대규모 언어 모델 기반 대학 입시상담 챗봇

이세훈*, 이웅회*, 김지웅*, 노연수O
*인하공업전문대학 컴퓨터시스템과,
O인하공업전문대학 컴퓨터시스템과
e-mail: seihoon@inhatc.ac.kr*, leewh0829@gmail.com*, gomwoong0619@naver.com*, quanternoh@gmail.comO

College Admissions Counseling ChatBot based on a Large Language Models

Se-Hoon Lee*, Ung-Hoe Lee*, Ji-Woong Kim*, Yeon-Su NohO
*Dept. of Computer Systems & Engineering, Inha Technical College,
ODept. of Computer Systems & Engineering, Inha Technical College

● 요 약 ●

본 논문에서는 대규모 언어 모델(Large Language Models)을 기반으로 한 입학 상담용 챗봇을 설계하였다. 입시 전문 LLM은 Polyglot-ko 5.8B을 베이스 모델로 대학의 입시 관련 데이터를 수집, 가공한 후 데이터 증강을 하여 파인튜닝 하였다. 또한, 모델 성능 향상을 위해 RLHF의 후 공정을 진행하였다. 제안 챗봇은 생성한 입시 LLM을 기반으로 웹브라우저를 통해 접근하여 입시 상담 자동 응답 서비스를 활용할 수 있다.

키워드: 입시상담 챗봇(Admissions Counseling ChatBot), 입시 LLM(Admissions LLM), 파인튜닝(Fine-turning), 데이터 증강(Data Augmentation)

I. Introduction

대학 모집 시기 각 대학은 입학 상담 부서의 운영, 온라인 상담, 챗봇 서비스 등을 활용하여 본 대학에 관심이 있는 학생들을 지원한다. 그러나 상담 부서의 운영 시간 제한 문제와 온라인 상담의 실시간 응답 불가능성으로 인해 지속적인 정보 제공에 어려움이 있고, 패턴 매칭 기반의 챗봇 사용은 사전에 작성된 규칙과 패턴에 의존하기에 모든 가능한 대화 시나리오가 고려되지 않는다면 정보를 제공할 수 없다는 문제가 존재한다.

본 논문에서는 이러한 문제 해결을 위해 대규모 언어 모델(LLM) 기반 입학 상담용 챗봇을 개발한다. 한글 베이스 모델을 이용해 대학 입학팀으로 부터 데이터를 수집, 증강하여 베이스 모델을 파인튜닝 하고 RLHF(Reinforcement Learning from Human Feedback)의 후공정을 진행하여 서비스를 한다.

II. Preliminaries

챗봇 개발에는 두 가지 접근 방식 즉, 패턴 매칭과 AI 기반 접근법이 있다. 앤런 튜닝이 기계의 지능에 대한 개념을 정립한 이후 챗봇은 사용자의 프롬프트와 스크립트 일치에 따라 응답하는 패턴 매칭 기술을 기반으로 시작되었다[1]. 표 1은 패턴 매칭 기반 챗봇의 개발 역사와 각 장단점 및 활용 사항이다.

Table 1. History of pattern matching-based ChatBot

ChatBot	개발 연도	특징 기술	약점	활용
Turing test	1950	문답법	-	대화를 기준으로 기계의 지능이 있는지 판별
ELIZA	1966	패턴 매칭 기술	제한적 시식, 긴 대화 유지 불가, 대화의 맥락 학습 불가	심리 치료에 사용
PARRY	1972	패턴 매칭 기술, Personality 반영, ELIZA에 비해 더 나은 통제구조	낮은 감정 표현 능력, 느린 응답 속도, 대화의 맥락 학습 불가	정신질환자 검사 도구로 활용
Jabberwacky	1988	첫 인공 지능 기술 적용, CleverScript로 작성, 문맥 매칭 기술	느린 응답 속도, 다수 사용자 동시 작업 불가	-
Dr.Sbaitso (Sound Blaster AI TTS Operator)	1992	사운드 카드 기반의 디지털 음성 표시	복잡한 상호작용 불가	심리학자의 역할 수행
ALICE (Artificial Linguistic Internet Computer Entity)	1995	휴리스틱 패턴 매칭 기술, AIML(XML 스키마로 작성, 대화의 규칙 정의	지적 기능의 부재, 감정과 태도 반영 불가	-
SmartChild	2001	대화를 통해 정보 시스템 DB에 연결되어 해당 정보를 응답	-	AOL(America Online) MSN(Microsoft)

챗봇은 Siri, Google Assistant, Alexa, Cortana에 이어 ChatGPT에 이르기까지 NLP와 AI 기술의 진보에 따라 감정분석, 의도 인식, 문맥 이해 등의 기술을 활용하여 보다 개인화된 정확한 답변을 제공하도록 발전했다.

III. The Proposed Scheme

대규모 언어 모델은 대량의 라벨링 되지 않은 텍스트 데이터를 학습한 언어 모델로 라벨 데이터를 기반의 파인튜닝을 통해 모델이 특정 작업에 대해 더욱 효과적으로 수행되도록 조정할 수 있다.

그림 1을 통해 입시상담 챗봇의 데이터 수집부터 파인튜닝까지의 과정을 확인할 수 있다.

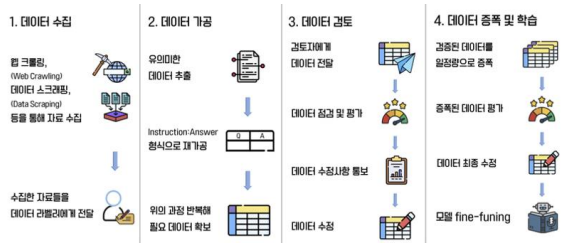

Fig. 1. Data Collection Flow

웹 크롤링과 데이터 스크래핑을 통해 입시 데이터를 수집하고, Question : Answer 형식으로 라벨링 한 후 입학팀을 통해 데이터에 대한 검토 및 수정을 거쳐 데이터 세트를 구축했다. 데이터 세트의 크기는 약 1000으로 작은 편이었기에 모델 학습 시 일반화 성능 향상을 목적으로 gpt-3.5-turbo model API[2]를 이용해 약 10,000 크기까지 데이터 증강을 진행했다.

텍스트 데이터 증강 시 문장의 의미를 유지하는 것이 중요하기에 표 2와 같이 지시문, 역할 부여, 목표 명시, 배경 정보 및 출력 문장의 포맷을 지정하는 등 프롬프트 엔지니어링을 통해 목적에 맞는 답변 생성을 요청한 후 생성된 데이터를 원본 데이터에 추가하여 데이터 세트를 재구축 하였다.

Table 2. Regulating prompts

생성 규제	요구 사항	
지시	Q	' 이 내용에 기반하여 동일한 의미를 가지지만 구조는 다른 다섯 개의 질문을 생성해 주세요.'
	A	' 이 내용에 기반하여 동일한 의미를 가지지만 구조는 다른 다섯 개의 답변을 생성해 주세요.'
역할 부여	Q	당신은 대학 입학에 관심이 있는 예비 대학생이라고 가정합니다.
	A	당신은 예비 대상생을 상대로 입학 상담을 진행하는 대학교 입학 상담관이라고 가정합니다.
목표 명시	각 문장들은 동일한 의미를 전달해야 하지만, 문장 구조는 서로 다르게 표현되어야 합니다.	
출력 문장의 포맷 지정	각 문장은 한 번의 줄 바꿈을 통해 구분되어야 합니다. 모든 분장들은 본분에 최대한 충실해야 합니다. 이는 원분의 수요 포인트를 반복하는 것, 또는 원문의 맥락을 이해하고 답변에 그것을 반영하는 것의 포함합니다.	
배경 정보	Q	-
	A	모든 문장들은 우리 대학교에 입학하고자 하는 잠재적인 학생들이 읽게 될 것이며, 그들이 우리 대학교에 주목할 수 있도록 해야 합니다.

파인튜닝 대상 모델은 한국어 서비스에서 좋은 성능을 내기 위해 한국어 언어 모델인 Polyglot-ko 5.8B을 베이스 모델로 사용했다.

입시상담 챗봇의 시스템 구성도는 그림 2와 같다. 사용자가 브라우저를 통해 서비스가 탑재된 웹에 접근하여 서비스를 요청하면, Back-End에서 모델을 서빙하는 Flask 서버로 RESTful API(POST)를 요청한다. 이후 Request를 받은 서버는 모델을 호출하여 답변을 응답받고, Response를 보냄으로써 사용자가 서비스를 이용할 수 있게 된다.

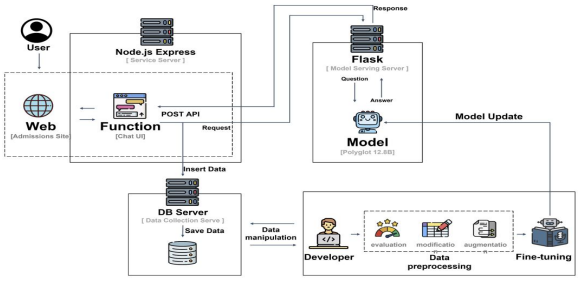

Fig. 2. System Architecture

IV. Conclusions

본 논문에서는 대규모 언어 모델을 기반으로 한 입학 상담용 챗봇을 개발하여, 자동응답 시스템을 입학 상담에 효과적으로 활용하는 것을 목표로 입시 데이터를 수집해 가공, 검토, 증강 및 파인튜닝하고, 이를 서비스하기 위한 Back-end 시스템을 구축했다. 현시점 챗봇은 사용자의 질문에 대해 이해하지만 상세한 응답 생성은 하지 못하는 문제점이 존재한다. 해당 문제점은 토픽별로 Knowledge Graph를 구축하여 전문화된 분야에서 응답을 생성할 수 있도록 개발할 예정이다.

본 논문에서 개발한 챗봇 시스템을 기반으로 대학의 입학 안내 사이트 메인 화면과 연동하여, 대학 입학에 관심이 있는 학생의 궁금한 점에 대하여 정확한 정보를 제공할 수 있는 입학 상담 AI로의 발전을 기대할 수 있다.

REFERENCES

[1] Eleni Adamopoulou, Lefteris Moussiades, "Chatbots: History, technology, and applications", Machine Learning with Applications, Vol. 2, 100006, Dec. 2020.

[2] OpenAI, "Language Models are Few-Shot Learners", arXiv:2005.14165v4, [cs.CL], Jul. 2020.