이화여자대학교

도배 하자 질의 응답 처리

DACON 한솔데코 시즌2 AI 경진대회

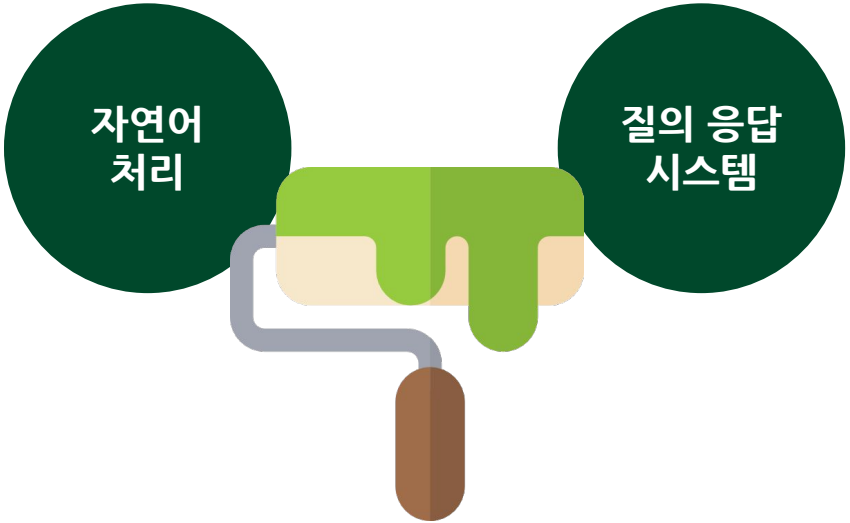
EURON 5기 중급 하자하자!팀 | 김유민 박은혜 이아영

EWHA,
THE FUTURE
WE CREATE

목차

EWHA,
THE FUTURE
WE CREATE

문제 정의 및 배경



한솔데코

데이터 수집 및 전처리

- 데이터 수집
  공모전 주최측에서 제공한 train, test 데이터 사용

train data

id	질문_1	질문_2	category	답변_1	답변_2	답변_3	답변_4	답변_5
TRAIN_000	면진장치기	면진장치이	건축구조	면진장치란	면진장치란	면진장치란	면진장치는	면진장치는
TRAIN_001	내진설계의	내진설계이	건축구조	내진 설계의	내진설계이	내진설계이	내진설계이	내진 설계에
TRAIN_002	철골구조의	철골구조의	건축구조	철골구조는	철골구조의	철골구조의	철골구조의	철골구조의
TRAIN_003	철골철근 ;	철골철근 ;	건축구조	철근철골콘	철골철근콘	철골철근 ;	철골철근콘	철골철근 콘

test data

id	질문
TEST_000	방청 페인트의 종류에는 어떤 것들이 있는지 알고 계신가요? 또한, 원목사이딩을 사용하
TEST_001	도배지에 녹은 자국이 발생하는 주된 원인과 그 해결 방법은 무엇인가요?
TEST_002	큐블럭의 단점을 알려주세요. 또한, 압출법 단열판을 사용하는 것의 장점은 무엇인가요?
TEST_003	철골구조를 사용하는 고층 건물에서, 단열 효과를 높이기 위한 시공 방법은 무엇이 있을

데이터 수집 및 전처리

데이터 전처리

for문을 활용해 질문과 답변이 1:1 대응이 되도록 데이터 포맷팅

	question	answer
0	면진장치가 뭐야?	면진장치란 지반에서 오는 진동 에너지를 흡수하여 건물에 주는 진동을 줄여주는 진동 ...
1	면진장치가 뭐야?	면진장치란 건물의 지반에서 발생하는 진동 에너지를 흡수하여 건물을 보호하고, 진동을...
2	면진장치가 뭐야?	면진장치란 지반으로부터 발생하는 진동 에너지를 흡수하여 건물에 전달되는 진동을 줄여...

토큰화/인코딩

데이터셋의 질문과 답변을 모델의 입력으로 사용하기 위해 텍스트를 토큰화하고 숫자로 변환

```
[tensor([[ 9411,  8265, 37765, 46651,  7991,   406,     1,  9411,  8265, 20725,
           7374,  9027,  7599,  9023, 14472, 15898, 14820, 33220, 36928, 10764,
           9166, 11818, 28037, 10090, 15898, 34062, 20725, 21154]]),
```

데이터 수집 및 전처리

- **데이터 증강**
  train data 양을 늘려 모델의 정확도를 개선하기 위해 2가지 방법으로 진행하여 데이터 증강

  - **MLM data augmentation**
    Masked Language Modeling 방식으로 학습 모델을 학습한 후에 새로운 문장의 일부에 마스킹을 적용하고 인퍼런스를 적용해 마스킹된 부분에 알맞는 새로운 토큰을 후보로 생성

기존 문장	증강된 문장
면진장치란 지반에서 오는 진동 에너지를 흡수하여 건물에 주는 진동을 줄여주는 진동 격리장치입니다.	면진장치란 지반에서 나오는 진동 에너지를 전달하여 건물에서의 진동을 줄여주는 진동의장치다.

데이터 수집 및 전처리

- **BERT augmentation**
  Bert based 모델을 활용하여, 의미상 자연스러운 토큰을 삽입하거나 대체(masking, insertion)하는 형식으로 문장 augmentation 수행

Masking	기존	면진장치란 건물의 지반에서 발생하는 진동 에너지를 흡수하여 건물을 보호하고, 진동을 줄여주는 장치입니다. 주로 지진이나 기타 지반의 진동으로 인한 피해를 방지하기 위해 사용됩니다.
	증강	면진장치란 건물의 지반에서 발생하는 진동 에너지를 막아 건물을 보호하고, 최대한 줄여주는데 주로 지진이나 기타 지반의 진동으로 인한 피해를 방지하기 위해 사용됩니다.
Insertion	기존	면진장치란 지반에서 오는 진동 에너지를 흡수하여 건물에 주는 진동을 줄여주는 진동 격리장치입니다.
	증강	면진장치란 지반에서 전해 오는 진동 에너지를 흡수하여 건물에 주는 진동을 최대한 줄여주는 진동 격리장치입니다.

모델 개요 및 학습 방법

KoGPT 모델

gpt-3 기반 한국어 언어 생성모델

- 한국어를 사전적, 문맥적으로 이해하고 이용자가 원하는 결과값을 보여줌
- 60억개의 매개변수와 2000억개 토큰(token)의 한국어 데이터를 바탕으로 구축
- 주어진 텍스트의 다음 단어 예측 가능

```
model = GPT2LMHeadModel.from_pretrained('skt/kogpt2-base-v2')
```

모델 개요 및 학습 방법

KoGPT 모델 하이퍼 파라미터

[하이퍼파라미터 선택]

- learning_rate(학습률) : 0.0001
- epoch(에포크) : 12 -> 5 (GPU 문제로 임시적으로 5로 학습)
- batch_size=32
- (cos_lr=True) : 학습률을 cosine 함수의 형태로 조절하여 초기에는 높은 학습률로 빠르게 학습을 진행하고, 점차적으로 학습률을 줄여가며 안정적인 학습을 가능하게 함

[최적화 알고리즘]

- AdamW로 설정 : 가중치 감쇠(weight decay)를 적용하여 더욱 효과적인 학습을 할 수 있도록 함

결과 분석 및 평가 지표

평가 산식 : Cosine Similarity

- 내부 곱 공간의 두 벡터간의 유사성 측정
- 생성된 답변을 **Sentence Transformer** 모델을 이용하여 512 차원의 Embedding Vector로 변환한 후, 변환된 벡터와의 코사인 유사도 계산
- (코사인 유사도 값이 0보다 작은 경우 0으로 간주)

결과 분석 및 평가 지표

결과

1차 시도

```
Epoch 1 - Avg Loss: 3.4892: 100%|██████████| 6440/6440 [08:29<00:00, 12.65it/s]
Epoch 1/5, Average Loss: 3.4892132444133668
Epoch 2 - Avg Loss: 2.2820: 100%|██████████| 6440/6440 [08:28<00:00, 12.66it/s]
Epoch 2/5, Average Loss: 2.2820136704013585
Epoch 3 - Avg Loss: 1.4637: 100%|██████████| 6440/6440 [08:26<00:00, 12.72it/s]
Epoch 3/5, Average Loss: 1.4637291784982505
Epoch 4 - Avg Loss: 0.9715: 100%|██████████| 6440/6440 [08:29<00:00, 12.63it/s]
Epoch 4/5, Average Loss: 0.9714992555020296
Epoch 5 - Avg Loss: 0.6692: 100%|██████████| 6440/6440 [08:34<00:00, 12.52it/s]
Epoch 5/5, Average Loss: 0.6692107720913724
('./hansoldeco-kogpt2/tokenizer_config.json',
 './hansoldeco-kogpt2/special_tokens_map.json',
 './hansoldeco-kogpt2/tokenizer.json')
```

하이퍼파라미터 튜닝 및 데이터 증강 후

```
Epoch 1 - Avg Loss: 2.7380: 100%|██████████| 25152/25152 [32:27<00:00, 12.92it/s]
Epoch 1/5, Average Loss: 2.7380280317021795
Epoch 2 - Avg Loss: 1.3808: 100%|██████████| 25152/25152 [32:22<00:00, 12.95it/s]
Epoch 2/5, Average Loss: 1.3807971935962746
Epoch 3 - Avg Loss: 0.9591: 100%|██████████| 25152/25152 [32:35<00:00, 12.86it/s]
Epoch 3/5, Average Loss: 0.959116224020319
Epoch 4 - Avg Loss: 0.7616: 100%|██████████| 25152/25152 [32:48<00:00, 12.78it/s]
Epoch 4/5, Average Loss: 0.7615894759303425
Epoch 5 - Avg Loss: 0.6424: 100%|██████████| 25152/25152 [32:40<00:00, 12.83it/s]
Epoch 5/5, Average Loss: 0.6423967647477012
('./hansoldeco-kogpt2/tokenizer_config.json',
 './hansoldeco-kogpt2/special_tokens_map.json',
 './hansoldeco-kogpt2/tokenizer.json')
```

- 1차 시도때보다 average loss가 0.02 정도 낮아짐
- 리더보드 제출 결과도 0.3591534998 -> 0.447102204로 높아짐

시스템 구현 계획 및 기술적 이슈

계획

- **다양한 하이퍼파라미터를 시도**하고 모델을 개선시켜 loss를 줄이고 모델의 성능을 더 높일 예정

기술적 이슈

- 학습하는데 **시간이 매우 오래 걸려** epoch를 키워서 테스트하는데 어려움을 겪고 있음
  이로 인해 하이퍼파라미터 조정이 어려움
- 증강한 데이터의 **정제** 작업

결론 및 제언

결론 및 제언

- 하이퍼파라미터 튜닝 및 데이터 증강한 결과 1차 시도때보다 높은 성능을 보였음
- 그러나, 코랩 환경의 GPU 한계로 epoch를 낮게 설정해 전체적으로 점수가 낮게 나온 것으로 보임
- 다음주까지 이 문제를 보완하고 하이퍼 파라미터 튜닝, 데이터 정제 작업을 진행해 성능을 더 높일 예정