대규모 언어모델의 한국어 이해 능력 평가 방법에 관한 연구

손기준[1], 김승현[2]
[1]오피니언라이브 AI데이터센터장
[2]한국지능정보사회진흥원 책임연구원
kijunson@opinionlive.co.kr, shkim@nia.or.kr

A Study on the Evaluation Method of Korean Comprehension Abilities of Large Language Model

Ki Jun Son[1], Seung Hyun Kim[2]
[1]Dept. of AI Data, Opinionlive
[2]Dept. of AI Data, National Information society Agency

요 약

최근 GTP4, LLama와 같은 초거대 언어모델을 활용한 서비스가 공개되어 많은 사람의 주목을 받고 있다. 해당 모델들은 사용자들의 다양한 질문에 대하여 유창한 결과를 생성하고 있지만 한국어 데이터에 대한 학습량이 부족하여 한국어 이해 및 한국 문화 등에 대한 잘못된 정보를 제공하는 문제를 야기할 수 있다. 이에 본 논문에서는 한국어 데이터를 학습한 주요 공개 모델 6개를 선정하고 5개 분야(한국어 이해 및 문화 영역으로 구성)에 대한 평가 데이터셋을 구성하여 한국어 이해 능력에 대한 평가를 진행하였다. 그 결과 한국어 구사 능력은 Bookworm 모델이, 한국어 이해 및 문화와 관련한 부문은 LDCC-SOLAR 모델이 우수한 것으로 확인할 수 있었다.

1. 서 론

대규모 언어모델((Large Language Model, 이하 LLM)은 지난 5년 동안 꾸준한 성장을 보였지만, OpenAI사에서 개발한 ChatGPT 서비스의 등장 이후 세계적으로 LLM의 활용 방안에 관한 다양한 시도가 이어지고 있다. LLM은 자연어를 이해하고 또한 생성할 수 있도록 만들어진 대규모 딥러닝 모델이며, 이를 개발하기 위해서 대규모 학습데이터와 모델링 작업이 수행되어야 한다. 따라서 데이터와 자본의 한계로 해외 유명 IT 기업을 중심으로 발전하고 있어 국내 IT 기업의 노력에도 충분한 한국어 능력을 기반한 서비스 제공에 한계가 있을 수 있다는 걱정도 존재한다.

이에 본 연구에서는 한국어를 학습한 LLM 모델들을 중심으로 한국어 이해 능력을 평가하기 위해 KBS 한국어 능력 시험 · TOPIK 등 한국어와 관련된 시험 문제를 중심으로 데이터셋을 구성하고 LLM의 성능 평가를 진행하였다. 이는 LLM의 한국어 이해 및 활용 능력을 객관적으로 평가하고 서비스 제공을 위한 가장 적절한 LLM의 선정 시 기본적인 정보를 제공할 수 있다.

2. 관련 연구

2.1 LLM 개념

LLM은 대규모의 언어모델을 의미하며, 거대한 데이터셋을 사용하여 훈련된 언어모델을 의미한다. 사전에 대규모의 언어 데이터를 학습하여 문장 구조, 문법, 의미, 단어 내에 내재된 다른 의미를 이해하고 이를 활용해 사람이 말하는 것과 유사하게 자연어를 생성할 수 있다. 최근 국내 주요 기업들도 초거대 인공지능 시스템 구축에 주력하고 있으며 이러한 LLM의 필요성은 더욱 커지고 있다[1][2].

2.2 트랜스포머(Transformer) 모델

2017년 트랜스포머 아키텍처의 등장과 함께 거대

언어모델이 등장하기 시작하였으며 트랜스포머 모델의 아키텍처는 RNN, CNN을 사용하지 않고 Positional Encoding을 사용하며 다수의 인코더(Encoder)와 디코더(Decoder)와 Attention 메커니즘을 사용한다. Postional Encoding을 사용하며 Attention 과정을 여러 레이어에서 반복 수행하면서 문장 내 문맥 간의 상호 작용에 대한 정교한 모델링이 가능하게 되었다. 이를 통해 번역이나 요약과 같은 작업에서 Attention과 Nomalization 작업을 통해 성능 향상을 가져왔으며 가장 중요한 정보를 강조하여 분석할 수 있게 되었다[3].

Google은 2018년 BERT (Bidirectional Encoder Representation from Transformers)를 공개하였다. BERT는 트랜스포머 모델의 인코더 아키텍처를 기반으로 양방향 해석을 통해 텍스트를 표현하는 학습모델이다[4]. LLM은 대부분 트랜스포머 아키텍처에서 파생된 AI 모델로, 사람의 언어, 코드 등을 이해하고 생성할 수 있게 되었다.

3. LLM 모델 평가

3.1 평가 모델 및 항목 선정

최신 기술의 발전과 더불어 다양한 LLM이 공개 및 개방되고 있으며 이에 사용자가 모델들의 성능을 객관적이고 세부적으로 판단할 수 있도록 다양한 평가 방안이 제시되고 있다. 가장 대중적으로 알려진 Hugging Face Open LLM Leaderboard는 <표1>과 같이 총 6가지 분야에 대한 성능을 평가하여 순위를 정하고 있다.

<표1> Hugging Face Open LLM Leaderboard 평가 항목

	항목	내용
1	ARC	초등수준의 과학문제
2	HellaSwag	특정 상황에서의 언어 추론 능력 평가
3	MMLU	사전학습 된 모델의 지식을 평가
4	TruthfulQA	환각현상 방지 능력
5	Winogrande	상식적 추론을 평가
6	GSM8k	다단계 수학적 추론 평가

또한 한국어 모델을 위하여 NIA-업스테이지에서 운영하는 Open Ko-LLM 리더보드가 운영되고 있으며 역사 왜곡, 환각 오류, 형태소 오류, 불규칙 활용 오류, 혐오 표현 등을 고려한 상식 생성 기준을 바탕으로 한국어 사용자가 가지고 있는 일반 상식에 부합하는지를 기준으로 모델의 성능을 평가하고 있다.

먼저 평가 모델을 선정하기 위하여 2024년 2월을 기준으로 Hugging Face Leaderboard의 상위 100위 모델이며 Open Ko-LLM 리더보드의 상위 30위에 중복으로 포함된 모델을 우선 선정하였다. 그 후 공공부문에서 활용성을 고려하여 한국어가 학습된 공개 모델이며, 평가에 활용할 수 있는 인프라의 성능을 고려하여 모델 사이즈가 12.8B 이하인 모델을 기준으로 검토하여 최종적으로 <표2>에서 제시한 6개의 평가 모델을 선정하였다.

<표2> 평가에 활용할 LLM 모델

	모델명
1	LDCC-SOLAR-10.7B[6]
2	Bookworm-10.7B[7]
3	SOLARC-M-10.7B[8]
4	Llama-2-13b-chat-hf[9]
5	Kullm-solar[10]
6	KoAlpaca-Polyglot-12.8B[11]

* 모든 모델은 2024년 3월 20일 자 모델을 다운로드함

SOLAR 모델은 AI 스타트업인 업스테이지에서 자체 개발한 LLM 모델이며 Hugging Face Leaderboard에서 1위에 오른 적이 있는 모델이다. LDCC-SOLAR 모델의 경우 롯데정보통신에서 SOLAR 모델을 바탕으로 자체 구축데이터를 추가 학습하여 만들어진 모델이며, SOLARC-M 모델의 경우 SOLAR-Instruction 모델에 Merge 기술을 이용하여 모델을 최적화한 모델이다.

Kullm-solar 모델은 SOLAR-Instruction 모델에 고려대학교에서 만든 구름 데이터셋을 학습한 모델로 Open Ko-LLM 리더보드에서 15위에 기록된 적이 있다. Bookworm은 야놀자에서 업스테이지의 SOLAR 모델을 베이스로 자체 제작한 데이터셋을 학습한 후 미세조정한 모델이다.

3.2 평가 데이터셋 구성 및 방법

앞서 설명한 Hugging Face Open LLM Leaderboard에서 선정한 6가지 분야의 성능 지표를 바탕으로 공공부문과 민간부문의 한국어 사용성을 고려하여 한국어 능력, 한국사, 한국지리, 문학, 문법 등을 폭넓게 평가할 수 있는 문제로 총 5가지 분야 431개 문항을 만들어 평가를 진행하였다.

<표3> 최종 선정된 평가 분야

	데이터명(문항)	내용
1	KBS 한국어 능력시험(88)	한국어 문장 구조 평가
2	TOPIK(203)	외국인 대상 한국어 평가
3	한국사 능력 검정시험(40)	한국의 역사 문화 평가
4	국내여행안내사(50)	한국 지리적 특성 평가
5	공무원 9급시험(50)	문학, 문법 등 한국어 평가

<표4> LLM 성능 평가 데이터셋 구성 예시

데이터셋	Question	Answer	유형	분류
KBS 한국어 능력 시험	Q : <보 기>에 제시된 단어의 발음이 표준 발음인 것끼리 묶인 것은? E : <보 기> ㄱ)의심[의심] ㄴ)본의[본이] ㄷ)닝큼[닝큼] ㄹ)무늬[무늬] O : 1. ㄱ,ㄴ 2.ㄱ,ㄷ 3. ㄴ,ㄷ 4. ㄴ,ㄹ 5. ㄷ,ㄹ	2	(단일)정답형	어법
TOPIK	Q : ( )에 들어갈 말로 가장알맞은 것을 고르십시오. 다른 사람과 대화를 할 때는 적당한 거리를( ) 한다. O : 1. 유지해야 2. 유지하는 3. 유지했고 4. 유지하니까	1	불완전 문장형	읽기
한국사 능력 검정시험	Q : (가)에 해당하는 인물로 옳은 것은? E : 이곳 경복궁은 조선의 궁궐로 (가)이/가 이름 지었대. 국왕과 백성이 만년토록 태평하며 큰 복을 누리기를 바란다는 의미가 담겨 있어. 그는 새 왕조의 통치 방향을 제시한 조선경국전도 저술하였지. O : 1.송시열 2.채제공 3.정몽주 4.정도전	4	(단일)정답형	기본
국내여행안내서	Q : 소재지와 동굴의 연결이 옳은 것은? O : 1. 경북 안동 - 성류굴 2. 강원 삼척 - 고씨굴 3. 전북 익산 - 천호동굴 4. 충북 단양 - 초당굴	3	짝짓기형	관광자원 해설
공무원 9급시험	Q: 관용표현 ㄱ~ㄹ의 의미를 풀이한 것으로 적절하지 않은 것은? - 그의 회사는 작년에 노사 갈등으로 ㄱ홍역을 치렀다. - 우리 교장 선생님은 교육계에서 ㄴ잔뼈가 굵은 분이 십니다. - 유원지로 이어지는 국도에는 차가 밀려 ㄷ입추의 여지가 없었다. - 그분은 세계 유수의 연구자들과 ㄹ어깨를 나란히 하는 물리학자이다. O : 1. ㄱ: 심한 어려움을 겪었다 2. ㄴ:오랫동안 일을 하여 그 일에 익숙한 3. ㄷ: 돌아서 갈 수 있는 방법이 없었다 4. ㄹ:비슷한 지위나 힘을 가지는	3	불완전 문장형	어휘

<표5> LLM 평가 결과

모델명	KBS 한국어 능력시험		TOPIK		한국사능력검정시험		국내여행안내사		공무원 9급 시험	
	정답수	비율(%)	정답수	비율(%)	정답수	비율(%)	정답수	비율(%)	정답수	비율(%)
LDCC-SOLAR-10.7B	20	23	146	72	18	45	23	46	28	55
Bookworm-10.7B	21	24	152	75	10	25	21	42	26	51
SOLARC-M-10.7B	15	17	141	70	14	35	20	4	23	45
Llama-2-13b-chat-hf	18	20	83	41	0	0	9	18	11	22
Kullm-solar	8	9	87	43	2	5	6	12	10	20
KoAlpaca-Polyglot-12.8B	13	15	33	16	8	2	12	24	6	12

3.3 평가 결과

선정된 LLM의 한국어 이해에 대한 평가를 진행하기 위하여 평가에 활용할 데이터셋<표4>을 객관식 문제 유형으로 구성하였다. 평가는 모델별 입력구조가 다양하여 입력 형식에 맞추어 프롬프트를 구성한 후 테스트를 진행하였다.

한국어 능력을 평가하기 위해서 시행한 'KBS 한국어 능력 시험'과 'TOPIK' 시험의 결과는 <표5>와 같다. 'KBS 한국어 능력 시험'의 경우 문법과 어휘를 중심으로 한국어 문장의 구조를 이해하는 능력을 평가하게 되며, 'TOPIK'의 경우 한국어를 모국어로 사용하지 않는 이를 대상으로 수행되는 시험으로 언어의 사용 및 이해도에 대한 평가로 Bookworm 모델이 가장 높은 정답률을 보였다.

'한국사능력검정시험'과 '국내 여행 안내사 시험'의 경우는 한국의 역사와 문화 지식 측정이 목표인 시험으로 LDCC-SOLAR 모델이 가장 좋은 성능을 보였다.

'공무원 9급 시험'의 경우 한국 문학, 문법, 한자 표현 등을 포함한 한국어 능력을 폭넓게 평가하여 문장의 맥락을 정확히 이해하고 사용할 수 있는지 평가하게 된다. '국내 여행 안내사 시험'과 '공무원 9급 시험'의 경우 LDCC_SOLAR 모델의 정답률이 가장 높게 나왔으며 Booookworm 모델이 근소한 차이로 두 번째로 좋은 정답률을 보이는 것을 확인할 수 있다.

모델별 평가 결과는 <표6>에서 확인할 수 있다. 롯데정보통신의 LDCC-SOLAR가 가장 좋은 결과를

나타냈으며, 야놀자의 Bookworm이 2순위를 기록하였으며 SOLARC-M 모델도 정답률이 49%로 준수한 결과를 보여주었다. 다만 나머지 3개 모델은 정답률이 30%에 미치지 못하여 상대적으로 낮은 수준을 나타냈음을 확인할 수 있다.

<표6> LLM 평가 결과 종합

순위	모델명	정답수	비율(%)
1	LDCC-SOLAR-10.7B	235	55
2	Bookworm-10.7B	230	53
3	SOLARC-M-10.7B	213	49
4	Llama-2-13b-chat-hf	121	28
5	Kullm-solar	113	26
6	KoAlpaca-Polyglot-12.8B	72	17

상위권을 기록한 모델들의 공통점은 베이스 모델을 모두 업스테이지가 개발한 SOLAR를 기본으로 만들어졌음을 확인할 수 있다. SOLAR 모델의 경우 LLM의 효율적인 확장을 위한 깊이 기반 스케일링과 지속적인 사전 훈련을 통해 모델을 구축하였으며, 롯데정보통신과 야놀자에서는 자체 구축데이터를 통해서 한국어와 관련한 다양한 분야의 추가 학습이 지속해서 진행되어 높은 정답률을 보일 수 있던 것으로 판단된다. 따라서 우수한 데이터를 지속해서 학습하는 것이 LLM의 성능을 높일 수 있는 가장 좋은 방법임을 확인할 수 있다.

4. 결 론

공공분야 및 민간분야에서의 LLM을 활용하여 다양한 서비스 제공을 위해 노력하고 있다. 이를 위해선 서비스의 목적에 맞는 모델을 선정하고 해당 모델의 장단점을 정확히 이해하는 것에서 시작된다고 할 수 있다. 이를 위해 본 연구는 한국어와 관련한 다양한 문제 풀이 방식을 통해 6개의 LLM에 대한 성능을 평가 및 분석하였다. 세부적으로 한국어 구사 능력은 야놀자의 Bookworm 모델이 가장 우수한 것으로 나타났으며, 한국사, 한국 지리 그리고 공무원 9급 시험 문제에서는 롯데정보통신의 LDCC-SOLAR 모델이 우수한 것으로 평가되었다. 향후 LLM 기술은 산업별 응용에서 중요한 역할을 차지할 것이며, 이를 위한 올바른 평가는 중요한 역할을 할 것으로 생각된다. 이에 초거대 언어모델의 한국어 이해 능력을 평가하기 위한 다양한 데이터셋 구축 및 평가 지표에 대한 개발이 필요할 것으로 생각된다.

참고문헌

[1] S. Lim and S. Lee “Research Trends in Artificial Intelligence Language Models”, Information and Communication Magazine, Vol 40, No. 3, pp.42-50, 2023.

[2] M. Shanahan “Talking about large language models”, Communication of the ACM, Vol 67, No. 2, pp68-79, 2024.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez et al., “Attention is All You Need”, Advances in Neural Information Processing Systems, pp5998-6008, 2017.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding”, North America Chaper of the Association for Computational Liguistics, pp4171-4186, 2018.

[5] A, Radfordm J, Narasimhan, T. Salimans, and I. Sutskever, Improving Language Understanding by Generarive Pre-training, OpenAI, 2018

[6] LDCC. (2024, February 28). LDCC/LDCC-SOLAR-10.7B. Hugging Face. https://huggingface.co/LDCC/LDCC-SOLAR-10.7B

[7] Yanolja. (2024, March 16). Yanolja/Bookworm-10.7B-v0.4-DPO. Hugging Face. https://huggingface.co/yanolja/Bookworm-10.7B-v0.4-DPO

[8] Dopeornope. (2024, January 15). DopeorNope/SOLARC-M-10.7B. Hugging Face. https://huggingface.co/DopeorNope/SOLARC-M-10.7B

[9] Meta. (2023, November 13). Meta-Llama/Llama-2-13b-Hf. Hugging Face. https://huggingface.co/meta-llama/Llama-2-13b-hf

[10] Heavytail. (2024, January 28). Heavytail/Kullm-Solar. Hugging Face. https://huggingface.co/heavytail/kullm-solar

[11] Beomi.(2023, May 3). Beomi/KoAlpaca-Polyglot-12.8B. Hugging Face. https://huggingface.co/beomi/KoAlpaca-Polyglot -12.8B