Mistral AI 모델을 활용한 특허 명세서의 자동 생성

이우석, 박소현, 김경재, 박영연, 이지석, 임현우, 정남주, 성영락, 박준석
국민대학교

wolfman352@kookmin.ac.kr, sohyunzzq@kookmin.ac.kr, economy02@kookmin.ac.kr,
parkyoungyeo@kookmin.ac.kr, runseok23@kookmin.ac.kr, dla418@kookmin.ac.kr,
plmnko1016@kookmin.ac.kr, yeong@kookmin.ac.kr, jspark@kookmin.ac.kr

Automatic Generation of Patent Specifications Using Mistral AI's Model

Lee Woo Seok, Park So Hyun, Kim Gyung Jae, Park Young Yeon, Lee Ji Seok,
Lim Hyun Woo, Jung Nam Joo, Seong Yeong Rak, Park Jun Seok
Kookmin Univ.

요 약

특허 등록 시 필수 요소인 특허 명세서는 전문성을 기반으로 한 다량의 내용을 작성해야 하므로 작성 시 많은 시간이 소요된다는 문제점이 있다. 본 논문에서는 GPT[1] 출시 이후 최근 주목받고 있는 LLM(Large Language Model)을 이용하여 앞서 설명한 문제점을 해결하고자 하였다. AI 를 학습시키기 위해 KIPRIS 와 KEYWERT 에서 특허 명세서에서 데이터를 크롤링하였다. 그 후, 학습에 적절한 데이터로 가공하기 위해 추출한 데이터를 2 차례 전처리 과정을 거쳤고, 가공된 데이터를 Mistral AI 에 Fine-Tuning 하여 특허 작성 AI 를 구현하였다.

Ⅰ. 서론

특허를 등록할 때에는 발명품에 대한 특허 명세서 작성이 필수적이다. 하지만 특허 명세서에는 <발명의 명칭>, <배경 기술>, <기술 분야>, <발명의 내용> 등 많은 내용을 작성해야 하므로 상당한 시간과 노력이 필요하다. 이러한 문제점을 해결하기 위해 본 논문에서는 최근 주목을 받고 있는 LLM(Large Language Model)[2], [3] 기반 AI 를 구현해 특허 명세서 작성에 도움을 주고자 하였다.

LLM 은 많은 양의 데이터를 사전 학습한 초대형 딥러닝 모델이고, 생성형 AI 는 이용자의 입력에 따라 결과가 생성해내는 AI 기술이다. 이를 이용해 이용자가 요구한 특허명세서를 작성하기 위해 데이터를 학습하여 이를 토대로 능동적으로 결과물을 제시한다[4]. 또한, 특허 문서 작성에 특화된 AI 모델을 구현해 전문가 수준의 내용도 작성할 수 있기 때문에 변리사의 특허 명세서 작성을 편리하게 해준다.

Ⅱ. 데이터 수집 및 전처리

KIPRIS 와 KEYWERT 에서 약 2000 개의 특허 데이터를 수집하였다. 크롤링한 데이터에서【0002】와 같이 문장의 위치를 나타내는 숫자는 특허와 관련된 내용이 아니므로 학습된 모델의 성능을 저하시킨다. 따라서 1 차 전처리 과정으로 크롤링 과정에서의 불필요한 데이터들을 불용어로 설정하여 제거한다. 다음으로, 특허 명세서에 <KSIC 분류코드>, <선행 기술 문헌>과 같이 모델 학습에 불필요한 항목들은 제거하고 <발명의 명칭>, <기술 분야>와 <배경 기술>만을 추출하여 2 차 전처리를 진행하였다. 최종적으로 [그림 1]의 전처리 전 데이터는 [그림 2]와 같이 가공되었다.

【0002】

현대 사회는 산업화가 진행됨에 따라 이동의 편의성 등으로 인해 수많은 자동차가 운행되고 있다. 일반적으로 도심지를 벗어난 국도변의 험한 굴곡도로나, 고속도로의 커브길, 또는 고속도로의 진출입 구간 등의 곡선도로상에서는 차량의 진행경로를 사전에 파악하기 위한 시야확보가 어려운 이유로 시선유도시설이 설치된다.

【0003】

…

【0016】

이에 따라, 사고 위험 구간에 대한 안내와 함께, 고속 및 저속 주행에 대한 위험성을 경고하여 안전 운전을 유도하고, 인식률과 식별력을 높인 도로 안내 표시장치의 개발 필요성이 요구되고 있다.

그림 1. 전처리 전 데이터 예시

현대 사회는 산업화가 진행됨에 따라 이동의 편의성 등으로 인해 수많은 자동차가 운행되고 있다. 일반적으로 도심지를 벗어난 국도변의 험한 굴곡도로나, 고속도로의 커브길, 또는 고속도로의 진출입 구간 등의 곡선도로상에서는 차량의 진행경로를 사전에 파악하기 위한 시야확보가 어려운 이유로 시선유도시설이 설치된다. … 이에 따라, 사고 위험 구간에 대한 안내와 함께, 고속 및 저속 주행에 대한 위험성을 경고하여 안전 운전을 유도하고, 인식률과 식별력을 높인 도로 안내 표시장치의 개발 필요성이 요구되고 있다.

그림 2. 전처리 후 데이터 예시

Ⅲ. Fine-Tuning

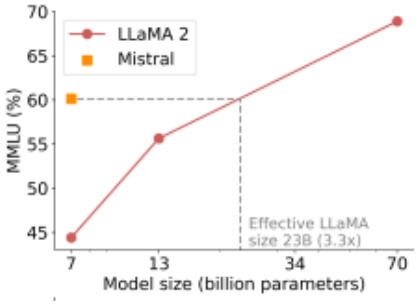

그림 3. 벤치마킹 비교[5]

Ⅲ장에서 전처리 과정을 거친 특허 데이터를 활용하여 Mistral AI 를 Fine-Tuning 하였다. Mistral AI 는 Window Sliding 기법을 사용하여 넓은 범위의 정보를 고려하여 Attention 을 수행하였고, 그로 인해 적은 파라미터 수에도 높은 성능을 이끌어낼 수 있었다[5]. [그림 3]은 Mistral AI 와 LLaMa2 의 MMLU(Massive Multitask Language Understanding) 벤치마킹을 비교한 그래프이다. [그림 3]처럼 Mistral AI 는 7B 의 파라미터로 LLaMa2 23B 파라미터와 성능이 유사하다. 본 논문에서는 <발명의 명칭>을 입력으로 설정하고, <기술 분야>와 <배경 기술>을 출력할 수 있도록 모델을 Fine-Tuning 하였다. 학습 시, 명확한 정답이 존재하지 않는 생성형 AI 의 경우 train 데이터에 비해 test 데이터의 비중은 높지 않다[6]. 따라서 본 논문에서는 train 데이터와 test 데이터를 9:1 의 비율로 분할하여 Fine-Tuning 하였다.

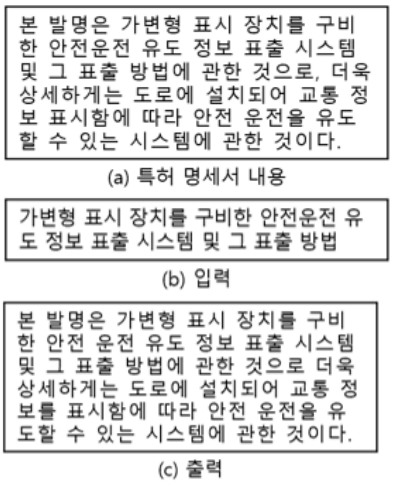

그림 4. 특허명세서 비교

[그림 4(a)]는 기존 특허 명세서에 작성된 <기술 분야>의 내용이다. [그림 4(c)]는 [그림 4(b)]를 입력으로 하여 생성된 <기술 분야>의 내용이다. [그림 4(c)]의 결과로부터, <기술 분야>의 말머리가 "본 발명은"으로 시작하는 특허 명세서만의 문체를 학습하였음을 확인하였다. 또한 [그림 4(a)]와 [그림 4(c)]를 비교하였을 때 문장이 유사하게 생성됨을 확인하였다.

Ⅳ. 결론

본 논문에서는 KIPRIS 와 KEYWERT 에서 특허 데이터를 수집 및 전처리하였으며, Mistral AI 를 활용하여 Fine-Tuning 과정을 통해 약 2000 개의 특허 데이터를 추가 학습하였다. 또한, 특허 명세서만의 특유 문체를 학습하였다. [그림 4]로부터 본 논문에서 구현한 모델을 통해 생성된 <기술 분야>와 특허 명세서에 작성된 <기술 분야>가 유사함을 확인하였다. 향후 충분한 데이터 확보가 이루어진다면 해당 모델의 성능을 높일 수 있을 것으로 확신한다.

참 고 문 헌

[1] T. B. Brown, et al. "Language models are few shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual,* H. Larochelle, M. Ranzato, R. Hadsell, M. Balcon, H. Lin, Eds., 2020.

[2] 박찬준, 이원성, 김윤기, 김지후, 이활석. (2023). 초거대 언어모델 연구 동향. 정보과학회지, 41(11), 8-24.

[3] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. "A Survey of Large Language Models," *arXiv preprint arXiv:2303.18223*, 2023.

[4] 양지훈, 윤상혁. (2023). ChatGPT 를 넘어 생성형(Generative) AI 시대로 : 미디어 · 콘텐츠 생성형 AI 서비스 사례와 경쟁력 확보 방안. 한국방송통신전파진흥원

[5] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, "Mistral 7b," 2023.

[6] 이지훈. "생성형 사전학습 및 데이터 검증을 사용한 한국어 질문-답변 데이터 합성." 국내석사학위논문 광운대학교 대학원, 2021. 서울