

2022

딥러닝과 자연어 처리를 통한 민원 상담 AI 챗봇 구축 프로젝트

AI Chatbot with DeepLearning and Natural Language Processing To Assist Nation

S4ItLux

노현진, 김지현, 박현태, 안수빈,

이힘찬

2022. 02 . 10

AI 챗봇 구축 프로젝트

목차

Contents

01

프로젝트 개요

1 민원 챗봇의 필요성과 사례
2 민원 챗봇 기대효과

02

데이터 수집 및 설명

03

프로젝트 흐름도

04

모델 구조

01 트랜스포머
02 셀프 어텐션, 포지셔널 인코딩

05

사용 도구 및 환경

06

데이터 전처리

07

모델 파라미터

08

실행 결과 및 결론

01 모델의 정확도, 실행 결과
02 결론 및 추가 제안

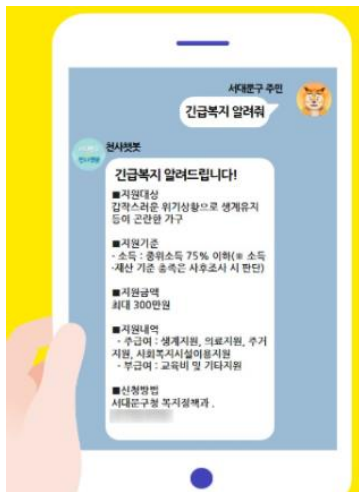
09

참고 문헌

01. 프로젝트 개요

민원 챗봇의 필요성과 사례

- 코로나19로 인하여 민원 서비스에서도 비대면 챗봇 기술 적극 활용하고 있으며 증가한 상담 수요에도 도움이 될 것으로 전망



▲서대문구 복지정보 천사챗봇 이미지

사례1) 제주도 민원 챗봇 제주톡

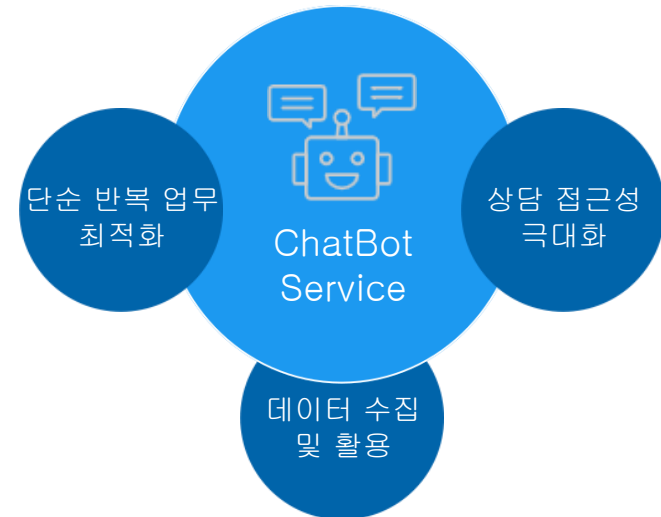
- 분실물, 관광지, 숙박시설, 버스 정류장 실시간 버스 도착 알림 등과 같은 정보 제공

사례2) 서대문구 복지정보 천사 챗봇

- 어르신, 장애인, 임신출산, 아동 등과 같은 복지 관련 정보 제공

민원 챗봇 기대효과

- 고객 응대와 같은 단순 업무 손쉽게 해결
- 사용자로부터 데이터 수집 및 추가 활용 가능
- 24시간 서비스를 제공하여 상담 접근성 극대화



02. 데이터 수집 및 설명

챗봇 학습 데이터 소개

- AI Hub의 민원(콜센터) 질의-응답 데이터 (<https://aihub.or.kr/aidata/30716/download>) 사용
- 챗봇의 예상 문답(Q-A) 페어 13,360개
- 문답 주제 관련된 label 변수는 학습에 반영하지 않음

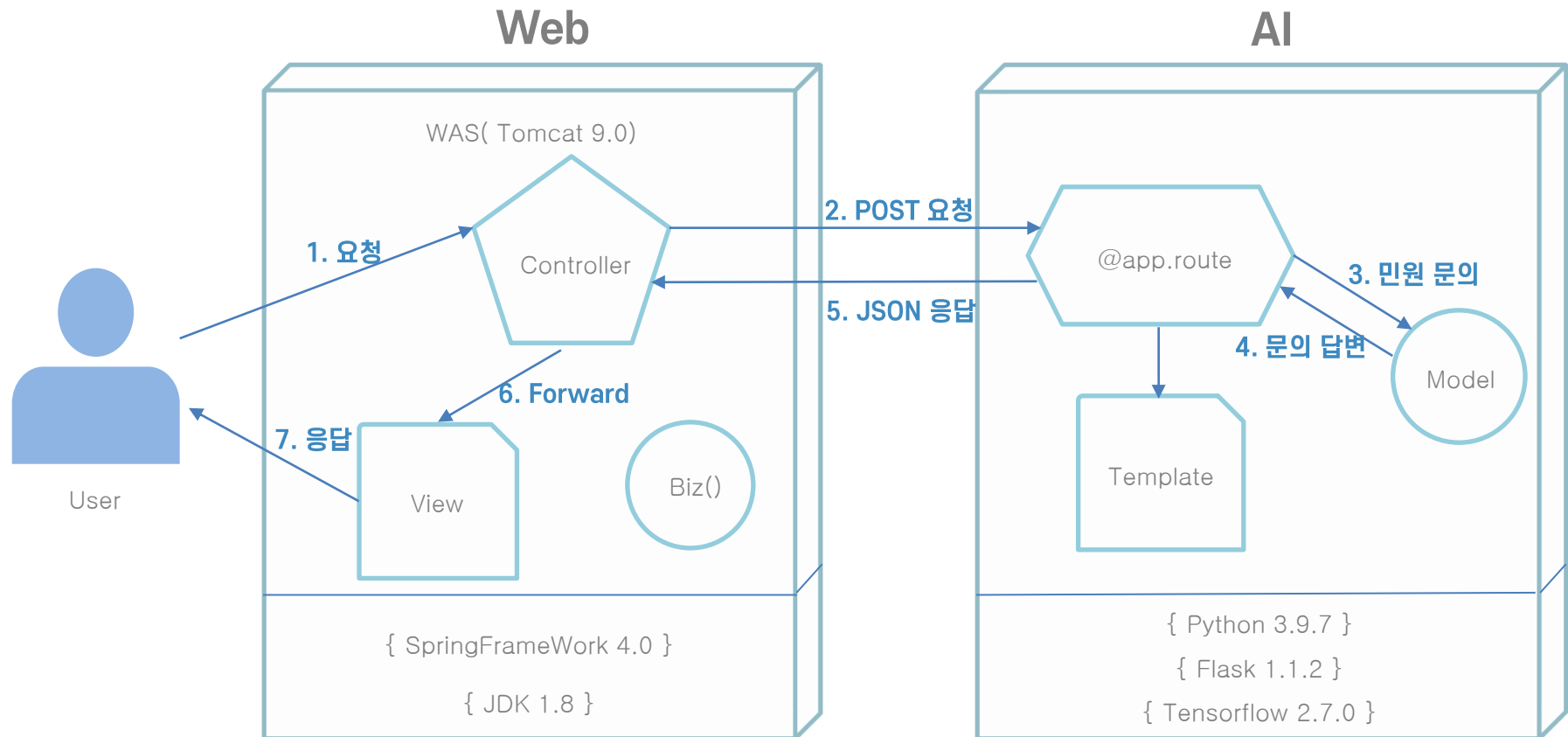
chatbotdata.csv

Q	A	label
운영 시간은 어떻게 됩니까?	오후 6시까지입니다.	0
페이지 수 제한은 없습니까?	A4 2~3장 내외입니다.	0
서울시에서 진행하는 평생교육 관련 문의드립니다.	네 신청하시는 겁니까?	0
정원이 몇명입니까?	정원은 20명 입니다.	0

shape = (13360, 3)

3. 프로젝트 흐름도

프로젝트 아키텍처

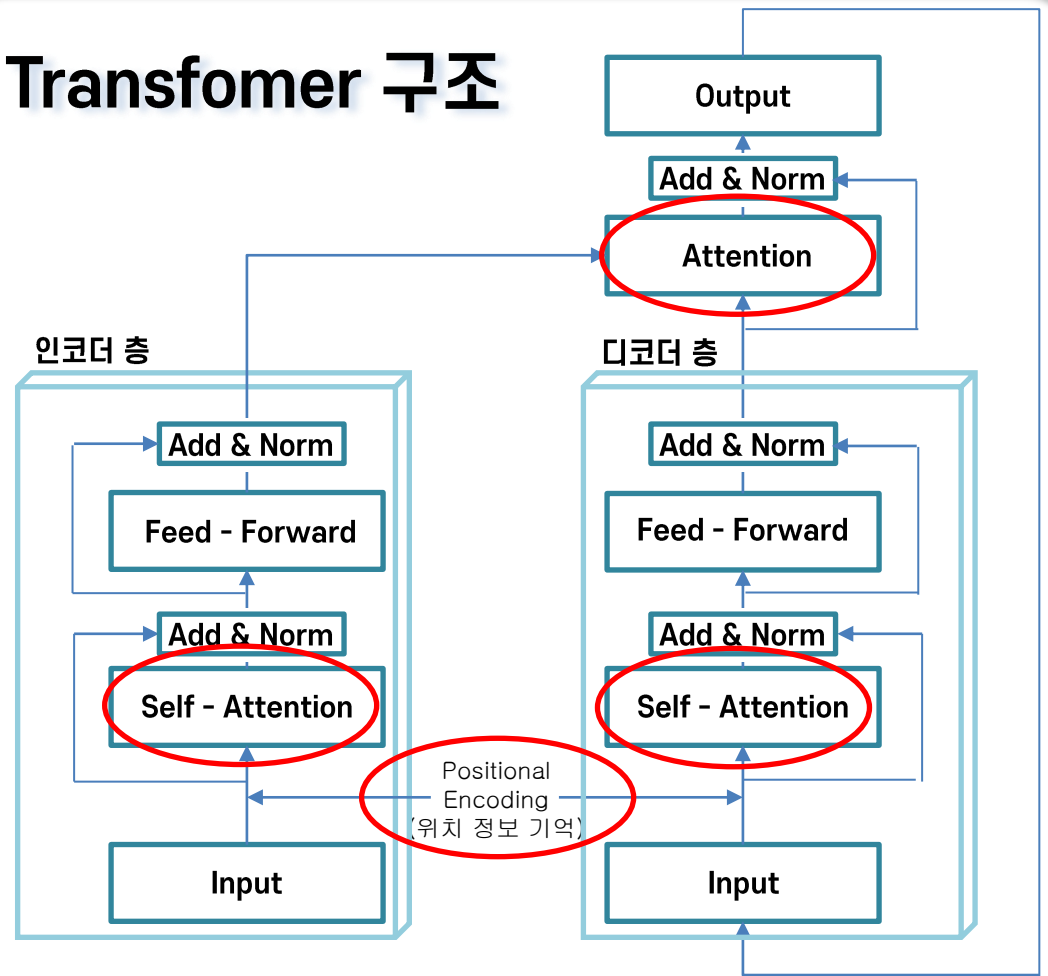


04 모델 구조

트랜스포머

- 챗봇과 기계번역에 주로 쓰이는 인코더-디코더 구조를 사용하여 질문 문장 입력하면 답변 문장으로 출력하도록 구현한 모델
- RNN 없이도 어텐션과 단어의 위치정보 벡터로 구조를 구현하여 RNN보다 효율적인 성능을 보임
- 트랜스포머는 단어를 순차적으로 입력받는 방식이 아니므로 각 단어의 위치 정보를 얻기 위해 **단어 임베딩 벡터 + 위치 정보**를 통해 모델에 입력
- 위치 정보를 활용하여 현재 처리 중인 단어에 대해 다른 연관 있는 단어들과의 맥락 파악
- 디코더의 출력을 예측하는 매 시점마다 인코더의 전체 입력 문장을 다시 한 번 참고하는데 입력 전부를 동일한 비율로 참고하는 것이 아닌 Attention을 통해 **예측 값과 연관**이 있는 부분을 집중적으로 참고

Transfomer 구조

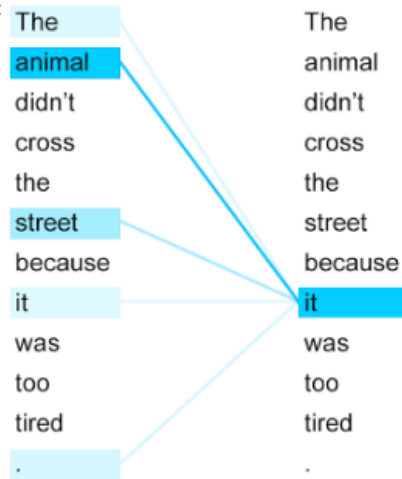


04 모델 구조

셀프 어텐션(Self-Attention)

- 주어진 단어와의 유사도를 파악하기 위해 사용
- 특히 인코더의 첫번째 어텐션층과 디코더의 첫번째 어텐션층은

셀프 어텐션을 사용하여 입력 문장 내 전체 단어간의 문맥을 파악하는 방식



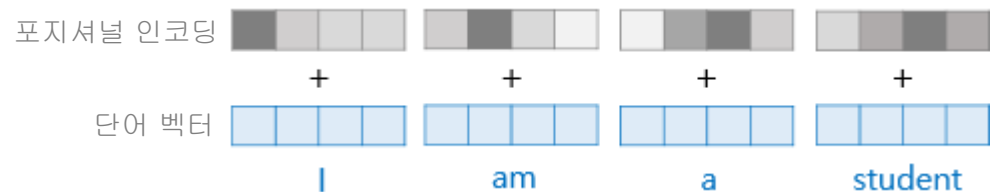
▲셀프 어텐션 사용 예

문장의 대명사(it)가 지시하는 것이 무엇인지 입력 문장 내 전체 단어를 참고하여 파악 가능

포지셔널 인코딩(Positional Encoding)

- 기존 인코더-디코더 구조는 단어를 순차적으로 입력받아 처리하는 RNN의 특성으로 단어의 위치정보를 기억
- 트랜스포머는 RNN 없이 구현한 모형이므로 단어의 위치정보를 기억하기 위해

Positional encoding 후 입력 벡터에 추가



▲포지셔널 인코딩 예

각 단어 벡터에 단어의 위치정보 인코딩 값을 더하여 위치를 반영

05 사용 도구 및 환경

OS 및 하드웨어



Intel(R) Core(TM)
i5-6500 CPU
@ 3.20GHz

RAM : 16.0GB

사용 언어



python™
3.9.7



1.8.0_301



1.8.0

도구



4.8.0



Visual Studio Code

1.64.1



1.58.0



2021.3

eGovFrame

3.9.0

프레임워크



spring

4.0



Flask

1.1.2

라이브러리



1.3.5



NumPy

1.19.5



matplotlib

3.2.2

TensorFlow

2.7.0



4.0.1



2.2.1

형상 관리

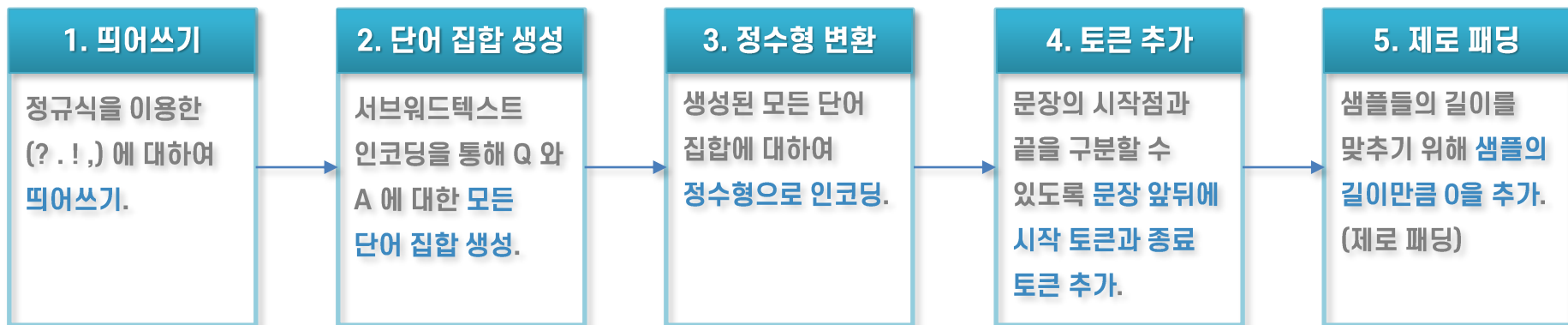


GitHub



06 데이터 전처리

예시를 통해 보는 전처리 과정



Ex)



07 모델의 파라미터

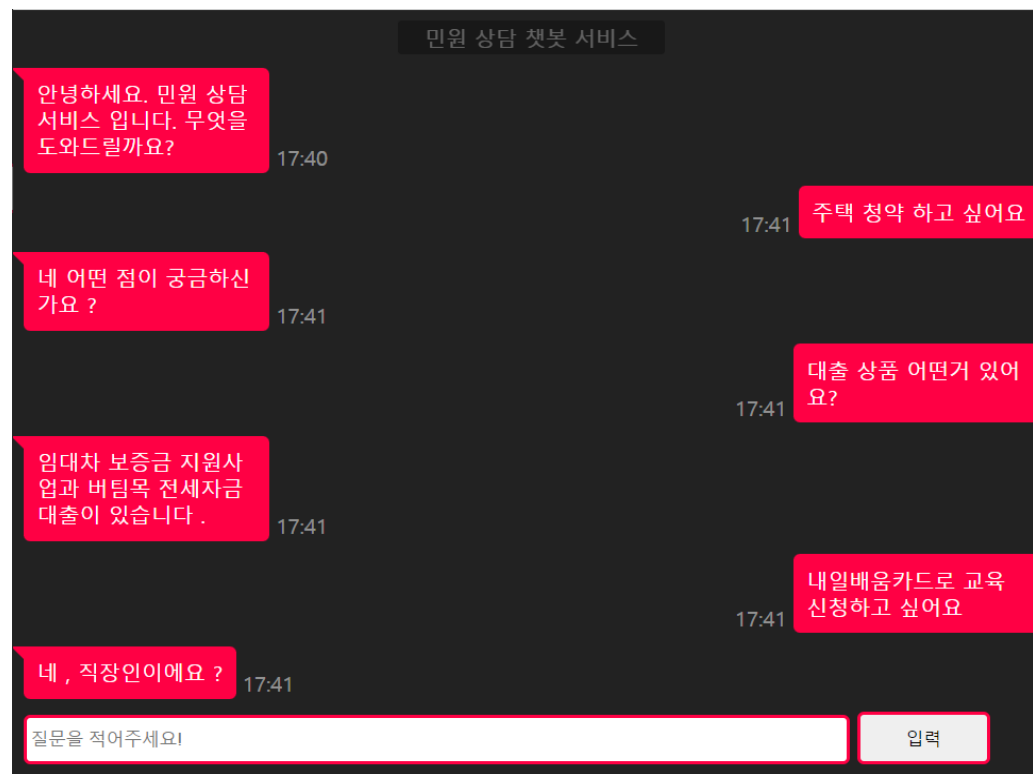
트랜스포머 모델 파라미터

파라미터	정의	지정 값
Vocab_size	단어 집합의 크기	7783
Num_layers	인코더, 디코더 층의 개수	2
Dff	피드포워드 은닉층 크기	512
D_model	인코더, 디코더의 입/출력 차원	256
Num_heads	멀티-헤드 어텐션에서 병렬적으로 사용할 헤드의 수	8
dropout	과적합 방지를 위한 일부 뉴런 탈락	0.1
Loss_function	손실 함수	Sparse Categorical Crossentropy
optimizer	활성화 함수	Adam

08 실행 결과 및 결론

모델의 정확도 : Accuracy - 0.2508

모델 실행 결과 :



08 실행 결과 및 결론

결론

민원(콜센터) 질의-응답 데이터를 활용하여 Transformer를 통해 AI 모델을 학습시킨 후 예측 결과를 웹 서버에 전달. 학습시킨 결과 Accuracy는 0.2508로 간단한 상담 업무의 경우 해결 가능. 상담사의 업무 부담 감소 및 효율성 향상 기대.

추가 제안

1. 하이퍼 파라미터 탐색을 통한 모델 최적화 과정이 필요.
2. 사용자가 초기 화면에서 민원 주제를 선택할 수 있는 주제 선택란 추가.
3. 추가적인 학습 데이터를 통한 모델 성능 향상 기대.

09 참고 문헌

참고 문헌

1. Attention Is All You Need

<https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>

2. 딥러닝을 이용한 자연어 처리 입문

<https://wikidocs.net/31379>