기계번역기의 성능 향상과 many-to-one 기계번역기의 유의미한 성능을 도출하였다.

주제어 : 한국어 방언 기계번역, 기계번역, 트랜스포머, 다언어 기계번역, 언어 융합

Abstract Based on the importance of dialect research, preservation, and communication, this paper conducted a study on machine translation of Korean dialects for dialect users who may be marginalized. For the dialect data used, AIHUB dialect data distributed based on the highest administrative district was used. We propose a many-to-one dialect machine translation that promotes the efficiency of model distribution and modeling research to improve the performance of the dialect machine translation by applying Copy mechanism. This paper evaluates the performance of the one-to-one model and the many-to-one model as a BLEU score, and analyzes the performance of the many-to-one model in the Korean dialect from a linguistic perspective. The performance improvement of the one-to-one machine translation by applying the methodology proposed in this paper and the significant high performance of the many-to-one machine translation were derived.

Key Words : Korean Dialect Translation, Machine Translation, Transformer, Multilingual Translation, Language Convergence

*This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C2004868).
*Corresponding Author : Yeongwook Yang(yeongwook.yang@gmail.com)
Received November 23, 2021 Revised January 3, 2022
Accepted February 20, 2022 Published February 28, 2022

1. 서론

방언은 표준어의 음운, 문법, 어휘 등의 영역에서 독립된 체계를 가지는 언어의 변종으로서 하나의 독립적인 언어로 간주할 수 있다. 주로 특정 집단에서 사용되며 각 지역의 사투리를 대표적인 예시로 들 수 있다. 즉 방언은 지리 및 사회적으로 구획된 특정 집단에서 사용하는 언어이며[1] 독립적인 언어로도 간주할 수 있다.

이러한 특수성에도 불구하고 방언 사용자를 위한 방언 서비스는 턱없이 부족하며 방언 사용자가 표준어 기반의 서비스를 혼용하는 사회적 현실에 놓여있다. Natural Language Process(NLP)의 발전으로 챗봇, 감정분석, 페르소나(Persona)를 활용한 대화 시스템, 맞춤법 교정[2]과 같은 태스크(task)들은 모두 표준어 데이터 기반으로 생성되어 방언 사용자들의 사용에서 제한이 있다. 현존하는 서비스를 방언 사용자가 원활하게 이용할 수 있기 위해서는 표준어를 기반으로 생성된 서비스를 사용자의 언어에 맞는 방언으로 번역하는 것이 필요하다. 이를 위해 사람이 직접 번역을 하는 것이 가장 정확할 수 있으나 번역은 번역가들의 배경 지식, 번역 실력에서의 편차를 가지고 있기에 일관된 품질의 번역을 기대하기 힘들다[3]. 또한 시간과 비용이 많이 사용되는 이면이 존재한다. 적은 비용으로 고품질의 번역 결과를 도출하기 위해서는 기계번역기의 사용이 불가피하다.

기술적 측면에서는 방언에서만 사용되는 어휘들은 기존의 상용화된 번역기가 이해하고 번역하기에는 한계가 있다. 구글번역기[1]의 경우 일부 방언 문장을 입력으로 사용하였을 때 방언을 오인하여 빈번하게 오번역이 발생하는 것을 알 수 있었다. 많은 도메인에서 사용되는 상용화 모델이 방언에 최적화 된 모델로 발전하는 것은 많은 어려움이 있다. 방언 번역기를 통하여 방언을 고품질의 표준어로 번역할 수 있다면 기존의 표준어를 기반으로 개발된 NLP 서비스의 언어적 한계를 뛰어넘어 다양한 NLP 서비스를 방언 사용자들에게도 제공할 수 있을 것이다.

방언 번역기는 상용화 모델로서도 사용될 수 있지만 방언 보존에 중요한 언어 데이터로서도 가치가 있다. 현재 사용되고 있는 방언지도, 방언사전과 같은 방법들은 단어-to-단어에 매핑되는 수준에 그친다. 방언 번역기는 기존의 단어별 번역에서 그치던 방언 보존 방식을 문장-to-문장 수준으로 발전시킨다. 문장간 번역을 통해 각 언어들의 화용론적 의미와 통사론적 구조를 반영함으로서 번역기가 방언과 표준어 각각의 다양성과 한국어라는 언어적 통일성을 번역 결과에 적용하여 출력한다.

본 논문은 방언 보존에 대한 방법을 새롭게 제시하며 이를 저자원 언어(Low-Resource Language, LRL) 번역 연구로 해석하고 지도학습 기계번역의 측면에서 다음과 같은 연구를 진행하였다.

첫째로 모델링 측면에서의 연구를 진행하였다. Vanilla Transformer 모델[4]과 Transformer모델에 카피 메커니즘(copy mechanism)[5]을 적용하였을 때의 성능비교를 1차적으로 진행하였다. 방언은 표준 한국어와 상당부분 단어의 토큰을 공유하면서 지역 방언 고유의 단어를 갖고 있다. 방언과 표준어 번역시에는 각 방언이 가진 다양성을 고려하며 표준어와의 언어적 통일성을 모두 포함해야한다. 본 논문에서는 이를 디코딩 과정에서 단어를 생성할 때 인코더의 입력을 참고 및 복사하여 문장을 생성하는 카피 메커니즘 사용과 학습된 방언 언어 모델의 혼용을 통해 두 요소를 만족할 수 있도록 설계하였다.

둘째, 효율성 측면에서의 연구를 진행하였다. 대부분의 기계번역 연구는 one-to-one 기반으로서 원시 언어 A와 목적언어 B 간의 번역만 가능하였다. 최근에는 다언어(Multilingual) 모델의 발전으로 many-to-one, many-to-many와 같은 서비스 측면의 장점을 가진 모델들이 연구되고 있다. 이러한 모델들은 기존의 one-to-one 번역 모델들에 비해 성능은 떨어지나 서비스 측면에서의 효율이 높은 장점을 가지고 있다[6,7]. 방언 번역기로서의 목적은 한가지의 방언에 대한 번역보다 다양한 방언을 적은 비용으로 서비스를 제공하는 것이 기업과 사용자가 원하는 방향일 것이다. Many-to-one 방언 번역기는 one-to-one 모델보다 적은 파라미터와 작은 사이즈를 통해 모델 관리와 계산속도에서의 이점을 통한 서비스 측면의 장점을 갖는다.

본 논문은 대한민국의 다양한 방언이 존재한다는 특성을 바탕으로 many-to-one 표준어-방언 기계번역기를 개발하였고 이를 one-to-one 번역기와 비교 분석을 진행하였다. 또한 many-to-one 기계번역기의 모델 수용성을 고려하여 적합한 파라미터로 튜닝한 many-to-one 번역기를 개발하여 성능 향상을 기대하였다.

1) https://translate.google.co.kr/?hl=ko

2. 관련연구 및 배경지식

많은 나라에서 수도에서 사용되는 방언을 표준어로 삼고 있으며 우리나라 또한 서울말을 표준어로 삼고 있다. 표준어가 모두 서울말로 구성되어 있는 것은 아니다. 표준어는 여러 방언에서 기준에 맞는 요소를 가려 뽑고 다듬은 말이기에 실제 서울말과는 다를 수 있다. 예시로는 경상도 사투리였던 '멍게'가 1988년 표준어 규정 개정으로 '우렁쉥이'와 함께 복수 표준어가 되었다[8]. 표준어는 방언들의 합성으로 구성된 것임으로 방언 연구와 보존은 표준어의 변화와 발전에 지대한 영향을 미친다.

2.1 자연언어처리 기반 한국어 방언 보존 연구

국내에서의 자연어 처리 기반의 연구는 매우 미비하며 방언 중 고유한 발음기호를 가진 제주어만 카카오 브레인에서 연구되었다. 카카오는 방언 연구와 함께 Jejueo Datasets for Machine Translation and Speech Synthesis[9] 데이터 또한 공개하였다. 해당 데이터는 17만개의 병렬 말뭉치와 13시간 47분 분량의 음성 데이터로 구성되어 있으며 제주어의 아래아 발음기호 적용과 제주어 발화 특성을 반영한 데이터를 통해 발전과 보존에 밑거름을 마련하였다.

2.2 해외 방언 기계 번역 연구

해외에서는 이미 방언 관련 연구들이 많이 진행되어 왔다. 해외에서의 방언, 유사한 언어 간의 번역은 규칙 기반 혹은 통계기반의 모델들이 주로 채택되었었다[10]. 유사 언어 번역으로서는 아일랜드어와 스코틀랜드어의 번역 모델인 ga2gd 기계번역 시스템이 존재하였다[11]. 이 외의 아랍어, 독일어, 포르투갈 등 방언을 보유한 나라에서 방언관련 연구가 진행되고 있다. 가장 활발한 연구가 진행된 대표적 방언으로는 아랍어[12]와 중국어를 꼽을 수 있다. [13]은 기존의 Neural Machine Translation(NMT)기반의 모델들이 입력 언어에 대한 단어 임베딩과 타겟 언어의 임베딩을 모두 학습하여야 했던 전통적인 방법과 달리 입력 소스와 타겟 언어에 같은 벡터 공간을 사용한다. 이는 번역 모델로 하여금 표준어를 기반으로 방언들의 동의어가 비슷한 벡터들을 갖도록 하였다.

중국어의 경우 비지도 학습을 통한 방언 번역의 사례가 존재한다. [14]은 파라미터 공유를 통하여 방언들 간의 잠재공간을 유사하게 만드는 방법, 기존의 Transformer 모델과 달리 낮은 층의 자질 또한 공유하여 구문 단위의 특징도 반영할 수 있게 하였다. 또한 Pivot-Private 임베딩을 통한 방언 문장의 유사성과 각 방언의 특징을 모두 반영한 연구들이 진행된바 있다.

3. 제안하는 한국어 방언 기계번역 모델

본 논문은 one-to-one기반 방언 기계번역기와 더불어 many-to-one 방언 기계번역기를 개발하여 이에 대한 비교 분석을 진행하였다.

3.1 One-to-one 기반 방언 기계번역기

One-to-one 기계번역은 하나의 모델이 입력 언어 A와 타겟 언어 B로 번역된 결과 값을 출력하는 것을 의미한다. 본 논문은 one-to-one 및 Transformer[4]와 카피 메커니즘[5]을 적용한 방언 기계번역 모델을 개발하였다. 제안하는 one-to-one 모델의 구조는 Fig.1.과 같다.

Transformer Transformer 모델은 sequence-to-sequence 기반의 자기회귀 모델로서 현재 NMT 분야 뿐 아닌 NLP분야에서 다양하게 사용되고 있다. Transformer의 특징으로서는 기존의 방법론들이 복잡한 회귀적 구조를 가졌던 반면 Transformer는 어텐션만을 사용하였다. 모델의 구조를 간단하게 하였지만 멀티헤드 어텐션을 통해 훈련 과정의 병렬화를 구현하여 학습의 성능은 향상하는 효과를 가졌다. 어텐션을 사용한 최종 출력은 다음의 수식으로서 표현될 수 있다.

$$Attention = (Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Transformer 구조는 인코더-디코더 구조를 통해 입력문과 출력문의 구조를 파악한다. 어텐션에서 사용되는 Q는 모든 시점에서의 디코더 셀에서의 은닉 상태들을 의미하며 K와 V는 모든 시점에서의 인코더 셀의 은닉 상태로 정의할 수 있다. 수식의 Q, K, V는 입력 문장의 단어벡터와 서로 다른 가중치 행렬을 곱하여 생성된 벡터이다. 입력 문장으로부터 구한 Q, K, V 벡터들은 어텐션 스코어를 구할 때 사용된다. 각 Q 벡터는 모든 K 벡터에 대해 어텐션 스코어를 계산하며 이를 V 벡터와 가중합하여 최종 어텐션 결과값을 출력하게 된다. 셀프

어텐션 구조는 가중치 행렬 연산 이전의 Q, K, V 값이 모두 같아 자신에게 어텐션을 수행하는 효과를 갖는다. 반면 디코더 두 번째 층의 어텐션은 K, V 값이 인코더 마지막 층의 결과값을 사용하는 점에서 셀프 어텐션과는 다르다.

Copy Mechanism 언어 모델이 학습 과정에서 습득할 수 있는 어휘는 무한하지 않다. 고유명사나 사람 이름의 경우 학습 데이터에 빈번하게 출현하지 않고 매우 드물게 포함된다. 이러한 토큰의 경우 출력 언어로의 문장을 생성할 때 필요한 어휘가 출력 사전에 존재하지 않는 문제가 발생할 수 있다. 카피 메커니즘은 out-of-vocabulary(OOV)문제를 해결하기 위해 고안되었다. 모델 학습에서 사용된 방언과 표준어 데이터 쌍은 한국어를 근간으로 생성되었으며 병렬 데이터 쌍의 문장들이 단어를 기준으로 많은 양을 공유하고 있음을 발견하였다.

입력문장을 번역할 때 언어적 변환이 필수적이지 않다면 모델의 언어생성보다 입력 문장을 복사하여 사용하는 것이 효과적일 수 있다. 이에 본 논문은 모델의 디코딩 과정에서 카피 메커니즘을 적용하여 번역 결과를 향상시키고자 한다. 카피 메커니즘은 디코딩 과정의 각 시점에서 단어를 생성할 때 출력 어휘 사전의 어휘와 입력 문장의 단어들을 카피 어텐션 스코어(Copy Attention Score)로서 함께 고려하여 단어를 선택하는 구조를 가지고 있다.

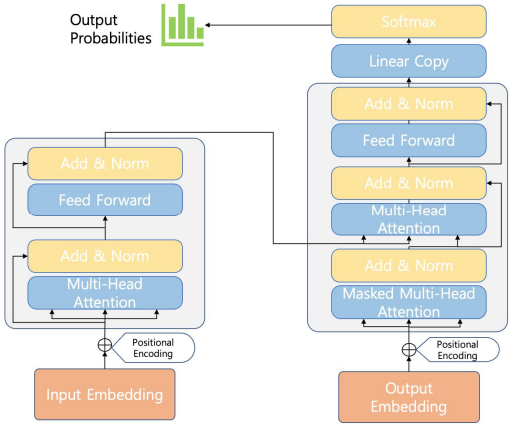

Fig. 1. One-to-One model architecture using copy mechanism

3.2 Many-to-one 방언 기계번역기

Many-to-one 기계번역 모델은 기존의 one-to-one 방식의 언어 다양성의 한계를 보완한 모델로서 다양한 원시언어를 입력으로 투입할 수 있다. Many-to-one 모델은 원시 언어의 개수만큼의 인코더를 사용하는 방식이 아닌 하나의 인코더 층에서 모든 언어에 대한 자질을 추출할 수 있도록 학습과정을 설계한다. 이후에는 기존의 one-to-one 방식과 비슷한 언어 생성의 과정을 갖는다. 이러한 다언어 모델의 사용은 번역 과정에서 많은 모델과 파라미터들이 사용되었던 one-to-one 방식에 비해 적은 파라미터와 용량으로 효율적으로 태스크들을 수행할 수 있게 되었다[6,7].

Many-to-one은 동일한 실험 환경에서 one-to-one 모델과 비교시에 성능이 소폭 하락하는 한계가 존재한다[15]. Many-to-one 번역 모델은 one-to-one에 비해 서비스 효율, 환경적 측면에서 이점을 더 갖는 모델이다. 예시로 번역의 서비스를 제공하는 기업의 경우 많은 언어들을 적은 파라미터를 통해 효율적으로 서비스하는 것을 목표로 한다. 기업의 입장에서 많은 모델들을 관리하는 것은 쉽지 않다. 다언어 모델을 사용한다면 하나의 모델로 다양한 언어를 서비스 할 수 있어 서비스 환경적 측면에서 많은 이점을 가질 것이다. 이러한 장점들을 바탕으로 최근 one-to-one 모델을 넘어선 many-to-one, one-to-many, many-to-many와 같은 다언어 NMT [7,16] 연구가 이루어지고 있다.

본 논문은 사용 가능한 방언-방언 병렬 말뭉치가 존재하지 않는 한계로 인해 다언어 모델의 모든 방법론을 실험하지 않고 다양한 방언언어와 표준어로의 번역이 가능한 many-to-one 모델을 개발하여 one-to-one 모델과의 비교 실험을 진행하였다. 제안하는 모델의 구조는 Fig. 2.과 같다.

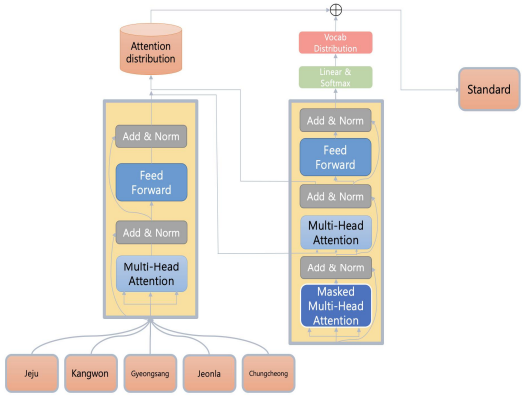

Fig. 2. Many-to-One model architecture

Table 1. Korean Dialect Data quantitative information

	Before Filtered	After Filtered	Average Length	Maximum Length	Minimum Length
Jeju	2,2215,694	909,758	22.049	338	5
Kangwon	1,850,036	966,714	21.202	203	5
Chungcheong	1,403,085	212,368	36.518	147	5
Gyeongsang	1,568,415	257,579	33.691	177	5
Jeonla	1,094,390	156,092	29.036	119	5
JIT	320,712	269,189	28.470	350	5

학습된 5개의 방언 중 하나의 언어 문장을 모델의 입력으로 투입하면 인코더 단의 층을 거듭하며 입력의 특징을 함축적으로 포함하게 되는 표현(representation)으로 추출된다. 인코더 마지막 층에서는 두개의 sub-layers를 통과하며 압축된 표현과 어텐션 스코어가 결과로 출력된다. 인코더의 최종 결과 값은 디코더의 입력으로 사용되며 생성 층에서는 카피 메커니즘에 사용되는 카피 어텐션 스코어가 함께 고려되어 언어를 출력하게 된다. 언어 생성시 각 시간별 출력 단계에서 입력 열의 단어 중 카피 어텐션 스코어가 가장 높은 단어와 디코더가 출력 어휘 사전에서 추론한 언어를 함께 고려하여 최종적으로 출력할 언어를 선택하게 된다.

4. 실험 및 실험결과

4.1 데이터 셋

본 논문은 실험을 위한 데이터로 AIHUB[2]에서 공개한 음성/자연어 데이터 셋 중 한국의 최상위 행정구역을 기준으로 분류한 한국어 방언 발화 데이터를 이용하였다. 데이터 셋은 10~60대 방언 사용자의 일상 대화 발화를 전사하여 음성 데이터와 매핑(Mapping)된 담화 텍스트 말뭉치로 구성되어있다. 총 5가지 데이터 종류를 포함하고 있으며 제주어, 강원어, 경상어, 전라어, 충청어의 다양한 방언 데이터를 실험에 사용하였다.

실험에 앞서 모델 성능 향상을 위해 모든 발화 데이터를 입력으로 사용하지 않고 불필요한 데이터 일부를 제외시켰다. 모든 데이터 셋이 발화를 바탕으로 생성되었기에 단답식의 문장이 다수 존재하였다. 평균 문장 길이와 큰 차이를 보이는 데이터는 모델의 학습에 큰 도움이 되지 않기에 5이하의 문장 길이를 가진 문장 데이터는 학습에서 제거하여 실험하였다.

첫째 제주어의 경우 두 가지의 데이터를 사용하였다. AIHUB의 방언 발화 데이터와 카카오 브레인의 JIT 데이터[9]를 사용하였다. AIHUB 제주어 데이터 셋은 총 909,758 문장으로 구성되어 있으며 평균 문장의 길이는 22.049를 가진다. JIT 데이터는 문장의 최대 길이가 2,689를 가진 데이터도 존재하였다. JIT 데이터 셋은 문장 길이가 평균적으로 제주어 AIHUB 데이터에 비해 길었으며, 제주어 데이터 셋의 최대길이를 기준으로 350 이상의 길이를 가진 데이터는 평균 문장에 비해 매우 큰 길이를 갖는 것으로 판단하여 역시 학습에서 제외하였다.

둘째 강원 데이터는 필터링 이전 50만 건 이상의 어절 데이터를 포함하고 있는 방언 발화 문장을 사용하였다. 필터링 이전의 문장 개수는 1,850,036 문장이며 필터링 이후의 실험 데이터 셋으로 966,714 문장을 사용하였다. 평균 문장 길이는 21.202로 전체 데이터 중 가장 작은 평균 문장 길이를 보유하였다.

셋째 충청어는 필터링 이전 1,403,085문장으로 구성되었으며 필터링 이후 212,368문장을 학습 데이터 셋으로 사용하였다. 문장 길이는 모든 데이터 중 가장 긴 평균 문장 길이를 보유하였다.

넷째 경상어는 한국어 발화 문장 1,568,415 개의 문장을 필터링 이전 보유하였으며 필터링 이후 257,579 문장을 학습 데이터로 사용하였다. 평균 문장 길이는 33.691의 평균 문장 길이를 보였다.

다섯째 전라어는 AIHUB 데이터 중 가장 적은 문장 개수를 보였다. 필터링 이전 1,094,390문장을 포함하였으며 필터링 이후 156,092문장, 29.036 평균 문장 길이를 보였다. 전반적인 데이터 셋에 대한 설명은 Table 1과 같다.

4.2 모델 및 세팅

One-to-one 기계번역 모델은 카피 메커니즘의 적용을 통해 언어적 통일성 측면에서의 방언-표준어 번역 성능 향상을 도모하고, one-to-one 번역 모델과 many-to-one 기반의 번역 모델들의 비교 실험을 진행하였다.

2) https://aihub.or.kr/aihub-data/natural-language/about

4.2.1 One-to-one 번역 모델

데이터 및 전처리 학습데이터와 검증 데이터 셋으로 AIHUB의 방언 데이터와 카카오브레인의 JIT 데이터를 전면 사용하였다. 데이터 전처리의 경우 32,000 크기로 설정한 Sentencepiece[17] 서브워드 분리기를 사용하였다.

모델 및 성능 평가 모델의 학습에 사용된 하이퍼파라미터는 다음과 같다. batch size 4096, optimization에 adam optimizer사용, Noam Decay와 5K warmup step을 사용하여 learning rate를 조절하였다. 6개의 인코더-디코더 층과 8개의 attention head를 사용하였다. Model dimension은 512로 설정하여 학습을 진행하였다. 모든 번역결과에 대한 성능평가는 BLEU 점수를 기준으로 진행하며 Moses의 multi-bleu.perl script[3]를 이용한다.

4.2.2 Many-to-one 번역 모델

데이터 및 전처리 JIT 데이터는 제주어와 같은 범주로 간주하여 many-to-one 모델 학습에 포함하지 않았다. 데이터 전처리의 경우 기존의 모델 대비 다양한 언어의 학습과 학습 데이터의 크기가 증가함에 따라 64,000크기로 Sentencepiece 서브워드 분리기를 학습시켜 사용하였다.

모델 및 성능 평가 기존에 one-to-one 번역 모델과 다르게 5개의 언어를 하나의 인코더가 학습하고 좋은 성능을 기대하기 위해 파라미터들을 증가시키어 실험하였다. 참고한 인용 모델[7]의 파라미터는 model dimension은 기존의 512차원에서 1024차원으로 증가하였고, feed-forward network의 hidden dimension 또한 2028에서 8192, Attention head는 8개에서 2배 증가한 16개로 증가하였다. 본 논문은 실험환경의 제약으로 인해 model dimension은 768과 attention head 16만 증가시킨 튜닝(Tuned)모델을 사용하였다. feed-forward network의 파라미터는 one-to-one모델과 동일한 파라미터를 사용하였다. 증가시키지 못한 파라미터에 대한 보상으로 layer층의 개수를 6개에서 7개로 증가시켜 실험하였다. Dropout rate는 모델의 학습량을 고려하여 0.2로 설정하였다. 성능평가는 마찬가지로 BLEU 점수를 기준으로 진행하였다.

4.3 실험 결과

4.3.1 Copy mechanism 비교 실험 결과

Table 2. Model Performance Comparison between according to Copy mechanism application

	Transformer + Copy	Transformer
Jeju	**69.11**	64.84
Kangwon	74.21	**78.34**
Gyeongsang	77.80	**79.01**
Jeonla	76.29	**78.67**
Cheongchung	80.35	**82.76**
JIT	43.85	**47.68**

카피 메커니즘을 적용했을 때의 실험결과는 Table 2와 같다. 실험결과 모든 언어에서 카피 메커니즘의 적용이 성능향상을 도모하지 않는다는 것을 알 수 있었다. 학습 이전의 전처리 과정에서 데이터의 구성 중 소스(Source)와 타겟(Target) 문장이 같은 데이터는 모델이 입력을 그대로 복사하는 것을 방지하기 위해 필터링하였었다. 그럼에도 불구하고 카피 메커니즘의 적용과 훈련량의 증가는 모델의 학습이 입력을 변경하기보다 그대로 유지하여 출력하는 쪽으로 모델의 학습이 이루어졌다. 모델의 학습 편향은 Table 3의 결과를 통해 확인할 수 있다. Table 3는 카피 메커니즘을 적용한 모델보다 카피 메커니즘을 적용하지 않은 모델이 뛰어난 성능을 보이는 언어들에서의 세부 모델 성능을 분석한 표이다. 모델이 예측한 번역문과 입력문 그리고 정답문간의 BLEU점수와 TER(Translation Error Rate)[4]점수를 비교 후 유사도를 계산하였다. BLEU점수는 번역문-입력문에서 번역문-정답문의 점수를 차감하여 기록하였다.

같은 언어의 Copy와 No Copy 모델의 비교에서 BLEU 점수가 낮은 모델이 입력을 그대로 복사하는 학습 편향이 나타났다고 할 수 있다. TER 점수의 경우 점수가 낮을수록 유사도가 높은 것으로서 비교한 모든 언어에서 번역문-입력문의 TER점수가 번역문-정답문의 TER점수에 비해 낮은 점수를 기록하였다. 본 실험을 통해 번역문과 입력문간의 품질 차이가 낮으며 모델의 학습과정에서 기존 입력문의 형태를 유지하려는 방향으로 학습이 이루어진 것으로 판단할 수 있다.

3) https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

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

Table 3. Similarity Comparison Score table

Data		BLEU		TER
		Copy	No copy	No copy
Chungcheong	Prediction-Target	2.79	3.59	20.10
	Prediction-Source			17.66
Kangwon	Prediction-Target	1.26	1.9	11.59
	Prediction-Source			10.23
Jeonla	Prediction-Target	0.21	0.25	16.27
	Prediction-Source			14.82

4.3.2 One-to-one 및 Many-to-one 비교 실험

5개의 AIHUB 방언 데이터들을 기반으로 5개 방언->표준어 기계번역 모델을 개발하였다. Many-to-one 모델을 같은 번역 태스크를 기준으로 one-to-one, one-to-one모델의 파라미터를 사용한 many-to-one, 적합한 튜닝을 적용한 many-to-one 모델들의 성능 비교를 진행하였으며 해당 실험 결과는 Table 4와 같다. 해당 실험 결과를 통해 세 가지 결론을 도출할 수 있었다.

첫째, 방언->표준어 번역이 표준어->방언 번역보다 높은 품질의 번역을 제공한다는 것이다. 이러한 결과가 도출된 이유에는 방언과 표준어의 단어 사용에 있다. 방언과 표준어의 관계에는 같은 의미를 가진 방언 단어와 이에 매핑되는 표준어 단어가 존재한다. 방언 문장에는 표준어와 지역 방언 고유의 단어가 혼용되어 출현한다. 지역 방언 고유의 단어가 있음에도 무조건적으로 방언 단어를 사용하지 않는다는 것이다. 이를 통해 방언 데이터 안에는 표준어가 함께 융합되어 있는 것을 확인할 수 있다. 방언 데이터와 달리 표준어 데이터 셋에는 방언 어휘가 표준어와 함께 사용되지 않는다. 표준어 데이터 셋에는 표준어만 존재하는 구조를 가지고 있다. 이러한 언어적 특성으로 인해 더 많은 어휘가 등장하는 방언 문장을 표준어로 번역하는 모델의 성능이 표준어를 방언으로 번역하는 모델에 비해 좋은 성능을 가질 수 있다는 것이다.

둘째, 다언어 번역과 적합한 튜닝이 이루어진 모델을 사용한다면 더 좋은 번역 성능을 가진다. One-to-one 번역에서 사용한 파라미터만으로도 한국어 다언어 방언 번역은 언어간의 상호작용을 통해 좋은 성능을 도출할 수 있었다. Many-to-one 모델에서는 파라미터의 증가로 인한 모델의 수용력을 증가시킬 수 있었으며 이에 따른 성능의 향상을 확인하였다.

셋째, 제안하는 many-to-one 모델은 one-to-one 모델보다 높은 성능을 보였다. 본 논문은 다언어 모델의 성능 향상에 대한 심층적 분석을 위해 언어적 유사성의 측면과 데이터 증강에 따른 언어모델의 향상에 대한 측면으로 접근하였다.

Table 4. Model Performance Comparison between One-to-One and Many-to-One

	One-to-One	Many-to-One	Tuned Many-to-one
Jeju	80.92	81.13	**82.12**
Kangwon	88.26	90.76	**92.05**
Gyeongsang	87.82	89.19	**91.50**
Jeonla	82.55	88.66	**89.24**
Chungcheong	87.67	91.60	**92.36**

Language Similarity 언어들간의 관계에서 형태적으로 유사한 언어들과 그렇지 않은 언어들이 존재한다. 대표적인 예로 영어와 프랑스가 있다. 영어 어휘의 1/3 정도는 프랑스어의 명사의 어원과 관련이 있으며 형태적으로도 유사한 명사들이 존재하며 유사성은 23.83%을 가진다. 영어와 독일어 프랑스어 3개의 언어 간의 유사성 또한 존재하는데 독일어와 영어의 유사성은 프랑스어보다는 낮은 10%정도의 유사성을 가지고 있다[18]. 이러한 유사성을 배경으로 인용한 논문[16]에서의 다언어 many-to-one 실험에서는 프랑스어-영어 간의 번역 성능이 독일어-프랑스어의 성능을 능가하는 지표를 보였다. 해당 지표들을 바탕으로 one-to-one모델보다 many-to-one 모델의 우월한 성능은 유사한 언어들을 더 많이 학습한 many-to-one 모델이 더 능숙한 한국어 언어 모델을 생성하여 성능의 향상으로 이어졌음을 추론할 수 있었다.

Large Language Model 데이터의 증강으로 인한 대용량 언어모델[19]의 학습과 파인튜닝 방법론은 사전학습되지 않은 언어모델보다 좋은 성능을 보인다. 표준어는 여러 지역의 방언들의 합성으로 이루어지며, 방언은 한국어 언어의 세부 언어라 할 수 있다. 따라서 한국어 방언 many-to-one 모델은 큰 틀에서 표준어->표준어 번역이라 할 수 있다. 본 논문은 다양한 방언의 학습

이 표준어의 관점에서 표준어 언어 모델의 데이터 증강으로 인한 대용량 언어모델로서의 발전으로 기여했다고 추론하였다. 이를 통해 many-to-one 인코더가 기존보다 효과적인 표현 추출로 인해 디코더의 언어 생성의 성능 향상으로 이어졌다.

다언어 모델은 기존의 one-to-one 모델들이 다량으로 필요하였던 문제를 해결하며 파라미터와 모델 개수에서의 효율성 측면에서의 긍정적 효과가 있다. 한국어 방언 다언어 번역 모델은 다량의 방언을 학습하며 효율성뿐만 아니라 성능 또한 개선되는 효과를 가진다. 한국어 다언어 번역 모델은 효율성과 성능 측면에서의 긍정적 효과를 통해 서비스 차원에서의 큰 이점을 one-to-one 모델에 비해 갖는다.

5. 결론

본 논문은 방언 연구의 중요성을 바탕으로 다양한 실험을 통해 한국어 방언 기계번역의 기반을 마련하였다. 같은 의미의 방언 문장과 표준어 문장이 공유하는 단어의 교집합이 크다는 사실을 바탕으로 카피 메커니즘 적용을 통해 기계번역 성능을 도모하였다. 카피 메커니즘을 적용한 모델은 훈련량이 증가함에 따라 학습의 방향이 통일성 측면에 편향된다. 카피 메커니즘은 방언과 표준어의 언어적 통일성과 다양성을 동시에 고려하여 적절하게 사용하여야 통일성 측면에 편향되지 않은 기계번역 모델의 성능을 기대할 수 있을 것이다. 방언 many-to-one 모델은 다양한 방언 언어의 학습으로 인한 대용량의 언어 모델로서의 발전으로 기여하였으며, 유사한 언어들의 학습이 기존 one-to-one 모델보다 좋은 성능 도출에 기여하였음을 검증 및 추론하였다. 개발된 방언 기계 번역기를 통해 표준어 기반 서비스만으로 소외되었던 방언 사용자들의 접근성을 높일 것으로 기대한다. 본 실험을 바탕으로 방언 기반의 one-to-many, many-to-many 와 같은 다언어 모델 개발과 방언과 한국어 이외 언어로 매핑되는 번역 모델을 향후 연구 과제로서 제안한다.

REFERENCES

[1] E. Benmamoun. (2000). *The feature structure of functional categories: A comparative study of Arabic dialects*. Oxford University Press.

[2] C. Park, K. Kim, Y. Yang, M. Kang & H. Lim. (2021). Neural spelling correction: translating incorrect sentences to correct sentences for multimedia. *Multimedia Tools and Applications*, *80(26)*, 34591-34608.

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

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

[5] J. Gu, Z. Lu, H. Li & V. O. K. Li. (2016). Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603. 06393*.

[6] S. R. Kudugunta, A. Bapna, I. Caswell, N. Arivazhagan, & O. Firat (2019). Investigating multilingual NMT representations at scale. *arXiv preprint arXiv:1909. 02197*.

[7] R. Aharoni, M. Johnson, & O. Firat (2019). Massively multilingual neural machine translation. *arXiv preprint arXiv:1903. 00089*.

[8] Korean Language Research Institute. (1988). Establishment of standard language regulation and Korean spelling. *Seoul : Ministry of Education, Republic of Korea*

[9] K. Park, Y. J. Choe & J. Ham (2019). Jejueo Datasets for Machine Translation and Speech Synthesis. *arXiv preprint arXiv:1911. 12071*.

[10] W. Salloum & N. Habash. (2012). Elissa: A dialectal to standard Arabic machine translation system. *Proceedings of COLING 2012: Demonstration Papers*, 385-392.

[11] K. P. Scannell. (2006). Machine translation for closely related language pairs. *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, 103-109.

[12] I. Guellil, F. Azouaou & M. Abbas. (2017). Neural vs statistical translation of algerian arabic dialect written with arabizi and arabic letter. *The 31st pacific asia conference on language, information and computation paclic*

[13] W. Farhan, B. Talafha, A. Abuammar, R. Jaikat, M. Al-Ayyoub, A. B. Tarakji & A. Toma (2020). Unsupervised dialectal neural machine translation. *Information Processing & Management*, *57(3)*, 102181.

[14] Y. Wan, B. Yang, D. F. Wong, L. S. Chao, H. Du & B. C. H. Ao (2020). Unsupervised Neural Dialect Translation with Commonality and Diversity

Modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34(5)*, 9130-9137.

[15] S. Lim, C. Park, J. Jo & Y. Yang (2021). Deep Learning based Korean Dialect Machine Translation Research. *Proceedings of the 33th Annual Conference on Human and Cognitive Language Technology*.

[16] M. Johnson et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339-351.

[17] T. Kudo & J. Richardson (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808. 06226*.

[18] S. Kim (2016). A Contrastive Analysis of the Noun Structure in German, English and French. *Yongbong Journal of Humanities*, *49*.

[19] 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*

임 상 범(Sangbeom Lim) [학생회원]

· 2021년 3월 ~ 현재 : 강남대학교 소프트웨어 응용 학부 소프트웨어 전공 (공학사)
· 관심분야 : Machine Translation, Deep Learning
· E-Mail : limsbeom@gmail.com

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, Data-centric AI, Grammar Error Correction, Deep Learning
· E-Mail : bcj1210@naver.com

양 영 욱(Yeongwook Yang) [정회원]

· 2009년 8월 : 한신대학교 소프트웨어학과 (이학사)
· 2011년 8월 : 고려대학교 컴퓨터교육과 (이학석사)
· 2018년 8월 : 고려대학교 컴퓨터학과 (공학박사)
· 2021년 3월 ~ 현재 : 한신대학교 컴퓨터공학부 조교수
· 관심분야 : Recommender system, Educational Data mining
· E-Mail : yeongwook.yang@gmail.com