유사도 알고리즘을 통한 챗봇 성능 개선에 관한 연구

강태현*, 김휘동*, 김현우*, 박주용*, 고석주**

A Study on Improvement of Chatbot Efficiency Using Similarity Algorithm

Tae-Hyun Kang and Hwi-Dong Kim and Hyun-Woo Kim and Ju-Yong Park and Seok-ju Koh

요 약

자연어 처리(NLP)가 큰 발전을 이룸에 따라 어플리케이션과 플랫폼 내에서 특정 서비스를 제공해주기 위한 챗봇이 활발하게 활용되고 있다. 기존의 거대 언어 모델(LLM)만을 이용하는 챗봇은 파인튜닝에 들어가는 비용과 사용자의 다양한 입력에 대한 일관된 응답에 있어서 어려움을 가진다. 본 논문에서는 LLM과 더불어 정적 답변 알고리즘을 함께 활용한 하이브리드 챗봇을 제안하여 이러한 어려움을 극복해낸다.

Abstract

As natural language processing (NLP) has seen significant advancements, chatbots that provide specific services within applications and platforms are being actively utilized. Chatbots that rely solely on large language model (LLM) faces challenges in terms of the cost of fine-tuning and providing consistent responses to diverse user inputs. In this paper, we propose a hybrid chatbot that leverages both LLM and a static response algorithm to overcome these challenges. By combining LLM with a static response system, the hybrid chatbot is able to provide more consistent and cost-effective solutions compared to relying on LLM alone.

Key words

large language model, chatbot, hybrid chatbot, text embedding, threshold, smilarity

Ⅰ. 서 론

인공지능 기술은 현대 사회에서 빠르게 발전하고 있으며, 특히 챗봇(Chatbot) 기술은 업무 및 서비스 분야에서 활발히 활용되고 있다. 다양한 어플리케이션과 플랫폼에서 챗봇은 채팅 인터페이스를 통해 정보를 주고받거나 서비스를 제공하면서 고객 응대, 자동화된 업무 처리, 정보 제공 등의 목적으로 이용되고 있다. 해당 어플리케이션이나 플랫폼에서 특정 서비스와 정보를 제공하기 위해서는 챗봇과 연동할 대규모 언어 모델(LLM)을 파인튜닝하는 과정이 꼭 필요하다. 파인튜닝은 전이학습(transfer learning)으로

*경북대학교 컴퓨터학부, 전자공학부
kang.1107.th@gmail.com, mat2408@naver.com, khw4420@gmail.com, claudeopk@gmail.com
**경북대학교 컴퓨터학부 교수(교신저자)
sjkoh@knu.ac.kr

모델을 특정 도메인에 적응시키기 위해 해당 도메인의 데이터에 대해 추가로 학습하여 세밀하게 조정하는 과정이다. 이를 통해 챗봇은 LLM이 가지는 언어적 능력에 더하여 특정 어플리케이션과 플랫폼에 특화된 서비스 능력을 갖출 수 있게 된다.

하지만, 파인튜닝된 LLM만을 통한 챗봇 구현에는 몇가지 문제점이 따른다. 첫 번째는 많은 비용이 발생한다는 점이다. 챗봇과 연동할 LLM으로 보통 OPEN AI같은 거대 기업이 제공하는 GPT 모델을 많이 사용하는데 이때 토큰 당 비용이 파인튜닝할 때 뿐만 아니라 사용할 때도 발생한다. 두 번째는 원하는 답변을 제공하기 어렵다는 점이다. 사용자로부터 들어오는 질문의 의미는 비슷하더라도 표현이 조금 달라짐에따라 파인튜닝한 원하는 답변이 제공될 수도 있고, 아예 의도하지 않은 답변이 제공될 수 있다. 아주 많은 학습 데이터로 다양한 표현을 모두 학습데이터로 사용하면 해결할 수 있겠지만 이는 결국 비용을 증가시키게 되어 첫 번째 문제점으로 귀결된다.

본 논문에서는 기존 LLM 기반 챗봇의 문제점을 보완하기 위해 LLM과 정적 답변(Static Response)을 모두 이용하는 하이브리드 챗봇(Hybrid Chatbot) 방식을 제안한다.

II. 하이브리드 챗봇

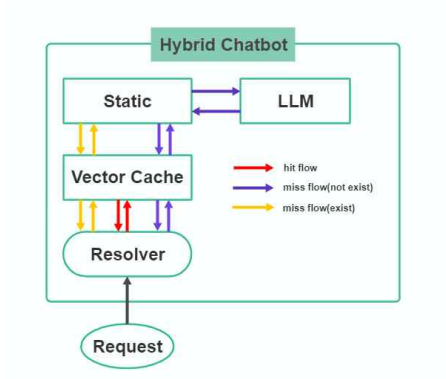

그림 1. 하이브리드 챗봇 구성도

사용자 입력을 바로 LLM에 전달하는 기존의 챗봇과는 달리 하이브리드 챗봇은 사용자의 입력을 벡터로 임베딩하여 유사도를 측정한 후 일정 기준을 만족하면 벡터 데이터베이스 내의 정적인 답변을 제공하고 그렇지 않을 경우에만 LLM에 전달하여 답변을 제공한다. 이를 통해 어플리케이션이나 플랫폼 내의 정보 같은 특정 작업 영역의 입력에 대해서는 의도한 답변 그대로를 제공할 수 있고 이외의 보편적인 입력에 대해서는 기존의 LLM이 동적으로 답변을 할 수 있게 된다. 전체 알고리즘은 다음과 같다.

1. 빈도가 높을 것 같은 사용자의 질문을 예측하여 미리 답변과 함께 텍스트 임베딩하여 벡터 데이터베이스를 준비.
2. 벡터 데이터베이스의 모든 질문 벡터간의 유사도를 계산한 후 그 값이 $T_{clustering}$ 이상인 유사한 특성의 질문 벡터들끼리 답변 벡터와 함께 군집화 진행
3. 사용자의 입력을 텍스트 임베딩 모델을 이용하여 벡터화
4. 3에서 얻은 임베딩 벡터와 벡터 데이터베이스 안의 각 군집들의 대표 임베딩 벡터들간의 유사도 측정
5. 4에서 측정한 유사도가 최대인 군집으로 가서 3에서 얻은 임베딩 벡터와 해당 군집의 벡터들 간의 유사도 측정
6. 5에서 얻은 유사도의 최댓값(Max Similarity)에 따라 case 분류
   (1) $Similarity_{max} \geq T_{dependent}$
   사용자 입력 벡터를 군집 내의 유사도가 최대인 벡터와 종속시켜서 준비된 답변 제공
   (2) $T_{update} \leq Similarity_{max} \leq T_{dependent}$
   사용자 입력 벡터를 군집 내에 새로 추가할 질문 벡터로 판단하여 따로 저장해두고 LLM에 입력을 전달하여 답변 제공
   (3) $Similarity_{max} < T_{update}$
   관련없는 질문으로 판단하여 사용자 입력 벡터를 벡터 데이터베이스에 추가하지 않고 LLM에 입력을 전달하여 답변을 제공
7. 6의 (2)의 경우에서 저장한 입력 벡터에 대해서 정적인 답변을 준비하여 벡터 데이터베이스를 업데이트

2.1 텍스트 임베딩 모델

텍스트 임베딩(text embedding)은 텍스트를 머신러닝이나 딥러닝 모델이 처리할 수 있는 고차원의 숫자 벡터로 변환하는 것을 말한다. 자연어처리(NLP) 분야의 급속한 발전 속에서 텍스트 임베딩 모델은 Word2vec 같은 단순한 단어 임베딩을 넘어 문맥 기반 단어 임베딩(Contextualized Word Embedding)으로 단어가 사용된 문맥까지 반영해서 임베딩하는 모델이 연구되었다. 최근 큰 주목을 받고 있는 BERT와 같은 최신 임베딩 모델들은 특정 단어를 예측하기 위해 이전의 단어와 함께 이후의 단어까지도 활용하는 양방향 Transformer 인코더를 사용하여 문장의 문맥과 단어의 의미를 매우 잘 추론할 수 있다.

하이브리드 챗봇의 텍스트 임베딩 모델로는 허깅페이스에서 제공하는 Pretraied ALL-MPNet-Base-V2 모델을 사용하였다. MPNet은 OpenAI에서 개발한 모델로 BERT와 같이 양방향 Transformer 아키텍처를 기반으로 하지만 Massively Parallel이라는 특징이 추가되어 학습 및 처리 과정에서 대규모의 병렬처리를 통해 효율성을 높였다.

2.2 벡터 데이터베이스

기존의 관계형 데이터베이스는 데이터의 크기가 매우 크고 고차원인 비정형 데이터일 경우 저장할 때 효율적이지 않고 각 개체 간의 의미적인 유사도를 측정하기 어렵다. 이에 반해 최근 이미지 비전과 NLP의 발전으로 수요가 늘어난 벡터 데이터베이스는 이런 비정형 데이터를 임베딩하여 저장하는데 최적화된 데이터베이스로, 이를 통해 효과적으로 데이터를 저장하고 유사도를 기반으로 검색을 할 수 있다.

빠르고 효율적으로 유사도 검색을 수행하기 위해 하이브리드 챗봇의 벡터 데이터베이스는 클러스터링(Clustering)을 통해 데이터를 그룹화하고, 그룹 간의 대푯값을 이용하였다.

2.3 유사도 측정 방법

유사도는 벡터 차원인 데이터간의 거리로 데이터가 얼마나 비슷한지를 측정한 것이다. 일반적으로 0~1 값을 가지며 이 값이 클수록 유사도가 높은 것이다. 유사도 측정 방법은 아래와 같은 방법들이 있다.

2.3.1 자카드 유사도(Jaccard Similarity)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

자카드 유사도는 두 개의 집합에서 두 집합의 합집합 중 교집합의 비율이다.

2.3.2 유클리드 유사도(Euclidean Distance)

$$U(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (2)$$

유클리드 유사도는 다차원 공간에서 계산한 두 개의 점 사이의 거리이다.

2.3.3 맨하탄 유사도(Manhatten Distance)

$$M(A,B) = \sum_{i=1}^{n}|a_i - b_i| \quad (3)$$

맨하탄 유사도는 L1 distance로, 다차원 공간의 두 점 a,b에 대하여 각 차원에서 두 점 사이의 차를 더한 것이다.

2.3.4 코사인 유사도(Cosine Similarity)

$$C(A,B) = \frac{A \cdot B}{\|A\| \, \|B\|} \quad (4)$$

코사인 유사도는 두 벡터 A, B의 코사인 값을 계산한 것이다.

다른 유사도 측정 방법들은 단순히 두 벡터간의 거리 측면에서 계산하는 반면, 코사인 유사도는 두 벡터간의 각도를 구하는 것이기에 방향성의 개념이 더해지므로 문장의 의미가 동일하면 표현이 달라지더라도 유사도를 잘 측정한다는 장점이 있다.

2.4 유사도 Threshold

전체 알고리즘에서 벡터 데이터베이스 내의 군집화 과정과 새로운 임베딩 벡터의 종속 또는 업데이트 여부를 판단하는 과정에서 기준이되는 $T_{clustering}$, $T_{dependent}$, T_{update} 값을 우리는 실험적으로 결정하였다. 질문 벡터간의 유사도가 0.93이상일 경우 해당 벡터들은 비슷한 특성을 가진 질문 벡터들로서 군집화가 가능하다고 판단할 수 있었고, 유사도가 0.95이

상일 경우에는 완전히 동일한 의미를 가지는 벡터이며 유사도가 0.8미만일 경우에는 아예 관련이 없는 질문 벡터라고 판단할 수 있었다.

$T_{clustering}$ = 0.93,

$T_{dependent}$ = 0.95,

T_{update} = 0.8

2.5 비교

미리 확보한 질문들에 대한 정적인 데이터를 군집화하여 새로운 질문들에 효율적으로 대응시켜 빠른 답변과 절감을 기대할 수 있는 알고리즘을 작성하여 그 효과를 측정하는 실험을 진행하였으며 정확도의 판단 기준은 다음과 같다.

(1) 정적인 데이터로써 이미 확보된 질문에 대한 올바른 정답을 반환한 경우.

(2) 정적인 데이터로써 미리 준비되지 않은 경우에 대해 LLM에 전가한 경우.

표본은 153개의 선별된 질문 데이터로 구성되었으며 표본공간 내 일부 대표성을 지닌 질문 데이터들을 다시 선별하여 실험을 진행한 결과 약 41.6%의 정확도를 보였고 표본공간 내 모든 질문 데이터에 대한 실험을 진행한 결과 37.9% (표준편차±7.69%)의 정확도를 보였다.

III. 결 론

본 논문에서는 정적답변과 LLM을 모두 이용하는 새로운 챗봇인 하이브리드 챗봇을 제안하였다. 기존의 LLM 기반 챗봇이 필연적으로 수행해야하는 파인튜닝 때문에 겪는 빅데이터 마련의 어려움과 학습 비용의 문제를 해결하였고 군집화된 벡터 데이터베이스를 이용하여 오버헤드를 최소화하며 벡터들간의 유사도를 측정할 수 있게 함으로써 표현 방식이 다른 동일한 의미의 질문에 대해 의도하지 않은 답변이 제공되는 문제 또한 해결할 수 있었다. 인공지능의 발전과 함께 챗봇의 활용도가 급격히 증가하는 시대에, 하이브리드 챗봇은 새로운 접근법으로 주목받을 것으로 기대된다.

ACKNOWLEDGEMENT

“이 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 통해 수행되었음“(2021-0-01082)

참 고 문 헌

[1] https://openai.com/chatgpt

[2] Kostka, Ilka, and Rachel Toncelli. “Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations.” TESL-EJ 27.3 (2023).

[3] Howard, Jeremy, and Sebastian Ruder. “Universal language model fine-tuning for text classification.” arXiv preprint arXiv:1801.06146 (2018).

[4] Haman, Michael, and Milan Školník. “Using ChatGPT to conduct a literature review.” Accountability in Research (2023): 1-3.

[5] Jeddi, Ahmadreza, Mohammad Javad Shafiee, and Alexander Wong. “A simple fine-tuning is all you need: Towards robust deep learning via adversarial fine-tuning.” arXiv preprint arXiv:2012.13628 (2020).

[6] Taipalus, Toni. “Vector database management systems: Fundamental concepts, use-cases, and current challenges.” Cognitive Systems Research (2024): 101216.

[7] Taipalus, Toni. “Vector database management systems: Fundamental concepts, use-cases, and current challenges.” arXiv preprint arXiv:2309.11322 (2023).

[8] Han, Yikun, Chunjiang Liu, and Pengfei Wang. “A comprehensive survey on vector database: Storage and retrieval technique, challenge.” arXiv preprint arXiv: 2310.11703 (2023).

[9] Han, Yikun, Chunjiang Liu, and Pengfei Wang. “A comprehensive survey on vector database: Storage and retrieval technique, challenge.” arXiv preprint arXiv: 2310.11703 (2023).