생성형 인공지능을 활용한 챗봇 학습 데이터 저작도구 개발

강연우, 박하늘, 박수연, 김영서, 김주성, 마창석, 염재민, 윤태복
서일대학교 AI융합콘텐츠학과
e-mail:tbyoon@seoil.ac.kr

Development of a chatbot training data generation tool using generative artificial intelligence

Yeon-Woo Kang, Ha-Neul Park, Su-Yeon Park, Young-Seo Kim, Ju-Sung Kim, Chang-Seok Ma, Jae-Min Yom, and Tae-bok Yoon
Dept. of AI Convergence Contents, Seoil University

요 약

최근 정보통신기술의 발달로 인공지능 기술을 접목한 챗봇이 다양한 분야에서 활용되고 있다. 특히 기업, 공공기관, 학교 등에서는 챗봇 서비스를 도입하여 업무 시간 외 뿐만 아니라 민원이 많아지는 시간에도 빠르게 민원인들의 민원처리를 해결하고 있다. 챗봇을 서비스하려는 기관에서는 기관의 특성을 고려하여 적절하게 응답 할 수 있도록 추가 학습하는 과정을 거쳐야하는데, 이 때 다양한 질문과 답변 데이터가 필요하다. 하지만, 질문과 답변을 생성하는 과정은 많은 시간과 인력을 필요로 하여 서비스의 어려움으로 작용하고 있다. 본 연구에서는 사전학습 모델의 미세조정(fine tuning)을 위하여 기관의 행정 문서를 질의 응답 형태로 변환해주고, 학습하는 저작도구를 제안하고자 한다. 실험에서는 대학의 규정류를 활용하여 학습데이터를 추출하고 학습에 활용하였으며, 적절하게 서비스하는 결과를 확인하였다.

1. 서론

최근 인공지능 기술의 발달로 여러 분야에 인공지능 기술을 접목시켜 사람들에게 더욱 편리한 서비스를 제공하는 사례가 증가하고 있다. 예를 들어 학교에서는 챗봇 시스템을 도입해 학생 또는 학부모들이 궁금해 하는 학사일정, 학교급식, 장학금, 수업시간, 성적처리 등의 질문들을 시간과 공간의 제약없이 서비스하고 있다. 챗봇 서비스 이전에는 재학생 커뮤니티나 학교 홈페이지에 마련되어 있는 질문 게시판에 글을 올려 담당자가 답변을 줄 때까지 기다려야 되는 시간의 소모가 컸다. 하지만 챗봇을 도입하고 나서부터는 업무 외 시간뿐만 아니라 대학 행정 서비스에 관한 질문이 가장 몰리는 3월과 입학 관련 질문이 몰리는 9월에도 학생 민원을 신속하게 해결할 수 있게 되었다[1].

전통적으로 사용되는 챗봇은 특정 질문, 특정 업무에만 특화되어 있거나 지정해 놓은 대화 흐름에 따라 제한된 답변 내용을 제공하여 제한된 대화 처리 능력을 지니고 있어 한계가 있었다. 하지만 최근에는 인공지능 기술의 발달로 사용자의 질문 내용을 자연어 처리(Natural Language Processing: NLP)와 거대언어모델(Large Language Model :LLM) 등과 같은 인공지능 기술을 통해 사용자의 의도를 파악하고 분석하여 사용자에게 더 정확하고 적합한 답을 제공하는 인공지능 챗봇 서비스가 활발하게 사용되고 있다.

그중에서도 OpenAI에서 개발한 “ChatGPT”는 인공지능 분야에서 큰 주목을 받고 있다. ChatGPT는 GPT(Generative Pre-trained Transformer) 모델을 적용한 생성형 AI로 인터넷에서 수집한 막대한 양의 텍스트 데이터를 기반으로 학습되었으며 사용자의 언어인 자연어를 이해하고 생성할 수 있다[2]. 이를 통해 다양한 주제와 분야에 대한 대화를 지원하여 학습한 내용을 기반으로 사용자의 질문에 대한 답을 하거나, 함께 문제를 해결하고 조언하는 등의 서비스를 제공할 수 있게 되었다[2,3].

특히 OpenAI에서는 GPT 모델을 기반으로 데이터를 학습시키는 방법은 미세조정(Fine-tuning)을 제공한다. 이에 사용자들이 특정 분야, 주제에 대한 데이터를 생성하여 학습하면 해당 내용에 대해 전문성 있는 챗봇을 제작할 수 있게 되었다. 하지만 데이터를 생성하고 학습시키는 과정이 컴퓨터 프로그래밍을 다뤄보지 못한 일반인들이 다루기에는 데이터 생성부터 학습시키는 과정이 복잡하다[4]. 따라서 본 논문에서는 데이터를 생성시키고 학습시키는 과정을 쉽게 처리할 수 있는 저작도구를 개발하여 다양한 사람들이 챗봇과 인공지능에 관심을 가질 수 있도록 하고자 한다. 이뿐만 아니라 학교에서 제공하는 규정집 내용의 데이터를 학습하여 학교 행정 특화 챗봇을 개발하고 테스트한다.

2. 챗봇 학습 데이터 저작도구 개발

대부분의 챗봇은 사용자의 질문 내용을 분석하여 사용자의 의도를 파악하고 해당 의도에 맞게 전문성 있는 내용으로 답변하는 과정을 가지고 있다. 하지만 가장 중요한 전문성 있는 답변의 생성 방법에 있어 너무 많은 시간을 소모하게 되며, 만들었다 하여도 챗봇에 사용자 데이터를 추가할 때의 방법인 프로그래밍과 인공지능 기법을 익히는데, 오랜 시간이 걸리므로 진입장벽이 높게 느껴진다. 이에 데이터를 학습하는 과정에서 효율성을 높이기 위한 방안이 필요하다. OpenAI에서 제공하는 학습 방법을 바탕으로 저작도구를 설계하였으며, 특히 대학의 규정집 관련 데이터 특화 저작도구로 설계하여 다양한 부가 기능을 추가하였다.

2.1 전체 시스템 설계

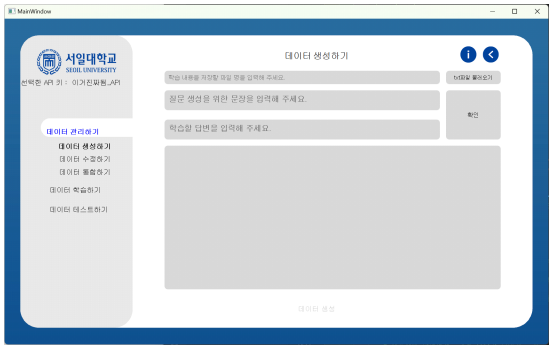
본 논문에서 제안하는 저작도구는 대용량 텍스트 파일을 데이터 학습 특화 파일로 변환하여 질문 생성 후 학습한다. 챗봇 제작에 사용된 모델은 OpenAI에서 제공하는 fine-tuning 관련 개발자 문서에 따르면 학습 데이터 세트는 질문-답변으로 구성되고 JSONL 파일로 저장되어 있어야 한다. 따라서 본 저작도구에서는 학교의 행정 문서의 파일 형식인 PDF 파일을 불러와 조항별로 텍스트를 나누고 조항마다 질문 10개의 학습 데이터를 생성한다. 생성한 데이터는 사용자가 확인 후 수정하고 저장할 수 있도록 제작하였으며, 학습하기 단계를 통해 사용자가 보다 쉽게 데이터를 학습할 수 있도록 하였다.

2.2 데이터 추출

데이터 생성은 학교 학칙 관련 데이터 특화 저작도구에 초점을 둬 학교 홈페이지에서 구할 수 있는 학칙 문서를 학습용 데이터를 생성하기 위한 형식 파일로의 변환과 데이터 생성하였다. 이를 위하여 데이터 추출을 위한 PDF 처리, 학습 데이터 생성, 제공 모델 비교와 최적 데이터의 개수, 생성형 인공지능의 환각 현상 처리 등의 단계별 절차를 통하여 데이터를 추출한다.

2.3 데이터 학습하기

기존의 OpenAI 개발자 문서에 따르면 학습 데이터 파일을 제공하는 명령어를 통해 업로드하고, 학습하는 과정을 가지고 있다. 하지만 OpenAI와 서버 연결 문제로 학습이 지연되는 현상이 있으며, API 키를 등록하는 과정에서도 빈번한 오류가 발생하였다. 특히 API를 등록할 때는 컴퓨터 내 환경변수를 지정하여 데이터 학습 및 GPT 사용을 연결할 수 있다고 기록되었으나, 이를 활용한 방법으로 실행 시 개발 프로그램이 환경변수로 지정된 경로를 읽지 못하는 등의 다양한 이슈가 발생하였다. 본 논문의 저작도구에서는 우회 방법인 명령 프롬프트를 활용한 OpenAI 학습 방법을 사용하여 일반적인 학습 방법보다 안정성 있는 연결을 도모하였다.



[그림 1] 데이터 생성하기 UI 화면

2.4 데이터 테스트

학습 완료 시 학습된 데이터를 테스트할 수 있는 테스트하기가 활성화된다. 학습을 완료하면 출력과 저장되는 엔진 키를 활용하여 챗봇처럼 질문할 수 있게 된다. 챗봇을 실사용하기 전에 간단하게 테스트하여 해당 데이터의 성능을 테스트해 볼 수 있을 것으로 예상한다.

3. 결론 및 향후연구

본 논문에서는 학교 행정 특화 챗봇을 제작에 필요한 학습 데이터와 데이터를 학습하기 위해 OpenAI에서 제공하는 GPT 모델과 미세조정(Fine-tuning)을 활용하여 데이터를 생성하고 학습시킬 수 있는 저작도구를 개발하였다. 본 저작도구를 통해 데이터 생성의 어려움 문제, OpenAI의 API 키 환경변수 저장 및 입력 오류 문제와 서버와의 연결이 끊김 문제를 해결할 수 있었다. 향후 연구로는 챗봇 프로그램의 대한 사용자 피드백을 활용하여 서비스 품질을 향상 할 수 있는 방법의 연구가 필요하겠다.

참고문헌

[1] Jeong Cheonsu, “A Study on the Service Integration of Traditional Chatbot and ChatGPT”, Journal of Information Technology Applications & Management , pp. 11-28, 2023.08.

[2] 서교리, “인공지능 기반 챗봇 서비스의 국내외 동향 및 발전 전망 분석”, 한국정보화진흥원(NIA), 2018.

[3] 석광호, “사람과 대화하는 챗봇 기술, 어디까지 왔나”, 한국뇌과학연구원, pp. 8-11, 2023.03.

[4] Meghan Rimol, “Gartner Predicts Conversational AI Will Reduce Contact Center Agent Labor Costs by $80 Billion in 2026”, Gartner, 2022.08.