자연어 처리를 활용한 메신저 정보습득 서비스

Messenger information acquisition service using natural language processing

문지원 Jiwon Moon
경희대학교 소프트웨어융합학과
laonm@khu.ac.kr

지은주 Eunju Ji
경희대학교 소프트웨어융합학과
ejji01@khu.ac.kr

최진유 Jinyu Choi
경희대학교 소프트웨어융합학과
zzang1850@gmail.com

요약문

보편적으로 이용하는 채팅방에서 지나간 내용들은 금방 사라지고 본인에게 필요한 내용은 일부에 불과하다. 이 문제를 해결하기 위해 정보를 더 쉽게 습득할 수 있는 채팅방위키 서비스를 제시하는 바이다. '채팅방 위키'는 위키 형식을 이용하여 특정 채팅방의 내용을 정리하고 요약하여 누구나 볼 수 있도록 한다. 또한 권한이 있는 사람은 내용을 수정할 수 있도록 하여 더 유용한 정보를 얻을 수 있도록 한다. 본 연구에서는 한국어 대화 데이터셋과 인공지능 문서 요약 모델을 이용하여 대화에 맞는 요약 모델을 구현하였다. 대화를 작은 단위로 나누고 대화 내용을 요약하고, 그 결과를 데이터베이스에 저장하고 사용자가 조회할 수 있도록 하여 채팅방 위키 서비스를 구현하였다.

주제어

자연어 처리, 요약모델, 대화요약, T5

1. 서 론

자연어 처리에서 문서 자동 요약 기술은 필요성이 높고 활용 범위가 넓어 활발한 연구가 진행되어 왔다. 문서 요약은 크게 추출 요약과 생성 요약으로 나눌 수 있는데 이 중 생성요약은 자연스러운 요약문을 얻을 수 있지만 원문 텍스트를 그대로 사용하지 않고 새로운 문장을 구성해야 하므로 추출요약보다 구현하기 어렵다. 그래서 생성요약은 높은 난이도로 인해 비교적 최근에 인공신경망 기반의 NLG 모델의 발전과 함께 여러 모델들이 개발되었다[1]. 대화 데이터 요약은 기존 문서 요약보다 더 어려운 과제로 알려져 있다. 데이터 특성상 구어적 속성을 가지기 때문에 문장 성분의 생략과 축약적 표현의 사용 등이 빈번히 일어나고 둘 이상의 화자 간 대화도 요약해야 하기 때문이다[2].

한편, 비대면 의사소통의 중요성이 주목받고 있는 요즘 채팅방에서는 유용한 정보들을 주고받는다. 그 중 필요한 정보들을 골라서 보기 힘들 정도로 실시간으로 여러 정보들이 쏟아진다.

본 연구에서는 채팅방의 대화를 효과적으로 요약하는 방법을 고안하고자 한다. 또한 요약 내용을 위키 형식으로 정리해 사용자들이 쉽게 정보를 찾을 수 있도록 하는 서비스를 개발하고자 한다.

학습 데이터셋으로는 오픈채팅방과 한국어 대화 요약 데이터셋을 이용하였고 T5 모델을 학습시켜 사용하였다. 또한 KeyBERT 을 사용하여 키워드를 추출해 태그로 이용하여 가독성을 높였다. 또한 기존 문서 요약모델을 대화에 맞게 fine tuning 하여 요약의 완성도를 높였다.

2. 연구 내용

2.1 대화 데이터와 전처리

대화 요약 데이터셋은 오픈 채팅방의 대화를 수집하였다. 맞춤법과 문법의 오류가 잦고 이모티콘, 사진, URL 과 같은 대화 텍스트 이외의 형태가 자주 사용된다. 또한 같은 자음의 반복과 자음으로만 표현하는 등의 특성이 있다. 또한 채팅방의 특성으로 채팅방을 관리할 때 나타나는 메세지도 많다. 따라서 학습 성능의 향상을 위해 특수문자 및 이모티콘을 제거하고, 자음으로만 된 문장인 경우 제거하였다. 불규칙하게 반복되는 자음의 경우 문장인지 여부에 상관없이 제거하였다. 채팅방 관리에 사용되는 문구와 홍보성 메시지, 사진 등도 제거하였다. URL 은 중요 정보를 담고 있을 가능성이 크므로 삭제하지 않고 따로 저장해 두었다.

대화 내용을 어떤 시간 단위로 요약할 때 가장 성능이 좋은 지 테스트하기 위해 여러 시간 단위로 나눠 요약을 진행하였다. 또한 여러 대화가 섞여 있는 대화를 더 잘 나누기 위해 대화의 시작과 끝을 나타내는 '반갑습니다', '감사합니다'등을 이용해 대화를 나눴다.

또한 모델의 학습시키기 위하여 AI Hub 에서 제공하는 한국어 대화 요약 데이터셋[3]을 이용하였다. 한국어 대화 원문 35 만건, 한 문장으로 요약된 생성 요약문 35 만건으로 원문 데이터에 대하여 정제 작업을 거쳐 대화 주제 분류와 요약문이 생성된 데이터셋이다.

2.2 대화 요약

T5 는 T5 모델 학습을 위해 제작된 텍스트 레이블이 되지 않은 거대한 텍스트 데이터셋 C4 로 사전 훈련된 모델이다. 다양한 다운스트림 작업에 맞게 조정이 될 수 있도록 설계되었다. T5 에서는 모든 NLP 작업의 입력 및 출력이 항상 텍스트 문자열인 통합 Text-To-Text 형식으로 재구성할 것을 제안한다. Text-To-Text 프레임 워크를 이용하면 기계번역, 문서요약 등을 포함한 모든 NLP 작업에서 동일한 모델, 손실 함수 및 하이퍼 파라미터를 사용할 수 있어 효과적인 전이 학습을 제공한다[4].

본 연구에서는 T5 모델 중 '문서요약'에 사전학습된 모델을 활용하였다. fine tuning 된 T5 모델을 이용해 여러 단위로 나뉜 대화 뭉치에 대한 요약을 생성하였다. 결과물을 확인한 후에, 기존에 문서 요약으로 사전 학습되어 있던 T5 모델을 AI HUB 의 대화 데이터셋을 이용해서 대화 요약을 할 수 있도록 다시 fine tuning 을 진행하였다. batch size 는 4, epoch 는 1, learning rate 는 1e-4 로 고정하여 진행하였다. 요약 내용이 원래 내용보다 길다면 원래 내용으로 대체하고 합친 내용의 길이가 30 미만이라면 중요하지 않은 내용으로 판단하고 삭제하였다.

2.3 키워드 추출

KeyBERT 는 BERT 언어 모델을 활용하고 트랜스포머 라이브러리를 사용한다. 많은 키워드 추출 기술의 단점인 통계적 특성에 기초하며 전체 의미적 측면을 고려하지 않는 문제를 해결하는 모델이다.

입력 텍스트는 사전 학습된 BERT 모델을 사용하여 내장된다. 텍스트를 문서의 의미적 측면을 나타내는 고정 크기 벡터로 변환한다. 키워드는 Bag Of Words 기법을 사용하여 동일한 문서에서 추출된다. 각 키워드는 문서를 포함시키는 데 사용된 모델과 동일한 고정 크기 벡터에 내장된다. 그 결과, 키워드와 문서가 동일한 공간에 표시되고 KeyBERT 는 키워드 임베딩과 문서 임베딩 사이의 코사인 유사성을 계산한다. 그 중, 코사인 유사성 점수가 가장 높은 키워드를 추출한다[5]. 원하는 내용을 쉽게 파악하고 찾기 위하여 위키의 태그로 사용될 키워드를 원형의 대화뭉치에서 추출하였다. KeyBERT 를 사용해 관련 키워드를 추출하였다.

3. 결과

3.1 요약모델 결과

활발하게 채팅이 이루어진 오픈채팅방을 기준으로 5 분 단위로 대화를 나눌 경우 가장 유의미한 요약 내용을 보였다고 판단하였다. 또한 대화 데이터에 대해 fine tuning 을 진행하였던 모델이 기존의 문서 데이터 기반 학습된 모델의 결과물보다 가독성이 좋지 않다고 생각하여 기존의 모델을 활용하였다.

결과물의 예시(아래의 Name 에 들어갈 내용)는 다음과 같다.

“오픈 AI 대표는 인간의 유기체적인 한계로 나오는 생산성의 차이가 극대화 될 경우 인간의 노동은 많은 분야에서 점점 더 가치를 잃는다고 말했다.”

“네트워크는 무엇이고 클라우드는 무엇인지 등 컴퓨터공학/컴퓨터과학이 무엇인지 개괄적으로 알아보고 싶은데 어떤 강의/책 등을 찾아봐야 할지 잘 모르겠다.”

표 1. 결과로 저장되는 내용

Text	Name	Tags	Index	URL	Title	Desc

특정 주제로 묶인 채팅내용을 전부 나타낸 것은 Text 에, 이를 요약한 것은 Name 에, 내용에서 언급된 주요 단어는 Tags 에, 대화의 순서를 나타내는 것은 Index 에, 만일 대화 내용에 URL 이 있다면 각각의 URL 주소는 URL 로, URL 의 메타 데이터는 각각 Title 과 Desc 로 나눠서 저장된다.

3.2 채팅방 위키 서비스

위 결과를 사용자가 편하게 볼 수 있도록 [그림 1]과 같이 웹서비스를 구현했다.

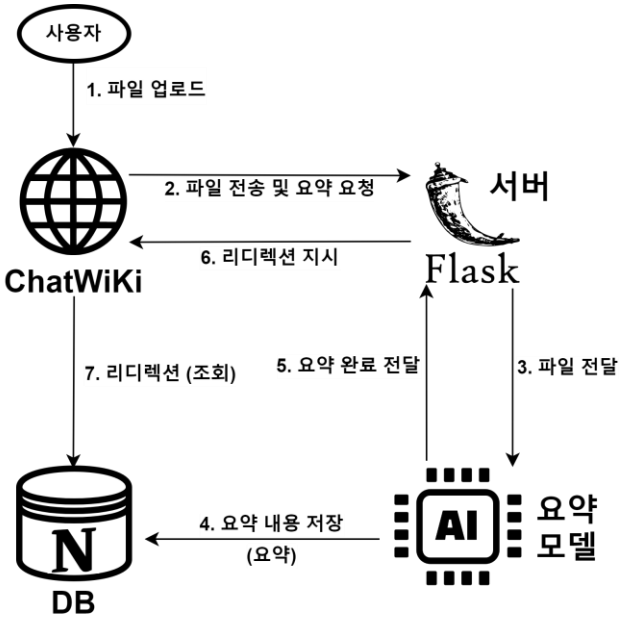

그림 1. 채팅방 위키 흐름도

채팅방위키는 서버, 요약모델, 채팅방위키 홈페이지, 노션[6] 데이터베이스로 구성된다. 서버는 파이썬 웹 프레임워크인 Flask[7]로 개발하였다. 홈페이지에서 사용자가 대화내용이 들어간 텍스트 파일을 올리면 서버로 전송되고 서버는 요약모델에 파일을 전달하고 요약을 요청한다. 요약모델이 결과를 데이터 베이스에 저장하고 완료 알림을 보내면 서버는 웹페이지에 리디렉션을 지시하고 노션 데이터베이 스를 조회하게 된다.

4. 결론과 향후 연구

본 연구에서는 인공지능 모델을 기반으로 대화 데이터셋을 적용하여 요약문을 생성하고 키워드를 추출하였다. 그 결과를 저장하고 웹페이지를 만들어 서비스를 제공하였다. 본 연구를 통해 사람들은 정보를 습득하기 위해 오픈채팅방의 수많은 대화들을 일일이 읽어보지 않아도 된다. 오픈채팅방의 대화 내용이 정리된 '채팅방 위키'를 읽음으로써 본인이 원하는 정보를 쉽게 습득할 수 있다. 이를 통해 사람들은 불필요하게 사용되던 시간을 줄이고 효율적으로 정보를 얻을 수 있다.

본 '채팅방 위키'의 위키 사이트를 이용한 정리된 문서는 접근성이 높아 유용하게 사용될 수 있다. 또한 특정 관심사로 모인 여러 사람이 사용하는 게시판에도 내용을 자동으로 정리할 수 있도록 활용하면 중복된 질문이 나오는 일을 방지할 수 있고, 질문하는 사람도 사전에 나왔던 질문인지 확인할 수 있다. 자동으로 정보를 저장하는 기능에 덧붙여, 그 문서를 사용할 권한이 있는 사람들이 문서를 자유롭게 수정할 수 있도록 하여 위키 형식으로 참여를 유도한다면 보다 유용한 정보를 얻을 수 있다. 또한, 사용자가 열람하는 정보들의 경향을 파악하여 추천 정보를 제공하는 기능도 제공하여 편의성을 높일 수 있다.

그러나 Notion 서비스에 종속적인 '채팅방 위키' 구현으로 인해, 위키의 장점을 완전히 활용하지 못한다는 점이 한계로 남아있다. 요약하는데 시간이 오래 걸려 사용자가 위키 내용을 확인하기까지의 시간이 오래 걸린다는 점도 한계점이다. 또한, 다양한 형태의 채팅방 데이터를 학습시키지 못하여 특정 채팅방에 오버피팅된 문제가 남아있다.

이런 문제점을 보완하기 위해 데이터베이스를 따로 구현하여 결과를 저장할 수 있게 하고 다양한 대화 내용을 학습시켜 일반화 성능이 좋아지도록 할 것이다. 또한 요약에 걸리는 시간을 줄일 수 있도록 모델을 경량화 할 계획이다.

참고 문헌

1. 김탁영, 김지나, 강형원, 김수빈, 강필성. 한국어 문서요약 및 음성합성 통합 프레임워크 구축. 대한산업공학회지. 제48권. 제1호. pp.80-90. 2022.
2. Xiyan Fu, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Changlong Sun, Zhenglu Yang. RepSum: Unsupervised Dialogue Summarization based on Replacement Strategy. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp.6042-6051. 2021.
3. AI Hub. 한국어 대화 요약 데이터. https://aihub.or.kr/aidata/30714 June 25, 2023.
4. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J.. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv preprint arXiv:1910.10683. 2019
5. MaartenGr. KeyBERT. GitHub Repository. https://github.com/MaartenGr/KeyBERT. June 25, 2023.
6. Notion. Notion. https://www.notion.so/ June 25, 2023.
7. Flask. Flask. https://flask.palletsprojects.com/en/2.2.x/ June 25, 2023