KorBERT 및 Transformer 모델을 동시 활용한 한국어 문장 요약 시스템

이민수, 유채림, 김선용

동서대학교

razzcat_min@naver.com, jchaezh@gmail.com, sykim@dongseo.ac.kr

Korean Sentence Summary System using KorBERT and Transformer Models

Minsu Lee, Chae Rim Yu, Sunyong Kim

Dongseo University

요 약

오늘날 독서량 감소 및 코로나 19로 인한 원격 수업 진행 등으로 인해 발생한 낮은 문해력 수준 문제는 계속 제기되고 있다. 이와 같은 문해력 저하 문제를 해결하기 위해 문해 능력이 부족한 사람이 문장을 이해할 수 있게 KorBERT와 Transformer Model을 활용한 한국어 문장 요약 시스템을 제안 및 구현한다.

Ⅰ. 서 론

1994년부터 1998년까지 OECD 국가를 포함한 20여 국가에서 실시한 국제 성인 문해조사 도구를 활용하여 대한민국 성인의 문해 수준을 조사한 결과 영역별 문해수준은 기존 조사에 참여한 국가들과 마찬가지로 낮게 나왔다 [1]. 또한 2020년 국가평생교육진흥원에서 조사한 「성인문해능력조사」에서 18세 성인 중 일상 생활에 필요한 문해 능력이 부족한 수준 1에 해당하는 자가 2백만 명으로 추정하고 있고, 2021년 국가 수준 학업성취도 평가 중등 3학년 국어 교과목에서도 기초학력 미달 비율이 6.0%가 나옴으로써 문해력 저하 문제가 과거부터 현재까지 꾸준히 제기되고 있다 [2]. 이와 같은 문해력 저하 문제들은 국내에만 국한된 것이 아닌 미국, 영국 등 해외 국가에서도 제기되고 있다. 미국에서는 문해력 증진 전략으로 '영어학습자를 위한 교육정책'과 '이중 언어 교육정책'을 마련하였다 [3]. 문해력 저하 문제를 해결하기 위해 본 논문은 한국전자통신연구원 한국어 BERT 언어모델인 KorBERT와 Transformer 모델을 활용하여 Encoder와 Decoder를 구축하고 Seq2Seq 모델의 구조를 활용하여 문장을 요약하는 시스템을 제안한다.

Ⅱ. 본론

본 논문에서는 한국어 문장 요약 시스템을 구현하기 위해 Seq2Seq 모델의 구조를 활용한다. Seq2Seq 모델은 Encoder와 Decoder로 구성되어 있는데 Encoder에는 KorBERT 모델을, Decoder에는 Transformer 모델을 활용하였다. 이와 같이 두 개의 모델을 동시 활용하면 단일 모델을 활용할 때보다 학습 시간도 줄일 수 있고 타 시스템 대비 수준 높은 언어 이해 능력을 기대할 수 있다. 그리고 Encoder에는 의미론적 유사성이나 문법 구조 등 언어적 특성에 더욱 초점을 맞추고, Decoder에는 Batch Size, Learning Rate를 조정하거나 Attention Mechanism 활용, 정규화 기법 사용을 통해 생성 능력을 강화시키는 등 Encoder와 Decoder를 별도로 최적화 및 튜닝할 수 있는 유연함을 가진다. 문장 요약을 위해서는 텍스트 전처리가 필수적인데 본 논문에서는 KorBERT에서 제공하는 Tokenizer를 활용하여 요약을 진행할 문장을 토큰화하여 개별 단어 및 토큰으로 분리하였다. 전처리된 텍스트는 KorBERT를 활용한 Encoder로 전달하여 각 토큰에 대한 임베딩 벡터로 생성된다. Encoder에서 변환된 벡터는 Transformer 모델을 활용한 Decoder로 전달되어 인코딩된 벡터를 바탕으로 디코더가 생성하고자 하는 시퀀스의 길이인 타겟 시퀀스(Target Sequence)의 길이와 배치 크기, Vocab 크기를 출력하고 이 모델의 예측을 통해 요약을 위해 어떤 단어를 선택할지 결정한다. 그림 2와 같이 요약 문장을 생성한 결과 토큰 인덱스가 반환되는 것을 확인할 수 있었다.

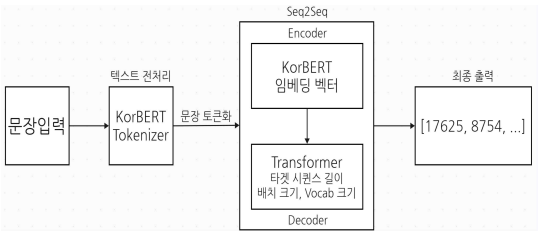

그림1. 제안하는 한국어 문장 요약 시스템

```
요약된 문장: [3012, 13918, 2124, 10483, 3012, 13918, 2124, 10483, 3012, 1391
8, 2124, 10483, 3012, 13918, 2124, 10483, 3012, 13918, 2124, 10483, 3012, 13
918, 2124, 10483, 3012, 13918, 2124, 10483, 3012, 13918, 2124, 10483, 3012,
13918, 2124, 10483, 3012, 13918, 2124, 10483, 3012, 13918, 2124, 10483, 3012
, 13918, 2124, 10483, 3012, 13918]
```

그림2. Encoder, Decoder 출력 및 문장 요약의 토큰 인덱스

Ⅲ. 결론

본 논문에서는 KorBERT와 Transformer 모델을 활용하여 필요한 문장을 요약하는 한국어 문장 요약 시스템을 설계 및 핵심 기능 일부를 구현하였다. 향후 연구로 생성된 요약의 토큰 인덱스와 실제 단어를 매핑하는 사전을 구축 및 적용할 계획이며, 문서 스캔 기능 및 OCR 처리 기능을 추가하여 요약이 필요한 영역을 선택하면 자동으로 문장을 요약할 수 있도록 보완할 계획이다.

ACKNOWLEDGMENT

"본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 지원을 받아 수행되었음"(2019-0-01817)

참 고 문 헌

[1] 이희수, 박현정, 이세정, "한국 성인의 문해실태와 OECD 국제비교 연구", *비교교육연구*, vol. 13, no. 2, pp. 193-219, 2003.

[2] 국가평생교육진흥원,「성인문해능력조사」, 2020.

[3] A. Boyle, et al., "Dual Language Education Programs: Current state policies and practices", *Office of English Language Acquisition*, 2015.