실시간 제지공정 운영 효율화를 위한 LLM 과 RAG 기반 대화형 AI 어시스턴트 설계

길기훈 [1], 이서영 [1], 백민우 [1], 유지오 [1], 최진영 [2], 도윤미 [2], 이상금 [1]*
*국립한밭대학교 [1], 한국전자통신연구원 [2]

{minegihun, syoung2353, bmw5779, uzo7383}@gmail.com, choij0@etri.re.kr, ydoh@etri.re.kr,
*sangkeum@hanbat.ac.kr

Design of an LLM and RAG-Based Conversational AI Assistant for Optimizing Real-Time Paper Manufacturing Process Operations

Gihun Gil[1], Seoyoung Lee[1], Minu Baek[1], Jio Yoo[1], Jinyoung Choi[2], Yoonmee Doh[2], Sangkeum Lee[1]*
*Hanbat National Univ.[1], ETRI[2]

요 약

본 논문은 제지 산업에서 발생하는 스팀 사용량의 이상치나 예측 모델의 정확도 추이 등을 분석하기 위해 각 분석 도구의 사용법을 전부 익히는 대신 Kanana-1.5-15.7B-A3B LLM(Large Language Model)을 활용한 자연어 기반 제지공정 맞춤 대화형 AI 어시스턴트를 제안한다. 대화형 AI 어시스턴트가 외부 제지공정 데이터를 참조할 수 있도록 MoE(Mixture of experts) 아키텍처와 RAG(Retrieval-Augmented Generation) 기술을 결합하여 실시간으로 발생하는 이상치들을 신속 분석이 가능하다. 또한 LLM 의 Function Calling 기능을 통해 단순히 이상 구간을 조회하는 것을 넘어 스스로 데이터 분석 함수를 호출할 수 있도록 설계하여 관리자들이 별도의 교육 없이 자연어를 사용하여 대화형 AI 어시스턴트와 소통하면서 이상구간 확인과 이상치 분석 등의 작업을 수월히 진행할 수 있다. 향후 비전-언어 모델을 결합해 텍스트와 이미지 데이터를 모두 활용하는 멀티모달 시스템으로 확장할 예정이다.

Ⅰ. 서 론

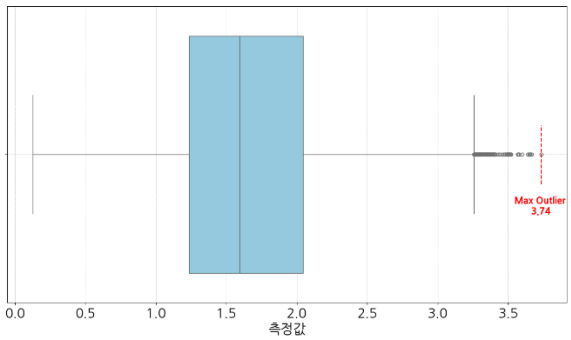

그림 1. 스팀 사용량 이상치 분석

제지 산업에서 안정적으로 고품질의 제품을 생산하기 위해 펄프 배합, 스팀 압력 등의 공정 변수들을 이상치를 실시간으로 탐지하는 일은 공정의 효율성을 높이기 위한 핵심 과제이다[1]. 그림 1 은 공정에서 사용된 실제 스팀 사용량이 예측치를 초과한 양의 분포로 대부분은 일정한 범위 내에서 관찰되나 IQR 을 벗어나는 이상치 또한 빈번히 발생한다. 그림 2 는 2022 년 동안 스팀 사용량에 대해 예측 모델이 실제 사용량을 얼마나 정확하게 예측했는지를 나타내며, 각 날짜의 최저 예측 정확도를 분석한 그래프로 전반적으로 95% 이상의 최저 예측 정확도를 보여주나 4 월부터 6 월 사이에는 이상치들로 인해 최저 예측 정확도가 90% 초반까지 떨어지는 모습을 보인다.

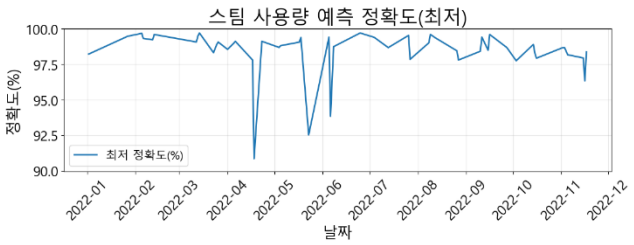

그림 2. 스팀 사용량 예측 최저 예측 정확도 그래프

관리자는 이와 같은 이상치에 대해 지속적인 모니터링이 필요하나 기존의 수많은 시계열 데이터들을 수동으로 비교하거나 데이터 분석 프로그램을 사용하기 위하여 별도의 교육을 실시해야 한다는 문제점이 발생한다. 본 논문에서는 LLM(Large Language Model)을 활용한 제지공정 맞춤 대화형 AI 어시스턴트를 제안한다. 대화형 AI 어시스턴트를 도입하여 운영자는 별도의 교육 없이 “공정 진행 중 현재 시점까지 발생한 이상 구간이 얼마나 되는지 알려주세요.”와 같은 대화형 자연어를 사용하여 데이터 조회 및 분석을 수행한다.

Ⅱ. 본론

1. 추론 속도 최적화를 위한 MoE 아키텍처의 적용

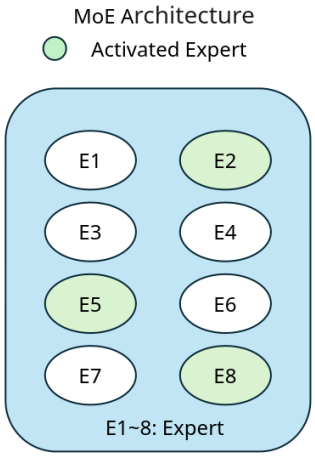

그림 3. MoE 아키텍처 개념도(설명을 위한 예시)

본 연구는 대화형 AI 어시스턴트로 Kanana-1.5-15.7B-A3B LLM 을 사용한다. 추론 시 모든 파라미터를 전부 활성화하는 기존 Dense 모델과 달리 해당 모델은 그림 3 과 같이 필요한 일부 Expert 들만 활성화하는 MoE(Mixture of Experts) 아키텍처를 채택하여 빠른 추론 속도를 유지한다[2]. 이는 제지공정 과정에서 실시간 응답을 필요로 하는 대화형 AI 어시스턴트를 구성하기 위한 최적의 조건이다. 일반적인 Kanana-1.5-15.7B-A3B 는 전체 64 개의 Expert 들중 8 개의 Expert 들을 활성화한다.

2. 대화형 AI 어시스턴트 기능 확장을 위한 LLM 에이전트 도입

2.1 RAG 를 사용한 LLM 지식보완

LLM 을 제지공정과 같은 특정 도메인에 적용하기 위해서는 해당 도메인에 대한 지식을 프롬프트에 제공하거나 파인튜닝하는 방법이 존재한다. 그러나 방대한 시계열 데이터를 전부 프롬프트에 입력한다면 LLM 이 한 번에 처리할 수 있는 컨텍스트 길이를 증가시켜야 하기 때문에 과도한 메모리 사용이 발생한다. 또한 파인튜닝의 경우 지식 컷오프가 발생하여 제지공정과 같이 실시간으로 시계열 데이터가 추가되는 상황에서는 적합하지 않다.

RAG(Retrieval-Augmented Generation)는 외부 문서에서 질의와 가장 관련성 높은 조각을 검색해 LLM 프롬프트에 결합함으로써, 컨텍스트 한계를 완화하고 답변 정확도를 높인다[3].

$$P(a|q) \approx \sum_{z \in Z_k} P(a|q,z) \cdot P(z|q)$$

RAG 는 기존 질문 q 와 답변 a 로만 이루어지는 과정 P(a|q)에 문서조각 z 를 추가하여 확장한다. P(z|q)는 RAG 의 Retriever 로 질문 q 의 의도와 가장 유사한 문서 조각 z 를 검색한다. P(a|q,z)는 RAG 의 Generator 로 문서와 원본 질문과 결합한 프롬프트를 기반으로 최종 답변 a 를 생성하여 연산효율과 답변의 정확도를 개선한다. 또한 새로운 시계열 공정 변수 데이터를 실시간으로 데이터베이스에 동기화함으로써 LLM 은 항상 최신의 공정 정보에 접근한다.

2.2 자연어 기반 데이터 분석을 위한 Function Calling 적용

실시간 데이터를 사용하여 새로운 이상치 분석을 하기 위해서는 RAG 로 얻은 공정 변수뿐 아니라 실제 분석을 실행할 함수실행이 필요하다. Kanana-1.5-15.7B-A3B 는 Function Calling 을 사용하여 스스로 사용자의 질문의 목적에 부합하는 함수를 알맞은 변수와 함께 호출문의 형태로 답변한다[4]. LLM 이 생성한 함수 호출 스펙을 실행한 뒤 결과를 다시 입력하면, LLM 은 실행 결과를 반영한 최종 답변을 산출한다. 이를 통해 관리자는 이상치 분석을 위해 별도의 프로그래밍 언어에 대해 배우지 않아도 자연어를 사용해 LLM 과 소통하여 원하는 분석 결과를 얻는다.

Ⅲ. 결론

본 논문에서는 MoE 아키텍처 기반의 LLM 에 RAG 와 Function Calling 기술을 결합한 제지공정 맞춤형 대화형 AI 어시스턴트를 설계하였다. 본 연구의 의의는 복잡한 데이터 분석 도구 또는 프로그래밍 언어에 대한 별도의 교육 없이 관리자가 자연어를 사용하여 필요한 정보에 실시간으로 접근하고 데이터 기반의 의사결정을 내릴 수 있는 새로운 방법을 제시함에 있다. 이는 공정 문제 해결 시간을 단축하며 운영 효율성을 높이는 데 기여한다.

향후 연구에서는 대화형 AI 어시스턴트를 비전 언어 모델과 결합하여 제지공정에서의 이미지를 활용한 멀티모달 시스템을 구축할 예정이다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원(No. 20202020900290) 및 2025 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음. (2022-0-01068)

참 고 문 헌

[1] Lee, Sangkeum, et al. "Factory Energy Management by Steam Energy Cluster Modeling in Paper-Making." Proceedings of the 2023 11th International Conference on Smart Grid (icSmartGrid), Paris, France. 2023.

[2] Shazeer, Noam, et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." arXiv preprint arXiv:1701.06538 (2017)

[3] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in neural information processing systems 33 (2020): 9459-9474.

[4] Schick, Timo, et al. "Toolformer: Language models can teach themselves to use tools." Advances in Neural Information Processing Systems 36 (2023): 68539-68551.