ChatPub: 검색 증강 생성 기반 청년 관련 정책 추천 서비스

김강산 [1], 박진호 [2], 양승빈 [3], 전창민 [4], 구형준 [5]
성균관대학교 (영어영문학과/소프트웨어학과 [1], 수학과/소프트웨어학과 [2],
데이터사이언스융합전공/소프트웨어학과 [3], 수학과/소프트웨어학과 [4]) 학부생
성균관대학교 교수 [5]
rkdtks419@g.skku.edu[1], jinho99@skku.edu[2], didtmdqls1@skku.edu[3] ,
jsjs1209@g.skku.edu[4], kevin.koo@skku.edu[5]

ChatPub: Retrieval Augmented Generation-based Service to Aid in Finding Relevant Policies for Korean Youth

요 약

본 논문은 검색 증강 생성 기법과 ChatGPT 를 결합한 사용자 맞춤 정책 추천 서비스인 ChatPub 을 소개한다. ChatPub 은 대한민국 청년을 대상으로 최소한의 개인 정보를 제공받아 적합한 정책을 추천해 주는 웹 서비스다. 정책 정보 사이트를 실시간으로 반영하는 데이터베이스를 참조함으로써 최신 정책 정보를 반영할 수 있으며, 사용자 친화적인 채팅 인터페이스를 통해 원하는 정책 정보에 쉽게 접근할 수 있다. 본 서비스를 통해 청년 정책의 접근성을 높이고 다양한 혜택을 쉽게 알림으로써 더 많은 기회를 제공할 수 있다.

1. 서론

2023 년 기준으로 대한민국 청년을 대상으로 한 청년 복지 관련 정책이 390 개가 선정되고, 예산이 25.4 조원이 책정되었다. [1] 하지만 청년이 혜택을 받기 위해 정부 정책 웹사이트에 공지된 4,000 개가 넘는 정책을 다 확인하기란 불가능하다. 이러한 문제점 때문에 많은 청년이 현재 실행 중인 정책에 대해 알지 못하며 자신에게 적합한 정책을 찾는데 어려움을 겪고 있다. 정책 접근성을 높이고 정보의 격차를 해소하기 위해 본 논문에서는 인공지능 기반의 웹 애플리케이션 챗봇인 ChatPub 을 제안한다. ChatPub 은 검색 증강 생성 (Retrieval Augmented Generation; RAG) [2] 기술을 활용한 챗봇 서비스이다. ChatPub 은 청년들을 대상으로 한 현재 시행 중인 정책 중 본인에게 맞는 정책을 쉽게 확인할 수 있도록 지원한다. 교육, 문화 교류, 여가 활동 등 다양한 분야로 확장할 수 있는 프레임워크 구조를 갖추고 있으며, 향후 다양한 연령층을 대상으로 한 챗봇 서비스로 발전될 것이다.

2. ChatPub 정책 추천 서비스 디자인

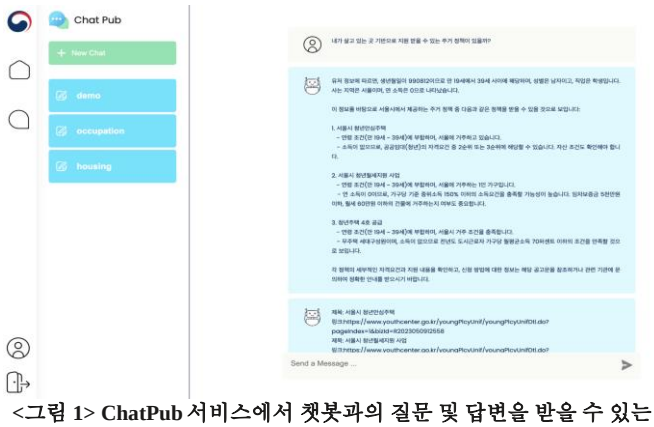
2.1 데이터 파이프라인

실시간으로 변경되는 정책 데이터를 효율적으로 관리하기 위해 Python, MariaDB, SQL 을 사용한 DDL 로 데이터 파이프라인을 구축하였다. 데이터 파이프라인은 총 3 단계로 구분된다. 1) EDA(탐색적 데이터 분석)는 데이터 파이프라인의 첫 번째 단계로, 데이터의 구조를 이해하는 것으로 시작한다. EDA 단계는 기존 데이터 세트에 대한 종합적인 이해를 제공하고 데이터 파이프라인 구축의 후속 단계를 용이하게 한다. 2) 다음은 크롤링 코드의 구현이다. 웹 크롤링 코드를 구현하기 위해 파이썬을 사용했으며, 웹 크롤링에 사용될 필요한 라이브러리와 모듈을 사용하였다. 3) 마지막은 데이터 전처리 및 통합이다. 크롤링한 데이터 및 텍스트는 모델이 아닌 사람을 위한 자연어 이므로, 모델에 맞게 데이터를 전처리해야 한다. 전처리 후에는 DB 와 통합시켜야 하므로 MariaDB 와 SQL 을 사용한 DDL 로 DB table 을 만들어 데이터베이스와 통합하였다. DB 는 **3.3** 의 검색엔진 구현부에서 유저의 질문과 가장 관련 있는 정보를 찾기 위해 사용된다.

2.2 검색 증강 생성 (RAG)

기능. ChatPub 은 사용자의 질문 의도를 이해하고 적절한 정책을 추천하기 위해 다음 세 가지 기능을 포함한다. 1) 사용자의 질문을 이해하고 자연스러운 답변을 생성한다. 2) 사용자에게 필요한 정책을 판단하고 추천한다. 3) 주기적으로 변경되는 정책 정보를 확인하고 답변 생성에 반영한다. 검색 증강 생성은 사전 훈련된 언어 모델과 외부 지식 소스를 검색하여 활용할 수 있는 능력을 결합한 자연어처리 기술로, 정보 검색 부분과 검색된 정보를 바탕으로 답변을 생성하는 생성 부분으로 구성된다.

장점. 검색 증강 생성은 외부 지식을 생성 모델에 반영하는 기존 미세조정 방식에 비해 크게 두 가지 장점이 있다. 1) 사전 미학습 정보 반영이 가능하다. ChatPub 은 주기적으로 삭제되거나 새로 생성된 정책을 반영해야 하는데, 미세 조정 방법론의 경우, 정책이 업데이트될 때마다 큰 LLM 모델을 새로 학습시켜야 하기 때문에 오버헤드가 높다. 반면, 검색 증강 생성의 경우, 추가 학습 없이 외부 지식 소스의 정보만 업데이트하므로 상대적으로 빠르고 용이하다. 2) Catastrophic Forgetting Problems 를 방지할 수 있다. 모델이 새로운 정보를 학습할 때 이전에 학습한 정보를 잊어버리는 문제가 발생할 수 있는데, LLM 에서 미세조정을 수행할 때 이러한 Catastrophic forgetting problems 발생할 수 있으나, 검색 증강 생성의 경우 추가 학습이 수행되지 않으므로 기존 지식을 보존할 수 있다.



<그림 1> ChatPub 서비스에서 챗봇과의 질문 및 답변을 받을 수 있는 웹페이지 사용자 인터페이스

3. ChatPub 정책 추천 서비스 구현

3.1 Front-End

Front-End 개발은 Figma 로 디자인하고, React 와 JavaScript 로 구현하였다. 비동기 처리와 데이터 요청에는 async/await 및 Axios 또는 Fetch API 를 사용하여 대응했고, React Router 로 라우팅과 내비게이션을 처리했다. <그림 1>은 실제 ChatPub 서비스의 사용자 인터페이스를 보여준다.

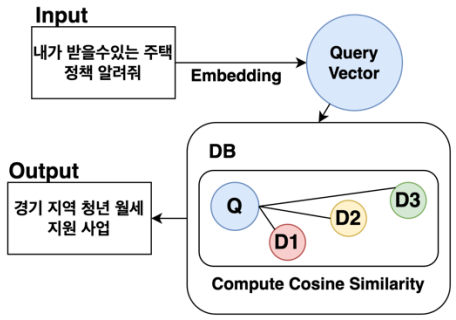
3.2 Back-End

FastAPI 는 빠른 속도, 직관성, 그리고 적은 버그를 가지고 있다. 이와 함께 사용되는 RESTful API 는 확장성, 재사용성, 그리고 유지보수의 편의성을 제공한다. MariaDB DBMS 를 사용하여 정책 데이터 및 사용자 데이터를 관리했고, FastAPI 웹 프레임워크를 통해 CRUD 에 따른 정보 관리를 수행했다.

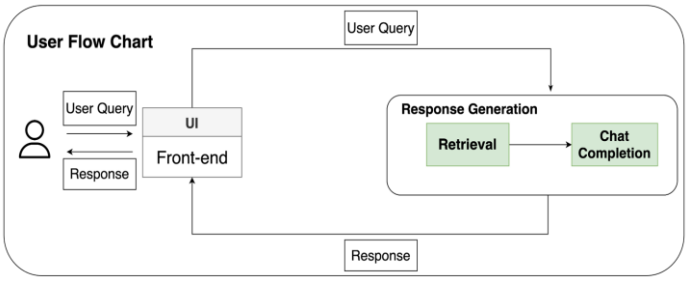
3.3 검색 증강 생성 기반 모델

검색 엔진. 검색 엔진은 사용자 질문과 관련된 정보를 **2.1** 에서 구축한 DB 에서 검색하는 부분으로 BERT [3] 기반의 언어 모델을 사용하고, 한국어 이해 평가(KLUE) [4] 데이터 세트에 사전 훈련된 Roberta [5] 모델에 풀링 계층을 추가하여 문장 트랜스포머를 구축했다. 주어진 질문과 관련된 정보를 지식 소스에서 추출하는 과정은 다음과 같다. 우선 DB 에 존재하는 텍스트 형식의 정보는 임베딩 되어 벡터형태로 저장된다. 입력 값으로 질문을 받으면, 쿼리 벡터로 변환한 후 구축한 DB 내 벡터와 코사인 유사도 연산을 수행하여 가장 유사한 벡터를 추출한다. 추출된 벡터의 인덱스를 통해 텍스트 형식으로 가장 유사한 정보를 추출한다. <그림 2>에서는 사용자의 질문이 임베딩 되어 가장 유사도 값이 높은 인덱스의 정보를 얻는 과정을 보여준다.

답변 생성 엔진. 최종 답변 생성 엔진은 사용자 질문과 DB 로부터 추출된 정보를 결합해 입력 후 답변을 생성한다. <그림 3>에서는 해당 입력 쿼리를 바탕으로 한 검색, 생성 프로세스를 보여준다. 본 논문에서는 답변 생성에 활용될 최적의 프롬프트를 찾기 위해 LLM 을 판단자로 사용하여 프롬프트를 평가하였다 [6]. 효율성과 객관성을 위해 GPT4 를 판단자로 활용해 생성된 답변 쌍을 받아 답변을 평가한다. 최종 프롬프트는 검색 증강 생성에서 일반적으로 활용되는 프롬프트를 비교군으로 하였을 때 가장 높은 성능을 보인



<그림 2> 사용자의 질문과 가장 유사한 정책 정보를 추출하는 모식도



<그림 3> 사용자의 입력 쿼리를 검색, 생성 프로세스로 거쳐 생성된 답변을 응답하는 모식도

프롬프트로 선정하였다. 구체적인 논문의 구현 코드는 깃허브 링크에서 확인할 수 있다. https://github.com/Chat-Pub/ChatPub

4. 결론

본 논문은 한국의 청년들이 지원 정책에 쉽게 접근하고 이해하며 신청할 수 있도록 하는 것을 목표로 하고 있으며, 이러한 솔루션은 정책 정보의 분산된 특성에서 비롯된 정보격차 문제점을 해결할 수 있다. 해당 기능의 중요성은 청년들이 정부 지원 정책에서 이익을 얻는 방식을 혁신할 수 있는 잠재력에 있다. 본 서비스를 통해 고성능 챗봇의 능력과 검색 증강 생성 방법론을 결합함으로써 정책 확인 과정을 단순화하는 것뿐만 아니라, 젊은 세대를 위한 포괄적이고 유익한 환경 조성의 목표에 기여할 수 있을 것이다.

ACKNOWLEGEMENTS

본 논문은 과학기술정보통신부 산하 정보통신기획평가원(융합보안대학원(성균관대학교))과 산학협력선도대학 육성사업 (LINC 3.0) 지원을 받아 수행한 연구임

참고문헌

[1] 중앙행정기관, 2023 년 청년정책 시행계획, 2023.03.29
[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K¨uttler, Mike Lewis, Wen-tau Yih, Tim Rockt¨aschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474, 2020.
[3] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), NAACL, pp. 4171–4186, 2019.
[4] Kim, S., Park, L., Oh, A., Ha, J.-W., and Cho, K. KLUE: Korean language understanding evaluation. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.
[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” CoRR, vol. abs/1907.11692, 2019.
[6] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging llm-as-a-judge with mt-bench and chatbot arena,” CoRR, vol. abs/2306.05685, 2023.