지역 정보 토큰을 활용한 KoBART 한국어 방언-표준어 번역 성능 개선 연구

황재성°[1], 양희철[2]

충남대학교 인공지능학과°[1], 충남대학교 컴퓨터융합학부[2]

Korean dialect-standard language translation using special token in KoBART

Hwang Jae Sung°[1], Heecheol Yang[2]
Department of Artificial Intelligence, Chungnam National Univ. °[1]
Department of Computer Science Engineering, Chungnam National Univ.[2]

요 약

본 논문에서는 KoBART(Korean Bidirectional Auto-Regressive Transformer)의 fine-tuning을 통해 한국어 방언-표준어 번역 성능을 개선하는 연구를 수행하였다. 기존 ChatGPT API와 KoBART 모델은 방언-표준어 번역에 있어서 낮은 성능을 보인다. 이를 해결하기 위해 [경상도], [전라도] 등의 지역 정보 토큰을 문장 앞에 추가하여 KoBART 모델을 fine-tuning 하는 방식을 사용하였다. 지역별 학습 데이터의 양과 구성에 변화를 주어 학습을 진행하였고, 경상도 방언에 대해 평가를 진행하였다. BLEU(Bilingual Evaluation Understudy) 점수를 통해 성능을 측정한 결과, 지역 정보 토큰을 통해 방언-표준어 번역 성능을 향상시킬 수 있음을 확인하였다.

Ⅰ. 서 론

신경망 기계 번역은 딥러닝 모델을 훈련시켜 언어 번역 문제를 해결한다. 다른 언어 사이의 번역 연구는 많이 진행되어 왔으나, 방언-표준어 번역에 대한 연구는 그 수가 많지 않았다. Transformer[1] 기술을 활용한 방언-표준어 번역 관련 논문[2]에서는 Transformer copy attention을 이용한 방언 번역을 연구하였다. 또한, 입력 문장에 라벨을 추가해주는 방식으로 번역 성능을 개선하는 연구[3]가 진행되었다. 본 연구에서는 Transformer와 지역 정보 토큰을 활용한 방언-표준어 번역 연구를 진행하였다.

기존의 zero-shot 기반의 ChatGPT 방식 또는 사전 학습된 KoBART (Korean Bidirectional Auto-Regressive Transformer) [4]모델은 방언-표준어 번역 성능이 매우 낮음을 확인하였다. 본 논문에서는 이를 개선하기 위해 KoBART 기반의 baseline 모델을 활용하였다. 또한, 한국어 방언 지역 정보 토큰을 활용한 fine-tuning을 수행함으로써 방언-표준어 번역 성능이 개선 될 수 있음을 확인하였다.

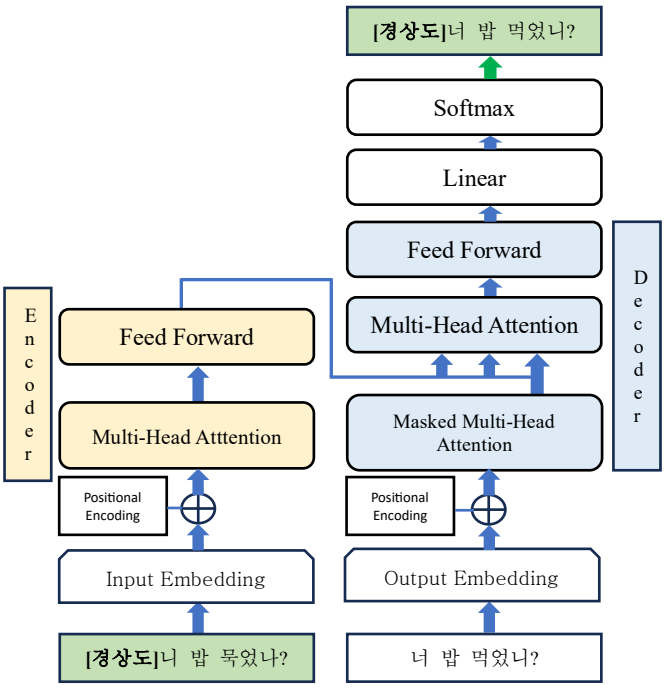
Ⅱ. 본론

1. 데이터셋

AI허브[5]에서 제공하는 한국어 방언 발화 5개 지역 경상도, 전라도, 강원도, 충청도, 제주도 데이터셋을 사용하였다. 텍스트로 저장되어 있는 json 파일을 csv 파일로 변환한 뒤, 중복된 문장, 공백 제거 등 후처리를 거쳤다. 학습 데이터셋은 각 지역별로 최대 50,000개를 사용, 평가 데이터셋은 경상도 10,000개의 데이터로 일관되게 진행하였다.

2. 모델 및 평가지표

BART는 bidirectional Transformer인 BERT와 auto-regressive Transformer인 GPT를 결합한 encoder-decoder 언어 모델이다. KoBART는 BART에서 사용된 text infilling 노이즈 함수를 사용하여 40GB 이상의 한국어 텍스트에 대해서 재학습한 한국어 encoder-decoder 언어 모델이다. 본 논문에서는 'HuggingFace'[6]에서 'gogamza/kobart-base-v2' 모델을 불러온 뒤, fine-tuning 하였다.



[그림 1] 지역 정보 토큰을 활용한 KoBART fine-tuning

[그림 1]에서 방언 데이터는 encoder의 입력으로, 표준어 데이터는 decoder의 입력으로 들어간다. 이중 입력 데이터에는 지역 정보 토큰을 문장 앞에 추가하여 데이터마다 방언의 지역 정보가 포함될 수 있도록 한다. Encoder로 들어간 방언 데이터는 encoder의 마지막 hidden layer에서 key vector, value vector로써 decoder의 multi-head attention 블럭의 입력으로 활용된다. Decoder의 각 단계에서는 <S>토큰을 시작으로 다음 토큰을 하나씩 예측한다. Encoder에서 온 key vector, value vector는 <S>토큰과 함께 시퀀스 생성을 시작하는 것을 학습하고, decoder의 입력으로 들어온 표준어 데이터는 'teacher forcing'에 사용된다. 즉, decoder의 입력으로 들어온 표준어 데이터는 'teacher forcing'의 target 데이터로 사용된다.

3. 실험 결과

Model			BLEU	
			지역 정보 토큰	
			w/o	w/
i)	base	GPT-3.5-Turbo	20.29	-
ii)		Pre-trained KoBART	44.15	-
iii)	fine tuned	G1+J1+GW1+C1+JJ1	90.51	90.59
iv)		G5+J5	88.05	88.10
v)		G2.5+J2.5	87.71	87.86
vi)		G1+J1	86.92	87.09
vii)		G0.5+J0.5	86.40	86.33
viii)		G5	87.82	87.93
ix)		G3	87.64	87.66
x)		G1	86.50	86.64

[표 1] BLEU-SCORE

본 연구에서는 BLEU(Bilingual Evaluation Understudy)[7] 점수를 모델의 번역 성능을 판단하는 평가지표로 사용한다. BLEU 점수는 번역된 각 문장에 대해 참조 표준어 문장과 비교하여 계산된다.

[표 1]은 각 평가에서의 BLEU 점수를 나타낸다. 경상도, 전라도, 강원도, 충청도, 제주도를 각각 'G', 'J', 'GW', 'C', 'JJ' 로 표기하고, 뒤의 숫자로 학습 데이터 수를 만 단위로 표기하였다 (ex. G1 = 경상도 방언 1만 문장). 또한 BLEU 점수 추출을 위한 평가 데이터는 모두 경상도 방언 1만 문장을 사용하였다. 단, ChatGPT API는 비용 문제 때문에 경상도 방언 300문장으로 진행하였다. [표 1]의 i)과 ii)를 통해 ChatGPT API와 KoBART의 방언-표준어 번역 성능이 낮음을 알 수 있다. 이를 개선하고자 KoBART를 해당 방언-표준어 번역에 맞게 지역 정보 토큰을 활용해 fine-tuning한 뒤 표준어로의 번역을 진행하고 BLEU 점수를 측정하였다. 방언-표준어 번역은 외국어 번역과 다르게 같은 언어에 대한 번역 작업이고, 방언-표준어 문장 사이에는 단어, 구절 차이만 일부 나타나므로 fine-tuning을 통해 높은 BLEU 점수를 얻는다.

문장의 시작 부분에 [경상도], [전라도]와 같은 지역 정보 토큰을 추가하여 훈련하면 BLEU 점수가 대부분의 경우에 향상되는 것을 확인 할 수 있다. KoBART의 encoder에서 input embedding과 positional encoding 연산 후 multi-head attention 블럭에 입력 방언에 대한 명시적인 지역 정보가 제공됨으로써 모델이 해당 방언의 지역과 관련된 특정 뉘앙스에 더 잘 집중하도록 훈련되기 때문이다.

또한, [표 1]의 iii), viii)을 통해 한 지역의 언어만 학습시키는 것보다 다양한 지역의 언어를 학습시키면 더 높은 BLEU 점수를 얻는 경향이 나타난다. 다만, [표 1]의 vii), x) 의 경우와 같이 데이터 양이 충분하지 않을 때는 한 언어만 학습시키는 것이 더 높은 성능을 얻을 수 있다.

Ⅲ. 결론

방언-표준어 번역을 위해 Transformer 모델을 fine-tuning을 할 때 지역 정보 토큰을 문장 앞에 삽입해서 입력데이터로 사용하면 더 높은 번역 성능을 얻을 수 있음을 확인하였다. 또한, 데이터가 충분하다면 다양한 지역의 문장들을 학습시키는 것이 한 지역의 언어만 훈련시키는 것보다 성능을 향상시킨다는 것도 알 수 있었다. 지역 정보 토큰 추가로 번역 성능 향상 및 특정 문화나 도메인에 특화된 번역 모델을 개발하여 해당 번역 작업에서 더 나은 성과를 얻을 수 있을 것으로 기대된다.

참 고 문 헌

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30 (NIPS 2017), California:USA, pp. 6000-6010, 2017. DOI: 10.48550/arXiv.1706.0376

[2] Lim S.B, Park C.J, & Yang Y.W (2022). Deep Learning based Korean Dialect Machine Translation Research Considering Linguistics Features and Service. 한국융합학회논문지, 13(2), 21-29, https://doi.org/10.15207/JKCS.2022.13.02.021

[3] Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. Multi-dialect Neural Machine Translation and Dialectometry, https://aclanthology.org/Y18-1001

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," arXiv:1910.13461, 2019.

[5] AI-Hub, available : https://aihub.or.kr/

[6] Huggingface, https://huggingface.co/gogamza/kobart-base-v2

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135