Journal of KIIT. Vol. 23, No. 6, pp. 11-20, Jun. 30, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.6.11

Human-in-the-Loop 적용 언어모델 파인튜닝을 통한 정밀 금속가공 업종의 암묵지 학습 방법

장성수*[1], 유동휘*[2], 권재영*[3], 이원희*[4]

Tacit Knowledge Integration in Precision Machining Industry with Human-in-the-Loop Fine-Tuning of Language Models

Seongsu Jhang*[1], Donghwi Yoo*[2], Jaeyoung Kwon*[3], and Wonhee Lee*[4]

이 논문은 2025년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임
(No.P0029861, 디지털전환 확산지원체계 구축 사업, 내역사업 : DX 실행모델 개발·적용)

요 약

대규모 언어모델의 고도화 및 확산에 따라 한국어가 적용된 언어모델 또한 다양한 분야에서 적용, 활용되고 있다. 이는 다양한 서비스 확장과 산업 응용 가능성을 증대시키고 있으며, 특히 전문 지식이 요구되는 분야에서 정보 제공 및 지원 도구로서의 역할이 점차 중요해지고 있다. 하지만 숙련요소의 데이터화는 전문가의 개입이 필요하며 많은 작업 공수가 필요하다. 본 연구에서는 HITL(Human-in-the-Loop) 접근 방식을 활용하여 정밀 금속가공 업종에서의 전문 지식 및 숙련공의 암묵지와 노하우를 효과적으로 모델에 반영하는 방법을 제안한다. 언어모델을 파인튜닝하고, 전문가의 반복적 검증과 피드백을 통해 모델 성능을 최적화하여 가공 공정 관련 가이드를 위한 파인튜닝 기반 챗봇용 언어모델 학습을 연구하였다. 본 연구를 통하여 언어모델 파인튜닝에서의 HITL 적용 시 모델의 특성에 따른 비언어적 특성을 포함한 데이터 학습 성능을 검증하였다.

Abstract

With the advancement and proliferation of Large Language Models, language models applied to the Korean language are also being applied across industrial fields. This trend has expanded into factory floor, service applications and other industrial areas, particularly in areas requiring tacit knowledge including know-how, where such models are increasingly significant as tools for information provision and support. However, digitalization of tacit knowledge often requires expert involvement and significant effort. This study proposes a method to effectively incorporate tacit knowledge of technicians and experienced workers in the precision machining industry into models using a Human-in-the-Loop (HITL) approach. By fine-tuning a Korean language-based model and employing iterative feedback from human experts, we studied a language model with tacit knowledge to support work guides.

Keywords

human in the loop, curriculum learning, tacit knowledge, Language model, KoGPT, KoBart

* 한국전자기술연구원(*[1]교신저자)
- ORCID[1]: https://orcid.org/0000-0002-9725-7990
- ORCID[2]: https://orcid.org/0000-0003-0244-6914
- ORCID[3]: https://orcid.org/0009-0007-5628-2878
- ORCID[4]: https://orcid.org/0009-0006-9762-8583

· Received: Jan. 23, 2025, Revised: May 16, 2025, Accepted: May 19, 2025
· Corresponding Author: Seongsu Jhang
Dept. of ICT Convergence Research Center, KETI
Tel.: +82-55-719-2352, Email: jss0221@keti.re.kr

Ⅰ. 서 론

최근 OpenAI의 ChatGPT와 구글에서 공개한 BARD 등 글로벌 AI 선도기업에서 앞다퉈 발표하고 있는 LLM(Large Language Model) 활용 서비스 이용자가 급격히 증가하고 있으며 그 활용 분야 또한, 빠르게 확장되고 있다. 특히, 일부 기업에서는 오픈소스로 개발된 소스코드와 모델들을 공개하여 Pre-trained Model을 조정하여 특정 목적에 맞는 모델로의 개발(Fine-tuning) 및 확산도 가속화되고 있다. LLM 서비스 중 하나인 챗봇(Chatbot)은 사용자와의 자연스러운 대화를 통해 정보 제공의 목적으로 개발되어 점차 복잡화, 고도화되고 있는 제조산업 분야의 설비, S/W 등의 운영, 유지보수에서의 활용도가 매우 높을 것으로 전망되고 있다.

국내의 경우, 고령화에 따라 제조 산업 내 숙련공들의 은퇴로 인해 전문 인력이 감소하는 상황에서, 작업 노하우(암묵지)가 필요한 공정 내 의사결정 요소, 설비 유지보수 등의 업무에 고령 또는 외국 국적의 숙련공들에 대한 의존도를 낮추기 위해 AI 도입에 대한 관심이 대두되고 있다. 그러나 대다수 중소, 중견기업에서는 핵심 공정에서의 암묵지와 노하우 요소 디지털화 및 AI 솔루션을 개발, 운영할 수 있는 인력이 부족한 실정이다. 그리고 다양한 유형의 데이터 활용을 위한 방안과 학습에 대한 검증과 사례연구 또한 부족하다. 본 논문에서는 대표적인 생성형 한국어 언어모델인 KoGPT-2와 KoBART를 활용하여 정밀금속가공 업종에서의 CNC(Computerized Numerical Control) 설비 운영, 유지보수 등에 대한 공정 단계별 기초지식, 암묵지 등을 데이터화하여 기존 모델을 파인튜닝 및 성능평가를 수행했다.

한편, 본 논문에서는 전술한 도전 과제를 해결하기 위해 HITL(Human-in-the-Loop) 접근법을 중심으로 한 언어모델 학습 방안을 제안하고 이를 실증적으로 검증했다. HITL 방식은 인간 전문가의 지식과 판단을 AI 모델의 개발과 학습 과정에 통합함으로써 모델의 신뢰성과 실효성을 높이는 방법론이다. 특히, 정밀금속가공 업종과 같은 고도의 전문성을 요구하는 분야에서는 암묵지와 경험적 지식을 효과적으로 디지털화하고 활용하는 데 있어 HITL이 중요한 역할을 한다.

2장에서 세부적으로 다루는 문헌조사 기반의 데이터 수집, 전처리와 전문가를 통한 데이터 보완 과정은 각각 HITL 커리큘럼 러닝 적용을 위하여 1, 2차로 구분하여 데이터를 확보하였다. 1차적인 데이터 수집 과정은 문헌 조사(국가직무능력표준 (NCS, National Competency Standards) 학습 모듈)를 통해 CNC 밀링, 선반 가공, 유지보수 및 CAM(Computer Aided Machining) 프로그래밍과 같은 핵심 작업 절차와 공정 데이터 수집했다. 그리고 2차 과정에서는 숙련공 및 전문가로부터 데이터를 보완하는 절차를 수행했다. 한편, HITL은 전술한 2차 과정으로부터 데이터의 추가 보완 작업이 수반된 데이터 분류, 파인튜닝에서의 Epoch 단계별 학습 과정에서 전체 데이터 일괄 학습과 커리큘럼 러닝에 기반한 단계적 학습으로 그 차이를 비교하는 방법론을 적용하고자 한다.

각 언어모델의 파인튜닝을 통한 학습 결과를 언어모델 성능 평가에 주로 사용되는 BLEU(Bilingual Evaluation Understudy)와 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)를 적용했다. 데이터 발췌를 위해 활용한 NCS 학습모듈은 금속가공 업종의 주요 작업에 대한 체계적인 매뉴얼과 절차를 포함하며, 전문 지식과 실무 경험이 반영된 자료로써 NCS 학습모듈 기반 다양한 업종, 분야에서 기초연구 활용 및 학습활용도 검증에 관한 연구가 다양하게 진행되었다[1]-[3].

Ⅱ. 데이터수집 및 전처리

2.1 데이터수집 및 커리큘럼 설계

HITL은 전문가의 도메인 지식 기반의 머신러닝 프로세스를 제어, 머신러닝과의 상호작용(Interaction)을 통해 능동학습을 수행하는 것을 의미하며 결과적으로 학습의 성능을 향상시키기 위한 방법론으로 정의된다[4]. 한편, HITL 중 커리큘럼 러닝(Curriculum learning)은 전문가의 판단을 통해 데이터셋을 구조화하고 학습 난이도를 설계하기 위해 주로 활용되고 있으며 이를 통해 학습 과정을 더욱

효과적으로 향상시키며 데이터 및 단계별 학습 난이도 등 응용 가능성을 높일 수 있는 방법이다[5].

본 논문에서 다루고 있는 중점 과제에서 가장 중요한 숙련공의 암묵지, 숙련 노하우 등의 데이터를 취득, 정제하기 위해 1차적인 데이터 수집 방법인 문헌정보 수집 결과를 기반으로 숙련공의 데이터 검토, 확인 절차를 적용한다. 이에 따라 언어모델의 순차적 학습 커리큘럼에 보완된 데이터셋을 반영하여 다중 작업 학습에서 작업 간 구조를 효과적으로 활용하기 위해 그림 1과 같이 2단계의 커리큘럼 러닝 구조를 설계하였다. 쉬운 작업에서 어려운 작업으로의 순서를 따르는 커리큘럼 학습이 임의 순서 또는 동시 학습보다 더 나은 성능을 보이기 때문에[6] 본 연구에서는 2단계 구조의 데이터별 커리큘럼 러닝을 적용하여 정밀 금속가공 업종에서의 언어모델 학습 성능을 극대화하기 위한 체계적인 접근법을 제안한다. 기존의 단편적인 작업 절차, 매뉴얼 등의 데이터에서 숙련공의 암묵지를 추가한 심층 데이터를 반영한 HITL인 커리큘럼 러닝에 적용한 연구 사례는 아직 드물다. 언어모델의 파인튜닝 과정에 이를 적용함으로써 학습 과정에서 언어모델이 해당 업종에 대한 전문 용어 및 작업 절차에 대한 매뉴얼 학습, 이후 커리큘럼에서의 암묵지 학습이 성능에 영향을 줄 수 있는지 실험적으로 확인 및 검증하였다. 이를 위하여 데이터의 품질을 기준으로 커리큘럼 설계 시 모델별 성능에 영향을 줄 수 있음을 전제로 커리큘럼에 반영되는 데이터 요소들을 분리하여 학습에 적용하는 방법론을 반영했다[7].

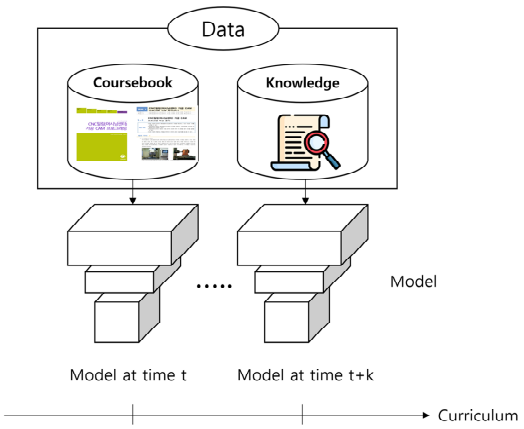


그림 1. 커리큘럼 러닝 개요
Fig. 1. Concept of curriculum learning

문헌 정보로부터 취득한 텍스트 데이터는 CNC 밀링, 선반 가공 작업, 유지보수, CAM 프로그래밍 등의 정밀금속가공 공정 전 주기에 대한 작업 절차, 유의사항 등 학습 내용을 포함한 NCS 학습 모듈(LM1502010408, LM1502010403, LM1502010405, LM1502010410, LM1502010409)에서 발췌하였다. 이 모듈들은 각 작업에 대한 작업 내용 정의(예: CNC 밀링 작업의 절차적 흐름), 작업 절차(예: 선반 가공 작업의 단계별 상세 설명), 그리고 각 작업 수행 시 유의사항(예: 유지보수 중 사고 방지 팁) 등이 포함되어 있다. 이러한 정보들을 정밀 금속가공 업종의 전문 지식과 노하우를 체계적으로 학습하는 데 적합한 기초 자료로 활용하여 252쌍의 데이터셋을 추출하였다. 그리고 그림 2에서와 같이 전문가 자문 및 현장 정보수집 방법을 적용하여 전문용어 및 기술적 표현의 정확성과 데이터셋에서 누락된 숙련 노하우 요소 보완 등 숙련공의 검토과정을 수행, 252쌍의 데이터를 보완하여 총 328쌍의 질문, 답변 학습 데이터셋을 구축하였다. 전문용어는 NCS 학습 모듈 등의 문헌에 포함된 용어(에어컷, 포스트 프로세서, 커터, 리머, 포켓, 홀 등 공구 명, 작업 용어 등)를 포함하며 세부적으로는 일부 암묵지 요소가 포함된 문맥 또는 문장(예: 공작기계 자동운전 실행 단계 : 공구 이송 남은거리 확인, 실제 거리 확인시 이상 없을 경우 single block 해제, coolant flood 동작 후 cycle start 진행) 등을 의미한다.

그림 2. 전문가 인터뷰(설문) 및 현장 공정 정보 수집
Fig. 2. Expert interviews(survey)) and on-site process information collection

2.2 데이터 전처리

데이터의 전처리는 Q&A 및 암묵지를 포함한 작업 절차에 대한 기술내용으로 구성된 데이터 쌍으로 JSONL(JavaScript Object Notation Lines) 포맷으로 구성하였다.

세부적으로는 키워드에 대한 세부 내용 설명과 키워드 중심의 질문, 답변 쌍으로 이뤄져 있으며 그림 3의 예시와 같이 작업 매뉴얼, 절차, 주의사항, 용어에 대한 정의 등에 대한 질문(또는 키워드)과 답변으로 구성된 데이터 쌍을 나타낸다. 데이터 포맷은 JSONL 포맷으로 각 Q&A 또는 Prompt, Completion 데이터 쌍을 독립적인 JSON 객체로 저장하며, 파일 내 각 줄은 하나의 JSON 객체를 표현한다. JSON 객체를 적용한 데이터 필터링 및 정제 과정에서는 중복된 질문, 불필요한 정보를 제거하여 데이터의 간결성과 유효성을 확보했으며 커리큘럼 2단계에 해당되는 숙련공의 데이터 검토, 보완 과정에서 실무에서 사용가능한 정보로 내용을 보충하거나 수정하였다. 이 과정에서 다음 장에서 다루는 HITL 기반 데이터 품질보증, 커리큘럼 러닝을 위한 데이터의 보완 과정이 진행된다. 또한, 이를 통하여 언어모델의 단계적 학습에 필요한 해당 분야, 업종에서의 숙련요소가 기존 NCS 학습모듈 기반의 데이터셋에 반영되는 절차를 거치게 된다.

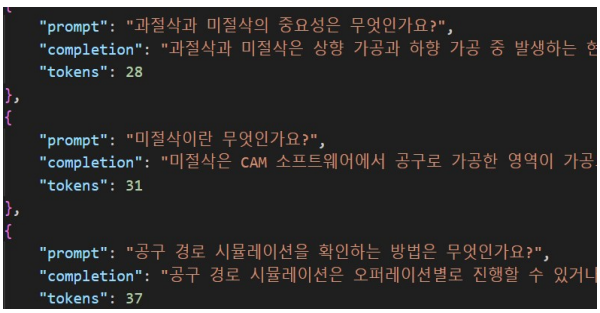


그림 3. 학습(Fine-tuning)용 데이터 쌍 예시 발췌
Fig. 3. Fine-tuning data examples

III. 한국어 언어모델 학습, 검증

3.1 언어모델 선정 및 파인튜닝 조건

파인튜닝 기반 학습에 사용한 Pre-trained 언어모델은 KoGPT-2와 KoBART 모델을 사용했다. KoGPT-2는 SKT에서 자체개발한 한국어에 특화된 딥러닝기반 언어모델로, GPT-2 아키텍처를 활용하여 한국어 위키피디아와 한국어뉴스 등에서 수집한 20 기가바이트 이상의 한국어 문장을 학습해 언어모델의 성능을 높였다[7]. 한편, KoBART는 SKT에서 세 번째로 발표한 한국어 버전의 모델이다. 사전학습에 한국어 위키피디아, 뉴스, 책, 모두의 말뭉치 등 이전보다 더 다양한 문장으로 학습되었다. BART(Bidirectional and Auto-Regressive Transformers) 언어모델은 facebook AI(Artificial Intelligence)에서 개발하였으며 BERT와 GPT를 합친 시퀀스 투 시퀀스 구조로 되어있다[8]. GPT와 BART형 모델의 경우, 각 구조별 특징에 따라 텍스트에 대한 이해도, 텍스트 생성에 높은 성능을 보였으며[9]-[11] 두 모델 모두 한국어 데이터로 사전 학습되어 한국어 자연어처리 작업에서 우수한 성능을 나타냈다[7][11]. KoBART의 경우, KoBERT에 비해 2개 단어 이상의 단어쌍에서 우수한 성능을 보여 문장 요약에 우수한 성능을 보이는 BERT형 구조에 비하여 텍스트 생성 기능이 상대적으로 더 요구되는 텍스트형 숙련요소 가이드 시스템에 적합함을 알 수 있다[12].

파인튜닝은 Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz, 32GB RAM, NVIDIA Quadro P2000 GPU 환경에서 진행되었다. 파인튜닝은 Epoch를 10회씩 수행하였으며, 각 파라미터는 최적화가 필요하지만, warm-up ratio는 0.1, learning rate는 0.00003으로 설정하였다. 나머지 변수들은 참고문헌 [7][13]에 제시된 검증된 환경을 반영하여 적용하였다.

한편, 커리큘럼 반영 여부에 따른 파인튜닝 결과를 비교하기 위해, 2단계 구조를 고려하여 전체 Epoch의 50%가 진행된 시점에서 데이터를 적용하고, 이에 따른 결과를 평가하였다.

3.2 파인튜닝 결과 분석

각 모델별 커리큘럼 러닝 적용 전, 후 Epoch, 학습률, 그리고 기타 하이퍼파라미터를 동일하게 설정한 상태에서 파인튜닝, 검증을 수행했다. 언어모델 파인튜닝 결과에 대한 성능 평가는 모델에 입력된 프롬프트(가공 공정에 대한 절차 키워드 또는 질문)로부터 생성된 텍스트(암묵지 요소를 포함하고 있는 작업 가이드 및 질문에 대한 답변)를 중심으로 자동 평가지표인 BLEU, ROUGE 지표를 사용하여 수행하였다.

BLEU는 IEC 24029에서 다루는 모델의 강인성 평가지표 중 하나로 생성된 텍스트와 참조

(Reference) 텍스트 간의 단어와 구문의 유사성을 측정하는 정량적인 평가 방법으로써 다양한 프롬프트, 여러 조건에서 얼마나 안정적으로 성능을 유지하는지 확인하는데 활용된다[14]. 구문의 길이(n-gram) 별 말뭉치의 유사성을 확인하는 방법론을 적용하여 생성된 텍스트와 비교 구문 텍스트 간의 유사성을 비교하여 수치화하는 방법론이다[15]. 아래의 식 (1)~(2)은 BLEU score를 산출하는 식으로 생성된 텍스트 c가 참조 텍스트 r보다 짧을 경우 발생하는 패널티 BP와 생성된 텍스트에서 참조 텍스트와 일치하는 n-gram의 비율 p_n의 기하평균을 결합하여 계산한다.

$$BP = \begin{cases} 1 & \text{if } c \geq r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad (1)$$

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n) \quad (2)$$

ROUGE 지표 또한 언어모델의 성능 평가 방법으로 활용되며, unigram, bigram 등 문장 간 중복되는 n-gram을 비교하여 평가한다. ROUGE-L은 식 (4)~(6)과 같이, LCS(Longest Common Subsequence) 기반을 활용하여 최장 공통 부분 수열을 기준으로 문장 간 유사도를 측정하는 방식이다. 식 (3)은 n-gram 기준으로 ROUGE 스코어를 구하는 식이며 Cnt는 카운트되는 수를 의미하며, 생성 텍스트와 기준이 되는 레퍼런스 텍스트 간의 매칭되는 비율 합으로 산출되고 식 (4)~(6)은 각각 재현율, 정밀도, F1-score를 산출하는 식으로 (4)와 (5)의 분모는 각각 참조, 비교군 텍스트의 모든 n-gram 수 이다[16]. 식 (3)은 참조 텍스트에 포함된 모든 n-gram 중에서 생성 텍스트와 일치하는 n-gram의 비율을 계산하는 것으로, ROUGE-N 점수를 정의한다. 이때 $Cnt_{match}(gram_n)$은 참조 텍스트와 생성 텍스트 간 일치하는 n-gram의 개수이며, $Cnt(gram_n)$는 참조 텍스트 내 전체 n-gram의 개수를 의미한다. 식 (4)는 생성된 텍스트가 참조 텍스트와 얼마나 많은 n-gram을 공유하는지를 나타내는 재현율(Recall)을 계산하는 수식이다. 분자는 생성 텍스트와 참조 텍스트 간 일치하는 n-gram의 개수이며, 분모는 참조 텍스트에 포함된 전체 n-gram의 개수를 의미한다. 이 지표는 참조 텍스트에 포함된 정보 중 얼마나 많이 회수되었는지를 정량적으로 평가한다. 식 (5)는 정밀도(Precision)를 나타내며, 생성된 텍스트가 생성한 n-gram 중 참조 텍스트와 실제로 일치하는 항목의 비율을 계산한다. 분자는 참조 텍스트와 일치하는 n-gram의 개수이며, 분모는 생성(Candidate) 텍스트 전체에 포함된 n-gram의 수이다. 이 지표는 생성된 텍스트가 얼마나 불필요한 정보 없이 참조와 정확하게 일치하는지를 평가하는 기준이 된다.

$$ROUGE-N = \frac{\sum_{(S \in Ref)} \sum_{gram_n \in S} Cnt_{match}(gram_n)}{\sum_{(S \in Ref)} \sum_{gram_n \in S} Cnt(gram_n)} \quad (3)$$

$$Recall = \frac{Matched\, n-grams}{Total_{ref}\, n-grams} \quad (4)$$

$$\mathrm{Pr}ecision = \frac{Matched\, n-grams}{Total_{can}\, n-grams} \quad (5)$$

$$F_1 = \frac{2\mathrm{Pr}ecision \cdot Recall}{\mathrm{Pr}ecision + Recall} \quad (6)$$

본 연구에서 두 가지 평가지표(BLEU와 ROUGE)를 적용하여 KoGPT-2와 KoBART 모델의 성능을 비교한 결과는 표 1과 같이 나타났다.

두 모델은 각각의 평가 지표에서 차별적인 성능을 보였으며, 특히 HITL(Human-in-the-Loop) 방식을 적용한 경우 각 모델 간의 차이가 더욱 명확하게 나타났다.

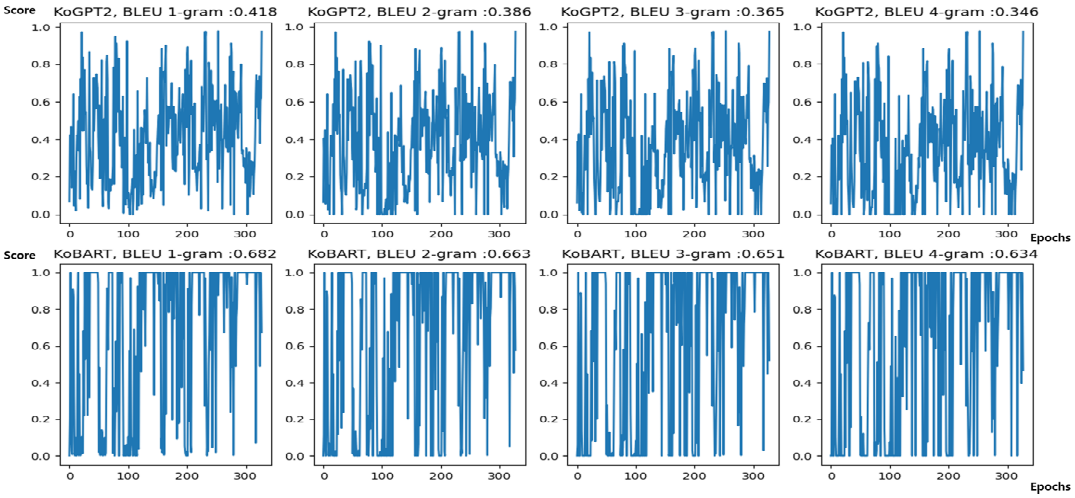
표 1. Pupil diameter 상관관계 상위 10개 변수
Table 1. Top 10 variables related to pupil diameter

Evaluation method		Without HITL		With HITL	
		KoGPT-2	KoBART	KoGPT-2	KoBART
BLEU 1-gram		0.418	0.682	0.439	0.775
BLEU 2-gram		0.386	0.663	0.407	0.763
BLEU 3-gram		0.365	0.651	0.385	0.755
BLEU 4-gram		0.346	0.634	0.365	0.739
ROUGE-1	Precision	0.378	0.383	0.329	0.446
	Recall	0.420	0.358	0.419	0.432
	F-1	0.371	0.357	0.330	0.429
ROUGE-2	Precision	0.244	0.188	0.208	0.239
	Recall	0.252	0.185	0.252	0.238
	F-1	0.229	0.183	0.199	0.235
ROUGE-3	Precision	0.154	0.077	0.137	0.122
	Recall	0.139	0.079	0.138	0.121
	F-1	0.136	0.075	0.118	0.119

KoBART의 경우, KoGPT-2에 비하여 두 지표 모두 높은 스코어를 보였으며 HITL 적용 시 테스트 프롬프트에 따라 생성되는 텍스트 말뭉치별(n-gram), 참조 레퍼런스 말뭉치와 비교하였을 때, 상대적으로 높은 유사도를 보여 더 나은 성능을 확인할 수 있었다. 결과적으로, HITL 방식을 통해 사람이 반복적으로 피드백을 제공하며 파인튜닝 데이터를 최적화한 결과, KoBART의 성능이 더욱 강화되었음을 확인할 수 있었다. 한편, KoGPT-2는 BLEU지표에서는 더 높은 스코어가 나타나는 경향이 있었지만 ROUGE에서는 근사하거나 더 낮은 수치를 보였다.

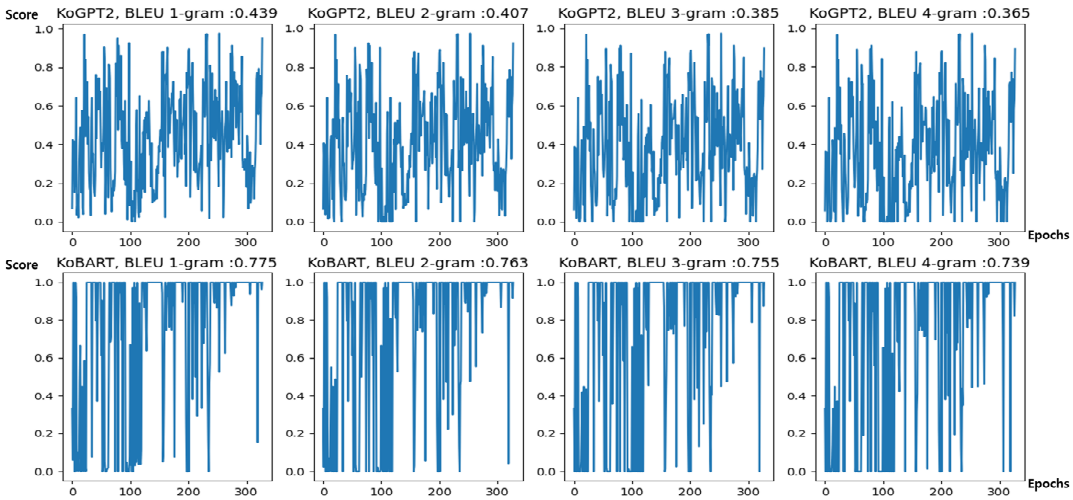
ROUGE 지표 내 재현율과 정밀도 측면에서는 대체적으로 KoGPT-2는 재현율, KoBART는 정밀도에서 더 높은 성능을 나타냈다.

전반적인 성능 수준은 KoBART 언어모델에서 생성한 텍스트가 레퍼런스 텍스트와 비교하였을 때 KoGPT-2 언어모델에서 생성한 텍스트보다 유사도가 더 높아 텍스트 형태의 숙련요소 학습 성능 측면에서 우수한 것으로 확인되었으며 HITL를 적용하였을 때, 각 모델별 성능향상에 대한 확인 결과는 평균 지표 기준, BLEU 스코어의 경우, KoGPT-2 모델은 5.34% 개선, KoBART 모델은 15.29% 개선 정도를 보였지만 ROUGE 스코어 기준, KoBART 모델은 26.31%로 파인튜닝시 성능이 개선됨을 확인하였다. 하지만 KoGPT-2 모델에서만 ROUGE 스코어 평균 -8.31% 성능 개악 결과가 나타났다.



(a) 커리큘럼 러닝 미적용 BLEU score (위: KoGPT-2, 아래: KoBART, X축: epoch, Y축: BLEU 점수)

(a) Normal learning (Top: KoGPT-2, Bottom: KoBART, x-axis: epochs, y-axis: score)

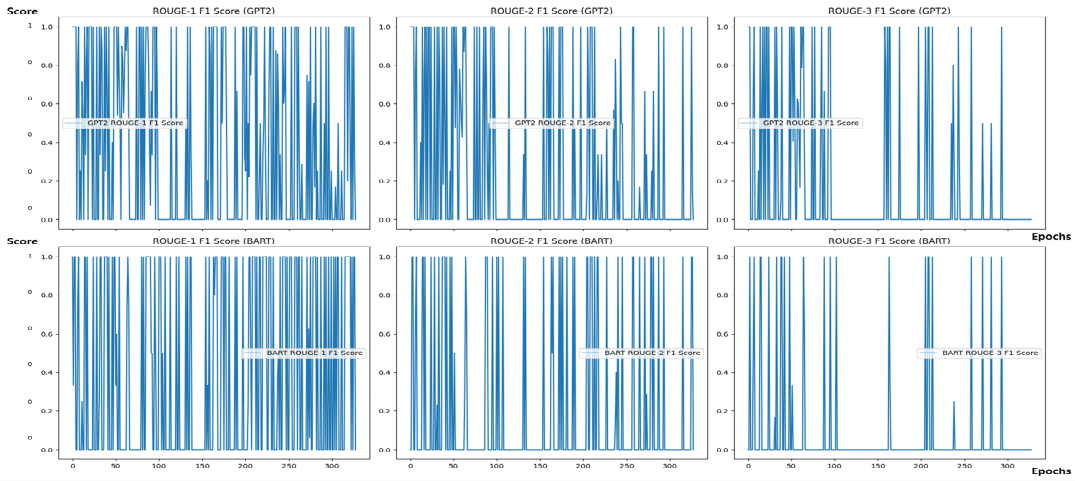


(b) 커리큘럼 러닝 BLEU score (위: KoGPT-2, 아래: KoBART, X축: epoch, Y축: BLEU 점수)

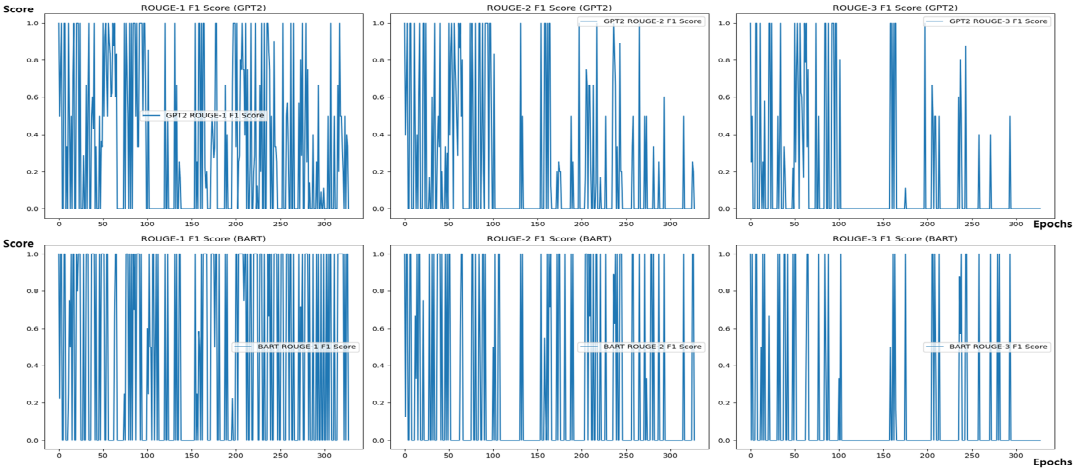
(b) Curriculumlearning (Top: KoGPT-2, Bottom: KoBART, x-axis: epochs, y-axis: score)

그림 4. BLEU(n-gram) score 결과

Fig. 4. BLEU(n-gram) score results



(a) 커리큘럼 러닝 미적용 F1 score (위: KoGPT-2, 아래: KoBART, X축: epoch, Y축: 점수)
(a) Normal learning F1 score (Top: KoGPT-2, Bottom: KoBART, x-axis: epochs, y-axis: score)



(b) 커리큘럼 러닝 F1 score (위: KoGPT-2, 아래: KoBART, X축: epoch, Y축: 점수)
(b) Curriculum learning F1 score (Top: KoGPT-2, Bottom: KoBART, x-axis: epochs, y-axis: score)

그림 5. ROUGE score 결과
Fig. 5. ROUGE score results

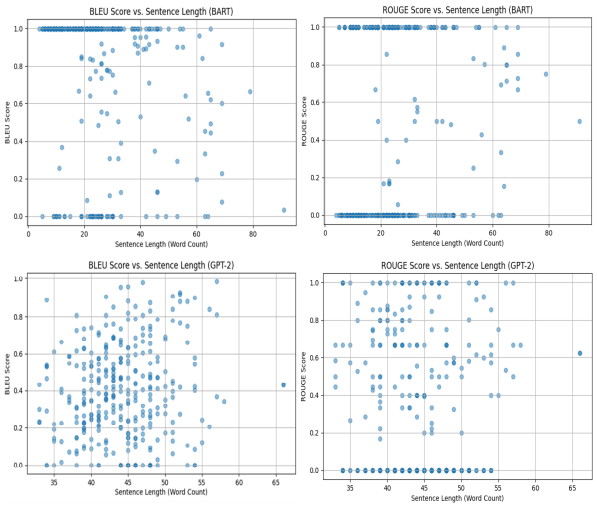


그림 6. 문장 길이별 지표별 점수 분포
Fig. 6. Score vs. sentence length (KoBART, KoGPT-2)

여기서 두 가지 자동 평가지표를 적용하여 성능을 비교하였을 때 KoGPT-2의 경우, KoBART에 비하여 HITL 적용 시, 성능이 저하되거나 더 적은 개선효과가 있었다. 이를 통하여 소량 데이터를 통한 파인튜닝에서 커리큘럼 러닝을 적용할 때 BART형 언어모델에서 그 효과가 더 큰 것을 확인할 수 있었다. 각 결괏값은 그림 4와 그림 5를 통해 확인할 수 있으며 파인튜닝 결과 BLEU, ROUGE 스코어를 나타낸다. x축은 각 프롬프트에 대한 모델의 생성 텍스트 데이터를, y축은 각 데이터에 대한 BLEU, ROUGE 스코어 산출 값을 의미한다.

그림 6은 문장 길이에 따른 BLEU 및 ROUGE 점수의 분포를 시각화한 결과를 나타낸다. 상단은

KoBART 모델의 결과이며, 하단은 KoGPT-2 모델의 결과이다. 전반적으로 KoBART의 경우 문장 길이의 분포 폭이 넓으며, 짧은 문장에서 BLEU 및 ROUGE 점수가 0 또는 1에 수렴하는 극단적인 분포를 보이는 경향이 관찰된다. 이는 비교적 짧은 문장에서 일부 생성 결과가 참조 문장과 완전히 일치하거나 전혀 일치하지 않는 경우가 존재함을 시사한다. 반면, KoGPT-2는 문장 길이 범위가 상대적으로 좁고, 전반적으로 점수가 중간 영역(0.3~0.8)에 밀집되어 있어 균일한 생성 성능을 보이는 것으로 판단된다. 특히, ROUGE 스코어의 경우 KoGPT-2는 문장 길이에 관계없이 일정 수준의 일관된 유사도를 보이는 반면, KoBART는 특정 길이(20단어 이하)에서 높은 점수를 빈번히 나타내는 경향이 있다. 이러한 결과는 모델 구조 및 사전 학습 방식의 차이에 기인한 것으로 해석될 수 있으며, 특히 KoBART는 짧은 문장 생성에 강점을 가지는 것으로 해석할 수 있다.

Ⅳ. 결론 및 향후 과제

머신러닝에서 전문가의 개입을 통한 모델의 성능 개선에 관한 연구가 활발히 이뤄짐에 따라 본 논문에서도 대표적인 한국어 언어모델을 대상으로 그 성능을 비교하였다. Fine-tuning 단계에서 데이터를 구분하여 적용한 HITL인 커리큘럼 러닝은 모델의 학습 효율성과 출력 결과의 품질을 높였음을 자동 평가지표인 BLEU와 ROUGE를 적용하여 비교했다. KoGPT-2와 KoBART 한국어 언어모델을 기반으로, 소량 데이터에 대한 파인튜닝을 진행하였으며, 언어모델의 파인튜닝 과정에는 커리큘럼 러닝의 후속 학습에서 숙련공의 정보가 담긴 데이터를 반영하였다.

이를 통하여 언어모델별 특성에 따른 학습 성능을 비교할 수 있었으며 자동 평가지표의 한계인 일치율 기준으로 평가하기 때문에 문장의 변형을 고려한 Human Feedback 방법론의 추가적인 적용이 필요함을 확인하였다.

HITL 방식의 도입은 단순히 모델 성능을 향상시키는 것을 넘어, 모델이 실제 현장에서 활용 가능하도록 하는 데 초점을 맞추고 있다. 본 연구에서 커리큘럼 러닝을 적용하여 파인튜닝된 KoGPT-2와 KoBART의 성능은 BLEU, ROUGE 모두 KoBART 모델이 KoGPT-2를 상회하는 결과를 보였다. 이러한 결과는 소량 데이터를 기반으로 한 파인튜닝에서 커리큘럼 러닝(HITL) 접근법이 BART형 모델의 성능뿐 아니라 실질적 활용성을 높이는 데 효과적임을 시사한다. 향후 연구에서는 암묵지 정보를 디지털화하여 데이터셋의 규모를 늘리고 HITL 접근법을 확장하여, 모델이 실시간 사용자 피드백을 기반으로 학습하고 개선될 수 있는 실시간 학습 메커니즘을 적용한 사례연구를 지속적으로 확장할 필요가 있다.

Acknowledgement

'2022년도 한국정보기술학회 추계종합학술대회에서 발표한 논문(아이트래커 기반 절삭공정 숙련도 관련 데이터 표준화)[17]을 확장한 것임'

References

[1] S. H. Seong, H. J. Min, and S. W. Park, "Analysis of Educational Needs for Vocational Counseling Services Using NCS Learning Modules", J. Vocational Education Research, Vol. 40, No. 5, pp. 51-75, Oct. 2021. https://doi.org/10.37210/JVER.2021.40.5.51.

[2] U. S. Kim and S. J. Jang, "The Impact of NCS Learning Module Application in Accounting on Job Competency of Human Resources in the Field", J. Human Resource Development Research, Vol. 21, No. 3, pp. 115-136, Sep. 2018. https://doi.org/10.24991/KJHRD.2018.09.21.3.115.

[3] S. H. Son and G. H. Yoo, "The Effect of Learning Satisfaction and Self-efficacy on Job Problem-solving Confidence in the National Job Competency Standard (NCS) Education of College Students Majoring in Skin Beauty", J. Korean Society of Cosmetology, Vol. 29, No. 6, pp. 1385-1395, Dec. 2023. https://doi.org/10.52660/JKSC.2023.29.6.1385.

[4] E. Mosqueira-Rey, et al., "Human in the Loop Machine Learning: A State of the Art", Artificial

Intelligence Review, Vol. 56, No. 4, pp. 3005-3054, Apr. 2023. https://doi.org/10.1007/s10462-022-10246-w.

[5] Y. Bengio, et al., "Curriculum Learning", Proc. 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada, pp. 41-48, Jun. 2009. https://doi.org/10.1145/1553374.1553380.

[6] A. Pentina, V. Sharmanska, and C. H. Lampert, "Curriculum Learning of Multiple Tasks", Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Boston, MA, USA, pp. 5492-5500, Jun. 2015.

[7] M. A. Lee, Y. J. Park, J. Y. Na, and C. B. Sohn, "Implementation of Review Sentiment Analysis Application Using KoBERT, KoGPT-2, and KoBART Optimized Hyperparameters", J. Digital Contents Society, Vol. 24, No. 11, pp. 2831-2840, Nov. 2023. https://doi.org/10.9728/dcs.2023.24.11.2831.

[8] H. R. Cho, Y. M. Lee, H. Y. Lim, J. W. Cha, and C. K. Lee, "Automatic Score Classification of Korean Learner's Writing Using Deep Learning-based Language Models – Focused on KoBERT and KoGPT-2", J. Korean Language and Culture, Vol. 18, No. 1, pp. 217-241, Apr. 2021.

[9] M. Asmitha, C. R. Kavitha, and D. Radha, "Summarizing News: Unleashing the Power of BART, GPT-2, T5, and Pegasus Models in Text Summarization", Proc. 2023 4th Int. Conf. on Intelligent Technologies (CONIT), Bangalore, India, pp. 1-6, Jun. 2024. https://doi.org/10.1109/CONIT61985.2024.10626617.

[10] O. W. Embarek and T. A. Benmusa, "Enhancing Deep Semantic Communication Systems (Deep-SC) Using the Pre-trained Models; BERT and BART", Proc. 2024 IEEE 4th Int. Maghreb Meeting of the Conf. on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA), Tripoli, Libya, pp. 515-519, May 2024. https://doi.org/10.1109/MI-STA61267.2024.10599693.

[11] N. G. Kim, H. J. Jo, J. H. Choi, B. K. Nam, and Y. C. Jung, "An Intensity Controlled Review Dataset Construction for Automatic Review Generation", J. Korean Society of Information Technology, Vol. 20, No. 2, pp. 157-165, Feb. 2022. http://dx.doi.org/10.14801/jkiit.2022.20.2.157.

[12] J. S. Jun and S. A. Lee, "Comparison of KoBART and KoBERT Models for Korean Paper Summarization", Proc. Annual Conf. on Human and Language Technology, Seattle, United States, pp. 562-564, Jul. 2022.

[13] S. W. Jeong, C. G. Kim, and T. K. Whangbo, "Question Answering System for Healthcare Information Based on BERT and GPT", Proc. 2023 Joint Int. Conf. on Digital Arts, Media and Technology with ECTI Northern Section Conf. on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Phuket, Thailand, pp. 348-352, Mar. 2023. https://doi.org/10.1109/ECTIDAMTNCON57770.2023.10139365.

[14] ISO/IEC 24029-1:2024, "Artificial Intelligence (AI) — Assessment of the Robustness of Neural Networks — Part 2: Methodology for the Use of Formal Methods", 2024.

[15] S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", Proc. 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311-318, Jul. 2002.

[16] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", Proc. Text Summarization Branches Out, Barcelona, Spain, pp. 74-81, Jul. 2004.

[17] H. J. Lee, S. S. Jhang, M. C. Song, Y. H. Song, J. Y. Kwon, J. P. Nam, and J. S. Lee, "Standardization of Eye-Tracker Based Data on Machining Skill Proficiency", Proc. KIIT Conf., Jeju, Korea, pp. 66-69, Dec. 2022.

저자소개

장 성 수 (Seongsu Jhang)

2017년 2월 : 광운대학교
전기공학과(공학사)
2019년 2월 : 광운대학교
전기공학과(공학석사)
2019년 2월 ~ 2021년 1월 :
(주)HD현대일렉트릭 연구원
2021년 2월 ~ 현재 :
한국전자기술연구원 선임연구원
관심분야 : ICT융합, 제조데이터, 산업 인공지능

유 동 휘 (Donghwi Yoo)

2021년 2월 : 광주과학기술원
기계공학과(공학사)
2023년 9월 : 광주과학기술원
기계공학과(공학석사)
2023년 9월 ~ 현재 :
한국전자기술연구원 전임연구원
관심분야 : 기계 예지보전,
고장진단, 데이터분석

권 재 영 (Jaeyoung Kwon)

2018년 2월 : 영남대학교
전기공학과(공학사)
2024년 2월 : 경남대학교
기계공학과(공학석사)
2020년 9월 ~ 현재 :
한국전자기술연구원 선임연구원
관심분야 : 공작기계, 데이터 표준,
OPC UA, 제조 데이터, 산업 인공지능

이 원 희 (Wonhee Lee)

2010년 2월 : 숭실대학교
정보통신전자공학부(공학사)
2013년 9월 : 숭실대학교
정보통신공학과(공학석사)
2013년 10월 ~ 현재 :
한국전자기술연구원 선임연구원
관심분야 : Wired/Wireless
Fieldbus system, 자율제조