KoAlpaca를 활용한 특허 명세서 작성을 위한 생성형 AI에 관한 연구

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

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

A Study on Generative AI for Writing Patent Specifications using KoAlpaca

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

요 약

특허 명세서는 발명의 보호를 위해 필수적한 권리서이다. 하지만, 특허 명세서를 작성하는 데에는 많은 시간과 노력이 소요된다는 문제점이 존재한다. 본 논문에서는 생성형 AI 언어 모델인 KoAlpaca를 활용하여 특허 명세서 내 “해결하려는 과제”와 “과제의 해결수단” 항목의 초안을 작성해 문제점을 보완하는 방식을 제안한다. 이를 위해 기존 특허 명세서에서 “해결하려는 과제”, “과제의 해결수단”, “발명의 명칭”, “청구범위” 추출하였고, 추출한 데이터를 불용어 제거와 독립항 분리 전처리과정을 거쳤다. 그 후, 데이터를 이용해 Fine-Tuning을 진행하였고, 모델을 통해 작성한 특허 명세서 초안과 실제 특허 명세서 초안을 비교하여 그 결과를 검토한다.

Ⅰ. 서 론

특허 명세서는 발명의 보호 범위를 명시하기 위해 권리서로의 역할과 발명의 내용을 공개하는 기술 문헌이다[1]. 하지만, 특허 명세서를 작성하는 데 있어, 많은 사람의 노력과 시간, 그리고 전문가 수준의 높은 지식이 요구된다. 이를 해결하고자 최근 연구가 활발히 진행되고 있는 LLM(Large Language Model)을 활용해 생성형 AI를 구현하였다. LLM은 방대한 양의 데이터로 사전 학습된 초대형 딥 러닝 모델이다. 이를 활용해 특허 명세서를 작성하기 위해 데이터를 학습하여 작성자가 원하는 결과물을 제시한다[2].

본 논문에서는 특허 명세서에 필수적인 항목 중 “해결하려는 과제”와 “과제의 해결수단”의 초안을 작성하도록 모델을 학습시켜 기존의 특허 명세서와 결과를 비교한다. 본 논문의 Ⅱ장에서는 특허 명세서에서의 데이터 추출 및 전처리과정을, Ⅲ장에서는 Ⅱ장의 데이터를 이용한 Fine-Tuning 방법에 관해 설명한다. Ⅳ장은 결론이다.

Ⅱ. 본론

특허 명세서의 청구범위는 특허로서 보호받고자 하는 사항으로 특허 권리의 범위를 정하기 때문에 특허를 허가하는 단계에서 중요한 역할을 한다[3]. 따라서, 특허 명세서로부터 “해결하려는 과제”, “과제의 해결수단”, “발명의 명칭”, “청구범위”에서 데이터를 추출해 학습을 진행하였다. 이때, 특허 정보검색 시스템인 KIPRIS와 KEYWERT를 활용해 약 2,000개의 데이터셋을 수집하였다.

청구범위는 해당 특허 내용에 관하여 가장 중요한 내용을 명시하는 독립항과 독립항을 구체화해 나타내는 종속항으로 이루어져 있다[4]. 작성자의 의도에 맞는 특허 명세서 초안을 작성하기 위하여, 청구항 내에서 독립항만 추출하는 전처리과정을 거쳤다. 또한, 특허 명세서 안에 “상기”, “청구항 1”, “[0001]”와 같은 학습 시에 불필요한 단어나, 인덱스를 의미하는 부분을 불용어로 설정하고 제거하였다. 그림 1은 기존 특허 명세서 안에서 추출한 “청구범위”의 전처리 전 데이터의 예시이고, 그림 2는 같은 “청구범위”의 전처리 후 데이터의 예시이다.

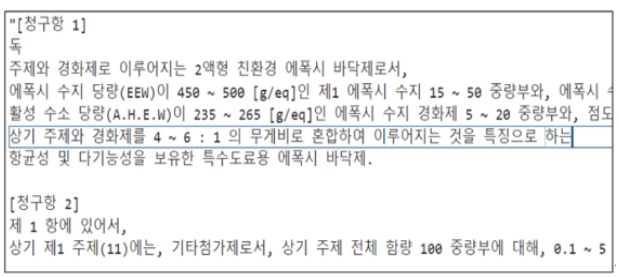

그림 1. 데이터 전처리 전 예시

"수제와 경화제로 이루어지는 2액형 친환경 에폭시 바닥제로서,
에폭시 수지 당량(EEW)이 450 ~ 500 [g/eq]인 제1 에폭시 수지 15 ~ 50 중량부와, 에폭시 수
합성 수소 당량(A.H.E.W)이 235 ~ 265 [g/eq]인 에폭시 수지 경화제 5 ~ 20 중량부와, 선도
주제와 경화제를 4 ~ 6 : 1 의 무게비로 혼합하여 이루어지는 것을 특징으로 한다.
양균성 및 다기능성을 부유한 특수도료용 에폭시 바닥제."

그림 2. 데이터 전처리 후 예시

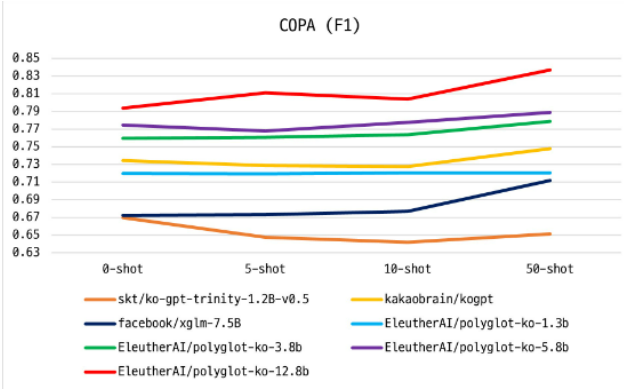

그림 3. 한국어 모델별 COPA score 비교

가변형 표시 장치를 구비한 안전운전 유도 정보 표출 시스템 및 그 표출 방법

(a) 입력 (발명의 명칭)

【0019】본 발명의 목적은 사고 위험 구간에 대한 안내와 함께, 고속 및 저속 주행에 대한 위험성을 경고하여 운전자의 안전운전을 유도하고, 인식률과 식별력을 높인 안전운전 유도정보표출 시스템 및 방법을 제공하는 것이다.

(b) 기존 특허 명세서 중 일부

본 발명은 상기와 같은 문제점을 해결하기 위한 것으로, 운전자의 시선에 따라 표시 정보들을 자동으로 변환하여 운전자의 안전운전을 유도할 수 있도록 구성된 가변형 표시 장치를 구비한 안전운전 유도 정보 표출 시스템 및 그 표출 방법을 제공하는 것을 목적으로 한다.

(c) 모델을 통해 작성된 내용 중 일부

그림 4. 해결하려는 과제

[청구항 1] 독 차량이 통행하는 도로에 설치되는 검지수단에 의해 상기 도로의 교통 상황을 수집하고 상기 검지수단에서 수집된 정보로부터 판단된 안내 정보를 운전자에게 제공하는 안전운전 유도 정보 표출 시스템에 있어서, 상기 도로에 설치되며 상기 안내 정보를 시각적인 이미지 형태의 안내 표지로 표시하기 위한 표시수단.

(a) 입력 (청구범위) 중 일부

【0021】 전술한 목적을 달성하기 위한 본 발명의 가변형 표시 장치를 구비한 안전운전 유도 정보 표출 시스템은, 검지수단에서 수집된 정보로부터 판단된 안내 정보를 운전자에게 제공하도록 구성되며, 이를 위해 도로에 설치되며 안내 정보를 시각적인 이미지 형태의 안내 표지

(b) 기존 특허 명세서 중 일부

상기한 목적을 달성하기 위한 자유롭게 가변 형상을 가지는 표시 장치를 구비한 안전운전 유도 정보 표출 시스템은, 이를 위해 도로에 설치되는 안내 표지를 시각적인 이미지 형태의 안내 표지로 표시하기 위한 표시수단을 제어하기 위한 제어수단을 포함한다. 조명유닛이 도로의 주행 방향을 상호 간격을 두며 연속적으로 설치되면서

(c) 모델을 통해 작성된 내용 중 일부

그림 5. 과제의 해결수단

III. Fine-Tuning

AI 모델은 데이터를 학습자료로 이용하기에, 해당 언어의 데이터가 적다면 좋은 성능을 발휘하기 어렵다[5]. 따라서 본 논문에서는 863GB 의 한국어 데이터로 학습된 EleuterAI 의 Polyglot-Ko 모델을 기반으로 한 KoAlpaca-Polygpot-Ko 12.8B 모델을 선택하였다. 그림 3 은 COPA (F1) score 로 한국어 모델별 성능을 비교한 그래프이다. Polyglot-Ko 12.8B 모델이 다른 한국어 모델보다 더 안정적이고 높은 score 를 기록하였다. [5]

기존의 KoAlpaca 에 전처리과정을 거친 약 2,000 개의 특허 관련 데이터를 추가 학습시켜 Fine-Tuning 을 진행하였다. 그 결과, 특허 명세서의 "발명의 명칭"과 "청구범위" 항목을 입력하였을 때, 출력으로 "해결하려는 과제"와 "과제의 해결수단"항목에 적합한 문장들이 생성되는 것을 확인했다. 또한, "본 발명은"과 같은 "해결하려는 과제"항목의 첫 머리말로 필수적인 문구도 출력되는 것을 확인하였다. 그림 4 는 "해결하려는 과제", 그림 5 는 "과제의 해결수단"의 기존 특허 명세서와 작성된 특허 명세서의 내용이다. 그림 4, 5 에서 기존 특허 명세서의 해당 내용과 생성된 문장을 비교하였을 때, 특허 명세서의 중요한 특정 단어와 내용을 포함하였고, 문체 또한 기존과 유사한 결과를 보였다.

IV. 결론

본 논문에서는 KoAlpaca 를 사용해 특허 명세서의 "해결하려는 과제"와 "과제의 해결수단"항목을 작성하는 모델을 제안했다. 특허 명세서의 "발명의 명칭"과 "청구범위"에서 데이터를 추출 후 특허 명세서의 문체나 키워드 같은 특징을 반영하여 전처리와 학습을 진행했다. 본 연구에서는 제한된 성능의 GPU 와 충분치 못한 데이터로 학습을 진행하였지만, 더 높은 성능의 GPU 와 추가적인 데이터 학습이 이루어진다면 더 좋은 성능을 보일 것이라고 예상한다.

참 고 문 헌

[1] 최병인 (2013), "특허명세서 작성 전략에 관한 연구.", 아주대학교 국내석사학위논문

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

[3] 특허청, (2010), 명세서 작성요령 및 청구범위 해석

[4] 송민호, (2018), "청구항 기반의 특허문서 의미구조 추출 및 시각화.", 한국항공대학교 일반대학원 국내석사학위논문

[5] Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon, …, & Kyubyong Park. (2023). A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models. 4-6, arXiv:2306.02254