DSBA

News Summary (feat. NLP)

Mini_Project
2023.11

4조 : 박종우, 박치수, 서형준
성문주 , 염규민, 장성웅

발표자 : 성문주

INDEX

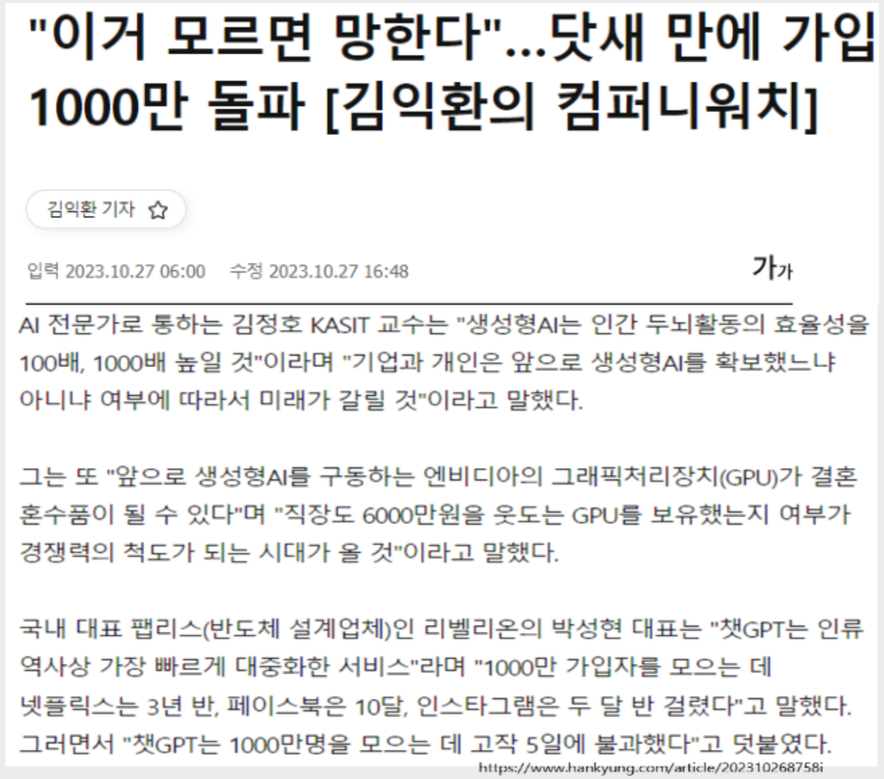
News Summary

목표 및 선정배경

- 뉴스의 핵심 요약 → 시간 절약, 방대한 양의 핵심 데이터 습득
- 한 두줄 내외의 요약문을 생성

News Summary

ex. 네이버 뉴스 요약봇



"이거 모르면 망한다"...닷새 만에 가입 1000만 돌파 [김익환의 컴퍼니워치]

김익환 기자

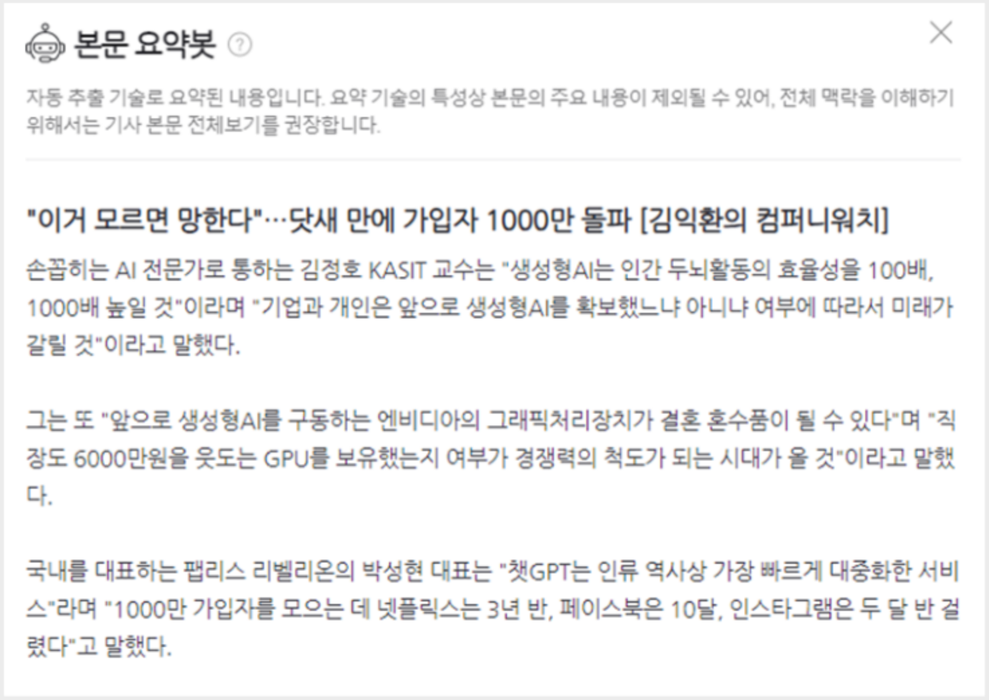
입력 2023.10.27 06:00 수정 2023.10.27 16:48 가가

AI 전문가로 통하는 김정호 KASIT 교수는 "생성형AI는 인간 두뇌활동의 효율성을 100배, 1000배 높일 것"이라며 "기업과 개인은 앞으로 생성형AI를 확보했느냐 아니냐 여부에 따라서 미래가 갈릴 것"이라고 말했다.

그는 또 "앞으로 생성형AI를 구동하는 엔비디아의 그래픽처리장치(GPU)가 결혼 혼수품이 될 수 있다"며 "직장도 6000만원을 웃도는 GPU를 보유했는지 여부가 경쟁력의 척도가 되는 시대가 올 것"이라고 말했다.

국내 대표 팹리스(반도체 설계업체)인 리벨리온의 박성현 대표는 "챗GPT는 인류 역사상 가장 빠르게 대중화한 서비스"라며 "1000만 가입자를 모으는 데 넷플릭스는 3년 반, 페이스북은 10달, 인스타그램은 두 달 반 걸렸다"고 말했다. 그러면서 "챗GPT는 1000만명을 모으는 데 고작 5일에 불과했다"고 덧붙였다.

https://www.hankyung.com/article/202310268758i



본문 요약봇

자동 추출 기술로 요약된 내용입니다. 요약 기술의 특성상 본문의 주요 내용이 제외될 수 있어, 전체 맥락을 이해하기 위해서는 기사 본문 전체보기를 권장합니다.

"이거 모르면 망한다"…닷새 만에 가입자 1000만 돌파 [김익환의 컴퍼니워치]

손꼽히는 AI 전문가로 통하는 김정호 KASIT 교수는 "생성형AI는 인간 두뇌활동의 효율성을 100배, 1000배 높일 것"이라며 "기업과 개인은 앞으로 생성형AI를 확보했느냐 아니냐 여부에 따라서 미래가 갈릴 것"이라고 말했다.

그는 또 "앞으로 생성형AI를 구동하는 엔비디아의 그래픽처리장치가 결혼 혼수품이 될 수 있다"며 "직장도 6000만원을 웃도는 GPU를 보유했는지 여부가 경쟁력의 척도가 되는 시대가 올 것"이라고 말했다.

국내를 대표하는 팹리스 리벨리온의 박성현 대표는 "챗GPT는 인류 역사상 가장 빠르게 대중화한 서비스"라며 "1000만 가입자를 모으는 데 넷플릭스는 3년 반, 페이스북은 10달, 인스타그램은 두 달 반 걸렸다"고 말했다.

https://n.news.naver.com/mnews/hotissue/article/015/0004907285?type=series&cid=1088958

DATA

- From AI - HUB : 요약문 및 레포트 생성 데이터

News Summary

Data Structure

- Training (17,300 files)
  - 라벨링
    - ■news.json
  - 원천
    - ■news.json
- Validation (4,300 files)
  - 라벨링
    - ■news.json
  - 원천
    - ■news.json

ex) passage - 원문, summary1 - 요약문


```
{
    "Meta(Acqusition)": {
        "doc_id": "REPORT-news_r-00089",
        "doc_category": "REPORT",
        "doc_type": "news_r",
        "doc_name": "[분양 포커스] 원석에서 보석으로 탈바꿈 '금관구'...트리플 프리미엄 중소형 아파트",
        "author": "채호연",
        "publisher": "중앙일보",
        "publisher_year": "2020",
        "doc_origin": "중앙일보"
    "Meta(Refine)": {
        "passage_id": "REPORT-news_r-00089-00001",
        "passage": "서울 서남권 부동산시장를 대표하는 영등포구와 이른바 '금·관·구'(이하 금관구)라고 불리
        "passage_Cnt": 1250
    },
    "Annotation": {
        "summary1": "금관구와 영등포구 일대는 굵직한 개발호재가 많아 투자자들에게 주목받고 있다.",
        "summary2": "서울 서남권 부동산시장를 대표하는 영등포구와 이른바 '금·관·구'(이하 금관구)라고 불리
        "summary3": null,
        "summary_3_cnt": null
```

Method

- 요약모델(자체), Pretrained_1(KoBART), Pretrained_2(Bert)

요약모델	Pretrained_1	Pretrained_2
• Seq2Seq + Attention 모델 • 전처리, train, test	• KoBART ○ ainize/kobart-news • test	• Bert ○ bert-base-uncased • test

- KoBART : SKT에서 공개한 한국어 BART 모델. 약 40GB 이상의 한국어 텍스트를 학습한 한국어 언어 모델.
  - https://github.com/SKT-AI/KoBART
- Bert : 구글에서 개발한 NLP 모델. 위키피디아(25억 단어)와 BooksCorpus(8억 단어)와 같은 레이블이 없는 텍스트 데이터로 사전 훈련된 언어 모델.

Train & Test

- 전처리과정
  - 요약모델에 한해 진행
- "딥러닝을 이용한 자연어 처리 입문"

News Summary

1. 불용어 제거

```
!pip install konlpy
from tqdm import tqdm
from konlpy.tag import Okt

stopwords = ['의','가','이','은','들','는','좀','잘','걍','과','도','를','으로','자','에','와','한','하다','을','이다','다']

def news_preprocessing(sentence):
    okt = Okt()
    tokenized_sentence = okt.morphs(re.sub(r'[^0-9가-힣Ws]', '', sentence), stem=True) # 토큰화
    stopwords_removed_sentence = [word for word in tokenized_sentence if not word in stopwords] # 불용어 제거
    result = ' '.join(stopwords_removed_sentence)
    return result

y_train=[]

for sentence in tqdm(data['summary1']):
  stopwords_removed_sentence = news_preprocessing(sentence)
  y_train.append(stopwords_removed_sentence)
```

2. 데이터 분리

```
print('훈련 데이터의 개수 :', len(encoder_input_train))
print('훈련 레이블의 개수 :',len(decoder_input_train))
print('테스트 데이터의 개수 :',len(encoder_input_test))
print('테스트 레이블의 개수 :',len(decoder_input_test))
```

```
훈련 데이터의 개수 : 17213
훈련 레이블의 개수 : 17213
테스트 데이터의 개수 : 4303
테스트 레이블의 개수 : 4303
```

3. 토큰화 & 정수 인코딩

```
src_vocab = 28251
src_tokenizer = Tokenizer(num_words = src_vocab)
src_tokenizer.fit_on_texts(encoder_input_train)

# 텍스트 시퀀스를 정수 시퀀스로 변환
encoder_input_train = src_tokenizer.texts_to_sequences(encoder_input_train)
encoder_input_test = src_tokenizer.texts_to_sequences(encoder_input_test)
```

Train & Test

News Summary

```
Layer (type)                 Output Shape              Param #   Connected to
==================================================================================================
input_1 (InputLayer)         [(None, 350)]             0         []

embedding (Embedding)        (None, 350, 128)          3616128   ['input_1[0][0]']

lstm (LSTM)                  [(None, 350, 256),        394240    ['embedding[0][0]']
                              (None, 256),
                              (None, 256)]

input_2 (InputLayer)         [(None, None)]            0         []

lstm_1 (LSTM)                [(None, 350, 256),        525312    ['lstm[0][0]']
                              (None, 256),
                              (None, 256)]

embedding_1 (Embedding)      (None, None, 128)         928384    ['input_2[0][0]']

lstm_2 (LSTM)                [(None, 350, 256),        525312    ['lstm_1[0][0]']
                              (None, 256),
                              (None, 256)]

lstm_3 (LSTM)                [(None, None, 256),       394240    ['embedding_1[0][0]',
                              (None, 256),                       'lstm_2[0][1]',
                              (None, 256)]                       'lstm_2[0][2]']

attention_layer (Attention   ((None, None, 256),       131328    ['lstm_2[0][0]',
Layer)                        (None, None, 350))                 'lstm_3[0][0]']

concat_layer (Concatenate)   (None, None, 512)         0         ['lstm_3[0][0]',
                                                                  'attention_layer[0][0]']

dense_1 (Dense)              (None, None, 7253)        3720789   ['concat_layer[0][0]']

==================================================================================================
```

- 생성된 요약 모델의 구조
- Encoder, Decoder 모델 + Attention Layer

Train & Test

- Pretrained_1,2

News Summary

```
from transformers import PreTrainedTokenizerFast, BartForConditionalGeneration
#  Load Model and Tokenize
tokenizer = PreTrainedTokenizerFast.from_pretrained("ainize/kobart-news")
model = BartForConditionalGeneration.from_pretrained("ainize/kobart-news")
# Encode Input Text
input_text = """
앵커] 국민의힘이 어제(16일) 주요 당직을 새롭게 인선했지만, 논란은 이어지고 있다고요.
국회 나가 있는 취재기자 연결합니다. 차승은 기자.[기자] 네, 국회입니다.이른바 '김기현 체제 2기' 지도부는 오늘 오늘 아침 국정감사 대책회의로 첫 공식 활동을 시작했는데요.
이만희 신임 사무총장은 "당이 변해야 한다는 민심의 준비를 겸허히 받들겠다", 유의동 정책위의장은 "이번 선거에서 보내준 민심의 경고를 외면하지 않겠다"고 밝혔습니다.
하지만 당대표와 원내대표에 이어 공천 실무를 맡는 사무총장까지 모두 영남권 의원이 차지한 점, 또 과연 대통령실과의 수직적 관계를 개선할 수 있을지에 대해 우려의 목소리는 계속되고 있습니다.
가령, 대표적인 비윤 정치인 유승민 전 의원은 CBS 라디오에 나와, "그동안 대통령의 실정에 한 마디도 못하다가 앞으로 바뀌겠다고 하면 국민이 어떻게 평가하겠느냐"며 김기현 대표 퇴진을 주장하기도 했습니다.
다만, 윤재옥 원내대표는 조금 전 기자들과 만나 용산과 당의 소통에 대해 "비춰지는 모습이 국민들 눈높이에 안 맞는 점이 있다면 고치려 노력하겠다"고 했습니다.또 사무총장직엔 지역 안배를 위해 영남권 의원을 앉힌 것이라며 "애를 썼지만,
현실적으로 적합한 인물을 찾는 데 어려움이 있었다"고 설명했습니다. [앵커] 국회 국정감사 8일째인 오늘은 12개 상임위에서 감사가 예정돼 있습니다.
법사위와 행안위의 피감기관이 각각 검찰청과 경기도인 만큼 이재명 더불어민주당 대표의 사법리스크를 두고 여야는 팽팽히 맞설 것으로 보이는데요.
[기자] 네, 그렇습니다.잠시 후인 오전 10시, 법사위는 서울 서초구 서울고검에서 서울중앙지검과 수원지검 등에 대한 국정감사를 진행합니다.
앞서 서울중앙지검은 이재명 더불어민주당 대표를 백현동 개발 특혜 의혹과 위증교사 혐의로 기소했죠. 수원지검은 이 대표의 쌍방울 대북 송금 의혹과 이 대표 배우자인 김혜경 씨의 법인카드 사적 유용 의혹을 수사 중인데요.
민주당이 그동안 검찰 수사를 두고 야당 탄압이다, 인권 침해다 비판해 온 만큼 여야는 격렬히 대립할 것으로 관측됩니다.경기도가 피감 대상인 행안위도 또 다른 격전지인데요.
여야는 이 대표의 경기도지사 시절 법인카드 유용 의혹과 대북 협력 사업 지원 의혹 등을 두고 맞붙을 것으로 보입니다.이밖에도 여야는 국토위에서는 서울-양평 고속도로 노선 변경 의혹, 과방위에서는 언론 장악 의혹 등에서 난타전을 이어갈 전망입니다.
지금까지 국회에서 전해드렸습니다. (chaletuno@yna.co.kr)연합뉴스TV 기사문의 및 제보 : 카톡/라인 jebo23
"""
input_ids = tokenizer.encode(article_text, return_tensors="pt")

# Generate Summary Text Ids
summary_text_ids = model.generate(
    input_ids=input_ids,
    bos_token_id=model.config.bos_token_id,
    eos_token_id=model.config.eos_token_id,
    length_penalty=0.1,
    max_length=142,
    min_length=56,
    num_beams=4,
)
# Decoding Text
print(tokenizer.decode(summary_text_ids[0], skip_special_tokens=True))
```

Input : 원문

Output : 요약문

```
You passed along `num_labels=3` with an incompatible id to label map: {'0': 'NEGATIVE', '1': 'POSITIVE'}. The number of labels will be overwritten to 2.
유승민 전 국민의힘 의원은 17일 시비에스 라디오 '김현정의 뉴스쇼'에서 박지원 전 국정원장이 유 전 의원과 이준석 전 대표가 내년 초쯤 중도 신당을 창당할 것이라고 예언한 것에 대해 12월쯤 당을 떠날지 남을지 선택할 것이라며 신당 창당 가능성도 열어뒀다.
```

Results

- 성능 평가 : Rouge-Metric

- Rouge (Recall-Oriented Understudy for Gisting Evaluation)
  - n-gram 기법을 사용하여 label(사람이 만든 요약문), summary(모델이 생성한 요약문)을 비교해서 얼마나 겹치는지 수치로 표시. 모델의 성능 평가 기준. → 높을수록 좋은 성능.

Rouge - N	Rouge - L
• N = 1 ○ 단어 1개를 그룹화 • N = 2 ○ 단어 2개를 그룹화	• 일치하는 시퀀스의 길이가 길수록 높은 점수

Results

News Summary

기사 원문 : https://n.news.naver.com/article/658/0000056776

통계청 '2023년 9월 산업활동 동향' 발표부산 광공업 생산 7개월 연속 감소 흐름소비·투자도 뚝...전국은 '트리플 증가' 연합뉴스지난달 부산지역 생산·소비·투자가 1년 전보다 모두 줄어들며 두 달 만에 다시 '트리플 감소세'를 기록했다.특히 생산과 투자가 각각 7개월, 8개월 연속 감소하면서 산업 활동 부진이 고착 국면에 진입한 게 아니냐는 우려가 제기된다.반면 전국은 이들 3대 지표가 모두 증가했다.31일 통계청과 동남지방통계청이 각각 발표한 '2023년 9월 산업활동 동향' 자료를 보면 지난달 부산 광공업 생산 지수는 95.0(이하 2020년=100)으로 1년 전 같은 달보다 12.8% 줄었다. 지난 3월(-4.7%) 이후 7개월 연속 감소세(전년 동월 대비)다. 전월과 비교해도 1.9% 줄었다.동남지방통계청은 의료정밀광학(-54.3%)과 전자부품·컴퓨터·영상음향통신(-41.0%) 등의 생산 감소가 큰 영향을 미쳤다고 설명했다. 1차 금속(15.5%)과 금속 가공(11.1%) 등은 늘었다.광공업 출하도 7.8% 감소했다. 제조업 재고는 전년 동월 대비 0.2% 줄었다.지난달 부산지역 대형소매점(백화점+대형마트) 판매액 지수는 117.1로 지난해 9월보다 0.4% 줄었다. 지난 8월에는 전년보다 2.0% 늘었으나 한 달 만에 다시 감소세로 돌아섰다.백화점은 4.1% 감소했고 대형마트는 6.7% 증가했다.건설 수주액도 지난해 9월보다 45.9% 줄어든 1조5873억 원을 기록했다. 지난 2월(-3.5%) 이후 8개월 연속 감소세다.동남지방통계청은 "재건축과 발전·송전 부문에서 수주가 급감해 전체 건설 수주액이 감소했다"고 설명했다.부산지역 산업활동 3개 지표가 모두 줄어든 것은 지난 7월 이후 2개월 만이다.이와 달리 전국은 호조세를 보였다.지난달 전(全)산업 생산(계절조정·농림어업 제외) 지수는 113.1로 전월보다 1.1% 증가했다. 광공업 생산은 1.8% 늘었다.반도체 경기 회복세도 뚜렷해졌다. 지난달 반도체 생산은 전월 대비 12.9%, 전년 동월 대비 23.7% 늘었다.소비 지표인 소매 판매는 전월보다 0.2% 늘었다. 음식료품과 화장품 등에서 판매가 증가했기 때문이다. 지난 7월(-3.2%)과 8월(-0.3%) 두 달 연속으로 감소했으나 3개월 만에 증가세로 전환됐다. 설비투자는 기계류와 운송장비 투자가 늘면서 전월보다 8.7% 증가했다. 산업 생산과 소비·투자가 모두 증가한 것은 지난 5월 이후 처음이다.

간략히 보기

Bert 요약문 :

통계청 '2023년 9월 산업활동 동향' 발표부산 광공업 생산 7개월 연속 감소 흐름소비·투자도 뚝...전국은 '트리플 증가' 연합뉴스지난달 부산지역 생산·소비·투자가 1년 전보다 모두 줄어들며 두 달 만에 다시 '트리플 감소세'를 기록했다.특히 생산과 투자가 각각 7개월, 8개월 연속 감소하면서 산업 활동 부진이 고착 국면에 진입한 게 아니냐는 우려가 제기된다.반면 전국은 이들 3대 지표가 모두 증가했다.31일 통계청과 동남지방통계청이 각각 발표한 '2023년 9월 산업활동 동향' 자료를 보면 지난달 부산 광공업 생산 지수는 95.0(이하 2020년=100)으로 1년 전 같은 달보다 12.8% 줄었다. 지난 2월(-3.5%) 이후 8개월 연속 감소세다.동남지방통계청은 "재건축과 발전·송전 부문에서 수주가 급감해 전체 건설 수주액이 감소했다"고 설명했다.부산지역 산업활동 3개 지표가 모두 줄어든 것은 지난 7월 이후 2개월 만이다.이와 달리 전국은 호조세를 보였다.지난달 전(全)산업 생산(계절조정·농림어업 제외) 지수는 113.1로 전월보다 1.1% 증가했다.

간략히 보기

- 5 ~7 문장 내외의 요약

4조 수제 요약문:

3 개월 만에 증가 세로 전환 돼다 설비 투자 기 계류 운송 장비 투자가 늘다 전월 보다 87 증가 산업 생산 소비 투자가 모두 증가 것 지난 5월 이후 처음

자세히 보기

- 단어 중심의 나열

Kobart 요약문:

통계청과 동남지방통계청이 발표한 '2023년 9월 산업활동 동향' 자료를 보면 지난달 부산 광공업 생산 지수는 95.0으로 1년 전 같은 달보다 12.8% 줄었으며 제조업 재고는 전년 동월 대비 0.2% 줄었다.

자세히 보기

- 1 문장 내외의 요약

Results

- 평가 데이터 : validation 中 100개의 원문과 요약문을 기준으로 평가.

News Summary

요약모델

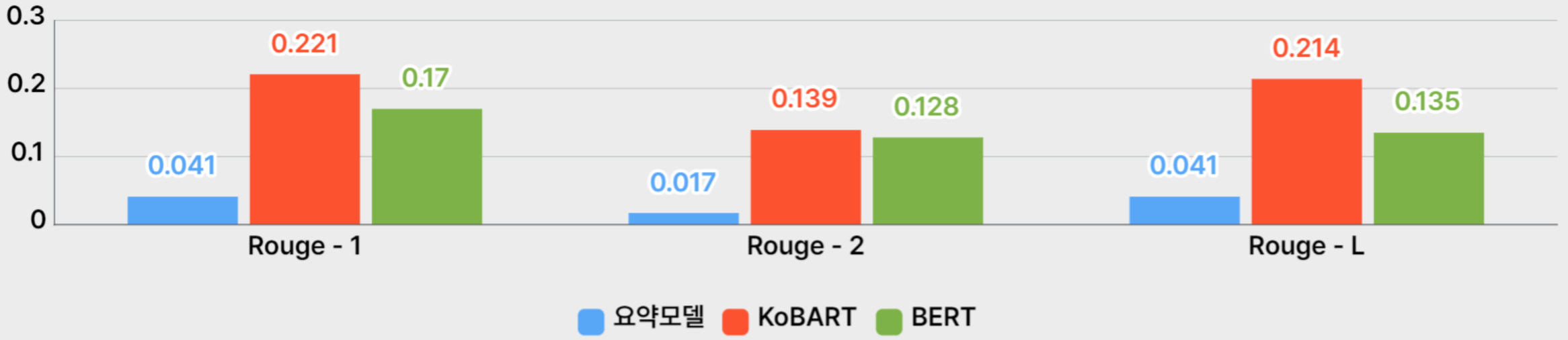
- Rouge-1 : 0.041
- Rouge-2 : 0.017
- Rouge-L : 0.041

KoBART

- Rouge-1 : 0.221
- Rouge-2 : 0.139
- Rouge-L : 0.214

BERT

- Rouge-1 : 0.17
- Rouge-2 : 0.128
- Rouge-L : 0.135



의의 및 한계점

- 정보의 간결한 전달
- URL을 활용한 편의성
  - 웹 크롤링 + NLP
- 플라스크를 통한 ,모델 간의 직관적 비교
  - 원문에 대한 Output 비교
- KoBART 성능 우수
  - 길이, 문장의 완성도 등 고려
  - ROUGE Metric 및 Human Evaluate

- ROUGE Metric 지표 한계
  - 보편적인 평가지표. 한국어 적용에 한계 존재
  - 한계 보완지표 RDASS
  - Human Evaluate를 통해 비교 및 보완
- 요약모델의 성능
  - 단어 위주의 요약으로 인한 문장 생성 한계
  - 문장을 매끄럽게 생성하는 기술적 보완 필요
- 요약의 한계
  - 요약의 특성상 중요 내용 제외 가능성

Member Roles

News Summary

박종우

- 요약 모델 코드 작성 및 모델링
- 요약 모델 학습 & 테스트

박치수

- 모델 플라스크 서빙
- 웹 크롤링 Input 변환
- html 작성 및 UI 수정

서형준

- 데이터 변환
- ROUGE-Metric 코드 작성
- 모델 성능 평가

성문주

- 발표 및 발표 자료 수집
- 요약 모델 코드 작성 모델링
- 요약 모델 학습 & 테스트

염규민

- 발표 자료 작성
- 프로젝트 작업 조율

장성웅

- 사전훈련모델 코드 수정
- ROUGE-Metric 코드 작성
- 웹 크롤링 코드 작성

Q & A

THANK YOU