학습데이터

구글 BERT(Bidirectional Encoder Representation Transformer)(Devlin, 2019)

사전학습 언어모델(Pre-trained Language Model)을 시작으로 텍스트 분류(Topic classification), 감성 분석(Sentiment analysis), 그리고 질의 응답(Question answering) 등의 다양한 자연어 처리 모델을 State-of-the-art 수준의 성능으로 학습시키기 위해서는 사전학습 언어모델의 역할이 필수적이게 되었다. 사전학습 언어모델은 비지도 방식을 활용하여 대량의 말뭉치를 기반으로 학습되며, 기존 영문 및 다국어(Multilingual) 말뭉치 기반의 사전학습 언어 모델은 한국어 자연어 처리 역량에 있어 한계점을 가진다. 이를 극복하기 위해 최근 오픈소스 커뮤니티 및 학계에서는 한국어 말뭉치를 기반으로 학습한 다양한 사전학습 언어 모델을 공개하였으며, KoELECTRA(Park, 2020), KLUE-RoBERTa(Parketal, 2021) 등의 한국어 사전학습 언어모델이 대표적으로 널리 활용되고 있다.

이러한 State-of-the-art 한국어 사전학습 언어 모델들은 위키, 뉴스 등의 범용적인 도메인에 대한 자연어 처리에서는 높은 성능을 보이지만 OOD(Out-of-Distribution)에 해당되는 의료, 법률, 금융 등의 특수 도메인에서는 취약한 면모를 보인다. 이는 해당 모델들이 위키 문서, 웹 페이지, 일반 뉴스 등으로 대부분 구성된 범용적인 말뭉치를 기반으로 학습되어 특정 도메인에서의 활용에 최적화되어 있지 않다는 한계가 있기 때문이다. 자연어 처리 모델에 있어 도메인 특화 능력의 부재로 발생할 수 있는 가장 일반적인 현상으로는 1) OOV(Out-of-vocabulary) 발생으로 인한 언어 이해 부족 2) 특정 도메인에서만 사용되는 단어 및 유의어 관계에 대한 이해 부족 등이 존재하는데, 이러한 문제들은 사전학습 언어 모델을 기반으로 학습된 다양한 자연어 처리 태스크(Downstream task)의 활용 시 최종 성능에 매우 큰 영향을 준다.

이러한 한계점을 극복하기 위해 의료, 법률 등 다양한 분야에서 해당 도메인에 특화된 사전학 습 언어모델을 학습하고 활용하기 위한 연구를 시도하였으며 그 필요성 및 효율성을 입증하고 있다. 금융 또한 이러한 도메인 특화 모델의 필 요성이 존재하는 분야의 하나로, 1) 일반적으로