최종 발표 2020.07.04

심심한 왑닉이

BERT로 챗봇 만들기

김승유 도유진 유승수 최민태 최종문

CONTENTS

01 프로젝트 개요

목표 1. YBIGTA를 위한 인공지능 챗봇을 만들자!

목표 2. 실제 사용 가능한 서비스로 만들자!

01

프로젝트 개요

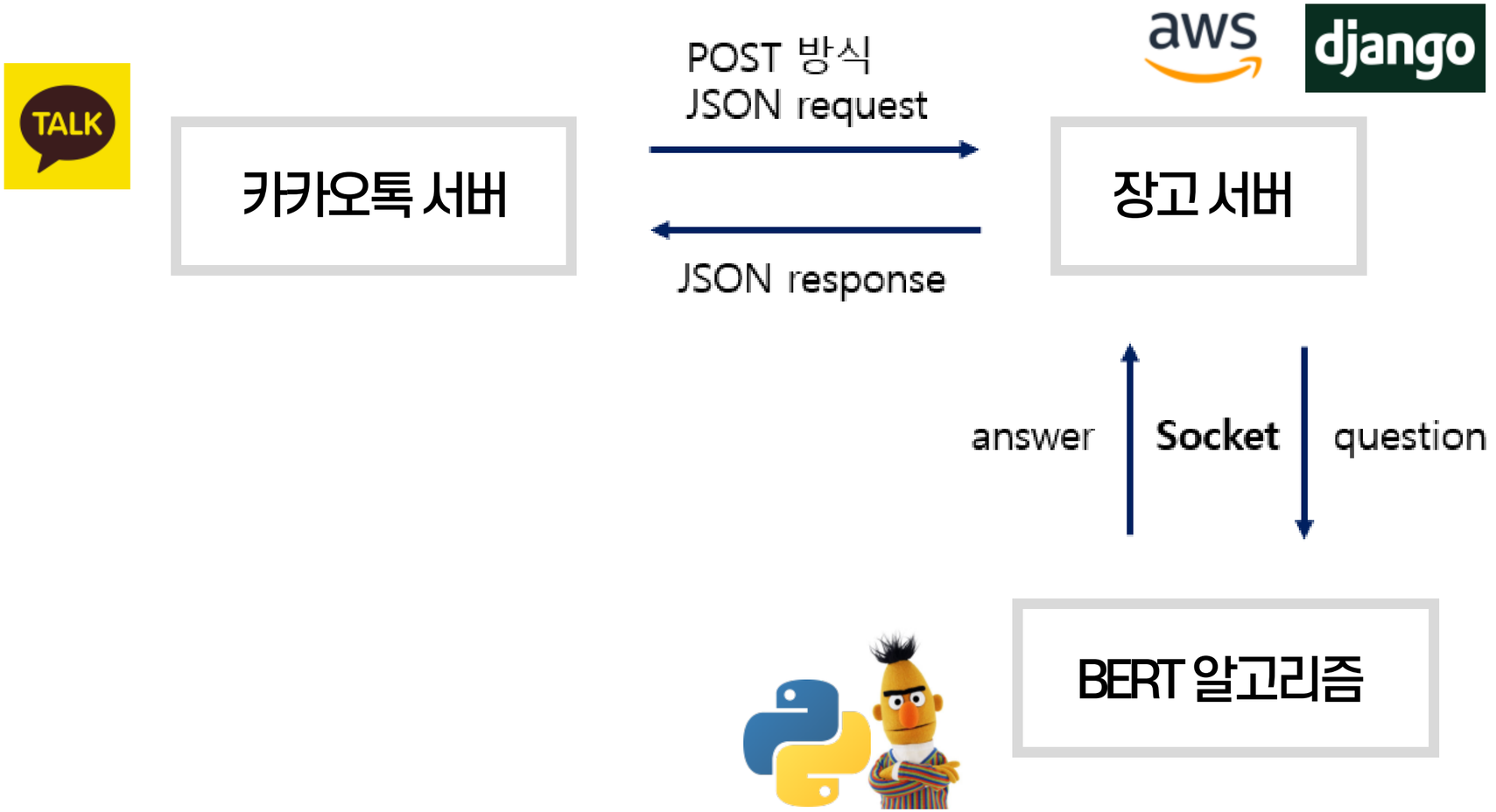
시나리오형 챗봇

- 사용자 발화를 지정하여 **미리 준비된 답변**을 출력
- 제공해야 할 서비스가 정해져 있을 경우에 용이
- **YBIGTA 17기 모집일정, 팀 소개** 등 답변을 준비

대화형 챗봇

- **딥러닝 모델**을 기반으로 한 자연스러운 대화 가능
- **엄청난 비정형 데이터 확보**가 필요
- 비용과 시간이 많이 필요

01 대화형 챗봇 프로세스 개요



02

데이터셋 구하기

일상적인 **대화 데이터**가 필요하다!

A : 밥 먹었어?
B : 아니, 너는 먹었어?
A : 나도 안 먹었어.
B : 같이 먹을래?
A : 싫어.

02

데이터셋 구하기

1. Chatbot_data_for_Korean v1.0

- 문답 페어 11,876개
- 사랑, 이별, 위로 관련 주제
- 일상적 대화 수집에 한계가 존재

11824 lines (11824 sloc) | 869 KB

We can't make this file beautiful and searchable because it's too large.

```
Q,A,label
12시 땡!,하루가 또 가네요.,0
1지망 학교 떨어졌어,위로해 드립니다.,0
3박4일 놀러가고 싶다,여행은 언제나 좋죠.,0
3박4일 정도 놀러가고 싶다,여행은 언제나 좋죠.,0
PPL 심하네,눈살이 찌푸려지죠.,0
SD카드 망가졌어,다시 새로 사는 게 마음 편해요.,0
SD카드 안돼,다시 새로 사는 게 마음 편해요.,0
SNS 맞팔 왜 안하지ㅠㅠ,잘 모르고 있을 수도 있어요.,0
SNS 시간낭비인 거 아는데 매일 하는 중,시간을 정하고 해보세요.,0
SNS 시간낭비인데 자꾸 보게됨,시간을 정하고 해보세요.,0
SNS보면 나만 빼고 다 행복해보여,자랑하는 자리니까요.,0
가끔 궁금해,그 사람도 그럴 거예요.,0
가끔 뭐하는지 궁금해,그 사람도 그럴 거예요.,0
가끔은 혼자인게 좋다,혼자를 즐기세요.,0
가난한 자의 설움,돈은 다시 들어올 거예요.,0
가만 있어도 땀난다,땀을 식혀주세요.,0
가상화폐 쫄딱 망함,어서 잊고 새출발 하세요.,0
가스불 켜고 나갔어,빨리 집에 돌아가서 끄고 나오세요.,0
가스불 켜놓고 나온거 같아,빨리 집에 돌아가서 끄고 나오세요.,0
가스비 너무 많이 나왔다.,다음 달에는 더 절약해봐요.,0
```

출처 : https://github.com/songys/Chatbot_data/blob/master/ChatbotData%20.csv

02

데이터셋 구하기

2. 카카오톡 채팅방 데이터

- 카카오톡 채팅 내보내기
- 채팅방 5개에서 40,000여 대화 수집
- 이모티콘, 사진, 송금 등 전처리

```
2020년 1월 25일 토요일
2020. 1. 25. 09:31, 😈오승은😈 : ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
2020. 1. 25. 09:31, 😈오승은😈 : 친구들
2020. 1. 25. 09:31, 😈오승은😈 : 새해복많이받으렴~
2020. 1. 25. 09:31, 😈오승은😈 : 해피설날~
2020. 1. 25. 09:33, 😈오승은😈 : 사진
2020. 1. 25. 09:34, 😈오승은😈 : 요 기여운 자식이 올해도 너네를 응원한대~
2020. 1. 25. 09:34, 😈오승은😈 : 기여운놈!!
2020. 1. 25. 09:37, 도유진 Yujin Doh : 사진
2020. 1. 25. 09:38, 도유진 Yujin Doh : 이모티콘 새해복 많이받으시게~^^*
2020. 1. 25. 09:38, 도유진 Yujin Doh : 요 예쁜 너석이 올해도 니네들 일 잘 풀리도록
2020. 1. 25. 09:41, 🇪🇸 거절전문가 정유빈선생님 : ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
2020. 1. 25. 09:41, 🇪🇸 거절전문가 정유빈선생님 : 다들 메리설되렴

2020년 1월 27일 월요일
2020. 1. 27. 22:05, 🇪🇸 떼디미 : 이모티콘 내 새로운 춘배 임티어때
2020. 1. 27. 22:05, 🇪🇸 떼디미 : 보고싶어아이드라
2020. 1. 27. 22:25, 도유진 Yujin Doh : 엉어.귀엽다아~~~
2020. 1. 27. 22:25, 도유진 Yujin Doh : 너 서울?
2020. 1. 27. 23:25, 🇪🇸 떼디미 : 내일 갑니다~~
2020. 1. 27. 23:25, 🇪🇸 떼디미 : 이모티콘

2020년 1월 29일 수요일
2020. 1. 29. 11:41, 도유진 Yujin Doh : 친구야!
```

02

01. 프로젝트 개요

02. 데이터셋 구하기

03. 알고리즘 구현

04. 서버 구축 및 카카오톡 플러스친구 연동

05. 챗봇 시연

데이터셋 구하기

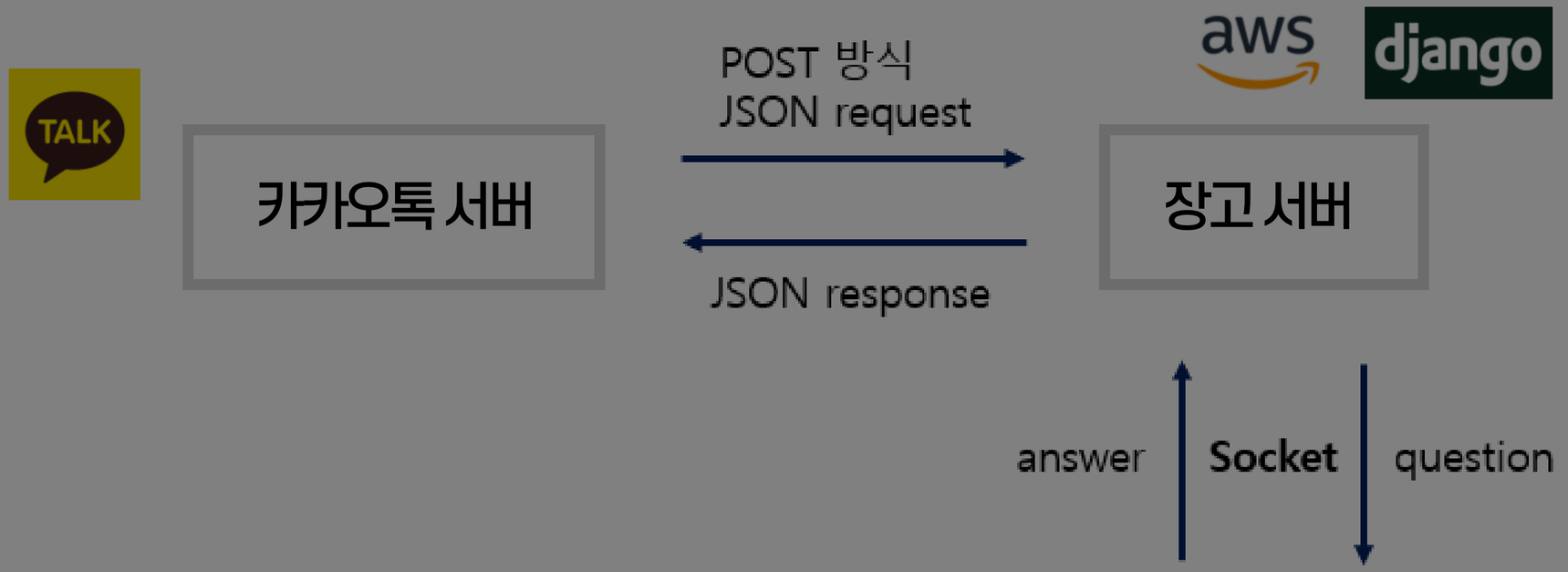
3. 코넬 대학 영화 자막 데이터(영어)

- 영어 대화 데이터
- 일일이 파파고로 번역
- 번역체 + 너무 긴 호흡의 대화

사용하지 않음!

5002	Good. Now the tenors. Fourth beat of the first measure - C. Con-fu-ta-tis. Second measure, fourth beat on D. Ma-le-dic-tis. All right?	Yes.
5003	Yes.	Fourth measure, second beat - F. Flam-mis a-cri-bus ad-dic-tis, flam-mis a-cri-bus ad-dic-tis.
5004	Now the orchestra. Second bassoon and bass trombone with the basses. Identical notes and rhythm. The first bassoon and tenor trombone -	Please! Just one moment.
5005	It couldn't be simpler.	First bassoon and tenor trombone - what?
5006	First bassoon and tenor trombone - what?	With the tenors.
5007	With the tenors.	Also identical?
5008	Also identical?	Exactly. The instruments to go with the voices. Trumpets and timpani, tonic and dominant.
5009	And that's all?	Oh no. Now for the Fire. Strings in unison - ostinato on all - like this.

프로세스 개요

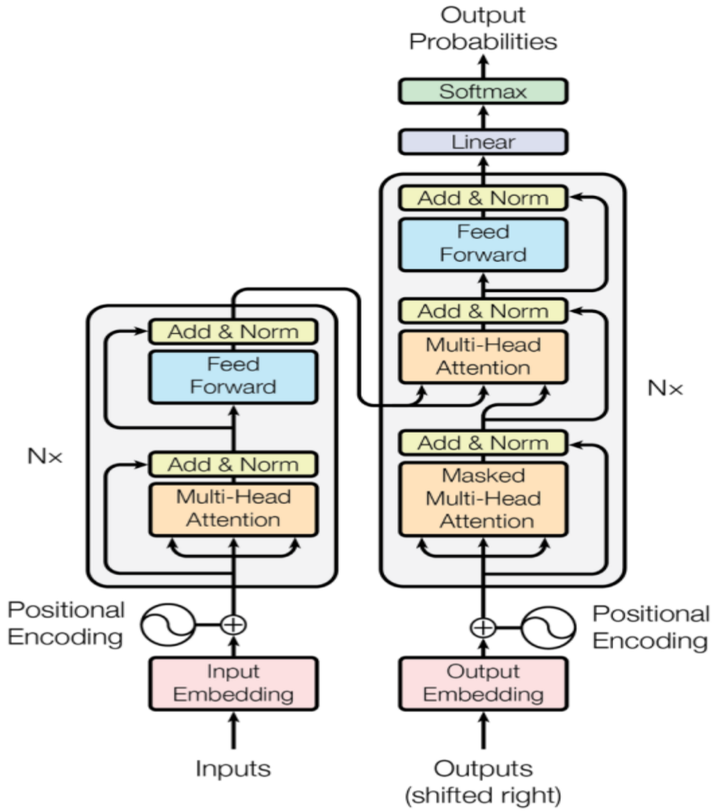

TALK
카카오톡 서버
POST 방식
JSON request
JSON response
aws
django
장고 서버
answer
Socket
question
BERT 알고리즘

03

Encoder & Decoder

Encoder : Transformer 기반의 KoBERT모델

Decoder : 단방향 GRU



* Encoder 의 hidden state 와 Decoder 의 hidden state 로 Attention 을 구하고
Attention 을 Decoder 의 hidden state 와 concat 한다

03

01. 프로젝트 개요

02. 데이터셋 구하기

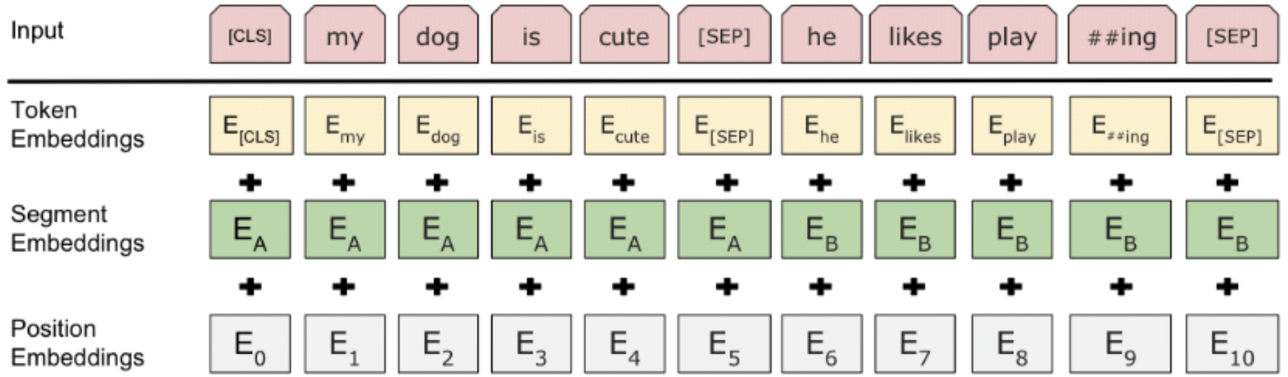
03. 알고리즘 구현

04. 서버 구축 및 카카오톡 플러스친구 연동

05. 챗봇 시연

Korean BERT pre-trained cased (KoBERT)

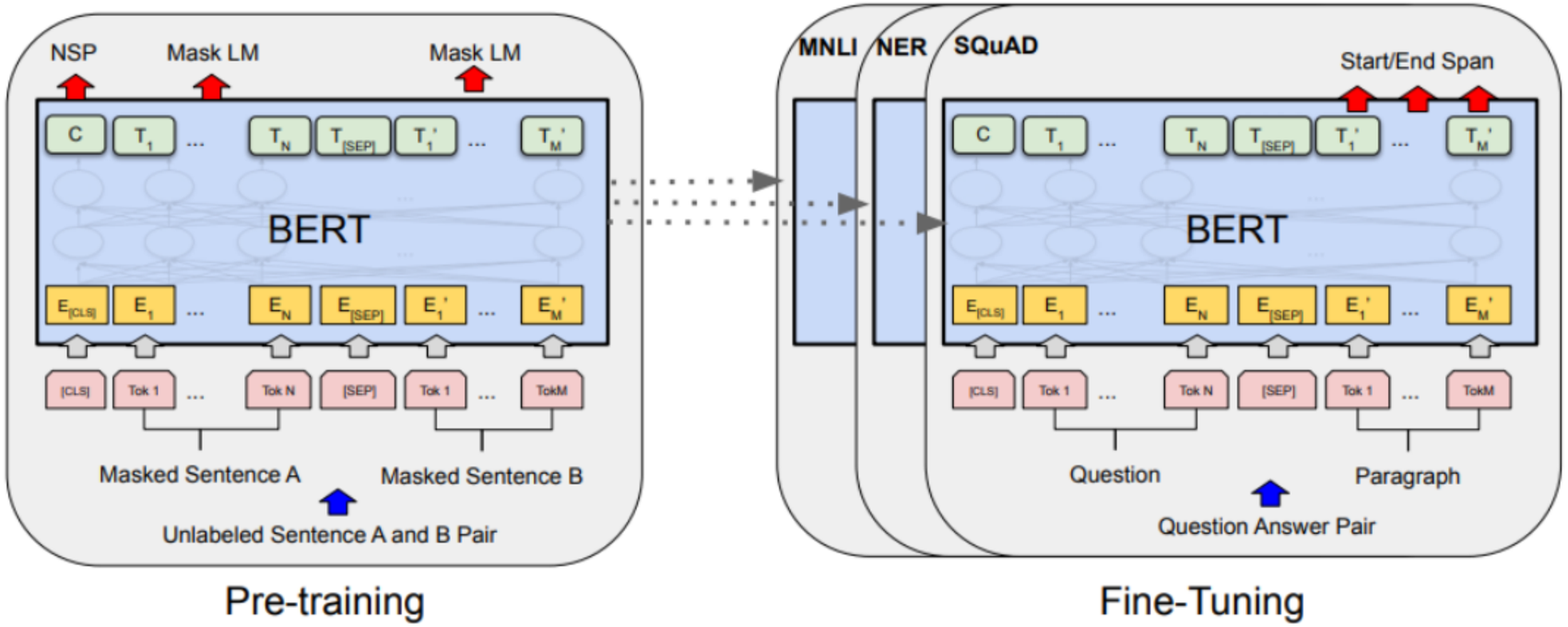
- SKT-Brain에서 개발한 KoBERT는 기존 BERT의 한국어 성능 한계를 극복하기 위해 개발
- 위키피디아, 뉴스 등에서 **수백만 개의 한국어 문장을 학습**
- 한국어의 불규칙한 언어 변화의 특성을 반영하기 위해 **데이터 기반 토큰화 기법**을 적용



문장 시작은 [cls], 끝과 문장 구분은 [sep], 패딩은 [pad], 그리고 형태소, 음절 단위로 짤라서 토크나이징

03

BERT: Pre-trainig of Deep Bidirectional Transformers for Language Understanding



BERT는 거대 Encoder가 입력 문장들을 임베딩 하여 언어를 모델링하는 **언어 모델링 구조 과정**과
이를 Fine-tuning하여 여러 자연어 처리 Task를 수행하는 **언어 모델 전이학습**으로 나뉜다.

03

01. 프로젝트 개요

02. 데이터셋 구하기

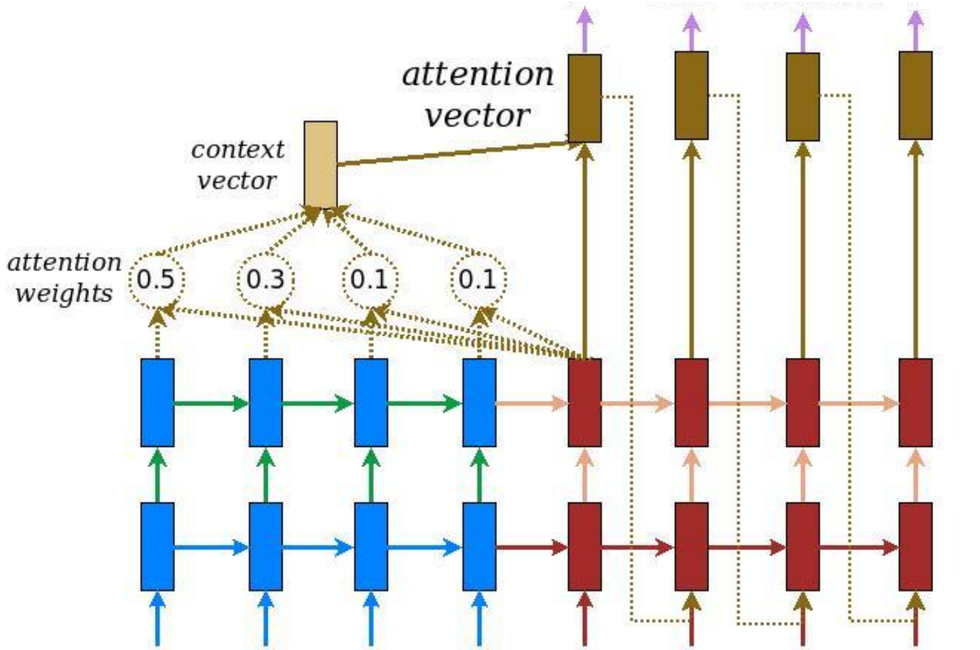
03. 알고리즘 구현

04. 서버 구축 및 카카오톡 플러스친구 연동

05. 챗봇 시연

Decoder (GRU)

[CLS] 김 승 유 입 니 당 [SEP]



[CLS] 와 이 빅 타 차 기 회 장 은 누 구 인 가 요 ? [SEP]

GRU는 **적은 파라미터로 학습**이 가능하기 때문에 적은 데이터셋에 대한 오버피팅을 방지할 수 있음.

03

01. 프로젝트 개요

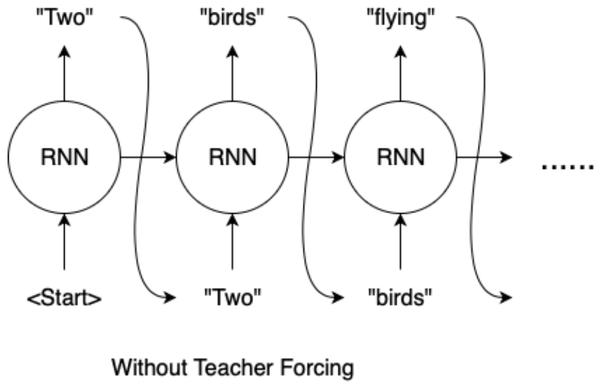
02. 데이터셋 구하기

03. 알고리즘 구현

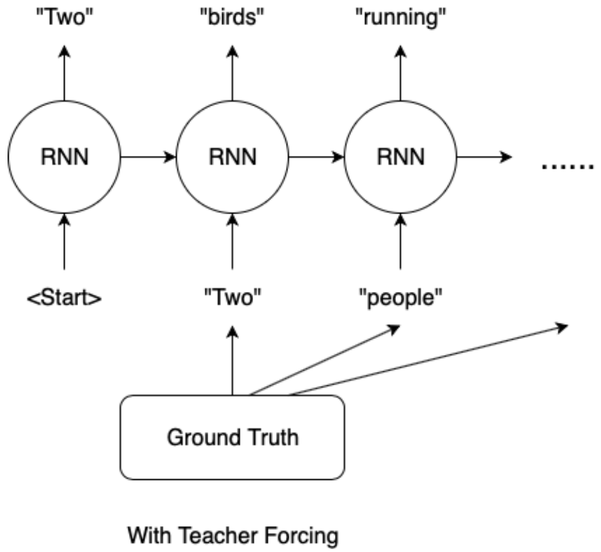
04. 서버 구축 및 카카오톡 플러스친구 연동

05. 챗봇 시연

Teacher Forcing



Without Teacher Forcing



With Teacher Forcing

Teacher Forcing 이란?

- 학습 중에 이전 Step에서 생성된 Token을 배제하고 Ground Truth를 넣어주는 방법

프로세스 개요


TALK
카카오톡 서버
POST 방식
JSON request
JSON response
aws
django
장고 서버
answer
Socket
question
BERT 알고리즘

04

AWS 서버 구축

왜 서버 구축이 필요한가?

기본적으로 챗봇은 사용자의 채팅을 받아 챗봇 서버로 요청을 보낸 후 응답을 반환하는 구조

챗봇이 정상적으로 24시간 작동하기 위해서는 항상 실행되고 있는 연동된 서버가 필요하고
따라서 AWS EC2를 이용한 "심심한 와빅이"용 서버를 구축하는 작업이 선행되어야 한다.

* 물론 챗봇을 위해 24시간 네트워크에 연결된 채 구동될 수 있는 컴퓨터가 있다면 이를 사용해도 되지만, 비효율적이다.

: 아마존닷컴의 세계 1위 클라우드 컴퓨팅 사업부

04

AWS 서버 구축

챗봇 서버용 AWS EC2를 개설한 후, 웹 서버 포트포워딩을 진행

Amazon EC2
Elastic Compute Cloud

: 독립된 컴퓨터에 가상서버를 구축해서 임대해주는 서비스

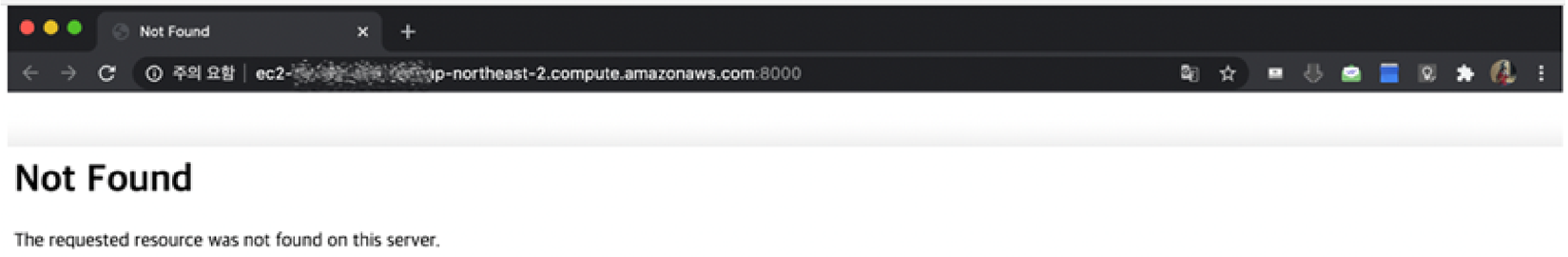
*웹 서버 포트포워딩 : 웹 서버와 카카오톡 간의 통신을 위해 지정한 포트를 열어주는 것 (ex. 80번 포트)

04

AWS 서버 구축

django : 파이썬으로 작성된 오픈소스 웹 애플리케이션 프레임워크

- django를 이용하여 AWS 서버에서 HTTP 통신을 담당할 웹 프레임워크를 구축
- AWS에서 클라우드 서버를 실행, Django 서버를 실행



AWS EC2에 Django가 배포되었다.

04

카카오 오픈빌더

: 인공지능 기술을 이용하여 카카오톡 채널 챗봇과 카카오 보이스봇을 동시에 설계할 수 있는 카카오 AI 설계 플랫폼.

04

01. 프로젝트 개요

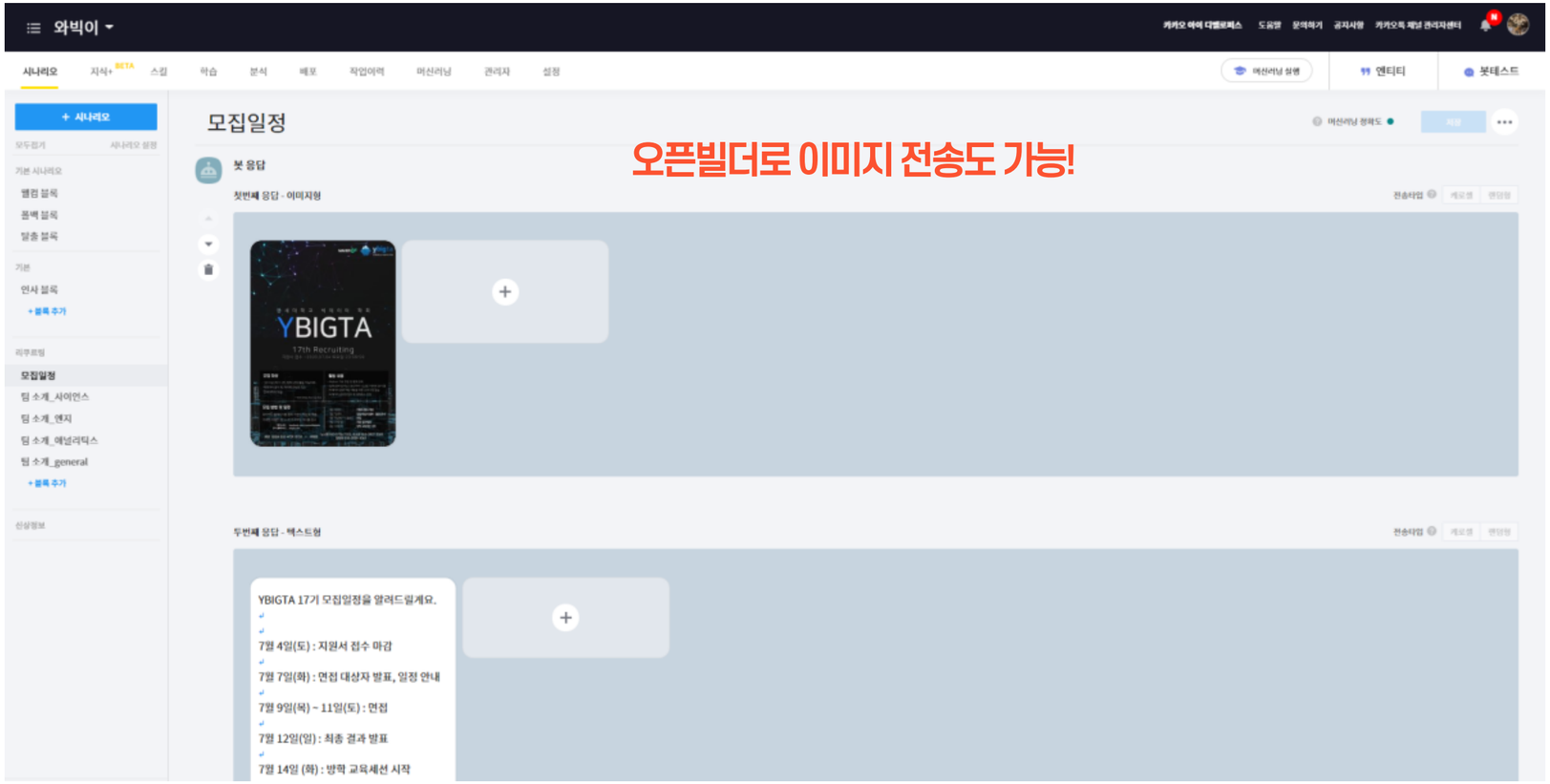
02. 데이터셋 구하기

03. 알고리즘 구현

04. 서버 구축 및 카카오톡 플러스친구 연동

05. 챗봇 시연

시나리오 작성



04

01. 프로젝트 개요

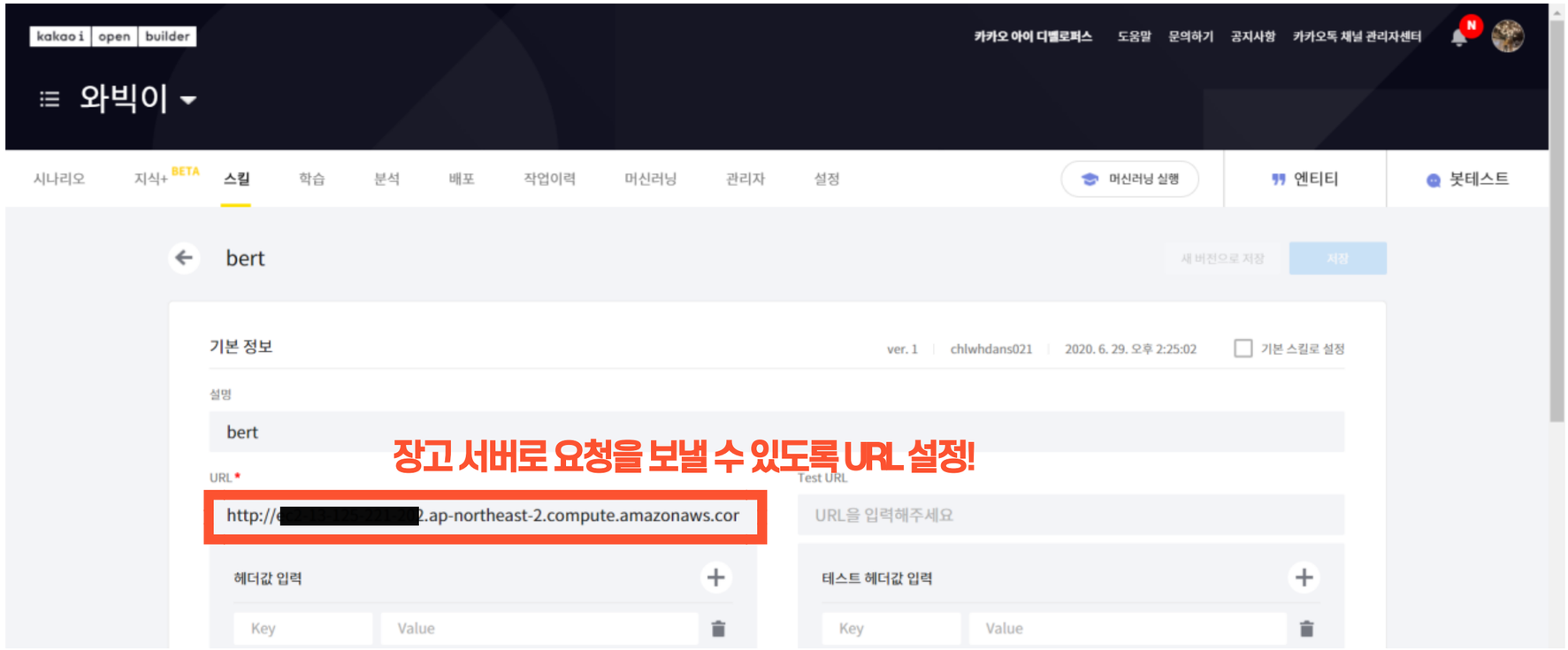
02. 데이터셋 구하기

03. 알고리즘 구현

04. 서버 구축 및 카카오톡 플러스친구 연동

05. 챗봇 시연

카카오톡 API 연동



- 블록에 URL을 연결하여 POST 방식의 요청을 전송
- 설정해둔 JSON 응답을 받아 형식에 맞게 응답을 지원

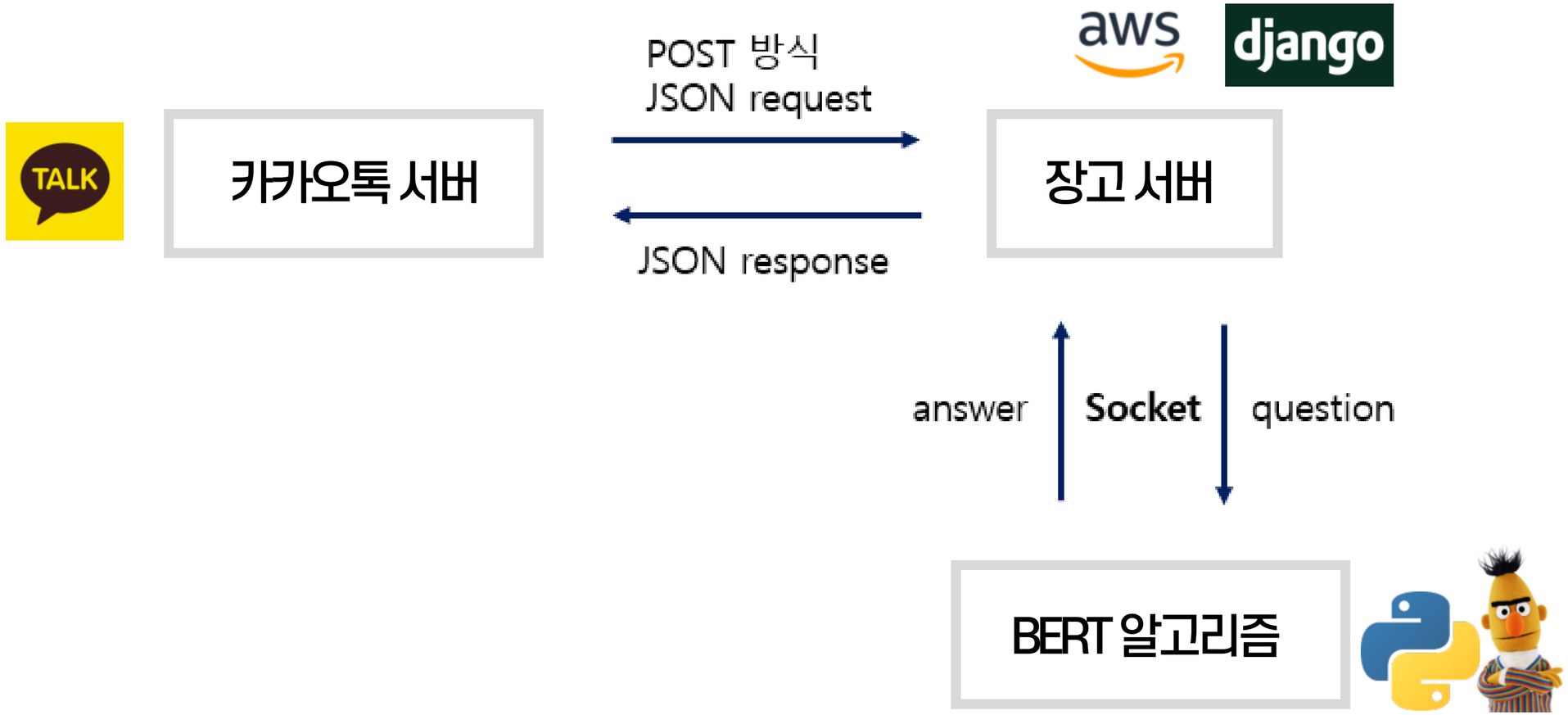
04

장고 프로세스와 알고리즘 연동

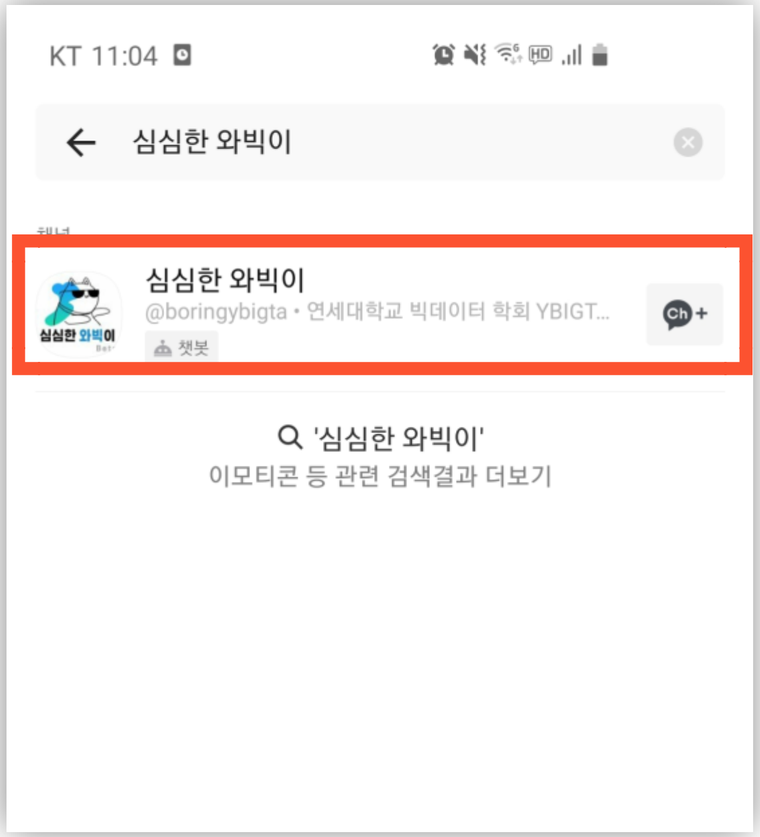
장고에 요청이 들어올 때마다 챗봇 모델을 실행시켜 응답을 얻고 프로세스를 종료하니까 모델을 불러오는데 딜레이가 존재

챗봇 모델을 백그라운드로 실행하여 소켓을 활용한 프로세스 통신을 이용하여 장고와 통신함으로써 속도 향상

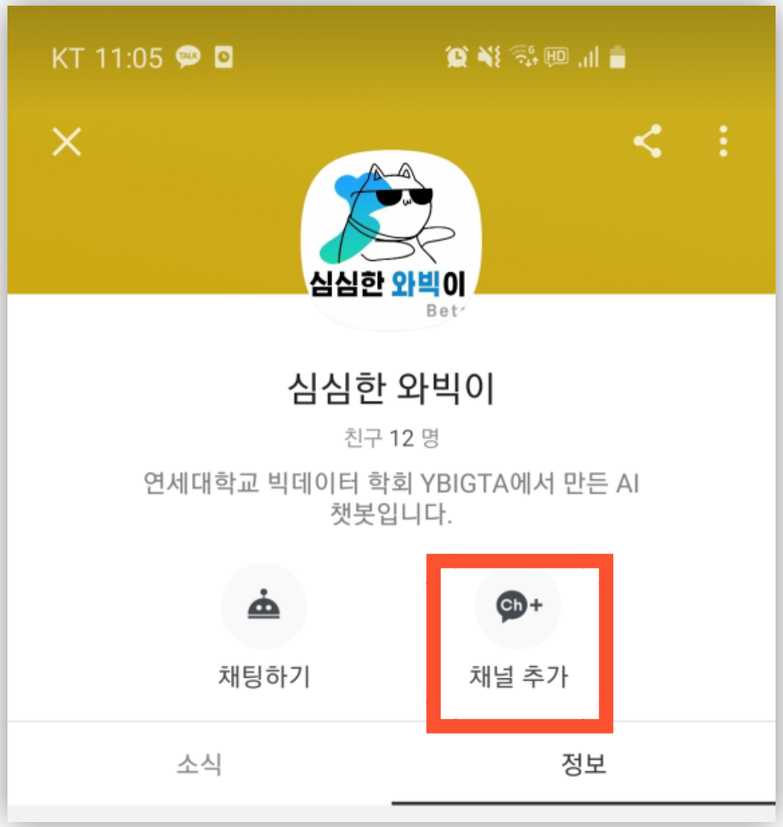
04 최종 프로세스 정리


aws
django
TALK
카카오톡 서버
POST 방식
JSON request
JSON response
장고 서버
answer
Socket
question
BERT 알고리즘

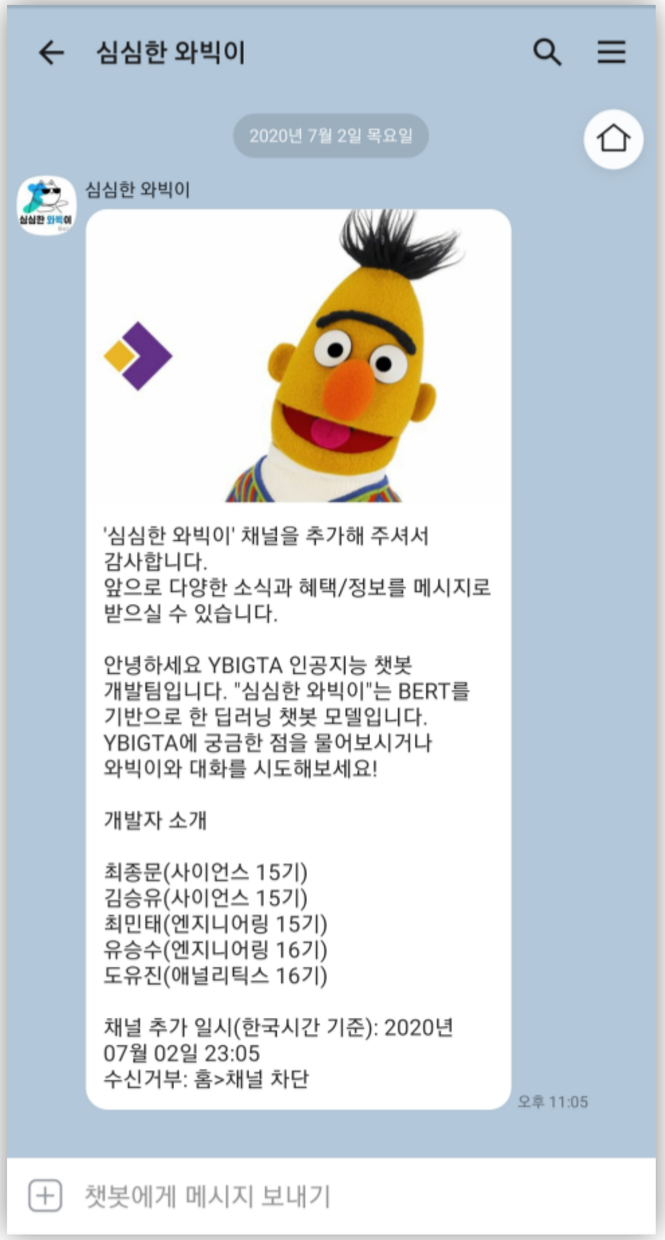
05. 챗봇 시연



1. 카카오톡에 들어가서 "심심한 와빅이"를 검색하세요.



2. 채널 추가를 눌러주세요.



3. 채팅에 들어가시면 와빅이의 웰컴 메세지를 확인할 수 있습니다.

4. 와빅이와 대화를 시도해 보세요!

감사합니다