원자력 도메인 특화 언어모델 개발 및 성능 평가: 맞춤형 데이터셋 구축을 통한 성능 비교 연구

염승돈, 최창수, 임경태, 유용균*

과학기술연합대학원대학교, 서울과학기술교육대학교, *한국원자력연구원

tmdehs77@gmail.com, {choics2623, ktlim}@seoultech.ac.kr, *ygyu@kaeri.re.kr

Development and Performance Evaluation of a Domain-Specific Language Model for Nuclear: A Comparative Study Using a Custom-Built Dataset

Yeom Seung Don, ChangSu Choi, Lim KyungTae, Yu Yong Gyun*

UST Univ., SeoulTech Univ., *KAERI.

요 약

본 논문은 특정 도메인에 최적화된 언어 모델의 필요성과 그 적용 방안에 대해 논의한다. 특히 데이터 부족 환경에서 도메인 특화 모델의 유효성을 높이기 위한 방법론을 제안한다. 그 대상으로 언어모델을 학습하기에 저 자원 상황인 원자력 분야에 특화된 언어 모델을 개발하기 위해 원자력 사전집을 학습 데이터로 가공하고, 지시어 튜닝을 통해 모델을 최적화했다. 또한, 도메인 특화 모델의 성능 평가를 위해 12개의 원자력 관련 질의응답 데이터를 제작하여 평가를 진행했다. 연구 결과, 본 논문에서 제안한 사전 증식 기반 사전학습과 Instruction Tuning 데이터를 활용해 최소 1%에서 최대 11.3%까지 성능 향상이 확인되었다.

Ⅰ. 서 론

언어 모델의 발전에 따라 특정 도메인에 최적화된 사전 학습 언어 모델이 활발히 활용되고 있다. 이러한 도메인 특화 언어 모델은 연구와 산업 분야에서 중요한 성과를 이루고 있으며, 특히 제조업, 교육, 학술 분야 등에서 높은 수요를 보이고 있다 [1, 2]. 이 모델들은 도메인 지식을 확장하기 위해 추가 사전 학습과 함께 새로운 작업을 수행할 수 있도록 지시어 튜닝(instruction tuning) 등의 추가 학습이 필요하다 . 그러나 매우 한정된 도메인에서는 학습 데이터의 수집과 가공 과정이 까다로울 수 있다. 특히 데이터가 부족한 환경에서 모델을 튜닝할 경우, 두 가지 문제가 발생할 수 있다 [3]: 첫째, 데이터의 편향성으로 인해 모델이 지나치게 특정 데이터에 적합화되고, 둘째, 기존에 학습된 일반적인 작업 수행 능력(번역, 요약, 질의응답 등)이 저하되는 지식 상실 현상이 발생할 수 있다. 이러한 문제는 모델의 신뢰성과 유효성을 저하시키는 주요 요인이 된다. 이를 해결하기 위해 소량의 고급 데이터를 증강하거나, 일반 학습 데이터와 도메인 데이터를 적절한 비율로 혼합하여 편향성과 지식 상실 문제를 최소화하는 방법이 제안되었다. 특히, 지시어 튜닝 과정에서는 기존 지식을 유지하면서도 새로운 작업을 학습할 수 있도록 기존 작업과 새로운 작업의 데이터를 혼합하는 전략이 효과적이다.

본 연구는 저자원 도메인 중 하나인 원자력 분야에 특화된 언어 모델을 개발하는 것을 목표로 한다. 이를 위해 첫째, 원자력 분야의 용어를 이해하기 위해 원자력 사전집을 추가 학습 데이터로 가공하여 모델을 학습시켰다. 이 데이터는 원자력 분야에서 자주 사용되는 용어를 모델이 보다 정확히 이해하도록 돕는다. 둘째, 지시어 튜닝을 통해 원자력 분야의 세부 내용을 반영하고 자주 사용되는 질의응답에 최적화된 언어 모델을 구축하였다.

마지막으로, 제안한 언어 모델의 성능을 어떻게 평가할 것인가가 중요한 문제이다. 본 연구에서는 정성 평가를 위해 원자력 도메인에 대한 이해가 필요한 12개의 질의응답형 데이터를 직접 제작하였다. 이 데이터는 추론, 수학, 코딩, 이해, 문법, 번역 등 다양한 작업으로 구성되었으며, 원자력 관련 지식이 없이는 답변하기 어려운 질의응답을 포함하였다. 평가 결과, 제안된 도메인 특화 학습 방법을 통해 최소 1%에서 최대 11.3%까지 성능이 향상됨을 확인할 수 있었다.

Ⅱ. 본론

1. Proposed Methods

Figure 1은 제안하는 원자력 도메인의 학습 과정을 도식화한 모형이다. 본 연구에서는 한국어-영어가 모두 소화 가능한 모델을 목표로 하고 있기 때문에 한국어-영어 이중 언어에서 가장 성능이 좋다고 알려진 Bllossom-8B모델을 활용해 추가 학습을 진행했다.

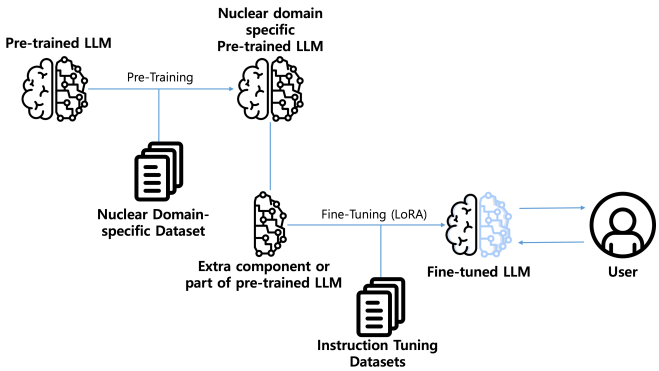

Figure 1. 원자력 도메인에 특화된 언어모델의 학습구조

1.1 학습 데이터 수집

Figure 1의 Nuclear Domain Dataset의 학습데이터는 원자력의 다양한 분야에서 수집되었다. 특히 원자력의 기본적인 용어 이해를 위해 한국 수원자력의 용어집과 법령 원자력안전위원회의 원자력 안전규제관련 용어사전

을 우선 수집했다. 사전데이터 이회에 한국 원자력연구원의 동향보고서와 원자력 wiki 및 원자력 분야별 안전정보에 대한 문서집을 수집했다.

1.2 언어 모델 학습

Figure 1의 Instruction Tuning Datasets은 Instruction Tuning데이터를 의미한다. 모델의 입력 데이터 구성으로는 아래 표1과 같은 질의응답 생성 과정을 거쳐 Instruction Tuning 데이터를 만들었다. 총 1만개의 데이터를 만들었는데 앞서 수집한 사전학습 학습데이터를 입력으로 GPT에게 자동으로 질의응답 데이터를 생성해 달라고 요청했다. 모델의 학습은 사용자의 입력 질의 부분을 제외하고, 출력 문장에 대해서만 loss를 계산하였으며, 이때의 Loss 계산은 [수식 1]과 같다. D_{atom}는 모델의 학습에 사용된 데이터 셋을 의미하며, 사용자 질의 x와, 이에 따른 답변 y의 쌍으로 이루어져 있다. $x_{\in st}$는 사용자 질의와 System Instruction이 포함된 전체 Instruction 샘플을 의미한다. 최종적인 모델의 입력 Instruction 샘플은 [Table 1]과 같다.

"[INST] <<SYS>> You are a helpful assistant. 당신은 원자력도메인에 특화된 AI 비서입니다. 주어진 질의에 정성스럽게 답변해주세요. <</SYS>> "원자력 발전시 비상 냉각 시스템의 종류와 필요성에 대해 알려줘" [/INST]"

Table 1. 모델에 입력되는 최종 Instruction 예시

$$L(\theta) = E_{(x,y)\sim D_{\text{atom}}}\left\{-\sum_{i=0}^{|y|}\log P(y_i|x_{\in \text{st}}, y_{<i};\theta)\right\},$$

$$where \quad (x,y)\in D_{\text{atom}}$$

수식 1. 모델의 SFT에 사용된 Loss 계산법

1.3 모델 평가를 위한 데이터셋 구축

한국어의 경우 언어모델에 대한 평가를 LogicKor 벤치마크와 같은 정성평가를 주로 진행하고 있다. LogicKor는 수리, 글쓰기, 코딩 등 8가지 항목의 평가 지표가 있기 때문에 한국어 언어모델의 일반적인 능력을 평가할 수 있다. 반면 원자력 도메인은 평가데이터가 없기 때문에 우리는 원자력 전문가의 검토를 토대로 총 12개의 평가 데이터를 구축했다. Table 3의 각 컬럼은 12개의 데이터가 포함된 평가 분야를 의미한다.

2. 실험환경 및 평가 결과

2.1 실험환경

모델 성능 평가는 Contextual Embedding을 활용하여 모델이 생성한 후보 문장과 평가 데이터셋의 레퍼런스 문장 간 의미적 유사성을 측정하는 BERTScore 지표를 사용하여 수행하였다. 성능 평가는 추론, 수학, 코딩, 이해, 문법, 번역 등의 분야에서 이루어졌으며, 총 12개의 평가 데이터셋을 구성하여 진행하였다. 또한, 비교 모델로는 최근 발표된 유사한 규모의 언어 모델인 'Llama3.1 8B', 'Qwen2.5 7B' 및 베이스라인 모델인 'Bllossom 8B'을 선정하여 성능 비교를 수행했다.

2.2 실험결과

[Overall] Table 3의 원자력 분야 평가 데이터셋에서 BERTScore 성능 결과를 통해 여러 언어 모델의 성능을 종합적으로 분석한 결과, 각 모델은 작업에 따라 성능 차이를 보였다. AtomGPT 8B는 대부분의 작업에서 뛰어난 성능을 보이며, 특히 추론(74.38), 수학(73.72), 이해(78.99), 문법(87.48), 번역(77.95) 작업에서 모든 모델 중 가장 높은 점수를 기록하였다. 반면, Blossom8B는 수학(71.35)과 코딩(91.73)에서 상대적으로 우수한 성능을 보였으나, 다른 작업에서는 AtomGPT에 미치지 못했다. Llama3.1 8B는 코딩(90.56)에서만 두각을 나타냈으며, 특히 번역(32.01)과 같은 작업에서는 매우 낮은 성능을 기록했다. Qwen2.5 7B는 코딩(92.09) 작업에서 가장 뛰어난 성과를 보였으나, 번역(32.91)과 수학(66.45) 작업에서는 낮은 성능을 보였다. 전반적으로, AtomGPT는 다양한 작업에서 균형 잡힌 성능을 보이며, 원자력 도메인에 최적화된 모델임을 입증하였다.

Table 3. 원자력 분야 평가 데이터셋에서 여러 언어 모델의 BERTScore 성능 결과

	추론	수학	코딩	이해	문법	번역
Bllossom8B	68.40	71.35	91.73	76.80	77.63	71.95
Llama3.1 8B	67.61	52.37	90.56	68.26	59.87	32.01
Qwen2.5 7B	66.92	66.45	**92.09**	77.76	86.90	32.91
AtomGPT 8B	**74.38**	**73.72**	91.88	**78.99**	**87.48**	**77.95**

Ⅲ. 결론

본 연구에서는 원자력 분야에 특화된 언어 모델을 개발하고 이를 평가하였다. 기존 언어 모델들이 다양한 도메인에 적용되기 위해 추가 학습과 지시어 튜닝이 필요함을 확인하였으며, 특히 저자원 환경에서 데이터 편향성과 지식 상실 문제가 발생할 수 있음을 지적하였다. 이러한 문제를 해결하기 위해 원자력 사전집을 학습 데이터로 활용하고, 기존 작업과 새로운 작업의 데이터를 혼합한 전략을 적용하였다. 연구 결과, 원자력 분야에 최적화된 언어 모델은 다방면에서 성능이 향상되었으며, 특히 추론, 수학, 코딩, 번역 등의 작업에서 우수한 결과를 보였다.

ACKNOWLEDGMENT

This work was supported in part by Korea Atomic Energy Research Institute R&D Program under Grant KAERI-524540-24.

참 고 문 헌

[1] B. Peng, C. Li, P. He, M. Galley, and J. Gao,"Instruction tuning with gpt-4," arXiv preprintarXiv:2304.03277, 2023.

[2] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu,H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E.Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: Anopen-source chatbot impressing gpt-4 with 90%*chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[3] H. Ko, K. Yang, M. Ryu, T. Choi, S. Yang, J. Hyun,S. Park, and K. Park, "A technical report for polyglotko: Open-source large-scale korean language models,"2023.