대규모 언어 모델을 활용한 음성 기반 대화 시스템의 가능성 연구

윤재석, 황승현, 김기응
한국과학기술원
{jake.yoon, steven1971, kekim}@kaist.ac.kr

Can Generative Large Language Models Perform Dialogue State Tracker for Spoken Dialogues?

Jaeseok Yoon, Seunghyun Hwang, Kee-Eung Kim
KAIST

요 약

최근 대규모 언어 모델(Large Language Models; LLMs)은 텍스트 기반 대화 시스템에서 뛰어난 성능을 보여주며, 기존의 접근 방식을 대체할 가능성을 보여주고 있다. 그러나 음성 데이터를 활용하는 시나리오에서 LLMs의 활용 방안과 가능성은 아직 충분히 탐구되지 않았다. 본 연구는 음성 관련 데이터가 추가된 MultiWOZ 데이터셋과 자동 음성 인식(Automatic Speech Recognition; ASR) 모델을 활용하여, LLMs가 음성 기반 데이터에서 대화 상태를 추적할 수 있는지를 대화 상태 추적(Dialog State Tracking; DST) 작업을 통해 분석한다. 이 연구는 LLMs가 음성 대화 시스템에서 어떠한 역할을 할 수 있는지 탐색하고, 다양한 크기의 ASR 모델과 함께 LLMs의 성능을 평가 및 분석함으로써, 음성 기반 대화 시스템에서의 잠재적인 효과와 역할을 밝힌다.

Ⅰ. 서론

최근 대규모 언어 모델(Large Language Models; LLMs)의 등장은 언어 이해 분야의 연구에 중요한 전환점을 가져왔다. LLMs은 고도화된 언어 이해 능력으로, 다양한 언어 처리 관련 기술을 처리할 수 있다는 것을 보였고, 언어 이해 능력뿐 아니라 In-Context Learning(ICL)과 같은 방법론이 제안되어 별도의 파라미터 미세조정 없이도 LLMs을 특정 작업을 위해 활용할 수 있게 되었다 [1].

특히 LLMs은 ChatGPT와 같은 텍스트 기반 대화 시스템에서 사용자 의도를 파악하고 적절한 반응을 생성하는 데 중요한 역할을 해왔다. 그뿐만 아니라 최근에는 Instruction Tuning 기법과 LoRA [2]를 활용하여 LLMs을 텍스트 기반 대화 상태 추적기로 활용할 수 있는지에 대한 연구가 진행되었다 [3]. 그러나 LLMs이 음성 데이터를 기반으로 한 대화 시스템에서 어떻게 활용될 수 있는지에 대한 연구는 아직 초기 단계에 있다.

본 연구는 음성 데이터를 기반으로 한 대화 시스템에서 LLMs이 어떻게 활용될 수 있는지에 대한 연구에 대해 새로운 통찰을 제공하고자 한다. 먼저, 우리는 자동 음성 인식(Automatic Speech Recognition; ASR) 모델로 전사문을 생성하고, LLMs를 음성 대화 상태 추적에 적용할 수 있도록 ICL을 활용한다. 이를 통해 음성 대화 상태 추적 작업에서의 LLMs의 활용 가능성을 확인한다. 이외에도 다양한 크기의 ASR 모델에 따른 LLMs의 성능을 분석함으로써, 음성 데이터의 처리와 대화 상태 추적에서 발생하는 실질적인 문제들을 제기하며 이 분야의 향후 연구 방향을 제안한다.

Ⅱ. 연구 배경

음성 기반 대화 시스템에서 대화 상태 추적은 주로 두 가지 접근 방식을 사용한다. 첫 번째는 ASR 모델을 활용하여 음성 데이터를 텍스트로 변환한 후, 텍스트 기반 언어 모델로 대화 상태를 추적하는 모듈화 방식이다. 이 방식은 ASR 과정에서 발생할 수 있는 오류를 별도의 보정 모듈을 통해 수정함으로써, 음성과 텍스트 데이터 간의 차이를 해결한다. 반면, 엔드-투-엔드 방식은 음성 데이터를 직접 처리하여 대화 상태를 추적하고자 하며, 이는 변환 과정에서의 데이터 손실을 최소화하기 위해 개발되었다. 그러나 이 방식은 주로 단일 턴 대화에 한정되어 연구되었다 [4].

DSTC11의 Speech-aware Dialog Systems 트랙 [5]에서는 텍스트 기반 대화 시스템에서 널리 사용되는 MultiWOZ 데이터셋을 확장하여 음성 기반 다중 턴 대화 시스템 연구를 가능하게 했지만, LLMs의 적용에 관한 연구는 아직 초기 단계에 있다. 본 연구는 기존 모듈화 방식과 DSTC11의 데이터셋, 그리고 LLMs를 통합하여 음성 기반 대화 시스템에 LLMs의 활용 가능성을 탐구한다.

Ⅲ. 본론

본 논문에서는 ASR 기술과 LLMs을 모듈화하여 음성 대화 상태 추적 작업에 적용하였다. ASR 시스템으로는 Whisper [6]를 사용하며, 대화 상태 추적을 위한 LLM으로 GPT-4를 활용하였다. 대화 상태 추적 작업은 기존 LLMs이 미세조정 없이 수행하기 어려운 작업이므로, GPT-4 에 In-Context Learning(ICL)을 적용하여 미세조정 없이도 프롬프트 디자인을 통해 모델의 파라미터를 활용할 수 있도록 하였다. 우리가 디자인한 프롬프트는 아래와 같은 상황을 고려했다 (그림 1).

작업에 대한 설명: LLMs이 음성 대화 상태 추적 작업에 대한 이해도를 높일 수 있도록 작업에 대한 설명과 사용자 입력 데이터에 노이즈, 인식 오류, 철자 오류가 포함될 수 있다는 점을 언급하였다.

온톨로지 데이터 활용한 슬롯 설명: MultiWOZ 데이터셋에서 제공된 온톨로지 데이터를 활용해, 대화 상태 추적의 대상이 되는 슬롯에 대한 슬롯 이름과 설명을 제공하였다. 특히 범주형 슬롯의 경우 후보 값을 제공하여 가능한 응답의 범위를 좁혀서 보다 정확한 추론을 할 수 있게 하였다.

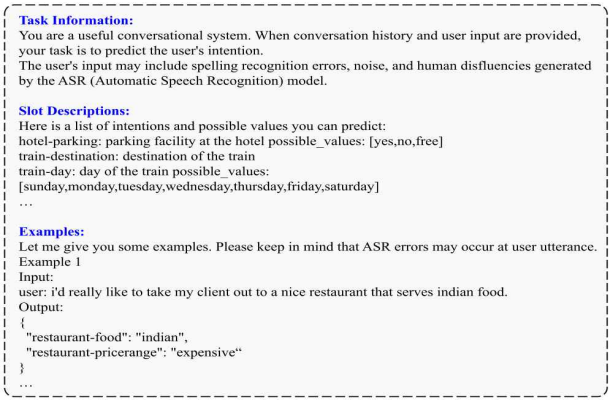

그림 1. 음성 대화 상태 추적을 위한 프롬프트 디자인 예시

다양한 시나리오의 예시: 모델 학습을 위해 제공된 데이터를 활용하여, 다양한 시나리오를 다루는 음성 대화 상태 추적의 예시와 답을 프롬프트에 추가함으로써 LLMs이 작업을 더 명확하게 이해할 수 있도록 하였다.

이렇게 디자인된 프롬프트는 LLMs이 음성 대화 상태 추적 작업에 대한 이해도와 문맥을 이해하는데 필요한 중요한 정보들을 제공하고, LLMs이 음성 데이터의 노이즈와 인식 오류를 효과적으로 처리할 수 있도록 돕는다. 결과적으로 음성 기반 대화 시스템의 성능 향상으로 이어지게 된다.

Ⅳ. 실험 및 결과

본 연구에서는 DSTC11의 Speech-aware Dialog Systems 트랙 [5]에서 제공한 데이터셋을 활용하였다. 표 1은 기존 텍스트 기반 MultiWOZ 데이터셋에서 우수한 성능을 보였던 TRADE [7]의 Joint Goal Accuracy (JGA) 성능이다. DSTC11에서 제공한 음성 기반 특징을 적용한 데이터셋에서는 성능이 30% 이상 크게 감소되는 것을 확인할 수 있다.

표 1. 데이터셋에 따른 TRADE Joint Goal Accuracy 성능 비교

Dataset / Model	MultiWOZ 2.1 (Written Conversation)	DSTC11 (Spoken Conversation)
TRADE	48.7	17.5

본 연구의 메인 실험은 두 단계로 구성된다. 첫 번째 단계에서는 Whisper를 ASR 모델로 사용하여 DSTC11에서 제공한 음성 입력 데이터를 텍스트로 변환하고, Word Error Rate(WER)로 변환된 텍스트의 정확도를 평가한다. 두 번째 단계에서는 변환된 텍스트와 Prompt를 결합하여 ICL기법으로 LLMs에서 대화 상태를 추적하고, Joint Goal Accuracy (JGA)와 Slot Error Rate(SER)를 사용하여 LLMs의 성능을 평가하였다.

표 2는 GPT-4를 LLMs으로 활용하고, Whisper ASR 모델 크기에 따라 성능을 비교한 결과이다. LLMs을 활용할 경우, 미세 조정되지 않았음에도 불구하고, 표 1의 TRADE 성능을 크게 뛰어넘는 것으로 나타났다. 또한 ASR 모델의 성능이 향상됨에 따라 WER이 감소하고 JGA가 증가하는 경향을 보여, ASR 모델의 성능이 LLMs의 음성 대화 시스템에서의 활용 가능성에 중요한 요소로 작용하는 것을 확인 할 수 있었다.

표 3은 대화 상태 추적 중에 발생한 오류 중 개체명 인식(Named Entity Recognition; NER) 오류로 인한 예시를 나타낸다. ASR 모델에서 개체명 인식 오류가 발생한 경우, 이를 그대로 예측값으로 발생시키는 것을 확인할 수 있었다. 즉 LLMs을 활용하더라도 NER 같은 오류의 영향이 여전히 중요하게 나타나, 이는 추가적인 오류 보정 기술이 필요함을 시사한다. 또한, LLMs은 과대 예측하는 현상을 보이기도 했는데, 예를 들어 사용자가 '좋은 호텔'을 언급했을 때 '5성급 호텔'을 원한다고 예측하였다.

표 2. LLMs 활용 시, ASR 모델 크기에 따른 성능 비교

Model Size	#params	WER(%)↓	JGA(%)↑	SER(%)↓
Base	74M	15.45	17.24	40.63
Small	244M	15.29	24.14	40.00
Large	1550M	14.55	31.03	37.50

표 3. 개체명 인식 오류 예시

Input	I'm interested in the hotel called **hillten**. (NER error)
Output	**hillten** hotel
Target	**hilton** hotel

V. 결론

본 연구는 음성 대화 시스템에서 LLMs의 역할을 탐구하고, ASR 모델과 ICL 기법을 사용한 대화 상태 추적의 가능성을 제시하였다. 실험 결과, LLMs는 기존에 미세 조정된 모델들의 성능을 크게 뛰어넘는 것으로 나타났으며, ASR 모델의 정확도가 대화 상태 추적의 효과성에 중요한 영향을 미치는 것을 확인했다. 그러나 LLMs를 활용함에도 불구하고, 고유명사의 인식과 스펠링 오류는 여전히 대화 상태 추적의 주요 도전과제로 남아 있으며, 이를 해결하기 위한 추가적인 후처리 및 오류 보정 기법 개발의 필요성을 시사한다.

ACKNOWLEDGMENT

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2022-0-00311, 일상적 물건들의 다접촉 로봇 조작을 위한 목적지향 강화학습 기술 개발, No.2020-0-00940, 안전한 강화학습 원천 기술 개발 및 자연어 처리에의 응용)

참 고 문 헌

[1] Brown, Tom, et al. "Language Models are Few-shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877 - 1901, 2020.

[2] Hu, Edward J., et al. "LoRA: Low-rank Adaptation of Large Language Models," *in International Conference on Learning Representations*, 2022.

[3] Feng, Yujie, et al. "Towards LLM-Driven Dialogue State Tracking," *in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 739 - 755, 2023.

[4] Saxon, Michael, et al. "End-to-End Spoken Language Understanding for Generalized Voice Assistants," *in Proceedings of INTERSPEECH 2021*, pp. 4738 - 4742, 2021.

[5] Soltau, Hagen, et al. "DSTC-11: Speech aware task-oriented dialog modeling track," in Proceedings of The Eleventh Dialog System Technology Challenge, pp. 226 - 234, Association for Computational Linguistics, 2023.

[6] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision," *in International Conference on Machine Learning*, pp. 28492 - 28518, PMLR, 2023.

[7] Wu, Chien-Sheng, et al. "TRADE: Transferable multi-domain state generator for task-oriented dialogue systems," i*n Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics 2019*, pp. 808-819, 2019.