

과학기술 도메인 특화 언어모델의 Instruction Tuning 기반 성능 향상

정창후, 장광선, 양동훈, 임찬욱
한국과학기술정보연구원 초거대 AI 연구센터

{chjeong, gsjang, yangdonghun3, chanuklim}@csu.ac.kr

Performance Enhancement of Science and Technology Domain-Specific Language Models through Instruction Tuning

Chang-Hoo Jeong, Gwangseon Jang, Donghun Yang, Chanuk Lim
Korea Institute of Science and Technology Information

요약

본 논문은 과학기술 도메인 특화 대규모 언어모델 KONI(KISTI Open Neural Intelligence)-4B-it의 instruction tuning 적용 결과와 성능 개선을 소개한다. KONI-4B-it는 KONI-4B-base 모델을 기반으로 과학기술 분야에 특화된 instruction tuning을 통해 개발되었다. 현재 한국어와 영어 이중 언어를 지원하며, instruction tuning 적용을 통해 사용자 명령어 이해 및 응답 품질이 크게 향상되었다. KONI-4B-it 모델은 다양한 한국어 및 영어 벤치마크에서 평균 57.16%의 성능을 기록하여 베이스 모델 대비 1.08%p의 성능 향상을 달성하였다.

I. 서론

최근 대규모 언어모델의 발전과 함께 instruction tuning의 중요성이 부각되고 있다. Instruction tuning은 모델이 인간의 지시사항을 정확히 이해하고 적절한 응답을 생성할 수 있도록 하는 중요한 후처리 과정이다[1]. 본 연구에서는 과학기술 도메인에 특화된 KONI-4B-base 모델[2]에 instruction tuning을 적용하여 KONI-4B-it 모델을 개발하였다.

II. 본론

KONI-4B-it는 과학기술 도메인에 특화된 continual pre-training이 적용된 KONI-4B-base 모델을 기반으로 개발되었다. 이 베이스 모델은 40억 개의 매개변수를 가진 트랜스포머 아키텍처로, 이미 과학기술 분야의 전문 지식을 상당 부분 포함하고 있다. 그러나 사전학습된 모델은 사용자의 구체적인 명령어를 이해하고 적절한 형식으로 응답하는 능력에서 한계를 보인다.

이러한 한계를 극복하기 위해 본 연구에서는 과학기술 도메인에 특화된 instruction tuning을 수행하였다. Instruction tuning용 데이터셋은 과학기술 분야의 질의응답, 문서 요약, 개념 설명, 문제 해결 등 다양한 태스크로 구성된 instruction-response 쌍으로 구축하였다. 특히 한국어와 영어 양 언어에서 균형잡힌 고품질 instruction 데이터를 확보하여 이중 언어 성능의 동시 향상을 도모하였다. 데이터셋 구축 과정에서는 과학기술 도메인의 특성을 반영한 응답 형식과 내용의 정확성을 중점적으로 고려하였다.

학습 과정에서는 Supervised Fine-Tuning(SFT) 방법론을 적용하였으며, H200 GPU 24개를 활용한 분산 학습 환경에서 수행되었다. 분산 학습 기법과 메모리 최적화를 통해 안정적이고 효율적인 instruction tuning이 가능하였으며, 학습 과정에서 모델의 과학기술 도메인 지식 보존과 instruction-following 능력 향상 간의 균형을 유지하는 데 중점을 두었다.

성능 평가는 한국어 벤치마크인 kmmlu, kobest, haerae, kormedmcqa와 영어 벤치마크인 mmlu, arc, hellaswag, scholar, aidabench 등에서 수행되었다. 평가 결과, instruction tuning 적용 후 KONI-4B-it 모델은 전체 평균 성능이 56.08%에서 57.16%로 1.08%p 향상되었다. 특히 주목할 만한 점은 haerae 벤치마크에서 66.00%에서 73.33%로 7.33%p의 큰 폭 개선을 보인 것이다. 이는 instruction tuning이 모델의 복합적 추론 능력과 한국어 이해 능력을 효과적으로 개선했음을 보여준다. 또한 arc_challenge에서도 2.52%p의 의미있는 향상을 달성하여 영어권 추론 태스크에서도 성능 개선이 확인되었다.

III. 결론

본 논문에서는 과학기술 도메인 특화 언어모델 KONI-4B-it의 instruction tuning 적용 효과를 제시하였다. KONI-4B-base 기반의 instruction tuning을 통해 평균 1.08%p의 성능 향상을 달성하였으며, 특히 복합적 추론 능력에서 큰 개선을 보였다. 본 연구의 방법론은 다른 전문 도메인의 언어모델 개발에도 응용 가능할 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 과학기술정보 특화 LLM, KONI(KISTI Open Neural Intelligence) 개발을 목표로 한 2025년도 한국과학기술정보연구원(KISTI)의 기본사업(과제번호: K25L1M1C1)으로 수행된 연구입니다.

참고 문헌

- [1] Ouyang, L., Wu, J., Jiang, X., et al. "Training language models to follow instructions with human feedback," Advances in Neural Information Processing Systems, vol. 35, pp. 27730-27744, 2022.
- [2] KISTI, "KONI-4B-base-20250819," 2025, (<https://huggingface.co/KISTI-KONI/KONI-4B-base-20250819>).