인공지능개론 기말과제

인공지능개론

어영정 교수님

2024년 12월 5일

연세대학교 글로벌인재대학을 위한 AI 챗봇 프로젝트

응용정보공학전공 2022106103 신규림

응용정보공학전공 2022106092 신영군

응용정보공학전공 2020106118 정예준

목차

1. 서론

1.1 챗봇 개발의 목적

하나의 웹사이트에서 정보를 찾는 것은 결코 쉬운 일이 아닙니다. 특히 글로벌인재대학에 입학하고 싶어하는 학생이나 신입생들에게는 더욱 힘든 일이다. 그러므로 저희는 글로벌인재대학에 관한 정보를 손쉽게 얻고 싶어하는 유저들를 위해서 AI 챗봇을 개발하기로 결정했다.

글로벌인재대학을 위한 AI 챗봇은 유저들에게 쉽게 정보를 제공해줌으로써 정보를 찾지 못하는 문제를 해결하고자 합니다. 또한 개인이 원하는 검색어를 직접 입력해서 정확한 정보를 제공해줍니다.

1.2 프로젝트 개발 절차

학생들에게 학사정보를 제공하기 위해 적절한 데이터를 수집하며 Language Model을 통해서 정확한 정보를 배출해야 합니다. 이 프로젝트를 구현하기 위해서는, 올바른 데이터 수집 및 이해, 데이터 전처리, 언어 모델 선정, 데이터 학습, 결과 배출의 과정으로 개발을 했다.

2. 데이터 처리

2.1 데이터 수집

본 프로젝트에서 사용한 데이터는 연세대학교 웹사이트에 있는 글로벌인재대학 학사규정과 연세대학교 교정집이다. 이 데이터는 연세대학교 규정집 웹사이트에서 웹 크롤링을 통해서 데이터를 수집했다: 우선 webdriver(chromedriver)를 사용하여 대상 컴퓨터에서 연결이 거부되지 않도록 설정합니다. 그리고 웹사이트 도메인에 따라 webdriver.Chrome()을 사용하여 각 HTML을 추출한다. (GLC의 경우: 링크는 https://ibook.yonsei.ac.kr/Viewer/glcrule_{i} for i in range(1,5). Yonsei의 경우: 링크는 https://rules.yonsei.ac.kr/search이다. 사전에 출력된 rule_pdf의 제목을 활용하여 search_box.send_keys(pdf_name)를 통해 원하는 규칙을 검색하고, driver.find_elements()를 사용하여 목표 규칙이 포함된 〈a〉 tag를 선택한다.) 그 후 정규 표현식을 사용하여 필요한 텍스트를 추출하고, 각 텍스트를 .txt 파일에 저장한다. 이 데이터들은 학생들에게 정확한 학사규정을 제공해줄 수 있다고 판단이 되며, 데이터 전처리를 통해서 언어 모델에게 학습할 예정이다.

2.2 데이터 전처리

본 프로젝트는 LlamaIndex 라이브러리를 활용하여 텍스트 데이터를 수집하고 처리했다. 데이터는 로컬 디렉토리에서 SimpleDirectoryReader 함수수를 사용하여 로드되었으며, 로드된 데이터는 총 개수를 출력하여 데이터가 제대로 수집되었는지 확인했다.

수집된 데이터는 품질을 보장하기 위해 전처리 과정을 거쳤습니다. 구체적으로, 아래와 같은 기준에 따라 데이터를 정제하였다:

- 텍스트에 “시행한다”, “삭제”, “부칙” 라는 문자열이 포함된 문서 제거.

전처리 후 데이터의 개수를 다시 출력하여 필터링 과정의 결과를 확인했다. 이렇게 정제된 데이터는 VectorStoreIndex.from_documents() 함수를 사용하여 먼저 RAG용 문장이 여러 작은 문자 블록으로 분리시키면서 벡터 데이터베이스(Vector DB)에 저장되었습니다. 이 단계는 검색 효율성을 높이기 위한 핵심 작업으로, 데이터베이스에 저장된 문서는 향후 검색 및 응답 생성 과정에서 사용된다.

3. 학습을 위해 사용된 모델

3.1 Hugging Face

Hugging Face는 LLAMA 모델과 함께 사용된 플랫폼으로, 고품질의 fine tuning된 언어 모델과 토크나이저를 제공합니다. 본 프로젝트에서는 “TheBloke/Mistral-7B-Instruct-v0.2-GPTQ” 모델을 사용하였으며, 추가적으로 fine tuning된 모델 "shawhin/shawgpt-ft"를 Peft를 통해 로드하여 성능을 향상시켰습니다. 이 모델은 사용자 질문에 대한 컨텍스트 기반의 정확한 응답을 생성하는 데 활용되었습니다.

3.2 RAG

RAG(Retrieval-Augmented Generation)는 검색 증강 생성 방식을 통해 대형 언어 모델의 성능을 향상시킨다. RAG는 검색된 문서에서 얻은 정보를 기반으로 모델의 응답을 생성합니다. 본 프로젝트에서는 다음과 같은 RAG 구성 요소를 사용하였다:

- VectorIndexRetriever: 입력된 질의(query)에 따라 벡터 데이터베이스에서 유사도가 높은 상위 3개의 문서를 검색합니다.
- RetrieverQueryEngine: 검색된 문서를 기반으로 질의 응답 엔진을 구성합니다.
- SimilarityPostprocessor: 유사도 기준이 0.5 미만인 문서를 필터링하여 응답 품질을 보장합니다.

3.3 Llama

LLAMA는 대형 언어 모델로, 본 프로젝트에서 데이터 임베딩과 자연어 응답 생성에 사용되었습니다. Hugging Face에서 제공하는 "BAAI/bge-small-en-v1.5" 모델이 임베딩 생성에 활용되었습니다. 이 모델은 주어진 텍스트 데이터를 벡터 형식으로 변환하여 문서의 의미적 표현을 생성합니다.

4. 모델 구현

4.1 데이터 Retrieval

검색 엔진을 설정하는 단계에서는 사용자가 입력한 질의에 적합한 문서를 검색할 수 있도록 시스템을 구성합니다. 먼저, 검색할 문서 블록록의 최대 개수를 top_k 변수로 설정하여 3개로 제한하였다. 그런 다음, VectorIndexRetriever 객체를 생성하여 벡터 데이터베이스에서 가장 유사한 문서를 검색할 수 있도록 구성하였다. 이후, RetrieverQueryEngine 객체를 통해 검색기와 후처리기를 결합하여 효율적인 검색 결과를 생성할 수 있도록 하였다. 마지막으로, SimilarityPostprocessor를 활용하여 유사도 기준(0.5) 이하의 문서를 필터링하였으며, 이를 통해 검색 결과의 정확도를 높였다.

검색 엔진을 활용하여 사용자가 입력한 질의에 대해 관련 문서를 검색하고, 검색된 문서의 내용을 바탕으로 응답을 생성하였다. 먼저, query 변수에 사용자가 입력한 질의를 설정하였으며, 예를 들어 query ="복수전공으로 인증 안된 전공은 뭐가 있어?" 로 설정하였다. 이후, query_engine.query(query)를 호출하여 검색 결과를 반환받았다. 검색된 문서의 내용을 바탕으로 응답을 생성하기 위해, 각 문서를 반복적으로 읽어 컨텍스트를 생성하였다. 이렇게 생성된 컨텍스트는 모델의 응답 생성에 활용되었습니다.

4.2 언어 모델

이 단계에서는 Hugging Face에서 제공하는 fine tuning된 언어 모델을 로드하였다. 먼저, AutoModelForCausalLM 클래스를 사용하여 "Mistral-7B-Instruct-v0.2-GPTQ" 모델을 불러왔다. 이 모델은 사용자 질의에 응답을 생성하는데 활용된다. 다음으로, PeftModel과 PeftConfig를 사용하여 "shawhin/shawgpt-ft" 모델의 fine tuning된 설정을 적용하였다. 마지막으로, AutoTokenizer 클래스를 사용하여 텍스트 데이터를 처리하기 위한 토크나이저를 로드하였다. 한국어 모델은 Bllossom/llama-3.2-Korean-Bllossom-3B라는 모델을 사용하였다.

4.3 결과 출력

사용자 질문에 응답을 생성하기 위해 로드된 언어 모델과 토크나이저를 활용하는 단계이다. 이 과정은 단순 프롬프트 응답 생성과 컨텍스트 기반 응답 생성의 두 가지 방식으로 진행된다.

먼저, 단순 프롬프트 응답의 경우 사용자의 질문만을 입력으로 받아 프롬프트를 생성합니다. 이를 위해 prompt_template 함수를 사용하여 사용자의 질문을 포맷팅합니다. 예를 들어, 사용자가 "what major cannot being the target of the double major?"라는 질문을 입력하면, 이를 기반으로 프롬프트가 생성된다. 생성된 프롬프트는 모델의 입력값으로 사용된다.

컨텍스트 기반 응답의 경우, 검색된 문서에서 얻은 정보를 추가적으로 포함하여 프롬프트를 생성합니다. 이 과정에서 prompt_template_w_context 함수를 사용하여 검색된 문서의 내용을 프롬프트에 포함시킨다. 이렇게 생성된 프롬프트는 사용자의 질문과 검색된 문서의 정보를 결합하여 보다 정교하고 구체적인 응답을 생성하는 데 활용된다.

응답 생성은 언어 모델의 generate method를 사용하여 수행된다. 먼저, 생성된 프롬프트는 Hugging Face의 토크나이저를 통해 토큰으로 변환되어 모델에 입력된다. 이후, 모델은 입력된 토큰을 바탕으로 응답 텍스트를 생성합니다. 마지막으로, 생성된 응답은 batch_decode method를 사용하여 텍스트로 디코딩된다. 이렇게 디코딩된 최종 응답은

사용자의 질문에 대해 적합한 답변으로 반환된다. 이 과정은 단순 질문부터 복잡한 컨텍스트 기반 질의까지 다양한 유형의 질문에 대응할 수 있도록 설계되었습니다.

5. 문제와 해결방안

문제1:프로젝트 개발 초기계단에서 예상된 방안은 다음과 같습니다: json형식으로 된 전통적인 물음과대답 데이터를 수집하고 open-source LLM한테 학습시킨다. 데이터 수집 후 문제점을 점점 나왔다: 데이터 양이 적을때 이런 훈련 방식은 편하지만, 학사 규정집의 모든 데이터를 얻은 후 PDF문서가 30개정도 나왔다. 큰 숫자이고 많은 양의 데이터를 모두 적합한 json형식으로 변환하면 시간상에서 너무 부담될 수 있는것 판단되었다. 더하고 전통적인 json형식인 데이터로 훈련시킨 경우 우리가 강조하고 있는 출처를 출력하려면, json데이터 내의 내용중에 출처를 하나하나 출력해야 한다.

해결책1: 여러문제를 예상하여, 종합적으로 고려하면서, 최종 우리가 RAG라는 방법을 사용하였다. 결과로보면, 이상 언급한 문제점이 해결 되었다고 할 수 있다.

문제2: 데이터가 블록으로 분리한 후, 특정한 단어를 근거로서 그 블록을 제거한다. 하지만 블록 중에 만약 필수한 내용을 같이 되있으면 같이 삭제된다.

해결책2: 코드내의 Settings.chunk_overlap 이용한다. 처음의 수치는 25였고, 직접50으로 증가 시킨 후, 내용을 잘 나온다. 원리가 이 parameter는 데이터 블록간 중복한 내용의 양을 설정하는 것이다. 예를 들면: 원본 데이터는 1234567이고, Settings.chunk_overlap = 25의 경우는 123,345,567로 나누는 것이다. 동시에 50으로 설정한다면 123,234,345,456,567로 나눈다. 유용한 내용이 4, 제거해야 하는 것은 3으로 가정하면, parameter가 25일때 4를 얻을 수 없는 데 50으로 설정하면 4를 얻을 수 있다.

문제3: 모델의 출력이 불안정한 상태이다, 어떤 시간에는 출처를 출력하고, 어떤시간은 출력하지 않다. 영어 모델은 글로벌인재대학이랑 언더우드국제대학의 영어이름을 틀리게 생성한다. 소용없는 내용을 가끔씩 생성한다.

해결책3: 모델의 prompt를 바꿨다.

6. 최종결과

6.1 결과

다음 결과는 한국어 언어 모델을 사용해서 구현했다. 질문으로는 “복수전공으로 인정 안된 전공은 뭐가 있어?”를 입력했다. 결과는 첨부된 결과에서 확인할 수 있듯이 정확한 답변을 주었으며, 잘 작동되고 있는 것을 확인할 수 있다. 그러나 하나의 문제점이 있다. 이 대답에 대한 출처를 제40조라고 하였지만, 제40조는 답변과는 상관이 없습니다. 즉, 저희가 시간상 문제로 data labeling을 하지 못했기 때문에 정확한 출처를 제공하지 못하고 있다.

this is the korean LLM with RAG ¶

```
# prompt (with context)
prompt_template_w_context = lambda context, comment: f"""[INST]너의 이름은 GLCPT다, 지금은 연세대학교 글로벌인재대학에 있는 학사지도교수이다. 학생들
해주고, 정보는 완벽하게 맞아야 된다. 모르는 정보는 그냥 모른다고 하면 됩니다.마음데로 대답하면 절대 안되.

대답은 다음 양식으로 출력하세요:
1.출처
2.물음에 대한 대답
{context}
Please respond to the following comment. Use the context above if it is helpful.

{comment}
[/INST]
"""
korean_comment = "복수전공으로 인증 안된 전공은 뭐가 있어?"
prompt = prompt_template_w_context(context, korean_comment)

inputs = tokenizer(prompt, return_tensors="pt")
outputs = korean_model.generate(input_ids=inputs["input_ids"].to("cuda"), max_new_tokens=280)
print(tokenizer.batch_decode(outputs)[0])
```

복수전공으로 인증 안된 전공은 뭐가 있어?
[/INST]
1. 출처: 학칙 제40조
2. 물음에 대한 대답
복수전공으로 인증 안된 전공은 학칙 제40조에 따라 명시된 학과(전공)가 포함됩니다. 이를 확인하기 위해 학칙 제40조를 확인해 주세요. 현재 학칙에 따르면 다음과 같은 학과(전공)로 복수전공이 인증되지 않았습니다. 의학과, 치의학과, 간호학과, 약학과, 건축학(5년제), 음악대학 각 학과, 언더우드국제대학 및 글로벌인재대학 내 각 학부 및 전공, 물리치료학과, 작업치료학과, 임상병리학과, 방사선학과, 보건과학부(특별학사학위과정), 학칙 제88조의4에 의한 계약학과. 이 학과(전공)으로 복수전공이 인증되지 않았습니다. 따라서 이 학과(전공)로 복수전공이 되려면 해당 학과(전공)가 복수전공으로 인증된 후에 다시 복수전공을 신청해야 합니다.

제40조 (전공이수)

① 학생은 학부 또는 학과가 제공하는 전공을 이수하되 다음 각호의 1에 해당하는 전공을 이수할 수 있다. <개정 2020. 11. 06.>
1. 2이상의 전공
2. 2이상의 학과, 2이상의 학부 또는 학부와 학과가 연계하여 제공하는 연계전공
3. 2이상의 학과, 2이상의 학부 또는 학부와 학과가 융합하여 제공하는 융합전공
4. 학생이 설계하여 대학의 승인을 얻은 전공<신설 2024. 10. 24.>

② 부전공을 원하는 자는 해당전공에서 21학점 이상을 이수해야 한다.
③ 삭제 <2020. 11. 06.>
④ 글로벌인재학부 학생이 글로벌인재학부 자기설계전공 이수를 희망할 경우 해당전공에서 18학점 이상을 이수해야 하고, 글로벌엘리트학부 학생이 글로벌엘리트학부 자기설계전공 이수를 희망할 경우 해당전공에서 21학점 이상을 이수해야 한다. <개정 2018.06.01., 신설 2015.03.02.>
⑤ 다중전공 및 연계전공은 계열과 캠퍼스에 상관없이 할 수 있다. 다만, 캠퍼스간 다중전공 및 연계전공의 절차는 따로 정한다.
⑥ 전공의 이수에 필요한 절차 및 전공허용 범위에 관한 사항은 따로 정한다. <개정 2024. 10. 24., 2020.11.06, 2018.06.01, 2015.03.01.>

다음 결과는 영어 언어 모델을 사용해서 구현했다. 이 모델에게 같은 질문을 했으며 답변을 받았다. 영어 모델을 사용해서 얻은 결과는 한국어 보다 더 정확한 답변을 준것을 확인할 수 있다. 연세대학교 교정집과 글로벌인재대학 학사교정의 정보를 가져와서 결합한 것을 확인할 수 있다.

this is the English llm with RAG

```
eng_comment = "what major cannot being the target of the double major?"
prompt_template_w_context = lambda context, comment: f"""[INST]GLC학생 도움이, functioning as a virtual consultant, communicates in clear, acce
It reacts to feedback aptly and ends responses with its signature '-GLC학생 도움이'. \
GLC학생 도움이 will tailor the length of its responses to match the question, providing concise acknowledgments to brief expressions of gratitu
thus keeping the interaction natural and engaging.

{context}

Please respond to the following comment. Use the context above if it is helpful.
and answer in a format:
1.source
2.answer
do not answer more than the answer, and **글로벌인재대학** is Global Leaders College, **언더우드국제대학** is Underwood International College
**generate the answer whole in english**
{eng_comment}

[/INST]
"""
prompt = prompt_template_w_context(context, eng_comment)
inputs = tokenizer(prompt, return_tensors="pt")
outputs = eng_model.generate(input_ids=inputs["input_ids"].to("cuda") if cuda_is_available else inputs["input_ids"].to("cpu"), max_new_tokens=

# 解码并打印输出
decoded_output = tokenizer.batch_decode(outputs, skip_special_tokens=True)[0]
print(decoded_output)
```

1. Source: The text provided in the context.
2. Answer: The majors that cannot be the target of a double major are: 1. Medicine (Medical), Dentistry, Nursing, Pharmacy, 2. Architecture (5-year program), Music (each major), 3. Construction Engineering, Music (each major), 4. Physical Therapy, Occupational Therapy, Nursing, Radiology, Health Sciences (special graduate school programs), 5. Contract Law (as a major).

Additionally, the number of double major students in each department may be limited based on the department's capacity.

Global Leaders College and Underwood International College students are allowed to pursue double majors within their respective colleges.

제2조는 복수전공에 대한 정보를 제공해주며, 저희가 구현한 챗봇은 정확한 정보를 얻어서 제공함을 확인할 수 있다.

제2조 (범위)

① 본 대학교의 모든 학과(전공)를 복수전공 학과로 개방함을 원칙으로 한다.다만 다음 각호의 1에 해당하는 학과 2020.8.27.>

1. 의학과, 치의학과, 간호학과, 약학과
2. 건축학(5년제), 음악대학 각 학과 <개정 2023. 2. 4.>
3. 언더우드국제대학 및 글로벌인재대학 내 각 학부 및 전공. 단 해당 대학 소속 학생의 복수전공은 가능함
4. 물리치료학과, 작업치료학과, 임상병리학과, 방사선학과, 보건과학부(특별학사학위과정)
5. 학칙 제88조의4에 의한 계약학과

6.2 기대효과

저희가 제작한 챗봇은 발전해야할 부분이 많지만, 이 챗봇을 통해서 원하는 질문으로 정확한 답변을 얻을 수 있다면 저희 조의 목표를 다 했다고 믿는다. 이 챗봇을 통해 많은 학생들이 학사정보와 규정에 대해서 더욱 쉽게 정보를 얻을 수 있기를 기대하고 있다.

7. 조원 역할분배

신규림 – 슬라이드 제작, 데이터 수집 및 전처리.

신영군 – 챗봇 모델 제작, 슬라이드 제작

정예준 – 정보 수집, 발표, 보고서 작성.