Zero-shot learning 기반 대규모 언어 모델 한국어 품질 비교 분석

허윤아[1,○], 소아람[1], 이태민[1], 신중민[1], 박정배[1], 박기남[1], 안성민[2], 임희석[1,*]
[1]고려대학교, Human-inspired AI 연구소, [2](주)오투오
{yj72722, aram, taeminlee, tlswndals13, insmile, spknn, limhseok}@korea.ac.kr[1], smahn@o2o.kr[2]

Comparative analysis of large language model Korean quality based on zero-shot learning

Yuna Hur[1,○], Aram So[1], Taemin Lee[1], Joongmin Shin[1], JeongBae Park[1], Kinam Park[1], Sungmin Ahn[2], Heuiseok Lim[1,*]
[1]Korea University, Human-inspired AI Research, [2]O2O Inc.

요 약

대규모 언어 모델(LLM)은 대규모의 데이터를 학습하여 얻은 지식을 기반으로 텍스트와 다양한 콘텐츠를 인식하고 요약, 번역, 예측, 생성할 수 있는 딥러닝 알고리즘이다. 초기 공개된 LLM은 영어 기반 모델로 비영어권에서는 높은 성능을 기대할 수 없었으며, 이에 한국, 중국 등 자체적 LLM 연구개발이 활성화되고 있다. 본 논문에서는 언어가 LLM의 성능에 영향을 미치는가에 대하여 한국어 기반 LLM과 영어 기반 LLM으로 KoBEST의 4가지 Task에 대하여 성능비교를 하였다. 그 결과 한국어에 대한 사전 지식을 추가하는 것이 LLM의 성능에 영향을 미치는 것을 확인할 수 있었다.

주제어: Large Language Model, Foundation Model, Instruction Finetuned Model, Zero-shot learning

1. 서론

언어 모델(Language Model, LM)은 단어 시퀀스에 대한 확률분포(probability distribution)로 특정 문장 또는 단어가 나타날 확률을 계산하는 통계를 이용한 방법에서 인공 신경망을 이용한 방법으로 발전하였다. GPT, BERT 등과 같은 트랜스포머 모델(Transformer model)은 문장 속 관계를 추적해 맥락과 의미를 학습하는 신경망으로 대규모 언어 모델(Large Language Model, LLM) 등장의 시초가 되었다 [1].

대규모 언어 모델(이하 LLM)은 대규모의 데이터로 사전 훈련(pre-trained) 후 미세 조정(fine-tuned)하는 대규모 범용 언어 모델로 대규모의 학습 파라미터가 존재한다. LLM은 언어 이해, 언어 생성 등 자연어 처리 분야 외에도 다양한 영역에 활용 가능하며, 이를 기반으로 다양한 산업에 효율성을 제공한다. 국외에서는 Open AI, Google, Meta 등 글로벌 주요 AI 기업들이 GPT-4, PaLM2, Llama2, Polyglot 등을 발표하였으며, 기존 서비스에 통합하여 ChatGPT, Bard 등 새로운 서비스를 개발 및 제공하고 있다. 이에 네이버, LG, NC 등 국내 기업들도 HyperCLOVA, koGPT, EXAONE, VARCO 등을 발표하였다. 초기 공개된 LLM은 영어 기반 모델로 비영어권에서는 만족할 만한 성능을 보이지 않았지만, 비영어권에 대한 수요 및 기술 선점을 위한 연구의 필요성이 증가되었으며, 특히, 중국과 한국 기업들은 자체적 언어 모델 개발에 총력을 기울이고 있다.

LLM은 Foundation Model(FM)과 Instruction Finetuned Model(IFM)과 같은 방법들이 있다. FM은 레이블(label)이 없는 데이터를 비지도 학습(unsupervised learning)하여 광범위한 작업에 응용이 가능하다 [2]. 하지만 학습에 사용되는 데이터에 내재된 편향성(bias) 증폭, 오류 전파 등과 같은 한계가 있고, zero-shot으로 명령어를 주었을 때, 성능 저하 및 특정한 지시에 따른 작업을 수행하지 못하는 한계가 있으며, 이러한 한계를 극복하기 위해 IFM이 제안되었다. IFM은 FM의 zero-shot 성능을 향상시키기 위해 자연어 지시사항(instruction)과 입력 예시(one-shot example)를 사용하여 다양한 작업을 학습하도록 돕는 학습 방법으로 FM과 IFM에 대한 내용은 2장 관련연구에서 다룰 예정이다.

본 논문에서는 LLM에 언어의 영향이 미치는지 비교하기 위하여 영어 기반 LLM과 한국어 기반 LLM을 Zero-shot learning으로 비교실험을 진행하였으며, FM과 IFM도 비교 평가하였다.

2. 관련 연구

2.1 Foundation Model

Foundation Model(FM)은 대규모의 비레이블된 데이터를 활용하여 학습한 모델로 [2] 하나의 모델이 다양한 다운스트림 작업(Downstream task)을 수행할 수 있는 유연성과 작업 수행력이 있으며, 기존 언어 모델들의 한계점을 개선하였다. BERT [3]와 GPT [4] 같은 대표적인 모델들은 대규모 텍스트 데이터를 활용하지만, FM은 텍스트에 제한되지 않고 DALL-E [5]와 같은 텍스트와 이미지의 통합하여 학습된다. FM은 자기 지도 학습 방식으로 다양한 데이터에서 학습되며, 주로 주석이 없는 데이터를 통해 복잡한 패턴을 학습한다. 이렇게 방대한 언어 데이터로 학습된 모델은 독립적으로 계산과 생성 능력을 습득할 수 있다. 이런 학습 방식은 단순하며 데이터의 양을 증가

시키는 것만으로도 높은 성능을 달성하며 다양한 작업을 수행할 수 있다. 그러나 FM은 사용자의 의도와 다른 결과를 생성할 수 있어 사전에 예측하할 수 없는 한계가 있고 [6], Zero-shot 기반 학습에서의 성능 저하, 명확하지 않은 작업을 수행하지 못한다는 한계도 존재하였다. 이러한 한계를 극복하기 위해 IFM이 제안되었다.

2.2 Instruction Finetuned Model

Instruction Finetuned Model(IFM)은 FM을 Instruction 데이터 셋으로 Fine-tuning하여 zero-shot 성능을 높이는 방법 [7]으로 명확하지 않은 작업(Task) 유형을 이해하고, 주어진 지시에 따라 작업을 수행하는 데 필요한 능력을 향상시킬 수 있다.

Instruction Finetuning의 핵심 목적은 FM의 zero-shot 학습 능력을 향상시키는 것으로, 이를 통해 다양한 작업들을 자연어 지시문으로 재해석하고 적용한다. 이 방법은 사전 학습된 기술과 미세 조정 기술을 통합하여 사용된다. 특히 [7] 연구에서는 137B 매개변수를 갖는 언어 모델인 LaMDA [8]를 60개 이상의 NLP 작업을 위해 자연어 지시문 템플릿으로 구성하여 학습한 모델이 FLAN(Fintuned Language Models are zero-shot Learners)이다. FLAN은 명확하지 않은 작업 유형에 대한 성능을 평가받았으며, GPT-3의 175B 버전보다 자연어 추론 및 QA 작업에서 더 우수한 성과를 보였다. 실제로 많은 작업에서는 지도 학습 기반의 모델과 유사한 성능을 보였다. 추가로 [9]의 연구에서는 Instruction Finetuning 전략을 더욱 확장 시켰는데, Instruction Finetuning task 갯수가 성능에 미치는 영향을 실험을 통해 분석하였다. 결과적으로 학습시킨 task의 수와 상관없이 Instruction Finetuning 방법론을 적용한 모델은 그렇지 않은 모델보다 더 우수한 성능을 보였고, 이를 통해 Instruction Finetuning 방법론이 효과적인 것을 다시한번 확인하였다.

3. 실험 설계

본 논문에서는 언어가 LLM에 미치는 영향을 분석하기 위해 한국어 기반으로 학습된 LLM과 영어 기반으로 학습된 LLM에 대해 비교 분석였다. 또한 FM과 IFM을 비교 평가하도록 실험을 진행하였다. 본 실험에서는 zero-shot learning 기반으로 학습에 사용되지 않은 한국어 데이터인 KoBEST 중 4가지 task에 대하여 실험하였다.

3.1 Dataset

KoBEST는 SK Telecom에서 공개한 한국어 벤치마크 데이터셋으로 COPA, HellaSwag, WiC, BoolQ, SentiNeg 5가지 task에 대한 Train, Dev, Test 데이터로 구성되어 있다 [10]. 본 논문에서는 COPA, HellaSwag, BoolQ, SentiNeg 4가지 task로 실험하였다.

3.2 Tasks

COPA 주어진 상황에 대해 2가지 대안 중 맞는 대안을 선택하는 task. 상황에 대한 전제와 2개의 대안이 주어지고, 원인/결과에 따라 어떤 대안이 옳은지 선택한다.

HellaSwag 주어진 문장에서 다음에 위치할 문장을 선택하는 task. 문장과 다음에 위치할 4가지 후보 문장이 주어지면 후보 문장 중 정답인 하나의 문장을 선택한다.

BoolQ 문장과 질의가 주어졌을 때 질의에 맞는 답변을 하는 task. 답변은 True와 False 두 가지로 나뉜다.

SentiNeg 문장에 대해 긍정인지 부정인지 결과를 도출하는 task.

3.3 Models

본 논문에서는 한국어 기반 LLM인 VARCO, Polyglot-ko, KULLM과 영어 기반 LLM인 komt, Llama 2으로 실험하였다.

VARCO VARCO는 NC에서 자체 개발한 한국어 기반 LLM이다[1]. VARCO는 한국어 기반의 전문지식 코퍼스, 대화 데이터, 뉴스 및 위키, 번역 총 4가지 카테고리에 대하여 사전학습하였다. 또한 NC에서 자체 구축한 Instruction 데이터셋으로 추가 학습한 IFM과 FM 모두 공개하였으며, 각 모델 사이즈(model size)는 1.3B, 6.4B, 13B로 구성되어 있다.

Polyglot-ko EluterAI의 Polyglot-ko는 다국어 거대 언어모델을 목표로 TUNiB와 함께 수집한 한국어 데이터 1.2TB를 수집 후 학습시켰으며 다른 다국어 언어모델과도 비교할 수 있다 [11]. 총 4개의 모델을 공개하였으며, 모델 사이즈는 1.3B, 3.8B, 5.8B, 12.8B로 구성된다. 본 논문에서는 Polyglot-ko 모델을 사용하였다.

KULLM KULLM은 고려대학교 NLP & AI Lab와 Human-Inspired AI research에서 공개한 한국어 기반 LLM이다 [12]. 베이스 모델로 Polyglot 12.8B을 LoRA(Low Rank Adaptation)를 사용하여 학습하였다. 모델 학습에 사용된 데이터셋은 GPT4ALL, Dolly, Vicuna 이며, KULLM 7B, 12.8B 모델을 공개하였다.

Llama 2 Meta에서 공개한 Llama 2는 영어 기반 LLM으로[2] 약 2조개의 토큰을 사전학습시켰으며 Llama 1보다 문맥 길이가 2배이다. 공개된 모델 사이즈는 7B, 13B, 70B이다.

komt komt는 Llama2 모델을 기반으로 한국어의 성능을 효과적으로 향상시키기 위해 한국어 데이터를 multi-task instruction 기술을 적용하여 개발하였다[3]. 한국어 학습데이터로

[1] https://ncsoft.github.io/ncresearch/varco-llm/
[2] https://ai.meta.com/llama/
[3] https://github.com/davidkim205/komt

표 1. LLMs기반 zero-shot learning 실험 결과

Models	FM Language	Model Sizes	Type	COPA	HellaSwag	BoolQ	SentiNeg
VARCO	Korean	13B	FM	87.12	**83.25**	56.75	80.84
			IFM	**88.68**	77.4	57.12	87.3
		6.4B	FM	78.67	58.0	54.91	39.42
Polyglot-ko	Korean	12.8B	FM	79.37	59.54	48.18	**91.17**
		5.8B	FM	77.45	59.76	43.56	33.94
KULLM	Korean	12.8B	IFM	78.67	57.8	34.65	83.1
		5.8B	IFM	76.53	44.25	33.71	33.5
Llama2	English	13B	FM	59.27	41.29	34.86	51.57
		7B	FM	56.2	41.36	35.21	34.75
	English(+)	7B	FM	74.36	45.62	**57.86**	45.94
komt	English	13B	IFM	61.56	38.45	33.43	39.64

는 AI Hub, 모두의 말뭉치, KISTI AI, ShareGPT를 사용하였으며, 총 1,642,421 데이터를 학습시켰다.

3.4 Experiments Results

본 논문에서는 한국어와 영어 기반 LLM을 비교하기 위해 KoBEST의 4가지 task인 COPA, HellaSwag, BoolQ, SentiNeg에 대해 zero-shot learning으로 실험하였으며, 총 5가지 LLM을 비교한다. task 평가는 marco F1로 하였다. marco F1은 각 클래스의 데이터 비중과 관계 없이 똑같은 가중치를 주고 평균을 내는 평가방법이다. 표 1과 같이 FM language는 한국어(Korean)와 영어(English)기반으로 학습한 언어모델로 분류하였으며, English(+)의 경우 한국어로 language transfer 방법을 추가 학습하였다.

각 LLM 별로 비교하였을 때 model size가 크면 클 수록 성능 향상에 영향을 주는 것을 확인 할 수 있다. VARCO 에서 model size 13B일 때 FM과 IFM을 비교해보면, HellaSwag 빼고 나머지 task 에서는 IFM이 대부분 높은 성능나타내는 것을 확인할 수 있다. Llama2와 Llama2를 기반으로 한국어 데이터 추가 학습시킨 komt의 경우에는 model size 13B일 때 비교해보면 COPA 에서 성능 향상된 부분을 볼 수 있지만 Polyglot-ko와 Kullm에서는 FM인 Polyglot-ko가 모든 task에서 성능이 높은 것을 볼 수 있다. 즉 동일한 조건인 VARCO에서는 IFM의 성능이 높은 것을 봐서 IFM이 FM보다 효과적인 것을 알 수 있지만 다른 모델에서 대부분의 FM이 높은 것을 보아 IFM을 어떻게 설계하느냐에 따라 성능에 영향을 미치는 것으로 분석된다.

Llama2에서 영어기반 학습 모델 7B와 한국어로 language transfer한 "(+)"의 실험 결과를 비교했을 때 영어 기반으로만 학습한 모델보다 한국어로 language transfer한 결과의 성능이 높았다. 이를 통해 한국어에 대한 사전 지식을 추가하는 것이 성능에 영향을 미치는 것을 확인할 수 있다.

각 task별 실험 결과를 보면 COPA와 HellaSwag에서는 VARCO가 성능이 가장 높게 나왔고, BoolQ에서는 한국어로 language transfer 한 Llama2가 가장 높으며, SentiNeg에서는 Polyglot-ko가 가장 높게 나왔다. 전반적으로 영어 기반으로 학습한 LLM은 한국어 기반으로 학습한 LLM보다 성능이 낮게 나왔다.

4. 결론

본 논문에서는 언어가 대규모 언어 모델에 영향을 미치는가에 대하여 한국어 기반 LLM과 영어 기반 LLM을 zero-shot learning 기반으로 새로운 task에 대한 비교실험을 진행하였다. KoBEST를 사용하였으며, 총 5개의 task 중 COPA, HellaSwag, BoolQ, SentiNeg 4가지 task에 대한 데이터를 사용하였다. 한국어 기반 LLM으로 VARCO, Ployglot-ko, KULLM, 영어 기반 LLM으로 Llama2, komt을 비교하였으며, 그 결과 한국어 기반 LLM이 영어 기반 LLM 보다 높은 성능을 보였다. 특히 Llama2의 경우, 한국어로 language transfer한 결과의 성능이 높게 나타났으며, 이는 한국어에 대한 사전 지식을 추가하는 것이 성능에 영향을 준다고 할 수 있다. 또한 FM과 IFM 비교 결과, 모델의 크기가 클수록 성능 향상에 영향을 주는 것을 확인할 수 있었으며, HellaSwag를 제외한 Task에서는 대부분 IFM이 높은 성능을 보였다.

감사의 글

본 연구는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 기초연구사업(NRF-2021R1A6A1A03045425)과 3단계 산학연협력 선도대학 육성사업(LINC 3.0)의 연구결과임.

참고문헌

[1] S. Ahn, J. Yoo, W. Cho, J. Noh, and H. Son, "Rise of hyper-scale llm(large language model) and issues," SPRI(Software Policy & Research Institute), Tech. Rep., 2023.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[5] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *International Conference on Machine Learning*, pp. 8821–8831, 2021.

[6] E. Caballero, K. Gupta, I. Rish, and D. Krueger, "Broken neural scaling laws," *ICLR 2023*, 2023.

[7] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[8] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[9] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[10] M. Jang, D. Kim, D. S. Kwon, and E. Davis, "Kobest: Korean balanced evaluation of significant tasks," *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3697–3708, 2022.

[11] H. Ko, K. Yang, M. Ryu, T. Choi, S. Yang, S. Park *et al.*, "A technical report for polyglot-ko: Open-source large-scale korean language models," *arXiv preprint arXiv:2306.02254*, 2023.

[12] "Kullm: Korea university large language model project," GitHub repository, https://github.com/nlpai-lab/kullm, NLP & AI Lab and Human-Inspired AI research, 2023.