성격 문장 기반 개인화 챗봇 개발을 위한 소형 언어 모델 미세 조정 연구

이청안
주식회사 에이엘아이

cheonganlee.ali@gmail.com

A Study on Fine-Tuning Small Language Models for Developing Personalized Chatbots Based on Personality Sentences

Cheongan Lee
ALI Co.,Ltd.

요 약

본 논문은 성격 문장들로 정의할 수 있는 개인화된 챗봇을 만들고 이를 통해 더욱 즐거운 대화 경험을 제공하는 것을 목표로 한다. 성격 문장에 기반한 대화 생성을 가능하게 하는 챗봇을 만들기 위해 대규모 언어 모델(LLM)을 활용하여 데이터를 수집하고, 이를 바탕으로 소형 언어 모델을 미세 조정하였다. 대화 생성 과정에서는 상황에 적합한 성격 특성을 선택하고 그에 따른 대화를 생성하였다. 데이터 수집에 두 가지 방식을 사용하여 다양한 데이터를 수집하였다. 연구 결과, 제안한 모델이 생성한 대화가 대규모 언어 모델보다 성격 특성을 잘 반영하는 것을 확인하였다. 또한, 대화의 즐거움에 대한 평가에서도 대규모 언어 모델보다 더 좋은 평가를 받았다. 이 연구는 소형 언어 모델을 효율적으로 미세 조정하여, 더 작은 모델로도 뛰어난 성능을 발휘하는 성격 설정이 가능한 챗봇 개발이 가능함을 보여준다.

Ⅰ. 서 론

최근 몇 년간 챗봇 기술은 급격히 발전하였으며 단순히 프롬프트에 성격을 작성하는 방식으로도 대규모 언어 모델은 성격이 반영된 대화를 생성할 수 있다. 그러나 대규모 언어 모델은 비용의 문제가 있으며 소형 언어 모델에서는 성격을 주어도 잘 반영되지 않는 문제가 있다.

이런 문제를 해결하기 위해 대형 언어 모델을 사용하여 두 가지 방식으로 다양한 데이터를 수집하고 소형 언어 모델을 미세 조정하였다. 또한, 상황에 맞는 적절한 성격 문장을 선택하고 대화를 생성하는 방법을 제안한다.

이런 방법을 통해 소규모 언어 모델로도 OpenAI 의 GPT-4o[5]보다 성격 특성을 잘 반영하는 즐거운 대화를 생성할 수 있었다.

Ⅱ. 본론

1. 데이터

1.1. 데이터 수집

성격 설정이 가능한 챗봇 개발을 위해 본 논문에서는 대화문, 성격 문장, 성격이 표현된 발화 세 가지로 구성된 데이터를 수집하였다. 대화문은 AI 허브의 주제별 텍스트 일상 대화 데이터를 사용하였다. 성격 문장과 성격이 표현된 발화는 OpenAI 의 GPT-4 Turbo 를 통해 두 가지 방법으로 수집하였다. 총 8,800 건의 데이터를 수집하여 학습으로 8,000 건, 검증, 테스트 데이터로 각각 400 건씩을 사용하였다.

첫 번째 방법은 대화문에서 성격을 추출하는 방법으로 프롬프트를 통해 대화문의 한 발화에서 성격을 유추하도록 하였다. 이 때 이 발화에서 성격이 강하게 드러나지 않을 수 있으므로 성격이 잘 드러나도록 개선하도록 요청을 하였다. 이를 통해 성격 문장과 성격이 표현된 발화를 수집할 수 있다.

두 번째 방법은 성격 문장으로부터 대화를 생성하는 방법이다. 첫 단계로 MBTI[1]를 바탕으로 여러 성격을 나타내는 문장을 생성하였다. 두번째 단계로 대화문과 성격 문장을 주고 그 이후 대화를 성격에 맞게 생성하도록 하였다. 그리고 주어진 성격이 가장 잘 표현된 발화를 찾도록 하였다. 주어진 성격 문장과 성격이 가장 잘 표현된 발화, 그 발화 앞의 대화문을 데이터로 사용하였다.

1.2. 수집된 데이터의 다양성

표 1. 데이터 다양성 지표

데이터	압축률	N-그램 다양성	긍정 비율	MBTI 엔트로피
대화로부터	**5.1**	**2.07**	99%	2.56
성격으로부터	11.5	0.60	**48%**	**2.67**

성격 문장 자체의 다양성을 확인하기 위해 diversity 라이브러리[2]를 활용하여 압축률 및 N-그램 다양성을 측정하였다. 두 가지를 기준으로는 대화로부터 수집한 데이터가 높은 다양성을 보였고 문장 자체의 다양성이 높은 것으로 볼 수 있다.

반면, 전체 문장 중 긍정적인 성격 문장의 비율을 나타내는 긍정 비율은 성격으로부터 수집한 데이터가 48%로 긍정과 부정 문장을 고르게 포함하고 있다. 또한, 각 문장에서 MBTI 를 인식하여 MBTI 분포의 엔트로피를 구했을 때 성격으로부터 수집한 성격

문장들이 더 높은 엔트로피를 가지고 있고 이는 각 MBTI 타입이 고르게 나타남을 의미한다.

2. 학습

대화문은 문맥으로만 사용하였고 성격, 성격이 나타난 발화 생성의 확률을 높이도록 학습하였다. 사전 학습된 모델로는 gemma-ko-7b[3]를 대화문만으로 자체적으로 학습한 모델을 사용하였다. 학습 방법으로 Qlora[4]를 사용하였고 4bit 양자화를 사용하였다.

3. 대화 생성

대화 생성은 두 단계로 이루어지는데 첫 단계에서는 주어진 성격 문장들 중 현재 대화 문맥에 맞는 성격이 어떤 문장인지 고르는 단계이다. 이 때 그냥 문장이 생성될 확률을 사용하면 긴 문장일수록 확률이 낮아지는 문제가 있어 문장의 길이로 확률을 정규화하였다.

두번째 단계에서는 대화문과 선택된 성격 문장을 바탕으로 발화를 생성한다. 샘플링은 사용하지 않았고 비슷한 발화가 계속 생성되는 것을 방지하기 위해 문맥이나 이미 생성된 발화에 있는 5-그램은 생성하지 못하도록 하였다.

4. 실험 결과

표 2. 모델 평가

모델	성격 표현력	대화 문맥 일관성	대화 즐거움 승률
GPT-4o	0.4854	94.77	37%
gemma-7b-it	0.3843	40.82	-
gemma-dialog	-	**94.80**	11%
gemma-personality	**0.5030**	94.69	**50%**

제안한 모델의 성능을 검증하기 위해 3 가지 실험을 진행하였다. 현재 가장 좋은 모델로 평가받는 OpenAI 의 GPT-4o, 파운데이션 소형 언어 모델인 구글의 gemma-7b-it, 대화 데이터로만 학습한 gemma-dialog, 본 논문에서 제안하는 gemma-personality 모델을 평가하였다.

모델이 생성한 발화가 주어진 성격 문장을 얼마나 잘 반영하는지 확인하기 위해 성격 표현력을 측정하였다. 성격 표현력은 주어진 대화문과 성격 문장을 바탕으로 발화를 생성하고 성격 문장을 제외한 대화문과 생성된 발화를 입력으로 GPT-4o 를 사용해서 다시 성격 문장을 생성한 후 발화를 생성할 때 입력으로 사용된 성격 문장과 새로 생성된 성격 문장을 문장 유사도 모델(sentence-roberta-large-klue-sts-all)[6]을 사용하여 Cosine 유사도 점수를 매겼다. 본 연구에서 제안한 모델이 GPT-4o 보다 더 높은 점수를 얻었다.

단순히 성격 표현력만 높은 발화문은 대화 문맥에 맞지 않을 수 있기 때문에 대화 문맥에 맞는 발화를 하는 것이 중요하다. 이를 GPT-4o 를 통해 평가하였고 gemma-7b-it 모델을 제외한 나머지 모델은 비슷한 성능을 얻었다.

모델들의 대화가 즐거운지를 평가하기 위해 어떤 MBTI 를 가진 대화 상대로 GPT-4o 를 사용하고 해당 MBTI 가 좋아하는 성격 문장들을 정의하였다. 이 대화 상대와 각 평가 대상인 모델들을 임의의 대화문을 시작으로 대화를 시키고 최종적으로 그 대화가 즐거운 대화였는지 GPT-4o 로 평가를 하였다. 평가의 입력으로는 대화의 상대방이 좋아하는 성격 문장들, 주어진 대화문, 두 모델이 각기 생성한 대화문이 두가지 주어졌고 이 두 대화문 중 어떤 대화가 더 즐거웠는지 평가하도록 하였다. 비교 대상은 본 논문에서 제안한 모델로 고정하여 나머지 모델과 비교를 하였다. GPT-4o 도 본 논문에서 제안한 모델과의 전체 대결 회수 중 37%에서밖에 승리를 거두지 못했다.

Ⅲ. 결론

본 논문에서는 성격을 가진 챗봇을 만들기 위해 데이터 수집, 학습, 대화 생성 방법을 제안하고 이를 통해 챗봇을 만들고 평가를 하였다. 소형 언어 모델을 통해서도 대형 언어 모델보다 더 좋은 성능을 가지는 성격을 가진 챗봇을 만들 수 있음을 확인하였다. 하지만 제안한 방법에서는 성격 문장을 하나만 선택하여 사용하는 방식을 채택하여 선택되지 않은 다른 성격 문장에 반하는 발화를 생성하는 경우가 있는 것을 확인하였고 이를 해결하기 위해 추가적인 연구가 필요하다.

ACKNOWLEDGMENT

본 연구는 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발사업 지원에 의하여 이루어진 것임 (과제고유번호: RS-2023-00267328)

이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 주제별 텍스트 일상 대화 데이터를 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

참 고 문 헌

[1] Myers-Briggs I. "The Myers-Briggs type indicator manual." Princeton, NJ: Educational Testing Service. 1962.

[2] Shaib C, Barrow J, Sun J, Siu AF, Wallace BC, Nenkova A. “Standardizing the measurement of text diversity: A tool and a comparative analysis of scores.” arXiv preprint arXiv:2403.00553. 2024 Mar 1.

[3] Lee J, Choi T. "gemma-ko-7b" [Internet]. 2024 [cited 2024 Aug 21]. Available from: https://huggingface.co/beomi/gemma-ko-7b. doi: 10.57967/hf/1859.

[4] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. "Qlora: Efficient finetuning of quantized llms". Advances in Neural Information Processing Systems. 2024 Feb 13;36.

[5] OpenAI. "Hello GPT-4o" [Internet]. OpenAI; 2024 [cited 2024 Aug 21]. Available from: https://openai.com/index/hello-gpt-4o/

[6] ys7yoo. "sentence-roberta-large-klue-sts-all" [Internet]. 2023 [cited 2024 Aug 21]. Available from: https://huggingface.co/ys7yoo/sentence-roberta-large-klue-sts-all