JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 8, pp. 2165-2174, Aug. 2024

잡음 정제를 통한 고응집도 군집과 후반부 요약을 이용한 효과적인 뉴스 요약

한 성 민[1] · 백 대 환[1] · 이 현 아[2*]
[1]국립금오공과대학교 컴퓨터소프트웨어공학과 학사과정
[2]국립금오공과대학교 컴퓨터소프트웨어공학과 교수

Effective News Summarization through High Cohesion Clustering with Noise Filtering and Second-Half Summary

Seongmin Han[1] · Daehwan Baek[1] · Hyunah Lee[2*]

[1]Bachelor's Course, Department of Computer Software Engineering, Kumoh National Institute of Technology, Gyeongbuk 39177, Korea
[2]Professor, Department of Computer Software Engineering, Kumoh National Institute of Technology, Gyeongbuk 39177, Korea

[요 약]

뉴스에서는 다양한 언론사가 동일한 사건을 다루는 기사를 중복으로 발행하므로 정보의 중복성이 크다. 유사한 기사를 군집화하고 그 내용을 적절히 요약하여 제공하면 사용자는 뉴스 동향을 빠르게 파악하고 자세하게 읽어야 하는 기사를 쉽게 선택할 수 있다. 본 논문에서는 효과적인 뉴스 요약을 제공하기 위해 HDBSCAN으로 최초 군집을 생성한 뒤에 Mean-Shift와 노이즈 정제를 적용하여 응집도가 높은 군집을 얻는 방법을 제안한다. 요약 생성에서는 군집 내 기사의 후반부에서 얻은 요약 중에서 군집을 대표할 수 있는 요약을 요약 평가 메트릭에 기반하여 선택하고 이를 기사의 리드 또는 전반부 요약과 결합하여 제공하는 방식을 제안한다. 실험 결과 Mean-Shift 적용과 노이즈 제거를 이용한 군집화와 후반부 요약을 추가한 요약 생성 모두에서 성능 향상을 얻어 뉴스 요약에서의 효율성을 확인했다.

[Abstract]

News articles often present high redundancy due to multiple outlets covering the same events. Clustering similar articles and summarizing them allows users to quickly understand trends and focus on articles requiring detailed reading. This paper proposes a method that begins with HDBSCAN clustering and applies Mean-Shift and noise refinement to enhance cluster cohesion. For summary generation, we introduce a strategy that selects representative summaries from the latter half of articles within clusters, evaluated using a summary metric. These are combined with lead summaries from the articles' first-halves. Experiment results demonstrate efficiency gains in news summarization with improved clustering performance and summary quality through the inclusion of second-half summaries.

색인어 : 뉴스 요약, 군집 정제, 후반부 요약, 리드, 자연어처리.

Keyword : News Summarization, Cluster Refinement, Second-Half Summary, Lead, Natural Language Processing (NLP)

http://dx.doi.org/10.9728/dcs.2024.25.8.2165

Received 14 July 2024; **Revised** 06 August 2024
Accepted 20 August 2024

***Corresponding Author; Hyunah Lee**

Tel: +82-54-478-7546
E-mail: halee@kumoh.ac.kr

Ⅰ. 서 론

자연어는 정보 생성과 교환에서의 가장 직관적인 소통의 도구로 활용되어 대부분 정보는 자연어를 통해 표현된다고 해도 무방하다. 이러한 자연어 텍스트로 구성된 뉴스 기사는 정보 전달에 가장 적합하도록 오랜 역사를 통해 발전해 왔으며, 인터넷의 발달 이후로는 양적으로도 비약적으로 성장했다. 한국언론진흥재단의 신문산업 실태조사[1]에 따르면, 2010년에는 1,126개의 인터넷신문 사업체가 있었으나 2023년에는 약 4배 늘어난 4,322개의 인터넷신문 사업체가 기사를 생산하고 있다. 인터넷 언론의 활성화로 뉴스에 대한 접근성이 좋아진 장점에 비해, 각기 다른 언론사가 같은 사건을 중복적으로 다루면서 뉴스 기사가 과다 생산되어 독자는 도리어 짧은 시간 안에 양질의 정보를 얻기 어려워졌다.

정보의 중복성이 높다는 점에서 뉴스 기사는 군집화(clustering)와 요약(summarization)에 대한 사용자 요구가 크다. 군집화를 통해 같은 사건을 다루는 기사들을 묶어주고 각 군집에 대해 효과적인 요약을 제시한다면, 독자들은 신속하게 최신 동향을 파악하고 그중 본인에게 필요한 기사를 선택하여 내용을 확인할 수 있다. 이러한 목적으로 뉴스를 대상으로 한 문서 군집화와 군집 내 기사들의 요약에 관한 다양한 연구가 시도되고 있으나, 기존의 연구들은 해당 연구에 적합하게 미리 정제된 군집만을 대상으로 하거나 기사 문장 중 일부분만 선택하는 추출 요약에 기반하는 등의 한계가 있다.

본 연구에서는 정제되지 않은 뉴스 기사들을 대상으로 문서를 군집화하고 군집의 대표 요약문을 생성하는 방법을 제안한다. 군집화에서는 HDBSCAN(Hierachical Density-Based Spartial Clustering of Applications with Noise)[2]으로 최초 군집을 얻은 뒤 Mean-Shift와 주제어 기반 군집 정제를 적용하여 군집과 요약 생성에 악영향을 미치는 노이즈 문서를 제거한다. 요약에서는 기존의 기사 요약이 제목과 기사 전반부 위주로 결과를 생성하여 충실한 정보를 제공하지 못했던 문제를 해결하기 위해 문장 후반부에 대한 요약을 추가하는 방식을 제안한다. 이 과정에서 딥러닝 기반 다양한 도구들을 활용하여 성능을 향상하고자 한다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보고, 3, 4, 5장에서는 본 연구의 제안 방식을 설명한다. 6장에서는 연구의 분석과 결과를 보이고 7장에서는 본 연구의 결론을 보인다.

Ⅱ. 관련 연구

2-1 문서 군집화

문서 군집화는 비슷한 특성을 갖는 문서들을 그룹화하는 자연어처리 기법으로 K-means부터 LDA(Latent Dirichlet Allocation)까지의 다양한 군집화 알고리즘이 적용되고 있다. [3]에서는 다양한 군집화 알고리즘의 결과 분석에서 군집의 개수를 사전에 지정하지 않는 HDBSCAN[2]이 세부적인 토픽 식별이 가능한 장점이 있음을 밝혔다.

HDBSCAN은 추출할 군집의 개수를 사전에 지정하지 않는 비모수 방식이면서 다양한 밀도 수준을 가진 문서 군집을 동시에 감지하는데 최적화된 계층적 군집화 알고리즘이다. HDBSCAN은 군집에 이웃을 포함할 최대 반지름 거리 ε을 변화시키며 계층적 군집을 식별한다. ε이 감소할수록 군집의 크기가 점점 작아지며 사라지거나 하위 군집으로 분할되는 전통적인 계층적 군집화의 모습을 보인다. 따라서 계층적으로 식별한 군집은 ε의 변화에 따라 상위 군집과 하위 군집의 관계성이 나타나는 트리로 표현된다. HDBSCAN은 트리에서 군집에 속한 각 객체의 밀도 수준을 고려하여, 군집이 존재할 수 있는 최소 ε와 최대 ε의 차이가 큰 군집을 선택한다.

문서 군집화에서 군집화 알고리즘과 더불어 성능에 큰 영향을 미치는 요소는 문서를 정형 데이터로 변환하는 방법이다. 비정형 데이터인 자연어는 군집을 생성하는데 직접적으로 사용할 수 없으므로, 자연어처리에서는 텍스트 문서를 정형 데이터인 벡터로 변환하여 사용한다. HDBSCAN에 기반하여 최근 좋은 성능을 보이고 있는 BERTopic[4]에서는 사전학습 언어모델인 SBERT(Sentence Bidirectional Encoder Representations from Transformers)[5]를 사용하여 문서를 벡터로 임베딩한다. 또한 이에 기반하여 문서를 군집화한 뒤 군집 기반의 TF-IDF(c-TF-IDF)를 활용하여 문서 군집의 주제어를 추출하는 방법을 제시한다.

뉴스 기사의 경우 하루에 생산되는 기사의 군집 개수가 매번 다르고 군집 간의 밀도도 서로 다르다. 특히 같은 사건을 다루는 기사를 군집화하기 위해서는 세부적인 토픽 식별이 중요하므로, 본 논문에서는 HDBSCAN과 BERTopic을 적용하여 군집과 군집의 주제어를 얻는다. 이렇게 얻어진 군집은 부적절한 기사를 군집에 포함할 수 있으므로 Mean-Shift와 주제어 기반 노이즈 제거를 적용하여 군집의 응집도를 향상시키는 방법을 제안한다.

2-2 문서 요약

문서 요약은 크게 추출 요약과 생성 요약으로 나뉘며, 최근 딥러닝 기술의 발전으로 생성 요약의 정확도가 크게 향상되었다. 딥러닝 기반 생성 요약 기술은 Seq2Seq 구조로 초기에는 순환 신경망에 기반한 방법[6]이 제시되었다. 이후 순환 신경망과 함께 어텐션(attention) 매커니즘을 사용하는 방법[7]의 유용성이 밝혀져 Seq2Seq 구조에서 어텐션 매커니즘만을 사용한 트랜스포머(transformer)[8]가 제안되고 이후 딥러닝 기반 생성 요약은 트랜스포머에 기반한 여러 모델이 개발되고 있다. 트랜스포머의 인코더를 사용한 양방향 언어모델인 BERT[9]와 디코더를 사용한 생성형 모델인 GPT[10]에 이어, BERT의 인코더와 GPT의 디코더를 사용한 BART[11]가 최근 요약 태스크에 특화된 모델로 우수한 성능을 보이고 있다. BART 모델을 기반으로 한국어 데이터를

이용하여 구축된 KoBART 모델[12]은 높은 성능으로 한국어 요약 태스크에 폭넓게 사용되고 있다.

뉴스 기사 요약에는 군집화 후 요약이 필요하므로 다중문서 요약을 적용한 여러 연구가 시도되었다. [13]은 군집화 후 이슈 및 단어 관련도에 기반한 문장 점수가 높은 문장을 선택하는 추출 요약을 제안하였다. [14]는 반도체 관련 기사에 한정적으로 휴리스틱 클러스터링을 수행한 후 문서 간의 유사도를 측정하여 대표 문서를 선정하고 선정된 대표 문서를 3개의 문단으로 임의로 나눈 뒤 문단별로 딥러닝 기반 요약 모델을 적용하였다. [15]는 딥러닝 기반 요약 모듈과 유사도 측정 모듈, 정보량 랭킹 모듈의 총 3개의 모듈로 다중문서를 요약하지만, 수동 군집화된 문서를 사용한다.

이런 기존 연구들은 학습 데이터에 기반하여 본문으로부터 요약을 추출하는 것을 목표로 한다. 이 방식으로 추출된 하나의 요약문은 본문 전체를 잘 대표할 수 있으나 뉴스 기사의 제목과 중복된 내용이 주로 포함되어 충분한 정보를 제공하지 못하는 한계가 있다. 본 연구에서는 기사를 전후반부로 분리하고, 전반부로부터 주요 내용을, 후반부로 추가 정보를 얻어 보다 충실한 요약문을 제공할 것을 제안한다. 실험에서는 전반부로부터 얻은 생성 요약과 기사 자체의 리드를 사용하는 결과를 비교한다.

Ⅲ. 데이터 수집 및 임베딩

3-1 뉴스 기사 수집

네이버 뉴스에서 2024년 4월 1일부터 2024년 4월 14일간의 종합 언론사 10개(경향신문, 국민일보, 동아일보, 문화일보, 서울신문, 세계일보, 조선일보, 중앙일보, 한겨레, 한국일보)의 뉴스 기사들을 수집하였다. 뉴스 기사는 각 언론사에서 자체적으로 분류하는 섹션 중에서 정치, 경제, 사회, 생활, 세계의 다섯 개의 섹션을 대상으로 수집하였다.

기사에서는 제목과 내용, 섹션, 발행일을 수집하였다. 제목과 내용의 BMP(basic multilingual plane)영역 이외의 특수문자는 제거하고, 음절이 50개 이하이거나 10,000개 이상인 뉴스 기사는 사용하지 않았다. 조건에 따라 수집된 뉴스 기사는 총 20,032개에 기사의 평균 길이는 262어절로 나타났다.

3-2 문서 임베딩

군집화 연구에서는 같은 주제에 대한 문서들은 그 임베딩 벡터 간 거리가 상대적으로 가까울 것으로 기대한다. 임베딩 벡터의 생성을 위해서는 다양한 사전학습 언어모델을 사용할 수 있으며, 문서의 임베딩 벡터가 문서의 특징을 얼마나 잘 표현하는지에 따라 클러스터의 품질은 향상될 수 있다. 본 논문에서는 BERTopic에서 사용하는 SBERT 모델 중에서 사전 학습된 한국어 언어모델 ko-sroberta-nli[16]를 사용하여 각 뉴스 기사의 본문을 모델 기본값인 768차원 벡터로 임베딩한다.

[17]의 연구에 따르면 데이터의 차원이 클수록 거리가 가까운 데이터와 먼 데이터의 거리 차이가 줄어드는 것으로 밝혀졌다. 이는 차원이 증가함에 따라 데이터 공간이 희소성을 띠게 되어, 각 데이터 간의 거리가 더 멀어지기 때문이다. 문서의 군집화를 위해서는 동일 주제와 그렇지 않은 주제 간의 거리 차이가 명확해야 좋은 성능을 보일 것이라 기대할 수 있다. 본 논문에서는 적절한 수준의 주제 간 거리 차를 얻기 위해 일일 단위의 데이터를 대상으로 섹션별로 차원 축소를 수행한다. 차원 축소에는 UMAP(Uniform Manifold Approxi-mation and Projection)[18]을 사용한다. nearest neighbors는 15, minimum distinct는 0으로 설정하였고, 유사도 측정에는 코사인 유사도를 사용하여 실험적으로 가장 좋은 결과를 보인 5차원으로 축소하였다.

Ⅳ. 군집화 및 정제

4-1 뉴스 기사 군집화

다양한 언론사에서 생성하는 동일한 사건을 다루는 뉴스 기사들을 군집화하기 위해 본 논문에서는 동일 날짜에 발행된 각 섹션의 모든 기사에 대해서 HDBSCAN으로 최초 군집을 얻은 뒤에 Mean-Shift 알고리즘과 추가적인 잡음 제거를 적용하는 뉴스 요약에 적합한 군집화 방식을 제안한다.

동일 사건을 다루는 기사라도 같은 단어나 문장으로 사건을 표현하지 않으므로 각 군집의 밀도 수준은 각기 다르다. 이러한 기사들에 비 계층적 군집화를 수행한다면 상대적으로 분산이 작은 군집에 잡음을 초래한다. 다양한 밀도 수준을 가진 문서 군집을 동시에 식별하기 위해서 계층적 군집화 알고리즘인 HDBSCAN을 사용한다. 그림 1은 수집한 뉴스 기사에 대한 HDBSCAN의 군집화 결과를 보인다. 결과에서 군집의 각 정점에서 중심까지의 거리에 대한 분산이 군집 별로 20배까지 차이가 나는 것으로 파악되어 각 군집의 밀도 차이가 큰 것을 확인할 수 있었다.

HDBSCAN은 밀도 기반으로 기하학적인 클러스터를 식별하여, 복합적인 주제의 기사가 서로 다른 주제의 군집 사이에 있는 경우에는 두 군집이 합쳐지는 문제가 발생할 수 있다. 예를 들어 표 1의 모든 기사는 HDBSCAN에서 레이블 0으로 같은 군집으로 예측되었지만 서로 다른 주제가 혼재되어 있다. 이를 하나의 군집으로 간주하여 요약문을 생성하면 정확하지 않은 정보를 전달하게 될 가능성이 크다. 본 연구에서는 이러한 군집을 분리하기 위하여 Mean-Shift 알고리즘을 사용한다.

Mean-Shift 알고리즘[19]은 데이터의 밀도가 높은 곳을 군집의 중심으로 판단하는 비 계층적 군집화 알고리즘이다. 다양한 밀도의 군집이 존재하는 전체 데이터에 Mean-Shift를 바로 적용하면 상대적으로 분산도가 낮은 군집에 잡음이 발생할 확률이 높아진다. 본 연구에서는 초기 군집 방식으로

HDBSCAN을 사용하여 밀도의 편차가 적은 초기 군집을 얻은 뒤에 각 군집별로 Mean-Shift 알고리즘을 적용하여, HDBSCAN에서 만들어진 기하학적 모양의 군집을 잘라 군집을 세분화하거나 노이즈를 제거한다. 그림 1의 HDBSCAN의 결과에 Mean-Shift를 추가로 적용한 결과인 그림 2에서 기하학적 모양인 28개 군집이 분리되어 37개의 클러스터로 분리된 것을 확인할 수 있다. 표 1에서는 Mean-Shift 알고리즘을 적용하여 비서실장과 영수 회담에 관련된 기사가 레이블이 0과 1인 두 군집으로 분리되고 노이즈 문서가 레이블 -1로 식별되는 것을 확인할 수 있다.

4-2 주제어 선정과 노이즈 제거를 통한 군집 정제

문서 임베딩에서 사용한 SBERT 모델에서 특징이 적절하게 추출되지 않는다면 군집에 잡음이 존재할 가능성이 있다. 이 단계에서는 군집의 정밀도를 높이기 위하여 통계 기반의 방법론으로 노이즈 문서를 제거한다. 군집의 정제 과정은 아래와 같다.

1) 주제어 선정 및 노이즈 주제어 제거

각 군집 별로 주제어 5개를 추출한다. 주제어 추출에서는 BERTopic의 c-TF-IDF를 사용하여 군집별 TF-IDF 수치

표 1. 정치 섹션 HDBSCAN과 HDBSCAN+Mean-Shift 결과의 예
Table 1. Example of HDBSCAN and HDBSCAN+Mean-Shift results on politics section

Title	Label		
	HDB-SCAN	+Mean-Shift	+Noise Filter
이재명·조국, 이제는 '법정의 시간' Lee Jae-myung and Cho Kuk, now it's 'time for the court	0	-1	-1
새 총리 후보에 주호영·김한길·박주선도 거론 Choo Ho-young, Kim Han-gil, and Park Ju-sun also mentioned as new prime minister candidates	0	0	-1
'차기 국무총리설'에 김부겸 "터무니 없는 소리 불쾌" Kim Boo-kyum on 'Next Prime Minister Rumors': 'Ridiculous and Unpleasant Remarks	0	0	-1
尹, 14일 비서실장 교체 전망… 원희룡·김한길·장제원 거론 Yoon expected to replace Chief of Staff on the 14th… Won Hee-ryong, Kim Han-gil, and Jang Je-won mentioned	0	0	0
尹, 이르면 14일 새 비서실장 임명… 대통령실·내각 개편 '신호탄' Yoon to appoint new Chief of Staff as early as the 14th… Signaling reshuffle of Presidential Office and Cabinet	0	0	0
尹 비서실장 후보로 원희룡 급부상… 이르면 14일 인선될 듯 Won Hee-ryong rapidly emerges as candidate for Yoon's Chief of Staff… Appointment expected as early as the 14th	0	0	0
이재명, 영수회담 가능성에 "尹 당연히 만나고 대화해야" Lee Jae-myung on the possibility of a leaders' meeting: 'Of course, Yoon and I should meet and talk	0	1	1
尹·이재명 '영수회담' 성사될까…대통령실 "체제 정비 前 대화 어렵다" [4·10 총선 이후] Will a 'leaders' meeting' between Yoon and Lee Jae-myung happen? Presidential Office: 'Difficult to have talks before reorganization' [After the April 10 general election]	0	1	1
尹, 일정 안 잡고 담화 내용·형식 고심… 李 "영수회담은 당연" Yoon, deliberating on the content and format of the address without setting a schedule… Lee: 'Leaders' meeting is a given	0	1	1

*To convey the original tone of the original article, Korean texts and its translations are paralleled in the left column.

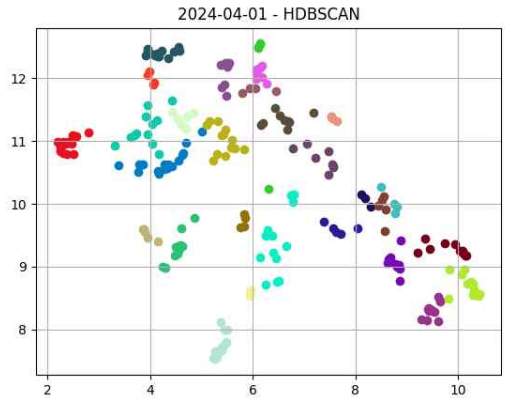


그림 1. 군집 평가 집합에 대한 HDBSCAN 군집화 결과
Fig. 1. Clustering result on evaluation set

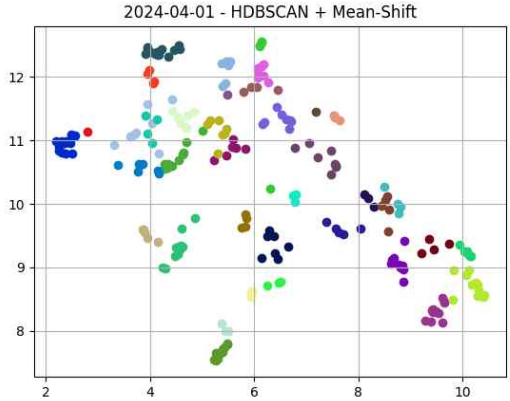


그림 2. 그림 1에 Mean-Shift를 추가 적용한 결과
Fig. 2. Results of applying Mean-Shift to Fig. 1

가 가장 높은 5개 단어를 선택한다.

선정된 주제어는 잡음을 포함할 수 있다. 수식 (1)에서는 군집 C의 5개의 각 주제어 tw_C를 대상으로 주제어가 클러스터 내부의 모든 문서 d에서 가지는 TF-IDF 값을 합산하여 주제어별 토픽 점수 $topicScore(tw_C)$를 계산한다.

$$topicScore(tw_C) = \sum_{d \in C} tfidf(tw_C, d) \quad (1)$$

토픽 점수가 높은 주제어는 해당 클러스터 내부에서 주요 토픽을 반영할 것으로 기대된다. 반대로 토픽 점수가 낮은 단어들은 주제 반영에 유효하지 않을 것으로 보고, 주제어들의 토픽 점수 평균 이하의 토픽 점수를 가진 단어는 클러스터와 관련도가 낮다고 판단하고 해당 단어들을 주제어에서 제외하여 정제된 주제어 tw'_C를 얻는다.

2) 노이즈 기사 제거 및 군집과 주제어 정제

각 문서 d에서 정제된 주제어 tw'_C의 TF-IDF를 합산하는 아래의 식으로 문서 점수 $docScore(d)$를 산정한다.

$$docScore(d) = \sum_{w \in tw'_C} tfidf(w, d) \quad (2)$$

노이즈 주제어 제거 단계에서와 마찬가지의 논리로 문서 점수가 낮을수록 상대적으로 클러스터 내부에서 주요하지 않은 문서라고 판단할 수 있다. 군집 내 문서들의 문서 점수 평균의 절반 이하의 문서 점수를 가진 문서는 잡음으로 판단하고 제거한다. 이때 군집에 남은 문서 개수가 2개 이하인 해당 군집은 잡음으로 간주한다. 노이즈 주제어 제거 과정을 거쳐 주제어의 수가 3개 미만으로 떨어지면 노이즈 기사 제거를 거친 군집의 문서들에 대해 클래스 기반 TF-IDF를 이용하여 토픽 점수 상위 3개의 주제어를 다시 추출한다. 이 과정을 잡음으로 제거되는 문서가 각 군집의 30% 이하일 때까지 반복한다. 아래 그림 3은 제안하는 군집화와 정제 과정을 그림으로 나타낸다. HDBSCAN으로 초기 군집 c_1, .., c_N을 얻고, Mean-shift를 통해 c_1은 c_{1-1}과 c_{1-2}, 잡음으로 군집이 세분화와 정제되며 c_N은 c_{N-1}과 잡음으로만 정제되고, 마지막 단계에서 주제어를 활용한 필터링을 통해 잡음을 반복 정제하는 과정을 그림으로 보인다. 최종적으로 얻어진 군집은 잡음이 최소화된 상태가 되며 이를 문서 요약의 입력으로 사용한다.

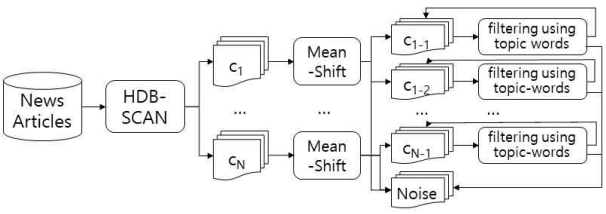


그림 3. 뉴스 군집화와 잡음 정제 과정
Fig. 3. News clustering and noise filtering process

V. 뉴스 기사 다중문서 요약

5-1 뉴스 기사의 구조

전문 기자가 작성하는 기사는 다른 문서와는 다르게 잘 정제되고 구조화되어 있다. 뉴스 기사 중 대부분을 차지하는 보도자료는 그림 4와 같이 역피라미드형으로 서술되어, 기사를 대표하는 문장인 제목(title)과 독자에 대한 흥미를 유발하고 기사에서 가장 중요한 사실을 포함하여 기사의 전체 내용을 요약하는 기사의 첫 문장인 리드(lead), 그리고 구체적으로 내용을 서술하는 본문(body)으로 이루어진다[20]. 이처럼 기사는 중요한 내용일수록 앞에 서술되고 구체적인 내용일수록 뒤에 서술되는 양상을 보인다.

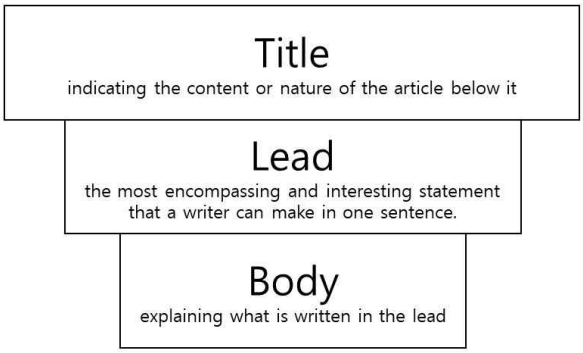


그림 4. 뉴스의 구조
Fig. 4. The structure of a news article

문서 요약 측면에서 다른 문서와 차별화되는 기사만의 가장 큰 특징은 제목과 리드가 있다는 점이다. 제목과 리드 모두 한두 문장으로 기사 전체 내용을 담고 있으며 기자가 직접 작성하므로 이를 해당 기사에 대한 정답 요약문으로 간주할 수 있다. 기존의 추출 요약 연구들에서는 문장 위치를 중요 정보로 사용하는 방식으로 리드를 활용해 왔다.

실제 기사에서 리드가 존재하는지 확인하기 위해 2024년 4월 10일 네이버 뉴스 기사 중에서 IT, 경제, 사회, 생활, 세계, 정치 섹션 별로 50개의 기사를 무작위 추출한 총 300개의 기사에 대해 첫 문장이 뉴스 전체의 내용을 요약하는지 수작업으로 확인하였다. 결과에서 300개의 기사 중 93%인 279개에서 첫 문장이 리드로 판단되었다. 이들은 일반적인 보도자료에 해당하였으며 리드만 읽어봐도 전체 뉴스의 내용을 이해할 수 있었으며 기자가 직접 작성하므로 문장의 질도 우수했다. 첫 문장이 리드가 아닌 7%의 경우들은 서론이나 인사말 등으로 시작하는 사설이나 분석, 비평, 인터뷰 기사들에 해당하였다. 리드가 존재하는 279개의 기사에 대해 KoBART로 얻은 생성 요약이 리드와 얼마나 유사한지 수동으로 분석한 결과에서는 90% 수준인 251개의 기사에서 요약이 리드 그 자체거나 리드에 준하는 문장으로 나타났다.

그림 5는 뉴스 기사의 예를 보인다. 첫 문장 즉 리드가 '리튬 이온 배터리 소재 개발 성공'이라는 전체 요약에 해당하며

이어진 기사 전반부는 제목과 리드에 연속되어 신소재 개발을 중심으로 기술되어 있다. 기사의 후반부에서는 기존 소재의 문제점과 함께 개발된 소재가 어떤 원리인지에 대한 추가적인 설명이 서술되어 있다. 이처럼 일반적인 뉴스에서 기사 전반부는 제목과 리드에 관련된 사항을 상세 서술하는 경향이 강하며 후반부는 추가 정보를 상세 서술하는 구조를 가진다. 이 기사에 대한 KoBART 생성 요약은 “UNIST는 고체 이온전도체를 개발하여 리튬 이온 배터리의 안전성을 높이는 소재를 개발하는 데 성공했다고 발표했다”를 결과로 제시한다. 이러한 요약은 제목과의 내용 중복성이 매우 높아 사용자가 제목이 아닌 요약에서 얻고자 하는 간결하지만 충분한 정보를 제공하지 못한다.

Title:UNIST's New Concept 'Lithium-Ion Battery' Significantly Reduces Explosion Risk and Enhances Performance

Contents:
UNIST(Ulsan National Institute of Science and Technology) announced on the 9th that it has successfully developed a material that enhances the safety of lithium-ion batteries.
Professors Sang-Young Lee and Sang-Kyu Kwak's team developed a 'solid ion conductor' concept that allows lithium ions to selectively and rapidly move through a straight, highway-like ion path.
Lithium-ion batteries are vulnerable to fire and explosion because they use flammable liquid electrolytes. As an alternative to address this issue, research is underway on 'all-solid-state batteries' and solid electrolytes, where the electrolyte between the anode and cathode is replaced with a solid. However, a drawback is that solid electrolytes generally have lower 'ion conductivity' compared to liquid electrolytes. Ion conductivity reflects the activity of charge movement due to ion migration, and higher ion conductivity improves battery performance.
Many solid electrolytes developed so far have complex and winding paths, making it difficult to enhance ion conductivity, thus limiting battery performance improvements. Since lithium ions are cations, they move with their counterpart anions. Unnecessary anion movement can cause undesired side reactions on the electrode surface, reducing battery performance.
The UNIST research team resolved this issue by using 'covalent organic frameworks (COFs)', a porous material with covalently bonded organic molecules, as a new ion conductor. Within this material, regularly arranged pathways are formed, which were designed to allow only lithium ions to pass through, significantly increasing ion conductivity.
Professor Lee stated, "We have laid the foundation for the development of 'high-performance solid electrolytes' necessary for the commercialization of next-generation batteries," and added, "This material can be widely used as a core material for lithium metal batteries, which have enhanced safety and efficiency."

그림 5. 뉴스 기사의 예시
Fig. 5. Example of a news article

본 연구에서는 본문을 전반부 본문과 후반부 본문으로 나누고, 후반부 본문에서 얻은 요약을 전반부 요약이나 리드와 결합하여 기사에 대한 요약으로 제시하여 더욱 구체적인 정보를 포함한 요약문을 제공하고자 한다. 전반부와 후반부는 문장 종결 기호를 기준으로 분리한 문장 중 앞쪽 $\lceil 1/2 \rceil$ 를 전반부로, 나머지를 후반부로 사용한다. 기사의 길이가 짧은 경우는 본문이 존재하지 않거나 존재하더라도 유의미한 정보가 없을 가능성이 크므로 기사의 길이가 200자 미만일 경우 기사 내용 전체를 요약문으로 취급한다.

5-2 제목과 리드를 활용한 대표 요약 선택

제안하는 시스템에서는 유사한 기사가 이미 군집화된 상태로 요약 단계에 입력된다. 군집 내의 기사들은 내용 중복이 매우 크므로 다중문서 요약을 수행하면 중복된 내용이 얻어질 수 있다. 뉴스 기사는 전문 기자들이 작성한 신뢰성 높은 문서이므로 군집 내 어떤 기사라도 전체 군집을 대표할 수 있다는 점에서 본 연구에서는 군집 내부의 모든 기사에 대한 요약을 추출하고, 그중 군집을 대표할 수 있는 요약을 결과로 선택하고자 한다.

같은 군집으로 묶인 기사들은 동일한 사실을 다루고 있어 뉴스의 리드 또는 전반부의 내용은 거의 유사한 것에 비해 각 기사의 후반부는 언론사에 따라 각기 차별적인 사실을 다룰 수 있다. 이 중 군집의 대표 요약은 가장 보편적인 것이 바람직하므로 후반부 요약이 군집 내 다른 문서와 얼마나 유사한지를 평가하여 대표성이 높은 후반부 요약을 선택하고, 해당 요약이 얻어진 기사의 전반부 요약과 결합하여 일관성 있는 요약을 생성한다. 후반부 요약 선정에서는 요약 태스크의 평가 지표로 널리 사용되는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[21]와 함께 최근 제시된 RDASS(Reference and Document Aware Semantic Score)[22]를 같이 사용한다.

아래의 식(3)~(5)는 요약 선정을 위한 점수를 계산한다. 각 식에서 $d \in C$ 는 군집 C에 속하는 각 문서 d, t_d는 문서 d의 제목(title), l_d는 문서 d의 리드(lead), b_d는 문서 d의 본문(body)을 나타낸다. s_{d_t}는 대상 문서 d_t에서 얻은 후반부 요약(summary)이다.

식 (3)의 $rgScore_C(s_{d_t})$는 대상 요약의 ROUGE 점수의 평균으로, 식 (4)는 $rdScore_C(s_{d_t})$는 RDASS 점수의 평균으로 각 요약의 점수를 계산한다. 식에서는 군집 내 다른 문서의 제목이나 리드, 본문과 유사성이 높은 요약이 높은 점수를 얻는다.

$$rgScore_C(s_{d_t}) = \underset{d \in C}{avg}(ROUGE\text{-}1(t_d, s_{d_t}) + ROUGE\text{-}1(l_d, s_{d_t})) \quad (3)$$

$$rdScore_C(s_{d_t}) = \underset{d \in C}{avg}(RDASS(b_d, t_d, s_{d_t}) + RDASS(b_d, l_d, s_{d_t})) \quad (4)$$

최종적으로 식 (5)를 통해 각 점수의 합이 가장 높은 점수의 요약을 클러스터의 대표 후반부 요약 $summary_C$로 판단한다. 시스템은 선택된 후반부 요약을 포함하는 문서 d_t의 리드 또는 전반부 요약을 후반부 요약과 결합하여 군집의 대표 요약으로 사용자에게 제시한다.

$$summary_C = \underset{s_{d_t} \in d_t \in C}{argmax}(rgScore_C(s_{d_t}) + rdScore_C(s_{d_t})) \quad (5)$$

VI. 분석 및 평가

6-1 평가 데이터

평가에서는 군집화와 요약의 성능을 각각 평가한다. 군집화에서는 본 논문에서 제안한 Maen-Shift 알고리즘과 잡음 정제의 적용이 HDBSCAN으로 얻은 초기 군집과 비교해 얼

마나 성능을 향상하는지 평가한다. 요약에서는 제안한 후반부 요약을 결합한 방식의 성능을 평가한다.

군집화 평가에서는 3-1절에서 기술한 데이터 중 2024년 4월 1일부터 2024년 4월 7일까지의 정치 섹션의 기사 1,520개에 대하여 수동으로 구축한 정답 데이터를 사용한다. 정답 구축에서는 HDBSCAN과 Mean-Shift를 적용하여 얻은 154개의 초기 군집을 수작업으로 새로이 분류하였으며, 결과로 143개의 정답 군집을 얻었다. 정답 군집은 일자별로 27, 25, 19, 26, 26, 8, 12개로 각기 다양하게 나타나 주말에 군집 개수가 줄어드는 양상을 관찰할 수 있었다. 군집의 평균 기사 개수는 7.51개였다.

요약 평가에서는 AIHub의 '요약문 및 레포트 생성 데이터'[23]의 뉴스 기사 300개를 사용하여 본 논문에서 제안한 후반부 요약 사용의 적절성을 평가하였다.

6-2 군집화 평가

평가 데이터에 대해서 HDBSCAN를 시행한 뒤 본 논문에서 제안한 Mean-Shift와 노이즈 제거를 단계별로 시행하고 각 결과를 평가하였다. 수동으로 구축한 143개 정답 군집은 평균 7.51개의 기사를 가진 것에 비해, HDBSCAN만을 적용한 결과는 군집 140개에 평균 9.29개의 기사, Mean-Shift를 적용한 결과는 군집 159개에 8.05개의 기사, 잡음 제거를 한 번 수행한 결과는 군집 143개에 6.30개의 기사를, 잡음 제거를 반복 수행한 결과는 군집 129개에 6.14개의 기사를 얻었다. Mean-Shift를 통해 군집이 세분화되었음을 확인할 수 있었으며, 잡음 제거를 통해 유사성이 큰 문서들만 군집으로 묶여 응집성이 높아졌음을 알 수 있다.

얻어진 결과를 대상으로 군집화 평가에 주로 사용되는 ARI(Adjusted Rand Index)[24]와 NMI(Normalized Mutual Information)[25], 정밀도(precision)를 이용하여 평가하였다. 본 연구는 같은 사건을 다루는 기사를 군집화하여 최종적으로는 요약을 제시하는 것이 목표이므로 노이즈가 최소가 되도록 응집도 높게 군집을 묶는 것이 중요하다. 따라서 평가에서도 정답 군집과의 유사성을 판단하는 ARI와 NMI와 함께 같은 군집으로 클러스터링된 기사들이 실제 같은 군집인지의 여부인 정밀도를 사용한다. 평가에서 사용하는 정밀도는 TP/(TP+FP)로 계산하며 시스템에서 같은 군집으로 판단한 두 문서가 정답에서 같은 군집이면 TP(True Positive)로, 다른 군집이면 FP(False Positive)로 판단한다. 정밀도가 낮아지는 경우는 노이즈 문서가 군집에 포함되는 경우이거나 노이즈가 아닌 문서를 정답이 아닌 군집에 분류한 경우이므로 시스템의 군집 결과가 응집도 높게 얻어져 기사 요약에 적합한지를 판단하는 좋은 기준이 될 수 있다.

평가에서는 일자별로 군집화를 30번 반복 시행하여 측정한 ARI와 NMI, 정밀도의 평균으로 성능을 측정하였다. 표 2는 평가 결과를 소수점 세 번째 자리에서 반올림하여 보인다. 결과에서 HDBSCAN의 초기 군집에 Mean-Shift를 추가 적용하여 모든 평가 지표에서 성능 향상을 얻은 것으로 나타났다. 이 결과에 잡음 제거를 한번 적용한 결과(+Noise removal)에서는 ARI와 정밀도가 향상되고, 노이즈 제거를 반복 적용(+Noise removal*)한 결과에서는 정밀도는 향상되었으나 ARI와 NMI는 하락하였다. 응집도 높은 군집을 구성하기 위한 잡음 제거 과정에서 군집에 포함되어야 하는 문서도 제거되어 ARI, NMI가 하락한 것으로 분석되었으며, 시스템의 최종 결과인 요약의 관점에서 가장 중요한 정밀도는 제안한 방식을 통해 가장 좋은 성능을 얻을 수 있었다.

표 2. 군집화 평가 결과
Table 2. Clustering evaluation result

	ARI	NMI	Precision
HDBSCAN	0.23	0.59	0.30
+ Mean-Shift	0.28	0.65	0.39
+ Noise removal	0.31	0.62	0.45
+ Noise removal*	0.27	0.59	0.46

6-3 문서요약

표 3은 요약 태스크에서 가장 많이 사용되는 지표인 ROUGE[21]와 RDASS[22]를 이용한 요약 평가 결과를 보인다. 표에서 ROUGE는 ROUGE-1, ROUGE-2, ROUGE-L의 평균이다. 표에서 각 행은 KoBART의 본문 요약만 사용한 경우와 KoBART의 전반부 요약과 후반부 요약을 결합한

표 3. 요약문 평가
Table 3. Summary evaluation

	ROUGE	RDASS
KoBART	0.15	0.80
KoBART 1st half summary + KoBART 2nd halfs summary	0.18	0.81
Lead+KoBART 2nd Half Summary	0.19	0.81

표 4. 평가 결과 예제
Table 4. Example of evaluation result

Title	UNIST's New Concept 'Lithium-Ion Battery' Significantly Reduces Explosion Risk and Enhances Performance
KoBART	UNIST announced that it has successfully developed a solid ion conductor, which improves the safety of lithium-ion batteries.
KoBART 1st+2nd half summary	UNIST announced that it has successfully developed a solid ion conductor, which improves the safety of lithium-ion batteries.They addressed this issue by utilizing a porous organic framework structure, which covalently bonds organic molecules, as a new ion conductor.
Lead + KoBART 2nd half summary	UNIST(Ulsan National Institute of Science and Technology) announced on the 9th that it has successfully developed a material that enhances the safety of lithium-ion batteries. They addressed this issue by utilizing a porous organic framework structure, which covalently bonds organic molecules, as a new ion conductor.

경우, 전반부 요약 대신 기사 자체의 리드와 후반부 요약을 결합한 경우를 보인다. 평가 결과에서 후반부 요약을 추가 사용하는 본 논문의 방식이 기사 전체 요약에 비해 높은 성능을 보였으며, 전반부 요약보다 리드를 사용한 것이 ROUGE에서 조금 더 나은 결과를 보였다.

표 4는 그림 5의 기사 예시에 대한 KoBART의 결과와 KoBART의 전반부와 후반부 요약의 결합 결과, 리드와 후반부 요약을 결합한 결과를 보인다. 본문 후반부 요약문을 포함해 기존 요약문에 비해 풍부한 정보를 제공하여 전반적인 뉴스 요약의 품질이 향상되었음을 알 수 있다.

표 1에서 제시한 예시에 대하여 시스템의 군집화를 거친 최종 요약 결과는 표 5와 같다. 표에서 Cluster title은 후반부 요약이 선택된 기사의 제목이다. HDBSCAN 이후 Mean-Shift를 적용하여 군집이 두 개로 분리되고 추가적인 잡음 제거로 얻은 응집도 높은 군집에 생성 요약을 적용하여 해당 이슈에 대한 적합한 요약을 얻음을 확인할 수 있다. 또한 후반부 요약을 결합하여 기사들이 다루는 충실한 정보를 제공하는 결과를 얻을 수 있었다.

표 5. 시스템의 최종 결과
Table 5. Example of the end result of the system

Cluster title	새 비서실장에 원희룡-이상민 등 물망… 野반발이 변수 Won Hee-ryong and Lee Sang-min are among the candidates being considered for the new Chief of Staff, but opposition from the opposition party remains a variable.
Cluster topic word	비서실, 원희룡, 교체 The Chief of Staff, Won Hee-ryong, replacement.
Cluster summary	4·10총선 참패 후 윤 대통령이 총리 교체 카드를 꺼내들었지만 192석을 확보한 범야권이 반대하지 않을 인사를 찾는 것이 1차적 관문이다. 윤 대통령의 의중을 잘 아는 몇몇 인사를 발탁할 경우 여당뿐만 야당의 반발 가능성도 고려해야 한다. Following the defeat in the April 10 general election, President Yoon has considered replacing the Prime Minister. However, the first hurdle is to find a candidate who will not be opposed by the opposition bloc, which holds 192 seats. If President Yoon selects individuals who are well-aligned with his intentions, the possibility of backlash from both the ruling party and the opposition must be taken into account.
Cluster title	尹·이재명 '영수회담' 성사될까…대통령실 "체제 정비 前 대화 어렵다" [4·10 총선 이후] Will a 'leaders' meeting' between Yoon and Lee Jae-myung happen? Presidential Office: 'Difficult to have talks before reorganization' [After the April 10 general election]
Cluster topic word	영수회담, 대화, 이재명 leaders' meeting, talks, Lee Jae-myung
Cluster summary	윤 대통령과 총선 참패 여파로 상처난 대통령실과 내각을 정비한 뒤 영수회담 여부와 시기, 방식에 대해 결정할 것으로 보인다. 야권발 특검법이 윤 대통령의 숨을 조이는 건 예견된 미래인 만큼 영수회담을 시작으로 정치적 돌파구를 마련해야 한다는 절박함이 대통령실 내부에 강해지고 있다. In the wake of the general election defeat, President Yoon is expected to reorganize the Presidential Office and Cabinet before deciding on the timing, manner, and whether to hold an inter-party summit. Given the anticipated pressure from a special prosecution bill initiated by the opposition, there is a growing sense of urgency within the Presidential Office to find a political breakthrough, starting with the inter-party summit.

*To convey the original tone of the original article, Korean texts and its translations are paralleled.

Ⅶ. 결 론

본 논문에서는 효율적인 뉴스 제공을 위한 응집도 높은 군집 생성이 가능한 군집 정제 방식과 후반부 요약을 결합한 정보성 높은 요약방식을 제안하였다. 군집 정제에서는 HDBSCAN으로 최초로 얻은 군집에 Mean-shift를 적용하여 군집을 분리하고 BERTopic의 c-TF-IDF을 이용하여 5개의 주제어를 얻은 뒤 주제어와 상관관계가 적은 문서를 군집에서 제거하는 방식을 적용하여 정밀도를 0.30에서 0.46으로 향상할 수 있었다. 요약에서는 상세 내용이 기술되는 기사 후반부의 요약을 추가 사용하여 ROUGE와 RDASS 모두에서 성능 향상을 얻을 수 있었다.

향후 연구로는 실시간으로 추가되는 기사에 대한 동적 군집화 및 토픽 모델링을 통한 군집화의 속도 향상 방법과 뉴스 기사에 대한 본 논문의 접근 방법을 회의록이나 논문 등에 대한 다중문서 군집화와 요약에 적용하는 연구를 계획하고 있다.

감사의 글

이 연구는 국립금오공과대학교 학술연구비로 지원되었음(2022년도)

참고문헌

[1] Korea Press Foundation. 2023 Newspaper Industry Survey [Internet]. Available: https://www.kpf.or.kr/front/research/realityListPage.do.

[2] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection," *ACM Transaction on Knowledge Discovery from Data*, Vol. 10, No. 1, 5, July 2015. https://doi.org/10.1145/2733381

[3] Y.-G. Lee and S. W. Kim, "A Comparative Study on Topic Modeling of LDA, Top2Vec, and BERTopic Models Using LIS Journals in WoS," *Journal of the Korean Society for Library and Information Science,* Vol. 58, No. 1, pp. 5-30, February 2024. https://doi.org/10.4275/KSLIS.2024.58.1.005

[4] M. Grootendorst, "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure," arXiv:2203.05794,

March 2022. https://doi.org/10.48550/arXiv.2203.05794

[5] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 3982-3992, November 2019. https://doi.org/10.18653/v1/D19-1410

[6] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, and H. Schwenk, Y. Bengio, “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1724-1734, October 2014. https://doi.org/10.3115/v1/D14-1179

[7] D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” arXiv:1409.0473, September 2014. https://doi.org/10.48550/arXiv.1409.0473

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, “Attention Is All You Need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach: CA, pp. 6000-6010, December 2017. https://doi.org/10.48550/arXiv.1706.03762

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language (NAACL-HLT 2019)*, Minneapolis: MN, pp. 4171-4186, June 2019. https://doi.org/10.18653/v1/N19-1423

[10] OpenAI. Improving Language Understanding by Generative Pre-Training [Internet]. Available: https://openai.com/index/language-unsupervised/.

[11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, ... and L. Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 7871-7880, July 2020. https://doi.org/10.18653/v1/2020.acl-main.703

[12] GitHub. SKT-AI/KoBART [Internet]. Available: https://github.com/SKT-AI/KoBART.

[13] H. Yu, S. Lee, and Y. Ko, “Incremental Clustering and Multi-Document Summarization for Issue Analysis Based on Real-Time News,” *Journal of KIISE*, Vol. 46, No. 4, pp. 355-362, April 2019. https://doi.org/10.5626/JOK.2019.46.4.355

[14] J.-Y. Kim, J.-E. Lee, J.-E. Lee, Y.-H. Lim, S. Lee, and M. Cho, “A Multi-Document Summarization System Using Heuristic Clustering: The Case of Semiconductor Industry,” *Journal of Digital Contents Society*, Vol. 23, No. 8, pp. 1437-1445, August 2022. https://doi.org/10.9728/dcs.2022.23.8.1437

[15] Y.-S. Lim, S. Kwon, B.-M. Kim, and S.-B. Park, “Multi-Document Summarization Use Semantic Similarity and Information Quantity of Sentence,” *Journal of KIISE*, Vol. 50, No. 7, pp. 561-572, July 2023. https://doi.org/10.5626/JOK.2023.50.7.561

[16] Hugging Face. jhgan/ko-sroberta-nli [Internet]. Available: https://huggingface.co/jhgan/ko-sroberta-nli.

[17] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When Is “Nearest Neighbor” Meaningful?,” in *Proceedings of the 7th International Conference on Database Theory (ICDT ’99)*, Jerusalem, Israel, pp. 217-235, January 1999. https://doi.org/10.1007/3-540-49257-7_15

[18] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” arXiv:1802.03426, September 2020. https://doi.org/10.48550/arXiv.1802.03426

[19] D. Comaniciu and P. Meer, “Mean Shift: A Robust Approach toward Feature Space Analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 5, pp. 603-619, May 2002. https://doi.org/10.1109/34.1000236

[20] J. Park, *Media Sentences and Reporting Methodology*, 2nd ed. Paju: Hanul Academy, 2005.

[21] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain, pp. 74-81, July 2004.

[22] D. Lee, M. Shin, T. Whang, S. Cho, B. Ko, D. Lee, ... and J. Jo, “Reference and Document Aware Semantic Evaluation Methods for Korean Language Summarization,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), pp. 5604-5616, December 2020. https://doi.org/10.18653/v1/2020.coling-main.491

[23] AI-Hub. Summary and Report Generation Data [Internet]. Available: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=582.

[24] L. Hubert and P. Arabie, “Comparing Partitions,” *Journal*

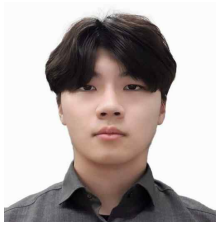
of Classification, Vol. 2, No. 1, pp. 193-218, December 1985. https://doi.org/10.1007/BF01908075

[25] L. N. F. Ana and A. K. Jain, “Robust Data Clustering,” in *Proceeding of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison: WI, p. II, June 2003. https://doi.org/10.1109/CVPR.2003.1211462

한성민(Seongmin Han)

2019년～현 재: 국립금오공과대학교 컴퓨터소프트웨어공학과 학사과정

※관심분야 : 자연어처리(Natural Language Processing), 데이터처리(Data Processing), 인공지능(Artificial Intelligence) 등



백대환(Daehwan Baek)

2019년～현 재: 국립금오공과대학교 컴퓨터소프트웨어공학과 학사과정

※관심분야 : 자연어처리(Natural Language Processing), 머신러닝(Machine Learning), 데이터과학(Data Science) 등

이현아(Hyunah Lee)

1996년 : 연세대학교 컴퓨터과학과 (학사)

1998년 : KAIST 전산학과 (석사)

2004년 : KAIST 전산학과 (박사)

2000년～2004년: ㈜다음소프트 언어처리연구소

2004년～현 재: 국립금오공과대학교 컴퓨터소프트웨어공학과 교수

※관심분야 : 자연어처리(Natural Lanauge Processing), 텍스트데이터마이닝(Text Data Mining) 등