KoBigBird를 활용한 수능 국어 문제풀이 모델

박남준 *김재광
성균관대학교
951011jun@naver.com *linux@skku.edu

Korean CSAT Problem Solving with KoBigBird

Park, Nam-Jun *Kim, Jaekwang
Sungkyunkwan University

요 약

최근 자연어 처리 분야에서 기계학습 독해 관련 연구가 활발하게 이루어지고 있다. 그러나 그 중에서 한국어 기계독해 학습을 통해 문제풀이에 적용한 사례를 찾아보기 힘들었다. 기존 연구에서도 수능 영어와 수능 수학 문제를 인공지능(AI) 모델을 활용하여 문제풀이에 적용했던 사례는 있었지만, 수능 국어에 이를 적용하였던 사례는 존재하지 않았다. 또한, 수능 영어와 수능 수학 문제를 AI 문제풀이를 통해 도출한 결괏값이 각각 12점, 16점으로 객관식이라는 수능의 특수성을 고려했을 때 기대에 못 미치는 결과를 나타냈다. 이에 본 논문은 한국어 기계독해 데이터셋을 트랜스포머(Transformer) 기반 모델에 학습하여 수능 국어 문제풀이에 적용하였다. 이를 위해 객관식으로 이루어진 수능 문항의 각각의 선택지들을 질문 형태로 변형하여 모델이 답을 도출해낼 수 있도록 데이터셋을 변형하였다. 또한 BERT(Bidirectional Encoder Representations from Transformer)가 가진 입력값 개수의 한계를 극복하기 위해 더 큰 입력값을 처리할 수 있는 트랜스포머 기반 모델 중에서 한국어 기계독해 학습에 적합한 KoBigBird를 사전학습모델로 설정하여 성능을 높였다.

1. 서론

오늘날 자연어처리 분야에서 기계독해 학습(Machine Reading Comprehension)과 관련된 연구가 활발히 이루어지고 있다. 그러나 이를 활용하여 문제풀이에 적용한 사례가 매우 적고, 특히나 한국어를 활용한 인공지능 문제풀이 사례는 찾아보기 힘들었다. 따라서 한국어 데이터셋을 활용한 문제풀이 모델의 필요성을 느꼈다.

본 논문은 기계독해에 특화된 자연어처리 모델에 오픈소스 기계독해 데이터셋을 학습시켜 수능 국어 문제풀이에 적용할 수 있는 모델을 제시한다. 그 중에서도 객관식 문항 중 올바른 정답 혹은 올바르지 않은 정답을 찾는 유형의 문제에 각 선택지를 학습모델이 결과값을 도출해낼 수 있도록 데이터셋을 변형하여 결과를 나타내도록 하였다. 논문의 구성은 다음과 같다. 먼저, 자연어처리 분야의 여러 기계독해 학습 모델을 제시한다. 둘째, 이러한 기계독해 학습 모델에 학습할 데이터셋을 제시한다. 셋째, 위에서 제시한 데이터셋을 학습시켜 이를 수능 문제풀이에 적용하여 도출해낸 연구 결과를 제시한다. 마지막으로, 연구를 통해 얻은 결론을 도출한다.

2. 관련 연구

2.1. BERT

그림 1이 보이는 바와 같이 BERT는 사전학습모델로 사용하도록 설계되고 훈련된 모델이다. BERT는 Bidirectional Encoder Representations from Transformers의 약자로 여러개의 Transformer 인코더 레이어를 갖고 있으나 Transformer와 달리 디코더는 사용되지 않는다. BERT는 문맥을 반영한 임베딩을 사용하기 때문에 기존 임베딩보다 문맥을 파악하는 능력이 뛰어나다[1].

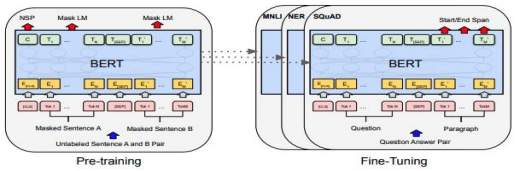

그림 1. BERT의 사전학습 및 Fine-Tuning 절차.

2.2. BigBird and KoBigBird

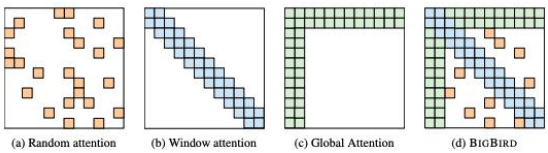

그림 2. BigBird 모델의 구조.

그림 2는 BigBird 모델의 구조를 보인다. 대부분의 Transformer 기반 모델은 계산량이 입력값의 길이의 제곱에 비례하기 때문에 입력 시퀀스 길이를 제한한다. 이러한 한계 때문에, 우리가 긴 문맥 시퀀스를 사용해야 할 때, 우리는 입력값을 여러개로 나눠서 입력해야 한다. 하지만 이 방법은 더 긴 시퀀스 입력을 효과적으로 처리할 수 없는데, BigBird는 이를 Sparse Attention 기법을 통해 해결한다. 일부 작업은 근처의 토큰뿐만 아니라 멀리 떨어진 토큰과의 관계도 포착 해야 하는데, 이를

Long Sequence Dependency라고 한다. 이러한 Long Sequence Dependency를 위해 BERT는 Full Attention을 통해 해결해야 하는데, 이는 긴 문장을 입력으로 사용할 때 Time Complexity와 Spatial Complexity의 비효율성을 초래한다. 이 문제를 해결하기 위해 Sparse Attention은 토큰을 세 가지 범주로 나누고, 그에 대한 Attention을 계산한다. 대상 문장의 시작과 끝을 담당하는 Global Attention, 근처의 토큰을 담당하는 Sliding, 그리고 무작위의 토큰을 사용하는 Random Attention을 사용 하여 Full Attention과 유사하게 Sparse Attention을 수행한다[2]. 본 연구에서는 BigBird와 구조가 동일하지만, 한국어 데이터셋 학습에 특화된 모델인 KoBigBird를 사용하였다[3].

3. 실험 및 결과

3.1. 데이터셋

본 연구에서는 기계독해 학습에 특화된 오픈소스 데이터셋을 사용하였다. 학습에 사용한 데이터셋은 두가지인데, 그 중 첫번째는 KLUE(Korean Language Understaning Evaluation) 데이터셋을 사용했다[4]. KLUE 데이터셋은 뉴스 자료와 위키피디아 자료들로 이루어진 한국어 데이터셋이다. 다음으로 사용한 데이터셋은 AI Hub의 도서자료 기계독해 데이터셋을 사용했다[5].

3.2. BERT 기반의 모델

연구 초기 단계에는 BERT를 사전학습모델로 선정하여 그림 3이 보이는 바와 같은 파라미터(Parameter)로 학습을 진행하였다.

```
torch.manual_seed(42)
config = BertConfig(
    vocab_size=indexer.vocab_size,
    max_position_embeddings=1024,
    hidden_size=256,
    num_hidden_layers=4,
    num_attention_heads=4,
    intermediate_size=1024
)
model = BertForQuestionAnswering(config)
# model.cuda()
optimizer = torch.optim.AdamW(model.parameters(), lr=2e-4)
```

그림 3. 학습에 사용한 파라미터(Parameter)

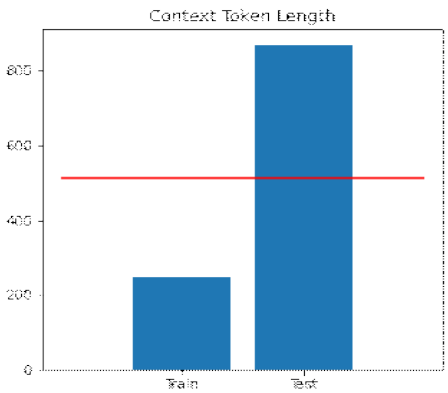

그림 4. 데이터셋의 평균 Context 토큰 길이.

그러나 그림 4가 보이는 바와 같이 수능 비문학 지문으로 이루어진 테스트 데이터셋의 Context의 평균 토큰 길이가 BERT의 최대 토큰 길이인 512를 넘어가는 문제점이 발생하였다.

3.3. KoBigBird 기반의 모델

따라서 BERT보다 더 많은 토큰을 입력값으로 받을 수 있고 한국어 기계학습 독해에 특화된 KoBigBird를 활용하여 모델의 성능을 높이고자 하였다.

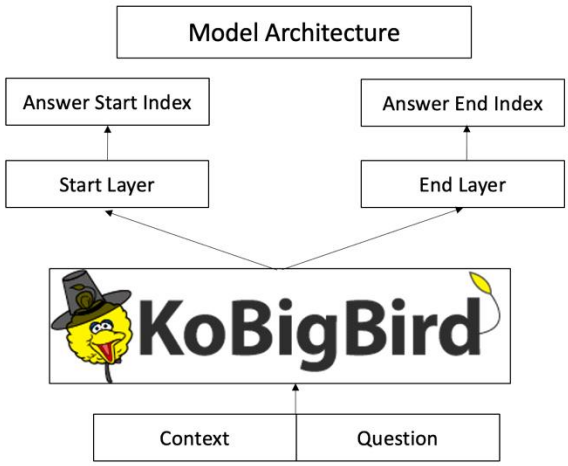

그림 5. KoBigBird 모델의 구조.

KoBigBird 모델 내부에 3개의 Attention Block이 존재하고, Attention Block은 Attention Layer, Intermediate Layer, 그리고 Output Layer로 구성된다. Intermediate Layer와 Output Layer 사이에는 모델이 최적의 Parameter에 수렴하도록 돕는 Layer Normalization Layer가 존재한다. 그리고 Output Layer 내부에는 모델을 훈련시키는 동안 과적합을 방지하기 위한 Dropout Layer가 존재한다. KoBigBird는 BERT보다 모델의 복잡성이 더 크기 때문에 이로 인해 KLUE 데이터셋으로만 학습을 진행할 경우 과적합이 발생했다.

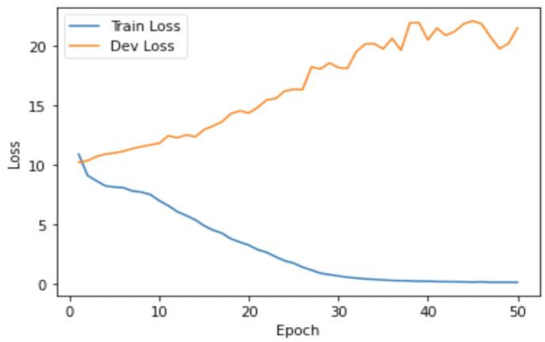

그림 6. KLUE 데이터셋만으로 학습을 진행했을 때의 학습 결과.

따라서 KLUE 데이터셋에 AI Hub 도서자료 기계독해 데이터셋을 추가하여 학습을 진행하였더니 과적합이 해소되는 모습을 다음의 그림

과 같이 확인할 수 있었다.

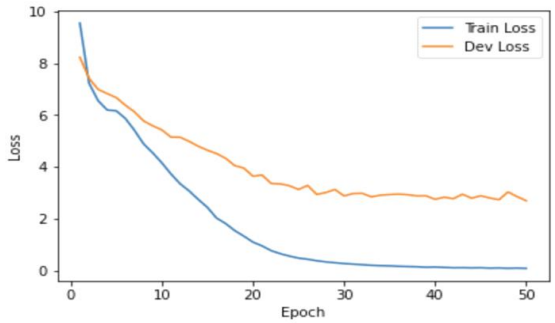

그림 7. KLUE + AI Hub 데이터셋으로 학습을 진행했을 때의 학습 결과.

이렇게 학습 완료한 모델에 수능 비문학 지문과 평문으로 이루어진 객관식 선택지를 질문형태로 변형하여 모델이 각 질문에 답변할 수 있도록 다음과 같이 구성하였다.

그림 8. 수능 비문학 지문과 객관식 선택지로 이루어진 테스트셋 예시.

테스트셋의 후처리 과정에서 GUID를 기반으로 출력 목록에서 정답을 찾아냈다. GUID의 마지막 8자리 숫자에는 문제의 정보가 포함되어 있으며, 각 숫자는 질문 번호, 질문 유형, 답변 번호 및 해당 선택 번호를 나타낸다. 예를 들어, 2번 문제가 올바르지 않은 선택지를 고르는 문제인데, 정답이 선택지 3번이고, 예측한 선택지 번호가 4번이라면 GUID의 마지막 8자리는 0201a304가 된다.

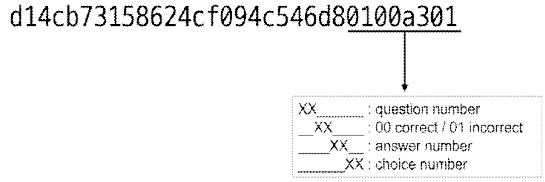

그림 9. GUID 구성 예시.

이렇게 구성한 테스트셋을 학습 모델에 적용시키면 다음과 같은 예측 결과값을 얻을 수 있었다.

	Id	Predicted
0	d14cb73158624cf094c546d80100a301	19세기 중반
1	d14cb73158624cf094c546d80100a302	화학자 분젠은 불꽃 반응에서 나타나는 물질 고유의 불꽃색에 대한 연구를 진행하고 있...
2	d14cb73158624cf094c546d80100a303	화학자 분젠은 불꽃 반응에서 나타나는 물질 고유의 불꽃색에 대한 연구를 진행하고 있...
3	d14cb73158624cf094c546d80100a304	화학자 분젠
4	d14cb73158624cf094c546d80100a305	창안

그림 10. 테스트셋 예측 결과 예시.

이렇게 얻은 예측 결과값을 ROUGE-L과 RDASS 두 가지의 평가지표를 활용하여 모델의 정확도를 측정했다[6][8]. ROUGE-L의 Precision과 Recall의 조화평균인 F-Measure를 활용하여 실제 정답 데이터셋과 모델이 생성한 예측 정답 데이터를 비교하여 정답으로 판단되면 1, 오답으로 판단되면 0으로 설정하였을 때, 총 20문항 중 2개의 문항을 정답이라고 판단하였다. RDASS도 위와 유사한 방식으로 본문, 실제 정답 데이터셋, 모델이 생성한 예측 정답 데이터셋 세 가지의 코사인 유사도를 활용하여 예측 모델이 생성한 정답이 맞으면 1, 틀리면 0으로 설정했을 때, 20문항 중 6개의 문항을 정답이라고 판단하였다. ROUGE의 경우 어근에 붙은 접사가 단어의 역할을 결정하는 한글의 특성을 반영하지 못하고, 빈번한 단어의 변형을 고려하지 못하기 때문에 RDASS로 모델을 평가했을 때 더 높은 점수를 나타냈다.

4. 결론

연구논문을 진행하면서, 많은 양의 기계독해 데이터셋을 활용해 모델 학습 과정에서 과적합을 해결하고, 객관식 질문의 답을 해결하는데 있어서 수능 비문학 지문으로 이루어진 테스트셋을 구성하여 정답을 얻기 위하여 확률적으로 접근했던 점이 긍정적이라고 생각한다. 하지만, 수능 문항에는 다양한 형태의 문제들이 존재하는데, 그 중 특수한 케이스(올바른 혹은 올바르지 않은 답 선택)에만 적용할 수 있기 때문 에 범용성이 떨어지고, 예상했던 정답률보다 정답률이 저조한 부분을 개선하기 위한 추가적인 연구가 필요하다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT명품인재양성 사업의 연구결과로 수행되었음 (IITP-2022-0-01821).

참고문헌

[1] J. Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 Oct. 2018.

[2] M. Zaheer et al., “Big bird: Transformers for longer sequences,” Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 17283-17297, Dec. 2020.

[3] Monologg. “KoBigBird: Pretrained BigBird Model for Korean.” GitHub, 8 Nov 2021. https://github.com/monologg/KoBigBird. accessed 21 May

2022.
[4] KLUE. “Datasets at hugging face”. Hugging Face, 2021., https://huggingface.co/datasets/klue. Accessed 21 May 2022.
[5] AI Hub. “도서자료 기계독해.” AI Hub, 18 Jun 2021. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=92. Accessed 21 May 2022.
[6] K. Ganesan. "Rouge 2.0: Updated and improved measures for evaluation of summarization tasks." arXiv preprint arXiv:1803.01937, Mar. 2018.
[7] Kim, Hyun-joong, Sungzoon Cho, and Pilsung Kang. "KR-WordRank: An unsupervised Korean word extraction method based on WordRank." Journal of Korean Institute of Industrial Engineers 40.1, pp. 18-33, Feb. 2014.
[8] D. Lee, et al. "Reference and document aware semantic evaluation methods for Korean language summarization." arXiv preprint arXiv:2005.03510, Apr. 2020.
[9] A. Vaswani et al. "Attention is all you need." Advances In Neural Information Processing Systems, 30, Dec. 2017.