COSE474-2023F: Final Project Proposal
Can BERT Understand Other Morphological Features?

Jongbin Won

1. Introduction

최근 별개의 프로젝트로 한국어와 이미지에 기반한 multi-modal Abductive reasoning(귀추추론)에 관한 연구를 진행하던 중, 한국어로 학습되지 않은 BERT 모델이 가장 적절한 대답을 고르는 다중택일 문제에 대하여 fine-tuning 후 성능이 향상된 것을 확인한 바 있다. BERT 모델의 어휘 사전(Vocabulary)에 한국어 어휘가 하나도 포함되어 있지 않다는 점을 고려할 때, 이러한 성능 향상은 흥미롭다고 할 수 있다.

따라서 본 프로젝트에서는 영어로만 학습된 BERT모델이 영어 외의 언어에 대하여 어떠한 성능을 보이는지를 파악하고자 한다.

2. Problem definition & chanllenges

본 프로젝트에서는 다음과 같은 주제들에 대하여 논하고자 한다.

1) mBERT(multilingual BERT)모델과 비교할 때, 오직 영어로만 학습된 BERT모델이 다국어를 이해할 수 있는가?
2) 만약 존재한다면, 문자 체계 및 형태론적 유사성에 따른 성능 차이가 존재하는가?
3) 최종적으로, BERT의 영어 외 다국어 이해 능력을 향상시킬 수 있는가?

3. Related Works

최근 BERT와 mBERT의 다국어 능력을 비교하고, 그 대안으로 독어-영어간의 이해에 집중한 BiBERT 모델을 제안한 연구가 진행된 바 있다.(Xu et al., 2021) 그러나 독일어와 영어의 문자 체계나 형태소 구성 등이 많이 다름을 고려할 때, 이는 언어적인 특성들을 적절히 활용하지 못했다고 할 수 있다.

한편 mBERT를 활용해 통사적인 측면에서 보편문법(Universal Grammar)를 찾으려는 시도 역시 있었으나(Chi et al., 2020), 이 역시 범언어적인 차원에서의 통사적 구조를 파악하고자 하였으며 문자 체계 및 형태론적 유사성에 기반한 분석이 이뤄지지는 않았다.

4. Datasets

본 프로젝트는 NLLB(No Language Left Behind)의 Seed 데이터, MD(Multi-Domain)데이터, 그리고 FLoRes-200 데이터를 활용하고자 한다(TeamNLLB, 2022).

NLLB Seed 데이터와 MD데이터는 NLLB 프로젝트의 훈련용으로 활용된 데이터셋이다. Seed 데이터셋은 위키피디아로부터 추출한 6000여 개의 문장들에 대하여, 전문 번역가들이 39개의 언어로 직접 번역 및 검수해 구축되었으며, MD 데이터셋은 위키피디아 외의 뉴스, 연설 등에서 추출한 3000여 개의 문장들에 대하여, 전문 번역가들이 5개의 언어로 직접 번역/검수하여 구축되었다.

한편, FLoRes-200 데이터셋은 위의 두 데이터셋으로 학습된 모델을 평가하기 위한 벤치마크 데이터셋으로, 위키뉴스, 위키저니 등에서 추출한 약 3000여 개의 문장들에 대하여, 전문번역가들이 200여 개의 Low Resource 언어들로 직접 번역 및 검수하여 구축되었다.

5. State-of-the-art methods and baselines

BiBERT는 $En \rightarrow De$ 태스크는 RoBERTa 모델에 비해 BLEU(Papineni et al., 2002)점수가 2.12점 높은 29.65점을 기록했으며, $De \rightarrow En$의 경우 GOTTBERT에 비해 4.06점 높은 37.58점을 기록했다.

6. Schedule

대략적인 프로젝트 진행 일정은 다음과 같다.

11월 중순: BERT 및 MBERT 성능평가
11월 하순: 문자 체계 및 어족별 성능 비교평가 및 개선방안 모색
12월 초순: 개선 솔루션 코드 작성 및 개선 후 성능 평가
12월 중순: 최종 보고서 작성

References

Chi, E. A., Hewitt, J., and Manning, C. D. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5564–5577, Online, July 2020. Association for Computational Linguis-

tics. doi: 10.18653/v1/2020.acl-main.493. URL https://aclanthology.org/2020.acl-main.493.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

TeamNLLB. No language left behind: Scaling human-centered machine translation. 2022.

Xu, H., Durme, B. V., and Murray, K. W. Bert, mbert, or bibert? A study on contextualized embeddings for neural machine translation. *CoRR*, abs/2109.04588, 2021. URL https://arxiv.org/abs/2109.04588.