

인간과 ChatGPT의 대화내용을 이용한 공개 대형 언어모델 LLaMA 한국어 대화 능력 개선

주호택, 이성하, 김경중
광주과학기술원

ureca87@gmail.com, shlee0414@gm.gist.ac.kr, kjkim@gist.ac.kr

Improving Korean conversational skills of LLaMA, a public large language model, using human and ChatGPT conversations

Ho-Taek Joo, Sung-ha Lee, Kyung-Joong Kim
Gwangju Institute of Science and Technology

요 약

최근 OpenAI의 ChatGPT, 구글의 Bard와 같은 대형 언어모델이 출시되었지만 이러한 모델들은 상업적 이유로 공개되지 않아, 일반 연구자들은 범용 인공 지능에 가까운 대형언어모델을 연구할 기회가 제한된다. 최근 메타(Meta)는 대형 언어모델인 LLaMA를 공개 및 배포하였고, 많은 연구자들은 이 모델을 기반으로 연구할 기회를 얻게 되었다. 하지만 LLaMA 모델이 학습한 20개의 언어 데이터 셋에는 한국어가 포함되어 있지 않기 때문에 우수한 한국어 대화 능력을 가지고 있으면서, 가중치가 공개된 (fine-tuning이 가능한) 대형 언어 모델은 없는 상황이다. 본 논문에서는 한국어 의사소통이 가능한 LLaMA 모델의 개발 및 활용 가능성을 탐구하였으며, 제안한 방법을 통해 학습한 한국어 대화 능력이 개선된 모델을 공개한다. (<https://huggingface.co/ureca07/korean-vicuna-7b-1.1>)

1. 서 론

최근 OpenAI의 대형 언어 모델인 ChatGPT는 인간과 경쟁할 수 있는 수준의 성과를 보여주며 범용 인공 지능에 가까운 모습을 보여준다. 그러나 이 모델은 상업적 이유로 일반 연구자들에게 공개되지 않아 대형 언어 모델(Large Language Model, LLM) 연구 기회가 제한된다. 대안으로 Facebook의 모회사인 Meta는 최신 AI 대형 언어 모델인 LLaMA를 공개하였다. LLaMA가 공개된 이후 많은 연구자들은 이 모델의 파라미터에 대한 미세조정 통해 많은 성과를 보이고 있다. 하지만 이 모델의 학습 데이터 셋에는 한국어는 포함되어 있지 않아 한국어 소통 능력이 매우 부족한 수준으로, 한국어 데이터로 연구하기 힘들다는 문제가 있다.

본 연구의 목적은 LLaMA 기반 모델이 한국어 의사소통 능력을 탑재할 수 있도록 학습 방안을 제안하는 것이다. 이를 위해 본 논문에서는 ChatGPT와 인간의 대화 내용을 수집했다. 그리고 이 대화내용을 한국어로 번역하고, 모델에 입력하여 한국어를 이해할 수 있도록 모델의 파라미터를 미세 조정(fine-tuning) 하였다. 또한 개발한 모델을 평가하기 위해 몇 가지 질문을 던졌고, 학습된 모델은 이 질문에 대답하도록 하였다.

본 연구를 통해 한국어 의사 소통이 가능한 LLaMA 기반 대형 모델의 개발 가능성을 탐구하고자 하였고, 이를 통해 한국어 대형 언어 모델 연구에 대한 기여하고자 한다.

2. 관련 연구

2-1. GPT-3 (Generative Pre-trained Transformer 3)

GPT-3는 2021년 OpenAI에서 개발한 대형 언어 모델로 범용 인공 지능의 가능성을 처음으로 제시한 모델이다.

대규모 데이터 세트에서 사전 학습된 GPT-3는 이전 모델인 GPT-2보다 훨씬 더 큰 규모인 175억 개의 파라미터를 가지고 있다. LLM이라고 불리는 엄청난 크기의 모델로 인해 자연어 생성, 번역, 요약, 질문 답변 등 다양한 NLP 작업에서 뛰어난 성능을 제공하였다.

2-2. FLAN (Few-shot Learning with Adaptive Networks)

2021년 10월 구글은 명령어 조정 학습(instruction learning)이라는 새로운 학습 개념을 제시하였다. 명령어 조정 학습은 사전 학습 종료 후 각 태스크에 맞는 명령어를 통해 추가적인 학습을 진행하는 방법이다. 하지만 기존의 추가 학습을 통한 특정 태스크 최적화와 달리, 단일 모델로 다양한 종류의 태스크를 수행할 수 있는 능력을 갖추었다.

2-3. ChatGPT

ChatGPT는 인공지능 연구의 발전에 따라 GPT-3와 FLAN의 강점을 결합한 대형 언어모델이다. ChatGPT는 GPT-3의 광범위한 지식과 성능을 기반으로 FLAN의 명령어 조정 학습 개념을 통합하여 다양한 태스크를 수행할 수 있는 높은 범용성을 갖추게 되었다.

2-4. LLaMA

2023년 2월 페이스북 모회사인 메타(Meta)는 대규모 언어 모델인 LLaMA를 배포했다. 메타의 CEO인 마크 저커버그는 AI연구자들의 연구 발전을 돕기 위해 LLaMA를 출시했다고 언급하며, 텍스트 생성, 대화, 자료 요약, 수학 풀이, 단백질 구조 예측, 코드 도움 등이 가능한, 이 모델을 AI 연구 커뮤니티에 제공하였다. 메타가 공개한 이 대형모델은 파라미터가 70억 개부터 650억 개에 이르는 모델 4종을 출시하였으며, 비상업적 라이선스에 따라 대학, 기관등에서 무료로 사용할 수 있다. 하지만 한국어

데이터 셋이 포함되어 있지 않은 LLaMA 모델은 한국어 대화 능력이 매우 낮으며, 이를 개선하기 위해 본 논문에서는 이 LLaMA 모델을 미세조정하여 한국어로 대화 가능한 수준으로 향상하는 것을 제안한다.

3. 학습 방법 및 실험

본 연구의 목적은 LLaMA 모델의 한국어 대화능력을 개선하는 것으로 인간과 ChatGPT의 대화내용을 수집하고, 수집한 데이터를 통해 이 모델의 파라미터를 미세 조정 하여 한국어 능력을 개선시켰다.

3-1) 데이터 수집

OpenAI가 제공한 ChatGPT는 사용자의 질문과 질문에 대한 ChatGPT의 대답을 저장하는 기능을 제공한다. ShareGPT 웹사이트는 이 대화 데이터를 수집하는 목적으로 만들어졌으며, 최근 수집한 데이터를 외부에 공개하였다. 본 연구에서는 사람과 ChatGPT의 공개 대화 데이터를 DeepL 번역 API를 이용하여 한국어로 번역하여 학습 데이터를 만들었다. 이 대화 데이터 셋은 약 60만 쌍의 대화문이 포함된다.

3-2) 데이터 전처리

데이터 전처리 과정은 세 가지 단계로 진행된다. 첫 번째는 데이터 전처리 단계이다. 이 단계에서는 중복된 질문 및 대답 삭제 그리고 html 문서나 markdown과 같은 인터넷 문서에서 발생하는 불필요한 문구(e.g. <div>)를 삭제한다. 두 번째 단계는 아래 예시와 같이 전처리 된 데이터를 학습된 모델이 잘 이해할 수 있도록 프롬프트를 이용하여 대화문을 만드는 것이다.

[표 1.] 프롬프트를 이용한 대화문

```
{
  "id": "1", "conversations": [
    {
      "from": "human", "value": "호기심 많은 USER와 인공지능 ASSISTANT 간의 채팅.", "ASSISTANT가 USER의 질문에 대해 유용하고 상세하며 정중한 답변을 제공합니다."
    },
    {
      "from": "USER", "value": "팝 음악 비디오 클레에 대한 아트 제안을 만들어주세요."
    },
    {
      "from": "Assistant", "value": "Assistant: {,
      "from": "gpt", "value": "팝 음악 비디오 클립을 위한 아트 시안을 소개합니다.\n\n제목: .... } ]
```

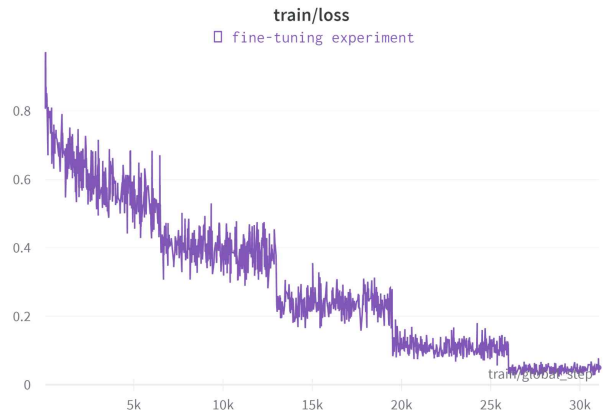
[표. 1] 대화문처럼 인간이 인공지능에게 어떤 질문을 했고, 어떤 대답을 듣기를 원하는지 적절하게 질문을 하는 방법을 “프롬프트” 라고 한다. 우리는 “호기심 많은 USER와 인공지능 ASSISTANT 간의 채팅.”, “ASSISTANT가 USER의 질문에 대해 유용하고 상세하며 정중한 답변을 제공합니다.” 과 같이 gpt에게 자세한 요구사항을 프롬프트로 미리 정의하여 모든 대화문 앞에 포함시켰다.

세 번째 전처리 단계는 대화 내용을 길이에 따라 나누는 과정이다. 언어 모델을 학습할 때 입력할 토큰 시퀀스가 길어지면 모델이 해당 문장을 학습하기가 어려워 성능이 떨어진다. 이에 본 연구에서는 토큰 길이를 3072에 제한을 두고 대화 내용을 나누었다.

3-3) 파인 튜닝 실험

LLaMA 기반 모델은 일반적으로 미세조정 하는 방법으로 메타에서 발표한 LoRA(Low-Rank Adaptation of Large Language Models)[1]라는 방법을 사용한다. 이는

모델의 파라미터의 대부분을 고정시키고, 트랜스포머(Transformer)모델의 주의 블록(Attention Block)에 해당하는 부분만 파라미터를 변화시키는 것이다. 하지만 이 방법을 실험해본 결과 한국어 의사 소통 능력이 크게 개선 되지 않았다. 이에 본 연구에서는 LLaMA 모델의 모든 파라미터를 미세조정 하는 방법을 사용하였으며, 손실함수는 교차 엔트로피(cross entropy)를 사용하였다. 손실함수에서 정답(label)은 ChatGPT가 대답한 내용이며, 예측 값(prediction)은 인간과 ChatGPT가 대화한 문장들이 나열 되어 있을 때 인간이 질문한 내용은 마스킹(masking) 하여 ChatGPT가 대답한 내용만을 예측할 수 있도록 하였다.



[그림 1.] 훈련시 손실 함수 그래프

실험에 대한 손실함수는 [그림 1.]처럼 모델을 훈련하는 동안 지속적으로 감소하는 것을 볼 수 있다. 본 논문에서는 이 실험을 위해 A100 GPU 8대를 사용하여 학습시켰으며, 학습 시간은 약 50시간이 소요되었다.

4. 실험 결과

본 연구에서는 한국어 능력을 개선한 모델과 LLaMA 모델의 한국어 대화 능력에 대해서 비교한다. 일반 질문, 코딩 질문, 역사 질문, 총 세 가지에 대해서 질문하였다.

[표 2-1.] 한국어 대화 능력이 개선된 모델 (예시 1)

```
USER: 하와이 여행에 대한 3일 계획과 예산을 알려주세요.
ASSISTANT:
1일차:
* 하와이 주립 공원 입장료: 성인 입장료 - $5, 어린이 입장료 - $3
* 하이킹: 히키 하이츠, 메인 리프, 브라운 레인지 등 하와이 주립 공원에는 멋진 하이킹 코스가 많이 있습니다. 일반적인 하이킹 루트는 성인 1인당 약 $10~$20, 어린이 1인당 $5~$10입니다.
* 렌트 앤 토크: 하와이에는 팝 루프, 브라운 레인지 등 아름다운 렌트 앤 토크 코스가 많이 있습니다. 일반적인 렌트 앤 토크 루트는 성인 1인당 $20, 어린이 1인당 $10입니다.
* 해변 탐험: 하와이는 일광욕, 수영, 일광욕 활동을 위한 아름다운 해변으로 유명합니다. 하와이에는 워터파크 및 수상 스포츠 시설, 산책로와 자연 명소가 많이 있습니다.
2일차:
```

* 골프: 하와이에는 하와이 주립 공원 및 매르헤 나르 국립공원과 같은 여러 골프 코스가 있습니다. 티어에 따라 가격은 성인 \$100-\$200, 어린이 \$75-\$150입니다.

* 롤리팝 코스: 하와이에는 하와이 주립 공원 및 매르헤 나르 국립공원과 같은 여러 공과 골프 코스가 있습니다. 티어에 따라 가격은 성인 \$100-\$200, 어린이 \$75-\$150입니다.

3일차:

* 하와이 디너레스토랑 입장료: 성인-\$20-\$30, 어린이- \$10-\$15

* 하와이 베이 스포츠 캠프에 참여하기: 하와이 베이 스포츠 캠프에서는 육상, 카약, 스노클링 등 다양한 스포츠 활동을 즐길 수 있습니다. 일반적으로 \$45부터 시작하는 스포츠 패키지 가격입니다.

이는 예시일 뿐이며 가격은 계절과 시즌에 따라 달라질 수 있다는 점에 유의하세요. 하와이에 도착하기 전에 항상 가격을 확인하는 것이 좋습니다.

[표 2-2.] LLaMA 모델 (예시 1)

1일 계획

* 오전 5시간에 하와이 도심을 둘러보게 여행을 시작합니다.

* 하와이 특별한 명소인 베르다운 폴리올레나 폭포, 덩크리텔링 레스토랑, 하와이 체인 등을 방문 감사합니다.

* 오후 1시간에 하와이 모델상품 방문 감사합니다.

* 하와이 도심을 돌보고 밤 베르다운 포토로스 호텔에 시작하여 3일 계획을 위한 간편한 체험을 취하세요.

2일 계획

* 오전 10시간에 베르다운 포토로스 호텔에 돌아오세요.

* 하와이 도심을 둘러보게 여행을 시작합니다.

* 하와이 맨 특별한 명소인 성도서였이 포토, 예약한 덩크리텔링 레스토랑, 베르다운 포토로스 호텔 등을 방문합니다.

* 하와이 도심을 둘러보게 여행을 종료합니다.

3일 계획

* 오전 5시간에 하와이 도심을 둘러보게 여행을 시작합니다.

* 하와이 특별한 명소인 발라드 센트럴, 카드 오브 빅게이더, 하와이 필리핀 등을 방문합니다.

* 오후 1시간에 하와이 특별한 체험 숙소인 베르다운 블랙엔톤 포토로스 호텔에 시작하여 3일 계획을 위한 간편한 체험을 취하세요.

[표 3-1.] 한국어 대화 능력이 개선된 모델 (예시 3)

USER: **삼국 통일 이후 신라에 대해서 설명해주세요.**

ASSISTANT: 기원전 450년 삼국 통일 이후, 신라는 불안정한 정치 상태에 빠졌습니다. 이때부터 신라는 왕권의 약화, 부패의 만연, 부족끼리의 충돌 등을 겪게 됩니다. 기원전 499년에는 백제의 침략을 받았고, 500년에는 일본에 의해 점령당했습니다. 이후 신라는 기원전 57년, 499년, 457년 등 여러번 백제의 침략을 받았으며, 이 시기에 신라는 왕권의 약화와 권력 비극을 겪게 되었습니다. 이러한 사건으로 인해 신라는 불안정하게 되었으며, 부패의 만연과 불안정한 정치 상황이 계속되는 시기를 거쳐 멸망하게 됩니다.

[표 3-2.] LLaMA 모델 (예시 3)

ASSISTANT: 통일 신라시대는 한국의 왕조로서 7세기부터 9세기까지 통치했으며, 시간이 너무 길 걸 확실히 없습니다. 그러나 몇 가지 출처에서는 시간이 더 작을 수 있으며, 57 BC에 시작되고 935년까지 지속되던 것으로 여겨집니다. 그러나 일부 출처는 1살에서 시작하고 10세기 이후에 끝났을 수도 있습니다.

[표 4-1.] 한국어 대화 능력이 개선된 모델 (예시 2)

```
USER: 피보나치 수열을 출력하는 파이썬 코드를 작성해줘.
ASSISTANT:
def fibonacci(n):
    if n <= 1: return n
    else: return fibonacci(n-1) + fibonacci(n-2)
for i in range(10):
    print(fibonacci(i))
```

[표 4-2.] LLaMA 모델 (예시 2)

```
ASSISTANT:
import numpy as np
data = np.random.normal(loc=10, scale=1, size=100)
data_column = data.reshape(10, 10)
print(data_column)
```

5. 실험결과 분석

[표 2.]는 일반 질문으로 하와이 여행 계획에 대한 질문을 하였다. LLaMA 모델 같은 경우, 사용자에게 정보를 제공해야하는 모델임에도 불구하고 “방문감사합니다, 여행을 종료합니다, 오전 10시간에” 와 같이 한국어 소통 능력이 떨어지는 것을 볼 수 있으며, 대화 내용의 질적인 측면에서도 많이 부족하다. [표 3.]에서는 역사질문으로, 제시한 모델이 한국어 능력은 나름 준수하였지만 허구가 섞인 대화 내용을 제시하였다. 반면 LLaMA 모델은 한국어 능력이 부족하며, 정보에도 허구가 섞인 것을 볼 수 있다. [표 4.]에서는 코딩 능력을 테스트 한 실험 결과로 제시한 모델은 한국어를 잘 알아듣고 정답 결과를 출력하는 코딩을 작성해주었지만, LLaMA 모델은 코딩 테스트 결과가 제대로 나오지 않았다.

6. 결론

현재 해외의 많은 연구자들은 LLaMA 모델을 바탕으로 연구가 한창이다. 법률, 의료 등 특정 분야에서 활용 가능한 소규모 언어모델 (Small Large Language Model, sLLM)의 활용성 탐구, 자가 성장가능한 AutoGPT, 다양한 입력이 가능한(멀티모달GPT) 등 다양한 방식으로, 다양한 분야에서 연구가 진행되고 있다. 하지만 한국에서는 소수의 대형 IT기업을 중심으로 대형언어 모델 연구가 제한적으로 이루어지고 있으며 일반 연구자들에게는 연구할 기회가 부족하다. 본 논문에서는 LLaMA 기반 대형언어 모델의 한국어 의사소통 능력을 개선시켰고, 이를 통해 LLaMA 모델이 수행할 수 있는 많은 작업을 한국어로 활용할 수 있도록 하였다. 본 연구를 활용하여 많은 연구자들이 한국어 대화 소통이 가능한 대형 언어모델을 학습하는 연구를 진행할 수 있을 것으로 기대한다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2021R1A4A1030075)

참고문헌

[1] Hu, Edward J., et al. “Lora: Low-rank adaptation of large language models.” arXiv preprint arXiv:2106.09685 (2021).