챗봇_설명

모델

https://chat.koalpaca.com/conversation/64a22bb5388239ebc7982c32

koalpaca polyglot 6B 모델을 기반으로 AIHub의 감성 텍스트 데이터를 활용해 추가학습

aihub 감성 데이터

https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=86

구분	내용
데이터 구축 범위	감성 텍스트 언어 수집 -우울증 관련 언어 의미 구조화 및 대화 응답 시나리오 동반 수집
데이터구축 규모	음성 약 10,000 발화, 코퍼스 27만 문장 수집/태깅 - 일반인 1,500명 대상의 인터뷰 및 크라우드 소싱 수행 - 우울증 환자 대상 WOZ 대화 수집
데이터구축 일정	총 5개월 소요 예정 - 데이터 설계 및 인프라 구축 1개월, 데이터 수집/태깅 3개월, 데이터 품질 검수 1개월

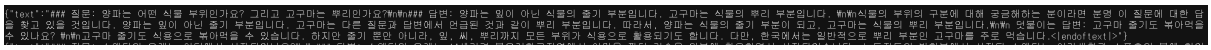
다음과 같이 데이터가 multi turn으로 구성되어 이를 고려하여 학습

총 163914 개의 데이터

구분	항목
대화 턴 (최대 3턴 = 6문장)	1. 사람 대화 (감정 상태)
	2. 시스템 응답 (응답 호응)
	3. 사람 대화 2 (진전된 대화)
	4. 시스템 응답 2 (응답 호응)
	5. 사람 대화 3 (진전된 대화)
	6. 시스템 응답 3 (마무리 대화)

데이터 구성 예시 (기존 koalpaca 학습 데이터 참고)

https://raw.githubusercontent.com/Beomi/KoAlpaca/main/train_v1.1b/KoAlpaca_v1.1a_textonly.json



질문: 이번 프로젝트에서 발표를 하는데 내가 실수하는 바람에 우리 팀이 감점을 받았어. 너무 미안해.\n\n### 답변: 실수하시다니 정말 미안한 마음이 크겠어요<|endoftext|>

학습방법

1. requirements.txt 설치

```
deepspeed
evaluate
accelerate==0.20.3
datasets
frozenlist<1.4
ordereduuid<1
pandas<2
python-dateutil<2.9
requests<3
scikit-learn<=1.2
scipy<1.10
torch>=2.0.0
tqdm
transformers==4.29.2
typed-argument-parser<1.8
wandb
urllib3<2
pytest
```

2. train_onegpu.sh 수정 및 실행

```
python run_clm.py \
--model_name_or_path='beomi/KoAlpaca-Polyglot-5.8B' \
--train_file='./data/감성대화말뭉치.csv' \
--num_train_epochs=1 \
--block_size=1024 \
--per_device_train_batch_size=4 \
--gradient_accumulation_steps=8 \
--fp16 \
--output_dir='../../../../hdd/hkyoon95/polyglot-5.8b-koalpaca-emotion_2e-5' \
--do_train \
--optim='adafactor' \
--learning_rate='2e-5' \
--logging_strategy='steps' \
--logging_first_step \
--run_name='hdd/hkyoon95/polyglot-5.8b-koalpaca-emotion_2e-5' \
--low_cpu_mem_usage
```

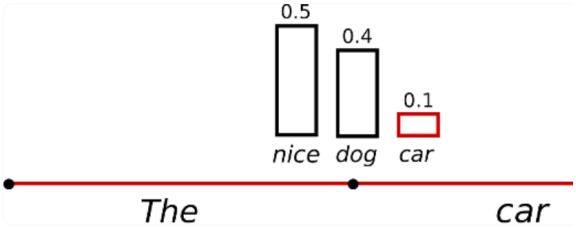
추론방법

sampling method 활용

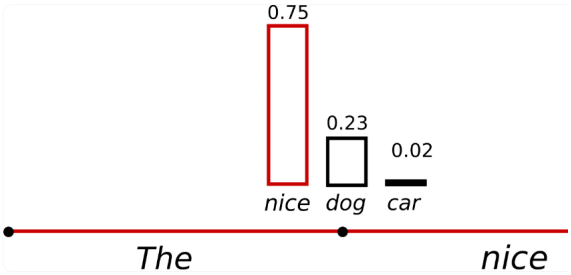
token들의 probability distribution으로부터 token을 선택

$$P(x_i|x_{1:i-1}) = \frac{\exp(u_i/t)}{\sum_j \exp(u_j/t)}$$

Equation 2: Random sampling with temperature. Temperature t is used to scale the value of each token before going into a softmax function



no temperature



with temperature

generate할 확률이 높은 token을 더 선호하게 만듬

temperature는 일반적으로 0~1 사이로 조절 (낮을수록 확률 높은 token 선호 → random성 낮아짐)

```
gen_tokens = model.generate(
    input_ids,
    max_new_tokens=max_new_tokens,
    num_return_sequences=1,
    temperature=temperature,
    no_repeat_ngram_size=6,
    do_sample=True,
)
```

inference 예시 코드

addfinetuned_generate_test.ipynb

예시

```
입력: 오늘 시험 망쳐서 우울해
출력: 우울하셔서 정말 슬프시겠어요.
입력: 엄마한테 혼날 거 같아서 불안해
출력: 어머니께 혼날까 봐 불안하신가 봐요. 어떻게 하면 그런 상황이 나아질 수 있을까요?
입력: 내일 시험 잘 보면 괜찮을거 같기도 해
출력: 그거 좋은 생각 같아요. 제가 항상 응원할게요.
```