LLM 기반 번역을 활용한 영화 영-한 자막의 의역 연구

나경원, 박지은, 성유진, 이예원
성신여자대학교 지식서비스공과대학 정보시스템공학과/AI 융합학부

요 약

AI 는 단순 번역에서 속도가 빠르고 정확도면에서 우수하다는 평가를 받는다. 그러나, 문화적 배경이나 문맥 등을 제대로 파악하지 못해 의역에 한계를 보인다. 본 연구에서는 영어-한국어 영화 자막 데이터셋을 활용하여 다양한 기계 번역 모델을 학습시키고, 성능 차이를 확인하였다. 성능을 평가하기 위해 정량적 평가를 수행하였으며, 실험 결과, Long-KE-T5, KE-T5-base, KoBART, NLLB-200 순으로 성능이 좋음을 확인 할 수 있었다.

1. 서론

단순 번역 AI 는 속도와 정확성 면에서 우수하다고 평가받고 있다. 그러나 이러한 AI 는 문맥을 깊이 이해하지 못한다는 한계가 있다. 동일한 단어나 구절이 다른 문맥에서 다른 의미를 가질 수 있지만, AI 는 이를 적절히 처리하지 못한다. 또한, 의역에서는 단순한 언어적 변환뿐만 아니라 문화적 뉘앙스를 반영하는 것이 중요한데, 기존 번역 AI 는 이러한 부분에서도 한계를 보인다. 특정 문화나 사회적 맥락을 이해하고 반영하는 데 부족함이 있는 것이다.
본 연구는 이러한 한계점을 극복하기 위해 한국어 데이터를 영어로 자연스럽게 의역하는 여러 방법을 제안한다. 특히, 영화나 드라마 데이터셋을 활용하여 기계 번역 모델을 학습함으로써 성능을 향상시키고자 한다.

2. 데이터셋

2.1 한국어-영어 번역(병렬) 말뭉치

한국정보화진흥원(AI Hub)는 AI 기술 및 제품·서비스 개발에 필요한 데이터 및 컴퓨팅 자원을 제공하고 있다. 그 중 한국어-영어 번역(병렬) 말뭉치는 인공지능 학습용 문장으로, 상황별 신조어, 약어, 은어, 관용적 의미와 어투까지 효과적으로 전달할 수 있는

인공신경망기계번역(Neural Machine Translation; NMT)용 한-영어/스페인어, 러시아어 통·번역 음성 및 텍스트 pair 데이터셋이다. 50 만 개의 라벨링된 한-영 통번역 번역말뭉치 데이터가 있다.

2.2 PAWS-X 데이터셋

PAWS (Paraphrase Adversaries from Word Scrambling) 데이터셋은 구글이 언어 구조 문제를 해결하기 위해 공개한 코퍼스(Corpus)로, 구조, 문맥, 그리고 단어 순서 정보의 중요성을 모델링하는 문제를 다루기 위해 설계되었다. PAWS-X 는 PAWS 의 유형학적으로 구별되는 6 개 언어(프랑스어, 스페인어, 독일어, 중국어, 일본어, 한국어) 다국어 버전으로, 사람이 번역한 23,659 쌍과 기계번역에 의한 296,406 쌍의 단어가 포함되어 있다.

3. 방법론

3.1 KE-T5

KE-T5 모델은 최초의 한국어 데이터 중심 T5 계열 모델이자, 언어의 의미 인식 특성과 표현 특성을 모두 포함하고 있는 범용 언어모델이다. T5 모델은 모든 자연어처리(NLP) 작업을 텍스트 입력을 텍스트 출력으로 변환하는 문제로 통일함으로써, 다양한 작업에서 일관된 성능을 보여준다. 이러한 T5 모델을 한국어와 영어 코퍼스를 이용하여 사전 학습하였다.

3.2 KoBART

BART 는 입력 텍스트 일부에 노이즈를 추가하여 이를 다시 원문으로 복구하는 autoencoder 의 형태로 학습이 된다. KoBART 는 문장이나 텍스트의 일부를 비워두고 이를 다시 복구하도록 모델을 학습시키는 방법인 Text Infilling 노이즈 함수를 사용하여 40GB 이상의 한국어 텍스트에 대해서 학습한 한국어 encoder-decoder 모델이다.

3.3 NLLB-200

NLLB-200 는 Meta AI 에서 개발한 다국어 번역 모델로, 언어 문제로 디지털 정보와 문화에서 소외되고 있는 사람들의 디지털 격차를 해소하겠다는 목적에 부합하게 전 세계적으로 통용되는 공용어나 주류 언어가 아닌 언어에서 비교적 높은 정확도를 보인다. 이 모델은

다국어 프레임워크를 기반으로 하여 다양한 언어 간 관계를 학습하고, 언어 간 공유되는 표현을 학습함으로써 각 언어의 특성을 효과적으로 파악한다.

3.4 long-KE-T5

long-KE-T5 는 T5 모델에서 확장된, 한국어와 영어 비정형 데이터를 이용하여 학습시킨 사전 학습 모델이다. 입력 길이를 늘리거나 모델 크기를 늘리면 transformer 기반 딥러닝 모델의 성능이 향상될 수 있다는 연구에 따라 제시되었으며, 학습 간 encoder 의 최대 입력 길이는 4K 토큰, decoder 의 최대 입력 길이 1K 토큰으로 제한된다.

4. 실험 및 결과

4.1 실험 셋팅

본 논문에서는 모델 학습 및 평가를 위해 GeForce RTX 3090 GPU 1 대와 PyTorch 프레임워크를 사용하였다. KE-T5 의 모델 중 영어와 한국어 쌍 데이터에서 미세 조정된 모델인 'KETI-AIR/ke-t5-base'와 한국어와 영어 비정형 데이터를 이용하여 학습시킨 사전 학습 모델인 long-KE-T5 의 모델 'KETI-AIR-Downstream/long-ke-t5-base-translation-aihub-bidirection', 마찬가지로 NLLB-200 의 모델 중 'dhtocks/nllb_350M_en_ko_v16', KoBart 의 모델 중 'dylanmengzhou/kobart-trans-en-ko-v2'를 사용하였으며, 전체 데이터(50000) 중 학습과 평가를 위해 데이터를 8:2 비율로 나누어 40000 개의 개의 데이터를 학습에 사용하였고, 배치 사이즈와 학습률은 최적의 번역 결과를 찾기 위해 값을 조정해가며 실험을 진행하였다.

4.2 평가 지표

생성된 결과가 얼마나 적합한지를 판단하기 위해서 기계 번역 평가 지표 중 하나인 BERTScore 를 사용하였다. BERTScore 는 BERT 모델을 기반으로 문장 수준에서 번역 또는 생성된 문장과 참조 문장(정답 문장) 간의 각 토큰에 대한 유사성 점수를 계산하는 방법으로, 문장 간의 정확한 일치가 아닌 문맥을 반영한 임베딩을 사용한다. 본 연구에서 의역은 의미는

같으나 문장이 달라질 수 있다는 것에 초점을 두어 신경망 기계 번역 결과와 정답 문장 간의 유사도를 확인하고자 하였다.

4.3 실험 결과

[표 1] BERTScore 을 사용한 정량 평가 결과

	KE-T5- base	KoBART	NLLB-200	Long-KE-T5
Baseline	0.73	0.74	0.71	0.76
Fine-tuned	0.77	0.76	0.73	0.81

5. 결론

본 연구에서는 단순 기계 번역이 아닌 의역이 가능한 신경망 기계 번역 모델을 위해 영화 자막 한영 병렬 말뭉치 데이터셋과 언어 구조 문제 해결을 위한 데이터셋을 기반으로 모델을 학습시켜 영어 텍스트에 대한 한국어로의 번역 연구를 수행하였다. 모델 학습을 통해 정량적으로 성능을 향상시켰으며 실제 정답과 유사한 의역에 가까운 번역 결과를 확인할 수 있었다. 향후 연구에서는 본 논문에 사용한 모델이 아닌 다른 모델들을 추가로 하여 연구를 진행하여 성능을 개선할 예정이다. 또한, 번역의 오류를 사용자가 찾아 피드백할 수 있도록 기능을 구현할 예정이다.

참고문헌

[1] 임다영, 김미숙. (2023-12-20). 의역을 위한 LLM 기반 번역 연구 : 영화 영-한 자막 데이터를 중심으로. 한국정보과학회 학술발표논문집, 부산.

[2] Bhavnick Minhas, Anant Shankhdhar, Vivek Gupta, Divyanshu Aggarwal, and Shuo Zhang. 2022. XInfoTabS: Evaluating Multilingual Tabular Natural Language Inference. In Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER), pages 59–77, Dublin, Ireland. Association for Computational Linguistics.

[3] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of machine learning research 21.140 (2020): 1-67

[4] 이용성.(2023.07.27).[데스크 칼럼] 헛점 많은 AI 번역이 가르쳐준 것. 조선비즈.

https://biz.chosun.com/opinion/desk_column/2023/07/22/P6ACMFKV3VEA7OZSWOEM7WUZYQ/

[5] 천진우, 구자환, 김응모. Transformer 를 사용한 영한 기계 번역 : English-Korean Machine Translation using Transformer. 한국정보처리학회 2020 년도 추계학술발표대회, 2020 Nov. 05, 2020 년, pp.912 - 915 .

[6] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675