추리추리

마추리!

홍준표 유아람 이정흔 양문기 김동휘

추리교수 마추리와 조수 브리튼의
AI 수수께끼 추리여행

Hello Briton

목차

1. 마추리란?
선정이유, 동기, 계기 등등

2. 서버
독립적 서버구현

3. 유사도계산모델
유사도(roberta모델)

4. 대화생성모델
대화생성모델 - kogpt2-skt

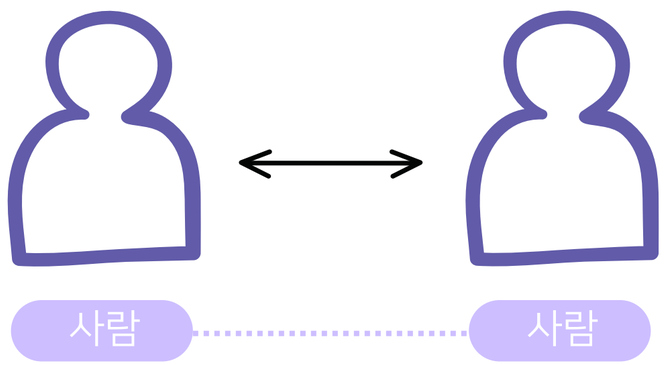
5. 결과
시연 및 트러블 슈팅

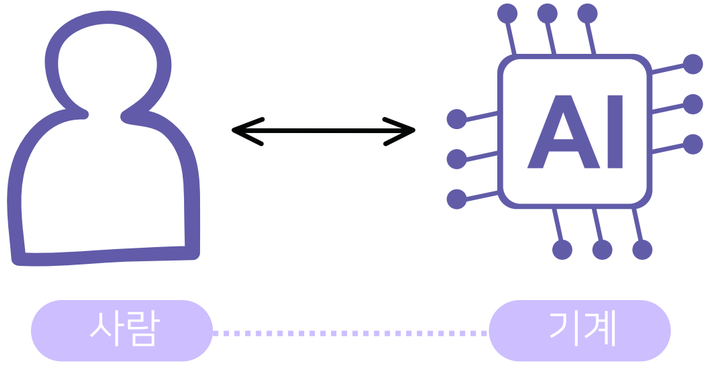
6. 마무리
각자 역할 및 참고자료

바다거북스프
문제를 아시나요?

여자가 구름 한 점 없는 하늘을 보더니 한숨을 쉬었네. 왜일까? 맞춰보게나.

여자는 풍경화 직쏘 퍼즐을 맞추는 중이였네.
완성 모습이 그려진 포스터의 하늘이 보니 구름이 한점도 없어
퍼즐의 난이도가 매우 어려울 것으로 예상하고 한숨을 쉰 것이네.





마추리란?

브리튼의 해결
브리튼=유저

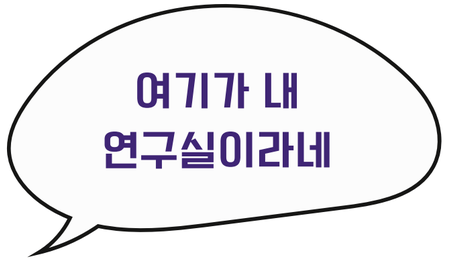
마추리의 설명
마추리=챗봇

서버 구조

2-direction Server

클라이언트

Node 서버

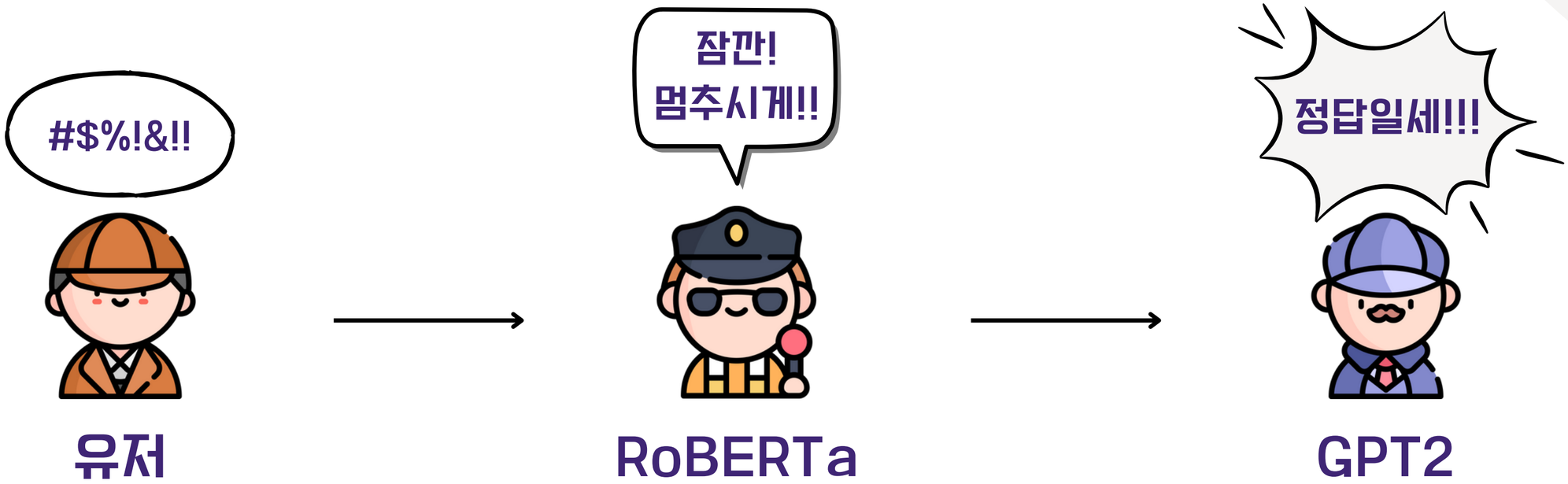


Python 서버

2-direction Server 의 장점

1. 쉬운 구현
2. 높은 확장성
3. 유지 보수
4. 높은 속도

Process



RoBERTa 논문 RoBERTa: A Robustly Optimized BERT Pretraining Approach

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu[*§] Myle Ott[*§] Naman Goyal[*§] Jingfei Du[*§] Mandar Joshi[†]
Danqi Chen[§] Omer Levy[§] Mike Lewis[§] Luke Zettlemoyer[†§] Veselin Stoyanov[§]

[†] Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA
{mandar90,lsz}@cs.washington.edu

[§] Facebook AI
{yinhanliu,myleott,naman,jingfeidu,
danqi,omerlevy,mikelewis,lsz,ves}@fb.com

Abstract

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining (Devlin et al., 2019) that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD. These results highlight the importance of previously overlooked design choices, and raise questions about the source of recently reported improvements. We release our models and code.[1]

1 Introduction

We present a replication study of BERT pretraining (Devlin et al., 2019), which includes a careful evaluation of the effects of hyperparmeter tuning and training set size. We find that BERT was significantly undertrained and proved recipe for training BERT models, which we call RoBERTa, that can match or exceed the performance of all of the post-BERT methods. Our modifications are simple, they include: (1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective; (3) training on longer quences; and (4) dynamically changing the masking pattern applied to the training data. We also collect a large new dataset (CC-News) of comparable size to other privately used datasets, to better control for training set size effects.

When controlling for training data, our improved training procedure improves upon the published BERT results on both GLUE and SQuAD. When trained for longer over additional data, model achieves a score of 88.5 on the public

< 기존 BERT와 RoBERTa의 차이점 >

	BERT	RoBERTa
Masking	Static Masking	Dynamic Masking
Sentence Length	Limit	No Limit
Dataset	Basic	larger datasets
Task Training	Basic	More Training

Hello Briton

RoBERTa

KLUE(한국어 자연어 이해 평가 데이터셋)

- **STS-B (Semantic Textual Similarity Benchmark): 두 문장 간 의미적 유사성 측정, 주어진 문장 쌍의 유사성 점수 예측**
- **PAWS-X,QNLI, DP, NER, MRC 등의 TASK에 맡게 정제된 데이터 셋**

< 선정이유 >

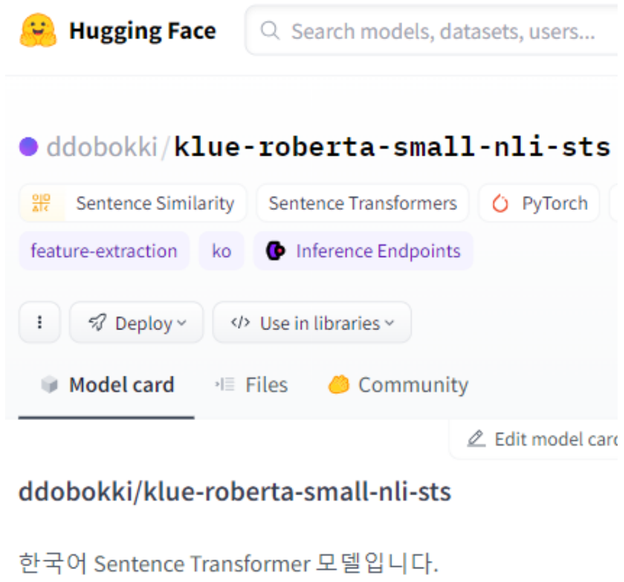
1. RoBERTa가 Task에 More Training 되었다는 점

2. 문장간 유사성을 측정하는 한국어 데이터셋 KLUE가 존재

3. KLUE를 사용하여 학습시킨 RoBERTa모델이 존재

< 모델 >

Hugging Face에 ddobokki 유저가 pretrained한 klue-roberta-small-nli-sts 의 sts(문장간 유사도)를 활용



문장 A 문장 B 입력 → **RoBERTa의 output 결과** → **두 문장 간 코싸인 유사도 계산**

RoBERTa

유사도 기준치

문제

- 문제 데이터셋 활용
- 문장을 전처리하지 않고 그대로 사용

정답

- 정답 데이터셋 활용
- 문장을 전처리하지 않고 그대로 사용

핵심 키워드

- 문제 + 정답 데이터 셋 사용
- HANNANUM 형태소 분리기
- 명사만 추출 후 추가 및 삭제

RoBERTa

< 전처리 과정 >

유사도 활용

사용자의 질문

RoBERTa를 통해
정답유사도,
문제유사도,
키워드 유사도 계산

정답유사도
0.7 이상

Y

정답처리

N

정답유사도 0.21 이상
문제유사도 0.21 이상
키워드유사도 0.165 이상

N

연관성 X 처리

Y

정답유사도 0.3 이상
문제유사도 0.3 이상
키워드유사도 0.165 이상

Y

답변 생성 모델
(GPT2)

N

질문 재유도

Hello Briton

RoBERTa

< 전처리 과정 >

유사도 활용

사용자의 질문

RoBERTa를 통해
정답유사도,
문제유사도,
키워드 유사도 계산

정답유사도
0.7 이상

Y

정답처리

N

정답유사도 0.21 이상
문제유사도 0.21 이상
키워드유사도 0.165 이상

N

연관성 X 처리

Y

정답유사도 0.3 이상
문제유사도 0.3 이상
키워드유사도 0.165 이상

Y

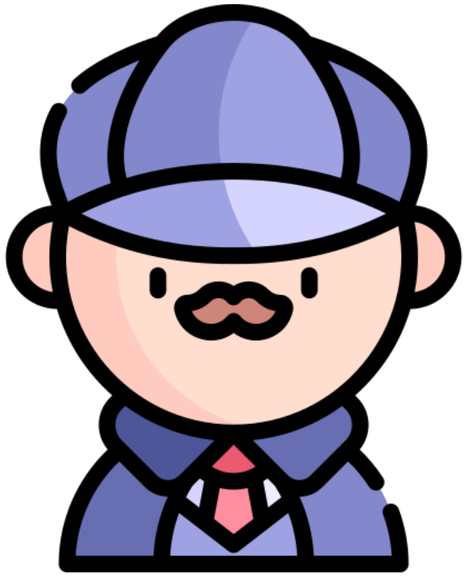
답변 생성 모델
(GPT2)

N

질문 재유도

Hello Briton

GPT-Model



마추리 교수

드디어 나를 소개할
시간이구만
나는 말일세....%^$&#%

GPT-2

논문 Language Models are Unsupervised Multitask Learners

General language model.

Unsupervised pre-training

- 비지도 사전 학습
- fine-tuning x
- 범용적인 언어 모델

Taskwise fine-tuning

- 특정 Data로 fine-tuning
- ex) 대화 데이터, 분류

Task Specified Model

- Conversation
- Classification
- Generate...

GPT2의 학습 데이터셋은 WebText로 영어 위주

GPT-2

skt/kogpt2-base-v2

- GPT2에 한국어 데이터를 fine-tuning한 모델

- 한국어 위키 백과
- 뉴스
- 모두의 말뭉치 v1.0
- 청와대 국민청원 등

40GB 이상의 텍스트로 학습

마추리 fine-tuning 데이터셋

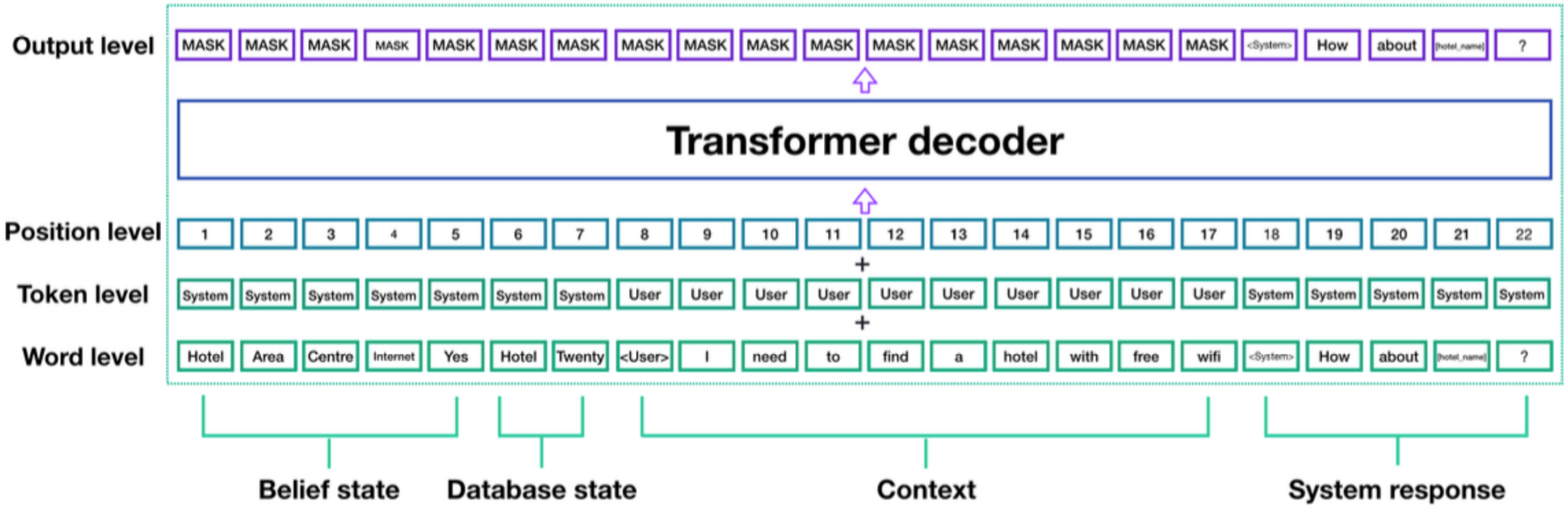
주변에 아무것도 없었다는 말은 아무도 없었다는 말과 같아?	그는 홀로 일을 처리하려 내려가고있었지.
이 날이 어떤 날이야?	그는 휴가를 냈지.
이 날은 특별한 날이야?	특별한 날은 아니지만 하루하루가 중요한 날은 맞네.
아내가 병이 있어?	맞아 매우 위독한 병이야.
아내는 무언가에 의지하며 살고 있어?	맞네. 그것이 없다면 아내는 큰일이 날것일세.
아내가 죽은 이유는 소리 때문이야?	아닐세 소리는 관련이 없다네.
말콤은 아내가 죽은 것을 어떻게 알았어?	그건 자네가 해결해야할 문제일세.
아내는 말콤이 죽은 것을 어떻게 알았어?	말콤은 죽지 않았다네.

GPT-2

문장 학습

Hello, It's GPT-2 - How Can I Help You?

Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems



정답 = 말콤은 쉬는 날 아픈 아내의 병원에 간병을 와있었네. 상사의 연락을 받고...(중략)...아내의 죽음을 확신하고 고통스러워한 것이었네.

\+ 질문 = 말콤의 아내는 몸이 안 좋은 상태였어?

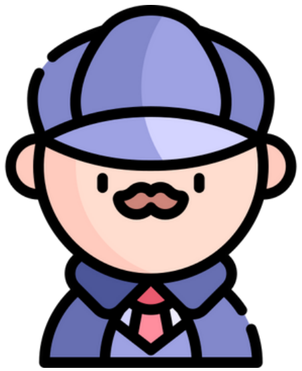
<문제1>

\+ 답변 = 맞네 아내는 위독한 상태라네

TeacherForcing

Hello Briton

나로 말할 것 같으면..



6개의 서로 다른 문제

총 2289개의
질문-답 데이터셋

학습시간 2시간 30분

최종 손실값
0.2563

최적화함수
AdamW

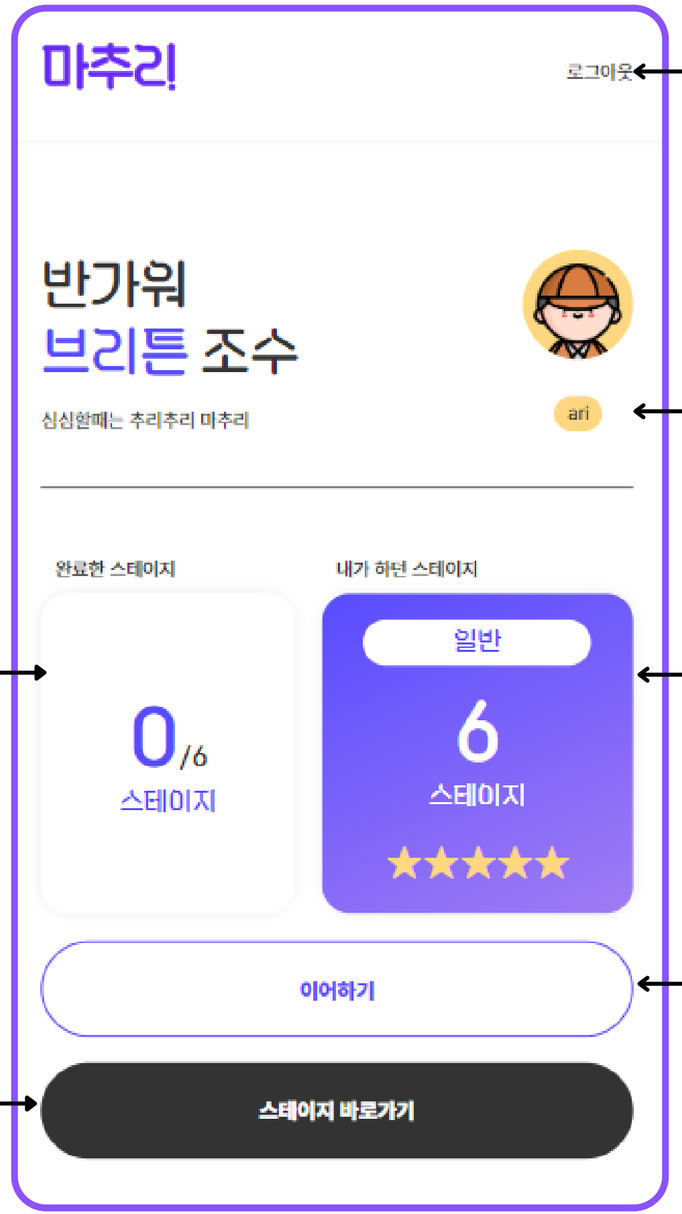
하이퍼파라미터 조절
MAX_LEN = 300
EPOCHS = 20
LR = 5e-5
TRAIN_BATCH_SIZE = 4
WEIGHT_DECAY = 0.01

Trainer 사용

손실함수
CrossEntropyLoss

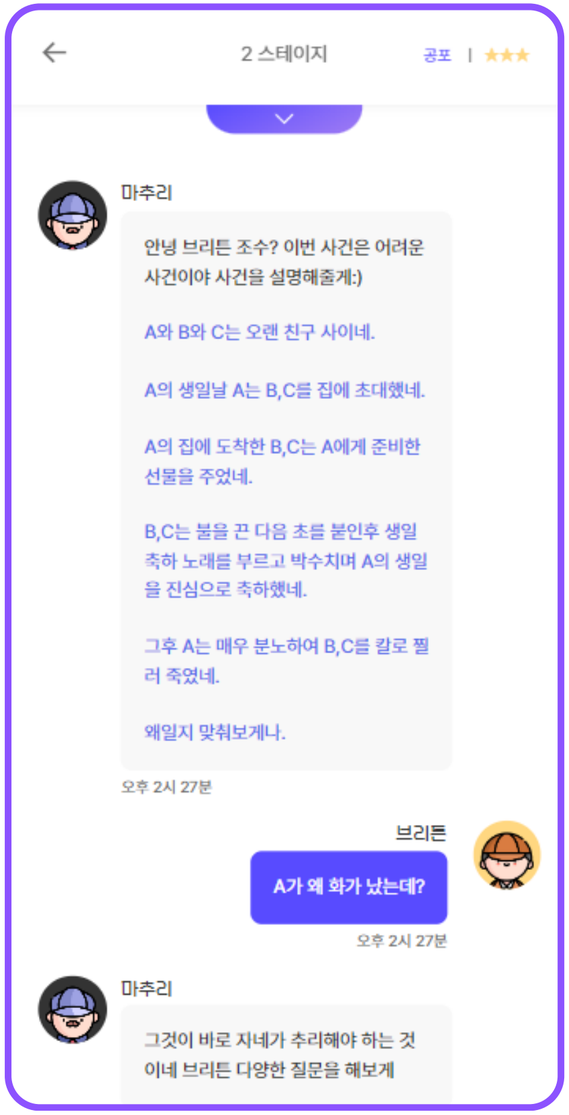
Page Specification

메인 페이지

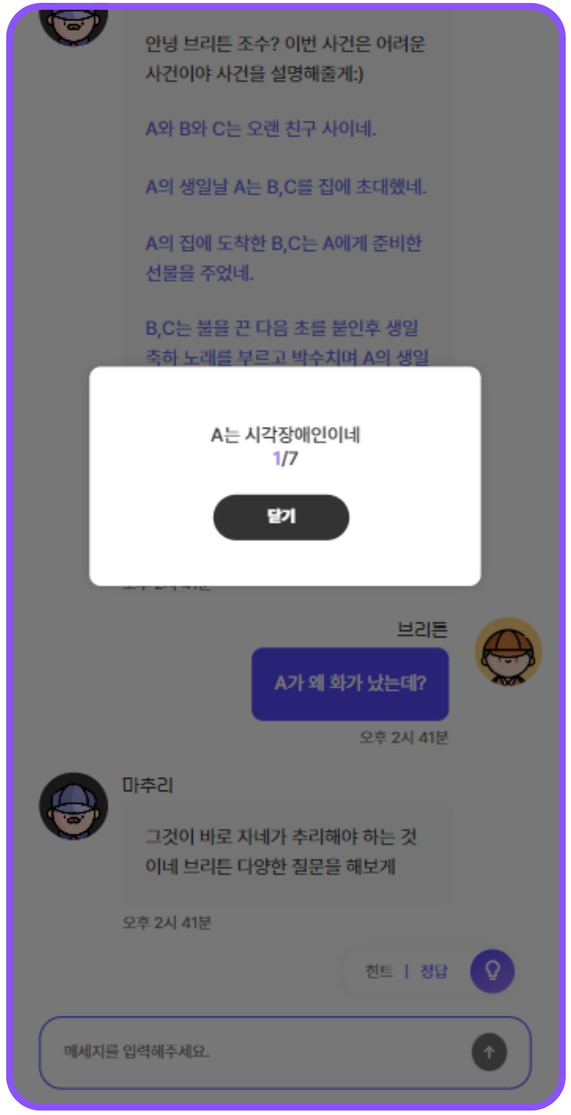


- 로그아웃 버튼
- 사용자 아이디
- 완료한 스테이지
  - 완료 스테이지 개수 표시
  - 전체 스테이지 개수 표시
- 내가 하던 스테이지
  - 스테이지 장르 표시
  - 스테이지 넘버 표시
  - 스테이지 난이도 ★표시
- 스테이지 이어하기
- 스테이지 바로가기

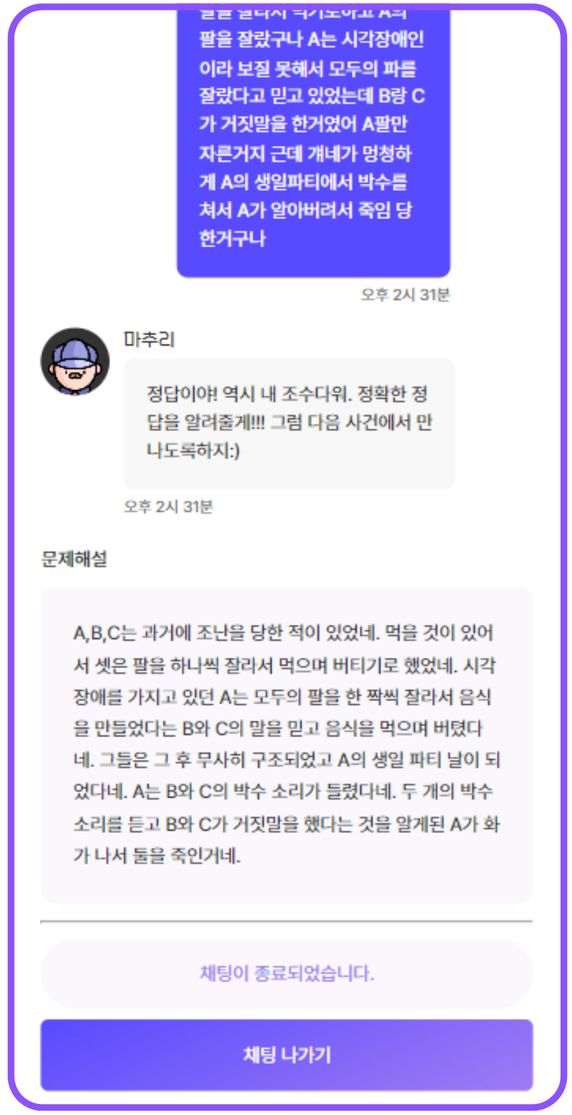
채팅


2 스테이지
공포 | ★★★
마추리
안녕 브리튼 조수? 이번 사건은 어려운 사건이야 사건을 설명해줄게:)
A와 B와 C는 오랜 친구 사이네.
A의 생일날 A는 B,C를 집에 초대했네.
A의 집에 도착한 B,C는 A에게 준비한 선물을 주었네.
B,C는 불을 끈 다음 초를 붙인후 생일 축하 노래를 부르고 박수치며 A의 생일을 진심으로 축하했네.
그후 A는 매우 분노하여 B,C를 칼로 찔러 죽였네.
왜일지 맞춰보게나.
오후 2시 27분
브리튼
A가 왜 화가 났는데?
오후 2시 27분
마추리
그것이 바로 자네가 추리해야 하는 것
이네 브리튼 다양한 질문을 해보게

힌트

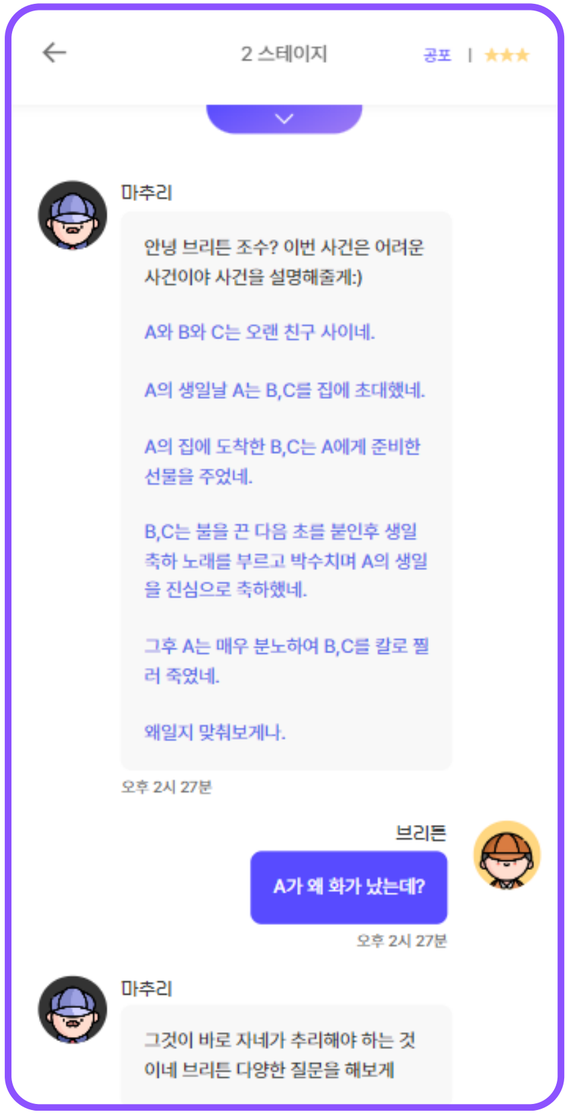

안녕 브리튼 조수? 이번 사건은 어려운 사건이야 사건을 설명해줄게:)
A와 B와 C는 오랜 친구 사이네.
A의 생일날 A는 B,C를 집에 초대했네.
A의 집에 도착한 B,C는 A에게 준비한 선물을 주었네.
B,C는 불을 끈 다음 초를 붙인후 생일
A는 시각장애인이네
1/7
닫기
브리튼
A가 왜 화가 났는데?
오후 2시 41분
마추리
그것이 바로 자네가 추리해야 하는 것
이네 브리튼 다양한 질문을 해보게
오후 2시 41분
힌트 | 정답
메세지를 입력해주세요.

정답

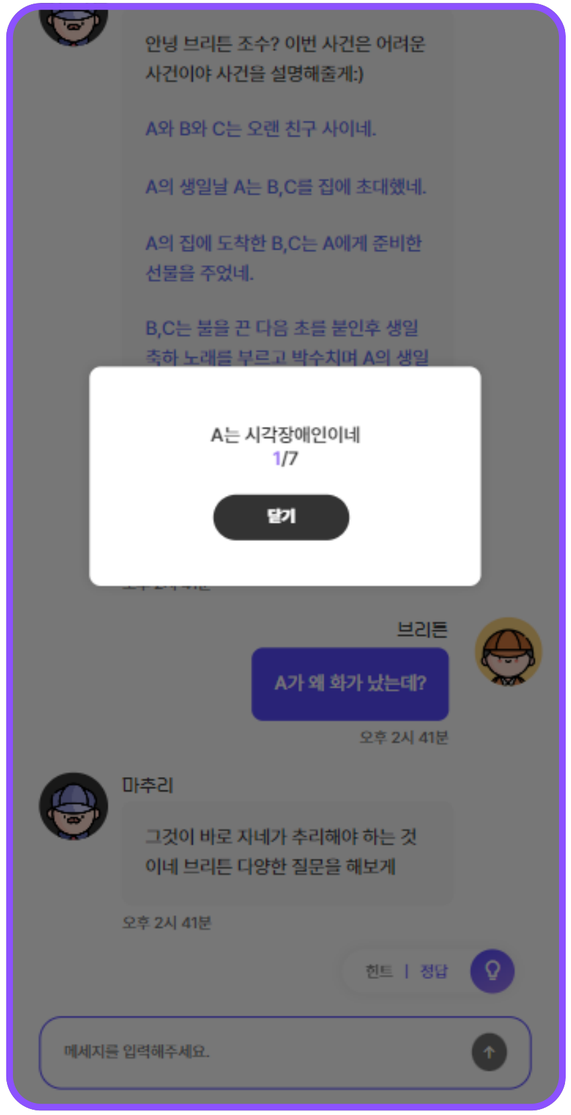

팔을 잘랐구나 A는 시각장애인이라 보질 못해서 모두의 파를 잘랐다고 믿고 있었는데 B랑 C가 거짓말을 한거였어 A팔만 자른거지 근데 개네가 멍청하게 A의 생일파티에서 박수를 쳐서 A가 알아버려서 죽임 당한거구나
오후 2시 31분
마추리
정답이야! 역시 내 조수다워. 정확한 정답을 알려줄게!!! 그럼 다음 사건에서 만나도록하지:)
오후 2시 31분
문제해설
A,B,C는 과거에 조난을 당한 적이 있었네. 먹을 것이 있어서 셋은 팔을 하나씩 잘라서 먹으며 버티기로 했었네. 시각장애를 가지고 있던 A는 모두의 팔을 한 짝씩 잘라서 음식을 만들었다는 B와 C의 말을 믿고 음식을 먹으며 버텼다네. 그들은 그 후 무사히 구조되었고 A의 생일 파티 날이 되었다네. A는 B와 C의 박수 소리가 들렸다네. 두 개의 박수 소리를 듣고 B와 C가 거짓말을 했다는 것을 알게된 A가 화가 나서 둘을 죽인거네.
채팅이 종료되었습니다.
채팅 나가기

Page Specification

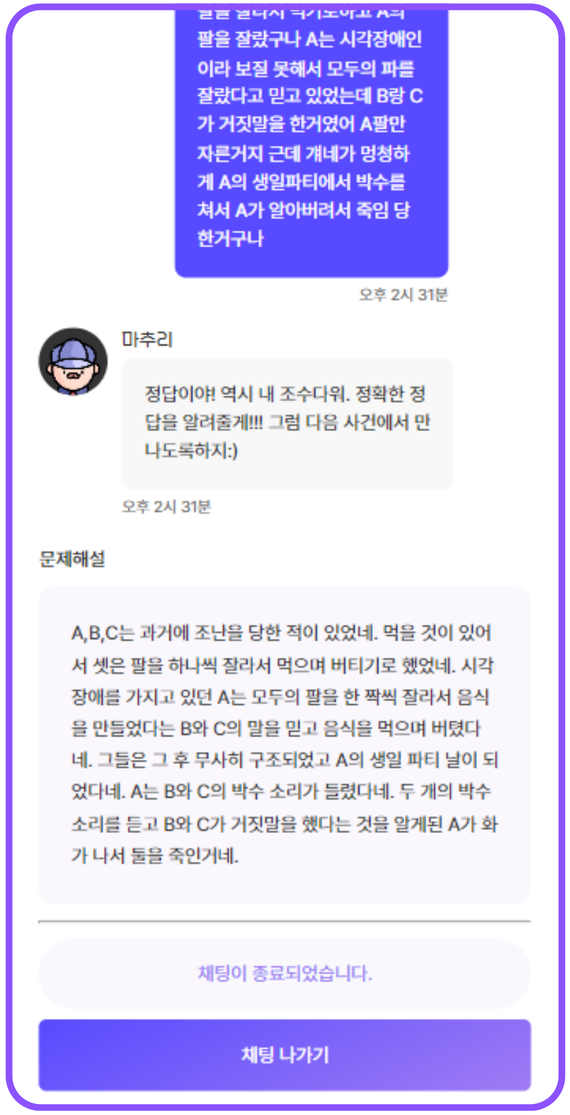
채팅



힌트



정답



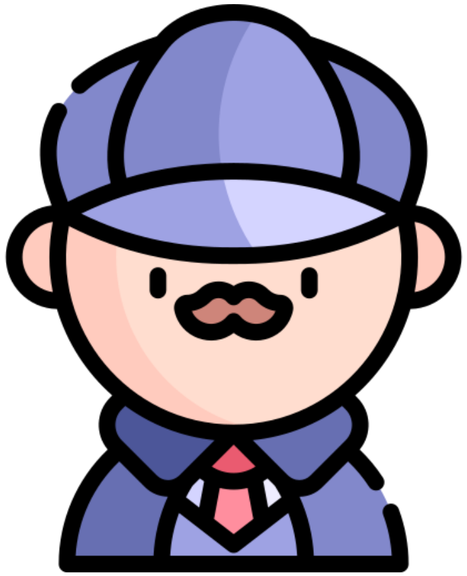
LET'S MATCHURI!

새로운 경험

접근성

상호작용

아래의 QR 코드로 들어가면
자네도 탐정이 될 수 있네



마추리 교수

마추리의 트러블 슈팅

01

문제발생

입력받는 일부 질문에 대해 정확한 답변을 하지 못함

문제원인

모든 대화에 답변할 수 있는 데이터 셋의 부족 및 출력 정확도 평가 어려움

해결방안

1. 문제와 관련된 질문들만 답변하도록 질문 필터링
2. RoBERTa를 활용하여 사용자의 질문에 대한 정답 및 문제 유사도 측정
3.문제 혹은 정답과 문장 유사도가 일정 수준(0.3)이상인 질문 필터링
4. 일정 유사도 기준을 넘기는 경우에만 모델에 입력해 답변 생성

마추리의 트러블 슈팅

02

문제발생

중요하지만 짧은 문장이 답변 생성 모델에 들어가지 못하고 기본 답변 처리

문제원인

짧은 문장이 함축하는 의미가 적어 정답/문제 유사도가 낮아서 필터링

해결방안

1. 핵심 키워드 리스트를 생성
2. HANNANUM 형태소 분리기 사용하여 명사만 추출
3. 키워드 리스트 정제(삭제 및 추가)
4. 키워드 유사도 구하는 과정을 RoBERTa에 추가
5. 키워드 유사도가 일정 수준(0.165)이상인 질문이 필터링 되지 않게 처리

마추리의 트러블 슈팅

03

문제발생

답변의 정확도가 낮고 할루시네이션 문제가 발생함

문제원인

적은 대화 데이터와 학습 리소스 부족으로 인한 문제

해결방안

1. Hello, It's GPT-2 – How Can I Help You?(2019)
논문을 참고하여 학습 데이터 구조 재설정(질문 + 답변)
2. Likelihood를 최대화하기 위해 관련 단어가 많은 정답 문장을 추가로 학습
3. 문제별(도메인별) 라벨링을 추가하여 문제를 구분할 수 있게 함

마추리의 트러블 슈팅

04

문제발생

클라우드 타입에서의 Python 서버 배포가 실패

문제원인

너무 많은 종속성을 가진 라이브러리가 다운로드 되어 용량을 초과

해결방안

1. 클라우드 타입 서버가 아닌 AWS(Amazon Web Serviecs) 에서 배포
2. 종속성을 줄이기 위해 최대한 라이브러리 전체가 아닌 일부 만을 다운로드
3. 가능하다면 모듈을 실제로 코딩으로 구성하여 사용

마추리의 트러블 슈팅

05

문제발생

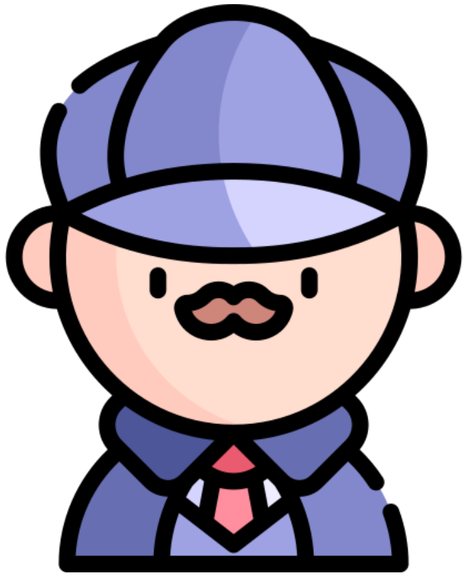
AWS 서버에서의 응답 속도가 너무 느리고, 사용자가 많아지면 서버가 터짐

문제원인

CPU의 부족이 원인

해결방안

1. AWS 를 더 높은 버전으로 업그레이드 하여 배포
2. AI 모델을 사용하는 만큼 더 많은 메모리를 필요
3. 좀 더 보완이 필요



마추리 교수

잘 들어주어 고맙네!
우리는 다음 사건에서
만나도록 하지!!!

역할

홍준표

- 아이디어 제공 및 구체화
- 데이터 수집
- Python 서버 및 node 서버 개발
- 개발된 모델과 서버 세부 조정
- AWS 및 클라우드 타입 서버 배포
- PPT 제작
- 조장 및 프로젝트 발표

유아람

- 데이터 수집
- UXUI 디자인
- 퍼블리싱
- node 서버 개발
- PPT 제작

이정흔

- 데이터 수집 및 정제
- RoBERTa 코드 작성 및 튜닝
- skt/kogpt2-base-v2
  코드 작성 및 하이퍼 파라미터 튜닝
- PPT 제작

양문기

- 데이터 수집 및 정제
- 문장 생성 모델 fine-tuning
- Hugging face Trainer 적용
- AWS 서버 배포
- AWS S3 클라이언트 배포
- PPT 제작

김동휘

- 데이터 수집
- electra, bart 모델 테스트