문장 난이도 측정을 위한 도메인 특화 언어 모델 연구

왕규현O, 오동규, 이수진

웅진씽크빅, 에듀테크연구소
wgh218@wjtb.net, lsjsj92@wjtb.net, odg0427@wjtb.net

Domain Specific Language Models to Measure Sentence Difficulty

Gue-Hyun WangO, Dong-Gyu Oh, Soo-Jin Lee
Woongjin Thinkbig, EduTech Labs

요 약

사전 학습된 언어 모델은 최근 다양한 도메인 및 응용태스크에 활용되고 있다. 하지만 언어 모델을 활용한 문장 난이도 측정 태스크에 대해서는 연구가 수행된 바 없다. 이에 본 논문에서는 교과서 데이터를 활용해 문장 난이도 데이터 셋을 구축하고, 일반 말뭉치로 훈련된 BERT 모델과 교과서 텍스트를 활용해 적응 학습한 BERT 모델을 문장 난이도 측정 태스크에 대해 미세 조정하여 성능을 비교했다.

주제어: 언어 모델, 문장 난이도 측정, 적응 사후 학습

1. 서론

문장 난이도 측정은 입력 문장의 난이도에 대해 난이도를 측정하는 것으로, 독자 수준에 맞는 적절한 수준의 텍스트 선정에 있어 필수적이다. 특히 교육 분야에서 아동에게 적절한 텍스트를 설정하여 독서 습관을 함양하고 문해력을 향상시키기 위해 주로 연구되고 있다. 해외의 경우, Lexile등 다수의 영어 텍스트 난이도 측정 지수가 널리 활용되고 있으며, 현재 사전 학습 언어 모델을 활용하여 텍스트 난이도 벤치마크 데이터 셋의 난이도를 측정하는 방안이 활발히 연구되고 있다[1]. 국내의 경우 다수의 연구진이 문장의 난이도를 측정하는 공식, 국어 텍스트 지표 등을 연구하고 있으나, 공식적인 데이터 셋의 부재로 언어 모델을 활용한 분석은 아직 널리 수행되지 않았다[2][3].

BERT[4]를 필두로 하는 사전 학습 언어 모델은 현재 다양한 도메인 및 응용태스크에 활발히 활용되고 있다. 사전 학습 언어 모델은 학습 언어의 분포를 학습하기 위해 위키피디아 등 다양한 주제의 대량의 말뭉치로 학습된다. 그러나 일반적인 말뭉치를 통해 학습된 언어 모델은 의료, 과학, 법률, 금융 등 전문적인 지식이 필요한 특정 도메인의 단어에 대해서는 잘 학습하지 못한다. 이에 특정 도메인의 대량의 문서로 훈련한 사전 학습 모델을 구성하는 연구가 활발히 진행되었다.

도메인 특화 사전 학습 언어 모델을 학습하는 것은 막대한 양의 데이터와 자원을 필요로 한다. 특히 언어 모델을 학습할 정도의 대량의 말뭉치를 구성하기 어려운 도메인의 경우, 도메인 특화 사전 학습 언어 모델 구성이 불가능하다. 이에, 기존 일반적인 말뭉치를 활용해 사전 학습된 언어 모델에 대해 비교적 적은 양의 도메인 말뭉치를 활용해 추가 학습을 진행하는 적응 사전 학습이 연구되었다.

아동 대상 문장 난이도 측정은 아동 수준의 문장 사이의 난이도를 측정해야 한다는 점에서, 일반적인 말뭉치로 훈련한 사전 학습 언어 모델이 적절히 측정하지 못할 가능성이 있다. 그러나, 기존 적응 사전 학습이 성능의 개선을 보인 전문 분야 말뭉치 대비 단어 학습의 어려움은 존재하지 않는 특성을 동시에 가지고 있다.

본 논문에서는 이러한 배경에서, 일반적인 말뭉치로 훈련 된 사전 학습 언어 모델과, 교과서 텍스트를 활용하여 적응 사전 학습을 진행한 언어 모델을 활용하여 아동 대상 국어 문장 난이도 측정 모델을 구성하고, 그 성능을 비교한다.

2. 관련 연구

적응 사전 학습은 사전 학습된 언어 모델에 도메인 말뭉치를 사전 학습과 같은 방식으로 훈련하는 것이다[5]. BERT의 경우, BERT 사전 학습의 목적 함수인 Masked Language Modeling(MLM)과 Next Sentence Prediction(NSP)를 모두 활용하거나, MLM 만을 활용하여 도메인 말뭉치를 추가 학습한다.

적응 사전 학습의 적용 도메인은 주로 의학, 과학, 중공업 등 전문적인 도메인의 말뭉치를 활용하며, 적용 결과 기존 언어 모델 활용 대비 성능의 향상을 보였다. BioBERT[6]는 생의학 분야의 말뭉치로 사전 학습하여 생의학 분야 응용 태스크에 대해 높은 성능을 보였다. SCIBERT[7]는 과학 분야의 태스크에 대해 최고 성능을 기록하였으며, HeavyRoBERTa[8]는 영어와 한글이 동시 등장하는 중공업 말뭉치 상에서 Perplexity와 zero-shot 유의어 추출 태스크의 성능 개선을 보였다.

전문적인 도메인 외에 문체에 따라서도 성능의 향상을 보일 수 있다. 감성 분류 태스크에 도메인이 같지 않더라도 같은 문체인 구어체 데이터를 활용하여 성능이 증

가한 바 있다[9].

3. 실험 설계

아동 대상 문장 난이도 측정 모델에 도메인 적응 사전 학습의 효과를 실험하기 위해, 교과서 텍스트를 처리하여 도메인 말뭉치를 구성했다. 활용 모델은 KLUE-BERT[10]를 이용했다. 언어 모델을 미세 조정하기 위한 문장 난이도 측정 공개 데이터 셋이 존재하지 않기 때문에, 교과서 데이터를 활용하여 문장 난이도 측정 데이터 셋을 구축했다.

3.1 교과서 말뭉치 구축

국어, 수학, 사회, 과학 교과서의 텍스트를 추출하여 언어 모델 훈련이 가능하도록 가공했다. 나이 별로 활용한 교과서 권수 및 포함 과목은 아래 표 1과 같다.

표 1. 나이별 교과서 정보

나이	포함 과목	교과서 권수
1	국어, 수학	10권
2	국어, 수학	10권
3	국어, 수학, 사회, 과학	15권
4	국어, 수학, 사회, 과학	15권
5	국어, 수학, 사회, 과학	16권
6	국어, 수학, 사회, 과학	16권

교과서 내에 존재하는 지시문, 문제등을 제외하고 본문의 텍스트를 추출했다. 구체적으로, 괄호가 포함된 문장, 공백 기준으로 분리한 토큰의 길이가 5개를 넘지 못하는 문장은 제외했다. 추출한 교과서 말뭉치의 크기는 다음 표 2와 같다.

표 2. 교과서 말뭉치 통계

문장 수	문장당 평균 길이	전체 토큰 수
79,778	34.83	694,617

3.2 적응 사전 학습

KLUE-BERT 모델을 활용하여 적응 사후 학습을 진행했다. KLUE-BERT 모델은 모두의 말뭉치, CC-100-kor, 나무위키, 뉴스, 국민청원 에서 추출한 63GB의 데이터로 학습되었다. 모델의 사전 크기는 32,000이며, 12개 layer로 구성되어 있다.

표 1의 가공된 교과서 말뭉치의 경우, 전체 토큰 수가 사전 학습 모델을 구성하기에는 충분치 않다. 이에 다양한 주제의 대용량 말뭉치로 훈련된 KLUE-BERT를 활용했다. KLUE-BERT 모델에 교과서 말뭉치를 사용하여 1 epoch의 적응 사전 학습을 수행하였다. 학습에는 Adam optimizer를 사용하였으며, 학습률은 0.00005, 배치사이즈는 64로 설정하였다. 적응 사전 학습 목적 함수로는 MLM과 NSP 중 MLM만을 활용했다.

3.3 문장 난이도 데이터 셋 구축

KLUE-BERT 모델과 교과서 말뭉치로 적응 사전 학습을 진행한 모델을 미세 조정 하기 위한 문장 난이도 데이터 셋 역시 구축하였다. 공신력 있는 구축을 위해, 교과서 텍스트 전체에서 단어장을 구축하고, 해당 단어가 포함된 문장을 학년에 따라 라벨링 했다.

단어장 구축은 총 6학년의 교과서 데이터에 따라 특정 학년에서 최초로 등장한 명사 및 동사에 해당 학년의 나이를 라벨링 했다. 라벨링한 단어의 수는 아래 표 3과 같다.

표 3. 단어장 나이 별 단어 수

나이	단어 수
1	1,919
2	1,199
3	2,359
4	1,542
5	1,455
6	478

구축한 단어장을 활용하여 각 나이 별 20,000개의 문장, 총 120,000개의 문장을 구성했다. 구성 시, 특정 과목에서 추출한 단어에 따른 주제 편중을 방지하기 위해 주제별로 문장이 균등히 분배되도록 했다. 구축한 데이터 셋은 108,000개의 훈련 데이터와 12,000개의 실험 데이터로 나누어 미세 조정 가능한 데이터 셋으로 구성했다.

4. 실험 결과

구축한 데이터 셋을 활용하여, 문장 난이도를 측정할 수 있도록 KLUE-BERT 모델과 교과서 데이터로 적응 사전 학습한 모델을 미세 조정했다. 총 6개 클래스에 대해 문장 분류하는 모델을 배치사이즈 64에 10 epoch로 훈련했다, 두 모델의 학습 곡선은 아래 그림 1, 그림 2와 같다.

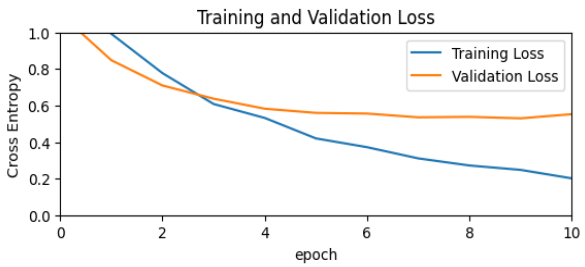

그림 1. KLUE-BERT 모델 학습 곡선

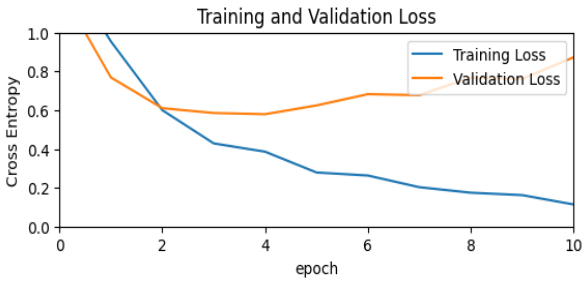

그림 2. 적응 사전 학습 모델 학습 곡선

그림 1과 그림 2의 학습 곡선을 보면, 교과서 말뭉치로 적응 사전 학습을 적용한 모델이 문장 난이도 데이터에 빠르게 과적합 되었다. 또한 validation loss 기준으로 KLUE-BERT 모델에 비해 낮은 성능을 기록함을 확인할 수 있다. 두 모델의 테스트 데이터에 대한 분류 정확도는 아래 표 4와 같다.

표 4. 테스트 데이터 분류 정확도

모델	정확도
KLUE-BERT	0.85
적응 사전 학습 KLUE_BERT	0.82

테스트 데이터 대상 분류 정확도 결과, 학습 곡선에 대한 결과와 같이 분류 성능이 오히려 떨어지는 결과를 보였다.

5. 결론

본 논문은 연구가 미흡한 언어 모델 활용 아동 대상 문장 난이도 측정 모델에 대해, 적응 사전 학습 방식을 적용하여 그 결과를 비교하였다. 모델 구성을 위해 적응 사전 학습 시 활용할 교과서 말뭉치를 구축하고, 교과서 데이터를 활용하여 6개 난이도로 구분되는 문장 난이도 측정 데이터 셋을 구축했다. 실험 결과, 교과서 데이터를 이용 적응 사전 학습을 실행한 모델이 훈련을 진행하지 않은 모델 대비 낮은 성능을 기록하였다. 이는 전문적인 도메인이 아닌 아동 말뭉치의 적응 사전 학습이 효용성이 없음을 의미한다. 또한, 난이도 측정 태스크의 특수한 분포를 일반적인 말뭉치로 훈련한 언어 모델이 미세조정을 통해 충분히 학습할 수 있음을 의미한다.

향후에는 아동 말뭉치의 특성에 대해 추가적인 연구를 수행하여, 난이도 측정 모델의 성능을 높일 수 있는 연구를 진행할 예정이다.

참고문헌

[1] Bruce W. Lee, Yoo Sung Jang. and Jason Lee, Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features, In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp.10669-10686, 2021.

[2] 조용구,이경남, 국어 텍스트 분석 프로그램(KReaD지수)의 개발, 독서연구, 56, pp. 225-246, 2020.

[3] 왕규현, 이수진, 정희원, 임누리, 사용자 이력을 활용한 앙상블 방식의 도서 난이도 연구, 독서연구, 65, pp.231-251, 2022.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, 2018.

[5] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A., Don't Stop Pretraining : Adapt Language Models to Domains and Tasks, arXiv preprint arXiv:2004.10964. 2020.

[6] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, Vol. 36, No. 4, pp. 1234-1240, 2020.

[7] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," arXiv preprint arXiv:1903.10676, 2019.

[8] 이정두, 나승훈, "HeavyRoBERTa: 중공업 특화 사전 학습 언어 모델", 제 33회 한글 및 한국어 정보처리 학술대회 논문집,pp 602-604, 2021.

[9] 이정훈, 김동화, 노영빈, 강필성, "구어체 적응 사전 학습을 통한 한국어 감정 분류 성능 향상", 대한산업공학회지, 47, 4, pp. 342-350, 2021.

[10] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh et al., "Klue: Korean language understanding evaluation," arXiv preprint arXiv:2105.09680, 2021.