KUBIG
DATA SCIENCE & AI

IT&과학 뉴스 요약 및 키워드 추출 프로젝트

NLP 2팀 | 16기 임정준 17기 서지민 19기 최주희

KUBIG
DATA SCIENCE & AI

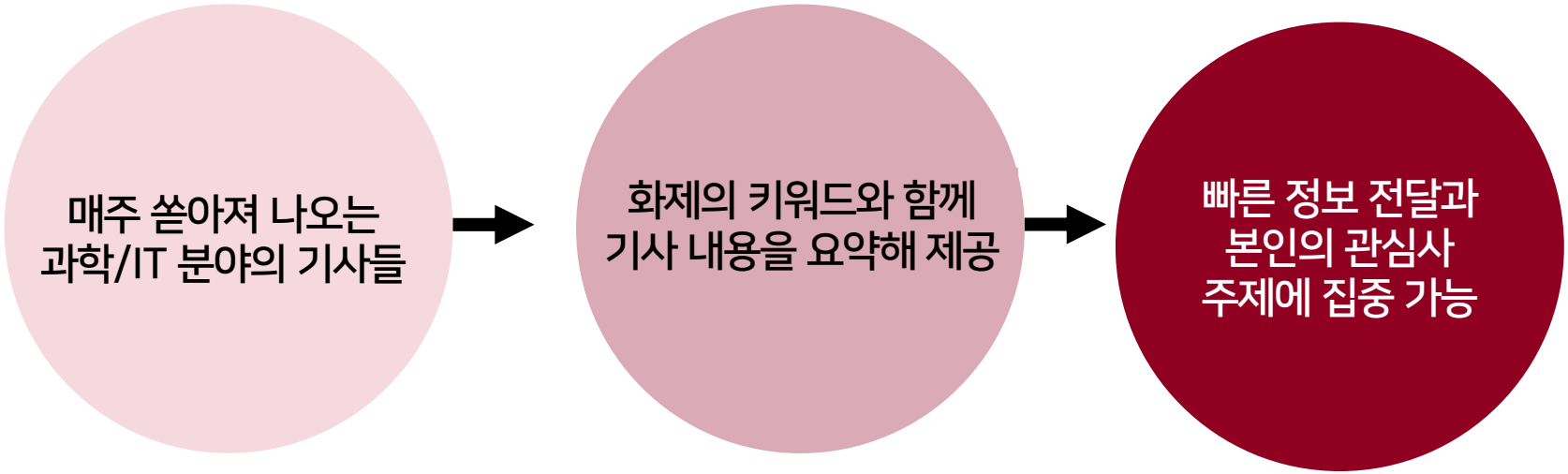
CONTENTS

KUBIG
DATA SCIENCE & AI

01. 주제 선정 및 개괄

01. 주제 선정 및 개괄

KUBIG
DATA SCIENCE & AI



01. 주제 선정 및 개괄

AI hub의
문서 요약 텍스트 데이터
활용

다음의 과학/IT 분야
뉴스를 크롤링하여
키워드 추출

그 주의 키워드와
키워드에 따른 뉴스
요약 내용 제공

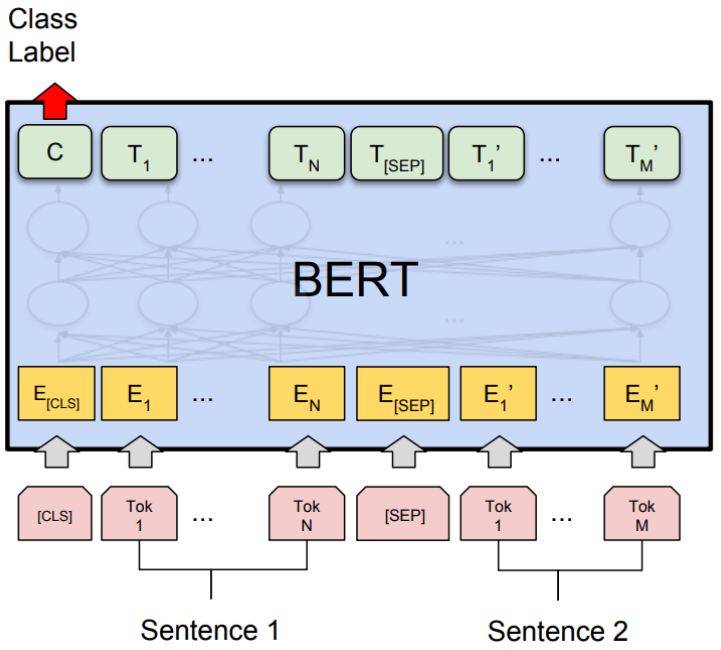
KUBIG
DATA SCIENCE & AI

02. 문서 요약 모델

02. KoBert

- 텍스트를 양방향으로 확인하여 문맥 정보와 문장 간의 선후관계를 학습해 자연어를 처리하는 모델인 Bert의 한국어 버전

- SKTBrain에서 위키피디아나 뉴스 등에서 수집한 수백만 개의 한국어 문장으로 이루어진 대규모말뭉치(corpus)를 학습시켜 한국어 처리를 용이하게 만든 모델

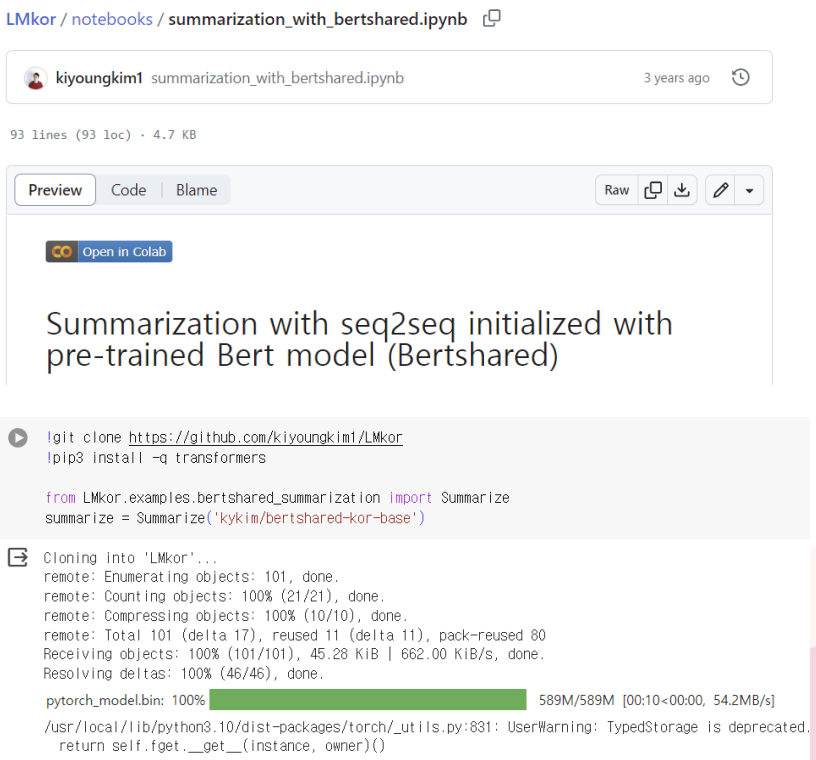


02. KoBertShared summarization

-https://github.com/kiyoungkim1/Lmkor

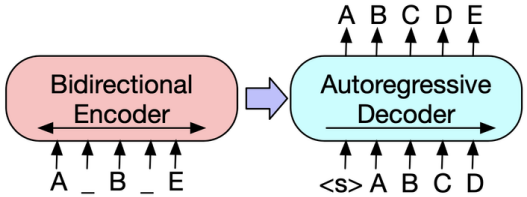
의 finetuning된 모델 사용

- EncoderDecoder 모델로 Encoder와 Decoder 간 파라미터 공유



02. KoBART-Summarization

BART



Masked documents들에 대하여 Autoregressive decoding하는 방식으로 Pre-trained parameter 학습

[BERT와의 주요 차이]

- Seq2Seq Transformer Architecture
- Activation Function: ReLU → GeLU
- Decoder의 각 Layer에서 Encoder의 마지막 Layer 이후 Cross-Attention 연산 수행
- Word-Prediction을 위한 FFN X
- 파라미터 수 10% 증량

KoBART (Ours)

SKT-AI에서 개발한, 40GB 이상의 한국어 텍스트를 활용하여 Pre-training Encoder & Decoder Model

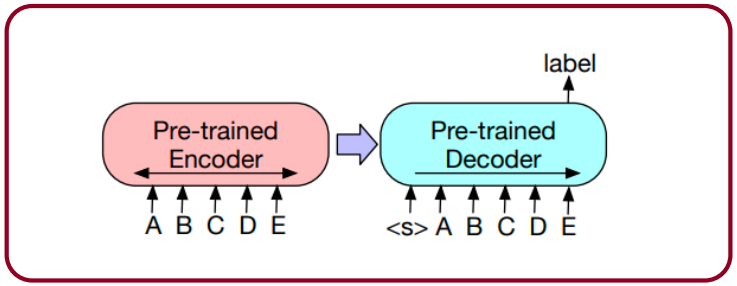
- 한국어 위키백과, 청와대 국민청원, 국립국어원 '모두의 말뭉치' 데이터 활용
- 3주간 1억 2400만개의 파라미터 사용
- 내부 GPU 자원 활용

모델 크기 특성상 Pre-Train 모델 기반으로 Fine-tuning 접근 (Fine-tuning에 친근화된 모델)

02. KoBART-Summarization

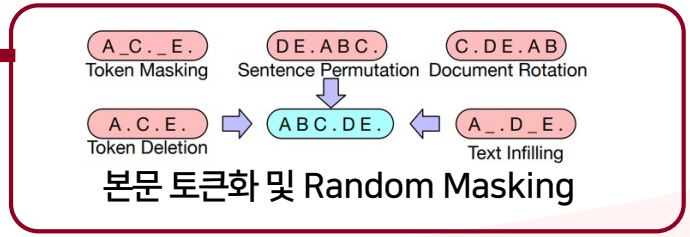
추출요약문

KoBART for Text-Summarization



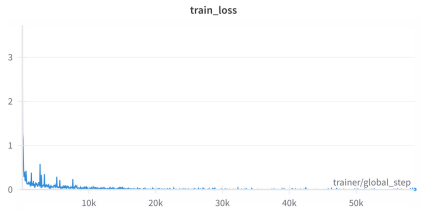
Fine-tuning koBART

본문



본문 토큰화 및 Random Masking

Fig. Loss graph for fine-tuning



- Training: 문서요약 텍스트 데이터셋 IT/과학 기사 1만 3천개
- Validation: 동 데이터셋 기사 2천개

02. Rouge Score - KoBART

Rouge-1
예측 요약본과 기존 요약본 사이 겹치는 Unigram 비교

Rouge-2
예측 요약본과 기존 요약본 사이 겹치는 Bigram 비교

Rouge-L
LCS 기반 최장길이로 매칭되는 문자열 유사도 측정

Example) Bigram 계산 +) Gram: NLP Basic Study 강의자료 참조

예측 요약본:
the cat was found under the bed

요약 답안:
the cat was under the bed

→

예측 요약본 (bigrams):
the cat, cat was, was found, found under, under the, the bed

요약 답안 (bigrams):
the cat, cat was, was under, under the, the bed

Recall: 4/5 = 0.8

Precision: 4/6 = 0.677

Recall	0.6078	0.537	0.6078
Precision	1	0.9667	1
F1-score	0.7561	0.6904	0.7561

KUBIG
DATA SCIENCE & AI

03. 키워드 추출

03. Daum News Crawling

Daum IT News

Crawling

Beautiful Soup

```
def daum_crawler(date_duration):
  dates = []
  today = datetime.today()-timedelta(hours=-9) #KST Time in colab

  for i in range(date_duration):
    today = today - timedelta(1) #Crawling until yesterday
    dates.append(today)

  news_all = []

  for date in tqdm(dates,desc = 'date loop'):

    date = date.strftime("%Y%m%d") #date setting
    page = 1  #page setting

    while True:

      url = f'https://news.daum.net/breakingnews/digital?page={page}&regDate={date}'
```

Articles CSV

- 2/20일부터 이전 7일간의 다음 IT 뉴스 크롤링
- 총 5337개의 뉴스 텍스트 크롤링 및 요약문 생성
- 키워드 추출 활용 결측치 및 짧은 보도성 뉴스 제거

	date	title	text
0	20240215	로봇진흥원, 5G기반 첨단제조로봇 실증 기술교류회 개최	[('지디넷코리아=신영빈 기자)한국로봇산업진흥원은(xx021일 대구 본원에서 5G ...
1	20240215	'세계 4위' 부자 다음주 10년 만에 한국 온다...尹대통령·이재용 만날 듯	['', '21일 IT 업계에 따르면, 저커버그 CEO는 아시아·태평양 순회 방문의...
2	20240215	무하유, AI 영상분석 전문가 이경주 CTO 영입	['', "인공지능(AI) 표절검사 서비스 '카피킬러' 운영사 무하유는 신임 최고개...
3	20240215	로보락, 지난해 국내 청소기전 매출 2천억원	[('지디넷코리아=신영빈 기자)청소기전 전문기업 로보락은(ya0지난해 국내 매출에이...
4	20240215	"1주에 최대 10개 의결권"…복수의결권주식 발행 1호 기업 탄생	[('지디넷코리아=방은주 기자)복수의결권주식 발행 1호 기업이 탄생했다. 작년 11...
...	...	...	...
5332	20240215	암 치료 근원 찾나...韓 연구진, 방사선 유발 DNA 돌연변이 첫 규명	['', "[서울=뉴시스]윤한성 기자 = 국내 연구진이 암 치료의 핵심 방법인 '방...
5333	20240215	세계 최초로 쇠고기발 먼들었다..'영양 만점'	['', '지금까지 다양한 형태의 배양육이 개발됐지만 곡물을 지지체로 이용한 방법은...
5334	20240215	방사선이 유발하는 DNA 돌연변이 특성 규명	[['아이뉴스24 최상국 기자] 암 치료의 중요한 방법 중 하나인 방사선 치료는 암...
5335	20240215	미세먼지 흡착력, 기존 에어필터 대비 200배	['', '와이즈랜드이즘(대표 우상학)은 'From Science to Engine...
5336	20240215	방사선 피폭되면 DNA 돌연변이 생긴다...국내 연구진이 첫 규명	['', '암 조직 성장을 억제하기 위해 사용되는 치료용 방사선이 인체 내 세포에서...

03. 뉴스 기사 군집화

DBSCAN clustering

밀도 차이 기반 알고리즘을 통해 중복된 기사들을 제거.

하루 동안 발행된 기사들에는 언론사만 다르고 내용이 같은 기사들이 많이 포함됨 -> 중복 기사 처리 필요.

- 유사한 제목을 가진 기사들의 군집을 생성 (기사 제목만 임베딩 후 군집화 진행)
- 해당 군집에서 대표 기사 하나만 추출하도록 함
- eps = 0.2, min_samples = 1

K-means clustering

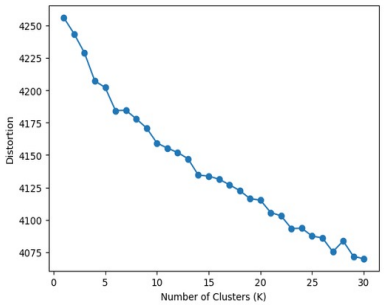
유사한 주제들의 기사를 묶어주기 위한 클러스터링 진행.

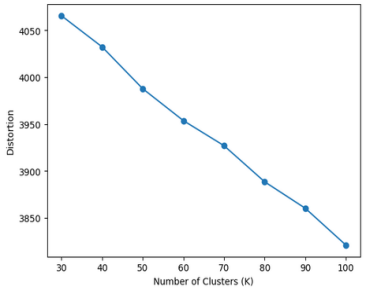
[최적의 k값 찾기]

- Elbow method: SSE 값의 감소 정도가 급격하게 줄어드는 지점의 k
- silhouette 계수가 높은 k

03. 뉴스 기사 군집화

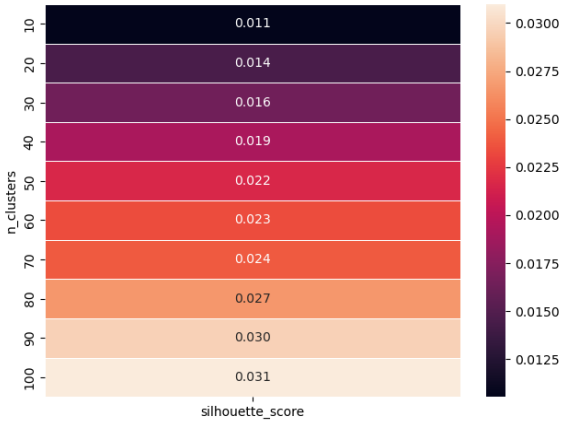
Elbow Method





K의 범위를 1~100까지 설정한 후
Elbow Method를 진행한 결과
SSE 값의 감소 정도가 급격하게 줄어드는 지점이 보이지 않음

Silhouette Analysis



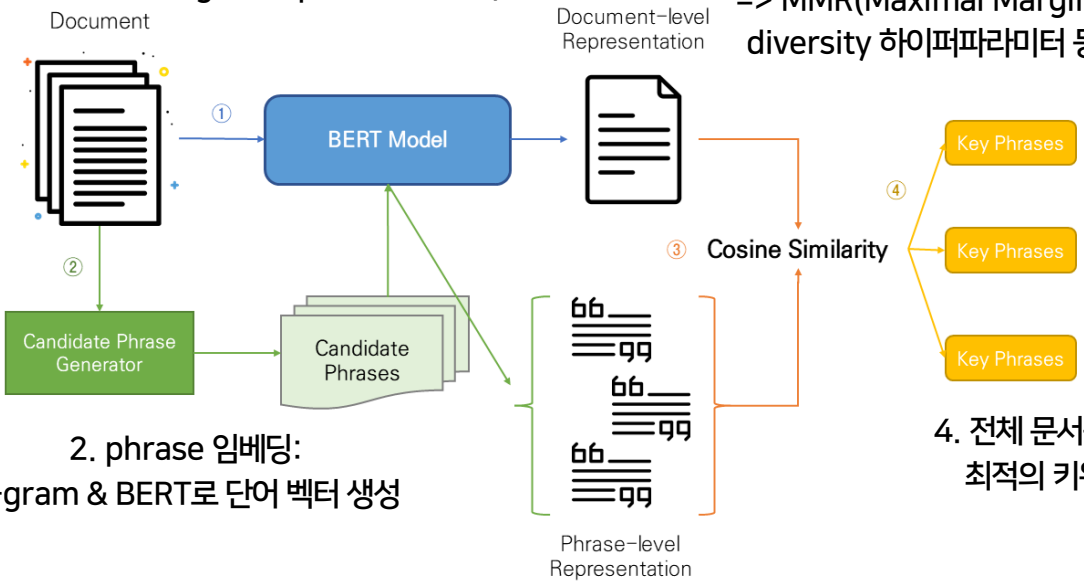
각 데이터 포인트가 속한 클러스터의 일관성을 측정하여
최적의 K값을 찾는 방법

k = 100일 때 실루엣 계수가 가장 높고 그 이상의 군집으로 나눌 시에는 키워드가 너무 많다고 판단해
→ k = 100으로 설정

03. 키워드 추출 - KeyBERT

KeyBERT: 토픽 모델링 중 키워드 추출을 위해 BERT를 사용한 오픈 소스 모듈

1. 문서 임베딩: BERT 모델 사용
(paraphrase-multilingual-mpnet-base-v2)



2. phrase 임베딩:
n-gram & BERT로 단어 벡터 생성

3. 문서와 가장 유사한 키워드 찾기
=> 코사인 유사도 사용
=> MMR(Maximal Marginal Relevance),
diversity 하이퍼파라미터 등 추가 튜닝 진행

4. 전체 문서를 설명하는
최적의 키워드 추출

03. 키워드 추출 - KeyBERT

키워드 추출, 어디에서? (제목 vs 요약문)

창작 국악
신임원장 디지털
운영 삼성강남서
네이버 학급밴드
갤럭시AI는 이제
듀얼모니터 노트북
출시 LG유플러스
쇠고기쌀
사이버훈련 DCM
삼성스토어 강의실
카톡 新성장
AI 기능
대전 스마트
졸업사진 기념사진
게임제작발표회 성료
한미家 소송전
로봇진흥원 5G기반
영림원소프트랩 ERP
AI폰은 갤S23
회장과 AI협력
AI가 주식매매
한전기술 원설본부
직무교육 구독서비스
출연硏 인재영입
LG유플러스 MWC서

해당되는 클러스터의 "제목"을 모두 붙인 후
이와 가장 유사한 키워드 추출

제목으로부터의 키워드 추출이
명사 다수 추출, 일목요연함 등
장점을 가지고 있기 때문에
제목으로부터 추출하기로 결정.

디렉터 작곡가
고령층의 디지털
게임 삼성
인게임에서 등장하는
모바일 AI시대를
스크린 노트북
구몬학습 빨간펜
개발했다 쌀에
육군 사이버작전센터
타타그룹은 아그라타스
감소해 수익성
아이폰16 모델로
관계자들이 양사의
로켓인 H3
NDM 게임제작발표회를
KAI가 사업비
5G기반 첨단제조로봇
태양전지 성능
갤럭시S23에도 조만간
한국을 방문한다
메리츠증권에 인공지능
일정으로 한국을
중소기업 임직원
연구자들을 유연하게
MWC2024 인공지능

해당되는 클러스터의 "요약문"을 모두 붙인 후
이와 가장 유사한 키워드 추출

KUBIG
DATA SCIENCE & AI

04. 결론 및 마무리

04. Summarization

제공된 요약

네이버가 인공지능(AI)기반 상품 추천 시스템 에이아이템즈(AiTEMS) 를 활용해 검색 결과에서도 관심사·선호도에 따라 쇼핑을 경험할 수 있는 '포유(FOR YOU)' 영역을 신설했다. 포유 영역은 '패션 류 및 잡화' 카테고리의 일부 품목에 우선 적용되며 최근 쇼핑 이력이 많은 사용자를 대상으로 노출된다.

결과

KoBert

네이버는 인공지능 (ai) 기반 상품 추천 시스템 에이아이템즈를 활용해 검색 결과에서도 관심사 선호도에 따라 쇼핑을 경험할 수 있는 포유(for you) 영역을 신설했다.

	Rouge1	Rouge2	Rouge-L
Recall	0.679	0.464	0.714
Precision	0.365	0.25	0.385
F1-score	0.475	0.325	0.5

KoBart

네이버는 인공지능 (ai) 기반 상품 추천 시스템 에이아이템즈를 활용해 검색 결과에서도 관심사 선호도에 따라 쇼핑을 경험할 수 있는 포유(for you) 영역을 신설했다.

	Rouge1	Rouge2	Rouge-L
Recall	0.867	0.867	0.867
Precision	1.0	0.929	1.0
F1-score	0.929	0.897	0.929

04. Summarization

제공된 요약

지난 25일 출시된 아이폰 11 시리즈의 첫날 개통량이 전작인 아이폰 XS·XR시리즈보다 30%가량 높은 것으로 나타났다.

KoBert

결과

지난 25일 출시된 아이폰 11 시리즈의 첫날 개통량이 전작인 아이폰 xs xr시리즈보다 30 % 가량 높은 것으로 나타났는데 이는 국내 이통사들의 5g 품질에 만족하지 못하는 고객이 많기 때문으로 보인다.

	Rouge1	Rouge2	Rouge-L
Recall	0.5	0.433	0.533
Precision	0.833	0.722	0.889
F1-score	0.625	0.542	0.667

KoBart

25일 출시된 출시된 아이폰 11시리즈의 첫날 개통량이 전작인 아이폰 XS·XR시리즈보다 30%가량 높은 것으로 나타났다.

	Rouge1	Rouge2	Rouge-L
Recall	0.474	0.459	0.474
Precision	0.947	0.944	0.947
F1-score	0.632	0.618	0.632

KUBIG
DATA SCIENCE & AI

Thank You