

AI 기반 데일리 뉴스 요약 챗봇 서비스



Team Leader

- 프로젝트 전체 총괄
- [Lim Heejin](#)
- dg961108@naver.com
- <https://github.com/heejvely>



Team Member

- 모델 설계 총괄 및 서비스 통합
- [Han A-Leum](#)
- hal0576@naver.com
- <https://github.com/zena-H>



Team Sub-Leader

- 모델 구축 프로세스 및 데이터 관리 총괄
- [Jo Hyunjeong](#)
- chchooobi@gmail.com
- <https://github.com/JoHyunjeong>



Team Member

- 서비스 배포 구현 총괄
- [Lee Jonghyeon](#)
- leejonghyeon819@gmail.com
- <https://github.com/Jjongu>



Team Member

- 모델 성능 개선 총괄
- [Bae Songyi](#)
- kksong13@gmail.com
- <https://github.com/kksonge>



Team Member

- 모델 성능 평가 총괄
- [Choi Yunjin](#)
- cyunjin@gmail.com
- <https://github.com/ete-llorona>

INDEX

1) 서비스 개요

2) 모델 개발

3) 서비스 배포

서비스 개요

서비스 개요



AI 기반 뉴스 요약기 구축 후
다양한 신문사의 기사를 요약하여
전달하는 챗봇 서비스 진행

서비스 개요

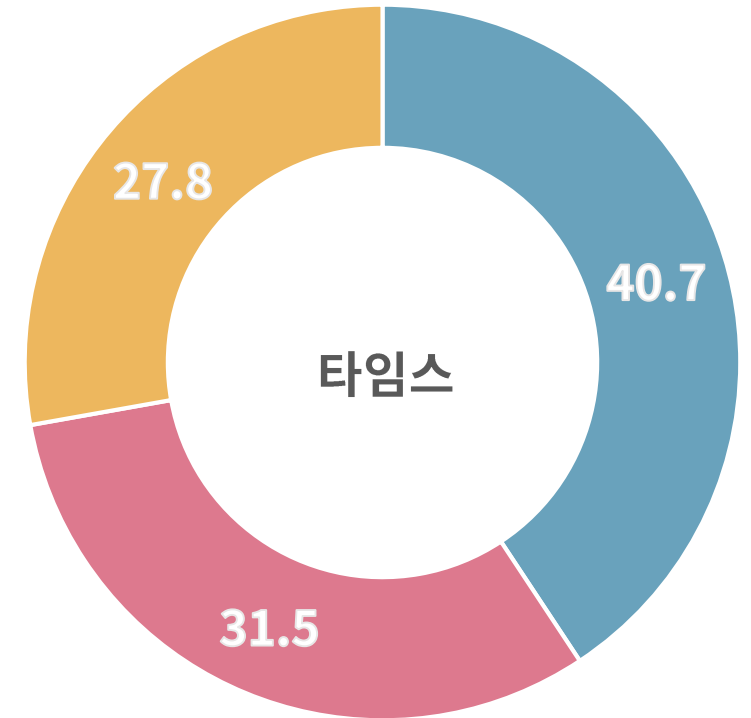
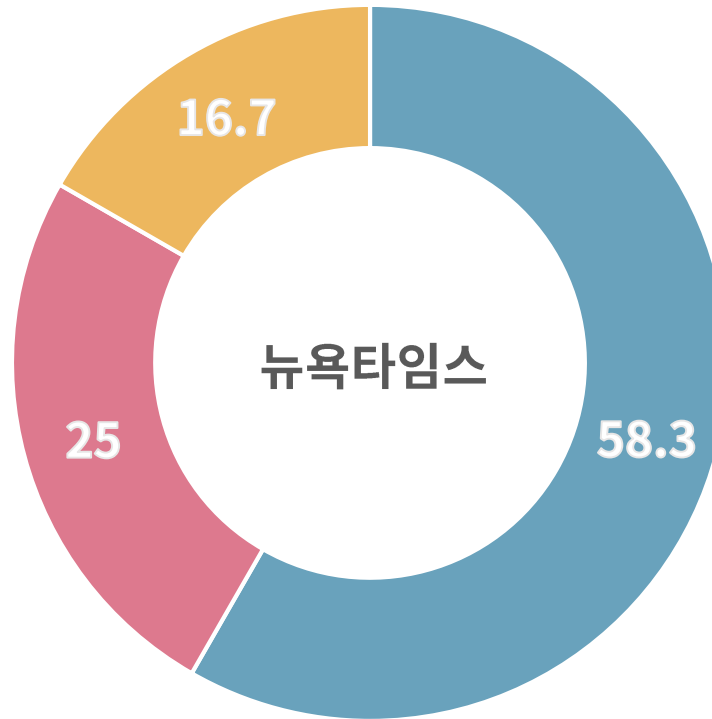
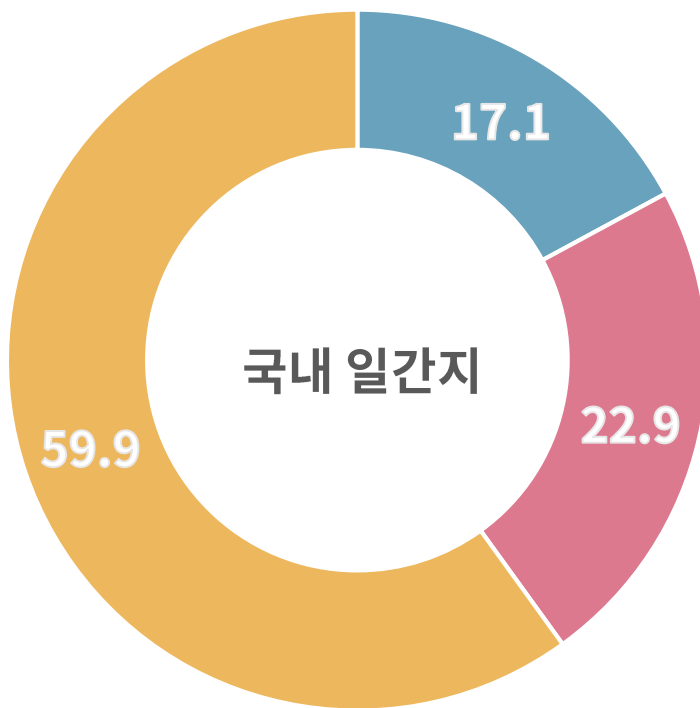
<부록> 경제활동인구조사 청년층 부가조사 개요

통계청은 7.16.(화) 2019년 5월 경제활동인구조사 청년층 부가조사 결과를 발표하였다.

- 청년층 인구는 907만 3천명으로 전년동월대비 8만 4천명(-0.9%) 감소하였음.
- 경제활동참가율은 48.4%로 전년동월대비 0.7%p 상승하였고, 고용률은 43.6%로 전년동월대비 0.9%p 상승하였음.
- 대졸자(3년제 이하 포함)의 평균 졸업소요기간은 4년 2.8개월로 전년동월대비 0.1개월 증가하였고, 휴학경험 비율은 45.8%로 1.4%p 상승하였음.
- 졸업(중퇴) 후 첫 취업 소요기간은 10.8개월로 전년동월대비 0.1개월 증가, 첫 직장 평균 근속기간은 1년 5.3개월로 전년동월대비 0.6개월 감소하였음.
- 첫 일자리에 취업할 당시 임금(수입)은 150만원~200만원 미만(34.1%), 100만원~150만원 미만(27.7%), 200만원~300만원 미만(18.1%) 순으로 나타남.
- 졸업(중퇴) 후 취업 경험자 비율은 86.2%로 전년동월대비 0.3%p 하락하였음.
- 청년층 비경제활동인구 중 취업시험 준비자 비율은 15.3%로 전년동월대비 2.2%p 상승하였음.

경향신문	첫 직장서 월 150만원 미만 저임금 청년 노동자 크게 줄었다
파이낸셜뉴스	청년 27.7%, 최저임금도 못 받는 '첫 일자리'
아시아 경제	첫 취업한 청년, 월급으로 150만 ~ 200만 가장 많이 받아

서비스 개요_ '관점의 다양성' 기사 비율



국내 일간지의 단일 관점의 기사 비율이 현저히 높음

서비스 개요_챗봇 서비스 기능



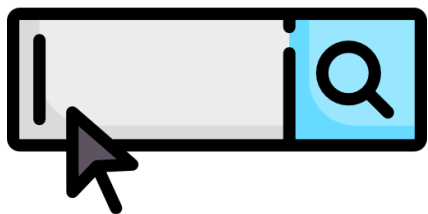
매일 아침 콕!

매일 오전 9시
헤드라인 뉴스 요약
구독 서비스



핵심만 콕!

4개 신문사별
동일 주제에 대한
헤드라인 뉴스 비교



원하는 것만 콕!

분야 선택, 키워드 검색,
링크 검색 등 사용자의
요구에 맞춘 뉴스 제공



실시간으로 콕!

실시간 업데이트된
최신 뉴스 제공

모델 개발

모델 개발_환경

Environment

Library &
Framework

Language

뉴스 크롤링



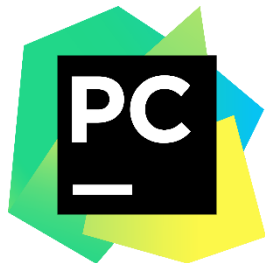
BeautifulSoup

딥러닝 모델 구축



PyTorch

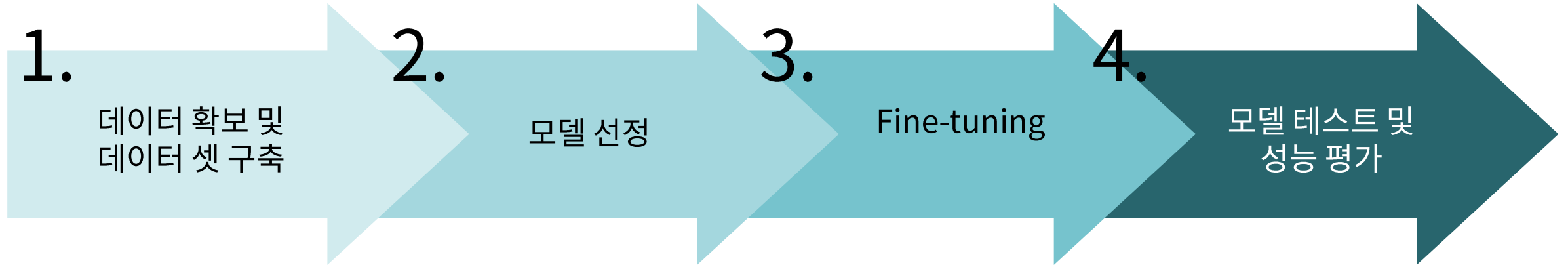
챗봇 서비스 구현



Flask



모델 개발_구축 과정



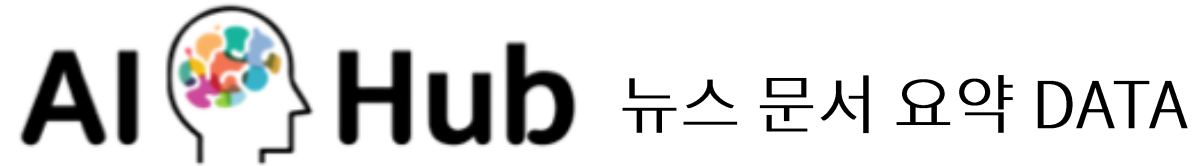
- 30만 건의 뉴스 데이터 확보
- Train/test 용 데이터 셋으로 구성

- Google-MT5, ETRI-ET5, KE-T5 테스트 후 생성 요약에 적합한 모델 선정

- 요약용을 위한 모델의 Fine-tuning 진행

- 요약문 생성
- BERT Score로 성능 평가

모델 개발_데이터 확보



구분	원문 건수
종합	9만 건
정치	3만 7,500건
경제	3만 7,500건
사회	3만 7,500건
문화	3만 7,500건
스포츠	3만 건
IT/과학	3만 건

총 30만 건 데이터 보유

모델 개발_데이터 구조

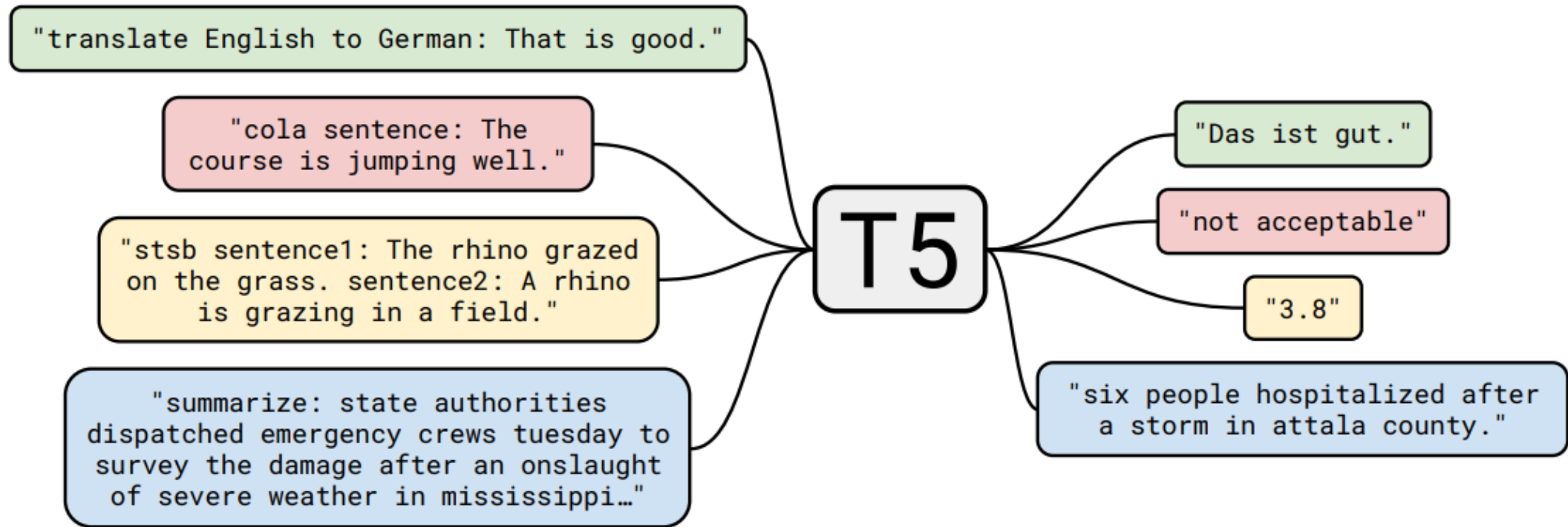
Original text
<p>오는 17일, 우수 공동체 사례 발표 및 활동 전시 공유김영신 기자 yskim@gynet.co.kr오는 17일 중마동 청소년문화센터에서 '2018 광양시 마을공동체 한마당'행사가 처음으로 개최된다.행사는 풍물동아리의 식전 행사에 이어, 행사 참여자들이 함께 박을 터뜨리며 마을공동체 어울림 한마당 행사의 문을 열 예정이다.특히'비벼봐 신나게!허형채 지사장은"이번 행사는 마을공동체의 활발한 활동과 성과를 보여주고, 시민들이 체험할 수 있는 다양한 참여형 행사로 준비했다"며 "마을공동체 활성화 기여에 도움이 되길 바란다"고 말했다.광양시 마을공동체지원센터 정회기</p>
Label text
<p>오는 17일 중마동 청소년문화센터에서 '2018광양시 마을 공동체한마당'행사를 개최하여 우수 공동체 사례를 발표하고 주민 참여형활동을 마련한다.</p>

- **Train: 290,000건**

- **Test: 10,000건**

모델 구축

모델 구축_T5 모델 소개(Text To Text Transfer Transformer)



: text를 입력하여 text로 출력해주는 모든 유형의 task 처리 가능

모델 구축_T5 특징

Architecture(구조)

1. Encoder-Decoder (NLP model)

- 문장 이해 및 생성 둘 다 가능
- 기존과 다른 Decoder 구조

2. Adapter layer

- 여러가지 task를 하나의 모델에서 가능하게 함

Training method(학습 방법)

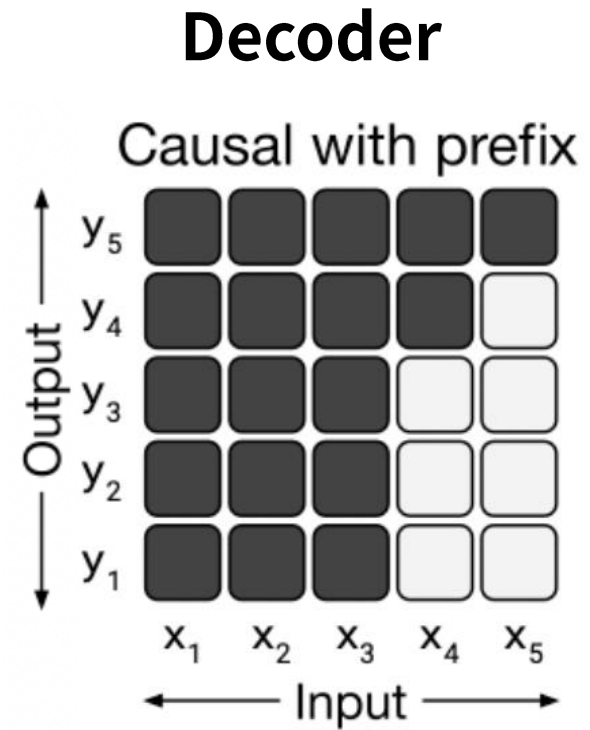
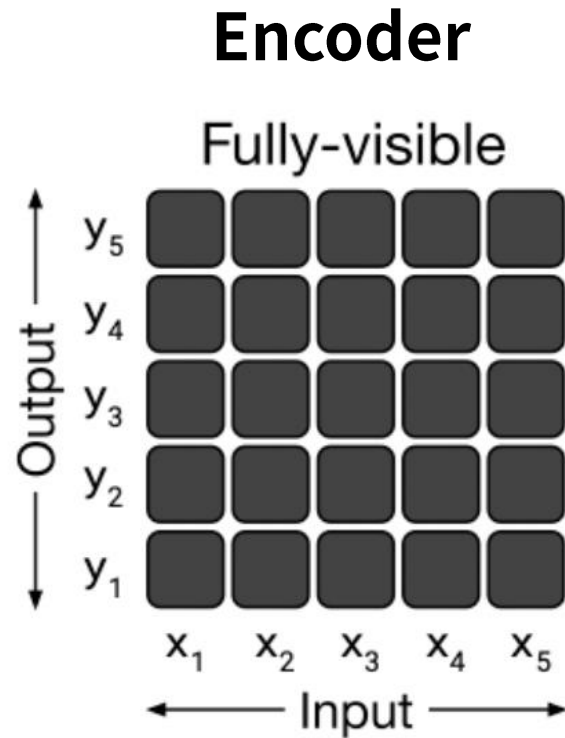
1. Multi-task pre-training

- 다양한 task 진행 가능하도록 학습

2. Denoising object

- 문장에 noise를 추가하여 학습

모델 구축_T5 특징(구조)



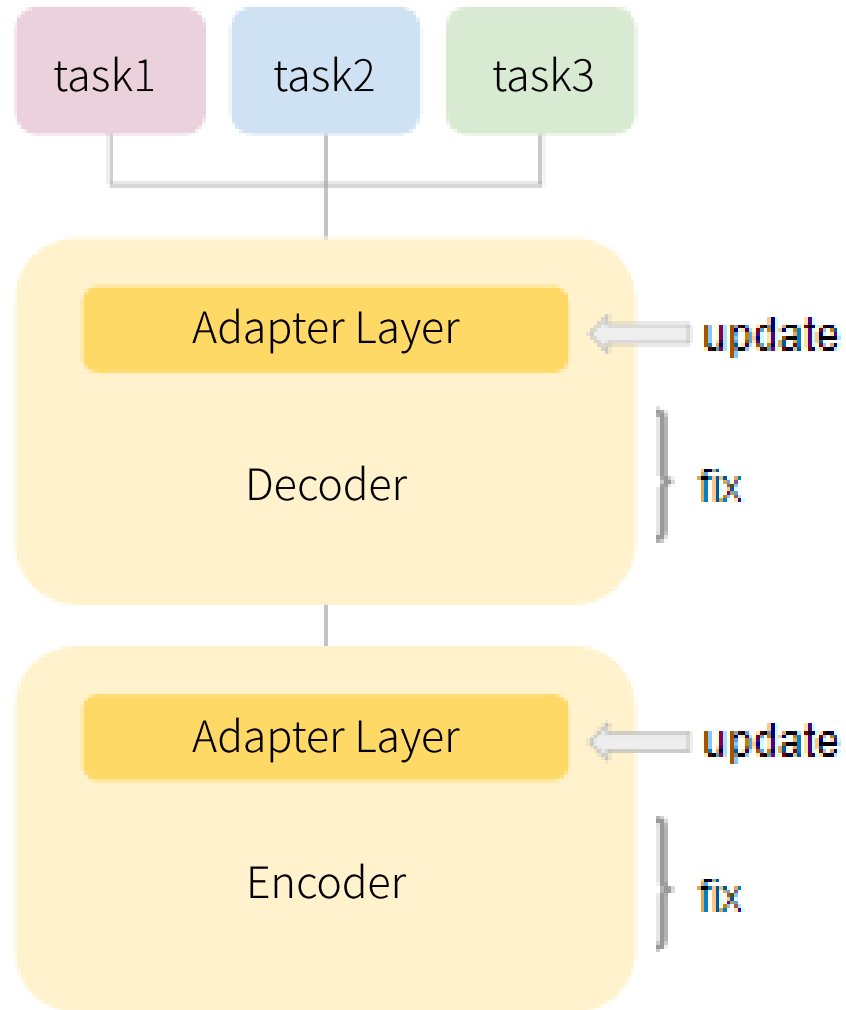
입력 문장

“Translate English to German : That is good.”
Prefix task(번역) text

출력 문장

“Das ist gut.”
text

모델 구축_T5 특징(구조 및 학습 방법)



- 입력 차원과 출력 차원 동일하게 함
- Loss, Hyper parameter, Weight 공유
- 하나의 신경망으로 여러 task 동시 수행

모델 구축_T5 특징(학습 방법)

Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

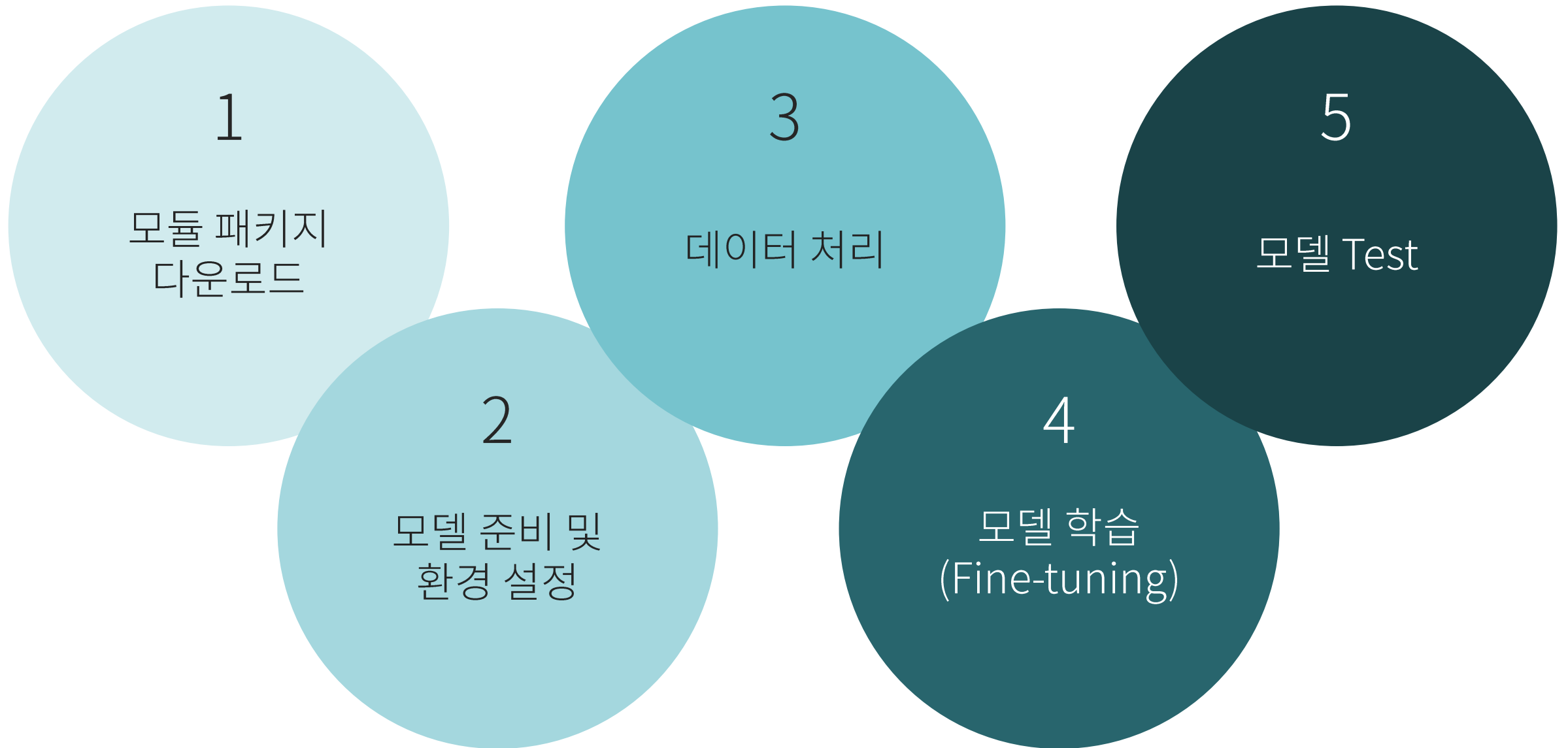
Targets

<X> for inviting <Y> last <Z>

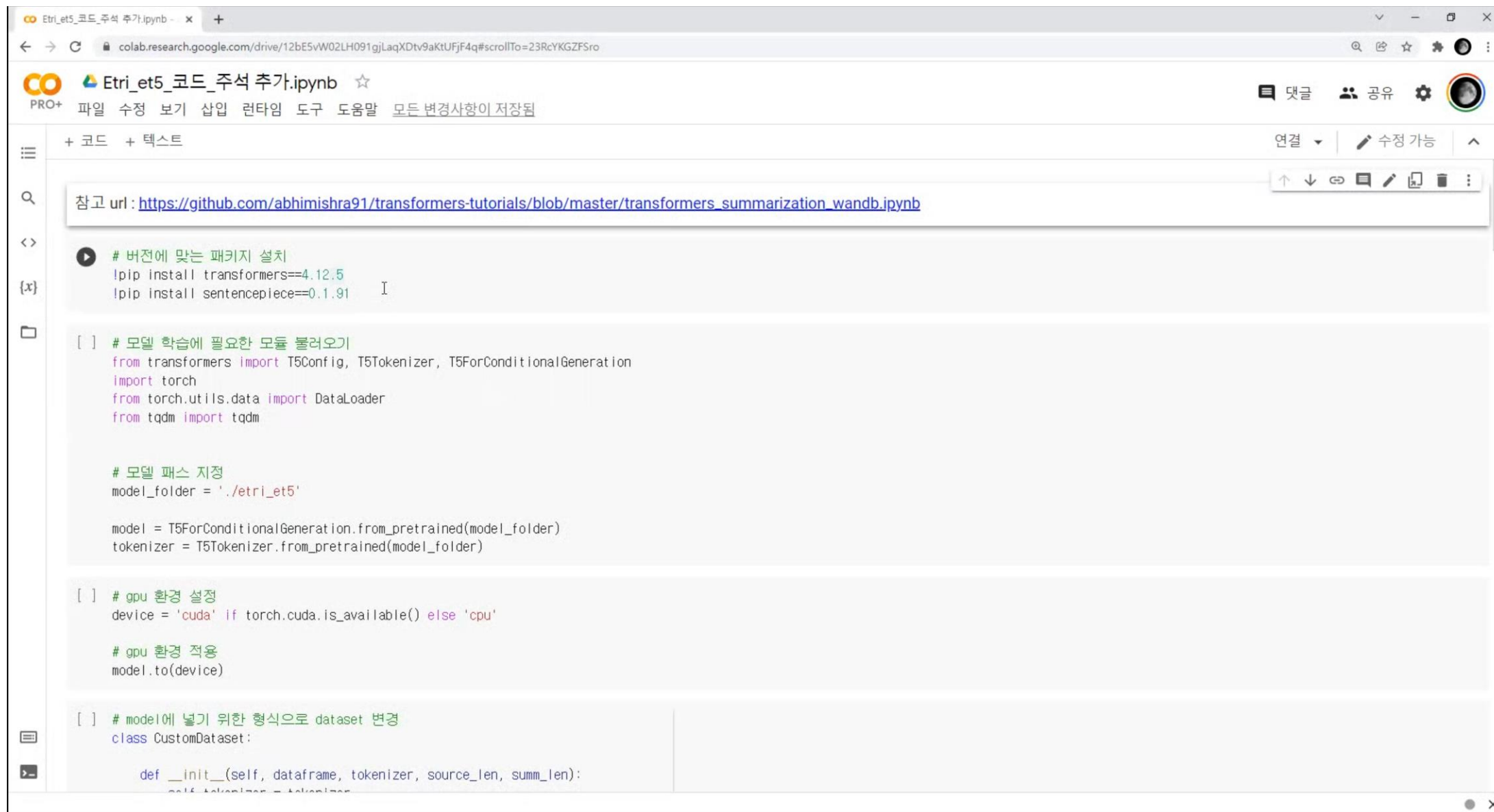
Bi-directional training

BERT-style (MLM objective) – Replace corrupted spans – 15%(rate) – 3(length)

모델 개발_구축 Process



모델 개발_구축 Process



The screenshot shows a Google Colab notebook interface. The title bar indicates the notebook is named 'Etri_et5_코드_주석 추가.ipynb'. The left sidebar shows a file explorer with a folder icon and a search icon. The main area displays the following code:

```
참고 url: https://github.com/abhimishra91/transformers-tutorials/blob/master/transformers\_summarization\_wandb.ipynb

# 버전에 맞는 패키지 설치
!pip install transformers==4.12.5
!pip install sentencepiece==0.1.91

[ ] # 모델 학습에 필요한 모듈 불러오기
from transformers import T5Config, T5Tokenizer, T5ForConditionalGeneration
import torch
from torch.utils.data import DataLoader
from tqdm import tqdm

# 모델 패스 지정
model_folder = './etri_et5'

model = T5ForConditionalGeneration.from_pretrained(model_folder)
tokenizer = T5Tokenizer.from_pretrained(model_folder)

[ ] # gpu 환경 설정
device = 'cuda' if torch.cuda.is_available() else 'cpu'

# gpu 환경 적용
model.to(device)

[ ] # model에 넣기 위한 형식으로 dataset 변경
class CustomDataset:

    def __init__(self, dataframe, tokenizer, source_len, summ_len):
        self.tokenizer = tokenizer
```

모델 개발_요약문 예시

기사 원문 (Original Text)	인피니트 장동우가 데뷔 첫 솔로 앨범 '바이(Bye)'의 트랙 리스트를 공개하며 컴백 카운트다운에 돌입했다. 장동우는 28일 밤 12시 공식 SNS를 통해 오는 3월 4일 발매되는 첫 번째 미니 앨범 '바이'의 트랙 리스트를 공개했다. 이번 미니 앨범의 타이틀곡은 '뉴스'로 인피니트의 히트곡 '텔미', '배드', '데스티니' 등을 작사, 작곡한 BLSSD의 곡이다. 앨범의 전체 분위기를 살펴볼 수 있는 1번 트랙 '아이 엠(I AM)'을 시작으로 타이틀곡 '뉴스'와 수록곡 '건(GUN)', '로미오(ROMEO)', '파티 걸(Party Girl)', '퍼펙트(PERFECT)', '썸띵 비트윈(Something Between)' 등이 담겼다. 앨범 명 '바이'는 올해 입대를 앞둔 장동우의 마지막 앨범이라는 의미 외에 'Beside You Every moment(모든 순간 너의 옆에)'라는 팬들을 향한 메시지가 포함됐다. 네이버 홈에서 '이데일리' 기사 보려면 [구독하기▶]꿀잼가득 [영상보기▶], 뽀침해소!청춘뉘우스~ [스냅타임▶]
참조 요약문 (label)	28일 인피니트 장동우는 공식 SNS에 3월 4일 발매되는 첫 번째 미니 앨범 '바이'의 트랙 리스트를 공개했는데 타이틀곡은 BLSSD가 작곡한 '뉴스'이며, 앨범 명은 그의 입대 전 마지막 앨범이라는 의미 외에 'Beside You Every moment'라는 팬들을 향한 메시지를 담았다.
생성 요약문 (generated text)	장동우는 28일 밤 12시 공식 SNS를 통해 오는 3월 4일 발매되는 각종 음원이 담긴 첫 번째 미니 앨범 '바이'의 트랙 리스트를 공개하였으며, 입대 전 마지막 앨범인 '바이'는 'Beside You Every moment(모든 순간 너의 옆에)'라는 팬들을 향한 메시지가 담겼다.

모델 개발_평가 지표 (ROUGE Score vs BERT Score)

ROUGE Score: n-gram에 기반하여 단어의 개수를 카운트

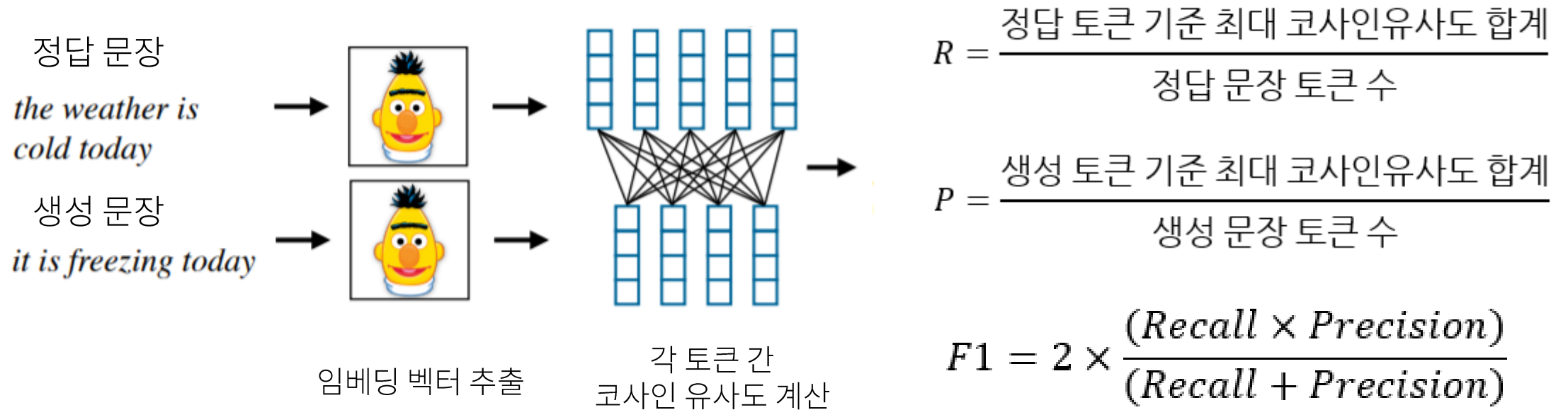
요약 모델의 성능 평가에 일반적으로 사용되나, **생성된 요약문과 정답 간의 ‘겹치는 정도’로 문장을 평가**

→ 제대로 된 평가가 어렵다는 단점




정답 문장	나는 사과를 먹는다	ROUGE-1 (F1)	ROUGE-2 (F1)	ROUGE-L (F1)	BERT Score (F1)
생성 문장1	나는 사과를 먹고 있다	0.5714	0.4	0.5714	0.9630
생성 문장2	나는 사과를 한다	0.6667	0.5	0.6667	0.9556

모델 개발_평가 지표 (BERT Score)

- BERT Score: 의미적 유사성을 고려한 문장 단위의 평가 방법
- 사전훈련된 BERT 모델을 통해 두 문장의 임베딩 벡터 간의 코사인 유사도를 계산하여 재현율(Recall), 정밀도(Precision), F1-score 를 산출
- 한국어 문장 유사도 평가 성능이 우수한 KoELECTRA 모델로 BERT Score 계산



모델 개발_T5 모델 비교

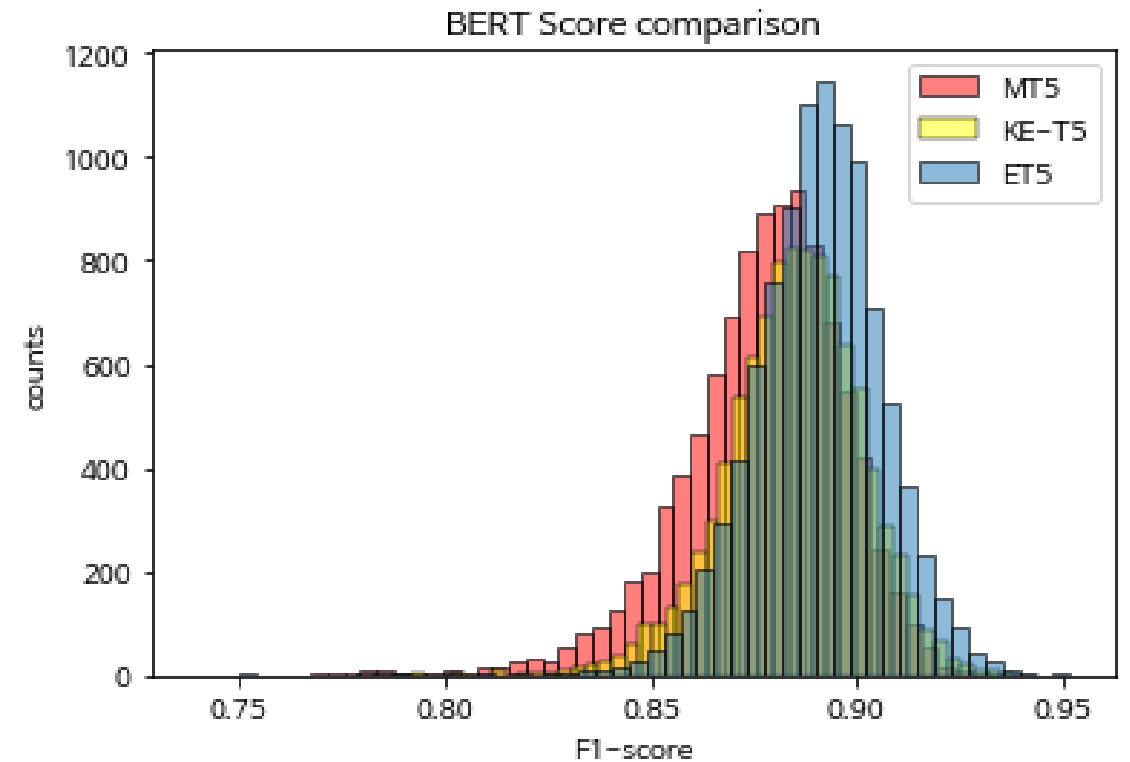
	MT5	ET5	KE-T5
	 Google AI		 한국전자기술연구원 Korea Electronics Technology Institute
Corpus	26TB(101개 언어) [Ko 100GB]	136GB	93GB
Vocab Size	250,112	45,100	64,000 (한국어:영어=7:3)
Feedforward layers dimension	2,048	3,072	2,048

ET5는 세 모델 중에서 한국어 말뭉치가 많고, 은닉층이 깊음

모델 개발_T5 모델 성능 비교

	ET5	MT5	KE-T5
훈련 데이터	24만 건	24만 건	24만 건
epochs	3	3	3
BERT Score(P)	0.9190	0.9144	0.9175
BERT Score(R)	0.8641	0.8428	0.8545
BERT Score(F)	0.8906	0.8770	0.8848
10자 미만 문장 생성 비율	0.0008	0.0077	0.0002

- 참조 문장: 기사 원문
- 비교 문장: 생성된 요약 문장
- 테스트 데이터셋 크기: 10,000건

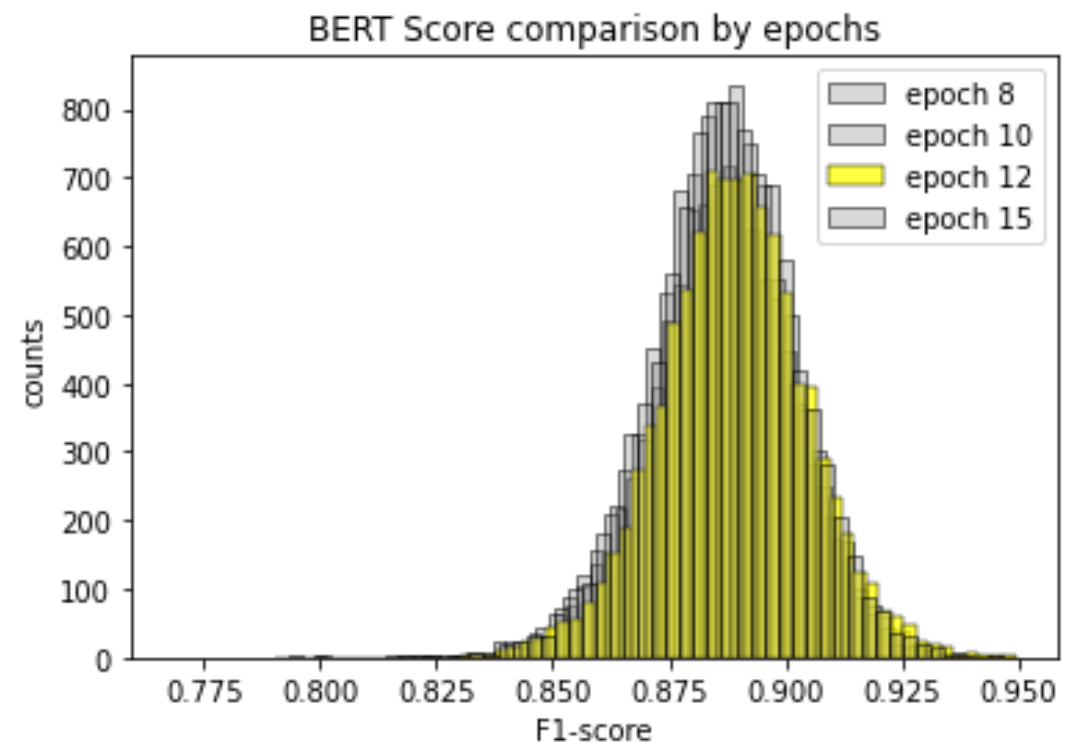


BERT Score가 가장 높은 ET5 선택

모델 개발_Epochs 변화에 따른 성능 비교

	ET5	ET5	ET5	ET5
훈련 데이터	29만 건	29만 건	29만 건	29만 건
epochs	8	10	12	15
BERT Score(P)	0.9138	0.9135	0.9156	0.9135
BERT Score(R)	0.8605	0.8607	0.8638	0.8617
BERT Score(F)	0.8863	0.8862	0.8889	0.8868
10자 미만 문장 생성 비율	0.0003	0.0001	0	0.0007

- 참조 문장: 기사 원문
- 비교 문장: 생성된 요약 문장
- 테스트 데이터셋 크기: 10,000건



BERT Score가 가장 높고, 10자 미만 문장 생성 건 수가 적은 epochs 12 model 선택

모델 개발_모델 구축 환경 한계

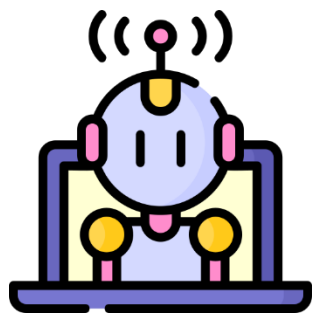


고성능 GPU 필요 → colab 사용

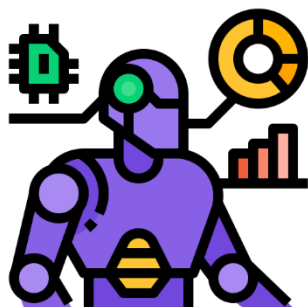
1. colab 런타임 24시간 제한
2. batch size 설정 2가 최대
3. 데이터 10만 건 epochs 30이 최대

모델 개발_추가 학습

Fine-tuning을 진행한 모델에 다른 데이터로 다시 학습 진행



Fine-tuning 끝난 Model



X N번

Data 계속 추가 학습

모델 패스 지정

```
model_folder = './etri_et5'
```

model, tokenizer 저장

```
tokenizer.save_pretrained('./fine-tuned/')
```

```
model.save_pretrained('./fine-tuned/')
```



모델 패스 지정 후 추가 학습

```
model_folder = './fine-tuned/'
```

학습 완료한 모델 경로로
모델 패스 지정 후 다시 학습

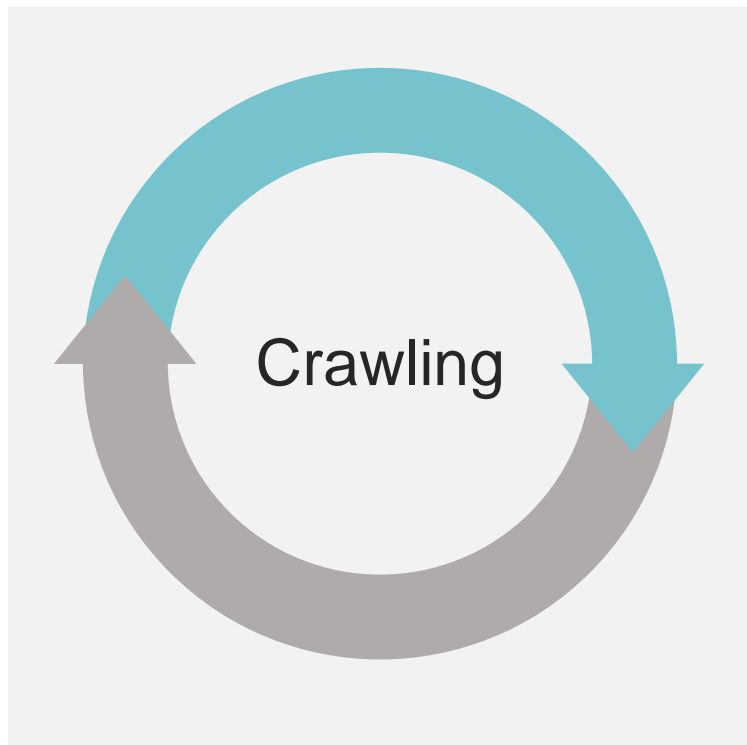
모델 개발_추가 학습

	(1) ET5	(2) ET5	(3) ET5	(4) ET5
데이터 크기	10,000건	20,000건	30,000건	30,000건
epochs	15	15	15	15
학습 방법	한번에 학습	(1) 학습 결과에 10,000건 추가 학습	(2) 학습 결과에 10,000건 추가 학습	한번에 학습
BERT Score(P)	0.7091	0.7113	0.7074	0.7085
BERT Score(R)	0.6335	0.6299	0.6337	0.6341
BERT Score(F)	0.6670	0.6678	0.6683	0.6690

2번 추가 학습한 (3) 과 한번에 학습한 (4)의 성능은 추가 학습한 모델이 약간의 성능 저하가 있음.
다만 그 차이가 미미하고, 데이터를 학습할 수록 성능이 향상되어 추가 학습 진행.

서비스 배포

서비스 배포



1. 포털 뉴스 기사 크롤링

ETRI_ET5

2. 모델을 통한 요약



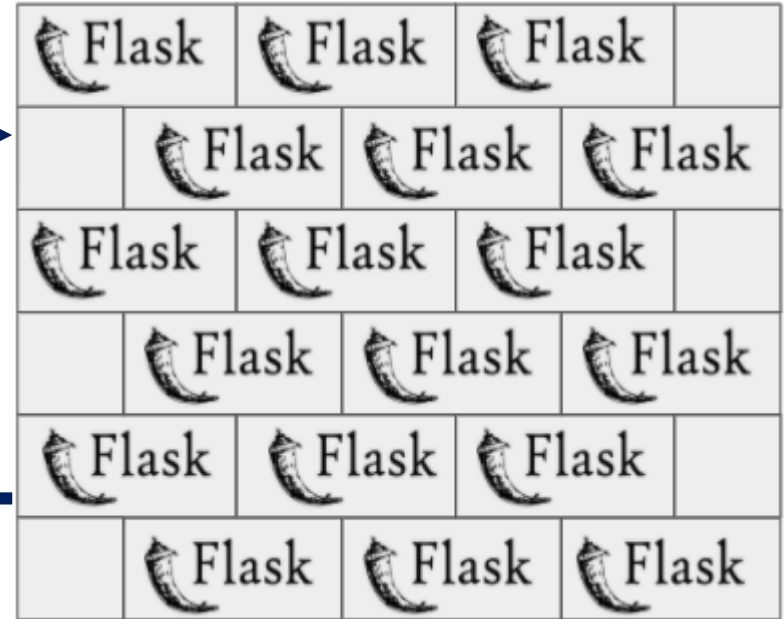
3. 챗봇을 통한 서비스

서비스 배포_챗봇 프로세스



[Chatbot]

[요청]

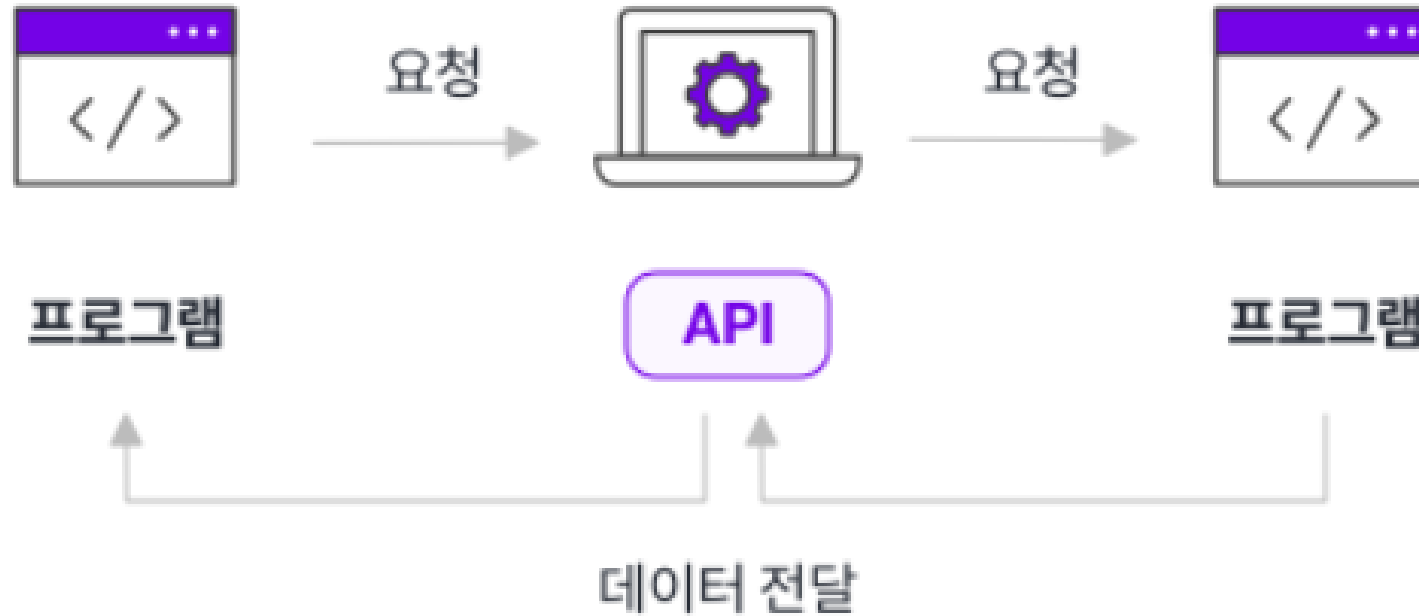


[응답]



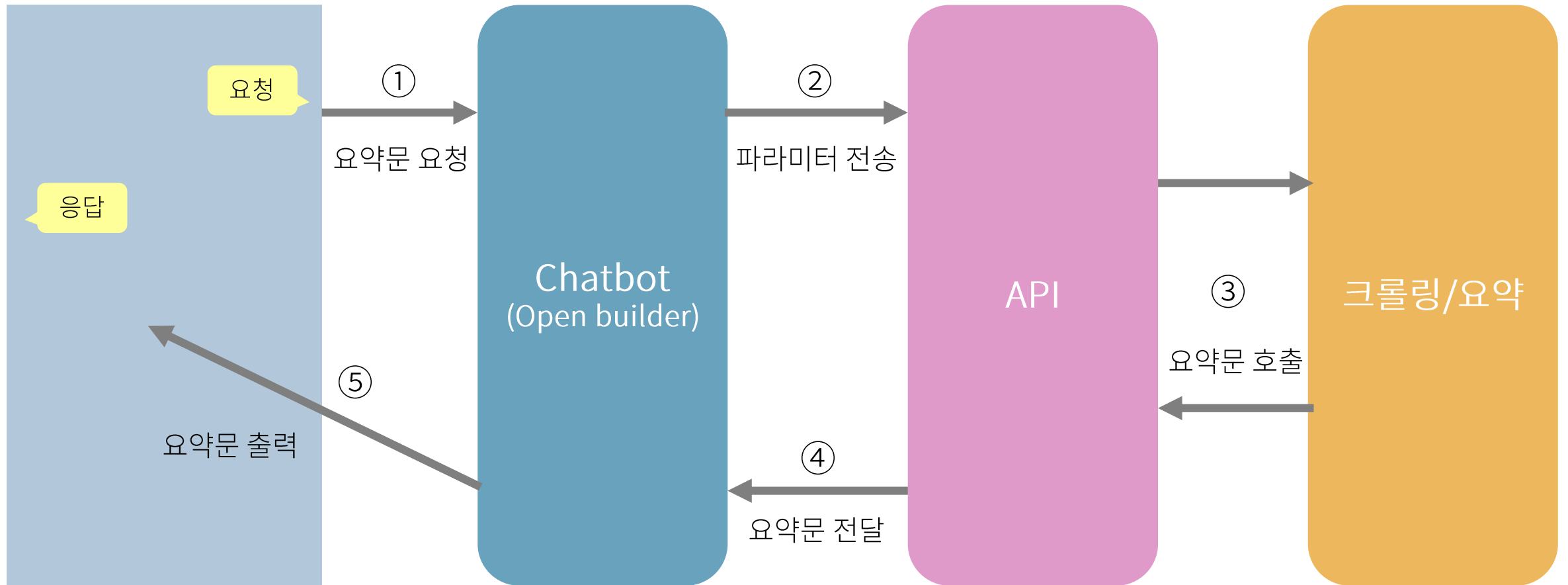
[API]

서비스 배포_API 역할

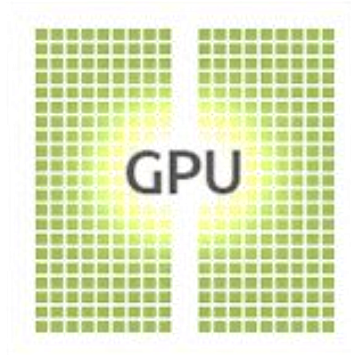
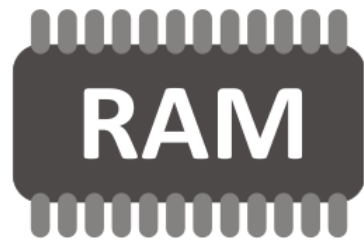


API = 모델요약 프로그램과 챗봇을 연결해 주는 역할

서비스 배포_전체 Process

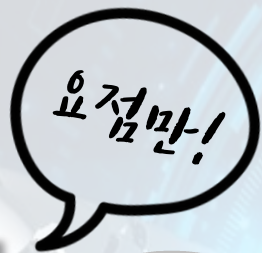


서비스 배포_서비스 구축 환경 한계



- 고용량 RAM & 고성능 GPU가 필요
- 성능 좋지 않을 시 서버 과부하 오류

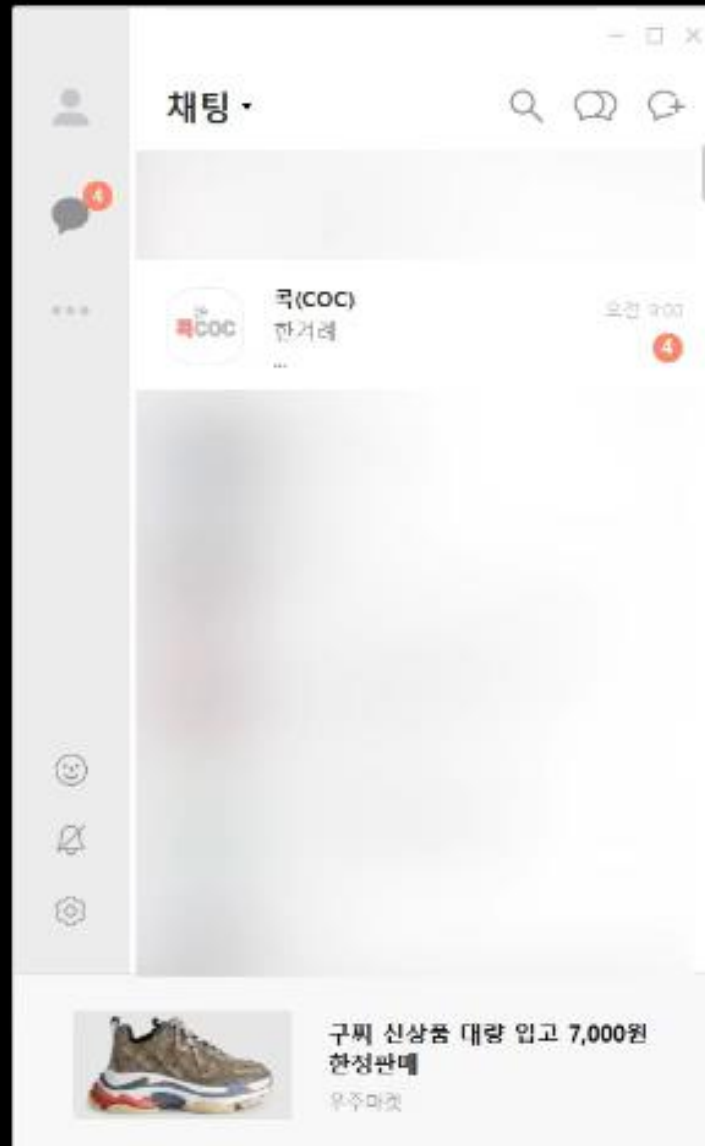
- ICT COC의 고성능 GPU를 탑재한 워크스테이션을 대여



coc

Crush On Context

서비스 배포_서비스 구현 영상





End of Document.



GitHub:
<https://github.com/Text-abstractive-summarization>