감성 극성 분류 모듈과 다중 에이전트 구조를 활용한 개선된 다중 문서 요약 기술 연구

조병웅, 봉한성, 홍주화*
성균관대학교

Whale1510@g.skku.edu, peterpen0110@g.skku.edu, j.hong@skku.edu*

multi-document summarization pipeline utilizing a sentiment polarity classification module and a multi-agent architecture.

Cho Byeong Ung, Han Sung Bong, Joo Wha Hong*
SungKyunKwan Univ.

요 약

하나의 문서만을 요약하는 단일 문서 요약과 달리, 다중 문서 요약은 상품 리뷰처럼 동일한 주제에 대한 수천 개의 문서 집합을 요약하는 과제이다. 이러한 다중 문서 요약은 요약 대상의 방대한 수와, 정보의 중복과 충돌 등의 요인으로 인해 단일 문서 요약 방법론의 확장으로는 해결할 수 없다. 그 때문에 ML 기반의 방법론을 적용하여 주요 정보를 추출하는 추출 요약과, 사전학습 Transformer 모델이나 LLM 을 활용한 추상 요약을 결합하여 다중 문서 요약을 수행하는 하이브리드 방법론이 등장하였다. 하지만 그러한 방법론은 추출 요약에서 발생하는 정보 누락 문제와 추상 요약에서 발생하는 환각 문제를 그대로 상속받으며, 리뷰에서 발생하는 긍정, 부정 의견 두 가지 극성을 제대로 반영하지 못한다. 이에 본 논문에서는 기존 방법론 구조에, 감성 극성 분류 모듈과 QA 기반의 feedback 을 수행하는 다중 에이전트 구조를 추가한 확장된 로직을 제안한다. 해당 로직을 적용했을 때, 기존 방법론보다 더 높은 성능을 달성함을 확인하였다.

Ⅰ. 서 론

온라인 상거래의 급성장과 함께 리뷰는 가격이나 상품의 매력에 버금가는 핵심 구매 결정 요소로 부상했으나, 방대한 리뷰를 소비자가 직접 모두 읽는 것은 현실적으로 불가능하다. 이에 따라 다수의 리뷰를 하나로 요약하는 Multi-Document Summarization(MDS) 기술이 주목되어 연구되어 왔다.

최근에는 Transformer 기반 모델이나 대형 언어 모델(LLM)을 활용한 추상 요약 방법론과 기존 ML 기반의 추출 요약 방식을 결합한 하이브리드 접근법[1][2]이 단순 LLM 기반 방법론에서의 토큰 제한 문제를 극복하고 리뷰 요약에 활용되고 있다. 그러나 이 방식은 여전히 LLM 의 환각, ML 의 정보 누락 문제를 상속하며 사용자 선호(양극성)를 반영하지 못하고 평균적인 요약문을 생성한다는 한계가 있다.

본 연구에서는 기존 하이브리드 방법론에서 (1) 긍/부정 리뷰를 분리하여 요약할 수 있도록 감성 분류 모듈을 추가하고, (2) Reflection 기법을 본 MDS 과제에 적용하고자 반복적 QA 기반 feedback 을 수행하는 논리적으로 분리된 다중 에이전트 구조를 추가하여 확장하는 기법을 제안한다.

Ⅱ. 관련 연구

Multi-Document Summarization

최근 LLM 의 등장은 자연어 처리 분야에서 큰 혁신을 가져왔고, 요약 분야에서도 기존 방법론들보다 LLM 을 활용한 요약이 기존 방법론들보다 더 뛰어나다는 것이 증명되었다.[3] 그러나 프롬프트에 원문을 탑재해야 하는 LLM 기반의 요약 방법론은 토큰 제한 문제로 인하여 수백~수천 개의 리뷰를 요약해야 하는 MDS 에서 적용하기 어렵다는 문제를 가지고 있다.

이를 극복하기 위해서 ML 기반의 추출 요약과 LLM 기반의 추상 요약을 결합하여 토큰 제약을 해결하려는 방법론[1][2]이 제시되었다. 그러나 이러한 방법론은 사용자 선호도를 반영하지 못하며 환각 및 정보 누락이 발생한다는 문제가 여전히 존재한다. 실제 MDS 의 환각 문제를 추적한 논문에서는 평균적으로 45%~75% 정도의 환각이 발생한다고 보고했다.[4]

Reasoning

LLM 의 환각 문제를 해결하기 위해 추론 능력을 활용하고자 하는 여러 연구가 진행되었다. LLM 추론을 활용하여 복잡한 문제를 해결하고자 하는 Reasoning 분야가 개척되었고, 이러한 발전 속에서 자가 피드백을 통해 출력을 개선하는 Reflection 방법론[5]이 새롭게 등장하면서 환각을 줄일 해결책이 등장하였다.

요약 과제에서도 그러한 Reflection 기법을 도입하여 환각과 정보 누락을 해결하고자 하는 연구[6]가 제안되었지만, 단일 문서에만 적용되며 단일 구조로 인한 자기 편향 문제가 발생한다는 한계가 있었다.

Ⅲ. 제안 기법

본 연구에서는 분류, 추출, 생성, 평가 단계로 구성된 다중 문서 요약 기법을 제안한다. 제안 기법은 분류

단계를 통해 긍정, 부정 리뷰 분류를 각각 수행하게 되며, 해당 리뷰들의 긍정, 부정 비율을 함께 결과로 제공한다.

이때 생성 단계와 평가 단계에서는 각각 단일 LLM API 인스턴스를 사용하되, '생성자'와 '평가자'로 프롬프트 구조를 명확하게 구분하여 논리적으로 독립된 역할을 갖고 상호작용하는 다중 에이전트 구조를 구축하여 환각과 정보 누락을 개선하면서도 자기 편향 문제를 해결하는 구조로 설계하였다.

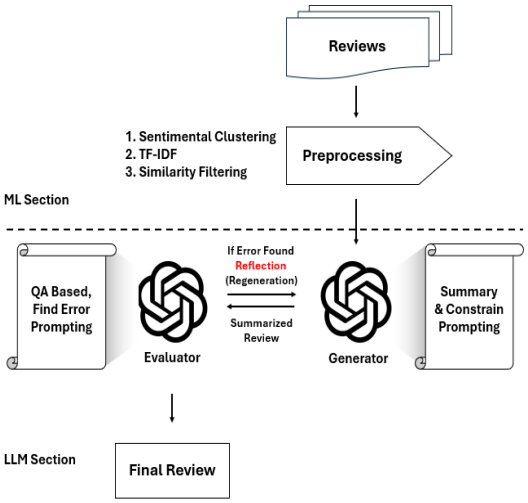

그림 1 Workflow Diagram

가. 분류 단계

분류 단계에서는 사전 훈련된 DistillBERT 를 활용한 감정 분류 모듈을 통해 리뷰 데이터셋을 긍정/부정 리뷰들로 분류하여 그룹화를 수행한다.

나. 추출 단계

추출 스테이지는 앞서 분류된 긍/부정 그룹에서 중요한 정보 추출을 수행한다. 각 그룹에서 TF-IDF 기법을 통해 주요 키워드를 추출하고, 해당 단어가 포함된 문장들을 선별하여 유사 문장 클러스터링을 진행한 후 각 중요 문장을 추출한다.

다. 생성 단계

추출 단계에서 추출한 그룹별 중요 문장들을 생성자에게 전달하여 자연스러운 요약문으로 생성하고 평가자에게 전달된다. 생성자는 GPT-4 모델을 사용하며, 제약조건이 주어진 프롬프트로 구성된다.

라. 평가 단계

평가 단계에서는 전달된 요약문을 평가자가 QA 기반 평가를 통해 환각 여부와 정보 누락 여부를 측정한다. 정밀도(Precision) 관점에서, 평가자는 요약문에서 뽑아낸 질의를 원문으로 검증하여, 요약문에 포함한 정보가 원문에서 사실적으로 근거하는지를 평가한다. 또한 재현율(Recall) 관점에서, 원문에서 생성된 질의를 요약문으로 검증하여, 원문에 포함된 필수 정보가 요약문에 누락되지는 않았는지를 평가한다. 이러한 QA 질의를 통해 MDS 의 원문과 요약문의 비교 평가를 간접적으로 수행하여 토큰 제약 문제를 해결하며, 평가자의 평가 결과에 따라 문제가 없다면 긍정, 부정 요약문을 각각 출력한다. 반면 평가자가 문제점을 확인한다면 이를 feedback 으로 제공하여 다시 생성자에게 재생성을 요청한다.

Ⅳ. 실 험

가. 데이터셋

본 연구에서는 리뷰와 관련된 다중 문서 후기 요약 데이터로 Amasum 벤치마크 데이터셋[7]과 Space 데이터셋[8]을 활용하였다. 해당 데이터셋들은 각 상품에 대해 주석자들이 작성한 정답 요약문과, 원문 리뷰 등으로 구성되어 있으며, Amasum 데이터의 정답 요약문은 Verdict(총평 요약), Pros(장점 요약), Cons(단점 요약)으로 개별화 되어있다.

나. 실험 구성

본 연구의 baseline 으로는 감정 분류와 평가 스테이지가 포함되지 않는 기존 연구의 하이브리드 방법론 구조를 차용하되, 기존 연구에서 사용한 BART 혹은 Transfomer 모델이 아닌 Openai 의 GPT-4 모델을 생성 요약 모델로 변경하였다. 이를 통해 baseline 과 제안 기법의 생성 모델을 같게 함으로써 생성단계에서 단순 LLM 의 추론 능력 차이에 따른 성능 차이를 배제하였다.

제안 기법 또한 baseline 과 마찬가지로 GPT-4 를 요약 모델로 사용하였으며, 평가 단계에서도 동일한 모델을 사용하였다. 평가 단계의 경우 최대 3 회까지 재생성이 가능하도록 제한하고 성능을 평가하였다.

다. 요약 성능 평가

공정한 비교를 위해 정답 요약을 Verdict·Pros·Cons 를 순서대로 연결한 단일 참조로 구성하였다. Baseline 의 단일 요약과, 본 연구 기법의 장점 요약+단점 요약을 연결한 요약을 각각 동일 참조에 대해 ROUGE 스코어로 평가·비교하였다. 각 데이터들은 Amasum 평가 데이터셋에서 50 개를 랜덤 샘플링하여 진행하였다.

.

	Rouge-1	Rouge-2	Rouge-L
Baseline	23.55	2.48	12.83
Ours	**27.30**	**3.17**	**13.74**

표 1 Amasum 데이터셋 요약 성능 평가 결과

Baseline 과의 비교에서, 모든 스코어에서 성능 향상이 발생하였다. 추상 요약의 특성상 빅그램(n-gram)에서의 성능이 낮게 측정되어 Rouge-2 에서는 큰 향상을 보이진 못했지만, Uni-gram 기반의 Rouge-1 에서는 두드러진 성능 향상이 나타났다. 이를 통해 기존 Baseline 보다 본 연구 기법이 더 개선된 요약을 생성함을 확인하였다.

라. 성분 분석 실험(Ablation Study)

QA 기반 피드백 평가 단계가 실제 요약 품질 향상에 기여하는지를 검증하고자, 감성 분류 단계만 제외하고 평가 단계는 유지한 변형 구조와, 평가 단계까지 모두 제거한 변형 구조 두가지를 구성하여 Space 데이터셋으로 실험을 진행하였다.

	Rogue-1	Rogue-2	Rogue-L
Non-Evaluator	32.73	7.62	20.6
Evaluator	**37.22**	**8.82**	**22.01**

표 2 평가 단계 효용성 검증 결과

실험결과, 평가 단계를 붙인 변형 구조의 결과가 더 성능이 향상되었으며, 본 기법에서 제안된 방식의 효과성을 입증하였다.

Ⅴ. 결론

본 연구에서는 기존 하이브리드 방법론의 한계점을 지적하고, 이를 해결하기 위해 긍정, 부정 그룹별 요약을 위한 감성 분류 모듈과 QA 기반의 피드백을 수행하는 다중 에이전트 구조를 활용하여 다중 문서 요약을 수행하였다. 이러한 기법을 통하여 기존 방법에서 발생하는 환각과 정보 누락 문제들을 개선할 수 있음을 실험을 통해 입증하였으며, 사용자의 선호도를 반영한 요약 결과에서도 더 성능이 향상됨을 확인하였다. 이러한 개선된 기법을 통하여, 리뷰 시스템이 적용되는 여러 분야에 개선된 리뷰 요약을 제공할 수 있을 것을 기대하며, 그동안 단일한 요약만을 제공해왔던 것을 넘어 사용자들에게 다양한 선호도를 대표하는 요약을 제공하여 리뷰 서비스에 있어 사용자 경험(UX)도 향상시킬 수 있을 것이라 기대한다. 향후 연구에서는 정보 누락과 환각을 식별·검증하기 어려운 ROUGE 스코어의 한계를 보완하기 위하여, QA 기반 다중 문서 요약 벤치마크 데이터셋 연구를 진행할 예정이다.

참 고 문 헌

[1] 이필원, 황윤영, 최종석, 신용태, "워드 임베딩 클러스터링을 활용한 리뷰 다중문서 요약 기법". 정보처리학회 논문지, 제 10 권, 제 11 호(통권 110 호), pp. 535-540, 2021 년.

[2] Bhaskar, A., Fabbri, A. R., and Durrett, G., "Prompted Opinion Summarization with GPT-3.5," in *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9282- 9300, July 2023

[3] Pu, X., Gao, M., and Wan, X., "Summarization is (Almost) Dead," arXiv preprint arXiv:2309.09558, Sep. 2023.

[4] C. G. Belém, P. Pezeshkpour, H. Iso, S. Maekawa, N. Bhutani, and E. Hruschka, “How LLMs Hallucinate in Multi-Document Summarization,” in Findings of the Association for Computational Linguistics: NAACL 2025, 2025, doi:10.18653/v1/2025.findings-naacl.293. ACL Anthology

[5] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language Agents with Verbal Reinforcement Learning,” in Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2023. NeurIPS Papers

[6] H. Zhang, X. Liu, and J. Zhang, “SummIt: Iterative Text Summarization via ChatGPT,” in Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 10644- 10657, 2023, doi:10.18653/v1/2023.findings-emnlp.714. ACL Anthology+ 1

[7] Braž inskas, A., Lapata, M., and Titov, I., “Learning Opinion Summarizers by Selecting Informative Reviews,” arXiv preprint arXiv:2109.04325, Sep. 2021.

[8] Angelidis, S., Amplayo, R. K., Suhara, Y., Wang, X., and Lapata, M., “Extractive Opinion Summarization in Quantized Transformer Spaces,” Trans. of the Assoc. for Computational Linguistics (TACL), 2021.