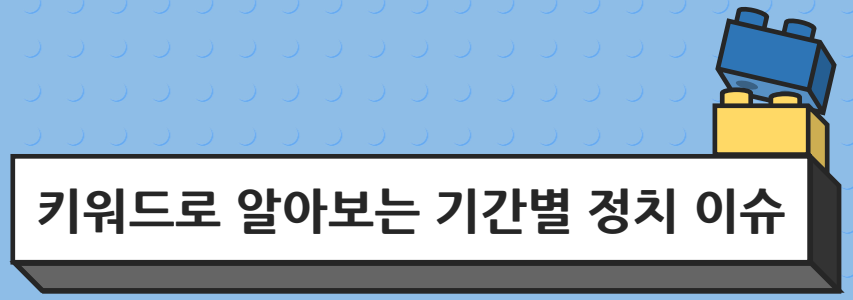


AIB_11기_CP1_2팀
김현승, 박동현, 박중혁, 신성윤

목차

01 프로젝트 개요

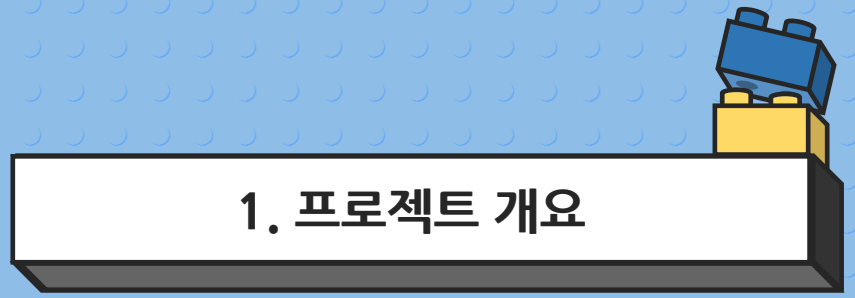
02 팀 구성 및 역할

03 수행 절차 및 방법

04 수행 결과

05 회고 및 마무리

06 참고자료 및 개발환경



1. 프로젝트 개요

키워드로 알아보는 기간별 정치 이슈

1. 주제 선정 배경

이용자들이 선호
할만한 컨텐츠

기간내에 수집할 수 있는
데이터는 무엇인가?

주제 회의만 8시간

유튜브 썸네일
추천

팀원 모두 NLP프로젝트 경험有

감성 분석

NLP

‘키워드로 알아보는 기간별 정치 이슈’

텍스트 요약

ComputerVision

KoBERT

색다른 주제

KoBART

챗봇

커뮤니티 감성
비교

교육사이트 후기
감성 비교

2. 프로젝트의 필요성

이 많은 뉴스를
언제 다 봅니까?

네이버 뉴스는 800여개의 언론사에서 일 평균 6만여건의 신규 기사가 생성되고,

누적으로는 1억 3천만건의 기사를 제공(2019, Naver Search & TECH)

3. 주요 기능

NAVER

네이버 뉴스 정치 기사

뉴스 댓글의 형태소 추출

특정 키워드를 포함한 뉴스 목록을 필터링하고 기사 댓글에서 많이 언급된 단어들을 워드 클라우드로 표현

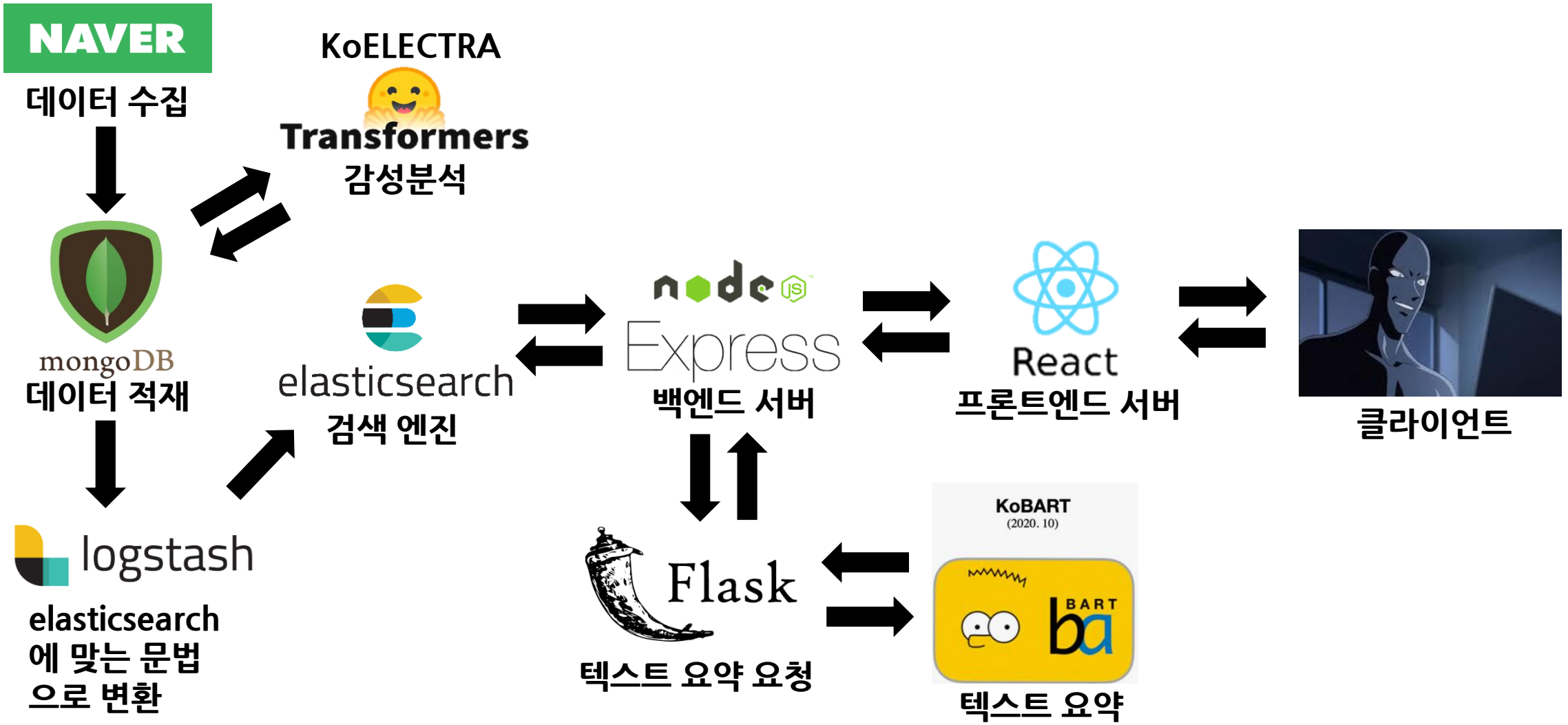
댓글 및 기사 감성 분석

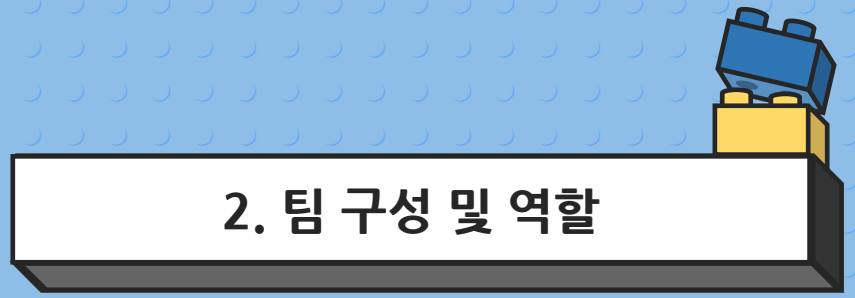
댓글에서 언급된 단어들이 어떠한 감성의 댓글에서 사용되었는지, 그리고 기간별 키워드를 포함한 기사들의 타이틀을 긍/부정 비율로 표현

기사 요약 서비스 구현

바쁜 현대인들을 위한 댓글이 많이 달린 상위 뉴스들의 요약본을 볼 수 있음.

4. 파이프라인


NAVER
데이터 수집
KoELECTRA
Transformers
감성분석
mongoDB
데이터 적재
elasticsearch
검색 엔진
node.js
Express
백엔드 서버
React
프론트엔드 서버
클라이언트
logstash
elasticsearch
에 맞는 문법
으로 변환
Flask
텍스트 요약 요청
KoBART
(2020. 10)
BART
텍스트 요약



2. 팀 구성 및 역할

키워드로 알아보는 기간별 정치 이슈

02. 팀 구성 및 역할

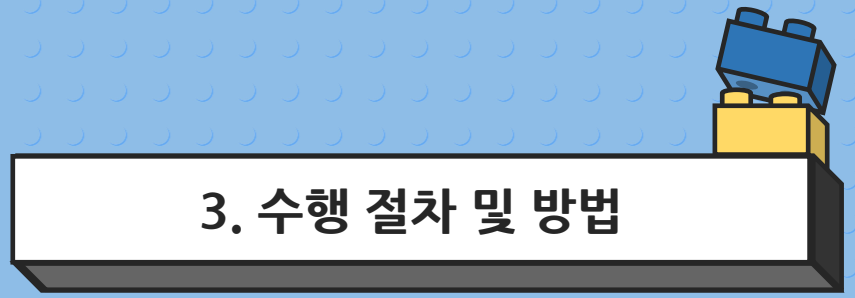
1. 담당 역할

박중혁
▶ 데이터 수집(Predict용) ▶ MongoDB ▶ ElasticSearch, Logstash ▶ React

신성윤
▶ 데이터 수집(fine-tuning용) ▶ 데이터 전처리 ▶ KoBERT - sentiment analysis ▶ 데이터 적재

박동현
▶ 데이터 수집(Predict용) ▶ 형태소 분석(KoNLPy) ▶ koELECTRA - sentiment analysis ▶ 발표 자료 제작

김현승
▶ 데이터 수집(fine-tuning용) ▶ 데이터 전처리 ▶ KoBART - summarization ▶ 데이터 적재

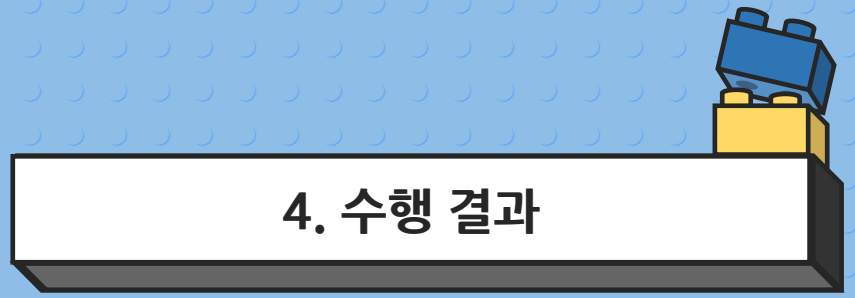


3. 수행 절차 및 방법

키워드로 알아보는 기간별 정치 이슈

03. 수행 절차 및 방법

구분	기간	활동	비고
사전 기획	▶ 6/24 ~ 6/26	▶ 프로젝트 기획 및 주제 선정 ▶ 기획서 작성	▶ 주제 확정
데이터 수집	▶ 6/26 ~ 7/1	▶ 네이버 및 다음 데이터 수집	▶ 기사 및 댓글
DB 설계	▶ 6/26 ~ 7/5	▶ Oracle Cloud에 MongoDB 설계(~6/28) ▶ Local MongoDB 스키마(~7/5)	▶ MongoDB
데이터 전처리 (모델 학습용)	▶ 6/28 ~ 6/29	▶ Fine-Tuning 데이터 정제	▶ NSMC, AI-HUB 등
모델링	▶ 6/27 ~ 6/29	▶ KoBERT 및 KoBART 공부 및 설계	
대체 모델링	▶ 6/30 ~ 7/1	▶ KoBERT 대체를 위한 KoELECTRA 설계	▶KoBERT로 뉴스데이터를 전부 처리하기에 모델이 느려서 경량화 모델로 구현
데이터 전처리 (서비스용) 및 적재	▶ 7/1 ~ 7/5	▶ Predict 데이터 형태소 분석(Elasticsearch) 감성분석(KoELECTRA)	▶감성분석한 결과를 DB에 적재
서비스 구축	▶ 7/1 ~ 7/6	▶ 웹 서비스 구축	▶ 최적화, 오류 수정



4. 수행 결과

키워드로 알아보는 기간별 정치 이슈

04. 수행 결과

서비스 데이터 소개

데이터 출처:
네이버 뉴스 - 정치 일반

수집 날짜 : 2019년 12월 ~ 2022년 1월

수집 데이터 : 뉴스 타이틀, 본문, 댓글

데이터 총 용량 : 20GB

뉴스 개수 : 20M

댓글 개수 : 100M

KoBERT(감성 분석)

BERT (Bidirectional Encoder Representations from Transformers)

모델 소개

- 2018년 구글이 공개한 **사전학습모델**
- 2018년 기준 11개의 자연어 처리 문제에서 SOTA 달성
- 기존의 사전학습모델(GPT-1,ELMo 등)과 달리 양방향성(bidirectional)을 가짐
- 2개의 문제 사전학습
  마스크드 언어 모델(Masked Language Model & 다음 문장 예측(Next Sentence Prediction)

모델 선택 이유

- 준수한 성능
- 풍부한 레퍼런스

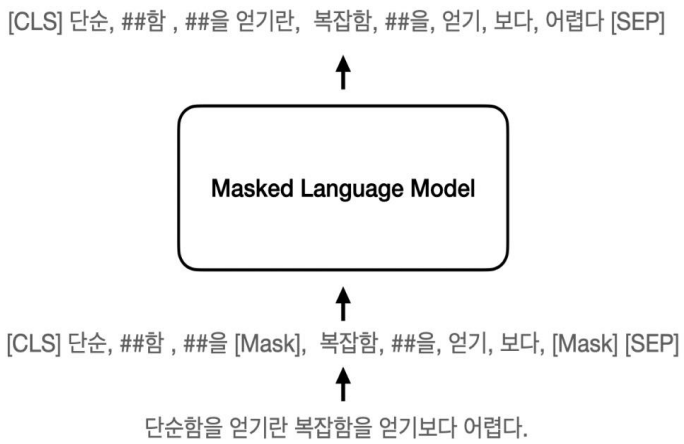
‘KoBERT’

04. 수행 결과

KoBERT(감성 분석)

(1) MLM(Masked Language Model)

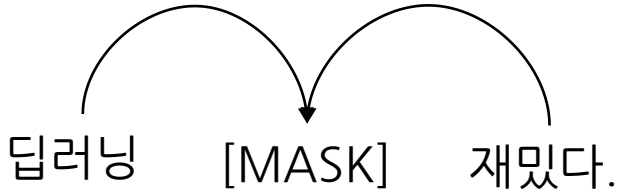
입력 문장의 일부 단어들을 마스킹해서 모델이 마스킹된 단어가 무엇인지 예측하게 합니다.



(2) Bidirectional

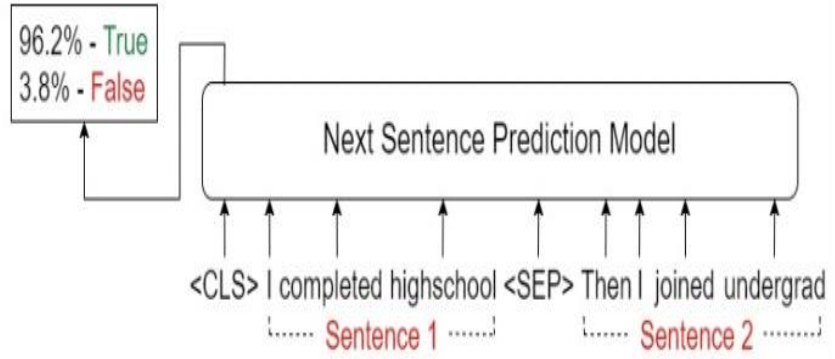
MASK 토큰을 예측하는 방법으로, GPT-1, ELMO 등은 MASK의 왼쪽 혹은 오른쪽 단어 하나로만 MASK를 예측한 반면 BERT는 양방향으로 추측한다.

예시) 딥러닝 공부 재밌다.



(3) NSP(Next Sentence Prediction)

두 문장의 관계를 이해하기 위해 BERT의 학습 과정에서 두 번째 문장이 첫 번째 문장의 바로 다음에 오는 문장인지 예측하는 방식입니다.



04. 수행 결과

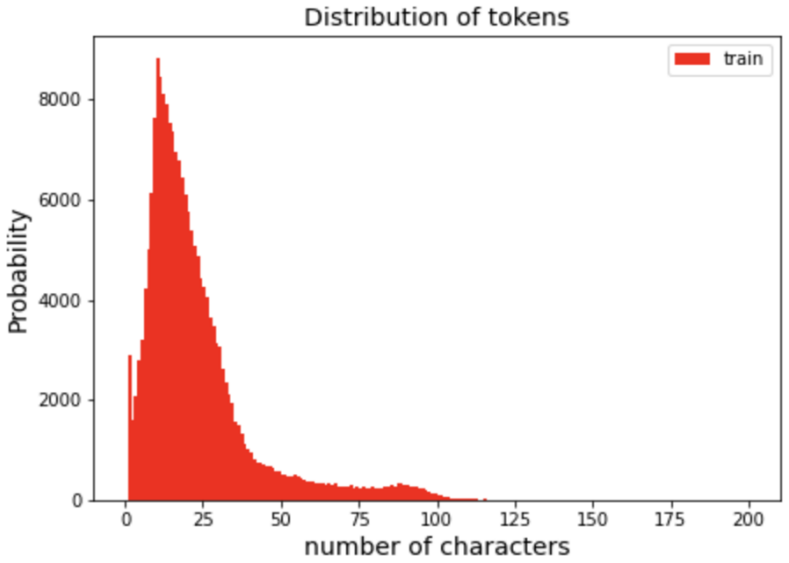
KoBERT(감성 분석)

Fine-tuning에 사용된 데이터

	Sentence	Emotion
0	굳 ㅋ	1
1	GDNTOPCLASSINTHECLUB	0
2	뭐야 이 평점들은.... 나쁘진 않지만 10점 짜리는 더더욱 아니잖아	0
3	지루하지는 않은데 완전 막장임... 돈주고 보기에는....	0
4	3D만 아니었어도 별 다섯 개 줬을텐데.. 왜 3D로 나와서 제 심기를 불편하게 하죠??	0

NSMC + AI-HUB(일상 대화 감성 데이터)
train dataset 약 20만건 / test dataset 약 6만건

Fine-tuning 주의사항




```
문장 tokens 개수 최대값: 225
문장 tokens 개수 평균값: 22.55202400793448
문장 tokens 개수 표준편차: 18.240284649330018
문장 tokens 개수 중간값: 18.0
문장 tokens 개수 1사분위: 11.0
문장 tokens 개수 3사분위: 27.0
문장 tokens 개수 99퍼센트: 93.0
```

BERT fine-tuning에 사용되는 최대 문장 길이는 토큰의 평균, 중간값, 최댓값을 고려해 적당한 최대길이를 수용해야 한다

04. 수행 결과

KoELECTRA(감성 분석)

ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)

모델 소개

기존 MLM 기반 모델의 **문제점**
=> 학습 데이터 크기에 비해 학습율이 낮다.

ELECTRA는 이를 해결하기 위해 Google에서 2020년에 개발한 Pre-trained 방법으로 설계된 모델이다.

특정 성능에 필요한 학습 시간

2018년 개발
GPT-2
120일

→

2020년 개발
ELECTRA
4일

Pre-training 과정

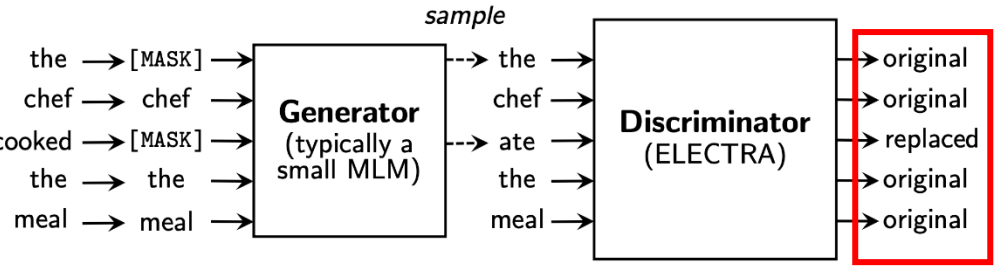


Figure 2: An overview of replaced token detection. The generator can be any model that produces an output distribution over tokens, but we usually use a small masked language model that is trained jointly with the discriminator. Although the models are structured like in a GAN, we train the generator with maximum likelihood rather than adversarially due to the difficulty of applying GANs to text. After pre-training, we throw out the generator and only fine-tune the discriminator (the ELECTRA model) on downstream tasks.

기존 MLM은 문장 내의 15% 단어에 대한 loss만 학습하는데에 비해 ELECTRA는 모든 단어의 loss를 구해 학습하게 되어 훨씬 적은 양으로 높은 학습이 가능하다.

KoELECTRA(감성 분석)

1) Pre-trained 데이터

본 프로젝트에서는 약 34GB 정도의 뉴스, 위키, 신문, 문어, 구어, 메신저, 웹 등의 한국어 데이터로 Pre-trained된 KoELECTRA-small를 사용하였다.

https://github.com/monologg/KoELECTRA

2) Fine-tuning 데이터

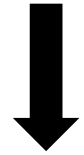
Fine-tuning에는 AI-HUB의 한국어 감정 정보가 포함된 단발성 대화 데이터셋의 총 7개 label에서 공포,분노,슬픔,혐오의 감정을 부정으로 행복을 긍정으로 변환하였고, NSMC(네이버 영화 리뷰 감성 긍/부정 데이터셋)를 병합하여 약 27만 개의 데이터를 fine-tuning에 사용하였다.

https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-009
https://github.com/e9t/nsmc

KoBERT

400MB

100개 문장 처리 속도 10초



KoELECTRA

54MB

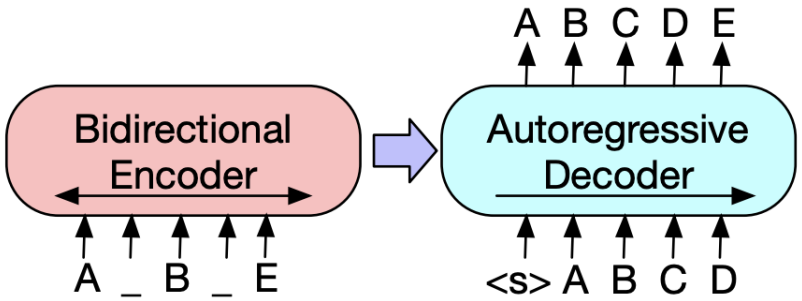
100개 문장 처리 속도 1초

04. 수행 결과

KoBART(텍스트 요약)

1) BART (Bidirectional Auto-Regressive Transformers)

BERT의 인코더와 GPT의 디코더 부분을 결합한 Transformers.



Task	
기계 번역	감성 분석
챗봇	**문서 요약**

2) Pre-trained 40GB 이상의 한국어 텍스트 데이터

예시) 안녕하세요 저는 코드스테이츠 AIB 11기 수강생입니다.

[MASK] 코드스테이츠 AIB 11기 수강생입니다. ➡ 안녕하세요 저는 코드스테이츠 AIB 11기 수강생입니다.

04. 수행 결과

KoBART(텍스트 요약)

3) Fine-tuning

Dataset: AI Hub 문서요약 텍스트

학습: 300,000
검증: 30,000

학습: 20,000
검증: 5,000

	news	summary
0	[박재원 기자] '대한민국 5G 홍보대사'를 자처한 문재인 대통령은 "넓고, 체...	8일 서울에서 열린 5G플러스 전략발표에 참석한 문재인 대통령은 5G는 대한민국 혁...
1	] 당 지도부 퇴진을 놓고 바른미래당 내홍이 격화되고 있다.바른미래당이 8일 연 최...	8일 바른미래당 최고의원 회의에 하태경 의원 등 5명의 최고의원이 지도부 퇴진을 요...
2	[홍윤정 기자] 8일 서울 올림픽공원 K아트홀.지난 3일 한국이 세계 최초로 5...	지난 3일 한국이 세계 첫 5세대 이동통신 서비스를 보편화한 것을 축하하는 '코리안...
3	] 박원순 서울시장(사진)이 8일 고층 재개발·재건축 관련 요구에 작심한 듯 쓴소리...	박원순 서울시장은 8일 서울시청에서 열린 '골목길 재생 시민 정책 대화'에 참석하여...
4	[임근호 기자] "SK(주)와 미국 알파벳(구글 지주회사)의 간결한 지배구조를 ...	주주가치 포커스를 운용하는 KB자산운용이 SK와 알파벳(구글 지주회사)의 모범적 ...

4) KoBART를 사용한 이유

- 인코더(BERT), 디코더(GPT)가 모두 존재

=> 양방향의 문맥정보를 반영, 일반화 능력 향상

- MLM을 사용하여 노이즈에 유연함

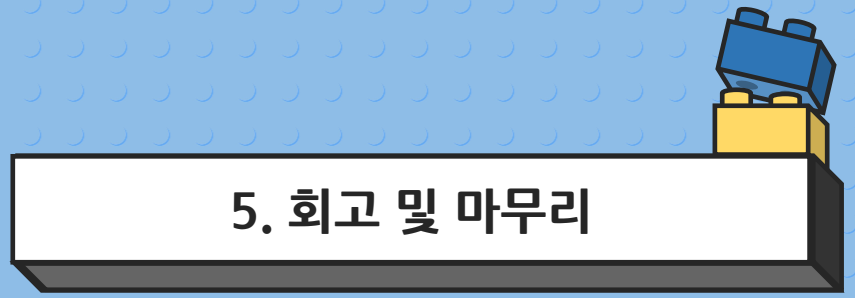
=> 토큰, 문장의 길이 변형에도 문장을 잘 생성

- 문장 생성에 특화 된 모델

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

04. 수행 결과

서비스 시연



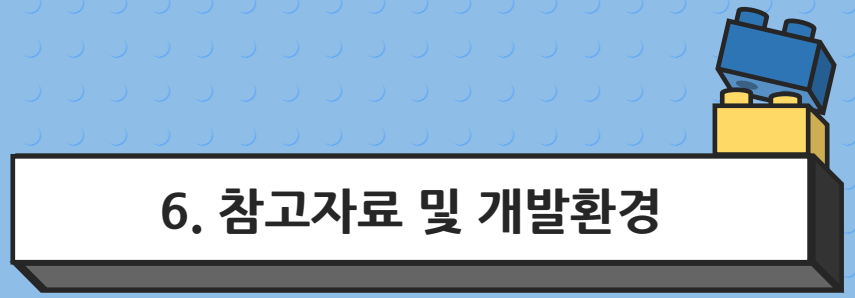
5. 회고 및 마무리

키워드로 알아보는 기간별 정치 이슈

한계점

한계점	해결방안
기존 계획에는 네이버를 포함한 여러 커뮤니티 데이터 그리고 정치 키워드 외에 여러 키워드를 사용하고자 했으나 DB에 데이터를 적재하는 시간과 서버 비용으로 인하여 기간내 수집할 수 있는 데이터의 한계로 네이버 뉴스 정치 섹션만 사용하게 되었다.	시간이 더 주어진다면 언론사 구분과 네이버 뉴스뿐만이 아닌 다양한 커뮤니티 데이터도 사용할 수 있을 것이다. 또한 정치 외의 키워드를 사용한 분석도 가능하여 보다 정확하고, 흥미로운 분석이 가능해 질 것 같다.
문서 요약 모델의 Fine tuning 시간이 오래 걸려서 많은 데이터를 사용하지 못했다.	좋은 GPU 환경이 주어진다면 더 깊게 학습시켜서 성능을 올릴 수 있을 것 같다.
감성분석에 중립 클래스를 사용하고 싶었지만, 중립 데이터는 긍/부정 데이터에 비해 상대적으로 작은 불균형 데이터였다.	모델 학습을 under-sampling 해보았지만 모델의 성능이 크게 떨어져 중립 데이터를 제외한 긍/부정 데이터로 학습하였다.

그리고 시간과 비용...



6. 참고자료 및 개발환경

키워드로 알아보는 기간별 정치 이슈

06. 참고 자료 및 개발 환경

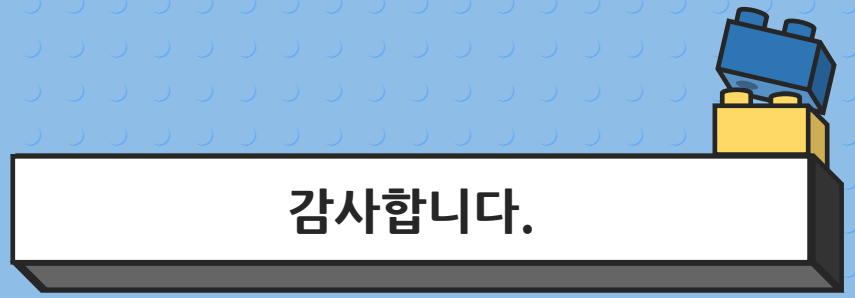
참고 자료

KoBART-summarization	KoBERT-sentiment analysis	PyTorch	Ko-ELECTRA -sentiment analysis
KoBART-summarization 구현예제1	KoBERT 모델 - GitHub	자주쓰는 Loss Function	fine-tuning 예제
KoBART-summarization 구현예제2	KoBERT 모델 - Huggingface	.detach().cpu().numpy()	논문 리뷰1
KoBART-summarization GitHub	다중분류모델 예제	Pytorch에서는 왜 항상 optimizer.zero_grad()를 해줄까?	**KoNLPy**
BART논문에 대한 블로그 글	fine-tuning 예제	loss.backward()	KoNLPy 공식
summarization 연구 동향 정리	BERT기반 모델들의 한국어 버전 Ko-RoBERTa ALBERT-Kor Ko-ELECTRA	**ElasticSearch**	품사표
요약 모델 성능 평가 척도(RDASS)	위 모델들의 논문 리뷰 블로그 글	Elastic 가이드	**크롤링**
Transformer 기반 Text Generation 옵션	Dialog-BERT	Elastic 기본개념과 특징	HTTP 요청 헤더
	Fine-tuning할 불균형 데이터 전후처리 성능비교	Docker로 elasticsearch와 연관 앱 설치	비동기(asyncio)

06. 참고 자료 및 개발 환경

개발 환경

Colab Pro, VS code, Python 3.8, Pytorch, Flask, Mongo DB, Elastic Search, React, Java Script, Transformers, Logstash, NodeJS Express



키워드로 알아보는 기간별 정치 이슈

AIB_11기_CP1_2팀
김현승, 박동현, 박중혁, 신성윤