과학 분야 다중 문서 요약을 위한 개선된 요약 및 평가 방안 모색*

김유빈[0], 남우성, 성무진**
경희대학교 컴퓨터공학부
kyb0314@khu.ac.kr, dntjd123kr@khu.ac.kr, mujeensung@khu.ac.kr

Towards Better Summarization and Evaluation for Scientific Multi-document Summarization

Yubin Kim[0], Woosung Nam, Mujeen Sung
School of Computing, Kyunghee University
kyb0314@khu.ac.kr, dntjd123kr@khu.ac.kr, mujeensung@khu.ac.kr

요 약

본 논문은 과학 분야의 체계적 문헌 고찰(SLR)을 자동화하기 위해 최신 언어 모델을 활용하여 다중 문서 요약의 정확성을 개선하며, 다양한 평가 지표로 모델의 성능을 검증한다. 기존의 문서 요약 모델과 최신 언어 모델이 생성한 요약을 표면적 유사성을 기반으로 한 평가 지표와 대형 언어 모델 기반의 평가 지표를 통해 성능을 비교한 결과, 표면적 유사성을 중점으로 한 평가에서는 기존의 문서 요약 모델이 더 높은 점수를 기록했지만, 대형 언어 모델 기반의 평가에서는 최신 언어 모델이 더 우수한 성과를 보였다. 이로써 대형 언어 모델 기반의 평가 지표가 맥락적 일치와 요약 품질을 보다 잘 반영할 가능성을 제시하였다. 과학 분야 다중 문서 요약에서 대형 언어 모델 기반의 평가 지표가 다양한 문맥과 의미를 반영할 가능성을 확인했으며, 이를 통해 과학 분야 문헌 요약 자동화의 정밀도 향상에 기여할 수 있음을 시사한다.

1. 서 론

체계적 문헌 고찰 (Systematic Literature Review, SLR)은 특정 연구 질문에 대한 관련 연구를 체계적으로 수집 및 요약하는 중요한 과정이다. 특히 과학 분야 중 생물의학 도메인의 SLR은 의료적 개입에 앞서 해당 연구의 효과성을 평가하기 위해 필수적이다. 그러나 문헌을 수작업으로 검토하고 요약하는 데는 상당한 시간과 비용이 소요되며, 급증하는 연구 문서의 양을 인간이 처리하는데 한계가 있다. 이를 해결하기 위해 SLR 과정을 자동화함으로써 시간과 비용을 절감할 방안이 필요하다.

이를 해결하기 위해 다중 문서 요약 (Multi-document Summarization, MDS) 연구를 적용하여 여러 문서에서 자동으로 핵심 정보를 추출하고, 요약함으로써 SLR의 일부 과정을 자동화하는 방법이 활발히 연구되고 있다.

기존의 체계적 문헌 고찰을 위한 다중 문서 요약 (Multi-document Summarization for Literature Review, MSLR) 연구들은 다양한 모델과 기법을 활용하여 자동화를 시도해 왔으나, 종종 요약의 질적 평가와 정확도에서 한계를 드러냈다 [1]. 이 연구에서는 최신 언어 모델로서 GPT-4o-mini를 사용하여 과학적 다중 문서 요약의 정확성과 일관성을 개선하고자 하며, GEval 평가 지표를 통해 평가를 체계화하여 더욱 정교하고 신뢰할 수 있는 요약 평가 방안을 제시한다.

2. 데이터셋 구축

본 연구는 무작위 대조군 연구 (Randomized Controlled Trials , RCTs)의 제목과 초록을 기반으로 생성 요약을 수행하는 것을 목표로 한다. 이를 위해 Cochrane 협회의 회원들이 작성한 4,528개의 체계적 문헌 고찰 (SLR) 데이터셋을 활용하였으며, 데이터셋에는 각 SLR 초록과 RCT의 제목과 초록이 포함되어 있다 [2]. 입력 데이터로는 각 RCT의 제목과 초록을 사용하였으며, 목표 요약으로는 SLR 초록의 저자 결론 부분을 설정하였다.

데이터셋은 무작위로 학습 (3,752개), 검증 (470개), 평가 (470개)로 나누어 구성하였다. 각 SLR은 평균적으로 10개의 RCT를 포함하며, RCT 초록의 평균 길이는 245단어, 저자 결론 부분은 평균 75단어로 이루어져 있다. 전체 데이터셋의 통계는 [표 1]에 표기되어 있다. 모델은 입력 데이터에 기반하여 요약을 생성하고, 이를 목표 요약과 비교하여 성능을 평가하였다.

	학습	검증	평가
SLR	3,752	470	470
RCT	40,497	5,033	5,678

[표 1] 데이터셋 통계

3. 실험

3.1 실험 세팅

baseline으로 사용한 모델은 Longformer Encoder-Decoder-base-16384로, Longformer Encoder-Decoder 모델에서 16K 토큰을 처리할 수 있도록 bart-base의 position embedding matrix가 16번 복사된 모델이다. 이 모델은 긴 문서 요약이나 질의응답에서 유용하게 사용될 수 있다. 학습 데이터셋으로 파인튜

* "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2024년도 SW중심대학사업의 결과로 수행되었음"(2023-0-00042)
** 교신저자 (Corresponding Author)

닝된 Longformer Encoder-Decoder-base-16384-cochrane 모델을 사용하였다 [3]. 생성 요약을 위해 사용한 모델은 GPT-4o-mini [4]로, OpenAI의 GPT 시리즈에서 파생된 소형 자연어 처리 모델이며 파라미터 수가 적어 메모리 요구량이 낮고 빠른 처리 속도를 제공한다. 설명의 용이성을 위해 본 논문에서 Longformer Encoder-Decoder-base-16384-cochrane 모델은 LED-16K로 GPT-4o-mini 모델은 GPT-4o로 표기한다.

3.2 실험 방법

기존 다중 문서 요약에서는 입력 데이터 길이가 길어 대부분의 모델이 입력 토큰 한계를 초과한다는 어려움이 있다. 이를 보완하기 위해 원본 데이터를 요약하기 전에 추출 요약을 진행하고, 추출 요약한 데이터와 원본 데이터를 비교하여 성능을 분석했다. 최종적으로 본 실험에서는 두 가지 데이터셋 (원본 데이터셋, 추출 요약한 데이터셋)을 사용하여 LED-16K와 GPT-4o 모델 각각에 대해 요약을 생성하고 성능을 평가하였다.

3.2.1 lecture summarizer를 사용한 추출 요약

추출 요약을 위해 MSLR 공유과제에서 가장 높은 점수를 얻은 SciSpace 팀이 사용한 Lecture summarizer 모델을 사용하였다 [5][6]. SLR 당 RCT의 제목과 초록을 입력 데이터로 하였으며, 초록이 없는 경우에는 제목을 대신 반복하여 추가하였다. 입력 데이터에 대해 0.5 길이의 추출 요약을 생성하도록 설정하였다.

3.2.2 LED-16K와 GPT-4o를 사용한 생성 요약

LED-16K와 GPT-4o 모델에서 원본 데이터셋과 Lecture summarizer를 사용하여 추출 요약한 데이터셋에 대해 요약을 생성하였다. LED-16K 모델의 최대 입력 토큰 수인 16,384가 넘어가는 입력 데이터의 경우 최대 입력 토큰 수만큼 입력 데이터로 사용한다. GPT-4o 모델은 원 샷 프롬프팅을 사용하여 요약을 생성하였고, 이때 사용된 예제는 학습 데이터셋 중 10개의 RCT를 포함하는 데이터를 선정하여 진행하였다. 프롬프팅에 사용된 프롬프트는 "Summarize this for a doctor: {abstracts}"이다. 두 모델 모두 최대 출력 토큰 수는 1,024로 지정하였다. 평가에 사용할 목표 요약 예시와 LED-16K, GPT-4o로 생성한 요약 예시는 [표 2]에 제시하였다.

3.2.3 요약 평가

모델이 생성한 요약을 평가하기 위해, 형태적 일치를 고려하는 ROUGE 지표와 의미적 유사성을 고려하는 언어 모델 기반의 BERTScore, 그리고 대형 언어 모델 기반의 GEval 지표를 사용하였다 [7]. ROUGE와 BERTScore는 목표 요약과 모델이 생성한 요약을 비교하여 측정하였고, GEval의 평가 지표는 요약 작업에 맞춤화된 지표인 FineSurE의 기준을 적용하여 신뢰성, 완전성, 간결성을 사용하였다 [8]. 신뢰성은 정보 왜곡 없이 입력 텍스트에 충실한지, 완전성은 요약에 핵심 사실이 모두 포함되었는지, 간결성은 불필요한 정보 없이 요약이 간결한지를 평가한다. GEval의 점수 산출은 각 요약에 대해 세 가지 지표마다 1~5점의 점수를 매긴 후, 출력된 토큰의 확률을 추출하여 가중합을 적용해 최종 점수를 계산했다. 마지막으로 각 SLR에서 산출된 점수를 모두 평균내어 최종 점수를 도출하였다.

입력 다중 문서 : Is milking and stripping chest tubes really necessary? The purpose of this study was to determine if chest tubes that are not milked or stripped occlude more frequently than milked or stripped tubes, and if the amount of drainage varies according to the treatment of the tubes. (생략) **목표 요약** : There are insufficient studies which compare differing methods of chest drain clearance to support or refute the relative efficacy of the various methods in preventing cardiac tamponade. Nor can the need to manipulate chest drains be supported or refuted by results from RCTs.	
LED-16K 요약	There is insufficient evidence from randomised controlled trials to support or refute the use of milking or stripping as a means of improving mediastinal drainage in patients undergoing surgery for myocardial revascularisation.
GPT-4o 요약	This study investigated whether manipulating chest tubes through milking or stripping affects occlusion rates and drainage volume. The findings suggest that neither milking nor stripping is necessary for effective chest tube care. Instead, proper positioning of the tubes to promote continuous drainage is recommended. Additionally, in patients undergoing myocardial revascularization, manipulation of chest tubes was not required during the first 8 hours post-surgery, as it did not improve drainage or prevent complications.

[표 2] 입력 데이터 및 생성된 요약 예시

3.3 실험 결과

요약에 사용된 모델에 따른 평가 데이터셋의 ROUGE, BERTScore, GEval 점수는 [표 3]와 [표 4]에 작성되어 있다. 표의 세로축은 요약에 사용한 모델들을 나타내며, 이때 Lecture summarizer를 활용하여 추출 요약 과정이 진행된 경우에는 LS로 표현하였다. Lecture summarizer을 통한 추출 요약을 제안한 SciSpace 팀은 BigBird PEGASUS 모델로 생성 요약을 진행하였고 생성된 요약은 제공되지 않아 ROUGE와 BERTScore 점수만 표에 추가하였다. 가로축의 ROUGE와 BERTScore은 SLR의 목표 요약과 생성 요약을 비교한 결과를, GEval은 RCT의 제목 및 초록을 기반으로 생성 요약을 평가한 결과를 나타낸다.

ROUGE와 BERTScore를 확인하면 LED-16K 모델의 요약이 더 좋은 점수를 보였고, 반면 GEval은 GPT-4o 모델의 요약이 더 좋은 점수를 보였다. 이때 lecture summarizer를 통한 추출 요약 과정은 점수에 크게 영향을 미치지 않는 것으로 확인되었다.

모델	ROUGE1	ROUGE2	ROUGEL	BERTScore(F1)
Scispace [5]	**0.262**	0.057	**0.197**	0.859
LED-16K	0.247	**0.064**	0.179	**0.872**
LS + LED-16K	0.241	0.062	0.176	0.870
GPT-4o	0.187	0.034	0.107	0.835
LS + GPT-4o	0.203	0.034	0.114	0.840

[표 3] 표면적 유사성 기반의 평가 점수

모델	신뢰성	완전성	간결성
LED-16K	2.161	1.413	3.066
LS + LED-16K	2.134	1.343	3.303
GPT-4o	**4.660**	**4.542**	**4.243**
LS + GPT-4o	4.601	4.488	4.141

[표 4] 대형 언어 모델 기반의 평가 점수

4. 분석

4.1 LED-16K와 GPT-4o의 질적 요약 분석

LED-16K 모델과 GPT-4o 모델이 생성한 요약을 확인해보면, LED-16K 모델은 입력 문장과 관계없이 특정 패턴과 표현을 반복하는 요약을 생성한다. 구체적으로, 생성된 요약은 "There is in sufficient evidence to" 또는 "There is no evidence to"로 시작하는 경향을 보인다. 반면 GPT-4o 모델은 요약 생성 시 특정 패턴을 보이지 않으며 SLR에 맞는 다양한 표현을 생성한다.

4.2 ROUGE, BERTScore 비교

ROUGE는 n-그램의 중복성과 정확성을 강조하여 생성된 요약이 목표 요약과 얼마나 일치하는지를 평가하는 데 중점을 두고, BERTScore는 BERT 모델을 통한 임베딩 기반의 유사성을 통해 생성된 요약과 목표 요약 사이의 의미적 유사성을 평가하는 데 중점을 둔다. 실험 결과 LED-16K가 ROUGE와 BERTScore 모두 GPT-4o보다 높은 성능을 나타냈다. LED-16K 모델은 학습된 데이터의 문장 패턴을 그대로 반영하는 경향이 강해 목표 요약과 유사한 문장 구조와 어휘가 반복되었기 때문이라고 예상할 수 있다. 이러한 경향 때문에 질적인 유사성이 적더라도 형태적인 유사성이 높아 높은 점수로 측정된 것으로 파악된다. 반면 GPT-4o의 경우 생성한 요약이 동일한 패턴을 보이지 않고, 단어를 반복하기보다는 다양한 표현을 생성했기에 낮은 점수로 측정된 것으로 파악된다.

4.3 GEval 점수 비교

GEval은 대형 언어 모델을 활용하여 텍스트의 품질을 평가하는 지표로, 다양한 지표로 평가가 가능하며 텍스트의 내용과 구조를 포괄적으로 분석한다. 본 논문에서는 신뢰성, 완전성, 간결성 지표로 생성된 요약을 평가하였는데, GPT-4o는 LED-16K보다 월등히 높은 점수를 기록했다. 이는 GPT-4o가 핵심 내용을 파악하여 문맥에 맞는 요약을 생성하는 데 강점이 있기 때문으로 파악된다. 반면, LED-16K는 고정된 패턴을 자주 반복하면서 중요한 내용을 생략하거나 불필요한 내용을 추가하는 경향이 있어 낮은 점수를 기록한 것으로 분석된다.

4.4 추출 요약에 따른 성능

Lecture summarizer를 사용한 추출 요약 과정을 추가한 데이터셋과 그렇지 않은 데이터셋 간의 성능 차이는 미미했다. 이는 LED-16K와 GPT-4o가 긴 입력 토큰을 처리할 수 있는 특성이 있어 Lecture Summarizer가 데이터를 압축하여 요약하는 과정이 크게 영향을 미치지 않았기 때문이라고 판단된다. 실제로 GPT-4o는 추출 요약하지 않은 데이터셋에 대해서도 모두 처리할 수 있었고, LED-16K의 경우도 16개의 SLR에 대해서만 입력 토큰이 초과했다. 또한 활용한 Lecture summarizer에서 잘못된 추출 요약이 발생한 경우 error propagation이 발생하여 최종 성능에도 영향을 미쳤을 가능성이 있다.

5. 결론

본 논문에서는 과학 분야에서 체계적 문헌 고찰의 자동화를 위해 최신 언어 모델을 사용하여 다중 문서 요약을 수행하고, GEval 평가 지표를 통해 더욱 신뢰할 수 있는 요약 평가 방안을 도입하였다.

LED-16K 모델과 GPT-4o 모델을 대상으로 형태적 일치를 고려하는 ROUGE, 의미적 유사성을 고려하는 BERTScore, 대형 언어 모델 기반의 GEval 평가 지표를 통해 생성된 요약을 평가하고 성능을 비교한 결과, ROUGE와 BERTScore에서는 LED-16K 모델이 우수한 성능을 보였으나, GEval 지표에서는 GPT-4o 모델이 더욱 높은 점수를 보였다. 이러한 결과는 GEval이 표면적 유사성을 평가하는 ROUGE와 BERTScore보다 맥락적 일치와 요약의 품질을 더 잘 반영할 가능성을 시사한다.

향후 연구에서는 더욱 적합한 언어 모델 기반 평가 지표와 방식을 개발하여, 인간 평가와의 상관성을 높이고 GEval의 타당성을 검증할 계획이다. 이를 통해 과학 분야의 문헌 고찰 자동화에 의미와 문맥을 반영하는 평가 지표를 제안하고, 요약 품질을 더욱 정교하게 평가할 수 있는 체계를 구축하여 실질적인 기여를 할 것이다.

6. 참고문헌

[1] Lucy Lu Wang., et al. "Overview of MSLR2022: A Shared Task on Multi-document Summarization for Literature Reviews", Proceedings of the Third Workshop on Scholarly Document Processing 2022

[2] Byron C. Wallace., et al. "Generating (Factual?) Narrative Summaries of RCTs: Experiments with Neural Multi-Document Summarization", AMIA Informatics Summit 2021

[3] https://huggingface.co/allenai/led-base-16384-cochrane

[4] https://openai.com/index/GPT-4o-mini-advancing-cost-efficient-intelligence/

[5] Kartik Shinde., et al. "An Extractive-Abstractive Approach for Multi-document Summarization of Scientific Articles for Literature Review", Proceedings of the Third Workshop on Scholarly Document Processing 2022

[6] https://github.com/dmmiller612/lecture-summarizer

[7] Yang Liu., et al. "G-EVAL: NLGEvaluation using GPT-4 with Better Human Alignment", EMNLP 2023

[8] Hwanjun Song., et al. "FineSurE: Fine-grained Summarization Evaluation using LLMs", ACL 2024