# Dialogue Summarization |
## 일상 대화 요약

Team 11. 전지재능

# 목 차


01 팀 소개

02 대회 소개 & 데이터 탐색

03 모델 학습
- Bart
- T5
- Llama3.1
- SOLAR
- Gemma2

04 결과 추론

05 프로젝트 결과

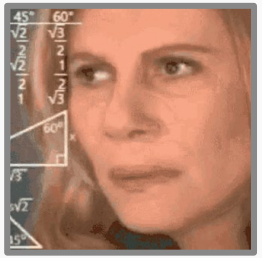
06 회고

07 Q&A

01

# 팀원 소개

# 팀원소개



## 이윤재

모델 학습
데이터 전처리

## 이재명

모델 학습
데이터 전처리

## 장은지

모델 학습
데이터 전처리

## 전백찬

모델 학습
데이터 전처리

## 최지미

모델 학습
데이터 전처리

# 협업방식

GitHub

코드 공유

Google
스프레드시트

실험 기록

Google
프레젠테이션

발표 자료 작성


zoom

미 팅


Slack

실시간 소통

- 미팅 일정 : 매일 오전 11시
- 기록 : 협업 툴을 이용하여 기록 및 공유

# 현황 공유

현황 기록 후 미팅 시 활용하여 아이디어 발굴

| | | |
|---|---|---|
| 9/3(화) | 어제한거 | @재명 Baseline 코드 분석. |
| | | @지미 t5 시도했지만 커널 터져서 실패ㅎ |
| | | @지미 https://huggingface.co/MLP-KTLim/llama-3-Korean-Bllossom-8B 라마3 기반인데 테스트 시 느리지만 꽤 잘돼요(1대화 2분) |
| | | @지미 deepL API 성능 좋지만 비쌈. 파일 번역도 유료해야할듯 |
| | | Yoonjae Lee google/flan-t5-xl one dialogue test시 결과가 좋았으나 메모리 부족으로 불가능, google/flan-t5-large 진행중인데 param조정 필요함 |
| | | Ej J 은지 베이스라인코드보다가 유의어교체해보려다 설치에서 막힌상태입니다. |
| | | @백찬 ko-gemma2 2B, 9B 모델 학습(QLoRA, SFT) \| do_sample = False (1EPOCH > 2EPOCH > 4EPOCH : 1~2 EPOCH에서 과적합 가능성) |
| | | |
| | 오늘할거 | @재명 외부 데이터 찾아보고 학습 해보기. |
| | | Yoonjae Lee google/flan-t5-large 제출해보기, SamSum 데이터셋으로 증강시도 |
| | | @재명 토크나이저를 학습해보고 서브미션 해보기. |
| | | Ej J 유의어교체 마저 해볼 생각입니다. |
| | | @지미 bllossom 모델 학습이랑 추론 확인해보고 진행? |
| | | |
| | tip | @지미 모델마다 설계된 input이 많이 달라서, 각 모델의 특성대로 넣어줘야 예측이 잘 나옴. 조금 안 맞으면 이상하게 나올 수 있음 |
| | | |
| | | |
| | 아이디어 | |
| 9/4(수) | 어제한거 | @재명 Baseline 을 보다 자세히 이해하기. (허깅페이스 Bart 문서 참고) |
| | | @재명 대화문에서 #Person..#: 패턴 지우고 학습 후 서브미션 해보기. Baseline final_result 기준으로 0.8355 점 하락. |
| | | @재명 aihub.or.kr 한국어 대화 요약 데이터셋 전체 학습 후 서브미션 해보기. Baseline final_result 기준으로 0.0501 점 상승. 학습 시간은 거의 10시간 정도 걸려서 |
| | | https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=117 |
| | | @지미 llama3 기반 모델 qlora 학습, 추론 테스트 |
| | | Yoonjae Lee flan-t5-xl 서버터짐, 서버재생성 |
| | | Yoonjae Lee token추가 32100 > 02021 para chat 추가 없는 듯 gpt도 별로 추가없을것같다 한 token이 중요한게 아닐것같 |

# 실험 기록

| 사용한 모델 | 변경한 내용 | 에폭 | 점수 |
|---|---|---|---|
| baseline | 기본 | 9 | 41.59 |
| | 기본+에폭수 늘림 | 14 | 41.59 |
| | 증강 0.5, 0.5, 0.5, 0.5<br>Synonym Replacement (SR) : 유의어 교체<br>Random Insertion (RI) : 임의 단어 삽입<br>Random Swap (RS) : 두 단어 위치 변경<br>Random Deletion (RD) : 임의 단어 삭제 | 7 | 24.78 |
| | 스페셜토큰 추가 | 9 | 41.23 |
| | 텍스트 정규화(Text Normalization)<br>text = re.sub(r"", '', text) # 큰따옴표 " 제거<br>text = re.sub(r'\s{2,}', ' ', text) # 2칸 이상의 공백을 1칸으로 변경<br>text = re.sub(r'\s\.', '.', text) # 공백 + 점(.)을 점(.)으로 변경<br>text = re.sub(r'\s,', ',', text) # 공백 + 쉼표(,)를 쉼표(,)로 변경<br>text = re.sub(r'ㅣ', 'l', text) # ㅣ를 l로 변경<br>text = re.sub(r'\?\?', '?', text) # ??를 ?로 변경<br>text = re.sub(r'\sl', 'l', text) # 공백 + 느낌표(l)를 느낌표(l)로 변경<br>text = re.sub(r'\s\?', '?', text) # 공백 + 물음표(?)를 물음표(?)로 변경<br>text = re.sub(r'\s;', ';', text) # 공백 + 세미콜론(;)을 세미콜론(;)으로 변경<br>text = re.sub(r',', '.', text) # 세미콜론(;)을 점(.)으로 변경<br>text = re.sub(r'[ㄱ-ㅎ&&[^ㅋ]]', '', text) # ㄱ~ㅎ 중 ㅋ를 제외한 모든 자음 삭제<br><br># 작은 따옴표 안의 텍스트를 제외하고 모음만 있는 경우 삭제<br>text = re.sub(r"(?<!')([ㅏ-ㅣ])(?!')", '', text) # 작은 따옴표로 감싸지 않은 모음만 삭제 | 9 | 42.05 |
| | 증강(RI,RS,RD) 0.5/ 전체 문단안에서 내용 섞임-Okt | 5 | 41.62 |
| | 증강(RI,RS,RD) 0.3/ 화자 고정, 문단 안의 문장 내에서 내용 섞음 | | 41.52 |

Google sheet를 이용해 실험 기록 공유

| 모델 | 기법 | EPOCH | RESULT | REMARKS |
|---|---|---|---|---|
| koBart | | 11 | 41.9858 | |
| koBart | special tokens, AUG | 11 | 41.8931 | 삽입, 삭제, 문장순서변경 |
| gemma2 2B | 4bit quantization | 4 | 41.7957 | |
| gemma2 2B | 8bit quantization | 1 | 43.0463 | |
| gemma2 2B | 8bit quantization, nosampling | 1 | 43.6519 | |
| gemma2 9B | FP16, nosampling | 1 | 45.5661 | |
| gemma2 9B | FP16, nosampling | 2 | 45.746 | |
| gemma2 9B | BF16, nosampling, lora size Up, drop out | 2 | 46.1901 | LoRA_RANK = 16(<-6), ALPHA=32 |
| gemma2 9B | BF16, nosampling, lora size Up, dropout tuning, validation+ | 2 | 46.6249 | DROPOUT=1.6 (<-1.0) |
| | | | | |
| exaone 3.0 | FP16, nosampling, lora size Up, dropout validation+ | 1 | 37.0855 | |
| openchat 3.5 | FP16, nosampling, lora size Up, dropout, validation+ | | 40.3 | |

| 모델 | 시도 | epoch | 결과 |
|---|---|---|---|
| digit82/kobart-summarization(baseline) | 기본 | | 41.9206 |
| | special token 추가, input format 정리 | | 42.0756 |
| | 전처리 : 중복 삭제, summary 오류 삭제, 영어 번역 삭제, 포맷 일치, 연속 공백 등 클렌징 | | 41.6266 |
| | 번역 증강(구글 번역기) | 17 | 41.8545 |
| | vocab 추가 | | 41.4904 |
| eenzeenee/t5-base-korean-summarization | 전처리 epoch | 8 | 42.7859 |
| | 번역 증강(구글 번역기) | 6 | 42.6897 |
| | vocab 추가 | | 39.767 |
| MLP-KTLim/llama-3-Korean-Bllossom-8B | 기본 | 1 | 37.6917 |

02

# 대회 소개 & 데이터 탐색

# 02 대회 소개 & 데이터 탐색

- 목표 : DialogueSum Solar로 번역한 한국어본을 활용, 일상 대화에 대한 요약을 생성하는 모델 개발
- 대회 기간: 2024년 8월 29일 (10시) ~ 2024년 9월 10일 (19시)

## 학습 데이터

- train : 12457
- dev : 499
- test : 499 (250, hidden-test : 249)
- 대화문에서 발화자는 #Person"N"#으로 구분되어 있습니다. 대화문에 존재하는 개인정보(예: 전화번호, 주소 등)는 다음과 같이 마스킹 예) 전화번호 -> #PhoneNumber#
- 대회 데이터셋: DialogSum Dataset:
- CC BY-NC-SA 4.0 license
- 단, 해당 데이터를 한국어로 번역하여 활용
- 원본: ttps://github.com/cylnlp/dialogsum

## Evaluation Metric

예측된 요약 문장을 3개의 정답 요약 문장과 비교하여 metric의 평균 점수를 산출

본 대회에서는 ROUGE-1-F1, ROUGE-2-F1, ROUGE-L-F1, 총 3가지 종류의 metric으로부터 산출된 평균 점수를 더하여 최종 점수를 계산

3개의 정답 요약 문장 중 하나를 랜덤하게 선택하여 산출된 점수가 약 70점 정도

ROUGE-1는 모델 요약본과 참조 요약본 간에 겹치는 unigram의 수, ROUGE-2는 bigram의 수, ROUGE-L: LCS 기법을 이용해 최장 길이로 매칭되는 문자열을 측정

ROUGE-F1은 ROUGE-Recall과 ROUGE-Precisioin 조화평균

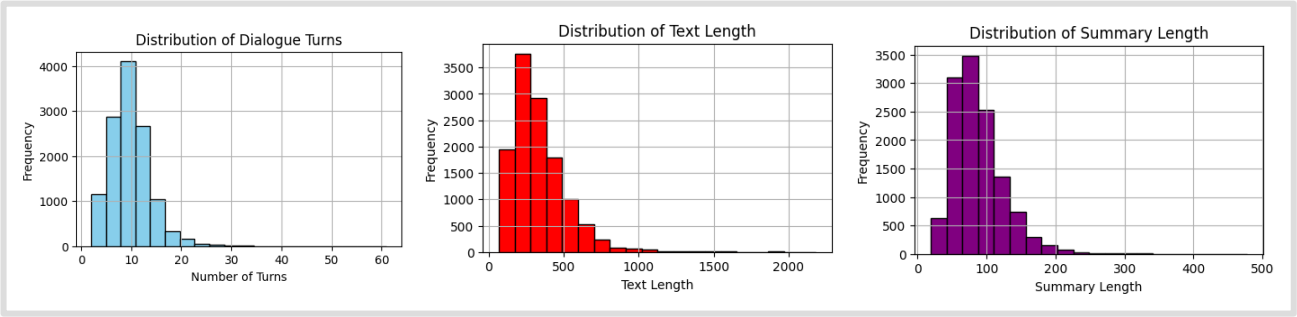
한국어 데이터 특성 상 정확한 ROUGE score 산출위해 문장 토큰화를 진행한 후 평가

# 02 대회 소개 & 데이터 탐색

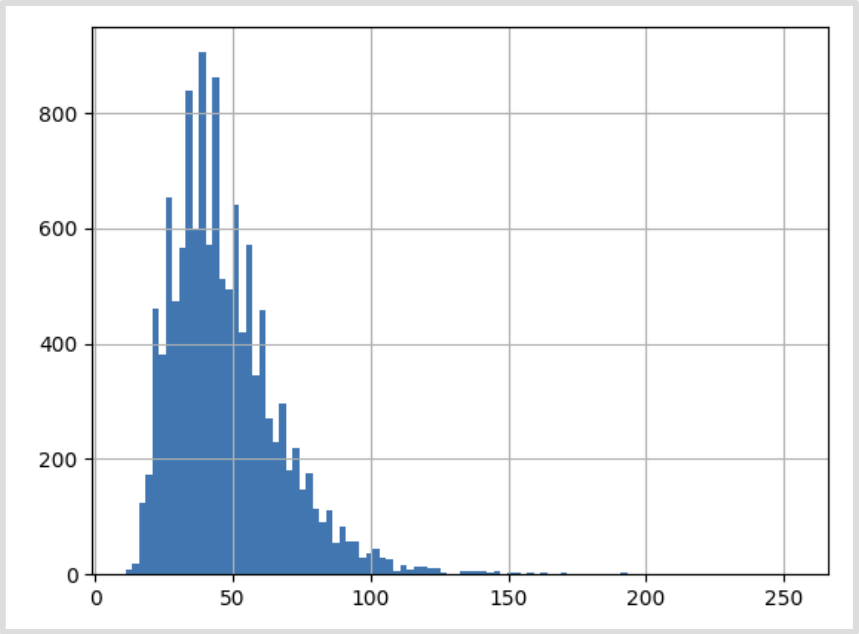
| train | Dialogue turns | Number_ Of_ speakers | Dialogue length (단순 len) | Summary Length (단순 len) | Summar /dialogue |
|---|---|---|---|---|---|
| mean | 9.49 | 2.01 | 335.39 | 87.40 | 0.26 |
| std | 4.15 | 0.13 | 187.74 | 37.64 | 0.20 |
| min | 2.00 | 2.00 | 67.00 | 19.00 | 0.28 |
| 25% | 7.00 | 2.00 | 211.00 | 61.00 | 0.28 |
| 50% | 9.00 | 2.00 | 294.00 | 80.00 | 0.27 |
| 75% | 12.00 | 2.00 | 418.00 | 106.00 | 0.25 |
| max | 61.00 | 7.00 | 2184.00 | 478.00 | 0.21 |

- dialogue summary length의 비율을 동적으로 계산하여 max_new_token에 적용할 수 있다.

- 모델별 tokenizer를 적용시 값은 달라질 수 있다.



# 02 대회 소개 & 데이터 탐색: train.csv 에서 summary 의 토큰 수 분포




대부분 20 ~ 100토큰 사이에 분포

# 03

# 모델 학습

# 03 모델 학습: BART

- 사용한 모델 : digit82/kobart-summarization
- KoBART는 한국어에 맞춘 BART(Seq2Seq 모델)로, 자연어 처리에서 주로 요약, 번역, 문장 생성 작업에 사용
- Dialogue Summarization 대회에서 제공되는 baselinecode

| 실험 내용 | Mid Result | Final Result |
|---|---|---|
| Baseline (epoch 9) | 41.5982 | 38.7166 |
| Baseline + epoch 추가 (epoch 14) | 41.5982 | 38.7166 |
| Train Data 내 빈도 높은 Vocab 1000개 추가 | 41.4904 | 39.2065 |
| Special tokens 추가 | 41.2387 | 38.7187 |
| Special tokens 추가, input format 정리 | 42.0756 | 39.4633 |
| 데이터 증강(Data Augmentation) : 번역 (구글 번역기) | 41.8545 | 39.5367 |
| 데이터 증강(Data Augmentation) p=0.5: 임의 단어 삽입, 두 단어 위치 변경, 임의 단어 삭제(Okt) | 41.6256 | 39.0742 |
| 데이터 증강(Data Augmentation) + Special tokens 추가 | 41.8931 | 38.6632 |
| 데이터 증강(Data Augmentation) p=0.3: 임의 단어 삽입, 두 단어 위치 변경, 임의 단어 삭제(Okt) | 41.5204 | 39.1807 |
| 텍스트 정규화(Text Normalization) : 불필요한 공백, 특수문자 및 모음, 자음 삭제 | 42.0593 | 38.6818 |
| 텍스트 정규화(Text Normalization) + 후처리 | 42.1058 | 38.7040 |

https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=117

# 03 모델 학습: T5

- 사용한 모델 : eenzeenee/t5-base-korean-summarization
- T5는 'Text-to-Text Transfer Transformer'의 약자로, 다양한 자연어 태스크를 처리하는 범용 언어 모델
- 이 한국어 모델은 요약 작업에 특화되어 fine-tuning 되었음

| 실험 내용 | Mid Result | Final Result |
|---|---|---|
| 기본 모델 - max_length:512, min_length:128<br>텍스트 정규화(Text Normalization) : 불필요한 공백, 특수문자 및 모음, 자음 삭제 | 42.7859 | 41.7716 |
| + 구글 번역 증강(총 24888) | 42.6897 | 40.9797 |
| + Train Data 내 빈도 높은 Vocab 1000개 추가 | 39.7670 | 38.3287 |
| + 역번역 한 dev dataset 포함해서 학습 | 43.1876 | 41.4760 |

# 03 모델 학습: Llama3.1

- 사용한 모델 : meta-llama/Meta-Llama-3.1-8B
- Llama recipes 코드를 수정해서 peft 학습 및 추론
  (https://github.com/meta-llama/llama-recipes/blob/v0.0.3/recipes/quickstart/finetuning/quickstart_peft_finetuning.ipynb)

| 실험 내용 | Mid Result | Final Result |
|---|---|---|
| train 데이터 학습<br>max_new_tokens=100<br>BitsAndBytesConfig load_in_8bit=True / LORA_CONFIG r=8 | 41.6496 | 40.3015 |
| train + dev 합친 데이터 증강학습<br>min_new_tokens=30, max_new_tokens=250<br>BitsAndBytesConfig load_in_8bit=True / LORA_CONFIG r=8 | 41.2613 | 39.5892 |
| train + dev 합친 데이터 증강학습<br>max_new_tokens=200<br>BitsAndBytesConfig load_in_8bit=True / LORA_CONFIG r=32<br>clean_up_tokenization_spaces=True | 42.3329<br>(+0.6833) | 40.6886 |

# 03 모델 학습: SOLAR

- 사용한 모델 : upstage/SOLAR-10.7B-v1.0
- Solar 는 Llama 기반이므로, Llama recipes 코드를 수정해서 학습 및 추론. (https://github.com/meta-llama/llama-recipes/blob/v0.0.3/recipes/quickstart/finetuning/quickstart_peft_finetuning.ipynb)

| 실험 내용 | Mid Result | Final Result |
|---|---|---|
| train + dev 합친 데이터 학습<br>min_new_tokens=30, max_new_tokens=250<br>Torch_dtype = torch.float16 을 torch.bfloat16 으로 대체 | 43.0265 | 41.3859 |
| min_new_tokens 삭제<br>max_new_tokens=175 로 줄임<br>추론한 문자열의 시작, 끝에 공백 문자 제거 | 44.6292<br>(+1.6027) | 42.8231 |

# 03 모델 학습: Gemma2

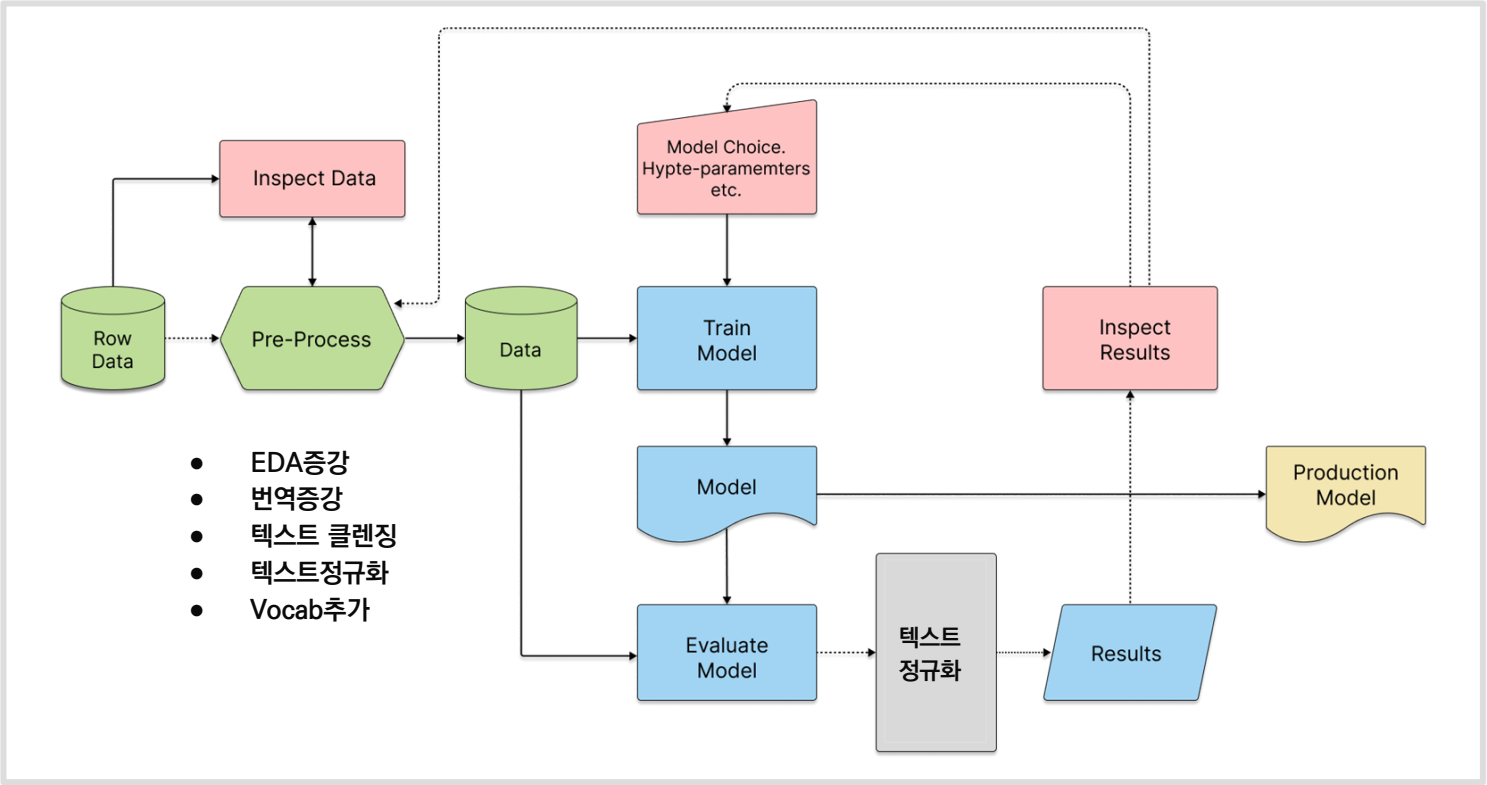
- 사용한 모델 : rtzr/ko-gemma-2-9b-it

| 실험 내용 | EPOCH | Mid Result | Final Result | REMARKS |
|---|---|---|---|---|
| 4bit quantization | 4 | 41.7957 | 40.6917 | |
| 8bit quantization | 1 | 43.0463 | 42.2578 | |
| 8bit quantization, nosampling | 1 | 43.6519 | 42.5284 | |
| FP16, nosampling | 1 | 45.5661 | 45.0033 | |
| FP16, nosampling | 2 | 45.7460 | 43.7476 | |
| BF16, nosampling, lora size Up, drop out | 2 | 46.1901 | 44.8012 | LoRA_RANK = 16(↙−6), ALPHA=32 |
| BF16, nosampling, lora size Up, dropout tuning, validation+ | 2 | 46.6249 | 44.6438 | DROPOUT=1.6 (↙−1.0) |

# 03 모델 학습: 기타 모델

| 모델 | 실험 내용 | Mid Result | Final Result |
|---|---|---|---|
| MLP-KTLim/llama-3-Korean-Bllossom-8B | FP16, nosampling, lora size Up, 1epoch | 37.6917 | 36.5837 |
| Exaone 3.0 | FP16, nosampling, lora size Up, dropout +validation, 2epoch | 37.0855 | 36.1048 |
| Openchat 3.5 | FP16, nosampling, lora size Up, dropout +validation, 2epoch | 40.3027 | 39.4023 |

# 03 모델 학습: workflow



# 04

# 결과 추론

# 04 결과 추론: 모델 별 추론

원문 Test_185,

"#Person1#: 실례합니다. 미스터 리를 찾습니다. 그분 소포입니다.

#Person2#: 아, 제 책상 위에 두세요. 여기 서명해야 하죠? 잠시만 기다리시겠어요? 제가 말씀드릴 것이 있어서요."

| Model | Summary |
|---|---|
| digit82/kobart-summarization | #Person1#은 #Person2#에게 미스터 리를 찾는 데 도움을 줍니다. |
| eenzeenee/t5-base-korean-summarization | #Person1#은 미스터 리를 찾고 서명을 요청합니다. |
| meta-llama/Meta-Llama-3.1-8B | #Person1#이 미스터 리의 소포를 #Person2#에게 건네주고 서명을 요청합니다. |
| upstage/SOLAR-10.7B-v1.0 | #Person1#은 미스터 리의 소포를 받습니다. |
| rtzr/ko-gemma-2-9b-it | #Person1#이 미스터 리에게 소포를 전달합니다. |

05

# 프로젝트 결과

# 05 프로젝트 결과: 모델 별 결과

| Model | ROUGE-1 | ROUGE-2 | ROUGE-3 | Mid Score | Final Score |
|---|---|---|---|---|---|
| digit82/kobart-summarization | 0.5159 | 0.3202 | 0.4271 | 42.1058 | 38.7040 |
| eenzeenee/t5-base-korean-summarization | 0.5217 | 0.3262 | 0.4356 | 42.7859 | 41.7716 |
| meta-llama/Meta-Llama-3.1-8B | 0.5118 | 0.3265 | 0.4317 | 42.3329 | 40.6886 |
| upstage/SOLAR-10.7B-v1.0 | 0.5371 | 0.3456 | 0.4561 | 44.6292 | 42.8231 |
| rtzr/ko-gemma-2-9b-it | 0.5557 | 0.3705 | 0.4725 | 46.6249 | 44.6438 |

06

# 회고

# 06 회 고: 가설 및 결과

## 텍스트 정규화

불필요한 공백, 특수문자 및 모음, 자음 삭제

점수의 등락이 아주 미세하게 있으나 큰 차이는 없음

## Output Data 후처리

미세하게 점수가 향상됨. 모델에 따라 결과가 다르기 때문에 처리할 내용도 달라지게 됨.

# 06 회 고: 가설 및 결과

## 데이터 증강

### 임의 단어 삽입, 두 단어 위치 변경, 임의 단어 삭제

증강된 데이터는 문장의 의미가 보존되지 않아서 점수가 낮아지거나, 변동이 없었음

### 역번역으로 데이터 증강(Google, Solar)

유의어가 대체되고 문장이 다양해져 학습 효과가 있을 것으로 기대했으나
점수가 등락이 미미해서 영향이 적음

# 06 회 고: 가설 및 결과

## 토큰 추가

### Train, dev, test data의 Special tokens 추가

- Test data에 있는 special tokens은 이미 Baseline코드에 있기 때문에 성능이 개선되는지 확인하기 어려웠음
- 모델에 따라서 원래의 토큰이 보존이 안되는 경우가 있었음

### data set에서 등장 빈도 높은 단어 추가

- Word2Vec, sentencepiece 를 활용해서 단어 추가 후 각각 테스트 진행했으나 새로운 단어가 생성되는 등의 부작용 발생
- 예측을 해치지 않는 단어들을 세밀하게 파악해서 추가하면 효과가 있을 것으로 예상됨

# 06 회 고: 가설 및 결과

## 모델별 최적화 요소 찾기

### Input 형식을 맞춰야 결과가 잘 나온다

모델들마다 가지고 있는 Input 형식을 따르지 않으면 결과가 크게 달라짐

### 적절한 max_new_tokens 값 찾기

값이 너무 작으면 요약문이 잘릴 수 있음. 그렇다고 값이 크면 정답과 차이가 더 커질 수 있음.

# 06 회 고: 인사이트

## LLM

요약문의 품질이 정성적으로 좋아보여도 정답과 스타일이 다르면 점수가 안좋을 수 있음

## 하이퍼파라미터 조정

LLM의 경우 하이퍼 파라미터 조정만으로 성능 향상을 이룰 수 있음

## 데이터 증강

번역 등의 증강을 해도 성능향상이 안됐지만, 외부데이터 사용하여 추가 학습시 성능 향상됨
의미 보전이 잘 되는 다양한 대화의 데이터를 추가하면 성능향상이 클 것으로 생각됨

# 진 행 소 감

**이윤재**

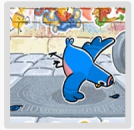
huggingface를 처음 접했다. 모델별 코드가 달라 난이도가 높고 참고 코드가 없이는 진행할 수 없었다. LLM에서는 Training에서 tokenizer의 조절이 불필요했다. inference시 tokenizer parameter를 변동하면 모델에 따라 큰 차이가 생겼다. Solar는 Llama기반이라 cookbook이 없는것인지? LLM세계에 발을 담궜다.

**이재명**

가설이 틀린 경우가 많아서 힘들었다. 그래도 팀원들로부터 새로운 아이디어를 많이 얻어서 계속 시도해볼 수 있었다.

**장은지**

성능은 아쉬웠지만 여러가지 기법들을 시도해 볼 수 있어서 재밌었다. 그리고 모델마다 구조가 다르고 LLM에서 GPU메모리 오류가 많이 발생했지만 든든한 팀원들의 도움으로 개선하기 위해 이것저것 시도해 볼 수 있었다.



**전백찬**

제한된 메모리 안에서 최대한의 효율과 성능을 뽑는 시도를 해보아서 유용했다.

**최지미**

LLM+하이퍼 파라미터의 효과를 확인할 수 있어 즐거웠다.
여러가지 기법들이 효과가 없어서 아쉬웠고, 방법들을 수정해서 다시 시도해보고 싶다.

# 07

# Q&A