한국어 논문 요약을 위한 KoBART와 KoBERT 모델 비교*

전제성○, 이수안

세명대학교 컴퓨터학부
jasonjun1121@gmail.com, suanlee@semyung.ac.kr

Comparison of KoBART and KoBERT models for Korean paper summarization

Jaesung Jun○, Suan Lee
Semyung University, School of Computer Science

요 약

통신 기술의 발전으로 일반인들도 다양한 자료들을 인터넷에서 손쉽게 찾아볼 수 있는 시대가 도래하였다. 개인이 접근할 수 있는 정보량이 기하급수적으로 많아 짐에 따라, 이를 효율적으로 요약, 정리하여 보여주는 서비스들의 필요성이 높아지기 시작했다. 본 논문에서는, 자연어 처리 모델인 BART를 40GB 이상의 한국어 텍스트로 미리 학습된 한국어 언어 모델 KoBART를 사용한 한국어 논문 요약 모델을 제안하고, KoBART와 KoBERT 모델의 한국어 논문 요약 성능을 비교한다.

주제어: 한국어 논문 요약, 논문 요약, Paper Summarization, BART, KoBART, BERT, KoBERT,

1. 서론

인터넷 기술의 발전으로 인터넷을 통하여 많은 사람들이 다양한 정보를 찾아볼 수 있게 되었고 그에 따라 Google, Yahoo, Naver 등의 검색 서비스를 제공하는 기업들이 생겨났다. 이러한 추세는 과학/연구 분야에도 이어져, Arxiv, Google Scholar 등의 출판 전/후 논문을 업로드하고 공개, 검색할 수 있는 서비스를 제공하는 기업, 단체 들이 생겨났다. 이러한 정보의 범람으로도 일컬어지는 현상이 일어남에 따라 사용자에게 필요한 데이터를 추천하고 큰 데이터를 축소, 요약하는 형태의 서비스의 필요성 또한 높아졌다. 본 논문에서는, 자연어 처리 모델인 BART를 40개 이상의 한국어 텍스트에 대해서 학습시킨 한국어 언어 모델 KoBART를 사용한 한국어 논문 요약 모델을 제안하고, KoBART 모델을 기존의 KoBERT를 사용한 한국어 논문 요약 모델과 성능을 비교하여 각 모델의 장단점을 분석하였다.

2. 관련 연구

문장 요약 기술은 자연어 처리 딥러닝 기술의 발전과 함께 성장해왔다. BART[1](Bidirectional Auto-Regressive Transformers) 구조가 연구되기 이전에는 주로 RNN(Recurrent Neural Network)과 LSTM(Long-Short Term Memory)를 활용한 Sequences-to-Sequences 구조를 활용하여 문장 요약 시스템을 구성해왔다. 2017년 Google Brain 팀에서 Transformer 구조가 발표되었을 때, 이를 활용한 GPT[2](Generative Pre-trained Transformer)와, BERT[3](Bidirectional Encoder Representations from Transformers) 가 개발되었다. GPT, BERT 모델은 각각의 특성을 가지고 있기에 특화된 작업이 다르다. GPT 모델의 경우, 문장 생성 작업에 강한 한편, BERT 모델은 문장 생성 작업에는 약하지만, 문장 전체의 의미를 파악하여 문장을 분류하는 작업에 강하다. 반면 본 논문에서 사용한 BART 모델은, 각각 모델의 장점만을 가져와, 문장 생성, 분류 두 작업 모두에 강하다. 따라서 문장 생성 및 단어 분류 작업에서 모두 성능이 보장되어야 하는 논문 요약 작업을 수행할 자연어 처리 모델로서 BART 모델을 사용하는 것이 적합하다고 판단하였다.

3. 한국어 논문 요약 모델

3.1 BART

BART(Bidirectional Auto-Regressive Transformer)는 2019년 Facebook AI에서 발표되었다. BART는 기존에 문장 요약, 문장 생성에서 사용되던 모델인 GPT와 BERT의 단

* 본 과제(결과물)는 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다.(2021RIS-001)

점을 보완하고자 고안되었다. GPT 모델의 단방향적 특성으로 인해 문장의 문맥을 제대로 파악하지 못하지만, 문장 생성에는 강하다는 점, BERT 모델의 양방향적 특성으로 인해 문장의 문맥을 잘 파악할 수 있지만, Regressive하지 않기 때문에 문장 생성에 약하다는 각 모델들의 장점을 살리고, 단점을 없애는 방향으로 개발되었다. BART 모델의 구조는 그림 1과 같으며, 인코더만 가진 BERT 모델과 디코더만 가진 GPT 모델을 결합한 인코더-디코더 모델이다. BART 모델은 데이터의 일부에 노이즈를 추가하여 노이즈를 원본 데이터로 바꾸는 방식으로 학습을 진행한다. 본 논문에서는 BART의 이러한 특성을 이용하여 논문 요약 작업을 수행하였다.

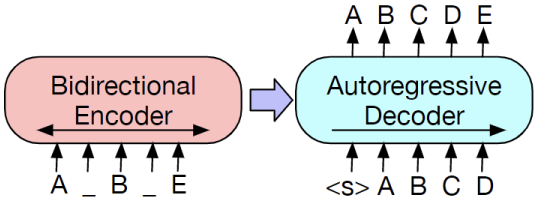

그림 1 BART의 구조

3.2 KoBART

KoBART는 앞서 소개한 BART의 한국어 버전 모델이다. 2020년 SKT AI 팀이 모델 구현 및 학습하였으며, 한국어 위키백과를 비롯한 뉴스, 책, 모두의 말뭉치, 청와대 국민청원 등의 다양한 데이터가 모델 학습에 사용되었다. 본 논문에서는 SKT AI 팀에서 미리 학습한 KoBART 모델을 미세 조정하여 한국어 논문 요약 모델을 구현한다.

3.3 모델 구성 및 학습 과정

본 논문의 학습 과정은 그림 2와 같이 총 3가지 단계로 구성되어 있다. (1)한국어 위키백과, 뉴스, 책, 청와대 국민청원 등 데이터로 미1리 학습되어 있는 KoBART 모델에 (2)Dacon에서 진행한 한국어 문서 생성 요약 AI 경진대회에서 사용된 데이터셋을 이용하여 학습시킨 가중치를 (3)AIHub의 논문자료 요약 데이터를 활용하여 배치 사이즈는 4, 에포크는 3로 미세 조정하였다.

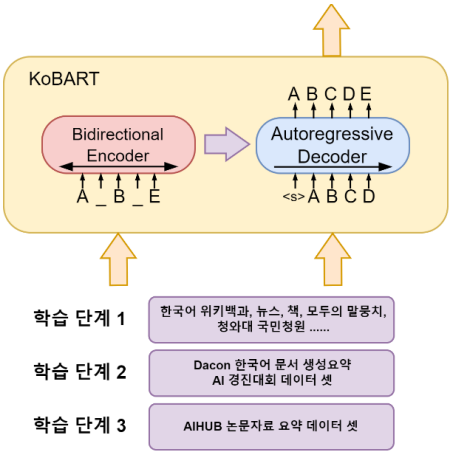

그림 2 KoBART의 학습 과정

4. 데이터 셋

본 논문에서는 KoBART를 논문 요약 작업에 활용하기 위해 기존의 미리 학습된 모델을 미세 조정하였다. 미세 조정에 사용한 데이터 셋은 AIHub의 논문자료 요약 데이터셋이다. AIHub의 논문자료 요약 데이터 셋은 학술논문 18만건과 특허 명세서 18만 건을 전체/섹션 요약 대상별로 요약한 요약문을 포함한다. 각 데이터는 결측치 제거 후, 원문/요약문 쌍으로 구성되도록 전처리 하였다. 전처리 완료 후 학습 데이터는 144,264개의 문장-요약 쌍으로 이루어져 있으며, 검증 데이터는 18,025쌍이다.

5. 실험 결과

평가 지표는 자연어 처리 분야에서 자연어 생성 모델의 성능을 측정하기 위해 사용하는 ROUGE를 사용하였다. 표 1에서 KoBART 모델은 KoBERT 모델에 비해 ROUGE-{1, 2, L}의 F-Score가 낮았다. 하지만, ROUGE-{2, L}의 Recall 값은 높은 경향을 띄었다.

ROUGE-{2, L}의 Recall값이 높은 이유는, KoBART 모델이 생성하는 요약이 데이터셋의 요약보다 문장 길이가 긴 경우가 많기 때문이다. 원본 요약의 문장 평균 길이는 약 160 글자였던 반면, KoBART 모델이 생성한 요약의 평균 문장 길이는 약 220 글자이다. 생성된 요약문이 길면 데이터셋의 요약문안에 포함된 단어를 생성된 요약문도 포함할 가능성이 크기 때문에, 일반적으로 Recall이 높아지는 경향이 있다. 또한 ROUGE-1에서는 높은 점수를 얻지 못하였지만, ROUGE-{2, L}에서 높은 점수인 것은, 데이터셋의 요약과 모델이 생성한 요약이 단어를 각각 하나씩 비교하였을 때에는 일치되는 부분이 적지만, 2개 단어 이상의 단어쌍으로 비교하였을 때는 KoBERT에 비해 일치하는 부분이 많다고 해석된다.

성능지표	KoBART	KoBERT
ROUGE-1 F-Score	37.16	**48.9**
ROUGE-2 F-Score	27.97	**29.8**
ROUGE-L F-Score	36.39	**37.05**
ROUGE-1 Recall	45.12	**46.07**
ROUGE-2 Recall	**35.19**	28.13
ROUGE-L Recall	**44.21**	36.31

표 1 실험 결과

6. 결론

본 논문에서는 BART 모델을 한국어 데이터로 학습시킨 KoBART를 AIHub의 논문자료 요약 데이터 셋으로 미세 조정하여 논문 요약 작업을 수행할 수 있는 모델을 제안하였다. 추후 한국어 자연어 처리 연구에서 더욱 뛰어난 모델이 나온다면 본 논문에서 전처리한 데이터를 활용하여 더욱 뛰어난 문장 요약 성능을 보여주는 모델을 학습시킬 수 있을 것이라 기대된다.

참고문헌

[1] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, arXiv preprint arXiv:abs/1910.13461, 2019.

[2] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving languageunderstanding by generative pre-training.URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/languageunderstanding paper. pdf, 2018.

[3] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1, 4171-4186, 2019.

[4] SKT-AI/KoBART : Korean BART, Github, 2022년 9월 20일 접속, https://github.com/SKT-AI/KoBART