대규모 언어 모델을 활용한 어휘력 맞춤형 뉴스 요약 플랫폼 개발

이민지, 우은식, 김영권, 채성수, 황경호*

국립한밭대학교 컴퓨터공학과

{ minjilee, EunsikWoo, youngkwonKim ,seongsu }@edu.hanbat.ac.kr

*gabriel@hanbat.ac.kr

Development of a Vocabulary-Adaptive News Summarization Platform Using Large Language Models

Min-Ji Lee, Eun-Sik Woo, Seong-Soo Chae,
Young-kwon Kim, *Gyung-Ho Hwang

Dept. Computer Engineering, Hanbat National University

요 약

정보 과잉의 시대에서 뉴스 소비자의 비판적 사고력과 이해도가 중요해지고 있으나, 청년층을 중심으로 뉴스 이용과 문해력은 꾸준히 하락하고 있다. 본 연구는 이러한 문제를 완화하기 위하여 대규모 언어 모델(LLM)을 활용한 뉴스 텍스트 기반 요약, 긍·부정 코멘트 통합 데이터셋을 구축하고 이를 활용한 뉴스 요약 플랫폼을 제안한다. LLM을 활용하여 요약문, 코멘트, 유·반의어 및 예문을 사전에 생성하고, 의미 보존과 중립성 검수를 통하여 품질을 확보한다. 본 연구에서 구축한 데이터셋은 한국어 문서 요약, 언어 교육 등 여러 분야에서 재사용이 가능하며, 특히 사용자의 이해도와 언어 능력 향상을 지원하는 응용 서비스의 핵심 자료로 활용될 수 있음을 시사한다.

Ⅰ. 서 론

최근 각종 디지털 플랫폼의 확산으로 정보가 과잉 생산·유통되면서 뉴스 소비자의 비판적 사고와 문해력은 사회를 이해하는 핵심 역량으로서 중요해졌다. 그러나 정확하고 다양한 뉴스를 꾸준히 접하지 못하면 이러한 역량을 유지하기 어렵다. 한국언론진흥재단 '2024 언론수용자 조사'에 따르면, 20대의 인터넷 기반 뉴스 이용률을 2021년 대비 13.3% 감소했으며, 같은 기간 온라인 동영상 플랫폼 뉴스 이용률도 9.4% 감소하였다[1]. 또한, OECD 국제성인역량조사(PIAAC) 우리나라 성인의 문해력 점수는 평균 249점으로 OECD 평균 260점보다 11점 낮다[2]. 이처럼 뉴스 이용 감소와 낮은 문해력 수준은 정치·세대·성별 갈등과 같은 민감한 사회 이슈에서 편향된 정보 소비를 유발하여, 사회적 분열을 가속화하는 요인으로 작용할 수 있다.

기존 뉴스 서비스는 이용자의 수준과 배경을 충분히 고려하지 못해, 개인별 이해도 차이를 완화할 수 있는 장치가 부족하다. 이에 따라 복잡한 사회적 문제를 단편적으로만 이해하게 되고, 균형 잡힌 뉴스 이해가 어려워지는 구조적 한계가 있다.

이에 본 연구는 사용자 수준에 맞춘 맞춤형 뉴스 요약 플랫폼인 "News Fit"을 제안한다. 본 플랫폼은 뉴스 기사의 핵심 내용을 사용자의 문해력 수준에 맞게 요약 및 제공함으로써, 정보 과잉 시대에 비판적 사고를 촉진하고 사회적 갈등 완화에 기여할 수 있는 가능성을 탐구한다.

Ⅱ. 본론

본 연구에서 제안하는 플랫폼은 네 가지 주요 기능으로 구성된다.

- 사용자 어휘력 수준 평가
- 어휘력 수준에 따른 요약 방식 분기(추상/추출)
- 뉴스 원문에 대한 긍·부정 코멘트 제공
- 어휘력 수준에 따른 단어 해석 및 예문 제공
- 어휘력 수준에 따른 유·반의어 제공

우선, 사용자 어휘력 수준을 진단해 언어 능력을 평가한다. 그 결과를 기반으로, 추상적 요약(Abstract)과 추출적 요약(Extractive) 중 적합한 요약문을 제공한다. 또한 뉴스 원문에 대한 긍·부정 코멘트를 함께 제공하여 원문을 다각도로 이해할 수 있게 한다. 마지막으로, 개인의 어휘 수준에 맞게 유·반의어와 예문을 제시하여 학습과 이해를 자연스럽게 연결하며 문해력 향상에 기여한다.

사전에 뉴스 원문에 대한 요약문, 긍·부정 코멘트, 유·반의어, 단어 및 예문 네 가지 주제에 대한 데이터셋을 구축한다. 데이터셋은 경제, 국제, 사회, 정치, 과학·IT의 다섯 가지 카테고리에서 수집한 총 26개의 뉴스 기사를 원문 데이터로써 활용한다. 이를 기반으로 52개의 긍·부정 코멘트와 156개의 단어 데이터를 생성한다.

```
당신은 텍스트 요약 전문 AI 어시스턴트입니다. 우리는 {data['content']}를 바탕으로 요약문을 생성하여야합니다.
해당 본문의 주요한 단어는 {data['feature']}입니다. 아래의 규칙을 따라주세요.

[본문]
{data['content']}

[주요 단어]
{data['feature']}

1. 요약문은 정확히 70-span으로 생성해주세요.
2. 주어진 <content>와 <feature>의 내용을 충분히 반영하도록 합니다.
3. 편향이란 부정적, 긍정적, 정치적 등 한 쪽으로 의견이 몰린 현상을 의미합니다. 요약문이 앞서 정의한 편향 현
   상이 일어나지 않게 객관적인 요약을 생성합니다.
4. <level>이 "Low"일 경우 해당 요약을 읽을 사용자는 낮은 어휘력을 갖고 있습니다. 추상 요약을 진행하세요.
   어려운 단어는쉬운 단어로 변경하여 생성합니다.
5. <level>이 "High"일 경우 해당 요약을 읽을 사용자는 높은 어휘력을 갖고 있습니다. 추출 요약을 진행하세요.
   어려운 단어는쉬운 단어로 변경하여 생성합니다.
```

그림 1 요약 프롬프트 예시

데이터 생성 시, GPT-4o 모델을 사용하였으며, 요약문과 코멘트 길이를 각각 70-span, 50-span 이내로 생성하도록 프롬프트를 설계한다. [그림 1]은 요약 프롬프트 예시이다.

생성한 데이터는 요약, 코멘트, 단어 세 가지 유형에 대하여 검수를 진행한다. 요약 데이터의 경우 원문과 의미적 등가성을 유지하는지, 길이가 기준을 충족하는지, 그리고 특정 이념이나 가치관에 치우치지 않는 중립성을 지니는지 검수한다. 코멘트 데이터는 원문과 일관성을 유지하면서도 필요한 이념•편향이 삽입되지 않았는지 검수한다. 단어 데이터는 단어의 사전적 정의와 일치하는지 검수한다.

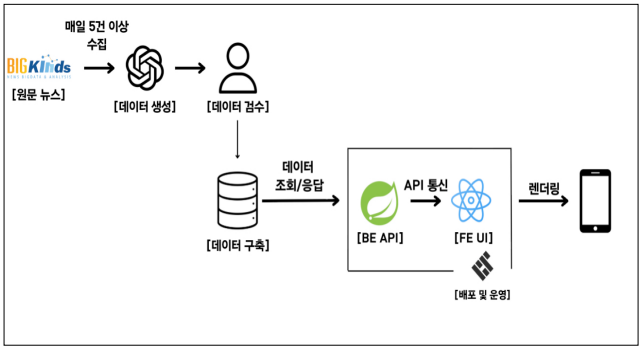


그림 2 시스템 구성도

[그림2]는 본 시스템의 구성을 나타낸다. 빅카인즈로부터 매일 5건 이상 수집한 뉴스를 기반으로 GPT-4o를 사용하여 데이터를 생성하고, 검수를 통하여 데이터베이스에 저장한다. 백엔드 API 서버는 Spring Boot 기반으로 구축되었으며, 사용자 인터페이스(UI)는 React 프레임워크를 통해 제공한다. 전체 시스템의 배포 및 운영은 Cloudflare를 기반으로 수행한다.

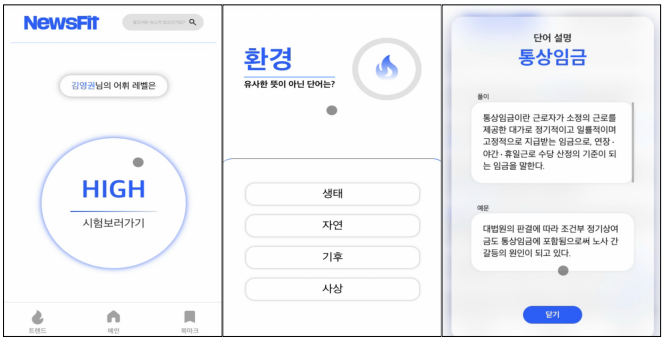


그림 3 어휘력 테스트 / 단어 정의 및 예문

[그림 3]은 사용자의 어휘력 수준을 측정하고 단어 정의 예문을 제공하는 화면이다. 어휘력 검증은 사용자가 특정 단어와 그 의미를 올바르게 매칭하는 형태로 진행하며, 총 10문항으로 구성한다. 임계값(threshold)은 5문항 이상 정답을 맞추는 것으로 설정한다. 검증 종료 이후, 해당 검증 결과를 사용자에게 Low, High 두 단계로 시각화하여 제시한다.

이후 제공되는 단어 및 예문은 단순 빈도 기반이 아닌, 뉴스 도메인 내 문맥 다양도•교육용 난이도•사용자 어휘력을 복합적으로 고려하여 선정한다. 예를 들어, 일상적으로 사용되는 단어는 Low 등급의 단어로 선정하고, 추상적 개념을 지칭하는 단어는 High 등급의 단어로 선정한다.

또한, 각 어휘는 유의어•반의어 및 예문 데이터를 함께 포함하도록 구축한다. 이는 사용자가 요약문 내에서 생소한 어휘를 접했을 때, 해당 단어의 의미를 맥락적으로 이해하고 어휘 확장을 유도하기 위함이다. 예를 들어, “기준금리”의 경우 “정책금리, 기준이율” 등 유의어를 함께 제시하고, “한국은행은 기준금리를 동결했다.”와 같은 예시 문장을 제공함으로써 학습자의 어휘 인지도를 강화한다. 이러한 구조는 본 시스템의 핵심 데이터셋으로 사용되며, [그림 3]의 어휘력 레벨별 어휘 차별화 제공 기능에 직접적으로 활용된다. 이후, 시스템은 설정된 어휘력 등급에 따라 뉴스 텍스트 원문에서 특정 단어 선택하면, 해당 단어의 정의와 실제 예문을 함께 제공하여 반복 학습과 문해력 향상을 지원한다.

[Level: High]
1분기 한국 경제는 -0.2% 성장으로 19개국 중 최하위였습니다. 정치적 불안과 미·중 교역 긴장, 관세 변수로 경기 둔화가 심화되며, 한국은행은 금리 인하를 검토 중입니다.

[Level: Low]
우리나라는 1분기에 -0.2% 성장해 19개 나라 중 꼴찌였습니다. 정치 혼란과 미국과의 무역 문제가 원인으로, 한국은행이 금리를 내릴 수도 있습니다.

그림 4 요약 레벨별 비교 예시

[그림4]는 한 뉴스 기사에 대하여 생성한 요약문의 예시이다. 2025년의 한국 경제성장에 관한 뉴스 요약문이며 각 어휘력 레벨에 따라, 사용 어휘가 상이함을 확인할 수 있다.

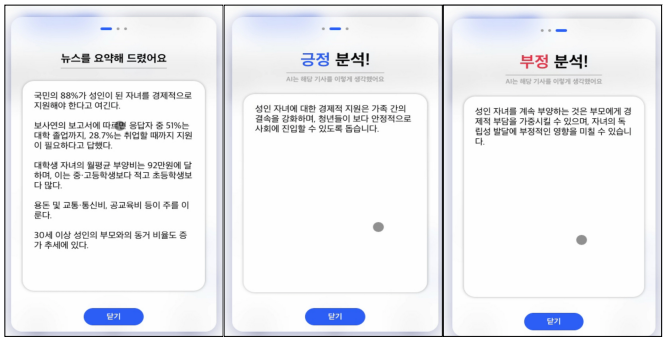


그림 5 뉴스 요약문 / 긍·부정 코멘트

[그림5]는 한 뉴스 기사에 대하여 생성한 요약문과 긍·부정 코멘트 제공하는 화면이다. 사용자는 요약문을 통해 핵심 정보를 빠르게 파악하고, 긍·부정 코멘트를 통해 원문에 대한 다양한 시각을 균형 있게 접함으로써 비판적 사고를 키울 수 있다.

.

Ⅲ. 결론

본 연구에서는 대규모 언어 모델을 기반으로 한 한국어 뉴스 요약 플랫폼을 설계 및 구현하였다. 제안한 시스템은 사용자의 어휘력 수준에 따라 맞춤형 요약을 제공하고, 긍·부정 코멘트 및 단어 해설을 통하여 중립적 가치관 형성과 문해력 향상에 기여할 수 있음을 시사한다. 향후 연구에서는 실제 사용자 집단을 대상으로 한 실험을 통하여 시스템의 효과를 정량적으로 평가하고, 보다 대규모의 뉴스 데이터셋을 구축할 예정이며, 교육 분야에서의 응용 가능성을 검토함으로써 연구의 실효성을 더욱 강화할 예정이다.

ACKNOWLEDGMENT

“본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음”(2022-0-01068)

참 고 문 헌

[1] 한국언론진흥재단, 「2024 언론수용자 조사」, 2024.

[2] OECD, *Programme for the International Assessment of Adult Competencies (PIAAC)*, 2023.