한국어 초거대언어모델 평가척도의 비교 분석

심주용, 이서영, 정다은, 김성환
국립한국교통대학교
stlaalsdyds@ut.ac.kr, 0721se@ut.ac.kr. jde3772@ut.ac.kr, seonghwan.kim@ut.ac.kr

Comparative Analysis of Evaluation Metrics for Korean Large Language Models

Ju Yong Sim, Seo Yeong Lee, Da Eun Jung, Seong Hwan Kim
Korea National University of Transportation

요 약

최근 국내의 여러 기업과 기관에서 한국어 기반의 오픈 소스 생성형 언어모델을 경쟁적으로 개발하고 있다. 우리는 한국어 언어모델의 평가에 주로 사용되는 평가척도들을 소개한다. 또한 각 평가 척도의 특성을 분석하고 평가척도들 간의 차이점이 드러나도록 비교 분석한다. 이를 통해 한국어 기반의 생성형 모델을 특정 목적에 맞게 미세 조정할 때, 해당 목적에 맞는 평가척도를 선택하고, 선택한 평가척도의 점수를 향상시키는 데 유용한 정보를 제공한다.

Ⅰ. 서 론

2022년 말에 출시된 초거대 언어모델인 ChatGPT-3.5는 자연스러운 대화 형식의 프롬프트를 텍스트 창에 입력하면 수 초 내로 전문가 수준에 가까운 답을 내놓아 전 세계적으로 큰 주목을 받았다. 뿐만 아니라 두 달 만에 서비스 가입자의 수가 1억 명에 달하는 파급력을 보여주었다. 현재는 각 분야의 전문가부터 대학생에 이르기까지 일상에서 널리 활용되고 있다. ChatGPT의 출현은 자연어 처리의 한 분야인 초거대 언어모델(Large-language model, LLM)의 발전에 기인한다. LLM의 활용은 기업과 개인의 생산성을 크게 증가시킬 것으로 예상되어 빅테크 기업뿐만 아니라 정부 기관들과 많은 중소기업들이 LLM의 도입을 고려하고 있다. 그러나 무료의 ChatGPT 서비스와 달리, LLM을 다른 프로그램과 연동하기 위해서는 유료 API (application programming interface)를 사용해야 하며, 이러한 추가 비용은 LLM의 도입을 어렵게 하는 요인이 되고 있다. 또한 기관이 내부 정보나 민감한 고객 정보를 외부 서버에 전송할 경우 기밀 유출에 대한 우려가 발생함에 따라 일부 기업에서는 직원들의 ChatGPT 사내 사용을 금지하고 있다.

이에 대한 대안으로 등장한 것이 메타에서 공개한 Llama 와 미스트랄 AI에서 공개한 Mistral 과 같은 오픈소스 기반의 LLM이다. 이러한 LLM들은 작게는 수십억에서 많게는 수백억 개의 매개변수를 갖고 있다. 이는 놀라운 창발능력을 보여주는 GPT3.5가 1750억 개의 매개변수를 갖는 것과 비교하면 크기 면에서 매우 작음을 알 수 있으며, 일반적으로 오픈소스 기반의 LLM은 GPT-3.5와 그 이상 수준의 LLM의 성능에 미치지 못하는 것으로 알려져 있다. 그러나 최근 크기가 작더라도 고품질의 데이터 집합으로 LLM을 미세조정할 경우 그 성능이 크게 증가할 수 있다는 관점이 주목받고 있으며 실제로 국내 기업 업스테이지가 발표한 수학 분야에 특화된 MathGPT라는 모델은 130억개에 불과한 매개변수로 GPT-4를 능가하는 성능을 최근에 보여주었다 [1]. 다른 여러 국내의 중소기업들도 사전 학습된 LLM을 특정 목적에 맞게 미세조정하여 자체적인 경량급 LLM을 개발하려는 시도가 이어지고 있다.

오픈소스 기반의 LLM 개발은 전 세계적으로 경쟁이 이루어지고 있으며, 허깅페이스가 운용하는 OpenLLM 리더보드를 통해 개발된 LLM들의 성능과 순위가 실시간으로 공개되고 있다[2]. LLM의 성능을 평가하기 위해 사용하는 데이터 집합을 벤치마크 데이터 집합이라 하며, 영어 버전인 Open LLM 리더보드는 ARC (AI2 Reasoning Challenge), HellaSwag, MMLU (Massive Multitask Language Understanding), TruthfulQA, Winogrande, GSM8K (Grade School Math 8K) 의 벤치마크를 사용한다. 한국어 모델의 성능과 순위는 AI 허브의 Open Ko-LLM 리더보드를 통해 공개되고 있으며, 영어 버전의 Open LLM 리더보드와 달리 다음 다섯 가지의 평가척도를 사용하고 있다: Ko-ARC, Ko-HellaSwag, Ko-MMLU, Ko-TruthfulQA, Ko-CommonGenn-V2. 이 벤치마크들은 언어모델을 평가하는 기준이자 언어모델 성능 개선의 방향이 된다. 언어모델의 공정한 평가를 위해, Open Ko-LLM 리더보드에 사용되는 테스트 데이터 집합은 비공개되어 있다. 그밖에 KoBEST (Korean Balanced Evaluation of Significant Tasks) 등의 벤치마크가 있다.

본 연구에서는 언어모델 평가에 사용되는 대표적인 평가척도들을 소개하고 이들의 차이점에 대해서 설명한다. 특히 한국어 언어모델의 평가를 위한 평가척도에 주목한다. 이를 통해 특정 다운스트림 태스크에 적합한 평가척도를 선택하고 선택된 평가척도의 점수를 향상시키는 데 있어 유용한 정보를 제공한다.

Ⅱ. 기존 자연어 모델 평가척도

이 절에서는 언어모델 평가에 주로 사용되는 평가척도들을 소개하며, 특히 한국어 버전의 평가척도에 대한 현황을 설명한다.

2.1 HellaSwag

상식적 자연 언어 추론 (Commonsense natural language inference) 이란 시간의 흐름이 있는 서술을 보고 바로 다음에 나올 적절한 문장을 보기에서 고르는 작업이다. 이를 위한 SWAG (Situation With Adverasrial Generation) 벤치마크가 있었으나 BERT (Bidirectional Encoder Representations from Transformers) 모델이 인간에 가까운 수준에 도달하여 더 어려운 데이터 집합이 필요하였다. 이에 따라 HellaSwag (Harder Endings, Longer contexts, and Low shot Activities SWAG) 벤치마크가 제안되었다[3]. 한국어 버전의 HellaSwag 벤치마크는 KoBEST 데이터 집합에 포함되어 공개되어 있으며 한국어의 주어가 생략되는 특징 등을 반영하였고 4지선다 문제로 구성된다[4].

표 1 한국어 언어모델의 평가척도들

평가척도	분야	답변 형태	훈련 데이터	공개된 한국어 버전	대표적인 작업
HellaSwag	일반 상식	객관식	벤치마크에 포함	O	시간의 흐름이 있는 문장들을 보고 다음에 올 문장 예측하기
CommonGen	일반 상식	문장 생성	벤치마크에 포함	O	(사진에서 포착된) 개념들을 보고 상식에 부합하는 문장 만들기
MMLU	57개 분야	객관식	벤치마크에 포함	X	STEM, 문학, 사회과학 등의 57개 분야 문제 풀기
TruthfulQA	38개 분야	객/주관식 혼용	없음	X	다양한 분야에서 편견과 오해로 틀릴 수 있는 문제에 답하기
ARC	과학	객관식	벤치마크에 포함	X	초중등 수준의 과학 분야 문제 풀기

2.2 CommonGen

한 장의 사진에 있는 객체들과 그들의 움직임 관찰하여 사진의 캡션을 만드는 작업이 있다고 하자. 여기서 영상 인식 절차를 생략하면 생성적 상식 추론 (Generative commonsense reasoning) 작업이 되는데, 즉 몇 가지 개념 단어를 보고 상식적이고 논리적인 문장을 생성해 내는 작업이다. CommonGen 벤치마크[5]는 언어모델의 생성적 상식 추론 능력을 평가하기 위해 제안되었으며, HellaSwag과 다르게 평가 대상인 언어모델은 몇 개의 개념을 입력받고, 이 개념들을 포함하는 문장을 생성하며, 생성된 문장은 CommonGen의 모법 답안과 비교하여 성능을 평가한다. 한국어 언어모델 연구를 위한 데이터 집합들은 대부분 자연어 이해가 목적이었기 때문에 생성적 상식 추론 연구에 필요한 자원이 부족하였는데, 논문 [6]은 이런 문제를 극복하기 위해 형태소 분절 등 한국어의 특성을 반영한 생성적 상식 추론 용 벤치마크인 Korean-CommonGen을 제안하였다. Open Ko-LLM 리더보드에서 사용하는 벤치마크는 Korean-CommonGen-V2로 비공개이며 논문 [6]의 데이터 집합과는 상이하다.

2.3 TruthfulQA

TruthfulQA은 다양한 분야에서 편견과 오해로 틀릴 수 있는 질문에 언어모델이 생성한 답변이 얼마나 사실에 부합하는 지 측정하기 위한 벤치마크이며 흔히 말하는 환각현상의 정도를 측정할 수 있는 평가척도이다 [7]. 논문에서는 언어 모델의 크기가 커질수록 TruthfulQA 점수가 하락하는 현상을 밝혔다. 공개된 한국어 데이터 집합이 없으므로 한국어 언어모델 연구를 위해 한국어 버전의 TruthfulQA 데이터 집합의 구축이 필요하다. 특히 한국인의 편견이 포함된 질문을 포함해야 할 것이다.

2.4 MMLU

MMLU 는 과학, 기술, 공학, 수학, 문학 등 총 57가지 주제에 대해 평가한다[8]. 논문에 의하면 평가 당시 오직 GPT-3 만 무작위 선택 수준의 성능을 넘길 정도로 언어모델이 답하기 어려운 질문으로 구성되어 있다. 언어모델이 특히 취약한 분야는 대학수준의 화학과 물리, 고교 수학, 도덕 등이다. MMLU의 공개된 한국어 버전이 없으므로 연구를 위해서는 한국어 버전의 MMLU 데이터 집합의 구축이 필요한 상황이다. 이는 한국사, 한국 정치, 한국형 사회 과목 등을 포함해야 할 것이다.

2.5 ARC

ARC는 만 8 ~13세 수준의 과학 분야 문제 7787개로 이루어져 있으며 대부분 4지선다로 이루어져 있다. 언어모델이 쉽게 맞출 수 없는 질문의 집합인 Challenge Set과 단순한 알고리즘으로도 맞출 수 있는 질문의 집합인 Easy Set으로 구성되어 있으며, 이를 구분하는 알고리즘이 제안되었다 [9]. TruthfulQA 및 MMLU와 마찬가지로 공개된 한국어 버전이 없다. 그러나 과학 분야만을 포함하므로 ARC 데이터 집합을 한국어로 번역하는 정도의 작업으로 한국어 버전 ARC를 만들 수 있을 것이다.

Ⅲ. 평가척도들의 비교 분석

표 1은 한국어 언어모델 평가척도의 요약을 보여준다. 표는 한국어 버전 데이터집합의 공개 여부를 나타내고 있으며, 한글 버전이 공개되어 있지 않은 벤치마크들은 한국어 언어모델 연구를 위해 구축될 필요성이 있음을 알 수 있다. 벤치마크의 질문이 객관식 형태인지 주관식 형태인지를 알 수 있으며 이에 따라 채점 알고리즘이 달라질 수 있다. 각 평가척도의 점수를 개선시키기 위해서는 적절한 훈련 데이터가 필요하며 TruthfulQA를 제외한 벤치마크들은 훈련 데이터 집합을 포함하고 있다.

Ⅳ. 결론

본 연구는 한국어 언어모델의 평가에 주로 사용되는 평가척도들을 소개하고, 평가척도들 간 차이점이 드러나도록 비교 분석하였다. 이를 통해 한국어 기반의 생성형 모델을 특정 목적에 맞게 미세 조정할 때, 해당 목적에 맞는 평가척도를 선택하고 선택한 평가척도의 점수를 향상하는 데 유용한 정보를 제공한다.

ACKNOWLEDGMENT

이 논문은 인공지능 중심 산업융합 집적단지 조성사업의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] “업스테이지-콴다-KT, 챗GPT 뛰어넘는 수학 특화 언어모델 개발,” 2024 (https://www.upstage.ai/newsroom/mathgpt-no1).

[2] Chanjun Park, et al. “Open Ko-LLM Leaderboard,” Upstage, National Information Society Agency, (https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard)

[3] Rowan Zellers, et al., “HellaSwag: Can a Machine Really Finish Your Sentence?,” *in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[4] Myeongjun Jang, et al., “KoBEST: Korean Balanced Evaluation of Significant Tasks,” *in Proceedings of the 29th International Conference on Computational Linguistics*. 2022.

[5] Bill Yuchen Lin, et al. “CommonGen: A constrained text generation challenge for generative commonsense reasoning,” *In Findings of the Association for Computational Linguistics: EMNLP 2020,* 2020

[6] Jaehyung Seo, et al., “A dog is passing over the jet? a text-generation dataset for korean commonsense reasoning and evaluation,” *in Findings of the Association for Computational Linguistics: NAACL 2022,* 2022

[7] Stephanie Li, et al., “TruthfulQA: Measuring how models mimic human falsehoods,” *arXiv preprint arXiv:2109.07958.* 2021.

[8] Dan Hendrycks, et al., “Measuring massive multitask language understanding,” *in International Conference on Learning Representations*, 2020.

[9] Peter Clark, et al “Think you have solved question answering? try arc, the AI2 reasoning challenge,” *arXiv preprint arXiv:1803.05457, 2018.*