Journal of The Korea Convergence Society
Vol. 12. No. 12, pp. 93-106, 2021

ISSN 2233-4890 / e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2021.12.12.093

모델, 데이터, 대화 관점에서의 BlendorBot 2.0 오류 분석 연구

이정섭[1], 손수현[2], 심미단[1,3] 김유진[1,4], 박찬준[2], 소아람[5], 박정배[5], 임희석[6*]
[1]고려대학교 Human-inspired AI 연구소 연구원, [2]고려대학교 컴퓨터학과 석박사통합과정
[3]경희대학교 소프트웨어융합학과 학생, [4]이화여대 과학교육과 학생, [5]고려대학교 Human-inspired AI 연구소 연구교수
[6]고려대학교 컴퓨터학과 교수

Empirical study on BlenderBot 2.0's errors analysis in terms of model, data and dialogue

Jungseob Lee[1], Suhyune Son[2], Midan Shim[1,3], Yujin Kim[1,4], Chanjun Park[2]
Aram So[5], Jeongbae Park[5], Heuiseok Lim[6*]
[1]Researcher, Human-inspired Computing Research Center, Korea University
[2]Master & Ph. D. Combined Student, Department of Computer Science and Engineering, Korea University
[3]Student, Department of Software Convergence, Kyung Hee University
[4]Student, Department of Science Education, Ewha Womans University
[5]Research Professor, Human-inspired Computing Research Center, Korea University
[6]Professor, Department of Computer Science and Engineering, Korea University

요 약 블렌더봇 2.0 대화모델은 인터넷 검색 모듈과 멀티 세션의 도입을 통해 실시간 정보를 반영하고, 사용자에 대한 정보를 장기적으로 기억할 수 있도록 함으로써 오픈 도메인 챗봇을 대표하는 대화모델로 평가받고 있다. 그럼에도 불구하고 해당 모델은 아직 개선점이 많이 존재한다. 이에 본 논문은 블렌더봇 2.0의 여러 가지 한계점 및 오류들을 모델, 데이터, 대화의 세 가지 관점으로 분석하였다. 모델 관점에서 검색엔진의 구조적 문제점, 서비스 시 모델 응답 지연시간에 대한 오류를 주로 분석하였다. 데이터 관점에서 크라우드 소싱 과정에서 워커에게 제공된 가이드라인이 명확하지 않았으며, 수집된 데이터의 증오 언설을 정제하고 인터넷 기반의 정보가 정확한지 검증하는 과정이 부족한 오류를 지적하였다. 마지막으로, 대화 관점에서 모델과 대화하는 과정에서 발견한 아홉 가지 유형의 문제점을 면밀히 분석하였고 이에 대한 원인을 분석하였다. 더 나아가 각 관점에 대하여 실질적인 개선방안을 제안하였으며 오픈 도메인 챗봇이 나아가야 할 방향성에 대한 분석을 진행하였다.

주제어 : 대화 시스템, 오픈 도메인, 블렌더봇, 검색 모델, 언어융합

Abstract Blenderbot 2.0 is a dialogue model representing open domain chatbots by reflecting real-time information and remembering user information for a long time through an internet search module and multi-session. Nevertheless, the model still has many improvements. Therefore, this paper analyzes the limitations and errors of BlenderBot 2.0 from three perspectives: model, data, and dialogue. From the data point of view, we point out errors that the guidelines provided to workers during the crowdsourcing process were not clear, and the process of refining hate speech in the collected data and verifying the accuracy of internet-based information was lacking. Finally, from the viewpoint of dialogue, nine types of problems found during conversation and their causes are thoroughly analyzed. Furthermore, practical improvement methods are proposed for each point of view, and we discuss several potential future research directions.

Key Words : Dialouge System, Open domain, BlenderBot, Retrieval Model, Language Convergence

*This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425) and supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).
*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received November 17, 2021 Revised November 30, 2021
Accepted December 20, 2021 Published December 28, 2021

1. 서론

최근 대화 시스템과 관련한 연구들이 다수 이루어지고 있으며, 특히 오픈 도메인(Open-domain) 대화 시스템 연구에서는 모델이 자연스러운 대화 흐름, 공감 더 나아가 지식 공유 등의 능력을 갖출 수 있도록 모델링하는 다양한 방향의 연구들이 발표되고 있다.

오픈 도메인 챗봇(Open-domain chatbot)은 모든 주제에 관해 대화할 수 있는 에이전트를 개발하는 연구 분야로 도전적인 자연어처리 분야이다. 이는 자연어 이해에서 사람 지능의 중요한 요소이고 여러 산업적 서비스에서 활용성이 높다. 해당 연구의 궁극적인 목표는 누구에게나 매력적인 응답을 생성할 수 있는 사람과 비슷한(Human-like) 즉, 휴먼 AI이다.

이러한 이유로 정확한 문서 지식을 바탕으로 응답을 생성할 수 있는 대화 모델링 기법을 제안한 [1] 및 [2] 등의 연구와, 대화모델이 사용자의 관심사를 기반으로 공감력 있는 응답을 생성하기 위한 대화 모델링 기법에 관한 연구를 제안한 [3] 및 [4] 등의 연구가 진행되었다. 이러한 연구들은 대화모델에 큰 성능 향상을 가져다주었으며, 모델은 지식을 통해 응답을 생성할 수 있었고, 상대방의 관심사가 무엇인지 파악하여 응답을 생성할 수 있었다. 그럼에도 불구하고, 오픈 도메인 대화모델이 사람다워지기까지 아직 여러 개선 사항들이 남아있다.

첫째로 대화 데이터를 사용하여 학습한 지식 문서 기반 모델은 최신의 정보를 반영하기 어렵다. 일반적으로 대화모델은 학습 당시 고정된 대규모의 학습 데이터셋을 통해 학습되고 학습 후의 고정된 파라미터를 사용한다. 따라서 데이터셋 구축 당시 고정된 지식만을 습득할 수 있고 후에 추가되거나 변화된 지식을 배울 수 없다. 기존 대화모델은 변화하는 실세계의 지식을 모델에 반영하기 위해 지속적인 데이터셋을 구축하고 학습해야 하는데, 이러한 작업은 많은 휴먼 리소스와 컴퓨팅 리소스가 요구되어 많은 어려움이 뒤따른다. 실시간으로 변화하는 지식을 반영하는 대화모델은 곧 실세계의 사람에 대한 모델링이며 이는 사람과 더 깊은 대화를 진행할 수 있음을 의미하고 궁극적인 목표인 휴먼 AI로서의 필수 요소이다.

둘째로 사람은 상대방의 관심사를 장기적으로 기억할 수 있지만, 대화모델은 비교적 짧은 턴의 대화만 기억한다. 예를 들어, 사용자와 모델 간의 대화에서 사용자가 취미에 대해 발화문을 작성하면 짧은 5-6턴(10-12 발화 문장) 대해서는 사용자의 취미를 기억할 수 있다. 하지만, 10턴(20 발화 문장) 이상 대화를 진행할 경우 사용자의 취미를 기억하지 못한다는 문제점이 있다. 즉, 턴이 누적되는 멀티턴(Multi-turn) 대화에서 이전에 대화했던 정보를 성공적으로 반영하지 못한다.

해당 문제로 인해 사용자는 동일한 정보를 반복적으로 발화해야 한다. 이는 대화 모델에서 개선되어야 할 중요한 연구 과제 중 하나로서, 오픈 도메인 챗봇에서 중요하게 평가되는 대화의 다양성을 해친다. 또한, 매력적인 주제에 대한 언급을 방해하며, 사용자에게 지루함을 유발한다.

블렌더봇 2.0은 위에서 언급한 문제점들을 인터넷 검색(Internet search)과 대화 사이에 시간적 공백이 존재하는 멀티 세션(Multi-session)을 도입하여 개선하였다. 변화하는 지식을 습득할 수 있도록 인터넷 검색에 접근할 수 있는 인터넷 검색 쿼리 생성으로 적절한 정보를 검색하여 응답을 생성할 수 있었고, 관심사 유지 및 대화 기록의 보다 효과적인 요약을 위한 멀티 세션을 구성하였다. 즉, 멀티 세션을 통해 비교적 긴 대화 기록과 관심사 두 가지 모두 반영하여 장기적인 기억을 갖출 수 있었다.

그러나 블렌더봇 2.0에서도 인터넷에 존재하는 정확하지 않은 정보를 바탕으로 답변을 생성하는 문제와 잘못된 인터넷 검색 등 완벽한 오픈 도메인 대화모델 발전으로 나아가기까지 해결해야 할 문제점이 많다.

본 논문은 블렌더봇 2.0의 문제점을 모델, 데이터 그리고 대화의 세 가지 관점으로 상호 분석하고 모델의 개선방안을 제안한다. 기존의 오픈 도메인 챗봇에 관한 연구들은 대화모델의 성능에 초점을 맞춰 다양한 모델링 기법을 연구하지만, 본 논문은 기존의 모델링 흐름을 벗어나 다양한 시각으로 현재 대화모델의 오류를 분석하고 이에 따른 개선방안을 바탕으로 추후 오픈 도메인 챗봇에 대한 다양한 연구 방향을 제안한다.

본 논문의 구성은 다음과 같다. 첫째로, 관련 연구 및 배경지식에 관해 서술한다. 해당 섹션에서는 연구에 대한 이해를 위해 오픈 도메인 챗봇의 개념과 블렌더봇 2.0의 초기 모델인 블렌더봇 1.0에 대해 간단히 서술하고 블렌더봇 2.0에 대한 모델 구조 및 모델 학습의 전반적인 내용을 서술한다. 둘째로, 모델, 데이터, 대화의 관점으로 블렌더봇 2.0의 문제점과 한계를 면밀히 조사하고 문제의 원인을 파악하여 서술한다. 셋째로, 각 관점으로 파악한 문제점을 해결할 개선방안을 제안한다. 본 연구로 오픈

도메인 챗봇에 대한 문제점과 한계점을 파악하고 오픈 도메인 챗봇의 목표인 휴먼 AI에 다가갈 수 있는 다양한 연구가 진행되어야 함을 제안한다.

2. 관련연구 및 배경지식

2.1 오픈 도메인 챗봇(Open-domain chatbot)

챗봇이란 사람들 사이의 비정형화된 대화를 모방하여 확장된 대화를 진행할 수 있는 시스템을 말한다[5]. 간단히 말하면, 챗봇은 상대방이 메세지 M을 보냈을 때 메세지 M과 대화 기록 C에 기반하여 응답 R을 생성하는 시스템이다. 이러한 시스템의 궁극적인 목표는 대화를 수행하는 과정에서 사람과 유사한 응답을 생성하는 것이다[6]. 특정 서비스를 제공하거나 정해진 주제에 대한 대화만 수행할 수 있는 목적지향 대화 시스템(Task-oriented dialogue system)과 다르게, 오픈 도메인 챗봇은 일상 대화를 목적으로 하여 다양한 주제에 대한 응답을 생성할 수 있어야 하기 때문에 개발 과정이 더욱 복잡하다.

1960년대의 간단한 대화형 프로그램 Eliza[7]를 시작으로, Apple의 Siri[1], Microsoft의 XiaoIce(또는 Little Bing)[8][2], [3]심심이, 이루다[4] 등이 국내외에서 서비스 되었던 오픈 도메인 챗봇의 대표적인 예시라고 할 수 있다.

최근에는 사전학습 모델을 활용한 오픈 도메인 챗봇 연구가 활발하게 이어졌고, 이는 인코더-디코더(Encoder-Decoder) 기반 방법이나 생성적 적대 신경망(Generative Adversarial Network, GAN)을 활용하는 다른 신경망 기반의 대화 생성 방법과 비교했을 때 일반적으로 가장 높은 성능을 보인다[9]. 대표적인 예시로 GPT-2[10]에 기반한 DialoGPT[11], 구글의 Meena[12], PLATO-2[13], 그리고 블렌더봇 1.0[14]이 존재한다.

2.2 블렌더봇 1.0

블렌더봇 1.0은 2020년에 Facebook AI Research(FAIR)에서 개발한 오픈 도메인 챗봇으로, 공감, 지식 등의 다양한 대화 스킬의 결합을 처음으로 시도한 챗봇이다. 블렌더봇 1.0 이전의 모델들은 파라미터 수를 늘림으로써 성능이 좋은 챗봇을 만들고자 했다면, 블렌더봇 1.0은 대용량(Large-scale)의 파라미터뿐만 아니라 사람의 대화 스킬을 잘 학습할 수 있는 데이터셋을 활용하여 모델의 성능을 개선하고자 했다.

블렌더봇 1.0 모델은 폴리 인코더(Poly Encoder)[15]를 활용하여 인코딩한 후 Retrieve and Refine(RetNRef) 방법으로 응답을 생성한다. Fig. 1은 해당 모델의 응답 생성 과정을 도식화한 그림이다. 사람들 사이에서 이루어지는 자연스러운 대화의 특성을 잘 학습할 수 있도록 소셜 미디어의 글, 댓글로 이루어진 Pushshift.io Reddit 데이터셋[16]을 사용하여 사전학습을 진행하였다. 또한, 모델이 페르소나(Persona)를 유지하면서도 매력적인 대화 주제를 제공하고, 지식을 활용하며, 공감을 잘 하는 챗봇으로 만들기 위하여 ConvAI2[17], Wizard of Wikipedia(WoW)[1], Empathetic Dialogues(ED)[18] 데이터셋을 모두 활용했으며, 각각의 대화 스킬을 적절하게 혼합하여 응답을 생성할 수 있도록 Blended Skill Talk(BST)[19] 데이터셋으로 미세 조정(Fine-tuning)했다. 90M, 2.7B, 9.4B 파라미터 크기의 모델을 구축하고 모델과 데이터셋을 오픈소스로 공개하였으며[5], 90M 모델의 경우 기존의 챗봇 모델보다 3.6배 많은 파라미터를 가진다.

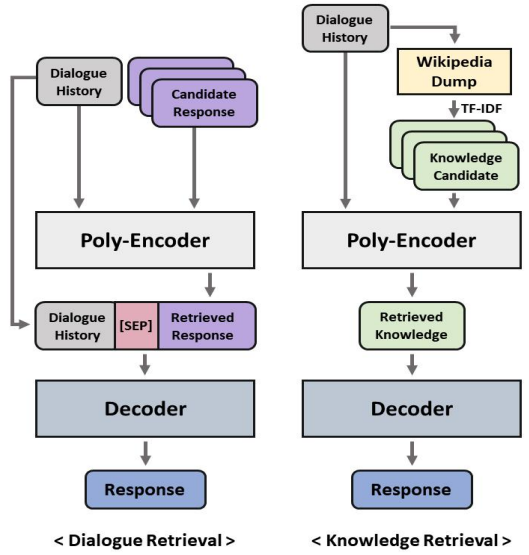

Fig. 1. Architecture of BlenderBot 1.0

그 결과, 구글에서 개발한 챗봇 Meena와 비교하여

1) https://www.apple.com/kr/siri/

2) https://www.xiaoice.com/

3) https://www.simsimi.com/

4) https://luda.ai/

5) https://parl.ai/projects/recipes/

휴먼 평가(Human evaluation)을 진행했을 때 더 많은 사람이 블렌더봇 1.0과의 대화를 더 매력적으로 느꼈으며 더 사람 같은 대화를 생성한다고 답했다. 하지만 사용자의 발언과 비슷한 응답을 반복하여 생성하거나, 앞선 대화의 내용을 기억하지 못하거나, 틀린 정보를 사실인 것처럼 이야기하는 등의 한계점이 여전히 존재한다.

2.3 블렌더봇 2.0

이전의 대화와 모순되는 발언을 하거나, 최신의 정보를 반영하지 않는 대화를 생성하는 것이 기존 모델인 블렌더봇 1.0이나 GPT-3[20]의 대표적인 문제점으로 지적되어 왔다. 블렌더봇 1.0을 개선한 블렌더봇 2.0[21,22] 모델이 2021년 7월에 공개되었다. 블렌더봇 2.0은 여러 세션에 걸쳐 일관된 대화를 수행하며, 인터넷의 있는 정보를 검색해 최신 정보에 관한 대화도 가능하다. 이는 기존 대화모델과 가장 큰 차이점이다.

이를 위해 블렌더봇 2.0은 블렌더봇 1.0의 장점을 그대로 가져가면서, 추가적으로 Wizard Of the Internet (WizInt) 데이터셋[22]과 Multi-Session Chat(MSC) 데이터셋[21]을 크라우드 소싱으로 구축하여 학습시켰다.

WizInt 데이터셋은 인터넷 검색 결과를 이용한 답변 생성 과정을 학습하기 위해 사용되었다. 데이터셋을 구축하는 과정에서 크라우드 워커는 대화의 순서가 돌아올 때마다 인터넷 검색 시스템에 접근할 수 있었고, 검색 결과로 나타나는 상위 K개의 문서를 바탕으로 응답에 활용할 문서를 선택한 후 이를 활용한 응답을 생성했다. 원하는 정보가 없다면 검색 결과를 무시하는 것도 허용되었다.

MSC 데이터셋은 여러 개의 세션으로 이루어진 대화에서 새롭게 알게 된 지식을 장기 기억에 저장하고 이를 활용한 응답을 생성하는 과정을 학습하기 위해 사용되었다. 해당 데이터셋을 구축하기 위해 크라우드 워커는 각 세션 사이에 몇 시간 또는 며칠, 몇 주의 시간 간격이 있는 것처럼 대화한다.

그 결과 WizInt 데이터셋은 주어진 페르소나를 해치지 않으면서 인터넷 검색 결과가 반영된 대화를, MSC 데이터셋은 대화 데이터셋에 기존에 없던 세션이라는 개념을 도입하여 긴 시간 간격이 있어도 자연스럽게 이어지는 대화를 생성 가능하게 하였다.

블렌더봇 2.0 모델을 살펴보면 이전 버전인 블렌더봇 1.0에 비해 크게 두 가지 구조가 추가되었다. 대화의 문맥을 입력으로 하여 관련된 검색 쿼리를 생성하는 인코더-디코더 구조는 두 구조에 공통적으로 활용되며, 이렇게 생성된 쿼리로 인터넷에서 관련 정보를 찾아 응답을 생성하는 부분과 대화와 관련된 정보를 사람마다 별도로 장기 기억에 저장하고 이를 활용한 응답을 생성하는 부분으로 나누어진다. Fig. 2 는 해당 모델의 응답 생성 과정을 도식화한 그림이다.

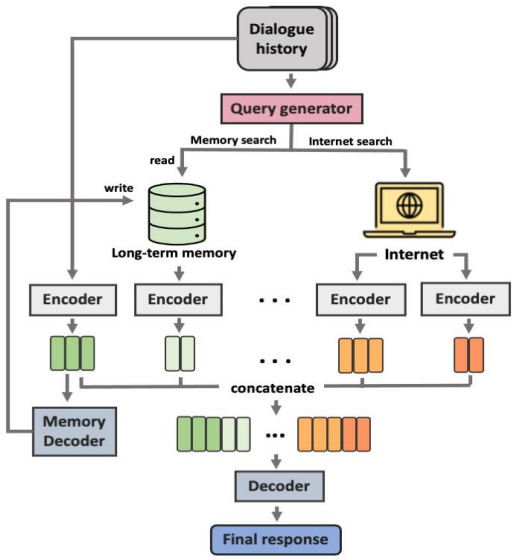

Fig. 2. Architecture of BlenderBot 2.0

인터넷의 정보를 이용한 응답을 생성하기 위해서는 쿼리 생성기의 결과를 통해 생성한 키워드를 이용하여 대화 내용과 관련된 상위 K개의 문서를 검색엔진을 통해 검색한다. 이전 세션의 정보를 기억하고, 이를 활용한 대화를 하기 위해서는 생성 요약기(Abstractive summarizer)로 대화 내용을 요약한 후, 요약된 내용이 기존의 메모리에 없는 새로운 정보를 포함하고 있으면 장기 기억 메모리에 기록하는 방식으로 과거 대화내용을 저장한다. 응답 생성 시에는 메모리에 접근하여 대화 문맥과 연관된 상위 N개의 문서를 검색한다. 인터넷과 메모리 네트워크에서 얻은 문서들은 개별적으로 인코딩된다. 각 인코딩 결과와 대화문맥 인코딩 결과를 연결한 후, 디코더에 입력으로 넣어 최종 응답을 생성한다.

이에 대한 휴먼 평가를 진행한 결과 평가자들은 블렌더봇 2.0을 블렌더봇 1.0보다 더 매력적으로 느꼈으며,

시간 간격을 둔 채 지속되는 대화에서도 앞선 대화의 내용을 기억하고 이를 활용하여 대화를 생성하는 것으로 나타났다. 또한, 지식 사용 능력을 테스트한 결과 인터넷 검색 기능이 환각 지식(Hallucinate knowledge)을 9.1%에서 3.0%로 줄이고 사실과 일치하는 발언을 더 자주 생성하는 데 도움을 주는 것을 발견했다. 해당 모델 역시 연구자들이 쉽게 재현할 수 있도록 코드와 함께 데이터셋이 공개되어 있다[6].

3. 모델 관점에서의 블렌더봇 2.0의 문제점

블렌더봇 2.0은 실시간 인터넷 검색 결과를 사용하여 최신 정보를 사용자에게 제공하고, 과거 대화 내용을 기억하는 대화를 가능하게 함으로써 기존의 오픈 도메인 챗봇의 한계를 극복하였다. 하지만 모델 관점에서 몇 가지 문제점을 가진다. 다음은 모듈 및 모델 구조를 중심으로 블렌더봇 2.0의 문제점을 제시한다.

3.1 인터넷 정보의 정확성 문제

블렌더봇 2.0은 Bing Search[7] 엔진을 사용하여 인터넷상의 정보를 검색한다. 하지만 블렌더봇 2.0에서는 사용하는 검색엔진이 현존하는 타 검색엔진과 비교하였을 때 최적의 검색엔진이라는 것을 증명하지 않았다. 더불어, 인터넷 검색 결과인 K개의 문서에서 쿼리에 대한 내용이 문서별로 상이할 경우, 어떤 정보를 우선적으로 사용해야 하는지에 대한 기준이 명시되어 있지 않다.

또한, 블렌더봇 2.0은 검색엔진을 통해 응답 생성에 필요한 정보를 가져오는 과정에서 검색한 정보에 대한 정확성을 검증하지 않는다. 블렌더봇 2.0에서 안전성 분류기(Safety classifier)로 생성한 응답을 정제하고 있지만, 이는 정보의 정확성이 아닌 응답에 편견이나 비방 등이 포함되어 있는지 확인한다. 검색한 정보를 검증하지 않고 사용하면 잘못된 정보를 포함하는 응답을 생성할 수 있으며, "과거 학습 시점의 정보가 아닌 현재의 정확한 정보를 반영하고자 한다"는 원래의 목적과 상반되는 결과를 야기할 수 있다.

3.2 서비스 속도 및 환경의 문제

블렌더봇 2.0 모델은 서비스 측면에서도 여러 한계를 가지고 있다. 챗봇 시스템에서는 메시지를 받은 직후 응답을 제공하는 데 걸리는 시간인 응답지연시간(Response latency)이 중요하다[23]. 하지만 블렌더봇 2.0은 인터넷 검색, 메모리 네트워크 사용으로 인한 응답지연시간에 대한 분석이 기술되지 않았다. 또한 블렌더봇 2.0은 모델 파라미터 개수가 많아, 이를 챗봇 서비스에 적용하기 위해서는 많은 양의 컴퓨팅 자원이 필요하다. 하지만, 이는 일반 기업이나 개인이 확보하는 것에 한계가 있다. 대화모델 및 챗봇의 목표가 사용자에게 원하는 서비스를 편리하게 제공하는 것임을 고려하였을 때, 성능이 우수한 모델이라도 크기가 지나치게 크면 상용화에 어려움이 존재한다.

4. 데이터 관점에서의 블렌더봇 2.0의 문제점

블렌더봇 2.0은 블렌더봇 1.0을 사전학습 모델로 사용하였다. Table 1에서와 같이 MSC 데이터셋을 통해 멀티 세션 대화가 가능하게, WizInt 데이터셋을 통해 대화에 인터넷 검색 결과를 반영 가능하게 함으로써 기존 대화 모델을 보완하였다. 그러나 사용한 데이터의 측면에서 크게 3가지의 문제점이 존재한다.

Table 1. Datasets used in BlenderBot 1.0 and 2.0

	Task	Dataset
BlenderBot 1.0	Pre-train	Reddit
	Fine-tuning	Blended Skill Talk (BST)
BlenderBot 2.0	Fine-tuning	Multi Session Chat (MSC)
		Wizard of the Internet (WizInt)

4.1 데이터 수집에서 통일된 기준 부재의 문제

첫 번째로 크라우드 소싱 과정에서 데이터 수집에 통일된 기준이 부재하다. MSC 데이터셋의 경우, 세션을 나누는 기준이 정해져 있지 않다. WizInt 데이터셋의 경우 인터넷 검색의 활용 여부의 기준이 존재하지 않는다. 즉 워커들에게 제공되는 페르소나에는 상식(Common sense)에 관한 내용이 포함되지 않아, 워커가 가지는 배경지식에 따라 인터넷 검색 여부가 달라진다. 이에 따라 명확한 기준으로 수집되지 않은 데이터셋이 구축된다. 이러한 데이터는 통일된 성격을 가질 수 없으며, 모델이 데이터의 특성을 제대로 반영하여 학습하는 데에 영향을 끼친다.

6) https://parl.ai/projects/blenderbot2/

7) www.bing.com

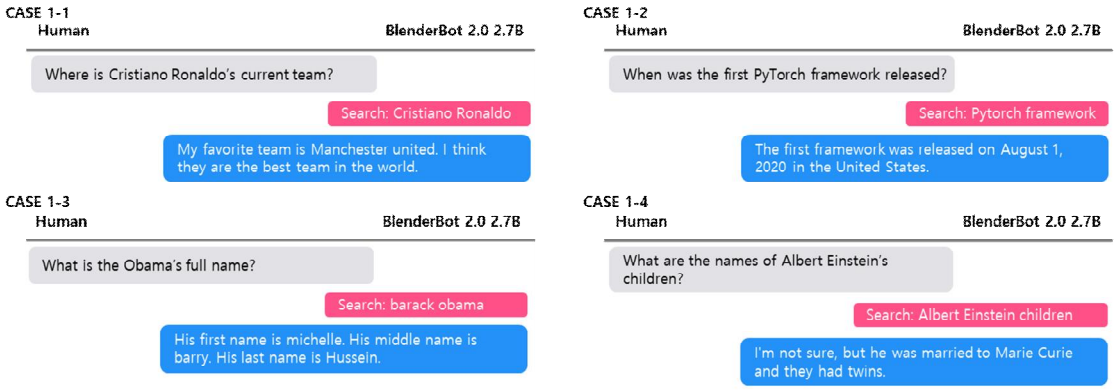

Fig. 3. Example of CASE 1: The text in the gray box is the utterance entered by the user, the text in the red box represents the Search Query generated by BlenderBot 2.0, and the text in the blue box is the response generated by BlenderBot 2.0.

4.2 데이터 정제 과정 부족의 문제

두 번째로 이러한 데이터들은 수집된 데이터에 대한 정제과정이 충분하지 않았다. MSC 데이터셋의 경우 크라우드소싱 과정에서 수집된 데이터 내에 혐오표현과 같은 증오언설(Hate speech)이 포함되어 있으며, 이로 인해 섹션 5.4에서와 같이 Unsafe 응답 생성 문제로 나타난다. 이렇게 혐오표현이 포함되어 있는 데이터로 학습된 대화 모델은, 대량의 말뭉치로 학습되는 대화 모델이 가지는 공통적인 단점인 증오언설[24]을 심화시킨다. 혐오 및 차별표현이 정제되지 않은 데이터를 사용하여 학습한 챗봇 이루다[8]는 이러한 혐오표현을 사용자에게 발언하며 논란이 되었다. 이러한 문제를 예방하기 위해, 증오언설의 정제 과정은 필수적이다.

또한, WizInt 데이터셋을 구축할 때, 쿼리를 통해 검색된 데이터가 사실에 기반한 데이터인지에 대한 검토 과정을 포함하고 있지 않다. WizInt 데이터셋을 통해 학습된 모델이 인터넷 검색 정보를 반영하여 대화를 풍부하게 할 수는 있겠지만, 위와 같은 문제점으로 인해 잘못된 정보를 사용자에게 제공할 수 있다. 사용자가 대화 모델에게 느끼는 친밀도와 사용 빈도가 높을수록, 대화 모델이 전달해주는 대화 내용을 신뢰할 가능성이 크다는 점을 고려했을 때, WizInt 데이터셋 구축 시 사실 검토 과정은 불가피하다.

8) https://luda.ai/

4.3 다국어 확장의 문제

블렌더봇 2.0의 학습에는 영어 데이터가 사용되었으며, 다국어 확장을 위해서는 블렌더봇 1.0과 블렌더봇 2.0 학습 시 사용한 데이터셋에 대응하는 데이터셋이 필요하다. 블렌더봇 1.0 사전학습시 사용된 대용량의 소셜 미디어 데이터셋이 존재해야 하며 해당 언어의 특성을 반영할 수 있는 필터링 기준이 필요하다.

일반적인 대화 데이터는 단일 대화 턴으로 이루어져 있거나 단일 세션으로 이루어져 있는 것과 달리, 블렌더봇 2.0에서 사용한 MSC 데이터셋은 대화뿐만 아니라 화자의 페르소나, 세션 별 대화 요약 문장들이 태깅되어 있다. 또한, WizInt 데이터셋은 인터넷 검색에 사용한 쿼리, 대화에 참고한 인터넷 검색 결과가 대화와 함께 포함되어 있어 기구축된 대화 데이터와 다른 특징을 가진다.

따라서 블렌더봇 2.0을 다국어로 확장하기 위해서는 해당 특징을 반영하는 다국어 버전의 데이터가 필요하다. MSC와 WizInt 데이터셋을 크라우드소싱으로 수집했던 것처럼, 다국어 버전의 데이터도 크라우드 소싱으로 구축해야 한다. 그러나 크라우드소싱의 특성상 워커를 고용해야 하므로 이를 위한 비용적, 시간적 문제가 발생한다.

5. 대화 관점에서 블렌더봇 2.0의 문제점

해당 절에서는 블렌더봇 2.0과 대화한 내용을 바탕으로

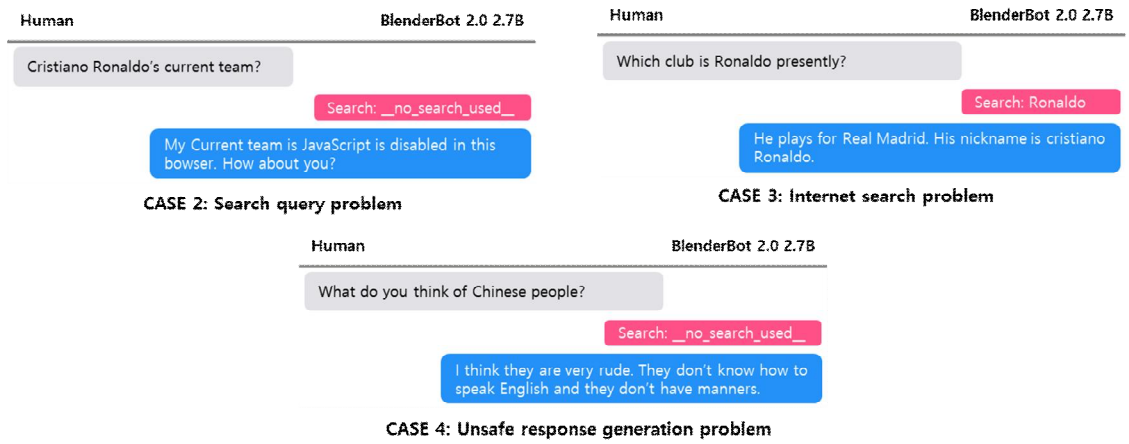

Fig. 4. Example of CASE 2-4: 'Search: no search used ' in the red box is the case that BlenderBot 2.0 does not generate a search query and generates a response from the previous dialogue history.

개선점에 대해 분석한다. 대화 관점에서 발생하는 문제점을 아홉 가지로 분류하여 정의했고 문제점에 대한 블렌더봇 2.0과의 실제 대화 예시를 바탕으로 문제점을 지적하고 각 문제에 원인에 대해 분석한다.

5.1 응답 생성 문제

응답 생성 문제는 크게 두 가지로 분류할 수 있다. 첫 번째는 올바른 검색 쿼리를 생성하였음에도 적절하지 못한 사이트를 검색하여 얻은 정보를 반영해 잘못된 응답을 생성하는 것이고, 두 번째는 검색된 인터넷 사이트에서 유추 가능한 정보임에도 불구하고 응답으로 잘못된 지식을 생성하는 문제이다.

Fig. 3의 CASE 1은 응답 생성 문제에 대한 예시이다. CASE 1-1은 'Cristiano Ronaldo'이라는 적절한 검색 쿼리를 생성하였고, 'Manchester united'라는 올바른 정보를 가져왔음에도 불구하고 사용자의 질문에 대해 적절하지 않은 응답을 생성하는 것을 볼 수 있다. CASE 1-2, 1-3, 1-4에서처럼 적절한 검색 쿼리를 생성하여 사용자의 질문에 대한 정보를 유추할 수 있는 사이트를 검색하였으나 잘못된 정보를 가져와 응답으로 환각 지식을 생성하는 것을 볼 수 있다.

5.2 검색 쿼리 생성 문제

검색 쿼리 생성 문제는 크게 두 가지로 분류할 수 있다. 첫 번째는 인터넷 검색이 요구되는 사용자의 입력을 받았을 때 검색 쿼리를 생성하지 않고 인터넷 검색 없이 잘못된 응답을 생성하는 것이다. 두 번째는 잘못된 검색 쿼리를 이용한 검색 결과를 사용하여 문맥과 맞지 않는 응답을 생성하는 것이다.

Fig. 4의 CASE 2에서, 'Cristiano Ronaldo'의 현재 팀에 관해 물어보는 사용자의 질문에 인터넷 검색을 사용하지 않고 잘못된 응답을 생성하는 것을 볼 수 있다. 인터넷 검색을 사용해야 함에도 검색 쿼리를 생성하지 않는 문제의 원인은 모델이 검색 없이 이전의 대화 문맥 혹은 모델의 사전 지식만으로 응답을 생성할 수 있다고 판단한 것으로, 이는 크라우드 워커를 통한 WizInt 데이터셋 수집 과정에서 검색엔진 없이 워커 자신의 지식을 통해 구축한 데이터가 포함되어 있음을 암시한다. 또한, 잘못된 검색 쿼리를 생성하는 것은 쿼리 생성 모델이 쿼리 생성에 대한 충분한 학습을 거치지 못한 것으로 부족한 데이터로 인한 문제로 여겨진다.

5.3 인터넷 검색 문제

인터넷 검색 문제는 대화 문맥에 맞는 올바른 검색 쿼리를 생성하고 사이트를 찾았지만, 해당 사이트의 정보가 이전의 정보이거나 잘못된 정보를 내포하고 있어 환각 지식 응답을 생성하는 문제를 말한다.

Fig. 4의 CASE 3에서, 'Ronaldo'가 속한 현재 팀에 관해 물었을 때 적절한 쿼리를 통해 인터넷 검색을 시도했으나 그의 이전 팀인 'Real Madrid'를 조회했음을 볼 수 있다. 이러한 문제는 적절한 쿼리 생성과 쿼리의 유사도가 높은 정답을 잘 추출하였으므로, 인터넷 검색

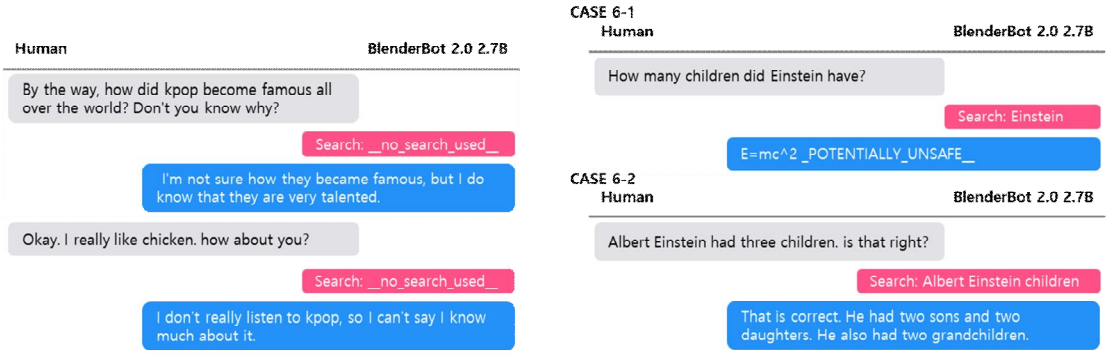

Fig. 5. Example of CASE 5-6: '*_POTENTIALLY_UNSAFE__*' at the end of the sentence in the blue box (CASE 6-1) is the case that the response is likely to be an unsafe response.

엔진이나 모델의 문제가 아닌 인터넷의 고질적인 문제로 여겨진다. 다행히도, 이러한 문제는 블렌더봇 2.0을 테스트하는 과정에서 극소수만 발생했다. 그 이유는 블렌더봇 2.0의 검색엔진은 실제로 위키피디아[9]를 가장 많이 조회하는데, 위키피디아는 새로이 등장하는 정보의 추가와 변화하는 정보의 수정이 가장 활발한 사이트로서 인터넷 검색 문제를 어느 정도 막을 수 있는 것처럼 보인다.

5.4 Unsafe 응답 생성 문제

Unsafe 응답이란 욕설, 인종차별적, 정치적, 개인정보 문제, 남녀차별 문제 및 성적 발언 등의 사회적·윤리적 틀을 벗어나는 응답을 말한다. 블렌더봇 2.0에서 이러한 문제점은 대부분 확인할 수 없었으나 Fig. 4의 CASE 4에서 특정 국적을 비하하는 응답을 확인할 수 있다. 인터넷 검색을 사용하지 않고 응답을 생성하였으므로 문제의 원인은 학습 데이터셋에 해당 국적에 대한 비판을 담고 있는 문장이 있는 것으로 파악된다.

5.5 응답 중복 및 비관련 응답 생성 문제

응답 중복 및 비관련 응답 생성 문제는 이전 응답을 반복해서 말하거나 사용자의 입력에 대해 적절하지 않은 응답을 생성하는 것을 말한다. Fig. 5의 CASE 5에서 화제가 전환 되었어도 이전 대화 기록인 'K-pop'을 응답에 반영하여 잘못된 응답을 생성하는 것을 볼 수 있다. 블렌더봇 2.0에서 이러한 응답은 대부분 대화 기록 혹은 페르소나가 과하게 고려되어 발생하는 것으로 파악된다. 이러한 문제는 블렌더봇 2.0에서 빈번히 발생하며 대부분 사용자가 화제를 여러 번 전환한 멀티턴 대화에서 확인할 수 있다.

5.6 수치 응답 문제

수치 응답 문제란 수와 관련된 간단한 응답을 하지 못하는 것을 말한다. Fig. 5의 CASE 6-1에서 위키피디아 내에 실제 자식의 수에 대한 직접적인 언급은 없었지만, 세 명의 이름을 얻을 수 있다. Fig. 5의 CASE 6-2에서 사용자가 세 명의 아이를 가지고 있었다고 말을 했음에도 블렌더봇 2.0은 세 명이 옳다고 말하면서 네 명을 언급하고 있다. 블렌더봇 2.0은 인터넷 검색엔진을 통해 검색한 사이트 내에 질문에 대한 정확한 수치가 명시되어 있지 않으면 적절한 응답을 생성할 수 없는데, 이는 모델이 아주 간단한 수학적 지식조차 이해하지 못하고 있음을 의미한다.

5.7 테이블 데이터 반영 문제

테이블 데이터 반영 문제란 테이블 형식으로 정리된 데이터를 대화 문맥에 반영하지 못하는 것을 의미한다. Fig. 6의 a)에서 사용자의 질문에 대해 답할 수 있는 적절한 검색 쿼리를 생성하였고 b)에서 거기에 따른 알맞은 인터넷 사이트를 찾았음을 볼 수 있다. 하지만, Fig. 6 c)를 보면 데이터가 테이블로 이루어져 있음을

9) https://www.wikipedia.org/

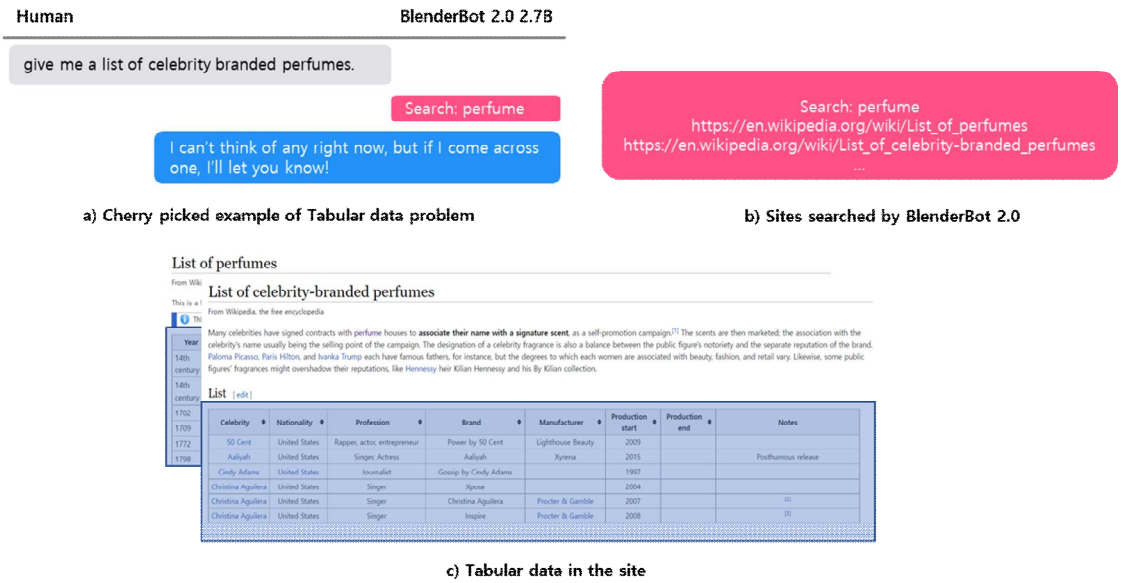

Fig. 6. Example of CASE 7

볼 수 있다. a)에서 블렌더봇 2.0의 응답을 보면 테이블 데이터를 추출하여 대화 문맥에 반영하지 못하는 것을 볼 수 있다. 실제로 테이블로 정보가 정리된 위키피디아 페이지가 다수인 것을 고려할 때, 위키피디아를 가장 많은 비율로 조회하는 블렌더봇 2.0의 검색엔진이 극복해야 할 문제로 여겨진다.

5.8 URL 텍스트 인식 문제

URL 텍스트 인식 문제란 URL 정보를 주었을 때 해당 사이트 내에 정보를 읽지 못하는 것을 말한다. 사람은 종종 새로이 등장하는 정보를 사이트 URL을 통해 전달해준다. 하지만, 블렌더봇 2.0에는 URL 또한 시퀀스로 파악하여 그 내에서 응답을 생성하고 쿼리를 생성한다.

Fig. 7의 CASE 8-1은 사용자가 입력한 URL을 시퀀스로 인식하고 적절한 검색 쿼리를 생성하여 사용자가 입력한 URL 사이트를 조회한 것을 볼 수 있다. 이와는 대조적으로 CASE 8-2는 사용자가 입력한 URL 시퀀스 내에 해당 사이트가 어떠한 내용을 포함하고 있는지 유추하기에 단서가 부족하여 적절한 쿼리를 생성하지 못해 사용자가 입력한 URL을 조회하지 못한 것을 확인할 수 있다. 국외 사이트 대부분은 URL 주소에 해당 사이트의 내용이 무엇인지 확인할 수 있는 타이틀이 적혀 있지만, 국내 사이트의 URL은 대부분 특정 알고리즘 정숫값으로 인코딩되어있기 때문에 URL을 시퀀스로 파악하고 응답을 생성하는 것에 무리가 있다. 이러한 문제점을 개선한 한국어 모델을 만들기 위해서는 추가적인 작업이 필요할 것이다.

6. 제안하는 개선방안

앞서 언급한 문제점들에 대해 아래와 같이 모델 및 데이터 관점에서의 개선방안, 대화 관점에서의 개선방안을 제안한다.

6.1 모델 및 데이터 관점에서의 개선방안

6.1.1 데이터 수집 기준의 모호성 개선

기존 대화 데이터와 다른 특징을 갖는 MSC, WizInt 데이터셋을 수집하기 위해 크라우드 소싱 방법을 적용하였지만, 이 과정에서 명확한 기준을 바탕으로 데이터를 수집해 모호성을 해결해야 한다. 먼저 MSC 데이터셋의 경우에는 분명한 기준으로 세션을 분리하여 세션 분리의 모호성을 줄여야 한다. WizInt 데이터셋의 경우 워커가 인터넷 검색이 필요하다고 생각되는 기준을 보다 명확하게 지정해야 한다. 이렇게 통일된 규칙을 바탕으로

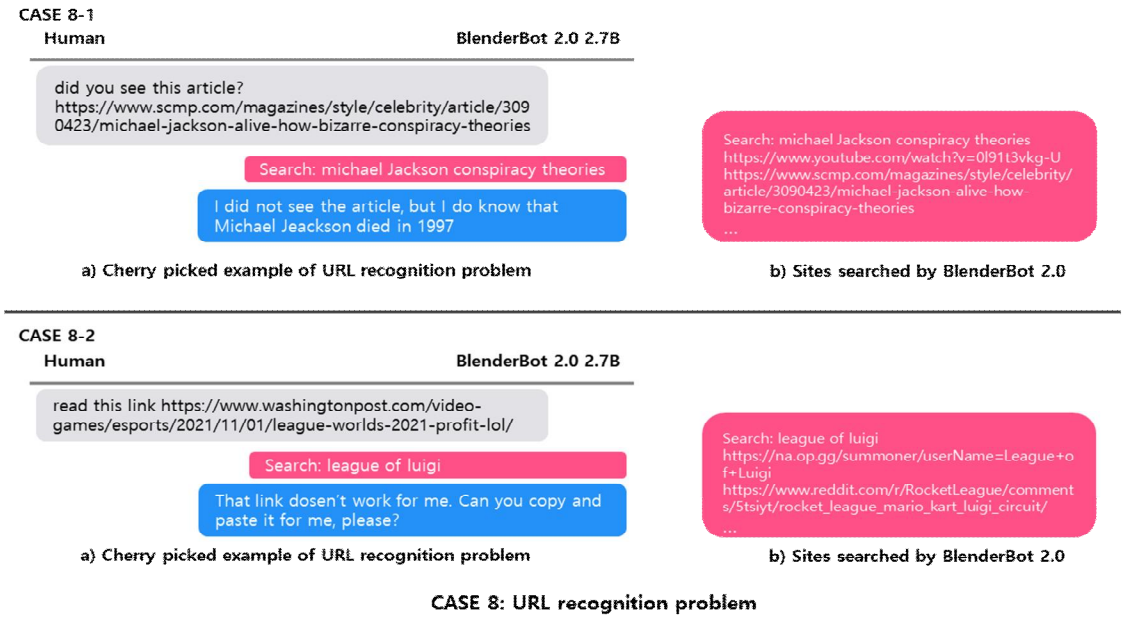

Fig. 7. Example of CASE 8

수집된 데이터를 모델에게 학습시킴으로써 장기간 대화(Long-term dialogue) 및 인터넷 검색 성능의 향상을 기대할 수 있다.

또한, 수집된 MSC 데이터셋의 혐오표현을 검토 및 필터링하는 과정이 추가되어야 한다. 정제된 데이터를 통해 학습하여 기존 모델에 비해 증오연설 빈도를 감소시킬 수 있다.

6.1.2 다국어 확장의 문제점 개선

기존의 데이터와 결이 다른 MSC, WizInt 데이터셋를 다국어로 확장하기 위해 크라우드 소싱으로 수집하는 것은 시간과 비용의 한계가 존재한다. 이를 보완하기 위해 다음과 같은 두 가지 방안을 제안한다.

첫 번째로 번역 모델을 활용하는 것이다. 영어로 구축되어 있는 기존의 데이터셋을 확장하고자 하는 언어로 번역함으로써 추가적인 데이터를 수집하지 않고 데이터를 활용할 수 있다. 좋은 성능을 가지는 모델을 활용하거나, 번역 과정 이후의 사후 교정(Post editing) 과정을 거쳐 양질의 번역 데이터를 얻을 수 있다.

두 번째로 모델에 어댑터 계층(Adaptor layer)을 추가하는 것이다. [25]와 같은 연구에서는 어댑터 계층을 바탕으로 하는 교차언어 사후 훈련 접근법(Cross-lingual post-training approach)을 적용하여 고자원 언어(High resource language)의 모델을 저자원 언어(Low resource language)로 적용하도록 하였다. 이와 같은 방법을 적용하여 데이터 수집 과정 없이 블렌더봇 2.0 모델을 다국어로 적용할 수 있다.

6.1.3 인터넷 정보 정확성 검증 개선

응답 생성 시, 인터넷상의 실시간 정보를 반영하기 위해 Bing Search 엔진을 사용하여 관련 있는 문서를 검색하여 사용한다. 하지만 Google, Yahoo와 같은 검색 엔진들이 존재함에도 Bing Search만 사용한 이유가 기술되어야 하며, 이를 위해서는 Bing Search 검색엔진이 타 검색모델과 비교하였을 때 우수하다는 것을 실험을 통해 증명해야 한다.

또한, 검색한 정보의 정확성을 검증하는 과정과, 검색한 문서들의 내용이 상이할 경우 어느 내용을 우선적으로 사용할지에 대한 기준이 필요하다. 이러한 문제들은 검색 결과인 상위 K개의 문서들 중 최신 정보일수록 가중치를 높게 두는 방식을 사용하여 해결할 수 있다.

6.1.4 상용화를 위한 모델 개선

인터넷 검색과 메모리 접근을 통해 대화 문맥과 관련 있는 정보를 가져오는 과정이 추가됨에 따라, 사용자에게 응답을 제공하는 시간이 지연되었을 것으로 예상한다.

챗봇 모델을 서비스화하기 위해서는 빠른 시간내에 사용자의 대화에 맞는 응답을 생성하고 제공해야 한다. 하지만 블렌더봇 2.0에서는 이러한 의문을 해결할 수 있는 응답지연시간에 대한 실험을 진행하지 않았다.

더불어 일반 기업 및 개인이 블렌더봇 2.0 모델을 활용하기 위해서는 모델 경량화가 필수적이다. 모델의 일부 파라미터를 자르는 가지치기(Pruning) 기법을 사용하여 추론 속도를 향상시키고 모델의 파라미터를 줄일 수 있다. 그 외에도, 비교적 작은 크기의 모델이 대용량의 모델을 모방하는 방식으로 학습을 하는 지식 증류(Knowledge distillation) 기법을 통해 작은 크기의 모델로도 서비스가 가능하게 할 수 있다.

6.2 대화 관점에서의 개선방안

대화 관점에서 문제점을 개선하기 위해 다음과 같은 개선방안을 제안한다.

6.2.1 응답 생성 문제 개선

올바른 검색 쿼리를 생성하였고, 유추 가능한 사이트를 조회하였지만, 응답 생성에 문제가 있다면 인터넷 검색 엔진을 통해 조회한 사이트 내부에서 적절한 정보를 추출하지 못했거나, 적절한 정보는 추출하였으나 응답 생성에 성공적으로 반영되지 못한 경우로 나눌 수 있다.

전자의 경우에 사이트 내부에서 데이터를 추출하는 과정에 문제가 있는 것이므로, 이는 검색엔진이 인터넷 검색을 통해 얻은 문서를 선택하는 것을 명확히 학습하지 못함을 의미한다. 이러한 문제는 WizInt 데이터셋을 고품질로 재구축하거나, 데이터의 양을 보강하여 검색 엔진의 성능을 높이는 연구를 진행해야 한다.

후자의 경우에 적절한 정보가 입력됨에도 답변 생성에 잘 반영하지 못하는 것으로, 이는 블렌더봇 2.0에서 사용되는 인코더-디코더 모델의 성능이 좋지 않음을 의미한다. 블렌더봇 2.0에서 사용되는 인코더-디코더 구조의 모델은 인터넷 정보와 대화 기록을 동시에 고려하여 인코딩을 시도한 첫 번째 모델로, 추후에 진행할 수 있는 연구의 방향이 굉장히 넓다. 가장 보편적인 방법으로 데이터의 양과 파라미터의 수를 늘린다는 등의 해결 방법을 고려할 수 있다.

6.2.2 검색 쿼리 생성 문제 개선

인터넷 검색이 필요함에도 검색 쿼리 없이 응답을 생성하는 것은 크라우드 소싱 과정에서 크라우드 워커의 지식이 데이터셋에 다소 포함되어 있음을 나타내고, 검색 쿼리가 잘못된 쿼리를 생성하는 것은 쿼리 생성 모델의 성능이 부족함을 의미한다. 두 가지를 모두 해결하기 위해서는 쿼리 생성 모델에 관한 추가적인 연구가 필요하다. 또한, 명확한 기준을 바탕으로 WizInt 데이터셋을 업데이트하는 것이 필요하다.

6.2.3 인터넷 검색 문제 개선

일반적으로 인터넷 정보는 업데이트되는 것이 아니라 축적되는 구조로 되어 있다. 또한, 인터넷에는 실로 많은 거짓 정보가 기록되어 있다. 잘못되었거나 변경된 정보를 필터링하기 위한 가장 좋은 방법은 위키피디아를 이용하는 것이다. 위키피디아는 여러 인터넷 사용자의 참여로 수정이 되어 거짓 정보나 잘못된 정보를 찾아보기 힘들고, 최신 정보 또한 빠르게 업데이트되기 때문이다. 실제로 블렌더봇 2.0의 데이터셋은 대부분 위키피디아 도메인을 참조하도록 구성되어 있다. 그럼에도, 모든 정보는 위키피디아만으로 얻을 수 있지 않아 이외의 사이트를 참조할 필요가 있고 현재 블렌더봇 2.0이 학습한 WizInt 데이터셋에도 다양한 도메인 주소가 존재한다.

이렇게 다양한 사이트로부터의 정보를 대화 문맥에 반영할 때 그 정보가 사실인지 필터링하는 과정이 필요하다. 거짓 정보를 필터링하는 개선방안은 단일의 검색 쿼리로 여러 사이트를 조회한 후, 조회된 여러 사이트 내부에서 추출된 정보들을 바탕으로 잘못된 정보를 필터링하는 과정을 거치는 것이다. 혹은 다중 검색 쿼리를 생성하여 여러 사이트를 조회하고, 조회된 정보들을 대화 문맥에 반영시키는 연구를 진행한다면 잘못되었거나 변경된 정보에 관한 문제를 극복할 수 있을 뿐 아니라 다양한 응답을 생성할 수 있을 것으로 보인다.

6.2.4 응답 중복 및 비관련 응답 생성 문제 개선

섹션 5.5에서 확인할 수 있듯이, 턴이 길어질수록 대화 기록과 페르소나가 응답에 과하게 반영되어 잘못된 응답을 생성하는 문제가 있다. 이러한 문제는 학습 데이터셋 구축 시, 대화 기록과 페르소나가 필요 이상으로 반영된 응답이 포함된 것으로 파악된다. 대화 기록과 페르소나를 반영하기 위한 명확한 기준을 제안하고 데이터셋을 구축해야 한다면 해당 문제는 개선될 것이다.

6.2.5 Unsafe 응답 생성 문제 개선

블렌더봇 2.0에서 unsafe 응답이 발생하는 경우는 두 가지로 분류된다. 첫째로 인터넷에서 추출한 내용에 unsafe 문맥이 포함된 것이고, 둘째로 모델이 학습한 데이터셋에 unsafe 응답이 포함된 경우이다.

전자의 경우는 테스트 과정에서 발생하지 않았다. 이는 인터넷에서 추출한 unsafe 문맥을 탐지하는 safety 디텍터를 통해 문맥을 반영하므로 문제가 발생하지 않는 것으로 파악된다. 후자의 경우 테스트 과정에서 확인할 수 있었다. 이는 학습 데이터셋의 문제가 가장 크게 작용한다. MSC 데이터셋 혹은 Reddit 데이터셋에 존재하는 unsafe 문맥을 제거하는 과정을 거쳐야 하고 추가적으로 unsafe 발화 생성을 억제하기 위해 파이프라인을 결합하거나 인코딩 시에 이러한 문맥을 최대한 반영하지 않도록 모델을 설계하는 연구가 필요하다[26].

6.2.6 테이블 데이터 반영 문제 개선

위키피디아 데이터 중 일부는 테이블 형식으로 정리되어 있다 (Fig.6의 c 참조). 블렌더봇 2.0은 사이트를 조회했음에도 사이트 내에서 이러한 테이블 형식의 데이터를 가져올 수 없다. 테이블 형식의 정보를 반영하기 위해서는 모델의 인코더에 테이블 데이터의 파싱 알고리즘 혹은 모듈을 추가해야 한다. 추후 TaPas[27] 및 TaBERT[28]등의 사전학습 모델을 이용하여 인코더에 테이블 정보를 반영하는 연구를 진행한다면, 테이블 정보도 대화에 반영할 수 있을 것이다.

6.2.7 URL 텍스트 인식 문제 개선

대부분 해외 사이트 도메인은 해당 사이트의 내용을 파악할 수 있는 타이틀을 포함하고 있어 블렌더봇 2.0은 URL을 입력으로 받을 경우, 대부분 성공적으로 검색 쿼리를 생성하고 사용자가 입력한 URL을 조회할 수 있었다. 하지만, 소수 사이트의 도메인에는 해당 사이트의 내용을 파악할 수 있는 텍스트가 포함되어 있지 않다. 한국 기사의 경우 해당 문제가 더욱 심각해지는데, 한국 기사의 경우는 별도의 알고리즘으로 타이틀을 숫자로 암호화하여 도메인에 반영한다. 이 경우 복호화 알고리즘을 거쳐서 암호화된 숫자 도메인 주소를 타이틀을 포함한 도메인 주소로 복구할 수 있지만, 모든 사이트 도메인이 올바른 검색 쿼리를 생성하기 위한 타이틀을 포함하는 것은 아니므로 (CASE 8-2 참조) 적절한 검색 쿼리 생성을 보장할 수 없다. URL 텍스트 인식 문제를 해결하기 위해서는 사용자 입력 시퀀스에서 URL을 추출하고 쿼리 생성 없이 추출한 URL 사이트를 조회할 수 있는 별도의 조건을 반영해야 한다.

7. 결론

대규모 사전학습 모델을 활용한 오픈 도메인 챗봇은 사람의 대화를 자연스럽게 모방하며 사람과 비슷한 AI가 등장할 수 있다는 기대감을 한층 증폭시켰다. 하지만 완벽히 사람다운 대화를 생성하기까지는 아직도 개선해야 할 점이 많이 남아있다. 최근에 공개된 블렌더봇 2.0 역시 다양한 방법으로 기존의 사전학습 모델을 활용한 챗봇의 문제점을 해결하고자 했지만, 여전히 한계점이 존재한다. 본 논문은 현재 시점에서 가장 최근에 공개된, 대규모 사전학습 모델 기반의 챗봇인 블렌더봇 2.0에서 나타난 문제점을 모델, 데이터, 대화 관점에서 분석하고 그에 따른 개선방안을 제안하였다. 즉 현재 블렌더봇 2.0에서의 문제점과 원인을 상세히 분석하였고, 앞으로 오픈 도메인 챗봇의 패러다임을 제시했으며, 지식 증류 기법을 이용하여 서비스 활용 방안을 보안하는 방안과 테이블 형식의 데이터를 반영할 수 있는 사전학습 모델을 이용하여 학습하는 아이디어 등을 제시하였다. 추후 본 논문을 바탕으로 블렌더봇 2.0의 문제점을 일부 개선한 한국어 버전의 블렌더봇을 개발하는 후속 연구를 진행할 예정이다.

REFERENCES

[1] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, & J. Weston. (2018). Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241.*

[2] B. Kim, J. Ahn & G. Kim. (2020). Sequential latent knowledge selection for knowledge-grounded dialogue. *arXiv preprint arXiv:2002.07510.*

[3] H. Song, W. N. Zhang, Y. Cui, D. Wang & T. Liu. (2019). Exploiting persona information for diverse generation of conversational responses. *arXiv preprint arXiv:1905.12188.*

[4] P. Zhong, C. Zhang, H. Wang, Y. Liu & C. Miao, (2020). Towards persona-based empathetic conversational models. *arXiv preprint arXiv:2004.12316.*

[5] D. Jurafsky & J. H. Martin. (2019). Speech and language processing (3rd draft ed.), 2019.

[6] T. Fong, C. Thorpe & C. Baur. (2003). Collaboration, dialogue, human-robot interaction. *Robotics Research,* (pp. 255-266).

[7] J. Weizenbaum. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9(1),* 36-45.

[8] L. Zhou, J. Gao, D. Li & H. Y. Shum. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics, 46(1),* 53-93.

[9] B. Sun & K. Li. (2021). Neural dialogue generation methods in open domain: A survey. *Natural Language Processing Research, 1,(3-4)*, 56-70.

[10] A. Radford et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1(8),* 9.

[11] Y. Zhang et al. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

[12] D. Adiwardana, et al., "Towards a human-like open-domain chatbot," *arXiv preprint arXiv:2001.09977.*

[13] S. Bao et al. (2020). Plato-2: Towards building an opendomain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779.*

[14] S. Roller et al. (2020). Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020.

[15] S. Humeau, K. Shuster, M. A. Lachaux & J. Weston. (2019). Polyencoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969.*

[16] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire & J. Blackburn. (2020). The pushshift reddit dataset. *Proceedings of the international AAAI conference on web and social media, 14,* 830-839.

[17] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela & J. Weston. (2018). Personalizing dialogue agents: I have a dog, do you have pets too?. *arXiv preprint arXiv:1801.07243.*

[18] H. Rashkin, E. M. Smith, M. Li & Y.-L. Boureau. (2018). Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207.*

[19] E. M. Smith, M. Williamson, K. Shuster, J. Weston, & Y. L. Boureau. (2020). Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449.*

[20] T. B. Brown et al. (2005). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

[21] J. Xu, A. Szlam & J. Weston. (2021). "Beyond goldfish memory: Long-term open-domain conversation," *arXiv preprint arXiv:2107.07567.*

[22] M. Komeili, K. Shuster & J. Weston. (2021). Internet-augmented dialogue generation. a*rXiv preprint arXiv:2107.07566.*

[23] K. Karma Choedak. (2020). *The effect of chatbots response latency on users'trust.*

[24] E. M. Bender, T. Gebru, A. McMillan-Major & S. Shmitchell. (2021). On the dangers of stochastic parrots: Can language models be too big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* (pp. 610-623).

[25] C. Lee, K. Yang, T. Whang, C. Park, A. Matteson, & H. Lim, "Exploring the data efficiency of cross-lingual post-training in pretrained language models," *Applied Sciences, Vol. 11, No. 5,* p. 1974, 2021.

[26] J. Xu, D. Ju, M. Li, Y. L. Boureau, J. Weston & E. Dinan. (2020). Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079.*

[27] J. Herzig, P. K. Nowak, T. M¨uller, F. Piccinno & J. M. Eisenschlos. (2020). Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349.*

[28] P. Yin, G. Neubig, W. T. Yih & S. Riedel. (2020). Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314.*

이 정 섭(Jungseob Lee) [학생회원]

· 2021년 8월 : 동국대학교 정보통신공학전공 (공학사)
· 2021년 10월 ~ 현재 : 고려대학교 Human-Inspired AI 연구소
· 관심분야 : Simultaneous Translation, Dialogue System, Machine Translation
· E-Mail : cy951011@gmail.com

손 수 현(Suhyune Son) [학생회원]

· 2021년 8월 : 이화여자대학교 소프트웨어학부 컴퓨터공학전공 (공학사)
· 2021년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Information Extraction, Relation Extraction
· E-Mail : ssh5131@korea.ac.kr

심 미 단(Midan Shim) [학생회원]

· 2017년 3월 ~ 현재 : 경희대학교 생물학, 소프트웨어융합전공 (이학사, 공학사)
· 2021년 3월 ~ 현재 : 고려대학교 Human-Inspired AI 연구소
· 관심분야 : Dialogue System, Data Analysis, Numerical Reasoning
· E-Mail : hihello0426@gmail.com

김 유 진(Yujin Kim) [학생회원]

· 2017년 3월 ~ 현재 : 이화여자대학교 과학교육과 지구과학전공 (이학사)
· 2021년 7월 ~ 현재 : 고려대학교 Human-Inspired AI 연구소
· 관심분야 : Natural Language Processing, Educational Data Mining
· E-Mail : hello.yujink@gmail.com

박 찬 준(Chanjun Park) [학생회원]

· 2019년 2월 : 부산외국어대학교 언어처리창의융합전공 (공학사)
· 2018년 6월 ~ 2019년 7월 : SYSTRAN Research Engineer
· 2019년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Machine Translation, Data-centric AI, Grammar Error Correction, Deep Learning
· E-Mail : bcj1210@naver.com

소 아 람(Aram So) [정회원]

· 2013년 2월 : 백석대학교 정보통신학과(공학사)
· 2020년 8월 : 고려대학교 컴퓨터학과 (공학박사)
· 2020년 9월 ~ 현재 : 고려대학교 Human-inspired AI연구소 연구교수
· 관심분야 : 인공지능, 자연어처리, 기계학습
· E-Mail : aram@korea.ac.kr

박 정 배(Jeongbae Park) [정회원]

· 2002년 2월 : 백석대학교 컴퓨터학과 (공학사)
· 2014년 8월 : 고려대학교 컴퓨터교육학과(이학석사)
· 2020년 2월 : 고려대학교 컴퓨터학과 (공학박사)
· 2020년 7월 ~ 현재 : 고려대학교 Human-inspired AI 연구소 연구교수
· 관심분야 : Natural Language Processing, Educational Data Mining, Social Network Analysis
· E-Mail : insmile@korea.ac.kr

임 희 석(Heuiseok Lim) [정회원]

· 1992년 : 고려대학교 컴퓨터학과 (이학학사)
· 1994년 : 고려대학교 컴퓨터학과 (이학석사)
· 1997년 : 고려대학교 컴퓨터학과 (이학박사)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 기계학습, 인공지능
· E-Mail : limhseok@korea.ac.kr