사전 학습 언어 모델을 이용한 한국어 문서 추출 요약 비교 분석

조영래[1], 백광현[1], 박민지[1], 박병훈[1], 신수연[2*]
[1]티쓰리큐(주), [2]한양대학교(서울) 창의융합교육원 교수
florenshio95@gmail.com, toiquen419@gmail.com, qkswnr0924@gmail.com,
warmpark@t3q.com, shinsy@hanyang.ac.kr

A Comparative Study on the Korean Text Extractive Summarization using Pre-trained Language Model

Young-Rae Cho[1], Kwang-Hyun Baek[1], Min-Ji Park[1],
Byung Hoon Park[1], Sooyeon Shin[2*]
[1]T3Q(주), [2]Center for Creative Convergence Education, Hanyang University(Seoul)

요 약

오늘날 과도한 정보의 양 속에서 디지털 문서 내 중요한 정보를 효율적으로 획득하는 것은 비용 효율의 측면에서 중요한 요구사항이 되었다. 문서 요약은 자연어 처리의 한 분야로서 원본 문서의 핵심적인 정보를 유지하는 동시에 중요 문장을 추출 또는 생성하는 작업이다. 이 중 추출요약은 정보의 손실 및 잘못된 정보 생성의 가능성을 줄이고 요약 가능하다. 그러나 여러 토크나이저와 임베딩 모델 중 적절한 활용을 위한 비교가 미진한 상황이다. 본 논문에서는 한국어 사전학습된 추출 요약 언어 모델들을 선정하고 추가 데이터셋으로 학습하고 성능 평가를 실시하여 그 결과를 비교 분석하였다.

1. 서 론

인터넷의 발달은 정보 생산자와 수용자 사이의 경계를 허물고 사람들은 원하는 정보에 쉽게 접근하고 얻을 수 있게 되었다. 그러나 과도한 정보량에 따라 정보를 효과적이고 효율적으로 획득하는 것이 중요한 문제가 되었다. 이를 위해 인공지능이 활용되고 있으며 문서 요약이 대표적이다. 원본 문서가 가지는 의미는 유지하는 동시에 문서의 복잡도는 줄이고 원본 문서보다 길이가 짧은 문서를 추출 또는 생성하는데 주목적을 가진다[1]. 문서 요약은 생성 요약과 추출 요약으로 구분된다. 생성 요약은 원문서에 대한 이해를 바탕으로 원문서에는 없는 단어 혹은 구를 활용하여 요약문을 생성한다. 추출 요약은 원문서의 문장들을 스코어링하고 이 가운데 가장 중요한 문장을 식별하여 요약문을 만든다. 추출 요약은 생성 요약과 달리 원문서의 문장을 그대로 가져오기 때문에 잘못된 정보 생성의 가능성이 적다. 또한 구조화된 글에 대해서는 효율적으로 요약문을 생성할 수 있다.

최근 한국어로 사전 학습된 언어 모형들은 활발하게 구축되고 있다. 특히 트랜스포머 기반 인코더(BERT, Bidirectional Encoder Representation from Transformers)[2][3]가 다양하게 개발되어 배포되고 있으며 우수한 성능을 보여주고 있다. 본 논문에서는 여러 트랜스포머 기반 인코더들을 활용하여 추출 요약에 관한 성능을 비교했다.

2. 문서 요약

2.1 사전학습 언어 모델

사전학습된 언어 모델로 알려진 BERT, RoBERTa, ELECTRA, BigBird를 한국어로 학습한 KoBERT, KLUE-RoBERTa, KoELECTRA, KoBigBird를 중심으로 본 논문의 실험을 위해 사용하고자 한다.

1) BERT는 Transformer 아키텍처 기반으로 텍스트의 양방향 컨텍스트를 동시에 고려하는 특징을 가지고 있다. KoBERT는 한국어 위키 54,000,000개의 단어 및 5,000,000개의 문장으로 학습되었으며

* 교신저자

Vocab의 크기는 8,002이다.

2) RoBERTa는 BERT에서 마스킹 방식을 동적 마스킹 방식으로 바꿨으며 배치 사이즈를 확장해 학습했다[4]. KLUE란 한국어 모델의 성능을 평가하기 위한 일련의 데이터셋으로, 이를 통해 RoBERTa를 사전 학습했다.

3) ELECTRA는 Replaced Token Detection이라는 학습 방식을 제시해 BERT의 학습 방식을 비용 효율적으로 개선한 모델이다. 이는 Generator가 마스킹된 토큰을 예측하고, Discriminator는 Generator가 생성한 토큰이 원본 문장의 토큰과 일치하는지 판별하는 방식의 학습법이다. 결과적으로 BERT에 비해 더 적은 파라미터와 훈련 시간으로 높은 성능을 달성한다[5]. KoELECTRA는 한국어 뉴스, 위키, 나무위키, 신문, 메신저, 웹에서 말뭉치를 모아 34GB의 한국어 텍스트를 학습했다.

4) BigBird는 BERT의 512개의 제한된 입력 시퀀스를 극복하기 위해 기존의 트랜스포머 모델과는 다른 어텐션 메커니즘이 제안된 모델이다[6]. 기존의 페어와이즈 어텐션은 문장의 길이에 따라 계산 복잡도가 제곱으로 증가하므로 긴 문서의 처리에는 비효율적이다. BigBird는 Sparse-attention을 기반으로 동작하며 Global Attention, liding Window Attention, Random Attention 세 가지 어텐션 패턴을 조합해 사용한다. BigBird는 최대 4,096개의 토큰을 입력 시퀀스로 받을 수 있다. 한편, KoBigBird는 모두의 말뭉치, Common Crawl, 뉴스 등의 다양한 한국어 데이터로 학습되었다.

2.2 추출 요약

추출 요약은 텍스트에서 중요한 텍스트 세그먼트를 추출하여 요약하는 기법이다. 이는 원문의 내용을 변형하지 않고, 원래 문장의 순서와 구조를 유지하면서 요약을 생성한다.

입력 시퀀스는 토큰화 과정을 거친 후 [CLS]와 [SEP]토큰으로 감싸져 WordPiece 임베딩을 통해 벡터로 변환되며 이를 input ids라고 한다. 이는 위치 임베딩 attention mask, 세그먼트 임베딩 token type ids와 함께 트랜스포머 레이어로 전달된다.

레이어를 통해 문장의 깊은 표현이 학습되면 출력된 [CLS]토큰은 문장의 전체 의미를 함축하고 있다. 이를 pooler output 벡터라고 한다. pooler output벡터를 추출 요약 레이어를 통해 스코어링하여 문장의 중요도를 점수로 할당한다. 이 점수를 기반으로 나열된 순위에 따라 상위 n개의 문장을 선택해 요약이 생성된다.

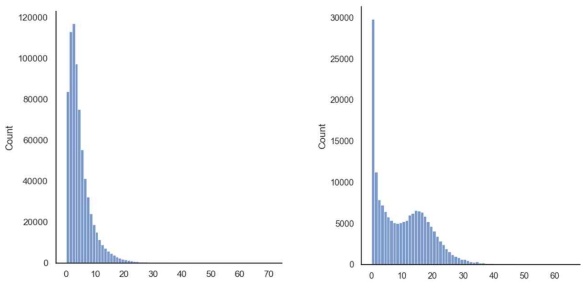
추출 요약은 세 가지의 목표 과제가 제시된다. i. 문서 내에서 중요한 문장을 판별하는 것은 주관적일 수 있으므로, 모델이 일관성 있게 중요한 정보를 판별할 수 있어야 한다. ii. 중요한 정보가 여러 문장에 반복되어 나타날 수 있으므로, 중복된 정보를 효과적으로 제거해야 한다. iii. 문장 간의 관계나 문맥을 고려하여 요약이 연속적이면서 의미가 모호하지 않게 해야 한다.

3. 실 험

3.1 문서 요약 데이터셋

본 논문에서는 AI hub 문서 요약 데이터셋 중 신문기사와 사설 데이터셋을 사용하여 추가 학습을 진행했다. 신문기사 데이터셋은 총 243,977개, 사설 데이터셋은 56,760개의 문서로 이루어져 있다. 신문기사 1개의 문서는 평균 14개 문장, 221개 단어로 구성되어 있고, 사설의 경우 평균 20개 문장, 273개 단어로 구성되어 있다. 정답으로 주어지는 추출 요약 문장은 신문기사 및 사설의 특성에 따라 (그림1)과 같이 앞쪽 문장으로 라벨링 되어 있다. (그림1)은 원문의 몇 번째 문장에 정답 요약문이 얼마나 많이 위치하는지 보여주는 분포 그래프이다.

총 데이터 가운데 신문기사와 사설 각각 10,000개, 20,000개의 문서로 학습 데이터를 구축했다. 데이터에 포함된 html 혹은 url 등의 정보와 특수문자, 정답값이 누락된 데이터는 제거하였다.

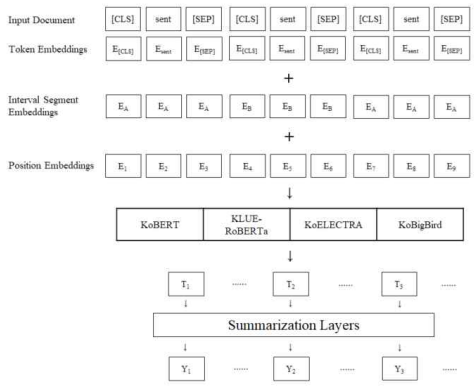


(그림 1) 신문기사(좌)와 사설(우)의 라벨링된 요약문의 문장 위치 분포

3.2 BERTSUM

BERTSUM 모델은 BERT를 추출 요약에 맞게 파인튜닝(fine-tuning)한 모델이다. BERT를 이용하여 입력 데이터를 임베딩하고, 추출용으로 기능하도록 추가한 층에 입력으로 사용된다. 문서 요약에 특화된 BERTSUM에서는 임베딩 단계에서도 기존의 BERT와 다른 형태로 데이터 입력을 받는다. 입력 데이터의 맨 앞에만 [CLS] 토큰을 추가한 기존 BERT와는 달리, 각각의 문장 앞에 [CLS] 토큰을 추가하여 각 문장들의 특징을 해당 토큰에 담을 수 있도록 수정했다. 또한 두 개 이상의 문장에 대해서도 세그먼트 임베딩을 진행할 수 있도록 했다. 이를 통해 문서 단위 입력에 대한 이해를 높인다. BERT 임베딩을 거치고 나면 각각의 문장 표현을 담은 [CLS] 토큰을 출력값으로 얻게 되고 이를 상단에 추가된 요약용 층에 입력한다. 요약층에서는 문장 간의 관계를 파악하기 위해 입력값에 각 문장의 위치를 나타내는 포지셔닝 임베딩을 더해주고, 시그모이드 함수를 통해 각 문장별로 요약에 포함할지를 결정한다[7].

본 논문에서는 임베딩 단계에서 한국어로 사전학습된 KoBERT, KLUE-RoBERTa, KoELECTRA,



(그림2) BERTSUM 모델의 내부 구조

KoBigBird 모델과 그리고 카카오에서 배포한 문서 추출 요약 모델인 Pororo와 비교 실험하였다. 각 사전학습 모델별로 호환되는 토크나이저를 사용한 것 이외에는 동일하게 학습을 진행했다. 학습의 경우 20,000개의 학습 데이터셋에 평균 문장 수를 곱하고 20 epoch로 나누어 배치 사이즈와 스텝을 선정하였다. 이후 <표1>과 같은 20,000개 데이터셋을 학습한 결과 중 우수한 결과를 달성한 모델에 대해 <표2>와 같이 신문 기사 243,977개 및 사설 56,760개로 이루어진 300,737개 전체 데이터셋을 학습하고 비교실험하였다.

4. 성능 평가

문서 요약의 성능 평가를 위해서 가장 많이 사용되는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)를 이용하였다. ROUGE-1과 ROUGE-2 는 각각 모델의 요약 결과와 정답 값 간 겹치는 unigram과 bigram 수를 보는 지표다. ROUGE-L은 LCS 기법을 이용해 최장 길이로 매칭되는 문자열을 측정한다[8]. 본 논문에서는 rouge-score 라이브러리를 사용하여 ROUGE 점수를 측정했다. 또한 한국어 기반 측정을 위하여 Mecab 형태소 분석을 사용했다.

<표1>의 결과에서 LEAD는 뉴스 기사의 특성상 글의 앞쪽에 중요 정보가 있다는 가정을 기반으로, 뉴스 기사의 첫 3문장을 정답 요약문으로 작성한 경우이다. ORACLE은 greedy하게 추출된 정답 요약문이다. 이는 존재하는 대부분의 데이터셋이 생성 요약을 기준으로 작성되었기 때문에 추출 요약 학습에 사용하기 위해 ROUGE-2를 최대화하는 문장을 자체적으로 생성한 것이다[7].

<표1> 사전학습 모델별 BERTSUM 결과-1
(20,000개 데이터셋 학습한 실험군)

Model	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	40.42	17.62	36.67
ORACLE	52.59	31.24	48.87
KoBERT	**59.29**	**38.71**	**54.36**
KLUE-RoBERTa base	52.71	27.82	47.26
KLUE-RoBERTa large	58.27	36.53	52.90
KoELECTRA	**57.56**	**36.27**	**52.86**
KoBigBird	**58.52**	**36.99**	**52.49**

<표2> 사전학습 모델별 BERTSUM 결과-2
(300,737개 데이터셋 학습한 실험군)

Model	ROUGE-1	ROUGE-2	ROUGE-L
KoBERT	39.81	37.43	39.62
KoELECTRA	63.59	45.98	58.10
KoBigBird	**63.97**	**46.41**	**58.35**
Pororo	**61.00**	**43.98**	**59.42**

KoBERT은 <표1>의 20,000개의 데이터셋으로 학습시킨 시킨 결과에서는 ROUGE 점수는 높게 나왔으나 휴먼 검증을 거치는 요약의 질적 평가에서는

성능이 다소 낮았다. <표2>의 전체 데이터셋으로 학습한 경우 ROUGE는 다소 낮았으나 요약의 질은 우수하였다. 이는 BERT의 입력 형식상 512개의 토큰을 넘을 수 없는데, 학습은 문서 단위로 입력되기 때문에 전체 문장을 표현하는 [CLS] 토큰의 개수가 전체 문장 수보다 적게 형성됨을 확인했다.

<표1>의 성능 결과중 KLUE-RoBERTa는 허깅페이스에서 배포한 모델이며 base와 large 모두 KoBERT 토크나이저와 동일한 것을 사용한다. 측정된 ROUGE 점수는 KoBERT 보다는 낮았지만, 요약의 질은 더 높았다. 이는 RoBERTa가 기존에 BERT에서 사용되던 NSP 학습을 제거하고 최대 512개의 토큰 제한을 극복한 것을 확인하였다. Pororo는 RoBERTa를 베이스로 하는 brain-RoBERTa를 사전학습한 모델을 사용하고 있으며 추출 요약을 포함하여 20개 이상의 기능을 제공하는데, 원하는 문장 수만큼 추출 가능하다.

KoELECTRA는 BERT의 MLM 학습을 제거하고 RTD(Replaced Token Detection task)를 이용하여 학습을 진행하였다. <표1>의 실험 결과에서는 KoBERT와 유사한 수준의 성능을 보이고 있으나, 기존 BERT 보다 효율적인 학습이 가능하고 성능이 좋다고 알려진 것처럼[5] <표2>의 실험 결과에서는 우수한 성능을 보여주고 있다.

KoBigBird의 경우, 문서 단위의 데이터셋을 입력 가능하고 BERT의 512개 토큰의 개수 제한의 한계를 극복하고 문장을 표현하는 [CLS] 토큰의 누락이 없이 최대 토큰을 적용할 수 있어 우수한 성능을 확인할 수 있었다.

5. 결 론

본 논문에서는 트랜스포머 기반 인코더를 적용한 BERT에 다양한 사전학습 언어 모델을 활용하여 한국어 문서 추출 요약을 진행하고 ROUGE를 이용하여 성능을 평가하여 그 결과를 비교하였다.

KoBERT의 경우 BERT 모델의 제한된 토큰 개수로 인해 문서 단위 입력에 있어 정보 누락이 발견되었으나, KLUE-RoBERTa, KoELECTRA, KoBigBird는 BERT의 토큰 제한 한계점을 극복하면서 성능이 향상됨을 확인할 수 있었다. 특히, KoBigBird의 경우는 최대 토큰 개수를 4,096개까지 늘리면서 문장 개수가 많은 문서에 대해서도 추출 요약이 가능하며 성능 평가에서도 상대적으로 높음을 확인하였다. 본 논문에서는 ROUGE를 이용하여 성능 평가를 실시 하였으나 ROUGE는 문장 간 겹치는 단어를 수치적으로 판단하여 점수를 내기 때문에 실제 추출 요약문이 의미적으로 원문서의 중요 정보를 추출했는지를 판단하기에는 여전히 어려움이 있다. 따라서 향후 계획으로 요약문의 성능을 평가하기에 더 효율적인 지표를 발굴하고 분석하고자 한다.

참고문헌

[1] 박재언, 김지호, 이홍철. “BERT 기반의 사전 학습 언어 모형을 이용한 한국어 문서 추출 요약 베이스라인 설계” 한국정보기술학회논문지 20, no.6 2022: 19-32.doi: 10.14801/jkiit.2022.20.6.19

[2] 유연준, 홍석민, 이협건, 김영운. “Transformer 기반의 언어모델 Bert와 GPT-2 성능 비교 연구” ASK 2022(29권 1호), 2022

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” arXiv:1810.04805, Oct 2018.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. “RoBERTa: A Robustly Optimized BERT Pretraining Approach” arXiv:1907.11692, Jul 2019.

[5] Kevin Clark, Minh-Thang Luong, Quoc V. Le and Christopher D. Manning. “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators” arXiv:2003.10555, Mar 2020.

[6] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang and Amr Ahmed. “Big Bird: Transformers for Longer Sequences” arXiv:2007.14062, Jul 2020.

[7] Yang Liu. “Fine-tune BERT for Extractive Summarization”, arXiv:1903.10318, Mar 2019.

[8] Kavita Ganesan. “ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks” arXiv:1803.01937, Mar 2018.