LLM과 RAG를 활용한 ChatGPT 기반 학습 지원 플랫폼 구축

ChatGPT based on LLM and RAG with ChatGPT learning support platform

신한대학교 소프트웨어융합학과
김민지[1], 김준성, 윤성호, 조하늘, 이태웅, 김민준, 신유림, 임상진, 황만수[2]

[1]신한대학교 소프트웨어융합학과
e-mail : mingdeee01@gmail.com

[2]신한대학교 소프트웨어융합학과
e-mail : mshwang@shinhan.ac.kr

Shinhan University
Min-Ji Kim, Jun-Seong Kim, Ha-neul Cho, Seong-Ho Yun, Tae-Woong Lee, Min-Jun Kim, Yu-Rim Shin, Sang-Jin Lim

요 약

본 논문은 LLM(ChatGPT)과 RAG를 활용해서 학습 지원 플랫폼 구축을 통해 대학생들의 학습 여건을 향상시키는 방법으로 이미지 형식의 자료를 텍스트화 하기위한 OCR 모델 선정, ChatGPT를 통한 데이터 전처리 기법을 사용해 RAG를 생성하고 해당 RAG 기반으로 문제를 생성하는 방법을 제안하였다.

또한 학생들의 학습 효율 향상을 위해 커뮤니티와 스터디그룹 기능을 구현하였다.

키워드 : RAG, LLM, 학습 지원, 커뮤니티

This paper proposed a method to improve the learning conditions of university students by building a learning support platform using LLM (ChatGPT) and RAG to select an OCR model for textualizing image-based materials, generate RAGs using data preprocessing techniques through ChatGPT, and create questions based on those RAGs.

We also implemented community and study group functions to improve students' learning efficiency.

감사의 글

"본 연구는 2023년도 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음" (2023-0-00089)

I. 서론

현재 대학 교육 과정은 학생들이 전공 개념을 충분히 이해하고 스스로 평가하기에는 한계가 있다. 전공 도서에서 제공하는 문제의 질과 양이 충분하지 않고, 정확한 답안 및 해설을 얻기에는 어려움이 있기 때문이다.

중·고등학생을 대상으로 한 AI 기반 학습 서비스들은 콴다, GetGPT(시험문제 만들기 AI) 등 이미 존재한다. 하지만 대학생들을 대상으로 한 서비스는 존재하지 않았다. 심지어 이미 구현된 콴다, GetGPT 마저 문제 생성의 기능을 완전히 갖추고 있지는 않은 상태였다. 콴다의 경우 기출 문제 제공과 문제 풀이에 초점이 맞추어져 있는 서비스이며, GetGPT의 경우 문제 생성 기능을 지원하고 있으나 문제만을 생성해주고 해답과 해설을 제공하지 않는다.

이에 본 프로젝트는 대학생들을 위한 AI 학습 시스템을 개발하고자 한다. AI를 통해 다양한 참고 자료를 이용하여 기존의 부족한 학습 자료를 보충하고 사용자가 학습 환경에 구애받지 않도록 한다. AI는 제공되는 학습 자료를 토대로 사용자를 분석하여 개인에게 맞추어진 양질의 문제를 생성하도록 한다. 문제 뿐만 아니라 해설과 해답까지 제공하며, 원하는 문제와 답안을 저장하여 추후에 다시 조회할 수 있는 오답노트 기능까지 지원한다.

추가로 [1] Louis Deslauriers가 제시한 논문을 참고하여 스터디그룹 기능, 커뮤니티 기능을 추가하였다. 해당 논문에는 학습 효율은 수동적인 학습보다 능동적인 학습 방법을 적용했을 때 효율이 상승한다는 사실이 서술되어 있었다. 따라서 학생들의 효과적인 학습을 위해 능동적으로 학습할 수 있는 스터디그룹이 필요하다고 판단하였다. 위 기능을 통해 동일한 학습 목표를 가진 사용자들이 정보를 공유하여 전공 개념에 대해 더 자세히 이해할 수 있게 하는 것을 목표로 하였다.

II. 본론

2.1 OCR

OCR은 광학 문자 인식이라는 뜻으로 문자를 기계가 해독할 수 있는 코드로 변환하는 기능을 한다. 이를 통해 컴퓨터가 종이 문서, PDF 파일 또는 이미지 내의 텍스트를 편집 및 검색할 수 있는 전자 텍스트로 변환하도록 한다. 본 시스템 개발을 위해서는 OCR 모델 채택이 필요해 보였다. 따라서 기존에 이미 개발되어 있고, 한국어 사용에 특화되어 있는 네이버 클로바 OCR 모델을 채택하였다.

2.2 ChatGPT

OCR 기능을 통해 얻은 파일에서 도서 정보, 머리말 등 불필요한 데이터를 삭제하기 위해 ChatGPT를 활용하였다. ChatGPT를 이용한 데이터 전처리 과정은 다음과 같다. 효과적으로 문제를 출제할 수 있도록 ChatGPT를 활용하여 마크다운 형식으로 변환한다. 마크다운 형식으로 변환된 자료는 문제 생성에 필요한 목차 추출에 효과적인 성능을 보였다. 따라서 변환된 자료를 다시 ChatGPT를 활용하여 문제를 생성하였다.

```
prompt = PromptTemplate(
    input_variables=["ocr_text","history"],
    template="""
        당신의 역할은 문제 생성을 지원하는 데이터 전처리 AI입니다.
        당신의 임무는 교사나 교수 AI가 문제 생성을 위해 사용할 수 있는 데이터를 정리하고 불필요한 내용 또는 단어를 제거하는 것입니다.
        가능한 한 원본 데이터의 형식을 유지하면서 텍스트를 마크다운 형식으로 변환해 주세요.

        다음은 OCR로 추출된 텍스트입니다. 불필요한 내용이나 단어를 제거하고, 문제 생성을 위한 필수적인 정보만 남겨주세요:

        ---

        {ocr_text}

        ---

        다음은 이전에 처리한 결과물입니다. 텍스트의 흐름이 자연스럽게 이어질 수 있도록 주의하여 작업해 주세요:

        ---

        {history}

        ---

        변환 규칙은 다음과 같습니다:
        1. 제목(Heading)은 '#', '##', '###' 형식으로 변환해 주세요.
        2. 번호가 있는 목록은 '1.', '2.', '3.' 형식으로, 번호가 없는 목록은 '-' 기호로 변환해 주세요.
        3. 강조된 텍스트는 *기울임* 또는 **굵게**로 변환해 주세요.
        4. 인용문은 '>' 기호를 사용해 변환해 주세요.
        5. 코드 블록은 ```로 감싸 주세요.
        6. 그림 설명과 관련된 내용은 삭제해 주세요.
        7. 문제 생성에 필요 없는 섹션(예: "도서 정보", "머리말", "저자 소개", "개정 내용")은 모두 삭제해 주세요.
        8. 섹션의 제목이나 형식은 유지하지 말고, 해당 섹션 전체를 삭제해 주세요.
        9. **중요**: 텍스트 전체를 코드 블록으로 묶지 말고, 코드 블록만 ```로 감싸서 사용해 주세요. 나머지 텍스트는 일반 마크다운 형식으로 작성해 주세요.

    """
)
```

그림1. 데이터 전처리 프롬프트

```
prompt_quiz = PromptTemplate(
    input_variables=["topic","referenced_header","key_keywords"],
    template="""
    당신의 역할은 선생님 혹은 교수입니다. 주어진 교재와 목차와 키워드를 바탕으로 객관식 문제를 생성하는 임무가 주어졌습니다.
    학생 수준에 따라 문제 를 출제해야합니다. 각 문제는 다음과 같은 형식을 따라야 합니다:
    1. 문제형식은 객관식(4개의 선택지 제공)
    2. 문제 목표:
    - 학생이 이해하기 어려운 개념이나 내용을 확인하는 문제
    - 학생이 실수할 가능성이 높은 함정문제
    - 주제에 대해 자주 놓치는부분을 상기시킬수있는 문제
    3. 문제 난이도 : 주제의 중요도에 따라 '상','중','하'를 조절하여 문제를 출제합니다.
    4. 오답선택지: 오답은 실제로 학생들이 혼동할수있는 내용으로 구성하여 단순히 틀린답을 제공하지않고 학습에 도움이 되도록 만듭니다.
    5. 문제수: 1개
    6. 학습 대상: 대학교 학부생

    다음 주어진 문제의 핵심 주제입니다.
    {topic}

    다음 주어진 교재의 핵심 키워드입니다.
    {key_keywords}

    다음 주어진 교재의 목차입니다.
    {referenced_header}

    응답은 다음 **반드시 JSON 형식**이어야 합니다:
    {{
        "question": "<여기에 문제를 입력하세요>",
        "options": [
            "1. <선택지 1>",
            "2. <선택지 2>",
            "3. <선택지 3>",
            "4. <선택지 4>"
        ],
        "explanation": {{
            "1": "<선택지 1에 대한 설명>",
            "2": "<선택지 2에 대한 설명>",
            "3": "<선택지 3에 대한 설명>",
            "4": "<선택지 4에 대한 설명>"
        }},
        "correct_answer": "<정답 선택지 번호>"
    }}

    """)
```

그림2. 문제 생성 프롬프트

2.3 Chroma DB

Chroma DB는 데이터를 클라우드 뿐만 아니라 로컬로 저장할 수 있다. 그리고 Chroma DB는 데이터를 컬렉션 단위로 관리한다. 따라서 학술 자료를 DB에서 손쉽게 로드 가능하다. 그리고 언어 모델을 랭체인 프레임워크를 통해 손쉽게 통합할 수 있다. 이러한 장점들을 활요하기 위해 Chroma DB를 사용하게 되었다.

Chroma DB가 사용되는 과정은 다음과 같다. 전처리한 데이터가 헤더별로 나뉘게 되고, 한 번 더 분할되어 Chroma DB에 저장된다. 이때 추후 문제 생성 시 빠르게 사용될 수 있도록 분할된 모든 문서 메타데이

터에 키워드와 목차를 추가한다. 추가된 키워드와 목차는 프롬프트 엔지니어링을 통해 문제의 주제를 선정할 때 사용된다. 위와 같은 과정으로 컬렉션 단위 정보가 저장되어 로드될 때 문제 목표, 난이도, 오답 선택지, 학습 대상 등이 저장된 사전정보를 토대로 사용자에 알맞은 주제가 선정된다.

2.4 React

React는 대규모 애플리케이션 제작의 유지보수와 사용자 경험에 쉽다. 특히 가상 DOM을 통한 빠른 렌더링과 컴포넌트 기반 아키텍처를 제공하는 것에 특화돼 있다. 따라서 개발의 효율성을 극대화하고 최적화할 수 있도록 axios와 fetch를 적절히 활용하여 백엔드의 api와 연결하였다. 이때 사용된 백엔드의 프레임워크는 SpringBoot를 사용하였고, DBMS는 MySQL을 사용했다.

2.5 RAG

RAG는 대규모 언어 모델의 출력을 최적화하여 응답을 생성하기 전에 신뢰할 수 있는 지식 베이스를 참조한다. 이를 통해 주제, 키워드, 목차를 검색 후 비슷한 문서를 찾아 AI가 참조하여 문제를 생성하도록 한다.

Ⅲ. 구현

구현에 사용된 시스템은 OCR, ChatGPT 기반 데이터 전처리, Chroma DB, React, RAG이다. 위 기능들을 통해 업로드된 학습 자료를 기반으로 사용자에게 알맞은 문제를 생성하도록 하였다. 해당 시스템은 사용자가 자료를 업로드할 때 OCR과 프롬프트 엔지니어링을 통해 텍스트 추출 및 데이터 전처리 과정을 수행한다. 이후 처리된 학습 자료를 벡터 DB에 저장한다. 저장된 학습 자료를 기반으로 핵심 키워드와 목차를 추출하여 문제를 생성한다. 이때 생성형 AI를 활용하여 키워드, 목차를 바탕으로 한 주제를 생성한다. 문제의 질을 높이기 위해 생성된 주제를 RAG 기술을 활용하여 외부 데이터베이스에서 검색하고 난이도, 학습 대상, 문제 유형을 정한다.

사용자가 업로드하는 자료의 정보 크기는 매우 방대하다. 따라서 문제 생성 시 LLM의 토큰 수 제한 이슈가 발생할 수 있다. 이 문제를 해결하기 위해 키워드를 추출하고 핵심 주제를 생성하여 DB에 추가적인 정보로 질 높은 문제를 생성하였다. 위 과정을 거쳐 생성된 문제는 사용자에게 제공된다.

키워드 추출 알고리즘은 [2]다양한 자료에서 키워드를 추출할 수 있으며 단어의 위치나 문맥을 고려한 키워드를 추출할 수 있는 성능을 가진 Yake 모델을 선택하였다.

Ⅳ. 결론 및 향후 구현 방향

구현된 학습 지원 시스템은 빠른 학습 피드백을 기반으로 사용자에게 큰 편의성을 제공해 줄 것이다.

본 시스템 이용을 통해 대학생들의 [3]학습 여건 개선을 기대할 수 있으며, 같은 학습 목표를 가진 사용자들의 소통으로 동기 부여와 능동적인 학습 분위기를 형성할 수 있을 것이다. 실제로 팀원들이 사용해본 결과, 해당 시스템으로 학습 편의가 올라간 것을 확인할 수 있었다.

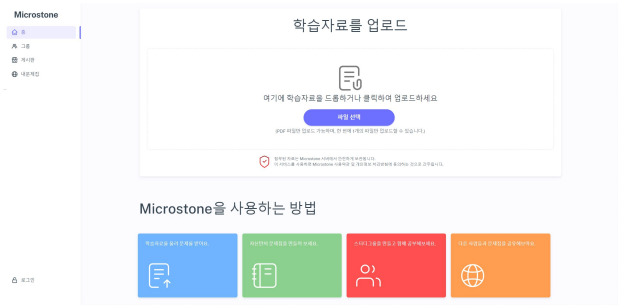

그림3. 최종 구현된 메인 페이지

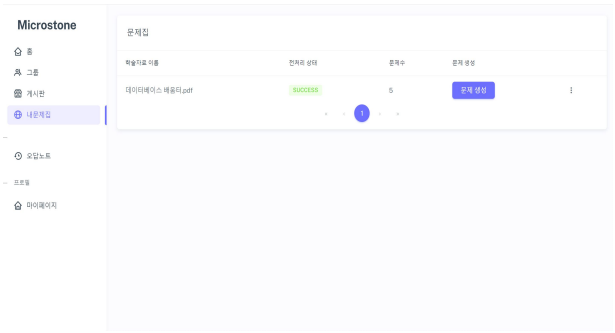

그림4. 문제 생성이 완료된 화면

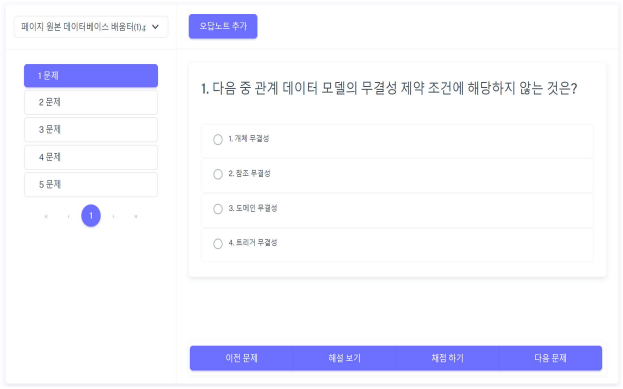

그림5. 문제 생성된 화면

계획했던 기능 중 일부분인 문제 생성까지는 구현이 완료되었으나, 사용자가 제공받은 문제를 다른 사용자들에게 공유하는 기능은 아직 구현되지 않았다.

향후 구현에서는 문제 공유 기능 추가 구현, 그리고 하나의 자료에서 문제를 생성하는 것이 아닌 여러 자료를 이용하여 문제를 생성하는 기능, 문제 유형이 객관식뿐만이 아닌 주관식, O/X문제, 서술형과 단답형 등으로 다양하게 늘려나갈 것이다.

참고문헌

[1] Deslauriers, Louis, et al.: Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. Proceedings of the National Academy of Sciences, vol. 116, no. 39, Sept. 2019, pp. 19251 - 57.

[2] Campos, Ricardo et al. “YAKE! Collection-Independent Automatic Keyword Extractor.” European Conference on Information Retrieval (2018).

[3] Castañeda, L., Selwyn, N. More than tools? Making sense of the ongoing digitizations of higher education. Int J Educ Technol High Educ 15, 22 (2018). https://doi.org/10.1186/s41239-018-0109-y

[4] 웹 프론트엔드 프레임워크 및 라이브러리 장단점 연구 : Angular, React, Vue 중심으로 = A study of web front-end framework and library pros and cons : focused on Angular, React, Vue (2019)

[5] Front-End Development in React: An Overview. Engineering International, Volume 7, No. 2 (2019)

[6] Chen, S., Thaduri, U. R., & Ballamudi, V. K. R. (2019). Front-end development in React: An overview. Engineering International, 7(2)

[7] Maratkar, P. S., & Adkar, P. (2021). React JS - An emerging frontend Javascript library. IRE Journals, 4(12), 99-102.

[8] Ritwik, C., & Sandeep, A. (2020). React.js and front end development. International Research Journal of Engineering and Technology (IRJET), 7(4), 3676 - 3679.