

도배 하자 질의 응답 처리 : 한솔데코 시즌 2 AI 경진대회
알고리즘 | 언어 | LLM | MLOps | QA | Cosine Similarity

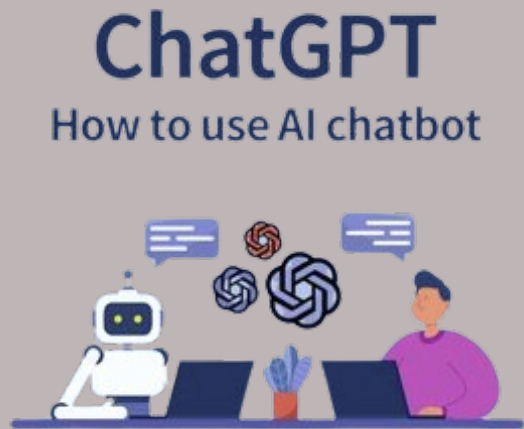
도배하자 질의응답 LLM 개발

구준희 김지원 박서진 신진섭 엄성원

contents

- 01 프로젝트 배경
- 02 EDA
- 03 Text Augmentation
- 04 Modeling & Inference
- 05 보완할 점 & 추후계획

Project Background



자연어처리

언어모델

추론

[목표]

자연어를 처리하여 언어모델에 학습시키고
특정 TASK에 적합한 TEXT를 생성하는 원리를 이해한다

Overall Progress

도배 하자 질의 응답 처리 : 한솔데코 시즌2 AI 경진대회

알고리즘 | 언어 | LLM | MLOps | QA | Cosine Similarity

₩ 상금 : 1000만 원

🕒 2024.01.29 ~ 2024.03.11 09:59

+ Google Calendar



Step 1

Transformer 이해
논문 공부

Step 2

NLP 도메인 및
공모전 목표에 대한 이해

Step 3

도배 하자 질의 응답
언어모델 개발 및 추론

EDA

주최 측 Train Data

- id, 질문 2개, 질문의 카테고리, 답변 5개의 column으로 이루어져 있음
- 질문 - 답변 set로 토큰화해서 input으로 입력

train data의 예시

	id	질문_1	질문_2	category	답변_1	답변_2	답변_3	답변_4	답변_5
0	TRAIN_000	면진장치가 뭐야?	면진장치에 사용되는 주요 기술은 무엇인가요?	건축구조	면진장치란 지반에서 오는 진동 에너지를 흡수하여 건물에 주는 진동을 줄여주는 진동...	면진장치란 건물의 지반에서 발생하는 진동 에너지를 흡수하여 건물을 보호하고, 진동을...	면진장치란 지반으로 부터 발생하는 진동 에너지를 흡수하여 건물에 전달되는 진동을 줄여...	면진장치는 건물의 지반으로부터 오는 진동 에너지를 흡수하여 건물에 전달되는 진동을...	면진장치는 건물에 오는 지반 진동의 영향을 최대한으로 흡수하여 건물에 전달되는 진동...
1	TRAIN_001	내진설계의 종류 좀 알려줘	내진설계에는 어떤 종류가 있는지 자세히 알려주실 수 있나요?	건축구조	내진 설계의 종류로 내진구조, 제진구조, 면진구조가 있습니다.	내진설계에는 내진구조, 제진구조, 면진구조가 있습니다. 내진구조는 건물 구조물이 지...	내진설계에는 주로 내진구조, 제진구조, 면진구조의 세 가지 종류가 있습니다. 이들은...	내진설계에는 주로 내진구조, 제진구조, 면진구조가 사용됩니다. 내진구조는 건물 구조...	내진 설계에는 다양한 종류가 있지만, 대표적으로 내진구조, 제진구조, 면진구조가 있...
2	TRAIN_002	철골구조의 장점이 뭐야?	철골구조의 장점을 알려줘?	건축구조	철골구조는 건물의 외벽에는 그다지 하중이 걸리지 않기 때문에 고층 건물의 건축이 가...	철골구조의 장점은 건물의 외벽에는 그다지 하중이 걸리지 않기 때문에 고층 건물의 건...	철골구조의 장점은 건물의 외벽에 하중이 적게 걸리기 때문에 고층 건물의 건축이 용이...	철골구조의 장점은 건물의 외벽이 하중이 걸리지 않아 공간 활용이 용이하고, 고층 건...	철골구조의 장점은 건물의 외벽에 하중이 크게 걸리지 않아 고층 건물을 건축할 수 있...

EDA

데이터의 한계에 따른 처리 방안

1. 답변이 주어진 질문에
대한 답이 아닌 경우

질문 1 A의 구조가 뭐야?

질문 2 A의 장점과 단점에는
무엇이 있을까?

답변 A의 구조는 ##입니다.

2. 오타 등 단순 텍스트 오류

case 1 벽면 손상이 발생한 경우
도배지 끝부분이 들“떨” 수
있을까요?

case 2 문장부호의 소실,
질문을 반복함,
영어 글자들이 삽입됨

3. 특정 키워드가
존재하지 않는 경우

도배하자 문맥 이해에 있어
중요한 키워드가 없는 경우

Ex) 겨울철 관리, 여름철 관리,
페인트 부작용 등등

EDA

데이터의 한계에 따른 처리 방안

1. 상이한 두 질문을
각기 다른 행으로 분리

질문 1 A의 구조가 뭐야?

답변 1 A의 구조는 ##입니다.

질문 2 A의 장점과 단점에는
무엇이 있을까?

답변 2 (text generation)

2. 오타 수정 및 전처리

case 1 벽면 손상이 발생한 경우
도배지 끝부분이 들뜰 수
있을까요?

case 2 알고리즘 처리 & 수작업

3. 키워드 생성

해당 키워드에 대하여
Prompt engineering을 통해
답변 생성

질문 1 겨울철 유의사항에
대해 알려줘

답변 1 (text generation)

Text Augmentation

Crawling & Prompt Engineering

1. 논문 및 레퍼런스 자료

- 전문성 있는 질문 소스 탐색



2. 실용적 앱 / 웹 사이트

- 사람들이 자주 묻는
질문적, 실용적 키워드 탐색



3. Train 데이터 키워드 기반 다양한 종류의 질문 생성

ChatGPT를 이용하여 질문에 대한 답변을 생성

Ex) 너는 도배업계에서 일하는 전문가야. 일반인을 대상으로 {질문}에 대한 전문적인 답변 5가지를 200글자 내로 생성해줘. {질문}

최종 train dataset

- 기존 : 644 rows * 2 * 5 = 6440 QA sets
- 증강 이후 : 8319 QA sets

A	B	C	D	E	F	G	H	I
id	질문_1	질문_2	category	답변_1	답변_2	답변_3	답변_4	답변_5
TRAIN_000_1	면진장치가 뭐야?		건축구조	면진장치란 지반	면진장치란 건물	면진장치란 지반	면진장치는 건물	면진장치는 건물
TRAIN_000_2	면진장치에 사용되는 주요 기술은 무엇인가요?		건축구조	면진장치의 주요	면진장치에 사용	면진장치의 핵심	면진장치에 사용되는 주요 기술로	
TRAIN_001	내진설계의 종류 좀 알려줘.	내진설계에	건축구조	내진 설계의 종류	내진설계에는 내	내진설계에는 주	내진설계에는 주	내진 설계에는 다
TRAIN_002	철골구조의 장점이 뭐야?	철골구조의	건축구조	철골구조는 건물	철골구조의 장점	철골구조의 장점	철골구조의 장점	철골구조의 장점
TRAIN_003	철골철근 콘크리트 구조가 뭐야?	철골철근 콘	건축구조	철근철골콘크리	철골철근콘크리	철골철근 콘크리	철골철근콘크리	철골철근 콘크리
TRAIN_004	철골구조는 어떤 방식이 있어?	철골구조의	건축구조	철골구조는 일반	철골구조는 일반	철골구조는 주로	철골구조는 주로	철골구조는 일반
TRAIN_005	커튼월이 뭐야?	커튼월이란	건축구조	커튼월은 건물의	커튼월은 건물의	커튼월은 건물의	커튼월은 건물의	커튼월은 건물의
TRAIN_006	내진구조가 뭐야?	내진구조에	건축구조	내진구조란 강한	내진구조란 지진	내진구조란 지진	내진구조는 지진	내진구조란 지진
TRAIN_007	중목구조 방식이 뭐야?	중목구조 방	건축구조	중목구조는 기본	중목구조는 건물	중목구조는 기본	중목구조란 주된	중목구조는 건물
TRAIN_008	기둥-보 구조 방식이 뭐야?	기둥-보 구2	건축구조	기둥-보 구조 방	기둥-보 구조 방	기둥-보 구조 방	기둥-보 구조 방	기둥-보 구조 방
TRAIN_009	블록구조가 뭐야?	블록구조가	건축구조	블록구조는 조적	블록구조는 건물	블록구조란 건물	블록구조는 건물	블록구조는 조적
TRAIN_010	철골구조의 단점이 뭐야?		건축구조	철골구조는 화재	철골구조의 단점	철골구조의 단점	철골구조의 주요	철골구조의 주요
TRAIN_011	콘크리트 구조는 어떤 방식이 있어?	콘크리트 구	건축구조	콘크리트는 철근	콘크리트 구조는	콘크리트는 건축	콘크리트 구조에	콘크리트 구조에

Modeling


Backbone Model 모델 탐색

Skt/kogpt2-base-v2 [from baseline code]

: skt에서 개발한 한국어 gpt2 모델 (125M)

42dot/42dot_LLM-SFT-1.3B

: 국내 최초의 한영통합 언어 모델 기반의 경량 생성형 언어모델

maywell/Synatra-42dot-1.3B 

: 위 모델을 기반으로 instruction tuning된 Pretrained LLM

: 다양한 3B 이내 경량 모델로 실험하던 중 GPU 메모리 아웃이 나지 않은 모델

Issues

1. 한정된 GPU 자원 (Kaggle Notebook, Colab T4 RAM 16GB 사용)
2. 파라미터 수가 큰 모델이면 한국어 데이터로 사전훈련된 모델이
아니더라도 높은 성능을 기대할 수 있지만...

Listup

1. Google/gemma-2b (new!!)
2. Microsoft/phi-2
3. etc ...

```
OutOfMemoryError: CUDA out of memory.
allocated memory 38.31 GiB is allocated
avoid fragmentation. See documentation
```

```
OutOfMemoryError: CUDA out of memory.
llocated memory 14.51 GiB is allocated
fragmentation. See documentation for
```

Modeling

Fine Tuning (SFT)

- Prompt : LLM에 instruction을 주어 특정 task에 알맞는 대답을 형성하는 것
- Fine Tuning : 사전 학습된 모델을 소규모의 특정 데이터 세트에 대해 추가로 학습시켜 특정 작업이나 도메인에서 기능을 개선하고 성능을 향상시키는 프로세스
generic task-specific

관절 통증, 피부 발진, 햇빛 민감성?



Base model



알려지?

관절 통증, 피부 발진, 햇빛 민감성?



Fine-tuned model



자가 면역 질환인 전신성 홍반성 루푸스일 수 있습니다.
류마티스 전문의에게 진찰을 받고 잠재적인 검사를 받는 것이 중요합니다.



알려지 증상
데이터



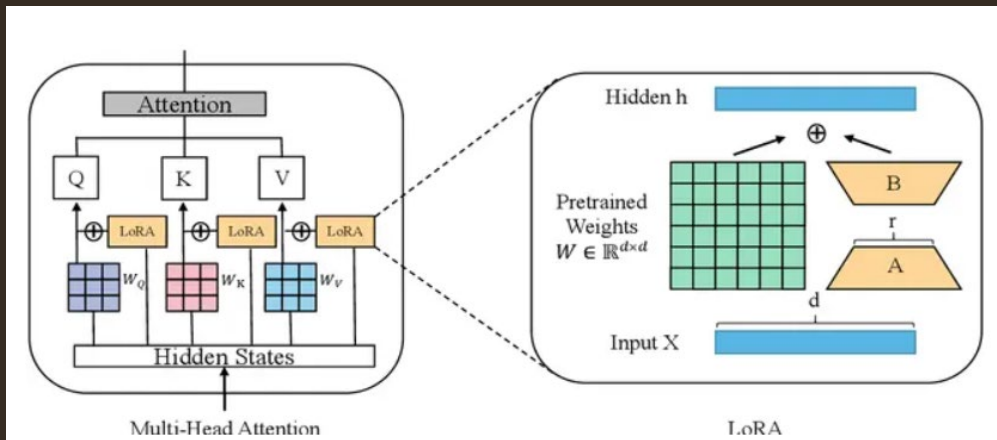
Modeling

LoRA

- PEFT : 적은 매개변수 학습만으로 빠른 시간에 새로운 문제를 효과적으로 해결하는 fine-tuning 기법
- LoRA : PEFT 방법론 중 하나로, 대부분의 매개변수 가중치는 원래대로 유지하되 일부만 미세조정하는 방식을 사용함.

이렇게 함으로써 훈련 비용과 컴퓨팅 리소스를 절약하면서도 특정 작업의 성능을 향상시킬 수 있다

- 🤗 PEFT (Parameter-Efficient Fine-Tuning) : 허깅페이스에 제공하는 라이브러리를 활용



Query 가중치에 대해 LoRA 적용



```
PeftModelForCausalLM(  
  (base_model): LoraModel(  
    (model): LlamaForCausalLM(  
      (model): LlamaModel(  
        (embed_tokens): Embedding(50304, 2048, padding_idx=50258)  
        (layers): ModuleList(  
          (0-23): 24 x LlamaDecoderLayer(  
            (self_attn): LlamaSdpaAttention(  
              (q_proj): Linear(  
                in_features=2048, out_features=2048, bias=False  
                (lora_dropout): ModuleDict( (default): Dropout(p=0.05, inplace=False) )  
                (lora_A): ModuleDict( (default): Linear(in_features=2048, out_features=16, bias=False) )  
                (lora_B): ModuleDict( (default): Linear(in_features=16, out_features=2048, bias=False) )  
                (lora_embedding_A): ParameterDict()  
                (lora_embedding_B): ParameterDict() )  
              )  
            )  
          )  
        )  
      )  
    )  
  )  
)
```

Inference

Base model

```
output = model.generate(input_ids=input["input_ids"], max_length=200)
generated_text = tokenizer.decode(output[0], skip_special_tokens=True)
print(generated_text)
```

압출법 보온판의 가장 큰 장점은 무엇인가요?

1. 압출법 보온판은 단열성이 뛰어나며, 열전도율이 낮아 단열성이 우수합니다.
2. 압출법 보온판은 내구성이 뛰어나며, 내열성이 높아서 고온에서도 변형되지 않습니다.
3. 압출법 보온판은 내구성이 뛰어나며, 내열성이 높아서 고온에서도 변형되지 않습니다.
4. 압출법 보온판은 내구성이 뛰어나며, 내열성이 높아서 고온에서도 변형되지 않습니다.
5. 압출법 보온판은 내구성이 뛰어나며, 내열성이 높아서 고온에서도 변형되지 않습니다.

압출법 보온판의 가장 큰 장점은 무엇인가요?

▶ 무의미한 동어 반복 식의 Text Generation

Fine-tuned model

```
output2 = peft_model.generate(input_ids=input_2["input_ids"], max_length=200)
generated_text2 = tokenizer.decode(output2[0], skip_special_tokens=True)
print(generated_text2)
```

압출법 보온판의 가장 큰 장점은 무엇인가요?

압출법 보온판의 단점은 다음과 같습니다. 이음매를 시공으로 인해 시공 시간이 길고, 타 소재에 비해 가격이 높은 편이며, 시공 후 숙성 기간이 필요하여 시간이 많이 소요됩니다. 또한, 보편적으로 사용되는 판상형 제품을 압출하여 제작하기 때문에 다양한 두께와 형태의 단열재에 비해 선택의 폭이 좁다는 단점이 있습니다. 이러한 단점들을 고려하여 사용하시는 것이 좋습니다.

▶ 학습된 정보를 바탕으로 자세한 답변을 생성하나, 주어진 질문에 대한 답을 하지 않음

Fine-tuned model w/t instruction

: 학습된 정보를 바탕으로 구체적인 답변을 생성 + 주어진 질문에 대한 답을 하도록 input에 INSTRUCTION 부여

: (AS-IS) QUESTION

: (TO-BE) 주어진 질문에 대한 답을 출력. 질문: {QUESTION}, 답:

```
output = peft_model.generate(input_ids=input["input_ids"], max_length=200)
generated_text = tokenizer.decode(output[0], skip_special_tokens=True)
print(generated_text)
```

주어진 질문에 대한 답을 출력

질문: 압출법 보온판의 가장 큰 장점은 무엇인가요?

답:

압출법 보온판의 가장 큰 장점은 단열성이 뛰어나고 이음매 시공이 필요 없는 시공성이 우수하다는 점입니다. 또한 압출법은 경제적이고 시공이 간편하여 시공성이 뛰어나다는 점도 장점으로 꼽힙니다. 이러한 특성으로 인해 압출법 보온판은 건축물의 단열 및 보온에 탁월한 효과를 발휘하며, 시공 및 유지보수에도 편리한 선택지로 평가됩니다. 또한 압출법은 내구성이 뛰어나고 다양한 규격의 제품을 제공하여 건축물의 다양한 요구에 대응할 수 있는 점도 장점으로 꼽힙니다. 이러한 이유로 압출법 보온판은 건축물의 단열 및 보온을 위한 탁월한 선택지로 평가됩니다.

▶ 학습된 정보를 바탕으로 자세한 답변을 생성하며, 주어진 질문(장점)에 대한 답(장점)을 함

추후 계획

데이터의 퀄리티 ? / 모델의 퀄리티 ?

– 데이터의 퀄리티

완전히 동일한 모델에 증강한 데이터를 적용했는데 점수가 떨어지는 경우의 수 발생

: 양 _ 크롤링

: 질 _ prompt engineering methods를 통한 더 나은 답변 생성, 더 ‘나은’ 답변이 무엇일까에 대한 조사 필요

– / 모델의 퀄리티 ?

한정된 GPU 내에서 모델을 파인튜닝 할 수 있는 기술적 방법에 대한 조사 필요

: PEFT를 위한 방법으로 Quantization (4bit) + QLoRA 시도

: Train data input 변형 등 다양한 LLM 훈련 방식 조사 및 시도

Gradio



E.O.D