의미 기반 하이브리드 접근을 통한 뉴스 3줄 요약 시스템
A Semantic Hybrid Approach for Three-Line News Summarization

함정민*, 김연태*, 김지섭*, 정현민*, 전태수**
Jeongmin Ham*, Yeontae Kim*, Jiseop Kim*, Hyunmin Jeong* and Taesoo Jun**
* 금오공과대학교

***Abstract*—온라인 뉴스의 폭발적 증가로 사용자가 핵심 정보를 신속히 파악하기 어려워졌다. 기존 추출형 요약은 원문 문장 선택의 안정성을 가지나 표현 다양성에 한계가 있고, 생성형 요약은 유창성은 우수하나 연산 비용과 품질 불안정 문제가 공존한다. 본 논문은 두 접근법의 장점을 결합한 의미 기반 하이브리드 뉴스 3줄 요약 시스템을 제안한다. [1]제안 시스템은 SBERT 임베딩과 문장 유사도 그래프 기반 커뮤니티 탐지로 입력 기사를 의미 단위로 분할하고, 각 의미 영역에서 KoBART를 이용해 생성형 요약을 수행한다. 생성형 요약이 비정상일 경우 TextRank 추출형 요약으로 자동 대체하는 안정성 강화로직을 포함하여, 최종적으로 핵심 1·2·3형식의 3문장 요약을 제공한다. 구현 관점과 사례 중심으로 시스템을 상세 기술하며, 실제 서비스 적용을 위한 설계적 고려사항을 정리한다.**

***Index Terms*—3줄 요약, KoBART , SBERT , TextRank, 하이브리드 요약, 의미 클러스터링**

I. 서론

뉴스 소비 환경은 초과잉 정보의 시대에 접어들었다. 사용자는 다수의 플랫폼에서 방대한 기사를 접하지만, 탐색 비용과 인지 부하가 커져 핵심 파악이 어렵다. 자동 요약 기술은 이러한 부담을 완화하지만, 추출형은 표현 유연성이 낮고, 생성형은 불안정성이 문제로 지적되어 왔다. 본 연구는 두 방식의 장점을 접목해, 의미 기반 입력 분할과 생성형 요약 + 추출형 대체를 결합한 하이브리드 3줄 요약을 제안한다.

- **의미 단위 분할**: SBERT 임베딩과 문장 유사도 그래프 기반 커뮤니티 탐지로, 기사 내 상이한 주제 축을 자동 분리
- **생성형-추출형 하이브리드**:[2][3] KoBART 생성형 요약을 기본으로, 실패/비정상 출력 시 TextRank 추출형으로 보완
- **3줄 구조화 출력**: 사용자 경험(UX)에 최적화된 핵심 1 · 2 · 3 형식(세 줄)으로 정보 계층화

II. 뉴스 3줄 요약 시스템

파이프라인은 (1) 전처리 & 문장화, (2) 의미 임베딩, (3) 의미 커뮤니티 탐지, (4) 의미 단위 요약(생성형 우선), (5) 추출형 대체, (6) 3줄 형식 통합으로 구성된다.

A. 전처리 및 문장 분리

기사에서 메타(기자명, 이메일, 매체 표기) 제거, 괄호 내 소스 코드/주석 제거, 개행/공백 정규화 등을 수행한다. 한글 문장 분리는 [4]KSS를 사용하며, 비정상 종결(.?"!″) 보정을 통해 문장을 안정화한다.

B. 의미 임베딩과 그래프 구성

[5]각 문장을 SBERT(KR-SBERT-V40K-klueNLI-augSTS)로 임베딩하고, 코사인 유사도로 문장 간 가중 그래프를 형성한다.[4]문장 유사도 그래프의 커뮤니티 탐지로 의미 축을 도출한다. 이는 뉴스의 주제 전환을 잘 반영하며, 군집 수 자동 탐지 후 상위 3개 커뮤니티를 선택하여 3줄 요약의 의미 캡슐로 사용한다.

C. 의미 단위 입력 생성

각 커뮤니티 내 중심성(centroid 유사도) 상위 문장을 대표 시드로 고르고, 정보 다양성을 위해 MMR을 적용해 중복을 억제한다. 필요 시 2–4문장을 연결해 의미 문단을 구성한다(길이 상한 관리).

D. 생성형 요약(기본 경로)

의미 문단을 KoBART에 전달하여 각각 1문장 생성형 요약을 수행한다. 빔 서치(beam search)를 4개의 후보 빔으로 수행하여 다양한 생성 가능성을 탐색하였으며, 동일한 구문 반복을 방지하기 위해 3-그램 단위 중복 제한을 적용하였다. 또한 문장의 비정상적 반복이나 과도한 공백 생성을 줄이기 위해 반복 패널티를 1.5로 설정하고, 최소 15단어에서 최대 128단어 사이의 길이 제약을 두어 출력의 안정성과 가독성을 높였다. 내부적으로는 의미 캡슐 단위 병렬 처리가 가능하다.

E. 추출형 대체(Fallback)

다음 조건 중 하나라도 만족하면 해당 의미 축은 TextRank 추출형 요약으로 대체한다: (i) 생성 실패/예외, (ii) 생성 길이 부족, (iii) 비문/반복률 과다.[6]TextRank는 Komoran 기반 토큰화로 단어 그래프를 구축하고, PageRank로 문장 중요도를 산정하여 최상위 문장을 채택한다. 이로써 *서비스 안정성*을 확보한다.

F. 3줄 통합 규칙

최종 출력은 핵심 1 · 2 · 3 형식으로 정렬한다. 기본은 (i) 사건/결정, (ii) 원인/영향, (iii) 전망/반응 순이며, 의미 축 라벨링(사전정의 키워드/간단한 규칙 기반)로 자연스러운 순서를 보장한다.

III. 시스템 구현

A. 개발 환경

Python 3.10, PyTorch, HuggingFace Transformers, SentenceTransformers, KSS, scikit-learn을 사용하였다. 서버는 CUDA 지원 환경에서 KoBART 추론을 수행하고, 나머지 전처리/그래프/추출형 로직은 CPU에서도 충분히 동작한다.

* 국립금오공과대학교 컴퓨터소프트웨어공학과, gkaejrdbs@kumoh.ac.kr, kyt0060@naver.com, kimzito@naver.com, hm990724@nate.com, ,
** 국립금오공과대학교 컴퓨터소프트웨어공학과, taesoo.jun@kumoh.ac.kr.

B. 모듈 구성

- **Preprocessor**: 규칙 기반 정규화, 문장 분리/보정
- **Embedder**: SBERT 임베딩, 유사도 계산
- **CommunityDetector**: 유사도 그래프 생성, Louvain 커뮤니티 탐지
- **CapsuleBuilder**: 중심성+MMR로 의미 문단 생성
- **Abstracter**: KoBART 생성형 요약
- **Extractor**: TextRank 추출형 대체
- **Integrator**: 3줄 형식 정렬/출력

C. 엔드투엔드 의사코드

Algorithm 1 Semantic Hybrid Three-Line Summarization

Require: Article D
S KSS_split(D)
E SBERT_encode(S)
G build_graph(S, E)
$\{C_1, C_2, \dots\}$ Louvain(G)
$\{C'_1, C'_2, C'_3\}$ top3_by_size_or_salience($\{C\}$)
for $i \in \{1\}$ **do**
  P_i build_paragraph(C'_i; centroid+MMR)
  y_i KoBART(P_i)
  if invalid(y_i) **then**
    y_i TextRank(C'_i)
  end if
end for
return order_as **핵심 1, 2, 3**: $\{y_1, y_2, y_3\}$

D. 서비스/UX 고려사항

요약 클릭 시 근거 보기로 각 핵심이 근거한 원문 문장 하이라이트를 제공하면 신뢰성이 향상된다. 또한 문서 길이와 평균 문장 수에 따라 커뮤니티 수를 2–4로 *동적 조절*하는 옵션을 두되, 사용자 인터페이스는 항상 3줄을 유지하여 일관된 경험을 제공한다.

IV. 사례 기반 적용 결과

정량 평가 대신, 실제 뉴스 도메인에 적용했을 때의 *동작 특성*을 사례로 정리한다.

- **다주제 기사**: 정책 발표(사건)–이유/배경(원인)–시장/여론(반응)의 세 축으로 자연 분할되어 3줄 구성이 명확해진다.
- **단일 주제+부연 기사**: 커뮤니티가 2개 이하로 형성될 수 있으며, 본 시스템은 *보조 세부*를 병합하여 3줄 형식을 유지한다.
- **인용이 많은 기사**: 전처리 단계의 인용 보호/복원과 커뮤니티 단위 요약으로 과도한 인용에 휩쓸리지 않고 *사실 축*을 보존한다.
- **길이 상이 기사**: KoBART 입력 길이 제어와 추출형 대체로 비정상 출력을 억제하여 *서비스 신뢰성*을 유지한다.
- **대화체(문답형) 기사**: 기자–인터뷰이 간 문답이 중심인 경우, 발화 단위가 짧고 다중 화자 맥락이 교차되어 본 시스템은 기본적으로 3줄 요약을 지원하지 않는다.

V. 한계와 고찰

커뮤니티 탐지의 품질은 임베딩 품질과 유사도 임계값, 그래프 희소성에 민감하다. 특정 도메인(예: 스포츠 실황, 연예 루머)에서는 사실성 대비 감성 표현이 과대 대표되어 요약 문체가 부정확해질 수 있다. 또한 KoBART의 생성 품질은 입력 문단의 길이/밀도에 따라 편차가 존재한다. 이를 완화하기 위해 (i) 의미 문단 길이 상한, (ii) 반복/비문 탐지 기반 재시도, (iii) 커뮤니티 수 동적 결정 후 3줄 표준화(병합/축약)를 적용하였다.

VI. 결론 및 향후 연구

본 논문은 SBERT 기반 의미 그래프와 커뮤니티 탐지로 기사 내 의미 축을 분리하고, 각 축에서 KoBART 생성형 요약을 수행하되 TextRank로 보완하는 **의미 기반 하이브리드 3줄 요약**을 제안하였다. 결과는 핵심 $1 \cdot 2 \cdot 3$ 형식으로 구조화되어 사용자 이해를 가속한다.

향후에는 프롬프트 기반 대형 언어모델(LLM)을 통합하여, 의미 축 라벨링과 문장 간 논리 전이를 지시(prompt)로 제어하는 방식을 연구할 계획이다. 예컨대, "사실–원인–전망" 프롬프트 템플릿과 출력 제약(3문장 · 각 25자 내)을 결합하면, 생성 품질과 일관성을 동시에 제고할 수 있다.

REFERENCES

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, P10008, 2008.

[2] Se U. Jung, "KoBART-summarization," GitHub repository. https://github.com/seujung/KoBART-summarization.

[3] Changhyun Cho, "TextRank 알고리즘 (문서 요약 및 키워드 추출)," *Excelsior 블로그* https://excelsior-cjh.tistory.com/93.

[4] J. Lee, "KSS: Korean Sentence Splitter," GitHub repository, 2020. https://github.com/hyunwoongko/kss

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of EMNLP*, 2019.

[6] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. of SIGIR*, 1998.