한국정보전자통신기술학회논문지(jkiiect)'23-8, Vol.16 No.4
논문 23-16-04-193 http://dx.doi.org/10.17661/jkiiect.2023.16.4.193

오픈 소스 기반의 거대 언어 모델 연구 동향: 서베이

주하영*, 오현택**, 양진홍°

A Survey on Open Source based Large Language Models

Ha-Young Joo*, Hyeontaek Oh**, Jinhong Yang°

요 약 최근 대규모 데이터 세트로 학습된 거대 언어 모델들의 뛰어난 성능이 공개되면서 큰 화제가 되고 있다. 하지만 거대 언어 모델을 학습하고 활용하기 위해서는 초대용량의 컴퓨팅 및 메모리 자원이 필요하므로, 대부분의 연구는 빅테크 기업들을 중심으로 폐쇄적인 환경에서 진행되고 있었다. 하지만, Meta의 거대 언어 모델 LLaMA가 공개되면서 거대 언어 모델 연구들은 기존의 폐쇄적인 환경에서 벗어나 오픈 소스화되었고, 관련 생태계가 급격히 확장되어 가고 있다. 이러한 배경하에 사전 학습된 거대 언어 모델을 추가 학습시켜 특정 작업에 특화되거나 가벼우면서도 성능이 뛰어난 모델들이 활발히 공유되고 있다. 한편, 사전 학습된 거대 언어 모델의 학습데이터는 영어가 큰 비중을 차지하기 때문에 한국어의 성능이 비교적 떨어지며, 이러한 한계를 극복하기 위해 한국어 데이터로 추가 학습을 시키는 한국어 특화 언어 모델 연구들이 이루어지고 있다. 본 논문에서는 오픈 소스 기반의 거대 언어 모델의 생태계 동향을 파악하고 영어 및 한국어 특화 거대 언어 모델에 관한 연구를 소개하며, 거대 언어 모델의 활용 방안과 한계점을 파악한다.

Abstract In recent years, the outstanding performance of large language models (LLMs) trained on extensive datasets has become a hot topic. Since studies on LLMs are available on open-source approaches, the ecosystem is expanding rapidly. Models that are task-specific, lightweight, and high-performing are being actively disseminated using additional training techniques using pre-trained LLMs as foundation models. On the other hand, the performance of LLMs for Korean is subpar because English comprises a significant proportion of the training dataset of existing LLMs. Therefore, research is being carried out on Korean-specific LLMs that allow for further learning with Korean language data. This paper identifies trends of open source based LLMs and introduces research on Korean specific large language models; moreover, the applications and limitations of large language models are described.

Key Words : Survey, Large Language Model, Open source, English, Korean

1. 서론

ChatGPT의 등장으로 거대 언어 모델의 시대가 열렸다. OpenAI가 개발한 ChatGPT는 사전 학습된 생성형 인공지능 챗봇으로 뛰어난 문장 생성 성능을 보여주며 출시 두 달 만에 월간 활성 사용자 수가 1억 명을 달성할 정도로 큰 화제가 되었다. 거대 언어 모델의 선두주자인 OpenAI, Google과 같은 빅테크 기업들은 거대 언어 모델이 답하는 결과의 편향성이나 환각(Hallucination; 없는 내용을 마치 사실인 것처럼 지어내는 것) 현상 등을 이유로 거대 언어 모델을 완전히 공개하는 것은 위험하다고 주장하며 기술들을 공개하

This work was partly supported by the KAIST-Megazone Cloud Research Center for Intelligent Cloud Computing Convergence grant funded by MEGAZONE CLOUD Corp. and partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(NRF-2022R1C1C2003437)
* KAIST-Megazone Cloud Research Center for Intelligent Cloud Computing Convergence Research (hyjoo@kaist.ac.kr)
** KAIST Institute for Information Technology Convergence & KAIST-Megazone Cloud Research Center for Intelligent Cloud Computing Convergence Research (hyeontaek@kaist.ac.kr)
° Corresponding Author : Department of Medical IT, INJE University (jinhong@inje.ac.kr)
Received July 31, 2023 Revised August 04, 2023 Accepted August 16, 2023

지 않았고 이로 인해 폐쇄적인 생태계가 만들어졌다. 거대 언어 모델에 대한 직접적인 접근이 제한되었으며 유료로 API를 제공하거나, 자사의 서비스를 통하여 이용해야 했다.

한편, 2023년 3월 Meta의 거대 언어 모델 LLaMA가 공개되며 거대 언어 모델 연구의 오픈소스화 및 대중화가 촉발되었다. Meta는 연구 목적으로 신청자에게만 소스 코드를 무료로 공개하였으나, 해당 코드가 유출되면서 누구나 거대 언어 모델을 자신의 컴퓨터에서 직접 실행시킬 수 있게 되었다. 해당 사건으로 LLaMA 모델을 변형한 다양한 언어 모델이 Github, Huggingface와 같은 플랫폼에 공개되면서 거대 언어 모델의 오픈 소스 환경이 조성되었으며, 현재는 Alpaca, Vicuna, Ployglot 등 경량화되고 뛰어난 성능을 가진 모델들이 등장하며 거대 언어 모델 생태계가 활성화되고 있다.

2023년은 거대 언어 모델 연구 개발의 오픈 소스가 활발해지는 중요한 해로 지난 6개월 남짓한 시간동안 다양한 오픈 소스 기반 프로젝트들이 공개되었다. 이러한 배경하에, 본 논문에서는 거대 언어 모델에 대한 개념과 생태계의 변화와 함께 오픈 소스 기반 거대 언어 모델의 동향을 중점적으로 살펴보고자 한다. 또한, 오픈 소스 기반의 한국어에 특화된 한국어 전용 거대 언어 모델과 관련 연구를 소개하고, 거대 언어 모델의 다양한 활용 방안 및 한계점에 관해 설명하고자 한다.

2. 거대 언어 모델

2.1 거대 언어 모델의 이해

거대 언어 모델(Large Language Model, LLM)은 방대한 데이터를 학습하고 수천억 개 이상의 파라미터 수를 갖는 대규모 언어 모델을 말한다[1]. 대규모 텍스트 데이터로 학습되어 다양한 언어 현상과 지식을 습득할 수 있어 자연어에 대한 이해가 높으며 새로운 문장을 생성할 수도 있다. 또한, 기본적인 언어 모델을 바탕으로 특정 작업과 도메인에 맞게 추가적인 학습이 가능하다.

거대 언어 모델은 매개변수(Parameter)의 개수가 많을수록 크다고 말하며, 매개변수는 인공지능이 연산을 위해 고려하는 다양한 변수로 이것이 많을수록 데이터로부터 다양한 정보를 학습할 수 있다[2]. 모델의 크기가 크다는 것은 더 많은 정보를 언어 모델이 가질 수 있다는 것이며 이는 정확도가 높아질 수 있음을 의미한다[3].

거대 언어 모델을 학습하기 위해서는 좋은 성능을 가진 컴퓨팅 연산 장치(예: GPU)와 대용량 데이터를 관리하는 저장 장치(예: GPU 메모리)가 필요하므로 거대 언어 모델을 구축하는 환경을 구성하는 것은 매우 힘들다[4]. 이처럼 거대 언어 모델이 발전하고 있음에도 불구하고 막대한 컴퓨팅 자원으로 인해 모델을 직접 개발하거나 운영하기는 쉽지 않다.

이러한 특성으로 인해, 오픈 소스 기반의 거대 언어 모델 프로젝트의 발전이 더뎠으나, 최근 다양한 경량화 기법의 발전으로 대규모 서버 기반의 모델부터 데스크톱 수준의 모델까지 다양한 크기의 모델이 연구되고 있으며, 더욱 자세한 내용은 거대 언어 모델과 관련된 서베이(Survey) 논문[5]에서 확인할 수 있다.

2.2 거대 언어 모델의 발전

2.2.1 거대 언어 모델 변천사

Zhao 등의 연구[5]에 따르면, 언어 모델의 연구는 N-gram 등으로 대표되는 통계적 학습 방법을 기반으로 한 통계적 언어 모델(Statistical Language Model), word2vec 등으로 대표되는 순환 신경망(Recurrent Neural Network)와 같은 신경망 기반의 신경 언어 모델 (Neural Language Model) 등의 순으로 발전되어 왔다. 이때까지는 언어 모델의 연구에서 “거대 언어 모델”이라는 표현은 사용되지 않았다.

2017년에는 거대 언어 모델의 출발점인 트랜스포머(Transformer) 아키텍처[6]가 공개됐다. 트랜스포머는 병렬 처리를 가능하게 하는 어텐션 메커니즘(Attention Mechanism)을 도입하여 자연어 처리 분야에서 뛰어난 성능으로 평가받고 있다.

2018년에는 Google이 BERT(Bidirectional Encoder Representations from Transformers)를 공개했다[7]. BERT는 인코더 전용 Transformer 아키텍처를 기반으로 단어를 양방향으로 해석하여 컨텍스트 표현을 학습한다. OpenAI은 GPT(Generative Pre-trained Transformer)을 공개했다[8]. GPT는 디코더 전용 Transformer 아키텍처를 기반으로 개발되었으며 비지도 학습 방식이다. 2019년에는 OpenAI가 GPT-2를 공개했다[9]. 이전 버전인

GPT와 비슷한 구조로 파라미터를 10배 늘린 15억 개며, 대규모 웹 페이지 데이터 세트인 WebText로 훈련했다. Google은 T5(Text-to-Text-Transfer Transformer)를 공개했다[10].

2020년에는 OpenAI가 GPT-3을 공개했다[11]. GPT-3는 1,750억 개(175B)의 매개변수로 약 1조 개 데이터로 학습되었다. 언어 관련 문제풀이, 번역, 코딩이 가능하며 인간과 비슷한 성능으로 문장을 생성한다. 2021년에는 Google의 LaMDA[12]와 EleutherAI의 GPT-Neo[13], GPT-J[14]가 공개되었으며, 2022년에는 OpenAI의 ChatGPT[15], GPT-3.5[16]와 Huggingface의 BLOOM[16]와 EleutherAI의 GPT-NeoX[17]와 Meta의 OPT[18], OPT-IML[19], Galatica[20]와 Google의 PaLM[21]가 공개되었다. 2023년에는 Meta의 LLaMA[22], LLaMA2[23]와 OpenAI의 GPT-4[24], Amazon의 Titan[25], Google의 Bard[26], PaLM2[27]와 Stanford의 Alpaca[28]와 MosaicML의 MPT[29], TII의 Falcon[30]과 Databricks의 Dolly2.0[31]와 EleutherAI의 Pythia[32], Vicuna[33]가 공개되었다.

거대 언어 모델의 개발은 기본적으로 매개변수 개수를 늘려 모델의 크기를 키움과 동시에, 학습에 사용되는 데이터의 양과 질을 늘림으로써 거대 언어 모델의 성능을 향상시키는 방향으로 연구 개발이 진행 중이다.

2.2.2 거대 언어 모델 학습

거대 언어 모델은 대용량의 텍스트 데이터 수집, 데이터 전처리 작업, 대규모 컴퓨팅, 시간 등의 리소스로 인해 모델을 처음부터 직접 구축하기에는 한계가 있다. 따라서 사전에 방대한 양의 텍스트로 학습한 사전학습 모델(Pretrained model)을 기반 모델(Foundation model)로 두고 이를 미세조정(Finetuning)하여 미세조정 모델(Finetuned model)을 만든다[34].

기반 모델(Foundation model)은 대규모 데이터로 사전 훈련되어 추가적인 훈련 없이 특정 처리 작업에 사용할 수 있도록 설계된 모델이다. 사전학습 모델(Pretrained model)은 대표적으로 BERT, GPT, LLaMA가 있으며 미세조정 모델(Finetuned model)은 Alpaca, Vicuna 등이 있다.

모델을 미세조정하는 방법은 별도의 작업(Task) 및 도메인 데이터를 활용하여 특정 목적에 맞게 모델을 추가로 학습시키는 것이다. Low-Rank Adaptation(LoRA)[35]과 같은 PEFT (Parameter-Efficient Fine-Tuning)기법[36]으로 작은 컴퓨팅 자원을 사용하여 미세조정한다. 이러한 미세조정 기법을 활용하여 비교적 적은 비용으로 다양한 분야에 적용할 수 있는 거대 언어 모델을 만들고 있다.

3. 오픈 소스 기반 거대 언어 모델

3.1 오픈 소스 기반 거대 언어 모델

3.1.1 오픈 소스 기반 거대 언어 모델 종류

오픈 소스 기반의 거대 언어 모델은 누구나 접근할 수 있고, 활용할 수 있어 비약적으로 발전하고 있다. 오픈 소스를 활용함으로써 기술 격차가 심화하거나 특정 조직에 의해 독점 및 권력화되는 등의 문제를 줄일 수 있으며 거대 언어 모델을 지속해서 관리, 개선해 나갈 수 있다. 또한, 최근에는 성능이 뛰어나면서 연구 목적뿐만 아니라 실제 서비스에 적용할 수 있는 상업적으로 사용이 가능한 모델들도 공개되고 있다. 표 1은 다양한 오픈 소스 기반 거대 언어 모델들의 개발 주체, 공개 날짜 및 상업적 사용 가능 여부를 나타낸다.

표 1. 오픈소스기반 거대 언어 모델의 상업적 이용 가능 여부
Table 1. Commercial Use of Open-source Large Language Models

Date	Model	Developer	Commercial Use
2022.05	OPT	Meta	X
2022.07	BLOOM	HuggingFace collaboration	O
2022.11	Galactica	Meta	X
2023.02	LLaMA	Meta	X
2023.03	Alpaca	Stanford	X
2023.03	Vicuna	LMSYS	X
2023.03	GPT4all	NomicAI	X
2023.04	Dolly	Databricks	O
2023.05	MPT	MosaicML	O
2023.06	Falcon	MosaicML	O
2023.07	LLaMA2	Meta	O

LLaMA는 Meta에서 공개한 언어 모델로 7B, 13B, 30B, 65B의 총 4가지 크기가 있다. LLaMA는 OpenAI의 GPT3(1,750억 개 파라미터; 175B)나 Google LLM보다 적은 매개변수를 가지지만, 고품질의 데이터 훈련으로 효율성을 높여 훨씬 적은 컴퓨팅 파워로도 뛰어난 성능을 낼 수 있어 주목받고 있다. 매개변수의 양을 키우기보다는 훈련 데이터의 양을 늘려 성능을 증가시켰다.

이후 공개된 LLaMA2는 7B, 13B, 70B인 총 3가지 크기 모델이 있다. 해당 모델은 2조 개의 토큰으로 훈련되었으며 이는 기존 LLaMA에 비해 40% 더 많은 데이터로 학습되었다. LLaMA2 는 기존의 오픈 소스 거대 언어 모델보다 성능이 우수하며 상업적으로도 이용할 수 있으므로, 가장 최근에 공개된 모델임에도 불구하고 활발하게 연구되고 있다.

Alpaca는 스탠포드 대학(Stanford University)의 한 연구실에서 공개한 LLaMA 7B 모델을 미세조정한 지시(Instruct) 기반 모델이다. GPT3.5 (text-davinci-003)를 사용하여 생성한 5만 2천여 개의 학습데이터로 미세조정(Finetuning)하였다. text-davinci-003과 비슷한 성능을 보이지만, 크기가 작고 저렴하게 학습 가능하다는 장점이 있다.

Vicuna는 University of California Berkeley, University of California San Diego, Carnegie Mellon University, Mohamed Bin Zayed Univ. of AI가 공동으로 개발한 오픈 소스 챗봇으로 LLaMA를 미세조정한 모델이다. 13B 매개변수이며 ShardGPT에서 수집한 약 70K 개의 사용자 대화로 학습되었다. GPT-3로 평가했을 때 기준으로, OpenAI ChatGPT 및 Google Bard와 비교하였을 때, 90% 이상의 성능을 달성했다.

Falcon은 Abu Dhabi의 Technology Innovation Institute에서 개발한 오픈 소스 거대 언어 모델이며, 매개변수의 개수는 각각 7B, 40B이다. 1조 개의 토큰으로 학습되었으며 상업적으로 이용할 수 있다.

Bloom은 HuggingFace의 창립자를 포함하여 1,000명 이상의 전문가가 참여한 공동 프로젝트의 결과로 공개한 다국어 언어 모델이며, GPT-3와 동등한 1,760억 개의 매개변수를 가지고 있다.

3.1.2 오픈 소스 기반 거대 언어 모델 비교

매일 수많은 거대 언어 모델이 공개되고 있으나, 각각의 모델들을 비교하거나 성능을 객관적으로 파악하기 어렵다. 이를 위해 인공지능 연구가 활발하게 공유되고 있는 HuggingFace에서는 거대 언어 모델들을 비교할 수 있는 다양한 리더보드들이 제공되고 있다. HuggingFace는 자연어 처리, 이미지 생성 모델 등 인공지능과 관련된 도구와 라이브러리를 제공하는 플랫폼이다. HuggingFace는 모델, 학습 데이터 등을 업로드할 수 있는 저장소와 모델의 정보 및 성능에 대해 다른 사람과 공유할 수 있는 커뮤니티 역할을 한다.

HuggingFace의 Open LLM Leaderboard[37]에는 ARC(AI2 Reasoning Challenge), HellaSwag, MMLU(Multitask Multilingual Understanding Evaluation), TruthfulQA이 포함된 4개의 벤치마크로 구성되어 모델들의 성능을 비교할 수 있다. 각 벤치마크의 구성은 다음과 같다.

- ARC: 다지선다 문제에서 언어 모델의 과학 추론 및 질문 답변 능력을 평가함
- HellaSwag: 언어 모델의 추론력을 확인하기 위한 것으로 사람은 쉽게 맞출 수 있는 질문이지만, 언어 모델은 잘 맞추지 못하는 테스트 문항들에 대한 평가함
- MMLU: 언어 모델이 다양한 작업과 다양한 도메인에서 얼마나 주어진 문제를 잘 해결할 수 있는지 평가함
- TruthfulQA: 사람이 작성한 글에서 발견되는 잘못된 내용으로 인해 발생하는 잘못된 답변 생성을 얼마나 잘 방지할 수 있는지 평가함

표 2는 위에서 언급된 벤치마크들을 활용한 F1-score의 평균(Average)을 기준으로 상위 10개 모델이 무엇인지 나타낸다. 2023년 7월 27일을 기준으로 벤치마크 성능 상위 10개의 모델 중 사전학습 모델(Pretrained model)은 1개, 미세조정 모델(Finetuned model)은 9개이다. 학습 모델은 Meta의 LLaMA 2 모델이며, 9개의 미세조정 모델은 모두 LLaMA를 기반 모델(Foundation model)로 두고 미세조정한 모델이다. 잘 설계된 아키텍처와 고품질 데이터 세트로 학습된 LLaMA2 뿐만 아니라, 이를 활용하

여 파생된 모델들도 우수한 성능을 보인다.

특히, LLaMA2 모델이 공개된 지 오랜 시간이 지나지 않았음에도 다양한 미세조정 모델들이 업로드된 것으로 보았을 때, 거대 언어 모델에 관한 연구가 활발하게 진행되고 있음을 알 수 있다. LLaMA2 모델은 우수한 성능뿐만 아니라 상업적 이용할 수 있으므로 활용 가치가 높아 앞으로 더욱 다양한 파생 모델이 공개될 것으로 예상된다.

3.2 오픈 소스 기반 한국어 거대 언어 모델

다양한 언어를 학습한 다국어 언어 모델이 공개되고 있으나 기존의 거대 언어 모델과 다국어 거대 언어 모델의 학습데이터는 영어에 편중되어 있어서 한글을 포함한 비영어권 언어에 대한 성능에서는 만족스럽지 못한 모습을 보인다. 따라서, 한국어에 특화된 모델을 만들기 위해 BERT, GPT, LLaMA와 같은 대규모 언어로 학습된 모델을 기반 모델(Foundation model)로 두고 한국어 데이터 세트를 재학습 시키는 미세조정 모델(Finetuned model)들이 만들어지고 있다. 본 논문에서는 오픈 소스로 공개되어 연구가 활발히 진행되고 있는 LLaMA 모델과 Polyglot 모델을 기점으로 파생된 모델들을 소개한다.

표 1. 오픈소스기반 거대 언어 모델의 상업적 이용 가능 여부
Table 2. HuggingFace Open LLM Leaderboard as of 27 July 2023

Model	Type	Foundation Model	Average (F1-score)
stabilityai/FreeWilly2	finetuned	LLaMA2	71.4
jondurbin/airoboros-l2-70b-gpt4-1.4.1	finetuned	LLaMA2	70.9
TheBloke/llama-2-70b-Guanaco-QLoRA-fp16	finetuned	LLaMA2	70.6
stabilityai/FreeWilly1-Delta-SafeTensor	finetuned	LLaMA1	68.7
TheBloke/gpt4-alpaca-lora_mlp-65B-HF	finetuned	LLaMA1	68.2
meta-llama/Llama-2-70b-hf	pretrained	-	67.3
upstage/llama-30b-instruct-2048	finetuned	LLaMA1	67
jondurbin/airoboros-65b-gpt4-1.2	finetuned	LLaMA1	67
TheBloke/guanaco-65B-HF	finetuned	LLaMA1	66.9
meta-llama/Llama-2-70b-chat-hf	finetuned	LLaMA2	66.8

Polyglot-ko[38]는 비영리 AI 연구단체인 EleutherAI에서 만든 대규모 한국어 자기회귀(auto-regressive language) 언어 모델이다. Polyglot 모델을 기반으로 하며 1.3B, 3.8B, 5.8B, 12.8B 매개변수들을 가지는 모델이 공개되어 있다. TuNiB AI에서 수집한 1.2TB 규모의 한국어로 학습되었으며 해당 데이터의 수집 방법은 한국 법률을 준수하며 Polyglot-ko 모델 학습을 목적으로 수집되었기 때문에 데이터 세트는 공개되지 않았다. Polyglot-ko-12.8B는 GPT-NeoX 프레임워크를 사용하여 256개의 A100 GPU에서 301,000단계에 걸쳐 1,670억 토큰에 대해 훈련되었다.

KoAlpaca[39]는 LLaMA 7B, 13B, 30B, 65B를 미세조정한 모델이다. 데이터 세트는 기본적으로 Stanford Alpaca에서 제공한 5만 2천 개 데이터 세트를 기반으로 한다. KoAlpaca를 학습시키기 위한 데이터 세트는 다음과 같은 방식으로 만들어졌다. 먼저, Alpaca 모델의 데이터 세트를 번역한다. Alpaca에서 제공한 데이터 세트는 Instruction, Input, Output으로 구성되어 있다. 그중 Instruction과 Input을 DeepL API 서비스를 사용하여 번역한다. Output은 OpenAI의 text-davinci-003 모델을 사용하여 생성한 데이터이기 때문에 별도로 번역하지 않는다. 그다음, 번역한 Instruction, Input으로 OpenAI text-davinci-003 모델을 사용하여 Output을 생성한다. 답변을 생성할 때 별도의 프롬프트(Prompt)를 제작하여 답변을 생성하였다.

Koalpaca-Ploylglot 5.8B, 12.8B은 EleutherAI의 Polyglot-ko를 기반 모델(foundation model)로 한 미세조정 모델(finetuned model)이며 자체 수집한 v1.1 데이터 세트로 학습되었다. 이 데이터 세트는 네이버 지식인 서비스의 베스트 질문 전체를 크롤링한 데이터이다.

KoVicuna[40]는 Vicuna 7B 모델을 한국어 데이터로 학습시킨 미세조정 모델이다. Vicuna를 학습시킨 ShareGPT 데이터 세트 62만 개의 대화문을 DeepL로 번역하여 학습시켰으며 Nvidia A100 GPU 8개로 15시간 동안 학습을 진행했다.

KULLM(구름)[41]은 고려대학교 NLP & AI 연구실

과 HIAI 연구소가 개발한 한국어 거대 언어 모델이다. KULLM은 기반 모델로 Polyglot-ko을 사용하여 학습되었으며, 모델 크기와 학습된 데이터 세트에 따라 3가지의 모델이 공개되어 있다. GPT4ALL 데이터 세트만으로 학습된 kullm-ployglot-12.8B-v1과 GPT4ALL, Dolly, Vicuna의 3가지를 병합한 데이터 세트로 학습된 kullm-ployglot-5.8B-v1 및 kullm-ployglot-12.8B-v2 모델이다. 학습된 사용된 데이터 세트는 모두 DeepL을 이용하여 한국어로 번역하였다. kullm-ployglot-12.8B-v1은 총 5 epoch 학습하였고, Nvidia A100 80GB 4대가 사용되었다. kullm-ployglot-12.8B-v2는 총 8 epoch 학습하였고, Nvidia A100 80GB 4대가 사용되었다.

표 3은 앞서 소개한 5가지 한국어 거대 언어 모델을 요약 정리한 것이다. 공개된 한국어 전용 거대 언어 모델을 살펴보면 대부분 기존의 거대 언어 모델이 학습한 데이터를 한국어로 번역하여 추가 학습시키는 방식이다. 이외에도 다른 모델들의 학습데이터로 학습하거나 Lora와 같은 미세조정 기법을 사용하여, 보다 적은 자원으로 미세조정하면서 계속해서 발전시키고 있다.

표 3. 한국어 기반 거대 언어 모델
Table 3. List of Korean LLMs

Model	Foundation Model	Parameters	Training Datasets
Polyglot-ko	Polyglot	1.3B, 3.8B, 5.8B, 12.8B	-
KoAlpaca	LLaMA	7B, 13B, 30B, 65B	Translated Alpaca training dataset
Koalpaca-Polyglot	Polyglot-ko	5.8B, 12.8B	Best questions from Naver Knowledge iN
KoVicuna	Vicuna	7B	Translated shareGPT training dataset
KULLM	Polyglot-ko	5.8B, 12.8B	GPT4ALL, Dolly, and Vicuna datasets

표 4는 HellaSwag 벤치마크를 활용한 10-shot 학습조건에서 성능 결과가 공개된 한국어와 영어의 거대 언어 모델 성능을 비교한 내용을 나타낸다 [24], [37], [38]. 앞선 표 2와 같이 결과는 2023년 7월 27일을 기준으로 정리되었다.

표 4. HellaSwag (10-shot) 기반 영어 및 한국어 성능표 (2023년 7월 27일 기준)
Table 4. English and Korean LLMs using HellaSwag benchmark (10-shot) as of 27 July 2023

Model	Target Lang.	HellaSwag (F1-score)
GPT-4	Multi.	95.3
GPT3.5	Multi.	85.5
LLaMA-65B	Eng.	84.2
LLaMA-30B	Eng.	82.6
LLaMA-13B	Eng.	78.9
Alpaca-13B	Eng.	77.6
LLaMA-7B	Eng.	75.6
polyglot-ko-12.8b	Kor.	60.98
polyglot-ko-5.8b	Kor.	59.79
polyglot-ko-3.8b	Kor.	56.7
polyglot-ko-1.3b	Kor.	52.78
KoAlpaca-Polyglot-5.8B	Kor.	35.6

표 4에 따르면, HellaSwag 벤치마크 10-shot 학습조건을 기준으로 가장 높은 점수를 보인 것은 OpenAI의 GPT-4와 GPT-3.5이다. 두 모델은 상용 서비스로 제공되고 있는 언어 모델이며 한국어와 영어뿐만 아니라 다양한 언어를 지원하는 모델이다.

한편, 한국어 기반 거대 언어 모델은 아직 영어 기반 거대 언어 모델과 비교하면 비교적 낮은 성능을 보여주었다. 기본적으로 한국어 언어 모델 연구는 매개변수 수가 13B를 넘기지 못하고 있어, 좋은 성능을 보여주는 데 한계가 있음을 알 수 있다.

한편, 매개변수 크기가 13B로 비슷한 환경에서도 영어모델인 LLaMA-13B(F1-score: 78.9)와 Alpaca-13B(F1-score: 77.6) 모델들이 비슷한 크기의 polyglot-ko-12,8b(F1-score: 60.98) 모델보다 F1-score 기준으로 10점 이상 앞서는 것을 알 수 있다. 이는 모델을 학습시키기 위한 한국어 데이터 세트의 규모가 영어의 그것에 비해 너무 작고, 질 좋은 데이터 세트를 수집하기도 어려움이 있으며, 한국어 토큰화 과정이 언어의 특성 상, 영어의 토큰화 과정과 비교하면 더 어렵고 비효율적이기 때문이다.

한국어 거대 언어 모델의 성능을 향상하기 위해서는 언어 모델 자체의 크기, 데이터 세트의 양과 품질, 그

리고 한국어 토큰화 과정의 효율화 등을 지속해서 개량할 필요가 있다.

4. 거대 언어 모델의 활용과 한계

4.1 거대 언어 모델의 한계

거대 언어 모델은 자연어 처리에 있어서 혁신적이지만 동시에 다양한 문제들이 존재한다. 거대 언어 모델을 효과적으로 훈련하기 위해서는 방대한 양의 고품질 데이터가 필요한데, 전문 도메인이나 특정 언어에 대한 접근이 어려울 수 있으며 이러한 데이터를 수집하고 전처리하는데 많은 시간과 자원이 소모된다. 또한, 데이터를 학습할 때 인종, 나이, 성별을 차별하거나 혐오나 선정적인 표현을 포함한 비윤리적이거나 잘못된 정보가 있는 데이터를 학습할 경우 이를 그대로 반영하거나 사용자의 의도와 일치하지 않거나 거짓된 정보를 생성하는 환각 현상을 보일 수도 있다[44]. 또한, 거대 언어 모델은 매개변수 개수가 매우 많아 학습과 추론에 많은 컴퓨팅 자원이 필요하고 이를 실제 서비스에 적용하여 운영하는 비용도 고려해야 한다.

4.2 거대 언어 모델의 활용과 미래

거대 언어 모델은 자연어에 대한 이해와 문장 생성 능력이 사람과 비슷할 정도로 뛰어나 의료, 금융, 마케팅 등 다양한 산업 분야에서 활용될 수 있다. 거대 언어 모델로 대화형 챗봇을 구축할 수 있어 자연어로 질의응답을 통해 고객 지원을 하거나, 업무 보조 도구로 활용할 수 있고 다양한 언어로 학습되어 있으므로 번역 작업도 할 수 있다. 또한, 고품질의 텍스트 생성할 수 있어 기사, 소설, 시나리오 등 콘텐츠를 제작할 수도 있으며 내용을 원하는 길이와 형태로 요약할 수도 있다. 코드를 학습한 모델은 텍스트뿐만 아니라 자연어를 기반으로 코드도 생성할 수 있다.

특히, 충분히 큰 거대 언어 모델은 텍스트나 콘텐츠 생성뿐만 아니라 그 원래 목적과는 다른 일반화된 문제를 해결하는데 사용될 수도 있는 것으로 알려져 있다. 예를 들어, 사칙연산에 관한 내용을 학습하지 않은 거대 언어 모델을 활용하여 사칙연산을 하거나, 체스판의 상황을 텍스트로 알려주고 다음 말의 움직임을 예측하거나 등의 다양한 현상을 일반 거대 언어 모델이 풀 수 있음이 밝혀지고 있다[42]. 이러한 현상은 확대되어 인간이 할 수 있는 어떠한 지적인 업무도 성공적으로 해낼 수 있는 기계의 지능인 “인공일반지능(Artificial general intelligence; AGI)”의 개발을 목표로 연구가 진행되고 있다[43].

한편, LLM 훈련에 너무 많은 비용과 자원이 소비되기 때문에, 같은 성능을 낼 수 있으면서, 크기가 작은(즉, 매개변수 수가 적은) LLM 모델을 만들기 위하여 다양한 방법들이 연구되고 있다[5]. 예를 들어, Meta의 LLaMA의 경우 65B 매개변수를 가진 모델이 GTP-3.5의 135B 매개변수를 가진 모델보다 더 좋은 성능을 보였는데, LLaMA의 경우 더 고품질의 훈련 데이터 세트를 사용함으로써 약 절반의 매개변수로 더 좋은 성능을 나타내었다[22].

또한, 추론용 LLM 모델에 필요한 메모리 크기를 줄이기 위하여 매개변수의 정밀도(precision)를 낮추면서 같은 성능을 낼 수 있도록 하는 양자화(quantization) 기법들도 다양하게 연구되고 있다[5]. LLM 모델에 사용되는 대표적인 양자화 기법은 16-bit 및 8-bit 양자화 기법으로 일반적인 LLM 모델의 정밀도가 32-bit인 것을 고려할 때, 이론적으로 각각 50%와 25%의 메모리만을 필요로 한다. 예를 들어, LLaMA-65B의 경우 16-bit 모델이 약 90~100GB의 메모리를 요구하지만, 8-bit 모델로 양자화할 경우 그의 절반인 40~50GB만을 요구하기 때문에 응용 관점에서는 요구하는 컴퓨팅 자원의 양이 획기적으로 줄어듦을 알 수 있다.

이렇듯 거대 언어 모델은 그 넓은 활용성을 바탕으로 성능 향상과 비용 최적화를 위한 다양한 방식들이 연구되고 있다.

5. 결론

본 논문에서는 오픈 소스를 중점으로 거대 언어 모델과 한국어 특화 언어 모델에 관한 연구 개발 동향을 살펴보았다. 대규모 데이터로 학습된 거대 언어 모델은 뛰어난 언어 생성 능력과 다양한 산업의 활용 가능성으로 최근 들어서 많은 주목을 받고 있다. 거대 언어 모델이 학습 및 추론하기 위해서는 막대한 컴퓨팅 자

원이 필요하므로, 빅테크 기업들에 의해 주도적으로 연구되었다. 이후 사전 학습된 모델들을 오픈 소스로 공개하는 개방적인 움직임으로 인해 모델, 데이터 세트, 학습 기법 등 모델의 성능을 개선하기 위한 연구가 활발하게 이뤄지고 있으며 오픈 소스 생태계가 활성화되었다. 공개된 언어 모델을 활용하여 한국어 데이터 세트로 추가 학습하는 한국어 특화 모델도 공개되고 있지만, 영어모델과 비교해서는 성능이 부족하므로 이를 극복하기 위해서는 한국어 데이터 세트와 전처리, 평가 방법에 대해 활발하게 연구될 필요성이 있다. 한편, 거대 언어 모델의 성능 향상과 비용 효율화를 동시에 추구하기 위해, 새로운 방식의 모델 연구뿐만 아니라 데이터 최적화 및 양자화 등 다양한 방식들이 연구되고 있다. 이러한 오픈 소스 기반의 거대 언어 모델은 상용 서비스 환경에서뿐만 아니라 자가구축(On-premise)형 환경에서도 실행할 수 있으므로, 데이터의 유출에 대한 걱정 없이 학습시킬 수 있어 다양한 서비스에 활용될 것으로 기대된다.

REFERENCES

[1] M. Shanahan, “Talking about large language models,” CoRR, vol. abs/2212.03551, 2022.

[2] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” CoRR, vol. abs/2001.08361, 2020.

[3] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, “Training compute-optimal large language models,” vol. abs/2203.15556, 2022.

[4] W. Fedus, B. Zoph, and N. Shazeer, “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” J. Mach. Learn. Res, pp. 1-40, 2021.

[5] Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223, 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, pp. 5998-6008, 2017.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, p. 4171-4186, 2019.

[8] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al., “Improving language understanding by generative pre-training,” 2018.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., “Language models are unsupervised multitask learners,” OpenAI blog, p. 9, 2019.

[10] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, and J. Schulman, “Webgpt: Browser-assisted question-answering with human feedback,” CoRR, vol. abs/2112.09332, 2021.

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. Mc- Candlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in Advances in Neural Information Processing

Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[12] R. Thoppilan, et al., "Lamda: Language models for dialog applications," CoRR, vol. abs/2201.08239, 2022.

[13] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman, "GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow," 2021.

[14] B. Wang and A. Komatsuzaki, "GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model,"https://github.com/kingoflolz/mesh-transformer-jax, 2021.

[15] "Introducing chatgpt," OpenAI Blog, November 2022.

[16] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," arXiv preprint arXiv:2303.10420, 2023.

[17] V. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeybi, and B. Catanzaro, "Reducing activation recomputation in large transformer models," CoRR, vol. abs/2205.05198, 2022.

[18] E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, Y. Zhou, S. Savarese, and C. Xiong, "Codegen: An open large language model for code with mtulti-turn program synthesis," arXiv preprint arXiv:2203.13474, 2022.

[19] S. Iyer, X. V. Lin, R. Pasunuru, T. Mihaylov, D. Simig, P. Yu, K. Shuster, T. Wang, Q. Liu, P. S. Koura, X. Li, B. O'Horo, G. Pereyra, J.Wang, C. Dewan, A. Celikyilmaz, L. Zettlemoyer, and V. Stoyanov, "OPT-IML: scaling language model instruction meta learning through the lens of generalization," CoRR, vol. abs/2212.12017, 2022.

[20] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," CoRR, vol. abs/2211.09085, 2022.

[21] A. Chowdhery, et al., "Palm: Scaling language modeling with pathways," CoRR, vol. abs/2204.02311, 2022.

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozi`ere, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, an G. Lample, "Llama: Open and efficient foundation language models," CoRR, 2023.

[23] H. Touvron, et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models", arXiv:2307.09288, 2023

[24] OpenAI, "Gpt-4 technical report," OpenAI, 2023.

[25] S. Wang, Y. Sun, Y. Xiang, Z. Wu, S. Ding, W. Gong, S. Feng, J. Shang, Y. Zhao, C. Pang, J. Liu, X. Chen, Y. Lu, W. Liu, X. Wang, Y. Bai, Q. Chen, L. Zhao, S. Li, P. Sun, D. Yu, Y. Ma, H. Tian, H. Wu, T. Wu, W. Zeng, G. Li, W. Gao, and H. Wang, "ERNIE 3.0 titan: Exploring larger-scale knowledge enhanced pretraining for language understanding and generation," CoRR, vol. abs/2112.12731, 2021.

[26] James Manyika, "An overview of Bard: an early experiment with generative AI", 2023

[27] Google, "PaLM 2 Technical Report", 2023

[28] https://crfm.stanford.edu/2023/03/13/alpaca.html

[29] https://www.mosaicml.com/blog/mpt-7b

[30] https://huggingface.co/tiiuae/falcon-40b

[31] https://www.databricks.com/blog/2023/03/24/hello-dolly-democratizing-magic-chatgpt-open-models.html

[32] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff et al., "Pythia: A suite for analyzing large language models across training and scaling," arXiv preprint arXiv:2304.01373, 2023.

[33] LMSYS The Vicuna Team, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," 2023.

[34] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A.

Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. Mc- Candlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[35] E. J. Hu, Y. Shen, P.Wallis, Z. Allen-Zhu, Y. Li, S.Wang, L.Wang, andW. Chen, "Lora: Low-rank adaptation oflarge language models," in The Tenth International Conferenceon Learning Representations, ICLR 2022, 2022.

[36] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, and S. Paul, "Peft: State-of-the-art parameter-efficient fine tuning methods,", 2022.

[37] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[38] https://huggingface.co/EleutherAI/polyglot-ko-12.8b

[39] https://github.com/Beomi/KoAlpaca

[40] https://github.com/melodysdreamj/KoVicuna

[41] https://github.com/nlpai-lab/KULLM

[42] Srivastava, Aarohi, et al. "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models." arXiv preprint arXiv:2206.04615, 2022.

[43] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of artificial general intelligence: Early experiments with gpt-4," vol. abs/2303.12712, 2023.

[44] J. Li, T. Tang, W. X. Zhao, and J. Wen, "Pretrained language model for text generation: A survey," in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, Z. Zhou, Ed. ijcai.org, 2021.

저자약력

주 하 영 (Hayoung Joo)

- 2021 2월: 인제대학교 BNIT대학 헬스케어IT학과 학사
- 2022 9월 ~ 현재: KAIST-메가존 클라우드 지능형 클라우드 융합기술 연구센터 연구원
- 2023 2월: 인제대학교 일반대학원 헬스케어IT공학 석사

〈관심분야〉 LLM, Generative AI, 클라우드

오 현 택 (Hyeontaek Oh)

- 2012년 2월: 한국과학기술원 전산학 (학사)
- 2014년 2월: 한국과학기술원 전기 및 전자공학 (석사)
- 2020년 2월: 한국과학기술원 전기 및 전자공학 (박사)
- 2020년 2월 ~ 현재: 한국과학기술원 IT융합연구소 지능화기술연구팀 팀장

〈관심분야〉 ICT트러스트, 마이데이터, 개인정보 생태계

양 진 홍 (Jinhong Yang)

- 2017년 2월 : KAIST 정보통신 공학 박사
- 2017년 2월~2018년 1월 : HECAS 최고기술책임(CTO)
- 2017년 10월~현재: 한국과학기술원 IT융합연구소 겸직교수
- 2018년 3월~현재 : 인제대학교 의료IT 학과 조교수

〈관심분야〉 데이터 컴플라이언스, 마이데이터, 헬스케어 데이터 활용