문맥 요약을 접목한 한국어 생성형 질의응답 모델 연구

남정재°[1], 김우영[1], 백상덕[1], 이원준[1], 김태용[1], 윤현수[2], 김우주*[1]
[1]연세대학교 스마트 시스템 연구실, [2]연세대학교 산업공학과
{njj97, timothy, eagleeagle, streo222, kasamdi5, hs.yoon, wkim}@yonsei.ac.kr
°제1저자, *교신저자

A Study on Korean Generative Question-Answering with Contextual Summarization

Jeongjae Nam°[1], Wooyoung Kim[1], Sangduk Baek[1], Wonjun Lee[1], Taeyong Kim[1], Hyunsoo Yoon[2], Wooju Kim*[1]
[1]Smart System Lab, Yonsei University, [2]Dept. of Industrial Engineering, Yonsei University
°First Author, *Corresponding Author

요 약

Question Answering(QA)은 질문과 문맥에 대한 정보를 토대로 적절한 답변을 도출하는 작업이다. 이때 입력으로 주어지는 문맥 텍스트는 대부분 길기 때문에 QA 모델은 이 정보를 처리하기 위해 상당한 컴퓨팅 자원이 필요하다. 이 문제를 해결하기 위해 본 논문에서는 요약 모델을 활용한 요약 기반 QA 모델 프레임워크를 제안한다. 이를 통해 문맥 정보를 효과적으로 요약하면서도 QA 모델의 컴퓨팅 비용을 줄이고 성능을 유지하는 것을 목표로 한다.

주제어: 질의응답(QA), 문맥 요약, 텍스트 요약, KoBART

1. 서론

질의응답(Question Answering, QA)은 자연어 처리 기술을 활용하여 주어진 질문과 해당 질문에 관련된 문맥에 대한 답변을 예측하는 작업이다. 질문은 일반적으로 한 문장으로 제공되지만, 문맥은 몇 문장에서부터 문단 수준의 텍스트까지 다양하게 입력된다. 문맥 정보를 질의응답 모델이 이해할 수 있도록 처리하는 과정에서 상당한 컴퓨팅 자원이 필요하다. 이러한 문제는 대규모 텍스트 정보를 처리해야 하는 QA 과정에서 한계로 작용한다.

문맥 정보를 QA 모델에 적용하기 위해 문서 요약 작업을 수행하면 문맥의 길이를 효과적으로 줄일 수 있다. 그러나 문서 요약은 문맥 정보의 일부 손실을 초래하고 성능 저하로 이어질 수 있다. 그럼에도 불구하고, 본 논문에서 제안하는 요약 기반 QA 모델 프레임워크를 통해 성능 저하가 미미함을 확인하였다. 이를 위해 Gensim[1]과 같은 TextRank 알고리즘을 기반으로 한 텍스트 요약 처리를 KoBART 생성형 QA 모델과 함께 사용하여 주어진 문맥 정보에서 핵심 정보만을 추출하였고, 결과적으로 QA 모델의 연산 비용을 감소시키고 성능을 최대한 유지하면서 장문의 문맥 정보를 효과적으로 처리할 수 있음을 입증하였다.

2. 이론적 배경

2.1 질의응답

질의응답은 사용자의 질문에 대한 답변을 생성하는 자연어 처리 작업이다. 이 작업에서 모델은 주어진 질문을 이해하고 관련 문맥을 이해하는 능력이 중요하다. 질의응답은 다른 자연어 처리 작업과는 다르게 지식을 필요로 한다. 사용자의 질문에 대한 답변을 생성하려면 모델이 해당 질문과 관련된 도메인 지식을 보유해야한다. 이러한 지식은 질문에 대한 답을 찾는 데 필요한 문맥 정보로 제공된다. 따라서, 질의응답 모델에는 이러한 문맥 정보가 포함되어야 한다. 그러나 모델에 들어가는 문맥 정보가 길어지는 경우가 많기 때문에 질의응답 작업은 많은 연산 자원을 필요로 한다. 질의응답 모델은 긴 문맥 정보를 처리하고, 이를 기반으로 질문을 이해하고 답변을 생성해야 하므로 상당한 컴퓨팅 자원을 필요로 한다.

2.2 KoBART

BART(Base Auto Regressive Transformer) [1]는 손상된 문서를 재구성하며 학습을 진행하는 식의 사전 학습 방법론을 이용한 인코더-디코더(encoder-decoder) 형태의 언어 모델이다. 사전 학습을 위해 토큰 마스킹(masking), 토큰 삭제, 토큰 채움(infilling) 등의 노이즈 기법을 사용하는데, 임의의 노이즈 함수로 텍스트를 손상시키고, 모델이 원본 텍스트를 재구축할 수 있도록 학습시킨다. BART는 특히 텍스트 생성형 질의응답과 같은 언어 생성 작업에 있어 효과적으로 작동한다.

KoBART[2]는 SKT에서 공개한 한국어 BART 모델로, 40GB 이상의 한국어 텍스트에 대해서 Text Infilling 노이즈 함수를 사용하여 학습된 모델이다. 학습 데이터로는 한국어 위키 백과 이외에도, 뉴스, 책, '모두의 말뭉치 v1.0', 청와대 국민청원 등이 활용되었으며, BPE [2] tokenizer를 활용해 학습되었다.

[1]https://radimrehurek.com/gensim/

[2]https://github.com/SKT-AI/KoBART

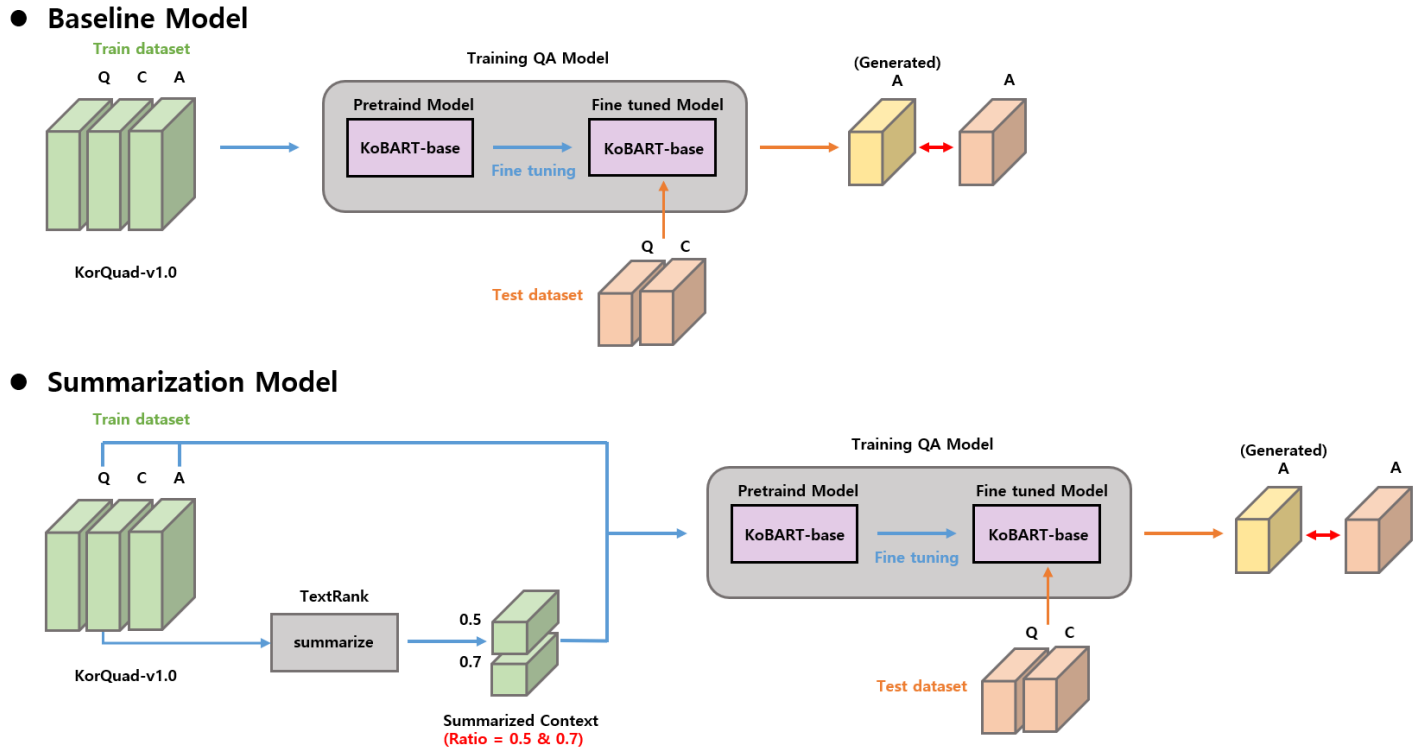

그림 1. 문맥 요약을 접목한 한국어 생성형 질의응답 모델 아키텍처

2.3 문서요약

텍스트 요약(Text Summarization)은 긴 문서를 간결하게 요약하는 과정이다. 텍스트 요약의 접근방식은 크게 추출 요약과 생성 요약으로 구분된다. 추출 요약은 원본 문서에서 중요한 문장 또는 구절을 선택하여 요약문으로 추출하는 방식이다. 이 방식은 주어진 문서의 핵심 내용을 보존하며, 원본 문서의 문장을 그대로 사용하기 때문에 단어의 재구성이나 문장 생성이 필요하지 않다. 생성 요약은 주어진 문서의 의미를 이해하고 새로운 문장을 생성하여 요약문을 구성하는 방식이다. 이 방식은 추출 요약과 달리 원본 문서에 없던 문장을 생성하며, 자연어처리 기술과 딥러닝 모델을 활용한다. 그렇기에 일반적으로 생성 요약 작업은 딥러닝 모델이 활용되어 컴퓨팅 연산자원을 상당히 필요로 한다. 하지만, 추출 요약 작업은 규칙 기반 알고리즘으로 동작하기에 생성 요약 작업에 비해 적은 컴퓨팅 자원을 필요로 한다.

2.4 TextRank

TextRank [3]는 키워드 추출, 문장 추출, 텍스트 요약 등의 자연어처리 작업에 활용되는 그래프 기반 키워드 추출 알고리즘이다. 이 알고리즘은 PageRank [4] 알고리즘을 텍스트 데이터에 적용한 방식으로서 중요한 키워드 또는 문장을 식별하는데 사용된다. TextRank의 핵심 아이디어는 문서 내에서 문장 간의 관계를 그래프로 표현하고, 그래프의 중요도를 계산하여 핵심 문장을 추출하는 것이다. TextRank는 문서를 문장 단위로 분리하고, 각 문장에 포함된 단어를 그래프의 노드로 표현하여 문장 간의 유사도를 측정한다. 우리는 문맥을 요약하여 적은 컴퓨팅 자원을 목적하기에 추출형 문서요약 모델인 TextRank를 사용한다.

3. 방법론

본 논문에서는 생성형 질의응답 모델 학습에 대해 기존의 문맥 정보를 그대로 사용하는 방법과 우리가 제안하는 TextRank 알고리즘으로 요약된 문맥 정보를 사용하는 요약-질의응답 모델 프레임워크를 비교 실험한다. 각 모델의 구조는 그림 1과 같다. 베이스라인 모델은 사전 학습된 KoBART 모델을 생성형 질의응답 작업에 미세 조정하여 사용한다. 학습이 완료된 KoBART 질의응답 모델에 대해 질문과 문맥을 입력하면 답변을 생성한다.

우리가 제안하는 모델은 그림 1의 Summarization Model로서 요약된 문맥을 통한 성능 비교를 위하여 TextRank와 KoBART를 결합한 구조의 모델이다. 데이터셋에서 추출한 질문, 문맥, 답변이 토큰화 전에 TextRank를 통해 문맥을 실험에서 설정한 비율에 맞게 요약한 후 사용한다. 이후 학습 방식은 베이스라인과 동일하다.

4. 실험

앞서 제안한 모델 아키텍처를 기반으로 문맥 요약을 접목시킨 한국어 생성형 질의응답 실험을 수행하였다. 본 실험에서는 한국어 생성형 질의응답 모델을 베이스라인으로 설정한 뒤, 베이스라인 모델과 문맥 요약을 접목한 모델을 비교한다.

4.1 데이터

실험 수행을 위해 사용한 데이터셋으로는 한국어 질의응답 데이터셋인 KorQuAD 1.0[3] 를 이용했다. KorQuAD 1.0은 한국어 질의응답을 위해 만들어진 데이터셋으로, 모든 질의에 대한 답변은 해당 위키피디아 문서 문단의 일부 하위 영역으로 이루어져 있다. 이 데이터셋은 영어 질의응답 데이터셋인 SQuAD [5] 버전 1.0과 동일한 방식으로 구성되어있다. 전체 데이터는 1,560개의 위키피디아 문서에 대해 10,645 건의 문단과 66,181 개의 질의응답 쌍으로, 학습 데이터셋은 60,407 개, 검증 데이터셋은 5,774 개의 질의응답 쌍으로 구분되어 있다. 해당 데이터셋을 사용하여 답변의 분포와 질문에 답하는데 필요한 근거의 유형을 파악할 수 있다. 본 실험에서 사용한 데이터셋에 대한 자세한 설명은 표 1에 나와 있다.

Datasets	**# Wikipedia Article**	**# Context**	**# QA pair**
KorQuAD 1.0	1,560	10,645	66,181

표 1. 한국어 질의응답을 위한 데이터셋

4.2 실험 세부사항

문맥 요약을 접목한 모델에서는 토큰화를 진행하기 전에 TextRank 알고리즘을 사용한 텍스트 요약 라이브러리인 gensim 3.8.3 버전을 사용하여 문맥에 대한 요약을 수행하였다. 해당 라이브러리를 사용하여 요약을 진행할 때 사용되는 하이퍼 파라미터 ratio 값을 지정할 수 있다. 해당 값은 요약할 문장 수 비율을 결정하는 0에서 1 사이의 실수 값으로 지정할 수 있고, 우리는 0.5, 0.7 두 개의 값으로 설정하여 요약 결과를 확인하였다. 실험을 위해 설정한 하이퍼 파라미터는 아래 표 2와 같다.

4.3 평가방법

평가는 총 두 가지 성능 지표를 기준으로 진행하였다. 첫 번째 성능 지표는 예측값에서 공통 단어 수를 전체 단어 수로 나눈 비율인 정밀도와 실제값에서 공통 단어 수를 전체 단어 수로 나눈 비율인 재현율의 조화평균인 F1-score로 선정하였다. 이는 어절 기준으로 얼마나 중복되는지를 고려한 점수이다. 두 번째

[3] https://korquad.github.io/

batch size	128
epoch	50
learning rate	0.00001
scheduler	StepLR
optimizer	AdamW
step size	500
gamma	0.1

표 2. 하이퍼 파라미터 설정

성능 지표는 주어진 질문과 대답 쌍에 대해 모델의 예측과 실제 정답을 비교하여 문자열이 완전히 일치하는지를 계산하는 지표인 Exact-Match(EM)로 선정하였다. 예측된 문자열이 실제 정답과 정확히 일치하는 경우, 해당 지표 값은 1, 그렇지 않으면 0으로 산정되어 생성형 질의응답 모델 전체 결과에 대한 성능을 확인하였다.

4.4 결과

본 실험에서 사용된 데이터셋의 평균 F1-score와 Exact-Match 결과는 표 3에 제시되었다. TextRank 알고리즘을 사용하여 ratio 값을 다르게 설정함에 따라 성능이 다르게 나온 것을 동시에 확인하였다. 요약 ratio 값을 0.7로 설정한 모델은 베이스라인 모델에 비해 F1-score와 Exact-Match 기준 약 0.17 가량 낮았고, 요약 ratio 값을 0.5로 설정한 모델은 베이스라인 모델에 비해 약 0.09 가량 낮음을 확인하였다. 이와 같이 TextRank를 사용하여 문맥 요약을 진행하였음에도 성능 하락폭이 크지 않음을 확인할 수 있다.

또한, 유의미한 시사점은 요약 비율이 0.5일 때보다 0.7일 때의 문맥의 정보가 더 많음에도 불구하고 성능 하락폭이 더 컸다. 이는 적절한 문맥 정보 선별하는 문맥 요약 작업이 질의응답 작업의 성능 향상에 도움이 됨을 보여준다. 더 나아가, 질의응답에 적절한 요약 방법론을 접목하기 위한 추가적인 연구가 필요하다.

	Baseline	**TextRank** (ratio = 0.7)	**TextRank** (ratio = 0.5)
F1-score	0.5980	0.4292	0.5128
Exact-Match	0.5062	0.3351	0.4186

표 3. KorQuAD 1.0의 생성형 질의응답 모델 평가 결과

5. 분석

본 논문에서 문맥 요약 작업을 접목한 요약-질의응답 모델 프레임워크의 효과성을 확인하고자 진행한 실험에서 모델이

Ground Truth	Baseline	TextRank (ratio = 0.7)	TextRank (ratio = 0.5)
'학생회관 건물 계단'	'경희대 내'	'경희대학교 내'	'경희대학교 내 어디인가'
'서울지방경찰청 공안분실'	'청량리경찰서'	'청량리경찰서'	'청량리경찰서에서 약 1시간'
'제89조'	'제89조'	'제89조'	'헌법 제9조'
'허영'	'허영석'	'헌법학부 장관'	'허영석'
'로널드 레이건 대통령'	'로널드 레이건'	'로널드 레이건'	'로널드 레이건'
'국무장관'	'국무장관관'	'알렉산 메이그스 헤이그 2세'	'국무장관'
'노터데임 대학교'	'노터데임 대학교'	'노터데임 대학교'	'노터데임 대학교'

표 4. Ground Truth와 Baseline 및 TextRank 생성 텍스트 비교

생성한 답변 결과를 분석한다. 표 4는 실제 정답과 베이스라인, 문맥 요약을 접목시킨 모델의 답변 생성 결과의 예시를 제시한다. 모델별로 생성된 답변 텍스트와 실제 정답을 비교해보면 실제 정답과 적절하게 일치하는 답변 결과도 있지만, 완전히 다른 답변이 나오는 예도 있다. 이는 모델을 평가하는 정량적인 성능 지표로 비교하는 것은 한계가 있어 아래와 같이 답변 생성 결과 비교를 통해 유의미한 결과를 확인할 수 있음을 시사한다.

실제 정답이 "학생회관 건물 계단"인 경우, 베이스라인 모델은 "경희대 내"라고 생성되었고, 실제 정답이 "국무장관"인 경우, "국무장관관"으로 실제 정답과 유사하지만 한 글자 차이로 다르게 생성되는 예시가 있다. 하지만, "제89조", "노터데임 대학교"와 같이 정답을 성공적으로 생성하는 예시도 존재한다.

실제 정답이 "제89조"인 경우, TextRank(ratio =0.7)은 "제89조"라고 완벽하게 생성해낸 반면, TextRank(ratio=0.5)의 경우 "헌법 제9조"와 같이 상당히 유사하지만, 결과적으로 생성된 텍스트에 있어서 아쉬운 부분이 존재함을 보여준다. 모두 정답인 경우는 "노터데임 대학교"의 예시가 있다.

6. 결론

본 논문에서는 한국어 생성형 질의응답 모델에 문맥 요약 작업을 접목하여 요약-질의응답 모델 프레임워크를 제시하였다. 문맥 요약 모델로 적용한 TextRank는 텍스트 내에서 중요한 핵심 단어 및 문장을 식별하고, 이를 기반으로 문맥 입력 정보에 대한 요약을 도출할 수 있다. 우리는 문맥 정보를 요약하여 활용한 질의응답 모델과 기존의 질의응답 모델의 성능을 비교한 결과 성능 차이가 크지 않음을 보였다. 이로써 질의응답 모델의 컴퓨팅 연산 비용을 효과적으로 줄이면서도 유사한 성능을 유지할 수 있음을 보였다. 또한, 문맥 요약 비율이 0.7일 때 문맥의 정보가 더 많음에도 불구하고 0.5일 때보다 성능 하락폭이 더 크다는 결과를 얻었다. 이는 질의응답 작업에서 효과적인 문맥 요약 방법론에 대한 연구가 추가적으로 필요함을 시사한다.

감사의 글

이 논문은 국토교통부의 스마트시티 혁신인재육성사업으로 지원되었습니다.

본 연구는 연세대학교 딥러닝이론및응용(IIE7721)의 연장으로 진행되었으며, 윤현수 교수님께 감사드립니다.

참고문헌

[1] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[2] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Aug. 2016. [Online]. Available: https://aclanthology.org/P16-1162

[3] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404–411, 2004.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, Vol. 30, pp. 107–117, 1998. [Online]. Available: http://www-db.stanford.edu/~backrub/google.html

[5] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension

of text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Nov. 2016. [Online]. Available: https://aclanthology.org/D16-1264