디지털 데이터 기반 검색 증강 생성을 활용한 독거노인을 위한 맞춤형 챗봇

김유청, 김동민*
순천향대학교, *순천향대학교

yc424k@gmail.com, *dmk@sch.ac.kr

Customized Chatbot for Elderly Living Alone Using Retrieval-Augmented Generation Based on Digital Data

Yu Cheong Kim and Dong Min Kim*
Soonchunhyang University

요 약

본 논문은 Meta Llama 3 모델을 활용하여 독거노인을 위한 챗봇에 대해 연구하였다. AI 에서 발생할 수 있는 오류를 최소화하기 위해 Retrieval_Augmented Generation 기법을 도입하였고, 독거 노인에게 실질적으로 도움이 되는 데이터에 집중하여 챗봇의 효율성과 정확성을 높였다. 본 연구는 독거노인의 일상생활을 지원하고, 그들의 삶의 질을 향상시킬 수 있는 기술적 접근을 제시하였다.

I. 서 론

본 논문은 자연어 처리와 인공지능 기술의 발전을 기반으로 하여, 독거노인을 위한 챗봇의 개발과 개선에 초점을 맞춘다. Meta Llama 3 모델[1]을 이용한 본 연구는 특히 대화 생성 모델의 한계를 극복하고자 Retrieval-Augmented Generation(RAG)[2]기법을 적용하여 실제적이고 신뢰성 있는 대화 생성을 목표로 한다. 최근의 딥러닝 기술은 대규모 데이터를 활용해 더욱 정교하고 진화된 텍스트 생성 능력을 보여주고 있지만, 이러한 모델들은 여전히 실제와 동떨어진 출력, 즉 'hallucination' 문제에 직면해 있다. 본 연구에서는 RAG 기법을 활용하여, 모델이 상황에 부적합한 정보를 제공하는 것을 최소화함으로써, 독거노인을 위한 챗봇이 보다 현실적이고 유용한 대화를 제공할 수 있도록 하였다. 이러한 접근은 지능형 홈케어 응용 분야에서도 자연어 처리 모델의 잠재력을 발휘할 수 있음을 보여준다. 본 연구는 딥러닝과 자연어처리 기술의 현실적인 지능형 홈케어 적용을 위한 가능성을 제시한다.

II. 본론

본 논문에서는 Meta Llama 3 모델을 활용하여 독거노인을 위한 챗봇을 개발하는 연구를 진행하였다. 특히, AI 에서 발생할 수 있는 오류를 최소화하고, 독거노인에게 실질적으로 도움이 될 수 있는 정확하고 신뢰할 수 있는 데이터를 생성을 목표로 RAG 기법을 도입하였다. 이 기법은 외부 데이터에서 정보를 효율적으로 검색하고 이를 생성 모델에 활용하여 답변의 정확도와 완성도를 높이는 방법으로, 본 연구에서는 이를 독거노인의 일상생활을 지원하고 그들의 삶의 질을 향상시킬 수 있는 챗봇 개발에 적용하였다.
본 연구의 데이터 처리 파이프라인은 다음과 같다(그림 1 참조).

1) 데이터 로드(Load)
RAG 의 첫 번째 단계는 외부 데이터를 시스템에 로딩하는 것이다. 본 연구에서는 PDF 파일 형태의 노인 데이터 파일을 입력 데이터로 사용하며, 이를 처리하기 위해 LangChain 라이브러리의 `PDFPlumberLoader` 를 사용하였다[3]. 이를 통해 PDF 파일을 구조화된 텍스트로 변환하고 이후 처리 단계를 위한 준비를 마쳤다.

2) 데이터 분할(Split)
소설과 같은 대용량 텍스트는 한 번에 처리하기 어려우므로 데이터 청크 단위로 분할한다. 본 연구에서는 Python 라이브러리 `RecursiveCharacterTextSplitter` 를 활용하여 최대 1024 자의 청크 크기와 80 자의 중첩을 두어 텍스트를 분할하였다.

3) 임베딩 생성(Embed)
각 청크는 `FastEmbedEmbeddings` 클래스를 사용하여 고차원 벡터로 표현되었다. 이러한 임베딩은 청크의 의미적 유사성을 수치적으로 나타내며, 이를 통해 유사성을 기반으로 한 검색이 가능해진다. 이 단계에서는 각 텍스트 청크에 대한 임베딩을 생성하고 저장하였다.

4) 임베딩 저장(Store)
생성된 임베딩과 원본 텍스트 데이터는 추후 검색 및 활용을 위해 벡터 데이터베이스에 저장된다. 본 연구에서는 `Chroma` 라이브러리를 사용하여 벡터 스토어를 구축하고, 이를 로컬 컴퓨팅 환경에 영구적으로 저장하였다.

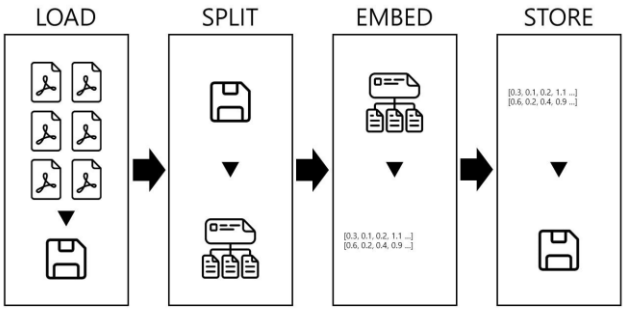

그림 1. RAG 기법 적용을 위한 데이터 처리 파이프라인

RAG 기법은 정보 검색과 생성 모델을 통합하여 질문에 대한 보다 정확하고 맥락에 맞는 답변을 제공한다. 이 과정은 그림 2 처럼 크게 Question, Retrieve, Prompt, LLM, Answer 의 다섯 단계로 구성된다. 본 연구에서는 Meta Llama 3 모델을 사용했으며 상세한 과정은 다음과 같다.

1) 질문(Question)
사용자로부터 질의 또는 질문이 주어진다. 본 연구에서 이 질문은 영어로 표현되어있다.
2) 검색(Retrieve)
입력된 질문을 바탕으로 벡터 스토어를 이용하여 유사한 문서를 검색한다. 벡터 스토어는 사전에 구축된 임베딩을 사용하여 유사성이 높은 텍스트 청크를 검색한다.
3) 프롬프트 구성(Prompt)
검색된 청크와 질문을 통합하여 프롬프트를 구성한다. 이 프롬프트는 질문에 대한 충분한 맥락을 제공하는 텍스트와 질문을 결합하여 언어 모델이 적절한 답변을 생성할 수 있도록 구성된다.
4) 언어 모델 처리(LLM)
Meta Llama 3 모델을 사용하여 프롬프트를 처리하고 답변을 생성한다. 프롬프트에 포함된 맥락과 질문을 바탕으로 언어 모델은 자연어로 답변을 생성하며, 이 단계는 생성된 답변의 품질과 관련된 중요한 역할을 한다.
5) 답변 생성(Answer)
언어 모델에서 생성된 답변은 시스템을 통해 사용자에게 반환된다. 또한 답변과 함께 참고한 문서나 정보의 출처도 함께 사용자에게 제공이 된다.

그림 2. 데이터를 모델에 전달하는 과정

독거노인을 위한 챗봇의 전체적인 동작은 그림 3 의 순서도와 같이 이루어진다. 초기에는 독거노인들을 위한 보편적인 데이터를 RAG 기법을 통해 적용한다. 이후 독거노인이 인공지능을 사용하면서 데이터가 축적되면, 이 데이터를 기반으로 다시 RAG 기법을 적용한다. 이 알고리즘을 사용하면 인공지능이 점점 더 개개인에게 맞는 대답을 하게 된다는 것이 확인되었다. 이와 같이 RAG 기법을 통해 인공지능에 개개인의 데이터를 적용하면, 각각의 독거노인을 위한 맞춤형 인공지능이 될 것으로 판단된다.

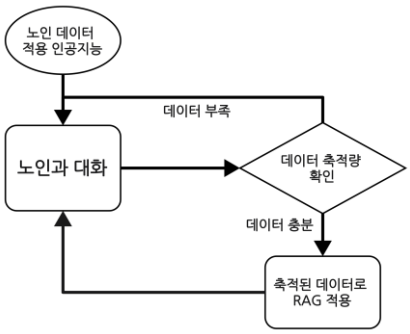

그림 3. 독거 노인 챗봇 순서도

III. 결론

본 논문에서는 LangChain 라이브러리와 RAG 기법을 활용하여, 독거노인을 위한 맞춤형 챗봇 개발의 가능성을 탐색하였다. 데이터 처리 파이프라인의 구축을 통해 임베딩된 데이터를 효율적으로 활용하였고, 특히 RAG 기법은 맥락에 부합하는 정보 제공을 최적화하여 사용자의 질의에 대한 더 정확하고 유의미한 답변을 생성할 수 있었다. 연구 과정에서 답변 생성의 품질에 대한 한계를 경험하였다. 주로 공공데이터에 의존한 결과, 데이터의 범용성은 확보되었으나 개인화된 대화 생성에는 제한이 있었다. 이는 독거노인과 같은 특정 사용자 그룹에 대해 더욱 세심하고 맞춤화된 응답을 제공하기 위한 데이터의 다양성과 깊이가 부족함을 나타낸다. 이러한 문제를 해결하기 위해, 추후 연구에서는 개별 사용자의 대화 내용을 실시간으로 수집하고 분석하여, 사용자 개인의 선호와 필요에 기반한 맞춤형 챗봇으로 RAG 기법을 확장할 계획이다. 이러한 접근은 독거노인이 직면한 고립감과 소외감을 완화하는데 기여할 뿐만 아니라 디지털 서비스가 사회적 연결고리를 강화하는 수단으로 작용할 수 있음을 시사한다. 이는 인공지능 기술이 단순한 기술적 진보를 넘어, 사회적 가치를 창출하고 인간 중심의 서비스를 제공하는 중요한 도구로 자리 잡을 수 있음을 보여준다.

ACKNOWLEDGMENT

이 논문은 2024 년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (P0012724, 2024 년 산업혁신인재성장지원사업). 이 논문은 2024 년 대한민국 교육부와 한국연구재단의 인문사회분야 일반공동연구지원사업(융복합연구)의 지원을 받아 수행된 연구임(NRF-2022S1A5A2A03052880).

참 고 문 헌

[1] Meta, “Introducing meta llama 3: The most capable openly available llm to date,” 2024, (https://ai.meta.com/blog/meta-llama-3/).

[2] P. Lewis, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in Proc. Advances in Neural Information Processing Systems 2020, Dec. 2020.

[3] LangChain, https://www.langchain.com/.