지식재산연구 19권 3호, 2024년 9월, 155-180면
The Journal of Intellectual Property Volume 19, Issue 3, September 2024, Pages 155-180.
Print ISSN 1975-5945
Online ISSN 2733-8487

RESEARCH ARTICLE

Large Language Model-based R&D Solution Analysis Approach Using Problem-Solution Information of Patents

Seunghyun Lee[1], Jiho Lee[2], Seoin Park[3], Jae-Min Lee[4], Hong-Woo Chun[5], Janghyeok Yoon[6*]

[1]PhD Candidate, Department of Industrial Engineering, Konkuk University, Republic of Korea
[2]Director, AI Lab, Neopons Inc., Republic of Korea
[3]Master's Student, Department of Industrial Engineering, Konkuk University, Republic of Korea
[4]Principal Researcher, Future Technology Analysis Center, Korea Institute of Science and Technology Information, Republic of Korea
[5]Director, Future Technology Analysis Center, Korea Institute of Science and Technology Information, Republic of Korea
[6]Professor, Department of Industrial Engineering, Konkuk University, Republic of Korea

*Corresponding Author: Janghyeok Yoon (janghyoon@konkuk.ac.kr)

ABSTRACT

Patents, i.e., the output of research and development (R&D) activities, are regarded as a concentration of Problem–Solution information. Despite various patent analysis studies aimed at solving problems, large language model (LLM)-based studies are scarce. LLMs, which are effective for natural language processing tasks, such as text summarization and generation, have been applied in numerous fields, including healthcare, finance, and law. By learning the Problem-Solution information of patents as an LLM instead of merely examining existing R&D solutions, one can generate new solutions applicable to a specified problem. Therefore, this study proposes an approach to generate and analyze new R&D solutions using LLMs. Our systematic approach involves 1) collecting numerous patents and constructing a database; 2) extracting Problem-Solution information from the Common Application Form section of patents and constructing a Problem-Solution dataset; 3) fine-tuning an LLM using the problem-solution dataset and generating R&D solutions; and 4) analyzing R&D solutions to present a technology concept portfolio map. This study extends beyond the existing R&D solution exploration, presents a new approach for generating solutions, and suggests technology concepts using LLMs. Therefore, this study contributes to the expansion of the available options and fosters innovation in R&D field.

KEYWORDS

Patent analysis, Problem-Solution information, R&D solution, Large language model, Fine-tuning

Check for updates

Open Access

Citation: Lee S, et al. 2024. Large Language Model-based R&D Solution Analysis Approach Using Problem-Solution Information of Patents. The Journal of Intellectual Property 19(3), 155-180.

DOI: https://doi.org/10.34122/jip.2024.19.3.8

Received: July 2, 2024
Revised: August 6, 2024
Accepted: September 3, 2024
Published: September 30, 2024

Funding: The author received manuscript fees for this article from Korea Institute of Intellectual Property.

Conflict of interest: No potential conflict of interest relevant to this article was reported.

원저

특허의 Problem-Solution 정보를 활용한 대규모 언어모델 기반 R&D 솔루션 분석 방법*

이승현[1], 이지호[2], 박서인[3], 이재민[4], 전홍우[5], 윤장혁[6*]

[1]건국대학교 산업공학과 박사과정, [2]네오폰스 주식회사 AI Lab 연구소장, [3]건국대학교 산업공학과 석사과정, [4]한국과학기술정보연구원 미래기술분석센터 책임연구원, [5]한국과학기술정보연구원 미래기술분석센터장, [6]건국대학교 산업공학과 정교수

*교신저자: 윤장혁 (janghyoon@konkuk.ac.kr)

차례

국문초록

문제 해결을 위한 연구개발(R&D) 활동의 산출물인 특허는 Problem-Solution 정보의 집약체로 간주된다. 문제에 대한 해결방법을 탐색하기 위해 Problem-Solution 기반의 특허분석 연구가 다수 수행되었지만, 대규모 언어모델을 활용하기 위한 시도는 미흡한 실정이다. 텍스트 요약 및 생성 등 자연어 처리 작업에 효과적인 대규모 언어모델을 통해 특허의 Problem-Solution 정보를 학습한다면, 이미 존재하는 R&D 솔루션을 탐색하는 것에서 나아가 주어진 문제에 적용가능한 새로운 솔루션을 생성할 수 있다. 따라서, 본 연구는 대규모 언어모델을 활용하여 새로운 R&D 솔루션을 생성 및 분석하는 방법을 제시한다. 구체적으로 1) 대량의 특허를 수집하여 데이터베이스를 구축하고, 2) 특허의 공통출원서식 텍스트로부터 Problem-Solution 정보를 추출 및 데이터셋을 구축한 후, 3) Problem-Solution 데이터셋을 활용하여 대규모 언어모델을 파인튜닝하고 R&D 솔루션을 생성한 다음, 4) R&D 솔루션을 분석하여 기술 컨셉 포트폴리오 맵을 제시한다. 본 연구는 기존의 R&D 솔루션 탐색을 넘어 대규모 언어모델을 활용하여 솔루션을 생성하고 기술 컨셉을 제시하는 새로운 방법을 제시한다. 따라서 본 연구는 R&D 분야의 문제 해결 프로세스에서 선택의 폭을 넓히고, 언어모델 기반의 R&D 혁신을 촉진하는 데 기여할 수 있다.

주제어

특허분석, Problem-Solution 정보, R&D 솔루션, 대규모 언어모델, 파인튜닝

1. 서론

특허는 문제 해결을 위한 연구개발(R&D) 활동의 산출물로, 발명이나 아이디어의 단순한 기록을 넘어 기술적 문제와 그에 대한 해결책이 체계적으로 제시되는 신뢰할 수 있는 기술 문서이다.[1] 특허는 개발된 기술이 해결하고자 하는 문제와 해당 문제를 해결하는 방법에 대한 상세한 설명을 제공하므로, 다양한 산업 분야의 의사결정을 위한 문제 해결 프로세스를 원활하게 하고 새로운 R&D 활동을 위한 아이디어를 제공한다.[2] 따라서, 특허는 해결하고자 하는 문제와 그에 대한 해결방법을 제공하는 Problem-Solution 정보의 집약체로 여겨진다.

특허로부터 문제 해결 패턴을 파악하여, 주어진 문제에 대한 해결방안을 탐색하는 Problem-Solution 정보 기반의 특허분석 연구가 다수 선행되었다. 대표적으로는 특허 문서 내에서 Subject-Action-Object(SAO)와 같은 텍스트 구조를 추출 및 활용하는 연구들이 존재한다.[3][4][5] SAO 구조는 어떠한 객체를 변화시키는 작용으로 정의되는 기능을 표현하는 방법이다.[6] 특허 텍스트로부터 추출되는 SAO 구조는 발명 기술을 구성하는 요소 간의 연관관계를 나타내며, 특허의 핵심적인 개념과 발명자의 전문성 및 지식을 포함한다.[7] 이때, SAO 구조 내에서 AO 구조가 해결하고자 하는 문제를 나타내고 S가 문제에 대한 해결방안을 의미할 경우에는 특허의 SAO 구조를 문제-해결 형식으로 정의할 수 있다.[8]

특허의 텍스트뿐만 아니라 공통출원서식(Common Application Formation; CAF) 정보, 기술 분류코드, 출원인 정보 등의 메타 데이터를 함께 활용한 연구들도 수행되었다. 특허는 다른 문서들에 비해 다양한 메타 데이터를 포함하므로, 메타 데이터와 특허 텍스트를 복합적으로 활용하는 것은 특허의 발명기술에 대한 이해를 용이하게 한다.[9] 특허의 CAF는 발명의 내용을 적는 명세서, 청구범위 등 출원서식의 기재항목 및 순서를 국제적으로 통일화한 것으로 한국, 미국, 중국, 일본, 유럽 등 CAF에 관해 협약한 IP5 국가들에 적용된다. 따라서, 특허 출원 시 출원인은 특허 전문에 '발명의 명칭', '배경기술', '해결하고자 하는 문제', '과제의 해결 수단' 등 지정된 항목에 발명 기술 관련 내용을 서술해야 한다. 다양한 CAF 항목 중에서도 '배경 기술'과 '해결하려는 과제' 항목은 특허가 해결하고자 하는 Problem에 대한 정보를, '과제의 해결 수단'은 특허가 제시하는 Solution 정보를 제공한다.[10] 또한, 특허의 Problem-Solution 정보를

* 이 논문은 2024년도 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다.(과제번호: (KISTI)K24L3M1C2)

1) Janghyeok Yoon & Kwangsoo Kim, "Detecting signals of new technological opportunities using semantic patent analysis and outlier detection", *Scientometrics*, Vol.90 No.2(2012), pp. 445-461.
2) Yuen-Hsien Tseng et al., "Text mining techniques for patent analysis", *Information processing & management*, Vol.43 No.5(2007), pp. 1216-1247.
3) Sungchul Choi et al., "An SAO-based text-mining approach for technology roadmapping using patent information", *R&D Management*, Vol.43 No.1(2013), pp. 52-74.
4) Janghyeok Yoon & Kwangsoo Kim, "Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks", *Scientometrics*, Vol.88 No.1(2011), pp. 213-228.
5) Xuefeng Wang et al., "Identifying R&D partners for dye-sensitized solar cells: a multi-level patent portfolio-based approach", *Technology Analysis & Strategic Management*, Vol.31 No.3(2019), pp. 356-370.
6) Semyon D. Savransky, *Engineering of creativity: Introduction to TRIZ methodology of inventive problem solving*, CRC press, 2000, pp. 45-53.
7) 윤장혁.김광수, "SAO 기반의 의미론적 특허 유사성을 활용한 특허맵 생성방법", 「Entrue Journal of Information Technology」, 제10권 제1호(2011), 19-27면.
8) Martin G. Moehrle et al., "Patent-based inventor profiles as a basis for human resource decisions in research and development", *R&D Management*, Vol.35 No.5(2005), pp. 513-524.
9) Georg Richter & Andrew MacFarlane, "The impact of metadata on the accuracy of automated patent classification", *World Patent Information*, Vol.27 No.1(2005), pp. 13-26.
10) 이지호 외 4인, "특허의 Problem-Solution 텍스트 마이닝을 활용한 기술경쟁정보 분석 방법", 「지식재산

추출하기 위해 기술 분류코드도 활용될 수 있다. 정재민 외 2인[11]은 특허 청구항에서 질병과 같은 문제 개체를 추출하여 문제요소로 정의하고, 특허의 메타 데이터 중 기술 분류코드를 특허가 해결하고자 하는 문제에 대한 기술요소로 정의하여 비즈니스 기회를 발굴하였다.

앞서 소개한 바와 같이 Problem-Solution 관점의 특허분석 연구가 지속적으로 수행되어왔지만, 특허의 Problem-Solution 정보를 활용한 대규모 언어모델 기반의 특허분석 연구는 아직 미흡한 실정이다. 최근 인공지능의 발전과 더불어 대규모 언어모델 역시 빠르게 발전해왔으며, OpenAI의 Generative Pre-trained Transformer(GPT) 기반 대화형 대규모 언어모델인 ChatGPT가 출시된 후 대규모 언어모델에 대한 관심이 더욱 크게 증가하였다.[12] 대규모 언어모델은 텍스트 요약 및 생성, 정보 검색 및 추출 등의 자연어 처리 작업에 효과적인 것으로 검증되었으며, 이로써 사회의 다양한 분야에서 문제 해결을 위해 대규모 언어모델이 활용되고 있다. 예를 들어, 의료 분야에서는 전자 의료 기록에서 핵심 정보를 요약하기 위해 대규모 언어모델을 활용하였으며,[13] 건설 분야에서는 건설 프로젝트의 계약서를 효율적으로 검토하기 위해 대규모 언어모델을 활용하였다.[14] 이외에도 헬스케어, 교육, 법 등 수많은 사회 분야에서 업무 효율화 및 문제 해결을 위해 대규모 언어모델을 사용하였다.[15][16][17]

대규모 언어모델을 활용하여 특허의 Problem-Solution 정보를 학습한다면, 이미 존재하는 R&D 솔루션을 탐색하는 기존의 연구방법에서 나아가 주어진 문제에 적용가능한 새로운 솔루션을 생성하여 제시할 수 있다. 따라서, 본 연구는 대규모 언어모델을 활용하여 특허의 Problem-Solution 정보를 학습하고, 새로운 R&D 솔루션을 생성 및 분석하는 방법을 제시한다. 본 연구에서 제시하는 방법은 1) 대량의 국내 출원 특허를 수집하여 데이터베이스를 구축한 후, 2) 수집된 특허의 CAF 텍스트로부터 Problem-Solution 정보를 추출 및 데이터셋을 구축한 뒤, 3) Problem-Solution 데이터셋을 활용한 GPT 언어모델의 파인튜닝을 통해 R&D 솔루션을 생성하고, 4) 생성된 R&D 솔루션을 분석하여 주어진 문제에 대한 솔루션 인사이트를 도출하는 기술 컨셉 포트폴리오 맵을 제시한다. 제시하는 방법의 활용성을 보이기 위해, 본 연구는 국내 사회문제의 R&D 솔루션을 생성 및 분석하는 사례연구를 수행하고자 한다.

본 연구는 다음의 세 가지 기여점을 갖는다. 먼저, 본 연구는 새로운 R&D 솔루션 창출 방법을 제시하는 이론적 기여점을 지닌다. 본 연구에서 제시하는 방법은 대규모 언어모델을 통해 대량의 국내 출원 특허의 Problem-Solution 정보를 학습하므로, 기존의 R&D 솔루션 탐색 방법을 넘어 새로운 솔루션을 생성 및 제시하는 새로운 연구방법이다. 다음으로, 본 연구는 문제 해결 프로세스에서 선택의 폭을 확대하는 데에 실무적 기여를 제공한다. 본 연구에서 제시하는 방법은 주어진 문제에 대한 R&D 솔루션을 새롭게 생성하고, 포트폴리오 맵을 통해 참신하고 적용가능한 기술 컨셉을 제시한다. 따라서, 대규모 언어모델을 활용하여 다각적인 해결책을 생성하

연구」, 제13권 제3호(2018), 171-204면.

11) 정재민 외 2인, "비즈니스 기회 발굴을 위한 문제-해결방법 기반의 특허분석 방법",「지식재산연구」, 제15권 제2호(2020), 187-222면.

12) Wayne Xin Zhao et al., "A survey of large language models", arXiv preprint arXiv:2303.18223, 2023.

13) Dave Van Veen et al., "Adapted large language models can outperform medical experts in clinical text summarization", *Nature Medicine*, Vol.30 No.4(2024), pp. 1-9.

14) Saika Wong et al., "Construction contract risk identification based on knowledge-augmented language models", *Computers in Industry*, Vol.157(2024), 104082.

15) Jonathan H, Choi et al., "ChatGPT goes to law school", *Journal of Legal Education*, Vol.71 No.3 (2021), p. 387.

16) Yu Gu et al., "Domain-specific language model pretraining for biomedical natural language processing", *ACM Transactions on Computing for Healthcare*, Vol.3 No.1(2021), pp. 1-23.

17) Zhe Zheng et al., "Pretrained domain-specific language model for natural language processing tasks in the AEC domain", *Computers in Industry*, Vol.142(2022), 103733.

고 분석함으로써, 본 연구는 문제 해결 과정에서 실무자가 더욱 다양한 솔루션을 고려하고 최적의 방안을 도출할 수 있도록 지원한다. 마지막으로, 본 연구는 언어모델 기반의 R&D 기술 혁신을 촉진하는 데 기여한다. 본 연구의 대규모 언어보델 기반 R&D 솔루션 분석 방법은 이미 존재하는 솔루션을 탐색하는 방법에 비해 더욱 효율적이고 신속한 솔루션 개발을 지원할 수 있으므로, 본 연구는 사회 전반에서 R&D 솔루션 기술의 발전과 혁신을 가속화 할 수 있다.

본 논문은 2장에서 Problem-Solution 기반의 특허분석을 수행한 선행연구들을 소개하고, 3장에서는 연구방법의 구체적인 절차에 관하여 설명한다. 다음으로 4장에서는 사회문제를 대상으로 한 사례연구 결과를 서술하고, 마지막 5장에서는 본 연구의 기여점과 추후 연구 방향을 제시한다.

2. Problem-Solution 기반의 특허분석 연구

R&D 활동의 최종 산출물인 특허는 발명의 배경 기술과 관련된 문제를 명확히 정의하므로, 이를 통해 특허가 다루는 기술적 문제에 대한 이해를 돕고 문제의 중요성 및 해결이 필요한 이유를 설명한다. 또한, 정의된 문제를 해결하기 위한 발명 기술의 작동 방법과 기술적 세부 사항을 설명한다. 즉, 특허는 해결하려는 기술적 문제와 그에 대한 해결책을 제시하는 Problem-Solution 정보의 집약체로 간주된다. Problem-Solution 정보를 활용한 특허분석은 특정 문제에 대한 솔루션을 분석함으로써, 기존 솔루션을 개선하거나 해결되어야 하는 문제를 탐색하는 등의 R&D 전략을 수립할 수 있다. 또한, 다양한 문제 및 솔루션의 모니터링을 통해 서로 다른 기술 분야의 아이디어를 융합하여 새로운 혁신을 창출하는 기술 기반의 비즈니스 기회를 발굴할 수 있다. 따라서, 다양한 분야에서 전문가의 의사결정을 지원하기 위해 Problem-Solution 기반의 특허분석 연구가 활발히 수행되었다. Problem-Solution 기반 특허분석 연구는 크게 텍스트 구조 기반 연구와 메타 데이터 활용 연구로 분류할 수 있다.

다양한 선행연구들이 특허 텍스트로부터 SAO 구조를 추출하여 분석하였다. SAO 구조 기반의 특허분석은 특허 문서 내에서 주체(Subject), 작용(Action), 객체(Object)를 식별하여 기술 구성요소 간의 관계를 파악하고 기술 관련 주요 정보를 추출 및 분석한다.[18] 간단한 예시로 “Soap cleans hands.”라는 문장이 주어졌을 때, “cleans”(A)는 “soap”(S)와 “hands”(O)의 관계를 나타낸다. 이를 특허에 적용해보면 특허 문서의 SAO 구조에서 S는 문제를 해결하려는 주체를, A는 문제를 해결하기 위한 기술적 방법 및 해결책의 효과를, O는 해결하고자 하는 문제를 나타낼 수 있다.[19][20] 따라서, 특허 문서에서 추출된 SAO 구조는 Problem-Solution 정보로 정의할 수 있다.[21] 따라서, 특허의 SAO 구조를 기반으로 기술적 내용을 분석하는 다수의 연구가 선행되었다. Wang, X. et al.[22]은 특허의 SAO 구조를 바탕으로 유사한 특허를 찾기 위한

18) Hyunseok Park et al., “Identifying patent infringement using SAO based semantic technological similarities”, *Scientometrics*, Vol.90 No.2(2012), pp. 515-529.

19) Youngho Kim et al., “Automatic discovery of technology trends from patent text”, In Proceedings of the 2009 ACM symposium on Applied Computing, Association for Computing Machinery, 2009, pp. 1480-1487.

20) Yi Zhang et al., “How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence:“problem & solution” pattern based semantic TRIZ tool and case study”, *Scientometrics*, Vol.101(2014), pp. 1375-1389.

21) Martin G. Moehrle et al., “Patent-based inventor profiles as a basis for human resource decisions in research and development”, *R&D Management*, Vol.35 No.5(2005), pp. 513-524.

22) Xuefeng Wang et al., “Measuring patent similarity with SAO semantic analysis”, *Scientometrics*, Vol.121(2019), pp. 1-23.

분석 프레임워크를 제시하였다. 그들은 특허 초록 내 SAO 구조를 추출한 후, SAO 구조의 가중치를 기반으로 SAO 구조 간의 유사성을 분석하여 유사한 특허를 찾는 프레임워크를 개발하였다. Kim, S. & Yoon, B.[23]는 특허의 SAO 구조를 활용하여 특허 침해를 자동으로 식별하는 접근법을 개발하였다. 그들은 서로 다른 특허의 청구항에서 각각 SAO 구조를 추출한 뒤 SAO 구조의 S와 O를 특허 구성요소로 정의하였으며, SAO 구조 및 특허 구성요소를 활용한 특허지표들을 제시하였다. 서로 다른 특허의 SAO 유사도는 두 특허가 해결하려는 기술적 문제가 얼마나 유사한지, 두 특허가 문제에 대한 기술적 해결책이 얼마나 유사한지를 나타내었다. 이외에도 특허 구성요소 간의 유사도, 특허 구성요소 간의 대체용이성 등 특허지표를 바탕으로 특허 침해 가능성이 높은 특허 쌍을 식별하였다. 또한, Kim, K. et al.[24]은 특허 기술의 목적과 효과를 심층적으로 검토할 수 있는 새로운 방법인 SAOx(Subject-Action-Object-others)를 개발하였다. SAOx는 'for'와 'to' 구문을 통해 SAO 구조로 얻을 수 있는 정보를 확장하고 의미있는 기술정보를 식별한다. 그들은 특허 문서로부터 SAOx 구조를 추출한 후 특허 기술을 구성하는 토픽을 식별하고, 산업 및 조직 수준에서 토픽 간의 관계를 통해 기술 기회를 도출하였다.

특허의 메타 데이터를 활용한 Problem-Solution 기반 연구도 다수 수행되었다. 이지호 외 4인[25]은 특허 문서로부터 SAO와 같은 텍스트 구조를 추출하는 연구들이 특허 텍스트를 구성하는 항목의 특성을 고려하지 못했다는 한계점을 제시하였다. 그들은 이를 해결하기 위해 기업이 보유한 특허의 CAF 텍스트로부터 해결하고자 하는 문제와 해결방법을 추출 및 분석하여 기술경쟁 동향을 파악하는 새로운 접근법을 제시하였다. 특허의 CAF 항목 중 '배경기술'과 '해결하려는 과제'는 문제로, '과제의 해결 수단'는 문제에 대한 해결방법으로 정의되었으며, 정의된 문제와 해결방법 텍스트에 토픽 모델링이 적용되어 각 기업이 지니는 문제와 해결방법이 토픽으로 식별되었다. 기업별 토픽들은 Problem-Solution 네트워크로 표현되었으며, 네트워크 간의 비교를 통해 기업 간의 기술경쟁 정보가 파악되었다. 또한, 정재민 외 2인[26]은 특허의 청구항 텍스트와 메타 데이터 중 기술 분류코드를 함께 활용하는 Problem-Solution 관점의 특허분석 접근법을 제시하였다. 그들은 청구항으로부터 특허가 해결하려는 문제요소를 추출 및 정의하고, 문제를 해결하기 위한 기술요소는 특허에 부여된 기술 분류코드로 정의하였다. 특허 출원인별 문제 및 기술 포트폴리오를 구축한 후, 출원인의 기술 역량을 바탕으로 비즈니스 기회를 도출하였다. 이처럼 특허를 구성하는 기술요소는 문제를 해결하는 해법을 의미하므로, 많은 연구자들이 IPC 코드, CPC 코드, F-term 등 특허의 다양한 기술 분류체계를 통해 출원인의 기술적 역량을 나타내었다.[27][28][29]

앞서 소개한 바와 같이 Problem-Solution 기반의 특허분석 연구가 다수 선행되었지만, 주

23) Sunhye Kim & Byungun Yoon, "Patent infringement analysis using a text mining technique based on SAO structure", *Computers in Industry*, Vol.125(2021), 103379.

24) Kyuwoong Kim et al., "Investigating technology opportunities: The use of SAOx analysis", *Scientometrics*, Vol.118(2019), pp. 45-70.

25) 이지호 외 4인, "특허의 Problem-Solution 텍스트 마이닝을 활용한 기술경쟁정보 분석 방법", 「지식재산연구」, 제13권 제3호(2018), 171-204면.

26) 정재민 외 2인, "비즈니스 기회 발굴을 위한 문제-해결방법 기반의 특허분석 방법", 「지식재산연구」, 제15권 제2호(2020), 187-222면.

27) Youngjin Seol et al., "Towards firm-specific technology opportunities: A rule-based machine learning approach to technology portfolio analysis", *Journal of Informetrics*, Vol.17 No.4(2023), 101464.

28) Jaewoong Choi et al., "Technology opportunity discovery under the dynamic change of focus technology fields: Application of sequential pattern mining to patent classifications", *Technological Forecasting and Social Change*, Vol.148(2019), 119737.

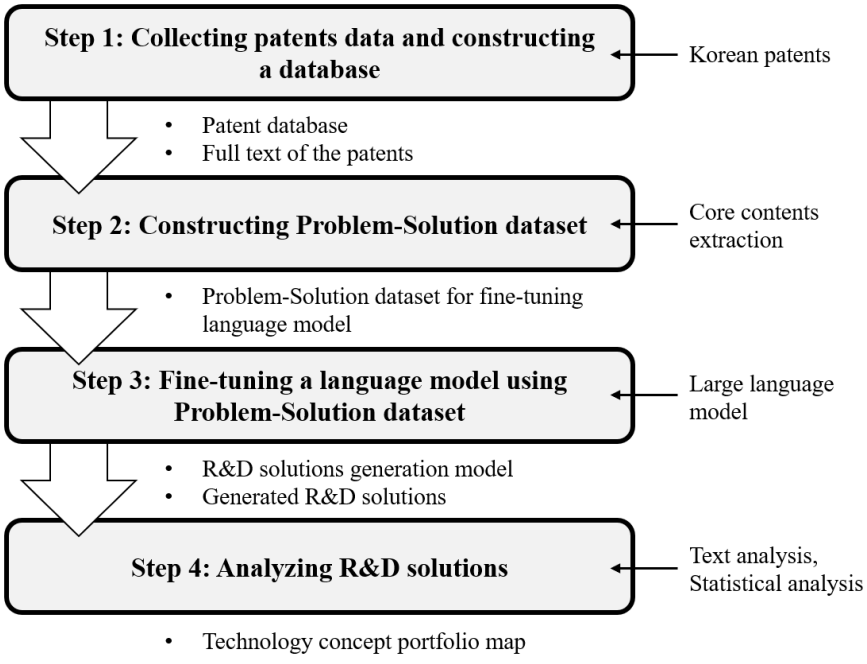
29) Jiho Lee et al., "An approach for discovering firm-specific technology opportunities: Application of link prediction to F-term networks", *Technological Forecasting and Social Change*, Vol.168(2021), 120746.

어진 문제에 대해 이미 정해져 있는 기존 해결책만 고려하는 한계점이 존재한다. Genrich Altshuller가 개발한 창조적 문제 해결 방법론인 TRIZ에 따르면, 다양한 문제들은 핵심이 되는 공통된 문제점으로 일반화가 가능하며, 다양한 해결방법 역시 일반화가 가능하다. 즉, 문제가 주어지면 어딘가에서는 이미 그 문제가 해결되었거나 유사한 문제의 해결방법이 존재하므로, 문제를 해결하려는 과정에서 다양한 방안을 고려해볼 수 있다. 따라서, 본 연구는 주어진 문제에 대한 기존 솔루션만 고려하는 것을 넘어, 대규모 언어모델을 활용하여 적용가능한 새로운 솔루션을 생성하여 제시하고자 한다.

3. 연구절차

본 연구가 제시하는 특허분석 방법은 다음과 같은 단계로 구성된다(그림 1). 1) 우선, 분석에 필요한 대량의 특허 데이터를 수집하고, 2) 수집된 특허의 CAF 항목을 구분짓고 CAF 텍스트로부터 Problem-Solution 정보를 추출한다. 3) 추출된 Problem-Solution 정보 데이터셋을 활용하여 대규모 언어모델을 파인튜닝하여 R&D 솔루션을 생성한 후, 4) 생성된 다양한 R&D 솔루션을 대상으로 텍스트 분석을 실시하여 R&D 솔루션을 구성하는 기술 컨셉을 도출한 뒤, 기술 컨셉에 대해 두 가지 평가지표 'Applicability'와 'Novelty'를 산출하여 기술 컨셉 포트폴리오 맵을 제시한다.

<그림1 연구절차>



3.1. 특허 데이터 수집 및 데이터베이스 구축

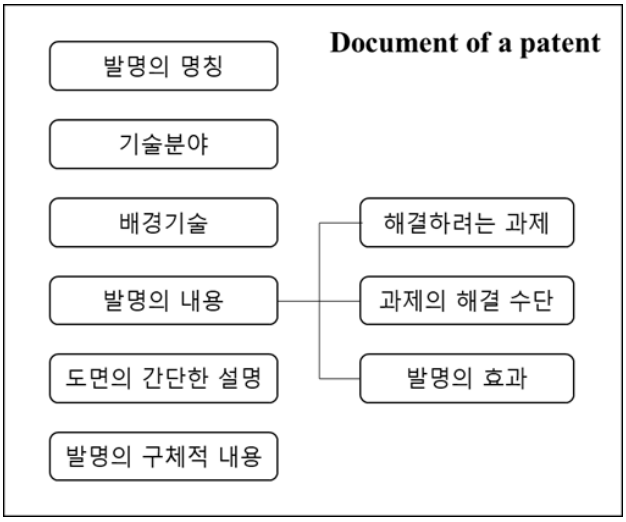
본 연구는 대량의 국내 특허 데이터를 수집하는 것으로 시작된다. 국내 특허 수집을 위한 방법으로는 Kipris Plus(https://plus.kipris.or.kr) 혹은 WIPS ON(https://www.wipson.com) 등의 온라인 특허 검색 및 활용 서비스를 활용할 수 있다. 수집 대상 특허는 기술 분류코드나

키워드 등으로 특정 기술분야에 제한되지 않고, 특허의 출원 날짜를 기준으로 선별된다. 한국, 미국, 유럽 등 IP5 국가들은 CAF에 관해 협약하였으며, 한국의 경우 2010년부터 특허 출원 시 출원인이 CAF 협약양식에 따라 '배경기술', '발명의 내용' 등의 지정된 항목을 특허명세서에 정형화된 형태로 작성하도록 규정하고 있다. 따라서, 본 연구는 특허의 CAF 항목 정보를 활용하므로 2010년 이후에 출원된 국내 특허를 대상으로 데이터를 수한다.

3.2. Problem-Solution 데이터셋 구축

본 단계에서는 수집된 국내 출원 특허의 특허명세서 내 CAF 항목 텍스트로부터 Problem-Solution 정보를 추출하고 데이터셋을 구축한다. 특허명세서를 구성하는 CAF 항목은 <그림 2>와 같으며, CAF 항목 중 '배경기술'과 '해결하려는 과제' 항목의 텍스트는 특허의 Problem으로, '과제의 해결 수단' 항목의 텍스트는 특허가 제시하는 Solution으로 정의할 수 있다.[30] '배경기술'은 선행 발명기술이 해결하지 못한 문제를 포함하고, '해결하려는 과제'는 본 특허의 발명기술이 해결하고자 하는 문제를 포함하기 때문에 Problem으로 정의될 수 있다. 또한, '과제의 해결 수단'은 특허가 Problem을 해결하기 위해 제시하는 발명기술에 대한 내용이므로 특허의 Solution으로 정의된다.

<그림2 특허 문서의 CAF 구조>



앞서 Problem과 Solution으로 정의된 CAF 항목의 텍스트 자체를 그대로 연구에 사용할 경우, 텍스트가 너무 길고 특허 별 텍스트 길이의 편차가 크기 때문에 분석 과정에 어려움이 존재한다. 따라서, 본 연구는 CAF 구조를 따르는 특허명세서의 문맥 정보를 바탕으로 특허의 Problem-Solution 정보를 추출하기 위해 사전학습 언어모델인 Bidirectional and Auto-Regressive Transformers(BART)를 활용한다. BART는 텍스트의 문맥 정보를 사전에 학습하여, 입력된 텍스트의 핵심 내용을 추출하는 언어모델이다.[31] 본 연구는 BART를 통해

30) 이지호 외 4인, "특허의 Problem-Solution 텍스트 마이닝을 활용한 기술경쟁정보 분석 방법", 「지식재산연구」, 제13권 제3호(2018), 171-204면.

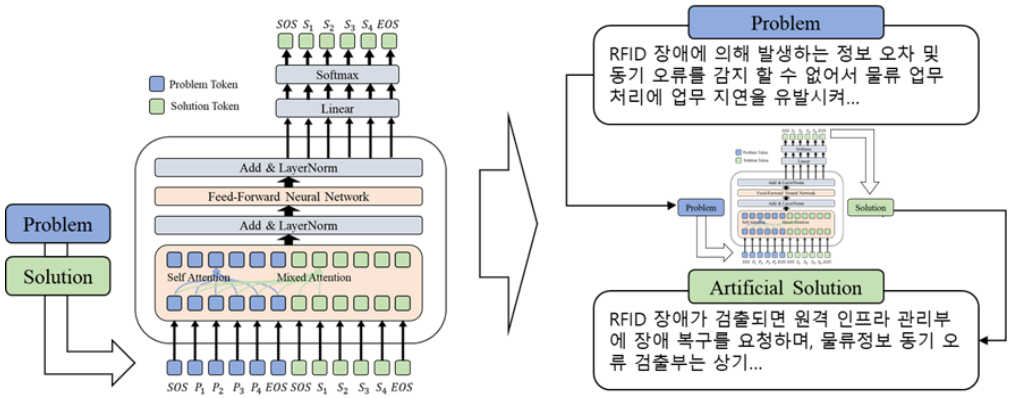
31) Mike Lewis et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension", arXiv preprint arXiv:1910.13461, 2019.

'배경기술'과 '해결하려는 과제' 항목의 핵심 텍스트를 도출하여 특허의 Problem으로, '과제의 해결 수단'의 핵심 텍스트를 특허의 Solution으로 정의한다. 최종적으로, 사전학습 언어모델 BART를 통해 각 특허의 Problem-Solution 쌍을 획득하여 대규모 언어모델의 파인튜닝을 위한 데이터셋을 구축한다.

3.3. Problem-Solution 데이터셋을 활용한 대규모 언어모델 파인튜닝

본 단계에서는 먼저 Problem-Solution 데이터셋을 활용하여 대규모 언어모델의 파인튜닝을 수행한다. 파인튜닝은 특정 도메인에 적합한 언어모델을 확보하기 위해, 사전에 학습된 대규모 언어모델에 해당 도메인의 데이터셋을 추가적으로 학습시키는 작업이다.[32] 본 연구는 Problem-Solution 데이터셋을 대규모 언어모델에 학습시킴으로써, Problem 텍스트가 입력되면 Solution 텍스트를 생성하는 R&D Solution 생성모형을 구축한다. Problem-Solution 쌍의 Solution은 기존재하는 특허로부터 추출되었으므로, 주어진 Problem에 대한 Existing Solution으로 정의할 수 있다. 그리고 R&D Solution 생성모형을 통해 도출되는 Solution은 입력된 Problem 텍스트의 문맥 정보를 바탕으로 생성되는 새로운 Artificial Solution으로 정의된다. 따라서, 본 연구는 Problem-Solution 데이터셋을 통한 대규모 언어모델의 파인튜닝을 통해 주어진 Problem에 대한 새로운 Artificial Solution을 생성한다(그림 3).

<그림3 Problem-Solution 데이터셋을 활용한 파인튜닝 및 Artificial Solution 생성>



다음으로, 본 연구는 R&D Solution 생성모형을 통해 해결하고자 하는 문제의 Artificial Solution을 생성하여, R&D Solution 생성모형의 텍스트 생성 성능을 평가한다. 보편적으로 생성모형의 성능을 평가할 때는 정답 텍스트와 생성된 텍스트 간의 일치하는 정도를 측정하는 BLEU, ROUGE, METOR 등의 평가지표가 활용된다. 하지만, 주어진 문제를 해결하려는 과정에서는 이미 정해진 기존의 해결방법뿐만 아니라 다양한 방안을 고려해볼 수 있으므로, 앞서 언급한 지표들을 통해 R&D Solution 생성모형을 평가하는 것은 적합하지 않다. 따라서, 본 연구는 생성된 Artificial Solution에 대한 정성적 평가를 위해 Sensibleness and Specificity Average(SSA) 지표를 도입한다.

SSA는 Google에서 발표한 자율 발화 모델에 대한 정성적 성능 평가지표로, 언어모델로부터 생성된 텍스트에 대해 Sensibleness와 Specificity 항목을 평가한다.[33] Sensibleness는 입력

32) 이치훈 외 2인, "사전 학습된 한국어 BERT 의 전이학습을 통한 한국어 기계독해 성능개선에 관한 연구", 「한국 IT 서비스학회지」, 제19권 제5호(2020), 83-91면.

된 텍스트에 대해 문맥적으로 합리적인 텍스트가 생성되었는지를 평가하여 0 또는 1의 값을 갖고, Specificity는 생성된 텍스트가 구체적인 내용을 포함하는지를 평가하여 0 또는 1의 값을 갖는다. 이때, Specificity는 Sensibleness가 1로 평가되었을 때 충분히 구체적인 텍스트가 생성되었는지를 살펴본다. 따라서, Sensibleness와 Specificity는 모두 0 또는 1의 값을 갖지만, Sensibleness가 0이면 Specificity도 0으로 설정된다. 생성된 텍스트에 대한 Sensibleness, Specificity 평가 예시는 <표 1>과 같다. 최종적으로, 평가자로부터 산출된 Sensibleness와 Specificity의 평균값을 계산하여 SSA 지표를 산출한다. 이로써 본 연구는 SSA 평가지표를 통해 R&D Solution 생성모형이 주어진 Problem에 대해 충분히 합리적이고 구체적인 Artificial Solution을 생성하는지에 대해 평가한다.

<표1 Sensibleness 및 Specificity 평가 예시>

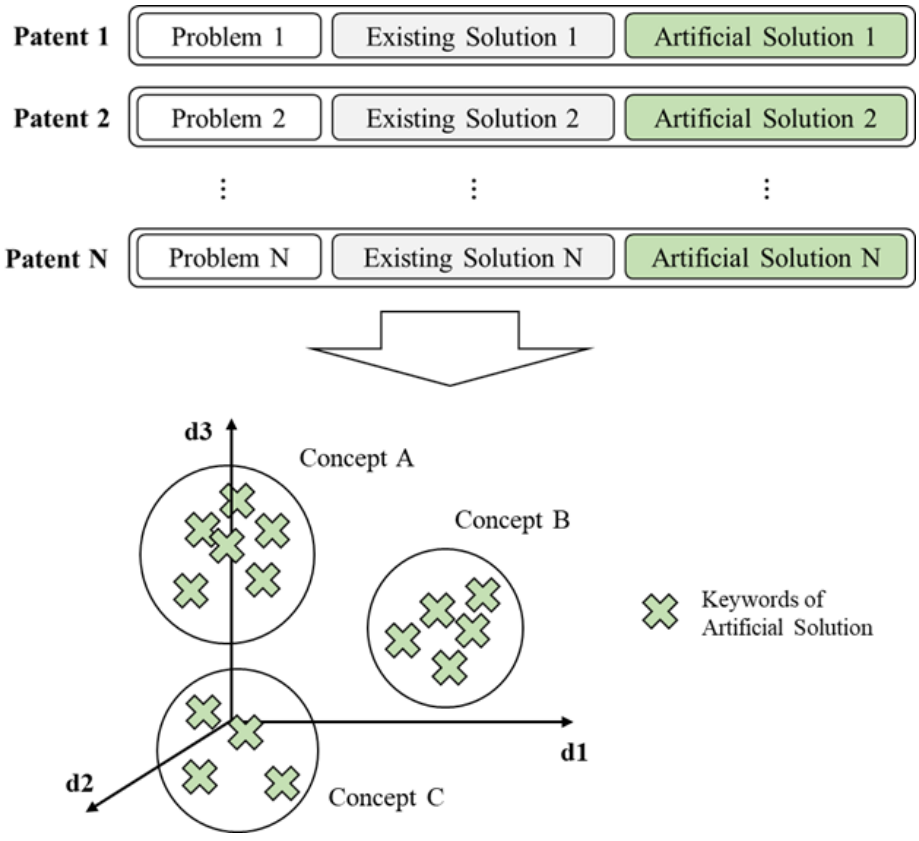
입력된 텍스트	생성된 텍스트	Sensibleness	Specificity
I love tennis.	That's nice.	1	0
	Tomorrow is my birthday.	0	0
	Me too, I can't get enough of Roger Federer!	1	1

3.4. R&D 솔루션 분석

본 단계에서는 해결하려는 문제에 대한 Artificial Solution을 효과적으로 살펴보고 문제 해결 과정에서 새로운 R&D 인사이트를 제공할 수 있도록, Artificial Solution을 구성하는 기술 요소의 그룹화를 통해 기술 컨셉을 도출한다(그림 4). 우선, 본 연구는 생성된 Artificial Solution의 기술 요소를 추출하기 위해 Solution들을 구성하는 명사구를 추출한다. 명사구는 Natural Language Toolkit,[34] KoNLPy[35] 등의 자연어 처리 기반의 키워드 추출 알고리즘을 적용하여 추출할 수 있다. 또한, 최근 텍스트의 문맥에 대한 이해도가 높아 정보 추출 작업에 활발히 사용되는 언어모델의 프롬프트 엔지니어링을 통해서도 명사구를 추출할 수 있다. 프롬프트 엔지니어링은 ChatGPT와 같은 대규모 언어 모델에 명확한 명령을 내리는 등 효과적인 대화를 통해 원하는 작업 결과를 얻도록 하는 기술이다.[36] 본 연구는 대규모 언어모델의 프롬프트 엔지니어링을 통해 Artificial Solution을 구성하는 다양한 명사구를 추출하여 기술 요소로 정의한다. 이어서, 추출된 기술 요소의 임베딩 벡터를 도출하고 벡터 간의 유사도를 측정하여, 유사한 기술 요소 벡터의 그룹화를 위해 클러스터링 알고리즘을 적용한다. 클러스터링은 분류되지 않은 데이터를 유사한 개체의 그룹(클러스터)으로 분류하는 대표적인 비지도 학습 기법이다. 본 연구에서 유사한 기술 요소들을 그룹화 한 각 클러스터는 Artificial Solution의 기술 컨셉으로 정의한다.

33) Daniel Adiwardana et al., "Towards a human-like open-domain chatbot", arXiv preprint arXiv: 2001.09977, 2020.
34) Edward Loper & Steven Bird, "Nltk: The natural language toolkit", arXiv preprint cs/0205028, 2002.
35) Eunjeong L. Park & Sungzoon Cho, "KoNLPy: Korean natural language processing in Python", In Proceedings of 26th Annual Conference on Human and Cognitive Language Technology, Special Interest Group of Human and Cognitive Language Technology, 2014.
36) Jules White et al., "A prompt pattern catalog to enhance prompt engineering with chatgpt", arXiv preprint arXiv:2302.11382, 2023.

<그림4 Artificial Solution의 기술 컨셉 도출>



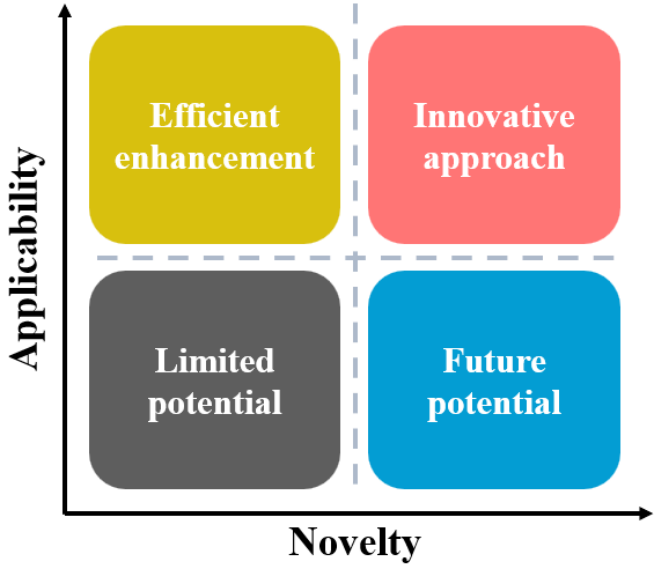
다음으로, Artificial Solution으로부터 도출된 기술 컨셉에 대해 두 가지 평가지표 'Applicability'와 'Novelty'를 산출하고 기술 컨셉 포트폴리오 맵을 제시한다. 두 가지 지표는 기술 컨셉을 구성하는 명사구가 추출된 Artificial Solution의 특허를 추적하여 산출된다. 기술 컨셉의 적용가능성을 의미하는 Applicability는 해당 기술 컨셉을 구성하는 기술 요소가 추출된 Artificial Solution의 수로 측정된다. 기술의 구성요소가 많을수록 해당 기술은 다양한 분야에 적용될 수 있다. 하지만 기술을 구성하는 수많은 요소가 오직 소수의 기술에서만 모두 등장했다면, 해당 기술이 높은 적용가능성을 지닌다고 해석하기에는 어려움이 존재한다. 따라서 본 연구는 기술 컨셉의 적용가능성을 더욱 정밀하게 산출하기 위해, 기술 컨셉을 구성하는 명사구가 실제로 등장한 Artificial Solution의 개수를 측정한다. 다음으로, 기술 컨셉의 참신성을 의미하는 Novelty는 해당 기술 컨셉의 기술 요소가 추출된 Artificial Solution과 Existing Solution 간의 벡터 거리의 평균으로 측정된다. 서로 유사한 의미를 갖는 텍스트의 임베딩 벡터일수록 벡터 공간에서 서로 가깝게 위치하며, 반대로 매우 이질적인 텍스트일수록 벡터 공간에서 멀리 위치한다. 따라서, 벡터 공간에서 Artificial Solution과 Existing Solution의 거리가 멀수록 기존 해결책과 유사하지 않은 참신한 해결방안이 생성된 것으로 해석할 수 있다.

기술의 포트폴리오 맵은 중요한 기술요소에 대한 정보를 제공하는 분석 도구이며,[37] 본 연구의 기술 컨셉 포트폴리오 맵은 Artificial Solution을 구성하는 기술 컨셉의 적용가능성과 참신성을 바탕으로 핵심 기술 컨셉을 나타낸다. 기술 컨셉 포트폴리오 맵은 기술 컨셉의

37) Sungchul Choi et al., "An SAO-based text-mining approach for technology roadmapping using patent information", *R&D Management*, Vol.43 No.1(2013), pp. 52-74.

Applicability와 Novelty의 평균값을 기준으로 4개 영역으로 구분된다(그림 5). Applicability와 Novelty가 모두 높아 Innovative approach 영역에 포함되는 기술 컨셉은 직면한 문제에 신속히 적용해볼 수 있으며, 동시에 기존 솔루션 대비 참신한 인사이트를 제공하므로 문제 해결의 혁신적인 접근법을 제시할 수 있다. Efficient enhancement 영역은 Applicability는 높지만 Novelty가 낮은 기술 컨셉이 포함된다. 참신성이 낮더라도 즉시 적용될 수 있으므로, 해결하려는 문제에 대한 효율적 개선이 가능하다. 반대로 Novelty는 높지만 Applicability가 낮은 Future potential 영역에는, 당장 적용하기 어려워도 연구개발을 통해 미래에 큰 잠재력을 가질 수 있는 참신한 기술 컨셉이 포함된다. 마지막으로 두 지표가 모두 낮은 Limited potential 영역은 현재 실질적인 응용이 어려우며 참신하지 않은 기술 컨셉들이 포함된다. 본 연구는 기술 컨셉 포트폴리오 맵의 결과와 도메인 전문가의 지식을 바탕으로 주어진 문제를 해결하기 위한 R&D 인사이트를 제공한다.

<그림5 기술 컨셉 포트폴리오 맵>



4. 사례연구: 사회문제 해결형 R&D

본 연구는 사회문제 해결형 R&D에 대한 사례연구를 수행한다. 사회의 급속한 발전으로 인해 다양한 사회문제가 등장하였으며, 과학 및 기술의 활용이 경제성장 뿐만 아니라 기후변화, 자원 부족 등의 사회문제를 해결하여 인간의 삶의 질을 향상시키는 방향으로 확대되었다.[38] 이에 따라 사회문제를 해결하기 위한 연구개발 활동인 사회문제 해결형 R&D가 주목받고 있다. 실제로 여러 국가에서 사회문제 해결형 R&D에 대해 공통된 관심을 갖고 있으며, 한국의 경우에는 '과학기술 기반 사회문제해결 종합계획'을 발표하여 43개의 사회문제를 제시하고 사회문제에 대한 실질적인 R&D 해결책을 강조하였다.[39] 따라서, 본 연구는 사회문제 해결형 R&D에 기여하기 위해, 우리나라의 43개 사회문제 중 사이버 범죄를 사례로 선정하여 관련된 Artificial Solution을 생성 및 분석한다.

38) Johan Schot & W. Edward Steinmueller, "Three frames for innovation policy: R&D, systems of innovation and transformative change", *Research policy*, Vol.47 No.9(2018), pp. 1554-1567.
39) 과학기술정보통신부, "제3차 과학기술 기반 사회문제해결 종합계획", 과학기술정보통신부, 2023, 1-101면.

4.1. 국내 특허 데이터 수집 및 Problem-Solution 데이터셋 구축

우선, 본 연구는 국내 온라인 특허 정보 검색 및 활용 서비스인 Kipris Plus (https://plus.kipris.or.kr)의 API를 활용하였다. 구체적으로, 최근에 출원된 특허들을 활용하여 R&D Solution 생성모형을 구축할 수 있도록 2020년부터 2022년까지 3년 동안 출원된 국내 특허 372,462건의 서지정보와 전문을 수집하여 데이터베이스를 구축하였다. 이때, 본 연구는 R&D Solution 생성모형이 다양한 분야에서 발생했던 Problem과 이를 해결한 Solution을 모두 고려하여, 주어진 문제에 대해 참신하고 새로운 솔루션을 제공할 수 있도록 기술 분야의 제한 없이 모든 출원 특허를 수집하였다. 수집된 특허는 모두 2010년 이후에 출원되었기 때문에 CAF가 적용되었다. 수집된 특허 전문 데이터는 XML 형식으로 제공되었으며, 본 연구는 XML 데이터의 파싱을 통해 CAF 항목별 텍스트를 식별하였다. CAF 항목 중 '배경기술'과 '해결하려는 과제'는 Problem으로, '과제의 해결 수단'은 Solution으로 정의되므로, 본 연구는 수집된 특허 전문 XML 데이터를 파싱하여 'background-art', 'tech-problem', 그리고 'tech-solution' 태그에 해당하는 텍스트를 수집하였다. 해당 태그명이 존재하지 않는 특허 전문 데이터는 제외하였다.

앞서 수집한 국내 출원 특허의 Problem-Solution 텍스트를 그대로 언어모델에 활용할 경우, 텍스트의 길이가 너무 길고 편차 역시 크다는 한계점이 존재한다. 따라서, 본 연구는 각 특허로부터 핵심적인 Problem-Solution 정보를 추출하기 위해 한국어 사전학습 언어모델 KoBART(https://github.com/SKT-AI/KoBART)를 활용하였다. KoBART는 40GB 이상의 한국어 텍스트를 사전에 학습한 언어모델로, 긴 한국어 텍스트의 핵심 내용을 추출할 수 있다. 우선 '배경기술'과 '해결하려는 과제'의 텍스트를 함께 KoBART에 입력하여 요약한 텍스트를 Problem으로 정의하였다. 다음으로 '과제의 해결 수단' 텍스트를 KoBART에 입력하여 도출된 요약 텍스트를 Solution으로 정의하였으며, 최종적으로 332,425건의 Problem-Solution 데이터셋을 구축하였다. 기존의 CAF 항목 텍스트와 구축된 Problem-Solution 데이터의 텍스트 길이는 <표 2>와 같으며, Problem-Solution 데이터셋의 예시는 <표 3>에 나타내었다.

<표2 CAF 항목 텍스트와 Problem-Solution 텍스트의 평균 길이>

CAF 항목	텍스트 길이 평균	구축된 데이터셋	텍스트 길이 평균
배경기술	2318.2	Problem	391.3
해결하려는 과제	704.9		
과제의 해결 수단	4551.2	Solution	816.6

<표3 Problem-Solution 데이터셋 예시>

출원번호	Problem	Solution
1020200010521	스마트폰의 보급이 증가하고, 스마트폰의 기술이 발전함에 따라 스마트폰을 이용한 링크연결이 자주 발생함에 따라 스마트폰의 링크연결을 이용한 범죄도 증가하고 있다.	본 실시에 따르면, 스미싱 방지장치가 스미싱을 방지하는 방법에 있어서 문자 메시지에 링크정보가 포함되어 있는 경우, 상기 문자 메시지에 대응하는 링크 에뮬레이터를 할당하고, 상기 링크정보를 링크 에뮬레이터에 대한 식별정보를 포함하는 보안 링크정보로 변환하는 변환과정; 및 상기 보안 링크정보를 포함하는 문자 메시지를 단말기로 전송하는 단말 통신부를 포함하는 것을 특징으로 한다.
1020200030977	급격하게 바뀌고 있는 산업화에 따른 결과물로서 자동차의 매연과 더불어 각종 공장에서의 매연과 더불어 미세먼지 등이 현대인들의 건강을 위협하고 있다.	상기의 목적을 달성하기 위해 본 발명은 렌즈가 장착되는 안구부와, 렌즈가 장착되는 안구부의 배면 측에 연결되는 안경다리부로 구성되는 에어커튼 기능을 가진 안경을 제공한다.
1020200108691	보이스 피싱은 전화를 통하여 신용카드나 계좌번호 등을 알아낸 뒤 이를 범죄에 이용하는 전화금융사기 수법으로 고령자들이 이러한 금융 사기의 주 대상자이고, 젊은 세대도 금융 사기 피해를 당하고 있다.	문자 메시지를 이용한 불법 금융 거래를 차단하기 위한 통신 단말은 경고 알림창을 표시하고, 문자 어플리케이션에 수신된 문자 메시지에 포함된 계좌번호를 비식별 처리할 수 있다.
1020210018162	전력 케이블은 지중, 공중에 위치하여 내부의 도체를 통하여 전력을 전달하기 때문에 케이블 내외부 각 매질의 열 특성 및 구조를 모두 고려해야 한다.	본 실시예에 의하면, 전력 케이블은 도체와 절연층 사이에 위치하는 내부 반도전층과, 절연층과 금속 시스층 사이에 위치하는 외부 반도전층을 포함한다.
1020220007594	건강기능식품이 각광받는 이유는 고령화 사회로의 진입, 식습관에 기인하는 고혈압, 당뇨병, 심혈관계 질환 등의 만성질환의 증가와 인스턴트 음식의 발달로 미량 및 건강 영양소의 섭취가 줄어들고 있기 때문이다.	본 발명에 따르면, 무청, 질경이, 백출, 겨우살이, 건강, 엉겅퀴, 다시마, 계피 및 차가버섯을 포함하는, 항염증 활성을 갖는 식품조성물이 제공될 수 있다.

4.2. R&D Solution 생성모형 구축 및 솔루션 생성

본 연구는 사전에 학습된 대규모 언어모델에 Problem-Solution 데이터셋을 추가적으로 학습시키는 파인튜닝을 진행하였으며, 사전학습 언어모델로는 한국어 텍스트 생성에 최적화된 KoGPT2(https://github.com/SKT-AI/KoGPT2)를 선정하여 사용하였다. Open AI의 GPT2는 주어진 텍스트의 다음에 위치할 단어를 예측할 수 있게끔 학습된 언어모델이다.[40] GPT2는 이전의 출력이 다음의 입력으로 사용되는 자기 회귀 모델로, 주어진 텍스트의 다음 단어를 예측하는 능력이 뛰어나다. KoGPT2는 이러한 GPT2의 한국어 성능을 보완하기 위해 국내 뉴스, 위키피디아, 국민청원 등의 한국어 텍스트를 학습하여 SKT AI 팀에서 공개한 오픈소

40) Alec Radford et al., "Language models are unsupervised multitask learners", OpenAI blog, <https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf>, 검색일: 2024. 06. 30.

스 언어모델이다. 다양한 한국어 사전학습 언어모델과 파인튜닝 API 서비스를 활용할 수 있지만, 본 연구는 수십만 건의 데이터셋을 활용하기 때문에 경제적 측면을 고려하여 오픈소스 모델인 KoGPT2를 활용하였다.

파인튜닝은 사전학습 모델을 새로운 목적에 맞게 변형하기 위해, 이미 학습된 모델의 가중치를 미세하게 조정하여 학습시키는 방법을 의미한다. KoGPT2는 한국어의 문장 구조가 사전 학습된 언어모델이며, 본 연구는 KoGPT2의 파인튜닝을 통해 주어진 Problem에 대한 Solution을 학습하여 R&D Solution 생성모형을 구축한다. 파인튜닝 과정에서 'KoGPT2-base-v2' 모델과 토크나이저가 활용되었으며, 자세한 하드웨어 및 소프트웨어 환경은 <표 4>와 같다.

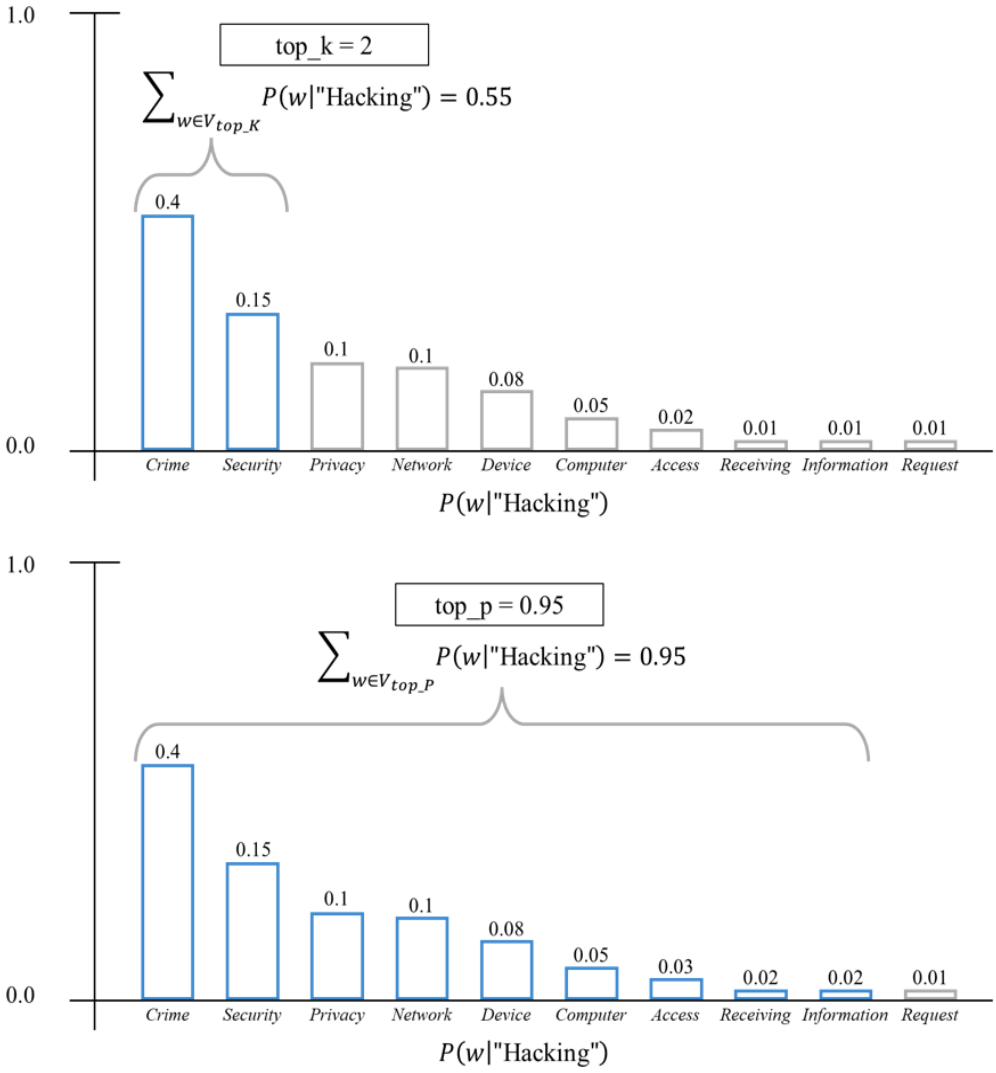
<표4 대규모 언어모델 파인튜닝을 위한 실험 환경>

구분		사용
하드웨어	CPU	AMD RyzenTM 9 7950 @ 4.5 Ghz
	GPU	NVIDIA GeForce RTX 4090
	RAM	64GB
소프트웨어	OS	Linux
	커널 버전	5.15.0-58-generic
	프로그래밍 언어	Python

먼저, 토크나이저를 활용하여 Problem과 Solution 텍스트를 각각 토큰 시퀀스로 변환하였다. Problem이 입력되면 그에 대한 Solution을 생성해야 하므로, 파인튜닝에 입력된 데이터는 <SOS 토큰 + Problem 토큰 시퀀스 + EOS 토큰 + SOS 토큰 + Solution 토큰 시퀀스 + EOS 토큰> 구조로 입력되었다. Start of Sequence의 약자인 SOS 토큰은 텍스트의 시작을 나타내고, 반대로 End of Sequence의 약자인 EOS 토큰은 텍스트의 끝을 나타낸다. 우선 <SOS 토큰 + Problem 토큰 시퀀스 + EOS 토큰> 구조를 통해 언어모델은 Problem의 토큰 시퀀스를 입력받고, 그 다음에 등장하는 <SOS 토큰>을 통해 Solution 토큰 시퀀스 생성을 시작해야 하는 위치를 식별한다. 따라서, 앞서 입력된 Problem에 대한 Solution 토큰 시퀀스를 생성하는 과정을 학습하고, 생성할 토큰이 없으면 <EOS 토큰>을 출력하여 종료한다. 즉, KoGPT2의 파인튜닝을 위해 Problem 토큰 시퀀스와 Solution 토큰 시퀀스가 함께 입력되었으며, Problem의 토큰 시퀀스가 끝나는 지점에서부터 함께 입력된 Solution 토큰 시퀀스를 생성하도록 학습되었다.

파인튜닝을 위한 옵티마이저(Optimizer)는 Adam이 사용되었으며, 학습률(Learning rate)은 1e-03으로 설정되었다. 거대한 학습 데이터의 규모를 고려하여 에폭(Epoch)과 배치(Batch) 사이즈는 모두 2로 설정되었다. 언어모델의 텍스트 생성 전략은 주어진 텍스트의 다음에 위치할 단어(토큰)를 선택하는 방식에 따라 Top-k 샘플링과 Top-p 샘플링이 존재한다(그림 6). Top-k 샘플링은 주어진 텍스트의 다음에 위치할 수 있는 모든 단어 중 가장 확률이 높은 K개의 후보 중 선택하는 방법이다. 그리고 Top-p 샘플링은 주어진 텍스트의 다음에 위치할 수 있는 모든 단어의 확률 분포를 계산한 후, 누적 확률 상위 p%에 해당하는 단어들을 대상으로 샘플링을 수행하는 방법이다. Top-p 샘플링의 p 값이 작을수록 생성모형은 더 높은 확률을 갖는 단어만 선택하여 일관된 텍스트가 생성되지만, p 값이 커질수록 생성모형은 더 많은 단어를 고려므로 다양한 텍스트가 생성될 수 있다. 따라서, 본 연구는 R&D Solution 생성모형이 Artificial Solution을 생성하는 과정에서 더 다양한 분야를 고려하고 새로운 R&D 해결방안을 도출할 수 있도록 Top-p 샘플링 기법을 적용하였다(p=0.95).

<그림6 텍스트 생성 전략 Top-k 샘플링 및 Top-p 샘플링 예시>



R&D Solution 생성모형의 Artificial Solution 생성 성능은 SSA 지표를 통해 평가되었다. 평가 대상 특허는 2010년 이후에 출원되어 CAF가 적용된 사이버 범죄 관련 특허이다. 본 연구는 사이버 범죄 관련 특허를 식별하기 위한 특허 검색식을 정의하였다. 우선 국가과학기술지식정보서비스의 사회문제해결플랫폼(https://www.ntis.go.kr/scisoplatform)에서 제공하는 사이버 범죄 관련 키워드를 수집하였다(표 5). 다음으로, 사이버 범죄와 관련된 기술 키워드를 추가적으로 수집한 후, 전문가 자문을 통해 사이버 범죄 관련 특허 검색식을 정의하였다(표 6). 검색식에서 '+'는 AND 연산, ' '(공백)은 OR 연산을 의미하며, 띄어쓰기가 포함된 키워드는 큰따옴표(" ")로 묶어서 표현되었다. 특허 검색식은 2010년 이후 출원된 국내 특허의 제목, 초록, 그리고 독립 청구항의 텍스트를 대상으로 수행되었다. 결과적으로 사이버 범죄와 관련된 212건의 국내 출원특허가 검색되었다.

<표5 사회문제 '사이버 범죄'의 핵심 키워드.

핵심 키워드 개수	핵심 키워드
15	사이버공간, 해킹, 침해사고, 피싱, 스미싱, 개인정보, 스팸메일, 불법콘텐츠, 사이버테러, 정보통신망, 사이버사기, 사이버명예훼손, 사이버폭력, 사이버음란물, 사이버도박

<표6 사회문제 '사이버 범죄'의 특허 검색식>

검색식
+((+사이버+(범죄 도박 테러 명예훼손 폭력 사기 침해 사고 괴롭힘 모욕 학대))(해킹 스파이웨어 랜섬웨어 웜 피싱 스미싱))+((스팸메일 불법콘텐츠 음란물 개인정보 프라이버시)(정보통신망 사이버공간 계산 이론 "계산 이론" 암호화 정보통신서비스 보안 소프트웨어공학 "소프트웨어 공학"))

앞서 Problem-Solution 데이터셋을 구축했던 과정과 동일하게, KoBART를 활용하여 사이버 범죄 관련 특허들의 Problem과 Existing Solution을 도출하였다. 그리고 R&D Solution 생성모형을 활용하여 사이버 범죄에 대한 Artificial Solution을 생성하였다. 도메인 및 특허기술 전문가 기반의 SSA 평가를 통해 R&D Solution 생성모형의 Artificial Solution 생성 성능을 평가하였으며, Sensibleness와 Specificity는 각각 0.67과 0.46으로, SSA는 0.57로 측정되었다. Sensibleness와 Specificity 모두 1이 부여된 Artificial Solution의 예시는 <표 7>에 나타내었다.

<표7 사이버 범죄에 대한 Artificial Solution 예시(Sensibleness=1, Specificity=1)>

출원번호	Problem	Artificial Solution
1020100005012	인터넷 금융 발달로 개인용 컴퓨터를 이용한 온라인 금융거래가 널리 보급되고 있는 가운데, 금융기관들은 보안문제를 해결하기 위한 방안으로 다양한 보안 프로그램을 컴퓨터에 설치하도록 고객에게 의무화하고 있다.	본 발명에 의한 처리 서비스 시스템은 보안 서비스에서 다수의 모듈 및 사용자 데이터 패킷을 송신하도록 구성된 데이터 송신 모듈을 포함한다.
1020140040192	스마트폰의 보급이 증가하고, 스마트폰의 기술이 발전함에 따라 스마트폰을 이용한 링크연결이 자주 발생함에 따라, 스마트폰의 링크연결을 이용한 범죄도 증가하고 있다.	본 발명에 따른 스마트 단말 어플은 통신 회로, 컴퓨터, 오디오 단말기, 및 기억 매체를 포함하고 상기 통신 회로의 동기화 장치들을 포함하는 것을 특징으로 한다.
1020140070938	생체 인식을 통한 사용자 인증은 사용이 편리할 뿐만 아니라 보안성 및 경제성이 뛰어나 현재 많이 상용화 되어 있지만, 최근에는 기술의 발전에 따라 휴대용 장치(mobile device)까지 쓰임이 확대되고 있다.	본 출원의 사용자 인증은 사용자의 휴대용 장치 및 전자 장치를 제안하고 있는데, 이 장치는 사용자의 전자 장치 본체, 및 사용자에 의해 형성된 적어도 하나의 신호를 수신하도록 구성된, 생체 인식 장치, 특히 사용자에 의해 형성된 자기장을 감지하도록 구성된, 생체 인식 장치 를 포함한다.
1020200157390	최근 각종 범죄나 긴급 상황의 발생 수가 급증함에 따라 각종 범죄나 긴급 상황을 대비한 다양한 보안장비들이 출시되고 있다.	본 발명의 다른 양태에 따르면, 본 발명은 상기 보안장치로 이루어지는 사용자의 개인 정보를 수신하고, 상기 보안도 정보의 발생을 위한 자동 제어기를 개시하며,

		상기 사용자 요청은 상기 사용자 단말이 사용자의 입력에 따라, 상기 사용자 등록 이 인증 절차를 실시하는 단계를 더 포함한다.
1020207002143	랜섬웨어는 컴퓨터 하드 드라이브와 같은 저장 매체에서 발견된 파일 세트를 암호화하고, 그 후 파일 소유자에게 각 데이터를 복구하기 위해 비용을 지불하도록 요구하는 위험한 유형의 멀웨어이다.	본 개시내용의 실시예는 복수의 파일 데이터로 구성된 디지털 저장 디바이스 및 사용자 장치와 통신하기 위한 통신 방법을 제공함으로써 현재 파일 저장 매체에 적합한 파일 데이터를 부여하도록 구성된다.

4.3. R&D 솔루션 분석 및 기술 컨셉 포트폴리오 맵 구축

앞서 생성된 사이버 범죄에 대한 Artificial Solution을 효과적으로 살펴보기 위해, Artificial Solution을 구성하는 명사구 키워드를 추출 및 그룹화하여 기술 컨셉을 도출하였다. 우선 명사구를 추출하기 위해 본 연구는 OpenAI의 GPT API 기반 프롬프트 엔지니어링을 활용하였으며, 언어모델에게 명확한 업무를 지시하고 풍부한 예시를 제공하는 등 선행연구를 통해 검증된 프롬프트 엔지니어링 전략을 적용하였다(표 8).[41]

<표8 Artificial Solution의 명사구 키워드 추출을 위한 프롬프트>

프롬프트
당신은 똑똑하고 지능적인 키워드 추출 시스템입니다. 입력되는 텍스트로부터 당신이 추출해야 하는 키워드는 과학기술 관련 명사 또는 명사구입니다. '발명', '상기', '실시'와 같이 보편적으로 등장하는 키워드는 지양합니다. 지켜야하는 출력 형식과 키워드 추출 예시를 함께 제공하겠습니다. 출력 형식: [키워드1;키워드2;키워드3;...;키워드N] 추출된 키워드가 없으면 'None'을 출력합니다. 예시: <문서 1> 상기의 목적을 달성하기 위해 본 발명은 렌즈가 장착되는 안구부와, 렌즈가 장착되는 안구부의 배면 측에 연결되는 안경다리부로 구성되는 에어커튼 기능을 가진 안경을 제공한다. <출력 1> [렌즈;안구부;배면;안경다리부;에어커튼 기능;안경] <문서 2> 문자 메시지를 이용한 불법 금융거래를 차단하기 위한 통신 단말은 경고 알림창을 표시하고, 문자 어플리케이션에 수신된 문자 메시지에 포함된 계좌번호를 비식별 처리할 수 있다. <출력 2> [문자 메시지;불법;금융거래;통신 단말;경고 알림창;문자 어플리케이션;계좌번호] 여기까지가 키워드 추출 예시이며, 입력된 문서에서 중요한 키워드를 최대한 많이 추출하기 바랍니다.

프롬프트 엔지니어링을 통해 728개의 명사 및 명사구 키워드가 추출되었으며, '개시', '사이', '양태' 등 기술적 내용을 나타내지 않는 키워드를 제거하여 최종적으로 650개의 키워드를 선별하였다. 본 연구는 Sentence-BERT(SBERT) 언어모델 'ko-sroberta-multitask'를 사용하여 선별된 명사구를 모두 임베딩 벡터로 변환하였다. SBERT는 벡터 공간에서 의미가 유사한 텍스트가 가깝게 배치되도록 조정하는 문장 임베딩 기법으로,[42] 단일 키워드뿐만 아니라 명사

41) Sondos Mahmoud Bsharat et al., "Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4", arXiv preprint arXiv:2312.16171. 2023.

42) Nils Reimers & Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", arXiv preprint arXiv:1908.10084, 2019.

구를 다루는 본 연구에 사용하기 적합하다. 다음으로, 추출된 명사 및 명사구의 벡터 유사도를 기반으로 그룹화하여 기술 컨셉을 정의하기 위해 K-means 클러스터링 기법을 사용하였다. 최적의 클러스터 개수는 클러스터 내 벡터 간의 코사인 유사도 평균을 측정하고, 유사도가 가장 낮을 때의 클러스터 수를 식별하여 14개로 설정하였다. 각 클러스터의 중심에서 가까운 명사구들에 기반하여 레이블링을 하고 기술 컨셉을 정의하였다(표 9).

<표9 정의된 기술 컨셉 및 주요 키워드>

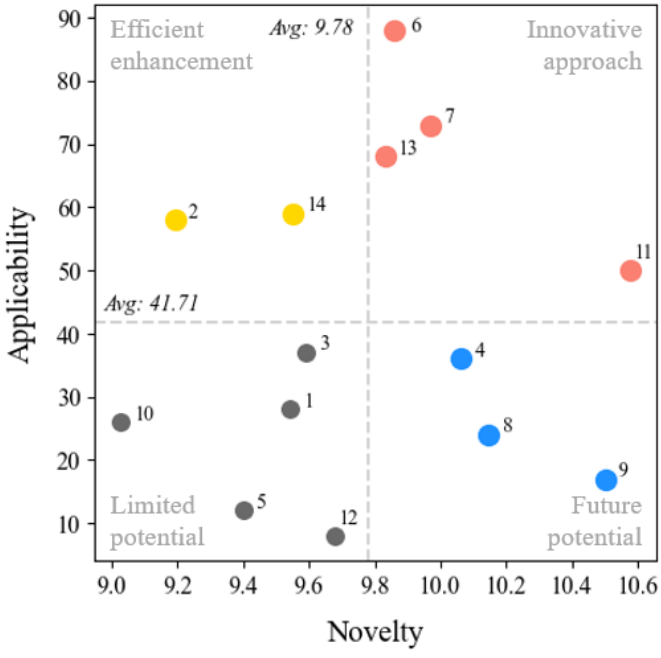
클러스터번호	기술 컨셉	주요 키워드
1	네트워크 통신망	네트워크 접속 장치, 통신망, 통신 네트워크, 무선 네트워크, 네트워크, ...
2	통신 기기 및 송수신 시스템	송수신 장치, 통신 장치, 송신 단말기, 통신 수단, 통신 디바이스, ...
3	데이터 처리 시스템	데이터, 데이터 처리, 데이터 링크, 데이터 스트림, 데이터 처리 장치, ...
4	서비스 정보 관리	서비스 정보, 서비스 장치, 서비스 요청 정보, 서비스 단말, 입력 정보 서비스, ...
5	카드 결제 시스템	카드 인증, 카드, 카드 번호, 카드 자동 식별 방법, 결제 장치, ...
6	제어 시스템 및 장치	제어 모듈, 조작 모듈, 제어 장치, 제어 회로, 제어 정보, ...
7	정보 처리 시스템	컴퓨터 디바이스, 컴퓨팅 인덱싱 장치, 컴퓨터 장치, 컴퓨팅 장치, 정보 처리 장치, ...
8	영상 및 이미지 처리	영상 처리, 이미지 데이터, 영상 데이터, 이미지 데이터 처리, 이미지촬영, ...
9	자동화 시스템	자동 등록, 자동 판독 모듈, 자동제어, 자동 식별 방법, 자동 변환...
10	보안 시스템 및 장치	보안 시스템, 보안 관련 정보, 보안, 보안장치, 보안 서비스, ...
11	기기 구성 요소	덮개, 본체, 획득부, 출력부, 연결부, ...
12	네트워크 노드 제어 시스템	입력 노드, 노드, 초기 노드, 입력 노드 분할, 네트워크 노드, ...
13	전력 관리 시스템	시그널링, 신호, 인가 출력, 셀, 전력, ...
14	개인정보 인증 시스템	식별 정보, 식별 코드, 인증 수단, 사용자 인증, 개인 정보 인증 장치, ...

마지막으로, 본 연구는 앞서 정의된 기술 컨셉의 Applicability와 Novelty를 산출하여 기술 컨셉 포트폴리오 맵을 구축하였다. Applicability는 기술 컨셉의 키워드가 등장한 Artificial Solution의 개수를 측정하여 산출하였다. Novelty의 경우, 앞서 활용한 'ko-sroberta-multitask' 모델을 통해 Existing Solution과 Artificial Solution의 임베딩 벡터를 도출한 후, 벡터 공간에서의 유클리디안 거리를 측정하여 산출하였다. 각 기술 컨셉의 Applicability와 Novelty는 <표 10>과 같으며, 이 결과를 바탕으로 기술 컨셉 포트폴리오 맵을 도출하였다(그림 7).

<표10 기술 컨셉의 Applicability와 Novelty>

클러스터 번호	기술 컨셉	Applicability	Novelty
1	네트워크 통신망	28	9.54
2	통신 기기 및 송수신 시스템	58	9.19
3	데이터 처리 시스템	37	9.59
4	서비스 정보 관리	36	10.06
5	카드 결제 시스템	12	9.40
6	제어 시스템 및 장치	88	9.86
7	정보 처리 시스템	73	9.97
8	영상 및 이미지 처리	24	10.15
9	자동화 시스템	17	10.50
10	보안 시스템 및 장치	26	9.03
11	기기 구성 요소	50	10.58
12	네트워크 노드 제어 시스템	8	9.68
13	전력 관리 시스템	68	9.83
14	개인정보 인증 시스템	59	9.55

<그림7 도출된 기술 컨셉 포트폴리오 맵>



도출된 포트폴리오 맵의 Innovative approach 영역에는 4개의 기술 컨셉이 포함되었다. 해당 컨셉들은 Applicability, Novelty가 모두 높은 값을 나타내므로, 사이버 범죄를 해결하기 위해 곧바로 적용할 수 있으면서도 이미 존재하는 R&D 솔루션 대비 참신한 기술 컨셉이다. 그중에서도 Applicability가 가장 높은 '제어 시스템 및 장치' 기술 컨셉은 제어 모듈을 활용한 네트워크 트래픽 실시간 모니터링, 제어 정보 기반의 지능형 방화벽 및 접근 제어 시스템 구축, 데이터 무결성 및 기밀성 보장 등 사이버 범죄의 해결과 예방을 위한 다양한 R&D 인사이트를 제공할 수 있다. 해당 기술 컨셉과 관련된 실제 Artificial Solution의 예시는 <표 11>과 같다. 사이

버 공간에서 디지털 콘텐츠의 저작권을 보호하기 위해 기존에는 보안영역에 저장된 코드 및 운영체제를 보호하는 해결방안이 제시되었다. 하지만, 새롭게 생성된 솔루션은 실시간 사용자 입력을 기반으로 한 보안 메커니즘을 통해 디지털 콘텐츠의 불법 복제와 무단 접근을 효과적으로 방지할 수 있다.

<표11 '제어 시스템 및 장치'의 Artificial Solution 예시(출원번호 1020180066614)>

항목	텍스트
Problem	디지털 컨텐츠는 저장 및 복제가 쉬운 특성을 가지고 있기 때문에 저작권을 보호하기 위한 방지책이 마련되어야 한다.
Existing Solution	보안영역에 저장된 코드에 대한 외부 접근을 제한함으로써, OS의 동작에 대한 외부 접근을 제한함으로써 보안이 유지된다.
Artificial Solution	스마트 디스플레이 제어 시스템은 터치 디스플레이의 사용자 터치 입력을 감지하여 적어도 하나의 타이밍 데이터는 디스플레이의 입력 포트를 결정하고, 타이밍 데이터의 적어도 한 입력 포트는 표시 모드를 결정한다.

Innovative approach 영역에서 Novelty가 가장 높은 기술 컨셉은 '기기 구성 요소'이며, 해당 컨셉은 기기 및 장치를 활용한 동적 시스템 기반 실시간 모니터링 및 대응과 같은 인사이트를 제공할 수 있다. <표12>에 나타낸 '기기 구성 요소' 컨셉의 Artificial Solution을 살펴본 결과, 새로운 솔루션은 기존의 정적이고 단계적인 보안 절차와 달리 이동체가 장착된 소형 차량용 무선 장치를 배치하여 활용하는 참신한 아이디어를 제공하였다. 이는 보안 대상물에 대한 전방위적인 감시와 동적 모니터링을 통해 보안 위협에 효과적으로 대응할 수 있는 능력을 강화할 수 있다.

<표12 '기기 구성 요소'의 Artificial Solution 예시(출원번호 1020160018484)>

항목	텍스트
Problem	항만시설에 야적된 컨테이너 화물을 포함한 각종 감시 대상물에 대해 의도적 침입, 훼손, 도난 등에 대해 전방위적으로 감시할 수 있는 보안체계의 필요성이 증대되고 있다.
Existing Solution	정보보호관리체계를 활용한 항만물류정보보안방법은 보안침해사고를 탐지 또는 접수하는 침해사고 탐지/접수단계(A); 상기 ... (생략) ... 보고하는 조치 완료결과 보고단계(G); 로 구성되는 것을 특징으로 한다.
Artificial Solution	본 발명에 따른 무선 장치는, 이동체가 장착되어 있는 소형 차량용 무선 장치로서, 이동체의 길이 방향 축이 되는 가이드 플레이트를 구비하여 이동하는 것을 특징으로 한다.

Efficient enhancement 영역과 Future potential 영역에는 각각 2개(통신 기기 및 송수신 시스템, 개인정보 인증 시스템), 3개(서비스 정보 관리, 영상 및 이미지 처리, 자동화 시스템)의 기술 컨셉이 포함되었다. 먼저 Efficient enhancement 영역의 기술 컨셉은 Applicability가 높아 즉시 적용해볼 수 있는 R&D 솔루션 정보를 제공하므로 주어진 문제에 대한 효율적 개선 및 해결을 가능하게 한다. 예를 들어 기술 컨셉 '개인정보 인증 시스템'의 경우, 2010년대 초반 스마트폰 상용화 이후 모바일 환경에서의 사이버 범죄를 예방하기 위한 사용자 인증 및 정보 보안 관련 특허가 지속적으로 출원되어왔다 (예. 출원번호 1020100011377: '모바일 단말기 정보를 이용한 사용자 정보 보안 방법', 출원번호 1020210118981: '생체 인식에 따른 랜섬웨어 동작 방지방법과 방지시스템'). 이처럼 Efficient enhancement 영역의 기술 컨셉과 관련된 솔루션은 오래전부터 최근까지 꾸준하게 제시되고 있으므로, 주어진 문제를 신속히 해결할 수 있는

R&D 인사이트를 해당 영역의 기술 컨셉으로부터 도출할 수 있다.

Future potential 영역은 당장 적용하기에는 어렵더라도 꾸준한 연구개발을 통해 미래에 큰 잠재력을 가질 수 있는 R&D 솔루션에 대한 아이디어를 제공할 수 있다. 해당 영역의 기술 컨셉 중에서도 Novelty가 높은 '자동화 시스템', '영상 및 이미지 처리'는 빠른 속도로 발전하는 머신러닝 및 인공지능 분야와 깊은 연관성이 있는 것으로 보여진다. 우선 '자동화 시스템'은 머신러닝과 인공지능 기술의 통합으로, 사이버 공격에 대해 보다 정교한 탐지 및 대응 능력을 갖출 수 있게 되었다. 이전에는 규칙 기반의 탐지 방법이 주를 이루었다면, 이제는 인공지능을 통해 비정상적인 패턴을 학습하고 새로운 위협을 자동으로 탐지할 수 있다. 또한, 최근 딥러닝 기술의 발전으로 '영상 및 이미지 처리'의 정확성과 효율성이 크게 향상되었다. 얼굴 인식, 객체 탐지, 행동 인식 등 다양한 분야에서 성능이 비약적으로 향상되고 있으므로, 실시간 영상 분석, 다중 인증을 위한 생체 인식, 이미지 포렌식 기술 기반의 조작 여부 판별 등 적용 범위가 넓은 R&D 솔루션들이 개발되고 있다. 따라서, Future potential 영역의 기술 컨셉을 통해, 즉시 문제에 적용하기보다는 추가적인 연구개발을 통해 미래에 더욱 효과적으로 문제를 해결할 수 있는 R&D 솔루션에 대한 아이디어를 도출할 수 있다.

5. 결론 및 추후 연구

본 연구는 대규모 언어모델 기반의 특허분석을 통해 주어진 문제에 적용할 수 있는 새로운 R&D 솔루션을 생성 및 분석하는 방법을 제시하였다. 우선, 대량의 국내 출원 특허 데이터를 수집하였으며, 특허의 CAF 텍스트로부터 Problem-Solution 정보를 추출하였다. 다음으로, Problem-Solution 정보를 활용한 대규모 언어모델의 파인튜닝을 통해 R&D Solution 생성모형을 구축하였으며, R&D Solution 생성모형의 텍스트 생성 성능은 SSA 지표를 통해 평가되었다. 분석 사례로는 국내 사회문제 중 사이버 범죄가 선정되었으며, 본 연구는 212건의 사이버 범죄 관련 특허와 R&D Solution 생성모형을 활용하여 사이버 범죄에 대한 새로운 Artificial Solution을 생성하였다. 생성된 Artificial Solution을 효과적으로 살펴보기 위해 키워드 추출 및 클러스터링을 활용하여 Artificial Solution을 구성하는 기술 컨셉을 도출하였으며, 각 기술 컨셉을 평가하기 위해 두 가지 지표 Applicability, Novelty를 산출하였다. 최종적으로, Applicability와 Novelty를 축으로 하는 기술 컨셉 포트폴리오 맵을 형성하였으며, 사이버 범죄에 적용가능하면서도 참신한 기술 컨셉 '제어 시스템 및 장치', '정보 처리 시스템', '기기 구성 요소', '전력 관리 시스템'을 식별하였다. 또한, 식별된 기술 컨셉의 주요 Artificial Solution을 살펴봄으로써 사이버 범죄 해결을 위한 새로운 R&D 해결 접근법을 확인할 수 있었다.

본 연구는 다음의 세 가지 기여점을 갖는다. 첫째, 본 연구는 최근 주목받고 있는 대규모 언어모델을 활용하여 R&D 솔루션을 창출하는 새로운 방법을 제시하였다. 국내 출원 특허의 Problem-Solution 정보를 활용한 대규모 언어모델의 파인튜닝을 통해 새로운 R&D 솔루션을 생성하고 기술 컨셉을 도출하였다. 이와 같은 방법은 기존의 R&D 지식과 더불어 창의적이고 혁신적인 문제 해결 방안을 제시할 수 있다. 둘째, 본 연구의 기술 컨셉 포트폴리오 맵은 문제 해결 프로세스에서 선택의 폭을 확대할 수 있다. 본 연구는 주어진 문제에 대한 새로운 R&D 솔루션을 생성하고, 생성된 솔루션을 구성하는 기술 컨셉의 적용가능성과 참신성을 정량화한 후 이를 포트폴리오 맵으로 시각화하여 제시하였다. 이를 통해 문제 해결 과정에서 더욱 다양한 접근 방향을 고려하고 최적의 방안을 도출할 수 있도록 지원할 수 있다. 셋째, 본 연구는 대규모 언어모델 기반의 R&D 기술 혁신을 촉진하는 데 기여한다. R&D 솔루션 생성모형과 같은 자체적인

대규모 언어모델을 개발한다면, 기존의 솔루션 탐색 방법보다 더욱 효율적이고 신속하게 솔루션 개발에 대한 인사이트를 도출할 수 있다. 따라서, 본 연구는 우리 사회의 다양한 분야에서 R&D 기술의 발전과 혁신을 촉진할 수 있다.

Problem-Solution 정보 기반의 특허분석 및 대규모 언어모델을 활용하는 연구분야는 지속적으로 많은 연구가 수행되는 분야이므로, 본 연구는 추후 개선될 수 있는 부분이 존재한다. 우선, 본 연구는 데이터 규모, 컴퓨팅 자원, 경제적 비용 등 다양한 측면을 고려하여 오픈소스 한국어 언어모델인 KoGPT2를 활용하여 R&D Solution 생성모형을 구축하였다. 더욱 풍부한 자원을 활용할 수 있다면, R&D Solution 생성모형 구축 과정에서 최근에 개발된 오픈소스 대규모 언어모델 혹은 언어모델 API의 파인튜닝 기능을 활용할 수 있을 것이다. 다음으로, 본 연구는 사회문제에 대한 새로운 R&D 솔루션을 생성하여 제시하였지만, 실제로 해당 솔루션을 적용하여 사회문제가 개선된 정도를 확인할 수 없었다. 본 연구에서 제시한 방법을 통해 다양한 사회문제의 해결을 위한 R&D 솔루션을 생성하고 적용하는 실질적인 시도가 요구된다. 마지막으로, 본 연구는 특허의 CAF 항목을 활용하여 Problem-Solution을 텍스트 형태로 도출하였다. 특허는 제목, 초록, CAF 항목과 같은 텍스트뿐만 아니라 출원인 및 발명자와 같은 서지정보, 도면 이미지 등 다양한 기술 관련 데이터를 제공한다. 추후에 이러한 정보들을 함께 활용하는 멀티모달(Multimodal) 기반의 연구가 수행된다면, 기업 맞춤형 문제 해결 전략을 제시하거나 R&D 솔루션 기반 제품 시뮬레이션 등의 폭넓고 다양한 R&D 결과물을 도출할 수 있을 것으로 기대된다.

참고문헌

단행본(서양)

Semyon D. Savransky, *Engineering of creativity: Introduction to TRIZ methodology of inventive problem solving*, CRC press, 2000.

학술지(국내 및 동양)

윤장혁.김광수, "SAO 기반의 의미론적 특허 유사성을 활용한 특허맵 생성방법", 「Entrue Journal of Information Technology」, 제10권 제1호(2011).

이지호 외 4인, "특허의 Problem-Solution 텍스트 마이닝을 활용한 기술경쟁정보 분석 방법", 「지식재산연구」, 제13권 제3호(2018).

이치훈 외 2인, "사전 학습된 한국어 BERT 의 전이학습을 통한 한국어 기계독해 성능개선에 관한 연구", 「한국IT 서비스학회지」, 제19권 제5호(2020).

정재민 외 2인, "비즈니스 기회 발굴을 위한 문제-해결방법 기반의 특허분석 방법", 「지식재산연구」, 제15권 제2호(2020).

학술지(서양)

Dave Van Veen et al., "Adapted large language models can outperform medical experts in clinical text summarization", *Nature Medicine*, Vol.30 No.4(2024).

Georg Richter & Andrew MacFarlane , "The impact of metadata on the accuracy of automated patent classification", *World Patent Information*, Vol.27 No.1(2005).

Hyunseok Park et al., "Identifying patent infringement using SAO based semantic technological similarities", *Scientometrics*, Vol.90 No.2(2012).

Jaewoong Choi et al., "Technology opportunity discovery under the dynamic change of focus technology fields: Application of sequential pattern mining to patent classifications", *Technological Forecasting and Social Change*, Vol.148(2019).

Janghyeok Yoon & Kwangsoo Kim, "Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks", *Scientometrics*, Vol.88 No.1(2011).

Janghyeok Yoon & Kwangsoo Kim, "Detecting signals of new technological opportunities using semantic patent analysis and outlier detection", *Scientometrics*, Vol.90 No.2(2012).

Jiho Lee et al., "An approach for discovering firm-specific technology opportunities: Application of link prediction to F-term networks", *Technological Forecasting and Social Change*, Vol.168 (2021).

Johan Schot & W. Edward Steinmueller, "Three frames for innovation policy: R&D, systems of innovation and transformative change", *Research policy*, Vol.47 No.9(2018).

Jonathan H, Choi et al., "ChatGPT goes to law school", *Journal of Legal Education*, Vol.71 No.3(2021).

Kyuwoong Kim et al., "Investigating technology opportunities: The use of SAOx analysis", *Scientometrics*, Vol.118(2019).

Martin G. Moehrle et al., "Patent-based inventor profiles as a basis for human resource decisions in research and development", *R&D Management*, Vol.35 No.5(2005).

Saika Wong et al., "Construction contract risk identification based on knowledge-augmented language models", *Computers in Industry*, Vol.157(2024).

Sungchul Choi et al., "An SAO-based text-mining approach for technology roadmapping using patent information", *R&D Management*, Vol.43 No.1(2013).

Sunhye Kim & Byungun Yoon, "Patent infringement analysis using a text mining technique based

on SAO structure", *Computers in Industry*, Vol.125(2021).

Xuefeng Wang et al., "Identifying R&D partners for dye-sensitized solar cells: a multi-level patent portfolio-based approach", *Technology Analysis & Strategic Management*, Vol.31 No.3(2019).

Xuefeng Wang et al., "Measuring patent similarity with SAO semantic analysis", *Scientometrics*, Vol.121(2019).

Yi Zhang et al., "How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence:"problem & solution" pattern based semantic TRIZ tool and case study", *Scientometrics*, Vol.101(2014).

Youngjin Seol et al., "Towards firm-specific technology opportunities: A rule-based machine learning approach to technology portfolio analysis", *Journal of Informetrics*, Vol.17 No.4(2023).

Yu Gu et al., "Domain-specific language model pretraining for biomedical natural language processing", *ACM Transactions on Computing for Healthcare*, Vol.3 No.1(2021).

Yuen-Hsien Tseng et al., "Text mining techniques for patent analysis", *Information processing & management*, Vol.43 No.5(2007).

Zhe Zheng et al., "Pretrained domain-specific language model for natural language processing tasks in the AEC domain", *Computers in Industry*, Vol.142(2022).

인터넷 자료

Alec Radford et al., "Language models are unsupervised multitask learners", OpenAI blog, <https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf>, 검색일: 2024. 06. 30.

기타 자료

과학기술정보통신부, "제3차 과학기술 기반 사회문제해결 종합계획", 과학기술정보통신부, 2023.

Daniel Adiwardana et al., "Towards a human-like open-domain chatbot", arXiv preprint arXiv:2001.09977, 2020.

Edward Loper & Steven Bird, "Nltk: The natural language toolkit", arXiv preprint cs/0205028, 2002.

Eunjeong L. Park & Sungzoon Cho, "KoNLPy: Korean natural language processing in Python", In Proceedings of 26th Annual Conference on Human and Cognitive Language Technology, Special Interest Group of Human and Cognitive Language Technology, 2014.

Jules White et al., "A prompt pattern catalog to enhance prompt engineering with chatgpt", arXiv preprint arXiv:2302.11382, 2023.

Mike Lewis et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension", arXiv preprint arXiv:1910.13461, 2019.

Nils Reimers & Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", arXiv preprint arXiv:1908.10084, 2019.

Sondos Mahmoud Bsharat et al., "Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4", arXiv preprint arXiv:2312.16171. 2023.

Wayne Xin Zhao et al., "A survey of large language models", arXiv preprint arXiv:2303.18223, 2023.

Youngho Kim et al., "Automatic discovery of technology trends from patent text", In Proceedings of the 2009 ACM symposium on Applied Computing, Association for Computing Machinery, 2009.