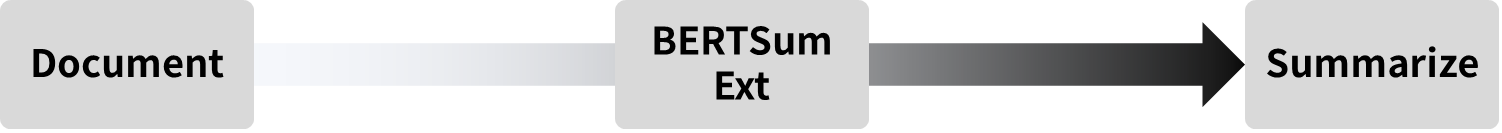
BERT based 문서요약 기술 개발

# 프로젝트 개요

장문의 텍스트 문서를 input으로, 해당 문서의 핵심 내용을 요약한 결과를 output으로 하는 BERT 기반의 문서요약 기술 개발

문서요약 기술에는 크게 생성요약(Abstractive summary)와 추출요약(Extractive summary)가 있으며 본 프로젝트는 추출요약에서 좋은 성능을 보인 BERTSum 모델의 베이스라인을 따라 진행되었으며, ETRI(<https://aiopen.etri.re.kr/service_dataset.php>)에서 공개한 KoBERT pre-trained model을 적용

# 활용 데이터

모델 학습 및 검증는 Aihub에서 공개한 문서요약 텍스트 데이터(<https://aihub.or.kr/aidata/8054>)를 활용하였음

텍스트이(가) 표시된 사진

자동 생성된 설명

신문기사, 기고문 잡지, 법률데이터 중 신문기사 데이터만을 활용하여 모델 학습 및 검증

# 데이터 개요

데이터는 json 포멧으로 제공되며 학습에 필요한 컬럼은 뉴스기사 본문(text), 중요한 문장 인덱스(extractive), 추출된 중요 문장(abstractive) 임

텍스트, 스크린샷, 모니터, 실내이(가) 표시된 사진

자동 생성된 설명

# 데이터 전처리

## Text 컬럼 전처리

텍스트이(가) 표시된 사진

자동 생성된 설명text컬럼은 리스트로 구성된 딕셔너리 구조로 되어있어 모델의 입력이 되는 문서 형태로 변환할 필요가 있음

train\_df = df[['id', 'category', 'media\_type', 'media\_sub\_type', 'media\_name', 'size',

       'char\_count', 'publish\_date', 'title', 'annotator\_id',

       'extractive', 'abstractive']]

for idx, row in df.iterrows():

  news\_dict\_list = row['text']

  sentence = ''

  temp\_list = list()

  for temp\_dict in news\_dict\_list:

    try:

      temp\_list.append(temp\_dict[0]['sentence'])

    except:

      pass

  train\_df.loc[idx, 'sentence'] = str(temp\_list)

## 형태소 분석 컬럼 추가

형태소 기반 모델에 적용하기 위해 형태소 분석(mecab 활용) 적용

import ast

test = train\_df

test['morphs'] = 0

for idx, row in test.iterrows():

  arr = []

  for SubSentence in ast.literal\_eval(row['article\_original']):

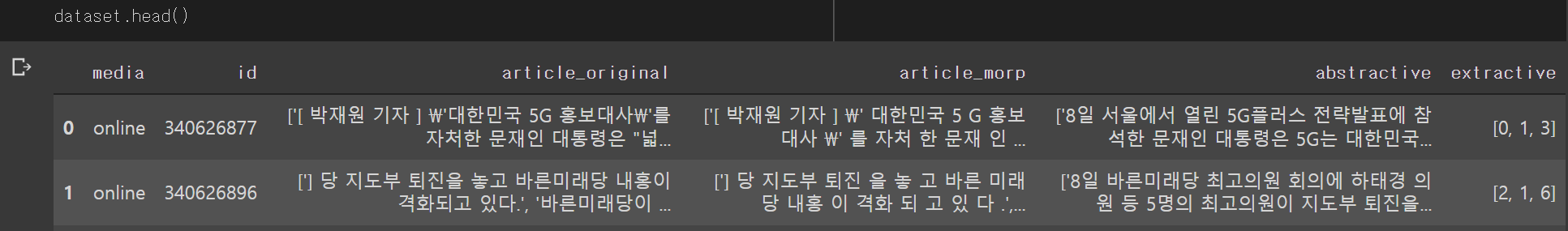
    # print(SubSentence)

    arr.append(' '.join(mecab.morphs(SubSentence)))

  # print(arr)

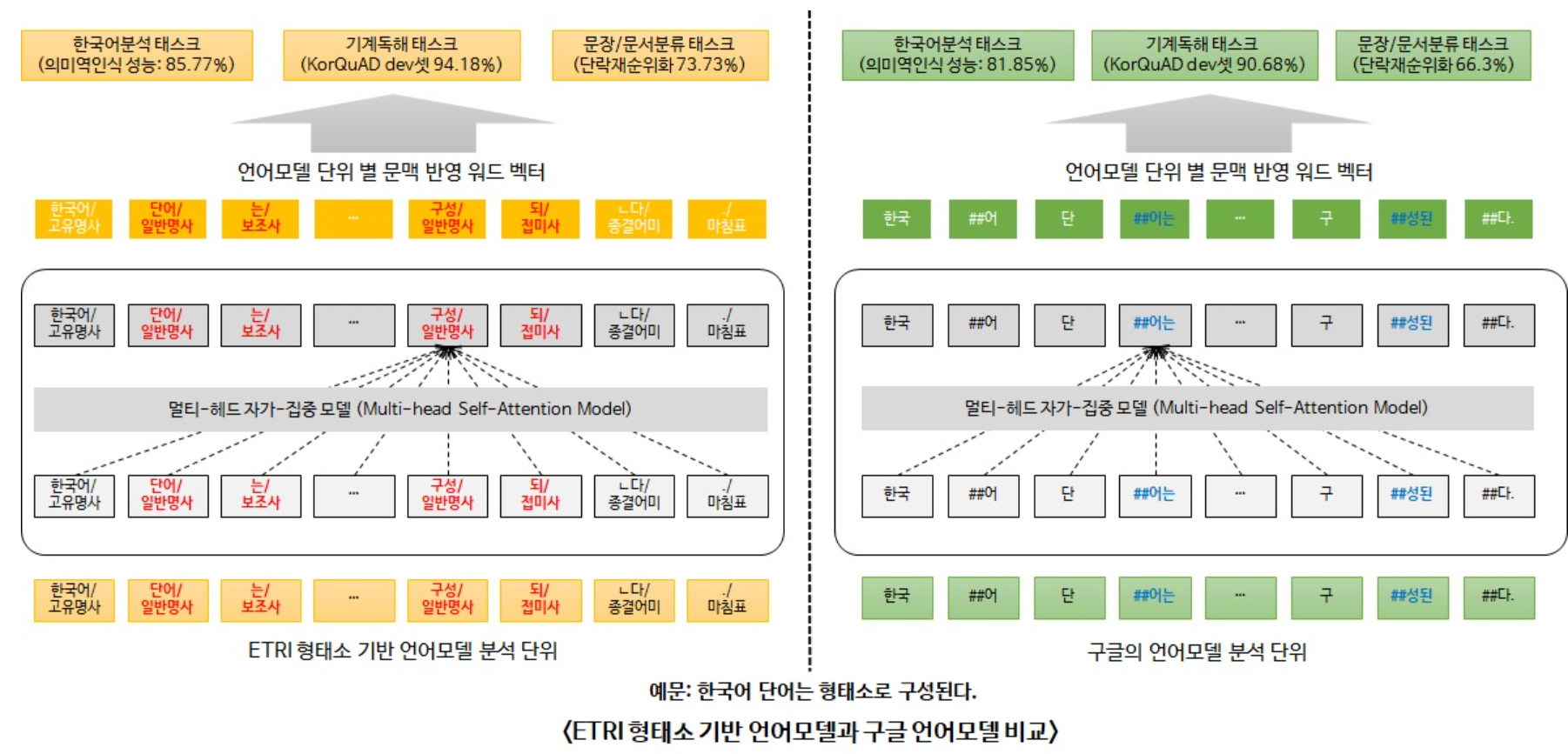
  test.loc[idx, 'morphs'] = str(arr)

test.to\_csv('/content/drive/MyDrive/Colab\_Notebooks/kobertsum/KorBertSum-master/data/output.csv')



# 모델 및 평가 지표

ETRI에서 제공하는 KoBERT는 과학기술정보통신부와 IITP의 프로젝트로 추진한 사업에서 한국어의 특성을 반영하여 개발한 BERT 모델로 의미역 인식, 기개독해, 단락 순위화, 문장 유사도 추론, 문서 주제분류 5종의 한국어 처리 태스크에서 구글이 배포한 모델보다 평균 4.5% 우수한 성능을 보였음



텍스트이(가) 표시된 사진

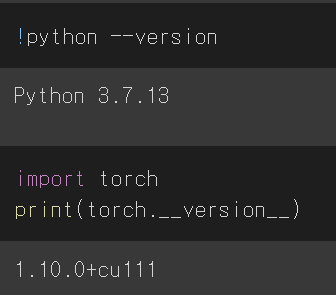
자동 생성된 설명

높은 성능으로 인해 문서 요약 태스크에도 본 모델을 활용하였으며, 기본적으로 pytorch 모델로 제공하기 때문에 torch Framework 사용

문서요약 모델의 평가지표로는 ROUGE Metric을 적용

# 학습환경

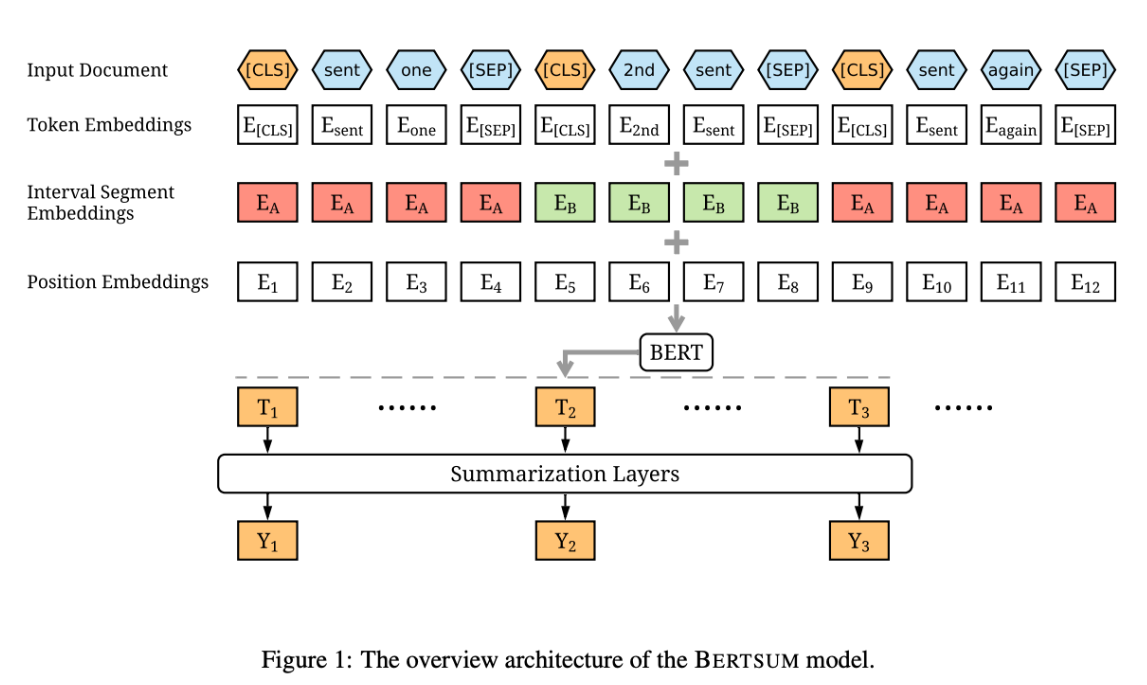
메모리 부족으로 인해 Colab PRO 환경에서 학습 진행하였으며 파이썬과 파이토치 버전은 아래와 같음.



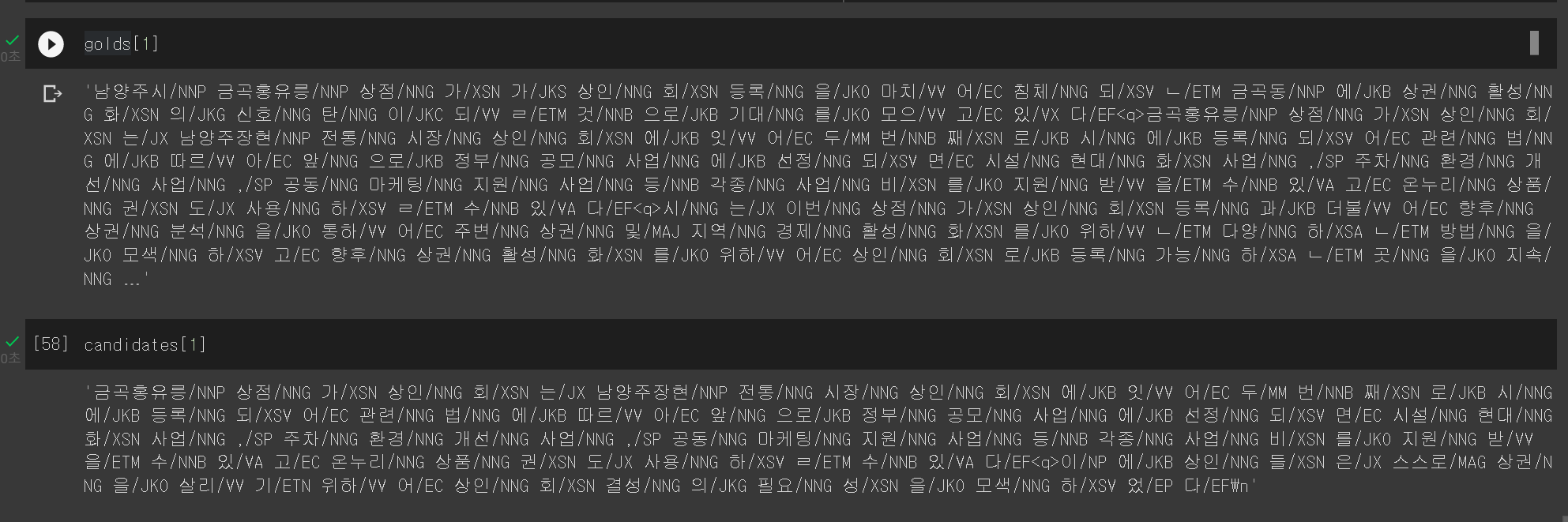
# 학습 및 결과

Pre-trained 모델에 Summarization layer를 추가하여 학습

BERTSum 논문(Fine-tune BERT for Extractive Summarization )에서는 Simple Classifier / Transformer / RNN 세 가지 encoder를 사용하였으며, 본 프로젝트에서는 단순 분류 레이어를 적용함.



## 결과



## 본문(처리 후)

남양주시 금곡홍유릉상점가가 상인회등록을 마치어 침체된 금곡동에 상권 활성화의 신호탄이 되ㄹ것으로 기대를 모으고 있다.

상점가상인회는 남양주 장현전통시장상인회에 잇어 두번째로 시에 등록되어 관련법에 따르아 앞으로 정부공모사업에 선정되면 시설현대화사업,주차환경개선사업,공동마케팅지원사업 등

각종사업비를 지원 받을수 있고 온누리 상품권도 사용하ㄹ수 있다는 이번 상점가상인회등록과 더불어 향후상권분석을 통하어 주변상권 및 지역경제 활성화를 위하ㄴ

다양하ㄴ 방법을 모색하고 향후 상권활성화를 위하어 상인회로 등록가능하ㄴ 곳을 지속적으로 발굴하ㄹ계획이다

## 추출요약 결과(처리 후)

금곡홍유릉상점가상인회는 남양주 장현전통시장상인회에 잇어 두번째로 시에 등록되어 관련법에 따르아

앞으로 정부공모사업에 선정되면 시설현대화사업,주차환경개선사업,공동마케팅지원사업 등

각종사업비를 지원 받을수있고 온누리 상품권도 사용하ㄹ수있다 에상인들은 스스로 상권을 살리기 위하어 상인회결성의 필요성을 모색하었다

참조

<https://github.com/nlpyang/BertSum>

<https://github.com/raqoon886/KorBertSum>

<https://github.com/kaniblu/hangul-utils>

프로젝트 노트북 깃허브 주소 : <https://github.com/simsimhae91/Seoul_ICT>