착수보고서

2023전기 졸업과제

응답하라 챗봇

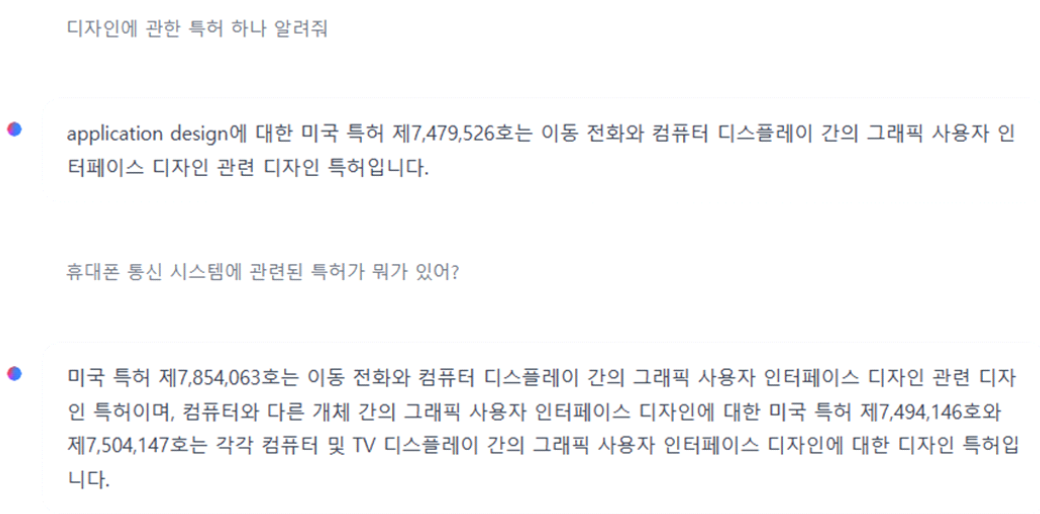
정보컴퓨터공학과 201624597 최상준
정보컴퓨터공학과 201824448 김정민
불어불문학과 201902163 김채원

목차

1. 과제 배경 및 목표

1-1. 과제 배경

2022년 ChatGPT가 큰 인기를 얻음에 따라 대화형 인공지능 서비스에 대한 관심이 높아지고 있다. 국내에서는 Stanford Alpaca 모델을 차용한, 한국어를 이해는 KoAlpaca 가 공개되어 주목을 받기도 했다.[1] 하지만 이를 실제로 사용해보았을 때 몇 가지 문제점이 있었는데, 문장 구성이 어색하거나, 문맥을 잘 파악하지 못한다는 점이다. 가장 취약한 부분은 전문성이 필요한 영역에서 대답의 질이 상당히 떨어진다는 점이다. 해당 서비스는 네이버 지식인을 크롤링하여 만들어졌다. 해당 웹사이트의 특성상, 의학, 법학, 어문학 등의 깊은 학문적 지식을 얻기는 어렵다. 다음의 사진과 같이, 특허 부문에 관한 질문을 하자 실제로 존재하지 않는 특허를 답변으로 제시하였다.



<사진 1> KoAlpaca의 문제점

이러한 배경으로, 특정 부문에 특화된 한국어를 사용하는 챗봇을 만들어보고자 하였다. 그 중 특화 도메인은 특허 부문으로 선정하였다. 한국어를 사용하는 특허 부문에 특화된 챗봇을 제작하여, 특허를 출원하고자 하는 사람, 특허 검색이 필요한 사람들이 보다 편리하게 원하는 정보를 얻을 수 있는 시스템을 구현하고자 한다.

1) https://github.com/Beomi/KoAlpaca

1-2. 과제 목표

본 과제는 자연어 처리를 통하여 챗봇을 학습시켜 한국어에 대한 이해도와 문장 구성력을 높이며 특허 부문을 집중적으로 학습시켜 특허와 관련한 정보를 얻고자하는 이용자에게 도움을 줄 수 있는 서비스를 개발하는 것에 목표로 한다.

a. 웹 배포
웹 플랫폼을 개발하여 이용자들이 언제든 해당 서비스에 접근할 수 있도록 한다.

b. 자연어 처리를 통한 한국어 학습
한국어 사용자들이 보다 쉽게 서비스에 접근할 수 있게 한다. 단순히 한국어로 문장을 구성할 수 있는 것을 목표로 하는 것이 아니라, 최대한 어색하지 않은 문장을 구성하는 것이 목표이다.

c. 특허 지식 학습
여러 특허들의 정보를 얻은 다음, 체계적으로 분류 및 요약하여 사용자들에게 간략하게 설명할 수 있도록 한다. 또한 사용자가 원할 시 더욱 구체적인 정보를 제공하거나, 비슷한 다른 정보를 제공하는 등의 행동을 할 수 있게 한다.

2. 요구 조건 분석

2-1 데이터 수집

a. 지식인

한국어 학습과 관련된 데이터의 질을 올리기 위하여 일반 글보다 조금은 선별되었다고 판단되는 '지식인 베스트' 문답을 크롤링 한다. 충분한 데이터 셋을 얻기 위하여 '지식인 전문가'의 답변 또한 선별적으로 크롤링 한다. 이때 욕설, 은어, 자음 등의 사용 불가한 언어는 제외한다.

b. 키프리스

키프리스 사이트에서 제공하는 엑셀 저장을 사용하여 데이터 셋을 얻는다. 웹사이트에서 크롤링 할 내용으로는, 특허의 요약 섹션이 있다.

F34

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	출원번호	발명의명칭	발명의명칭	상태	IPC분류	출원인	대리인	발명자	출원일자	등록번호	등록일자	공개번호	공개일자
2	102022703	현상제 보	DEVELOP	공개	G03G 15/0	캐논 가부	장수길|이	무라카미	2018.09.21			102022015	2022.11.22
3	102022014	내용물 용	CONTAINE	등록	A45D 40/2	한국콜마주	해움특허법	이진성|정	2022.11.10	102532511	2023.05.10		
4	102022005	사출연신불	TURN TAE	등록	B29C 49/3	주식회사 L	서평강|윤	유병전|송	2022.08.08	102497146	2023.02.02		
5	102021011	액체 분사	CONTAINE	공개	B05B 11/0	주식회사	특허법인은	이경창	2021.08.27			102023003	2023.03.07
6	102021011	재활용 가	Pumping T	공개	B05B 11/0	임동선	조홍규	임동선	2021.08.23			102023002	2023.03.03
7	102021004	고양이 배	Separating	공개	A01K 1/01	강갑주	이익배	강갑주	2021.04.13			102022014	2022.10.20
8	102020703	히터 어셈	HEATER A	등록	A24F 40/4	니쏜 다바	특허법인율	야마다 마	2019.04.23	102532401	2023.05.10	102020014	2020.12.22
9	102020005	미용기	Cosmetic I	공개	A61F 7/00	주식회사	문용회|오	민선웅|김	2020.05.14			102021002	2021.03.09
10	102019703	밀폐된 용	IMPROVE	공개	B01J 2/30	야라 인터	윤의섭|김	르두, 프랑	2018.06.22			102020002	2020.03.04
11	102019015	샴푸 용기	PUMP DIS	등록	B05B 11/0	강민구	박재환	강민구	2019.11.26	102122160	2020.06.05		
12	102019008	펌프 용기	Pump vess	등록	B05B 11/0	강민구	우덕근	강민구	2019.07.16	102120010	2020.06.01		
13	102019008	펌프 용기	Pump vess	등록	B05B 11/0	강민구	우덕근	강민구	2019.07.11	102120009	2020.06.01		
14	102018010	가스공급용	NOZZLE P	등록	H01L 21/6	주식회사	특허법인주	김홍열	2018.09.04	102113276	2020.05.14	102020002	2020.03.12
15	102018003	젓가락이	Food conta	등록	B65D 81/3	방윤정	신진만	방윤정	2018.03.27	102012590	2019.08.13		
16	102017004	화장도구오	Make up c	소멸	A45D 33/2	(주)아모레	김희소	홍성수	2017.04.14	101886144	2018.08.01		
17	102022014	내용물 용	CONTAINE	등록	A45D 40/2	한국콜마주	해움특허법	이진성|정	2022.11.10	102532512	2023.05.10		

<사진 2> 키프리스 엑셀

<사진 3> 키프리스 웹사이트의 요약 섹션

2-2 데이터 전처리

자연어 처리를 위하여 사용 가능한 데이터를 선별해내는 작업이다. 어떤 언어를 필터링 해내고 어떤 언어를 사용할 것인지 선별하는 것이 중요하다.

2-3 자연어 처리

컴퓨터를 이용하여 사람이 사용하는 자연어를 분석하고 처리하는 작업이다. 형태소 분석, 통사 분석 등을 사용할 수 있다.

2-4 서비스 배포

웹 플랫폼을 통해 제작한 챗봇을 배포할 수 있도록 한다.

3. 진행 방안

3-1. 개발 환경

개발 언어: Python (크롤링 및 자연어 처리)
개발 도구: VSCode, Jupyter Notebook
실행 환경: PC, 모바일 등의 web 접속이 가능한 플랫폼

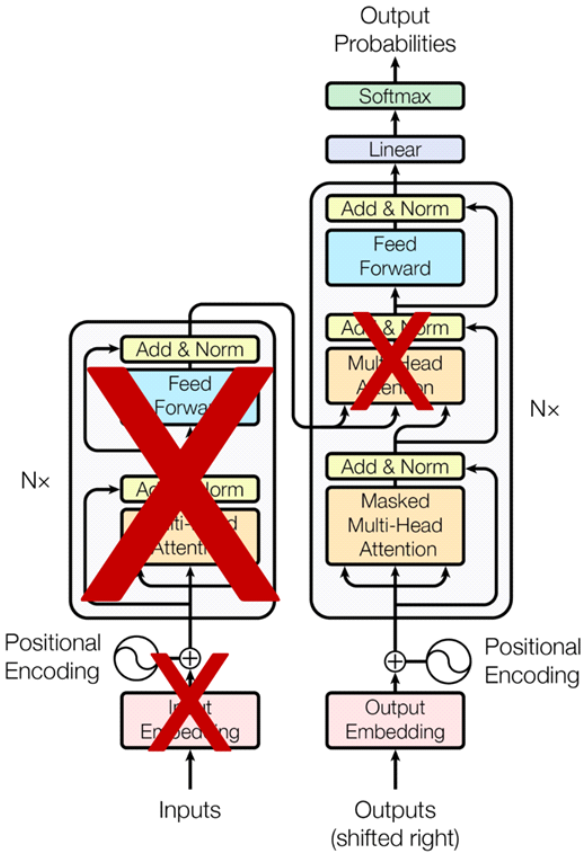
3-2. 사용 기술

a. 웹 크롤링

Python의 beautifulSoup 라이브러리를 이용한다.[2] 이는 HTML의 태그를 파싱해서 필요한 데이터만 추출하는 함수를 제공하는 오픈소스 라이브러리이다.

b. gpt 모델 학습

gpt 모델로는 BERT와 GPT가 존재하는데 해당 과제에서는 GPT를 사용한다. 이는 트랜스포머의 디코더 구조와 유사한 구조를 사용했다. 앞의 단어들을 활용하여 해당 단어를 예측하는 전통적인 언어 모델 방식을 사용한다는 것이 특징이다.



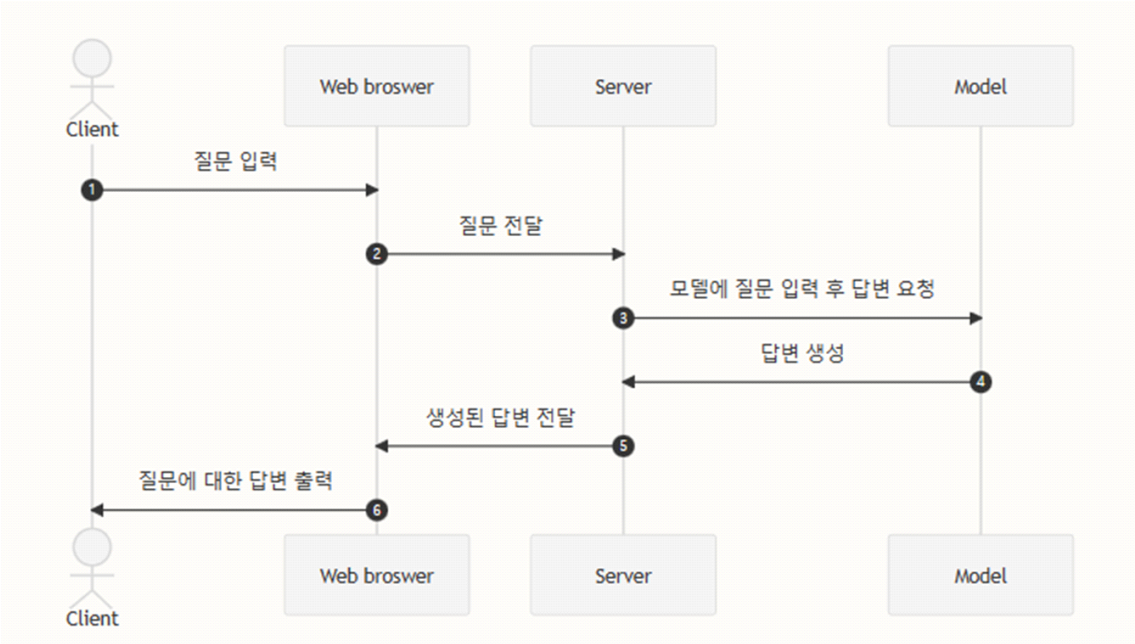
<사진 4> 트랜스포머 모델에서 인코더를 제외한 디코더만 차용한 구조

이를 위해서는 PyTorch[3] 오픈소스 소프트웨어 라이브러리를 사용한다. 비슷한 역할을 하는 TensorFlow에 비하여 직관적이고 최적의 성능을 끌어낼 수 있다는 장점으로 PyTorch를 채택하였다. 이때 모든 신경망 모델은 torch.nn 패키지를 통하여 생성한다.

---

2) https://www.crummy.com/software/BeautifulSoup/bs4/doc/
3) https://tutorials.pytorch.kr/beginner/basics/intro.html

3-3. 플로우차트



4. 문제점 및 해결방안

4-1. 개발시 예상되는 어려움 및 해결방안

a. 크롤링 방지 시스템

네이버같은 대형 포털사이트의 경우 일정 시간 내에 한 컴퓨터에서 많은 접속이 일어날 경우 해당 크롤링을 차단하기도 한다. 이는 크롤링을 방해하는 대표적인 요소인데, 이는 헤더 설정으로 크롤링 탐지를 우회할 수 있다. 헤더에서 user-agent 설정을 변경해주는 것으로 해결할 수 있다.

b. 사용할 단어의 필터링 기준 선정

지식인의 경우, 답변의 질의 떨어지거나, 비속어가 포함되어있거나 문법 파괴가 포함된 문장 등이 있을 수 있다. 이를 어떤 기준으로 얼마나 선별해낼 것인지가 중요하다고 판단된다. 현재는

- 비속어, 은어, 비표준어
- 부적절한 답변 (광고 및 오답)

등을 걸러내는 것을 목표로 하고 있다.

4-2. 현실적 제약 사항 및 해결 방안

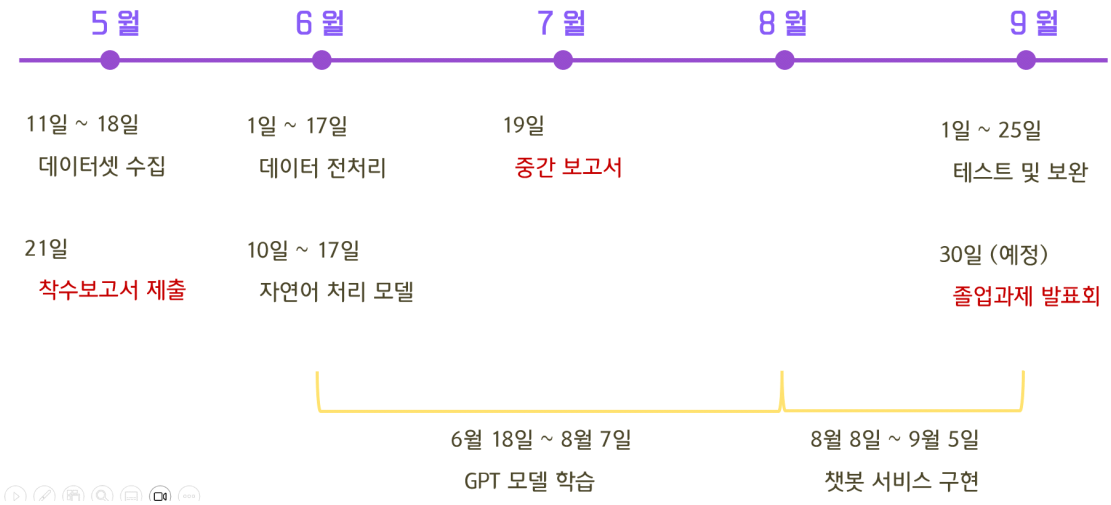
a. 답변의 최신화가 어려움

특허 분야는 정보의 최신화가 중요하다. 기업의 특허 업무로는 특허, 상표 출원 등도 있지만 경쟁사의 특허를 모니터링하여 현 상황을 파악하는 것 또한 기본적 업무다. 본인의 사업에 문제를 줄 수 있는 특허를 조기에 발견하는 것이 중요한 이유는, 발견한 속도가 빠를 수록 해당 특허에 대처할 수 있는 해결법이 다양해지기 때문이다. 또한 새로 출허된 특허로 새로운 아이디어를 만들어내거나 시장의 동향을 파악하기 용이하다.

이런 점으로 보아 특허 부문에서는 특히 최신화가 중요하다는 것을 알 수 있는데, 현재 수행 중인 과제는 데이터 수집이 끝나고 나면 데이터가 더이상 업데이트되지 않는다. 그러므로 개발이 끝난 이후에 상용화를 하게 된다면, 자동적으로 정보를 수집하여 최신화 할 수 있는 기능을 구축하는 것이 도움이 될 것으로 보인다.

5. 개발 일정 및 역할 분담

5-1 개발 일정



5-2 역할 분담

최상준	키프리스 데이터 수집 데이터 전처리 web 챗봇 서비스 구현
김정민	지식인 데이터 수집 자연어 처리 모델 생성 프롬프트 학습
김채원	지식인 데이터 수집 데이터 전처리 web 챗봇 서비스 구현