Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 2, pp. 176-192, April 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.2.176

대규모 언어 모델을 활용한 비주석 기술 텍스트 기반 도메인 특화 지식 그래프 구축

이수연 · 박소형 · 김현종 · 조성준†

서울대학교 산업공학과

Constructing Domain-Specific Knowledge Graphs from Unannotated Technical Texts using LLMs

Suyeon Lee · Sohhyeong Park · Hyunjong Kim · Sungzoon Cho

Department of Industrial Engineering, Seoul National University

Building knowledge graphs in specific domains presents significant challenges when domain expertise is limited. The primary obstacles include the lack of annotated datasets and domain-specific models. Although large language models (LLMs) enable flexible extraction from unstructured text, their direct application to technical domains often leads to domain mismatch and inconsistent relation representations. This paper presents a hybrid framework that combines a domain-specific extraction model with LLM-based reasoning to build knowledge graphs from unannotated patent abstracts. Using semiconductor patents as a case study, domain-relevant entities are first identified and then refined through iterative LLM prompting to extract relational triplets. The resulting relations are normalized and integrated into a unified knowledge graph. Experimental results indicate improved textual faithfulness and more coherent relation structures compared to domain-only and LLM-only baselines, demonstrating a practical approach for scalable knowledge graph construction in unannotated technical domains.

Keywords: Knowledge Graph Construction, Domain-Specific Knowledge Graph, Large Language Models

1. 서 론

지식 그래프(Knowledge Graph)는 현실 세계의 지식을 개체(entity)와 그 관계(relation)로 구조화하여 표현함으로써, 다양한 출처에 분산된 정보를 통합하고 활용할 수 있게 하는 효과적인 지식 표현 방식이다(Zou, 2020; Ji *et al.*, 2021; Hogan *et al.*, 2021). 특히 자연어처리 기술의 발전으로 비정형 텍스트에 내재된 정보를 자동으로 구조화할 수 있게 되면서, 학술 문헌이나 기술 보고서와 같은 비정형 문서로부터 지식 그래프를 구축하려는 연구가 증가하고 있다(Ji *et al.*, 2021) 여러 문서와 출처에 흩어진 지식을 통합된 하나의 지식 그래프로 통합하면, 개별 문서 단위에서는 드러나지 않는 개념 간의 구조적 연결성과 잠재적 관계를 체계적으로 파악할 수 있어 지식의 구조화와 확장 측면에서 다양한 이점을 제공한다(Ji *et al.*, 2021; Zou, 2020).

이러한 지식 그래프는 질의응답, 추천 시스템, 정보 검색, 대화형 인터페이스 등 다양한 응용 분야에서 활용되어 왔으며(Zou, 2020; Schneider *et al.*, 2022; Zhong *et al.*, 2023), 최근 검색 증강 생성(RAG)과 결합되어 지식 그래프를 외부 지식 기반으로 활용하려는 시도도 증가하고 있다(Peng *et al.*, 2023; Peng *et al.*, 2024). 기존의 지식 그래프 구축 방법은 주로 개체명 인식(NER)과 관계 추출(RE)로 구성된 정보 추출 파이프라인에

이 논문은 삼성종합기술원의 지원을 받아 수행되었음.
† 연락저자 : 조성준 교수, 08826 서울시 관악구 관악로 1 서울대학교 공과대학 산업공학과, Tel : 02-880-7025, Fax : 02-889-8560, E-mail : zoon@snu.ac.kr
2025년 12월 5일 접수; 2026년 1월 19일 수정본 접수; 2026년 2월 9일 게재 확정.

기반하는데, 이는 구조적 일관성과 해석 가능성 측면에서 강점을 가지는 반면, 대규모 레이블링 데이터나 사전 정의된 스키마나 온톨로지에 대한 의존도가 높다는 한계를 지닌다(Hogan *et al.*, 2021; Ji *et al.*, 2021; Zhong *et al.*, 2023). 이러한 접근은 WebNLG(Gardent *et al.*, 2017)나 TACRED(Zhang *et al.*, 2017)와 같은 일반 도메인 데이터셋뿐만 아니라, 과학 도메인의 SciERC(Luan *et al.*, 2018), 금융 도메인의 REFinD(Kaur *et al.*, 2023), 재료 과학 도메인의 MatKG(Venugopal *et al.*, 2024) 등 다양한 특수 도메인 데이터셋을 기반으로도 활발히 연구되어 왔다(Schneider *et al.*, 2022). 그러나 도메인마다 엔티티와 관계의 정의 및 구조가 크게 상이하기 때문에, 한 도메인에서 학습된 모델을 다른 도메인에 그대로 적용하기는 어렵다(Abu-Salih, 2021). 이에 따라 효과적인 지식 그래프 구축을 위해서는 도메인 특화 접근이 요구되지만, 해당 도메인에 대해 레이블링 된 데이터나 정형화된 지식 자원이 존재하지 않는 경우 기존 방법론을 적용하기에는 한계가 존재한다.

이러한 문제는 특히 기술 발전 속도가 빠르고 개념 체계가 지속적으로 변화하는 산업 분야에서 더욱 두드러진다. 반도체 분야와 같이 공정, 재료, 장비, 공정 조건 등 다양한 기술 요소들이 상호 의존적으로 결합된 복합적인 기술 체계를 지닌 산업 분야에서는, 엔티티와 관계의 범주를 사전에 정의하거나 대규모 레이블링 데이터를 구축하는데 현실적인 제약이 따른다. 이로 인해 특허와 논문에 산재된 반도체 기술 지식을 단순한 키워드 검색이나 빈도 기반의 통계적 분석만으로는 기술 요소 간의 구조적 패턴이나 의미적 관계를 체계적으로 파악하거나 활용하기 어렵다(Kim *et al.*, 2025). 실제로 기존 반도체 특허 분석 연구들은 기술적 문제와 해결 간의 관계, 또는 공정, 재료, 장비 간의 연관 구조를 기반으로 네트워크 분석을 수행하며, 반도체 기술 요소들이 상호 연결성과 의미적 연관성을 가진다는 점을 강조해 왔다(Chen *et al.*, 2024; Kim *et al.*, 2025). 이러한 맥락에서 반도체 기술 분석, 공정 최적화, 경쟁 기술 탐색, R&D 전략 수립과 같은 다양한 응용 과제를 효과적으로 수행하기 위해서는 여러 기술 단계와 구성 요소 간의 관계를 명시적으로 표현할 수 있는 지식 그래프가 중요한 역할을 수행할 수 있다. 그러나 기존 특허 기반 지식 그래프 구축 방법들은 주로 규칙 기반 마이닝이나 전문가 레이블링에 의존해 왔으며(Chen *et al.*, 2020; Sarica *et al.*, 2020; Siddharth *et al.*, 2022), 이는 확장성과 도메인 적응성 측면에서 한계를 가진다.

이러한 배경에서 최근에는 대규모 언어 모델(LLM)을 활용하여 레이블 없이도 텍스트로부터 지식 그래프를 구축하려는 연구가 활발히 이루어지고 있다(Yang *et al.*, 2024; Zhu *et al.*, 2024). LLM 기반 접근은 사전 정의된 관계 집합이나 고정된 스키마에 의존하지 않고 다양한 관계 표현을 추출할 수 있다는 점에서, 기존 정보 추출 기반 지식 그래프 구축 방법 대비 유연성을 제공한다(Pai *et al.*, 2024; Bian *et al.*, 2025). 또한 프롬프트 기반 추론을 통해 모델을 재학습하지 않고도 새로운 개념이나 관계를 구조화된 지식으로 비교적 유연하게 확장할 수 있다는 장점을 가진다(Khorashadizadeh *et al.*, 2023; Dagdelen *et al.*, 2024). 그러나 다수의 기존 연구는 Wikipedia 기반의 일반 도메인 데이터에 집중되어 있으며(Zhang *et al.*, 2024), 특수 도메인에 적용되는 경우에도 레이블링된 데이터나(Zhang *et al.*, 2024; Lairgi *et al.*, 2024; Orlando *et al.*, 2024) 온톨로지(Ding *et al.*, 2024)와 같은 추가적인 지식 자원에 의존하는 경우가 많다. 또한 LLM은 그럴듯한 응답을 생성할 수 있으나, 사실적 근거가 부족한 환각(hallucination) 문제로(Andriopoulos *et al.*, 2023; Lavrinovics *et al.*, 2025) 인해 전문 용어의 누락, 비기술적 일반 명사의 엔티티 오인, 관계 표현의 비일관성과 같은 문제로 이어질 수 있다. 이는 도메인 특화 지식 그래프 구축에서 신뢰성과 재현성을 저해하는 요인으로 작용한다.

이러한 문제를 해결하기 위해, 레이블이 제공되지 않은 도메인 텍스트를 대상으로 도메인 특화 모델과 LLM을 결합한 하이브리드 지식 그래프 자동 구축 방법론을 제안한다. 제안하는 방법은 LLM을 단독으로 활용해 원시 텍스트에서 삼중항(triplet)을 직접 생성하는 기존 접근과 달리, 먼저 도메인 특화 모델을 통해 엔티티 후보를 식별하고 이를 LLM 추론의 가이드로 활용함으로써 환각 현상을 완화하고 도메인 중심의 엔티티 및 관계 추출을 가능하게 한다. 또한 사전 정의된 온톨로지 없이, LLM 기반의 반복적 정제 과정을 통해 다양한 관계 표현을 점진적으로 통합함으로써, 문서 전반에 적용 가능한 일관된 관계 체계를 자동으로 유도한다.

본 연구의 핵심 기여는 LLM을 개별 문장 단위의 삼중항 생성 도구로 사용하는데 그치지 않고, 관계 표현들의 집합을 반복적으로 정제하고 통합하는 메커니즘으로 활용함으로써 관계 공간 자체를 점진적으로 구조화한다는 점에 있다. 이 과정에서 세부 주제 내에서는 의미적으로 유사한 관계 표현을 통합하고, 세부 주제 간에는 상이한 명명 체계를 정렬함으로써, 레이블이나 온톨로지 없이도 응집력 있는 관계 사전을 자동 구축한다. 이는 기존 LLM 기반 지식 그래프 구축 연구들이 주로 문장 수준의 추출 정확도에 초점을 맞춘 것과 대비되는 차별적 접근이다.

제안한 방법론을 반도체 특허 문서에 적용하여, 공정, 재료, 장비, 기능 간의 복잡한 기술 관계를 구조화된 지식 그래프로 자동 구축한다. 이는 기존 특허 마이닝 기반 그래프나 수작업 레이블링 기반 지식 그래프에 비해 확장성, 도메인 적응성, 관계 표현의 일관성 측면에서 우수한 대안을 제공한다. 실험 결과는 제안한 하이브리드 프레임워크가 LLM 단독 방식이나 기존 도메인 모델 대비 더 높은 엔티티 정확도와 더 응집력 있는 관계 체계를 생성함을 보여준다.

2. 관련 연구

2.1 도메인 특화 지식 그래프 구축

자연어 처리 기술의 발전과 함께 비정형 텍스트로부터 구조화된 지식을 추출하여 지식 그래프를 구축하려는 연구가 활발

히 진행되고 있다(Wang *et al.*, 2017; Zou, 2020). 이러한 비구조화된 텍스트 기반 지식 그래프 구축은 일반적으로 개체명 인식(NER), 관계 추출(RE) 등의 정보 추출 기법을 결합하여 지식을 식별하고, 링크 예측(Link prediction)이나 엔티티 예측(Entity prediction) 등을 통해 누락된 삼중항을 보완하는 지식 그래프 완성 기법(Knowledge graph completion)이 함께 활용되기도 한다(Peng *et al.*, 2023).

특히 도메인 특화 지식 그래프 구축은 의학, 재료과학, 금융, 법률 등 전문 분야의 텍스트에서 의미 있는 엔티티와 관계를 추출하여 지식 구조를 체계화하려는 연구로, 여러 분야에서 활발히 연구되어 왔다(Abu-Salih, 2021). 이러한 연구에서는 도메인 전문가가 정의한 온톨로지를 바탕으로 엔티티 유형과 관계 스키마를 설계하는 방식 등이 활용되었다. 예를 들어, 의학 분야에서는 UMLS나 MeSH와 같은 정교하게 구축된 온톨로지를 활용하여 PubMed 문헌으로 부터 엔티티와 관계를 추출하는 연구들이 수행되었다(Bodenreider, 2004; Zhang *et al.*, 2019). 이와 같은 접근은 높은 정확도를 보장하지만, 전문가 의존성이 크고 새로운 도메인으로 확장하는 데 많은 비용과 시간이 소요된다는 한계가 존재한다.

최근에는 딥러닝 기술의 발전과 함께 도메인 텍스트에 특화된 사전학습 언어 모델을 활용하여 자동으로 지식을 추출하려는 연구가 증가하고 있다(Zhong *et al.*, 2023). 생명 과학 분야에서는 도메인 데이터로 사전 학습된 BioBERT(Lee *et al.*, 2020)가 엔티티 및 관계 추출과 같은 다양한 정보 추출 과제에 활용되어 왔다. 재료과학 분야에서도 재료 과학 문헌에 학습된 MatBERT(Tshitoyan *et al.*, 2019)와 같은 모델들이 개발되었으며, 과학 기술 분야에서는 SciNER 기반 모델(Kulkarni *et al.*, 2022)을 활용한 정보 추출이 수행되었다. 이러한 도메인 특화 사전학습 모델은 범용 언어 모델에 비해 전문 용어 처리 능력과 기술적 문맥 이해도가 뛰어나 지식 그래프 구축의 성능 향상에 큰 기여를 하는 것으로 알려져 있다. 그러나 도메인 특화 모델 기반 접근 방식은 여전히 도메인별로 사전 정의된 온톨로지 또는 레이블링된 도메인 데이터에 크게 의존하고, 각 도메인에서 관계 체계와 개념 정의가 상이하다는 특성(Sharma *et al.*, 2022)으로 인해 다른 분야에 직접적으로 적용하기 어렵다는 한계를 갖는다.

2.2 LLM을 활용한 지식 그래프 생성

최근에는 LLM을 활용하여 지식 온톨로지 및 지식 그래프를 자동으로 구축하려는 연구가 활발하게 진행되고 있으며, 이는 기존 방식이 지닌 한계를 보완하기 위한 새로운 접근으로 주목받고 있다(Bian, 2025).

AutoKG(Zhu *et al.*, 2024)는 멀티 에이전트 시스템을 도입하여 도메인 전문가 역할을 수행하는 에이전트와 웹 검색을 담당하는 에이전트가 협력적으로 지식을 수집·정제함으로써 지식 그래프를 자동으로 구축하는 방법을 제안하였다. 이 접근법은 방대한 온라인 자원을 활용하며 반복적 피드백을 통해 지식 그래프의 품질을 점진적으로 향상시키는 특징을 갖는다. TKGCon(Ding *et al.*, 2024)은 지식 온톨로지 및 지식 그래프 구축 범위를 특정 세부 주제 단위로 축소하여, 특정 주제에 특화된 지식 그래프를 자동으로 생성하는 방법을 제안하였다. 본 접근은 위키피디아 카테고리 구조와 LLM을 결합해 주제 중심의 테마 온톨로지를 구성함으로써 효율적이고 확장 가능한 온톨로지 구축을 가능하게 한다.

또한, Extract, Define, Canonicalize(EDC) 프레임워크(Zhang *et al.*, 2024)는 개방형 정보 추출(Open Information Extraction), 스키마 정의, 정규화(Canonicalization)의 세 단계를 통해 LLM 기반 지식 그래프 구축의 체계적 절차를 제시하였다. EDC는 LLM의 추론 및 언어 이해 능력을 적극 활용하여 스키마 정의 및 엔티티 정규화를 자동화함으로써, 복잡하거나 대규모 스키마 구조를 처리하는 기존 방식의 한계를 완화하고 다양한 도메인에의 적용 가능성을 보여 주었다. 또한 iText2KG(Lairgi *et al.*, 2024)는 후처리 과정을 최소화한 plug-and-play 방식의 zero-shot 및 few-shot 지식 그래프 구축 방법을 제시하여, 과학 논문이나 웹 문서 등 다양한 형태의 비정형 문서로부터 지식 그래프를 점진적으로 확장할 수 있음을 보였다.

이처럼 LLM 기반 접근법은 수작업 중심의 온톨로지 설계나 도메인별 레이블링된 코퍼스에 대한 의존이라는 전통적 한계를 극복할 수 있는 잠재력을 보여주고 있다. 다만, 여전히 대규모 데이터셋에 대한 확장성, 노이즈 관리, 특정 도메인 지식의 일관적 통합 등의 과제가 남아 있어, 이를 보완하기 위한 지속적인 연구가 필요하다.

3. 방법론

3.1 개요

(1) 문제 정의

본 연구의 목적은 레이블링되지 않은 반도체 도메인 특허 텍스트로부터, 사전 정의된 온톨로지 없이 지식 그래프를 자동으로 구축하는 것이다. 입력은 반도체 분야의 특허 문서 코퍼스 D이며, 출력은 엔티티 집합 E^*, 관계 집합 R^*, 그리고 이들로 구성된 삼중항 집합 T^*로 이루어진 최종 지식 그래프 $G^*=(E^*, R^*, T^*)$이다.

각 삼중항 $(s, r, o)\in T^*$는 텍스트로부터 추출된 두 엔티티 $s,o\in E^*$와 엔티티 사이의 의미적 관계 $r\in R^*$를 나타낸다. 이러한 삼중항 집합을 자동으로 구성함으로써, 비정형 특허 텍스트에 내재된 기술 지식을 구조화된 그래프 형태로 표현하는 것을 목표로 한다.

(2) 방법론 개요

반도체 도메인의 특허 문서를 수집하여 데이터셋을 구축한

후, 특허의 분류 체계인 Cooperative Patent Classification(CPC)를 기반으로 세부 주제(subtopic)를 구성한다. 각 세부 주제에 대해 도메인 엔티티 추출 모델을 활용하여 후보 엔티티를 식별하고, LLM을 활용하여 엔티티를 정제하고 관계를 추론함으로써 삼중항(triplet) 형태의 하위 그래프(subgraph)를 생성 한다. 이후 세부 주제별로 도출된 관계들을 표준화하고 통합하여 일관된 관계 체계를 구축한 뒤, 이를 바탕으로 모든 하위 그래프를 하나의 최종 지식 그래프로 통합한다. 전체 프레임워크는 <Figure 1>에 제시되어 있다.

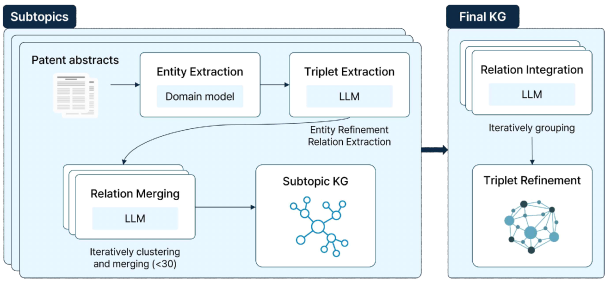


Figure 1. Overall Framework of the Proposed Methodology

3.2 데이터셋 구축

도메인 특화 데이터를 확보하기 위해, 미국 특허청(USPTO)에서 반도체 분야의 주요 영문 특허 문서를 수집하였다. 반도체 제조 공정을 포괄적으로 반영하기 위해, CPC에 기반하여 반도체 분야를 8개의 세부 주제로 구분하고, 2020년부터 2023년까지 공개된 특허를 선별하였다. 각 세부 주제의 CPC 코드와 구체적 범주는 <Table 1>에 제시하였다.

Table 1. CPC classification codes by subtopic.

Subtopic	CPC Code
(1) Silicon wafer fabrication	H01L 21/02, C30B
(2) Oxidation	H01L 21/31, C23C
(3) Photolithography	H01L 21/027, G03F 7/00
(4) Etching	H01L 21/3065, H01L 21/768
(5) Ion implantation	H01L 21/265, H01J 37/32
(6) Deposition	H01L 21/285, H01L 21/768
(7) Chemical mechanical polishing	H01L 21/304, B24B 37/00
(8) Packaging	H01L 21/56, H01L 23/00

전체 문서 코퍼스 D는 수식 (1)과 같이 정의된다. 여기서 $k=8$이고 $S=\{s_1, s_2, \cdots, s_k\}$는 세부 주제 집합이다. 각 세부 주제 s_i에 대응하는 문서 집합 D_{s_i}는 CPC 코드를 기반으로 수집된 특허 초록들의 집합 $D_{s_i}=\{a_{s_i,1}, a_{s_i,2}, \cdots, a_{s_i,n}\}$으로 구성된다.

$$D=\bigcup_{i=1}^{k} D_{s_i} \tag{1}$$

엔티티 및 관계 추출은 세부 주제 s_i 단위로 독립적으로 수행하였다. 이는 동일한 상위 반도체 분야에 속하더라도 세부 주제별로 고유한 기술적 특성과 문맥적 차이가 존재한다고 가정하였기 때문이다. 이러한 특성을 반영하여 관계 정제 및 표준화를 세부 주제 수준에서 수행함으로써, 보다 일관적이고 응집력 있는 관계 체계를 구축할 수 있도록 설계하였다.

3.3 엔티티 추출

반도체 분야에는 레이블링된 데이터가 거의 존재하지 않기 때문에, 해당 도메인에 특화되어 학습된 엔티티 추출 모델이 부재한 상황이다. 또한 도메인마다 엔티티와 관계의 정의가 상이하므로, 재료과학 등 다른 과학 분야의 데이터로 학습된 기존 엔티티 추출 모델(Friedrich *et al.*, 2020; Yamaguchi *et al.*, 2020)을 반도체 특허 데이터에 직접 적용하는 데에는 구조적 한계가 존재한다.

이를 확인하기 위해 재료과학 분야의 개체명 인식(NER) 모델(Kim *et al.*, 2023)을 수집한 반도체 특허 데이터셋에 그대로 적용한 결과를 <Figure 2>에 제시하였다. 해당 모델은 TiO_2, nanocrystals 등 재료과학의 엔티티 유형에 최적화되어 있어, 반도체 도메인에서 핵심적인 개념인 “wafer”, “polishing system”, “etching chamber”와 같은 기술적 단위를 일관되게 식별하지 못하였다. 실제 출력 결과에서도 “polishing system”과 같은 복합 기술 용어를 하나의 의미 단위로 인식하지 못하고 “polishing”과 같이 어절 단위로 분절하는 경향이 확인되었다.

Recently, the high integration of **semiconductors** has increased the processing and **storage capacity** of information per unit area. This has led to demands for large diameter **semiconductor wafers**, miniaturization of **circuit line width**, and **multilayer wiring**. In order to form a **multi-layered** wiring on a **semiconductor wafer**, high-level flatness of the **wafer** is required, and a wafer flattening process is required for such high-level flatness. One of the **wafer** flattening processes is a **wafer polishing** process. The **wafer** polishing process is a step of **polishing** the upper and lower **surfaces** of the **wafer** with a **polishing** pad. The **wafer polishing** process is carried out using a **polishing** system having a **polishing** unit provided with an upper plate, a lower **plate** and a means for supplying **polishing** slurry to the **polishing** unit. A pipe connected to the **polishing** unit for supplying the **slurry** to the **polishing** unit may be provided in the **polishing** system.

Figure 2. Result of Applying a Materials-science NER Model to sEmiconductor Patent Abstracts

또한 LLM을 직접 활용할 경우에도 도메인에 특화된 기술적 특성을 반영한 엔티티를 안정적으로 추출하기 어렵다(Hajikhani *et al.*, 2024; Ling *et al.*, 2025). 이에 의미 있는 기술 엔티티(technical entity)를 추출하기 위한 1차 단계로, 과학 기술 문헌에 특화된 키프레이즈 추출 모델인 KBIR(Keyphrase Boundary Infilling with Replacement)(Kulkarni *et al.*, 2022)을 LLM 의 입력 가이드로 활용하는 하이브리드 구조를 설계하였다. 즉, KBIR이 식별한 키프레이즈의 경계 정보를 LLM 프롬

프트에 제공하여, LLM이 비기술적 일반 명사 대신 도메인 특화 기술 개념을 중심으로 엔티티 정제 및 관계 추출을 수행하도록 한다. 본 실험에서는 과학 논문 초록으로 구성된 Inspec 데이터셋(Hulth, 2003)으로 학습된 KBIR-Inspec 모델을 사용하였다. 해당 모델은 반도체 분야에 특화되어 학습된 것은 아니지만, 과학기술 문서에서 핵심 개념을 구문 경계 단위로 추출하도록 설계되어 있어 후속 LLM 정제 단계에 입력되는 초기 기술 엔티티를 안정적으로 제공하는 데 적합하다.

이를 수식적으로 표현하면, 각 문서 $d \in D$에 대해 수식 (2)와 같이 도메인 특화 엔티티 추출 함수 f_{ent}를 적용하여 후보 엔티티 집합 $\tilde{E}_d$를 얻는다. 여기서 $\tilde{E}_d$는 KBIR-Inspec 모델이 문서 d로부터 추출한 키프레이즈 기반의 기술 엔티티 후보 집합을 의미한다.

$$f_{\text{ent}}(d) \rightarrow \tilde{E}_d \tag{2}$$

추출된 후보 엔티티들은 대소문자 통일, 단수 · 복수형 정규화, 관사 제거 및 특수 기호 정리 등의 기본적인 정규화(normalization) 과정을 거쳐 일관성을 확보하였다. 이때의 정규화는 동일 개념이 표기만 다르게 등장하는 경우를 완화하기 위한 전처리 단계로, 도출된 $\tilde{E}_d$는 최종 엔티티 집합이 아니라, 이후 LLM 기반 엔티티 정제 및 관계 추출 단계의 입력으로 사용되어 추가적인 의미적 보정과 정제가 수행된다.

3.4 LLM 기반 엔티티 정제 및 관계 추출

원문 문장과 KBIR-Inspec 모델을 통해 1차적으로 추출된 엔티티 후보 목록 $\tilde{E}_d$을 포함한 프롬프트를 LLM에 입력하여 엔티티 정제와 관계 추출을 수행하였다. 이 과정에서 LLM은 $\tilde{E}_d$에 포함된 초기의 후보 엔티티를 기본으로 삼되, 문맥상 동일 개념의 표기 변형을 통합하거나 불필요한 수식어를 제거하여 간결하고 기술 중심적인 엔티티 라벨로 정제하였다. 또한 과도하게 길거나 “method”나 “process”와 같은 절차적 의미에 가까운 표현은 핵심 기술 개념 중심으로 재표현하여, 보다 의미 있는 삼중항을 도출하고 최종 지식 그래프의 활용도를 높이고자하였다. 필요할 경우에는 문장 의미를 보존하는 범위 내에서 엔티티의 수정이나 보강을 제한적으로 허용하였다. 최종적으로 정제된 엔티티와 이들 간의 관계를 바탕으로(subject, predicate, object) 구조의 삼중항을 생성하였으며, 사용된 프롬프트는 부록 A의 <Figure 10>에 제시하였다.

이를 수식적으로 나타내면, LLM 기반 함수 f_{LLM}에 원문 d와 후보 엔티티 집합 $\tilde{E}_d$를 입력으로 하여, 정제된 엔티티 집합 E_D와 원시 삼중항 집합 $T_{d,\text{raw}}$를 생성한다.

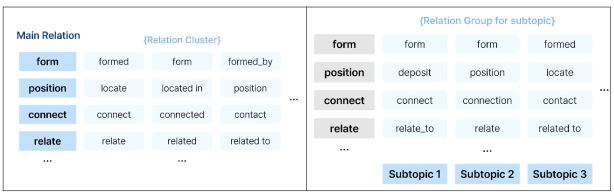
$$f_{\text{LLM}}(d, \tilde{E}_d) \rightarrow (E_d,\ T_{d,\text{raw}}) \tag{3}$$

여기서 원시 삼중항 집합은 $T_{d,\text{raw}} = \{(s, r_{\text{raw}}, o) \mid s, o \in E_d\}$로 정의되며, r_{raw}는 LLM이 생성한 비표준화된 자연어 형태의 관계 표현을 의미한다. 즉, 동일한 의미의 관계가 문장에 따라 서로 다른 표기로 생성될 수 있으므로, 관계 표현의 일관성은 후처리 과정에서 별도로 보정될 수 있다. 또한 엔티티의 경우에도 문장 단위 정제 결과는 문서 전체 관점에서 표기 변형이 남아 있을 수 있으므로, 필요시 문서 수준에서 엔티티 매핑(entity mapping) 및 정규화를 수행하여 동일 개념의 엔티티가 하나의 노드로 일관되게 표현되도록 한다.

3.5 관계 정제 및 표준화

LLM을 활용한 관계 추출 과정에서는 동일하거나 의미적으로 유사한 관계가 다양한 자연어 표현으로 생성되는 문제가 발생할 수 있다. 이를 해결하기 위해 세부 주제별로 생성된 원시 관계 표현들을 군집화하고 통합하여 표준화된 관계 체계를 구축하는 관계 정제 및 표준화 절차를 수행하였다.

<Figure 3>의 (a)는 세부 주제 내에서 관계 표현을 군집화하고 병합하는 과정을, (b)는 세부 주제별로 구축된 관계 그룹을 통합하여 전체 데이터에 대한 관계 사전을 구축하는 과정을 예시적으로 보여준다.



(a) Clustering and merging relations within each subtopic

(b) Integrating relation groups across subtopics

Figure 3. Example of the Relation Refinement Process

(1) 세부 주제 내 관계 정제

각 세부 주제 $s_i \in S$에 대해, 해당 주제에 속한 문서 집합 D_{s_i}에서 생성된 원시 삼중항 집합 $\{T_{d,\text{raw}} \mid d \in D_{s_i}\}$으로부터 원시 관계 표현들의 집합을 다음과 같이 정의한다.

$$R_s^{\text{raw}} = \bigcup_{d \in D_s} \{r \mid (s, r, o) \in T_{d,\text{raw}}\} \tag{4}$$

세부 주제 내 관계 정제 연산자 g_{intra}는 <Figure 4>의 Algorithm 1에 제시된 절차에 따라 수행되며, 전체 과정은 저빈도 관계 제거, 반복적 의미 기반 정제, 최종 관계 집합 도출의 세 단계로 구성된다.

먼저, 초기 단계에서는 R_s^{raw}에 포함된 관계 표현 중 한 번만 등장하거나 빈도수가 낮은 관계 표현을 노이즈로 간주하여 제

거함으로써, 관계 표현의 등장 빈도를 기준으로 초기 관계 집합을 구성한다. 이를 통해 이후 의미 기반 정제 과정에서 발생할 수 있는 불필요한 군집화를 사전에 방지한다.

이후 핵심 단계에서는 초기 단계에서 필터링된 수백 개 규모의 관계 표현들을 대상으로, LLM 기반 단계별 프롬프팅(step-by-step prompting)을 활용하여 현재 관계 집합을 의미·맥락적 유사도에 따라 군집화(clustering)하고, 각 군집에 대표 관계명을 부여함으로써 관계 집합을 점진적으로 병합(merging)한다. 이러한 군집화 - 대표명 부여 과정은 관계 집합의 크기가 사전에 정의된 목표 크기 K 이하가 될 때까지 반복적으로 수행되며, 이를 통해 관계 체계가 점진적으로 응집된다.

마지막 단계에서는 반복 정제 과정을 통해 도출된 관계 집합을 세부 주제 s_i에 대한 정제된 관계 사전 $R_s = g_{\text{intra}}(R_s^{\text{raw}})$으로 반환한다. 이 과정을 통해 각 세부 주제 내에서 중복되거나 유사한 관계 표현들이 일관된 대표 관계로 통합된다. 사용한 프롬프트의 예시는 부록 A의 <Figure 11>에 제시한다.

Algorithm 1: Intra-subtopic Iterative Relation Refinement (g_{intra})

Input: Raw relation set $\mathcal{R}_s^{\text{raw}}$, frequency threshold τ_{freq}, target size K
Output: Refined relation dictionary $\mathcal{R}_s$

$\mathcal{R}^{(0)} \leftarrow \text{FilterLowFrequency}(\mathcal{R}_s^{\text{raw}}, \tau_{\text{freq}})$;
$k \leftarrow 0$;
while $|\mathcal{R}^{(k)}| > K$ **do**
  $\mathcal{C} \leftarrow \text{LLMCluster}(\mathcal{R}^{(k)})$; // Semantic clustering via step-by-step prompting
  $\mathcal{R}^{(k+1)} \leftarrow \emptyset$;
  foreach *cluster* $c \in \mathcal{C}$ **do**
    $r_{\text{rep}} \leftarrow \text{SelectRepresentativeLabel}(c)$; // LLM or heuristic
    $\mathcal{R}^{(k+1)} \leftarrow \mathcal{R}^{(k+1)} \cup \{r_{\text{rep}}\}$;
  $k \leftarrow k + 1$;
$\mathcal{R}_s \leftarrow \mathcal{R}^{(k)}$;
return $\mathcal{R}_s$;

Figure 4. Pseudo-code of Intra-subtopic Iterative Relation Refinement (g_{intra})

(2) 세부 주제 간 관계 통합

세부 주제별로 구축된 관계 사전 $\{R_{s_i} | s_i \in S\}$에는 서로 다른 세부 주제에서 생성되었으나 의미적으로 유사한 관계 표현들이 중복되어 포함될 수 있다. 이에 이러한 중복 관계들을 통합하여, 전체 데이터에 대해 일관된 표준 관계 체계를 구축한다. 이를 위해 LLM 기반 관계 통합 연산자 g_{inter}를 적용하여 최종 관계 사전 R^*을 다음과 같이 정의한다.

$$R^* = g_{\text{inter}}(\{R_s | \ s \in S\}) \tag{5}$$

관계 통합 연산자 g_{inter}의 절차적 구현은 <Figure 5>의 Algorithm 2에 제시되어 있으며, 전체 과정은 세부 주제별 관계 집합의 통합, 반복적 의미 기반 군집화 및 대표 관계 선택, 최종 관계 집합 도출의 단계로 구성된다.

구체적으로, 각 세부 주제에서 정제된 관계 집합들을 하나의 통합 관계 집합으로 병합한 후, LLM 기반 의미 군집화를 반복적으로 수행하여 각 군집을 대표하는 관계 표현을 선택함으로써 관계 집합을 점진적으로 축소한다. 이 과정은 더 이상 관계 집합이 변화하지 않을 때까지 반복되며, 의미적으로 안정적인 대표 관계 집합이 도출되면 이를 최종 관계 사전 R^*로 반환한다.

이를 통해 서로 다른 세부 주제에서 생성된 중복되거나 유사한 관계 표현들은 일관된 대표 관계로 통합되며, 전체 코퍼스에 대해 공통적으로 적용 가능한 표준화된 관계 체계가 구축된다. 이때 사용한 프롬프트의 예시는 부록 A의 <Figure 12>에 제시한다.

Algorithm 2: Inter-subtopic Relation Integration (g_{inter})

Input: Set of refined relation dictionaries $\{\mathcal{R}_s \mid s \in \mathcal{S}\}$
Output: Unified relation dictionary $\mathcal{R}^*$

$\mathcal{R}_{\text{all}} \leftarrow \bigcup_{s \in \mathcal{S}} \mathcal{R}_s$;
repeat
  $\mathcal{C}_{\text{all}} \leftarrow \text{LLMCluster}(\mathcal{R}_{\text{all}})$;
  $\mathcal{R}_{\text{all}} \leftarrow \emptyset$;
  foreach *cluster* $c \in \mathcal{C}_{all}$ **do**
    $r_{\text{rep}} \leftarrow \text{SelectRepresentativeLabel}(c)$;
    $\mathcal{R}_{\text{all}} \leftarrow \mathcal{R}_{\text{all}} \cup \{r_{\text{rep}}\}$;
until $\mathcal{R}_{all}$ *stops changing*;
$\mathcal{R}^* \leftarrow \mathcal{R}_{\text{all}}$;
return $\mathcal{R}^*$;

Figure 5. Pseudo-code of inter-subtopic Relation Integration (g_{inter})

Algorithm 1과 Algorithm 2에 적용된 LLM 기반 프롬프트는 원시 텍스트로부터 개별 삼중항을 직접 생성하는 도구가 아니라, 관계 표현들의 집합에 반복적으로 적용되어 의미적으로 유사한 관계 표현을 군집화하고 통합함으로써 관계 표현을 점진적으로 정규화하는 역할을 수행한다. 이 과정에서 LLM은 서로 다른 표기로 생성된 관계 표현들을 하나의 대표 관계로 병합하여, 세부 주제 내에서는 유사한 관계 표현의 중복을 제거하고, 세부 주제 간에서는 상이한 명명 체계를 공통된 관계 체계로 정렬한다. 이러한 반복적 의미 압축 과정을 통해, 사전 정의된 온톨로지 없이도 문서 집합 전체에 적용 가능한 응집력 있는 관계 사전을 자동으로 구축할 수 있다.

(3) 관계 매핑

LLM이 생성한 원시 관계 표현 r_{raw}는 표현상의 중복과 변형을 포함하므로, 이를 표준 관계 사전 R^*에 매핑하기 위한 관계 변환 함수 ϕ를 다음과 같이 정의한다.

$$\phi : r_{\text{raw}} \rightarrow r, \ \ r \in R^* \tag{6}$$

이때 함수 ϕ는 의미적 유사도를 기준으로 각 원시 관계 표현을 관계 사전 R^*내 가장 적합한 대표 관계로 할당한다.

이를 통해 원시 삼중항 (s, r_{raw}, o)는 표준화된 삼중항 $(s, \phi(r_{\text{raw}}), o)$로 변환되며, 최종 지식 그래프에서 관계 표현

의 일관성과 해석 가능성이 확보된다.

3.6 최종 지식 그래프 구축

관계 정제 및 표준화 과정을 통해 구축된 최종 관계 사전 R^*을 기준으로, 각 문서에서 추출된 원시 관계 표현들을 표준 관계로 매핑하여 최종 삼중항을 생성한다. 이때 정제된 표준 관계의 총 개수는 30개이다. 이를 통해 세부 주제별로 생성된 하위 그래프들을 통합하고, 전체 문서에 대해 하나의 일관된 최종 지식 그래프를 구축한다.

최종 삼중한 집합 T^*는 다음의 수식 (7)과 같이 정의된다. 여기서 (s, r_{raw}, o)는 각 문서 $d \in D$로부터 생성된 원시 삼중항이며, ϕ는 관계 정제 및 표준화 단계에서 정의된 관계 매핑 함수이다.

$$T^* = \left\{ (s, \phi(r_{\text{raw}}), o) \mid (s, r_{\text{raw}}, o) \in \bigcup_{d \in D} T_{d,\text{raw}} \right\} \quad (7)$$

최종 지식 그래프 구축 절차는 <Figure 6>의 Algorithm 3에 제시되어 있으며, 전체 과정은 원시 삼중항 통합, 관계 매핑을 통한 삼중항 표준화, 그래프 구성 요소 집합 도출의 세 단계로 구성된다.

먼저, 초기 단계에서는 전체 문서 집합 D로부터 생성된 원시 삼중항 집합들을 모두 병합하여 전체 원시 삼중항 집합을 구성한다. 이를 통해 문서 단위로 분산되어 있던 관계 정보를 하나의 통합된 삼중항 공간에서 처리할 수 있도록 한다.

이후 핵심 단계에서는 각 원시 삼중항 (s, r_{raw}, o)에 대해, 관계 정제 및 표준화 단계에서 정의된 관계 매핑 함수 ϕ를 적용하여 원시 관계 표현 r_{raw}를 표준 관계 $r \in R^*$로 변환한다. 이를 통해 모든 삼중항은 $(s, \phi(r_{\text{raw}}), o)$ 형태의 표준화된 삼중항으로 일관되게 변환된다.

마지막 단계에서는 표준화된 삼중항 집합에 등장하는 엔티티들을 수집하여 엔티티 집합 E^*을 구성하고, 이를 표준 관계 사전 R^* 및 최종 삼중항 집합 T^*과 함께 최종 지식 그래프 $G^* = (E^*, R^*, T^*)$로 정의한다. 이 과정을 통해 관계 표현의 일관성과 해석 가능성이 보장된 통합 지식 그래프가 구축된다.

Algorithm 3: Final Knowledge Graph Construction

Input: Raw triplets $\{T_{d,\text{raw}}\}_{d \in \mathcal{D}}$, final relation dictionary $\mathcal{R}^*$, relation mapping function ϕ

Output: Final knowledge graph $\mathcal{G}^* = (\mathcal{E}^*, \mathcal{R}^*, \mathcal{T}^*)$

$\mathcal{T}_{\text{raw\_all}} \leftarrow \bigcup_{d \in \mathcal{D}} T_{d,\text{raw}}$;
$\mathcal{T}^* \leftarrow \emptyset$;
foreach $(h, r_{raw}, t) \in \mathcal{T}_{raw\_all}$ **do**
  $r \leftarrow \phi(r_{\text{raw}})$; // Map to standardized relation
  $\mathcal{T}^* \leftarrow \mathcal{T}^* \cup \{(h, r, t)\}$;
$\mathcal{E}^* \leftarrow \{e \mid e \text{ appears in } \mathcal{T}^*\}$;
return $\mathcal{G}^* = (\mathcal{E}^*, \mathcal{R}^*, \mathcal{T}^*)$;

Figure 6. Pseudo-code of Final Knowledge Graph Construction Via Relation Mapping (ϕ)

4. 실험

4.1 실험 설정

모든 실험 과정에서 LLM은 GPT(Achiam *et al.*, 2023) 계열 모델을 사용하였다. 구체적으로, 엔티티 정제 및 관계 추출 단계에서는 GPT-4.1(*gpt-4.1-2025-04-14*) API를 활용하였으며, 비교 및 평가 단계에서는 경량화된 GPT(*gpt-5-mini*) API를 사용하였다. 모든 생성 과정에서 temperature 값은 1.0으로 고정하여, 생성 다양성을 유지하면서도 반복 실험 간 일관성을 확보하였다.

실험은 3장에서 설명한 반도체 특허 초록 데이터셋을 대상으로 수행되었으며, 비교 대상 방법들에는 동일한 입력 문서와 동일한 평가 절차를 적용하였다. 또한, 제안한 프레임워크에서 수행되는 세부 주제 내 관계 정제 단계의 목표 관계 수 K는 30으로 설정하였다.

4.2 비교 실험 설계

제안한 하이브리드 프레임워크의 성능을 분석하기 위해 서로 다른 세 가지 지식 그래프 구축 방식을 비교 대상으로 정의한다. 비교 대상은 Domain-only 방식, LLM-only 방식, 그리고 Proposed Hybrid 방식으로 구성된다. 이들 세 방법은 각각 규칙 기반 정보 추출, 순수 LLM 기반 추출, 그리고 두 접근을 결합한 하이브리드 전략을 대표한다. 이들 방법은 정보 추출에 대한 가정과 제약 조건이 상이하므로, 반도체 특허와 같은 도메인 특화 텍스트에서 생성되는 지식 그래프의 특성을 비교하기 위한 기준으로 활용된다.

Domain-only 방식은 LLM을 사용하지 않고, 과학 기술 문헌에 특화된 키프레이즈 추출기와 전통적인 자연어 처리 기법만을 이용하여 지식 그래프를 구축하는 방법이다. 먼저 특허 텍스트로부터 기술 관련 엔티티를 추출하기 위해 KBIR-Inspec 모델을 적용하며, 이 단계에서 추출된 키프레이즈는 추가적인 의미적 정제 없이 엔티티로 직접 사용된다. 이후 엔티티 간 관계는 의존 구문 분석(Dependency Parsing) 결과를 기반으로 도출된다. 도메인 특화 학습 데이터나 관계 추출 모델이 존재하지 않는 상황을 고려하여, 학습 기반 접근 대신 문장의 구문 구조에 기반한 규칙 기반 접근을 적용한다. 문장 내 엔티티 쌍 사이의 의존성 트리 상 최단 경로와 동사 중심의 주어-동사-목적어(SVO) 구조를 분석하여 관계 후보를 생성하며, “includes”, “comprises”, “consists of”와 같은 빈번한 표현과 나열 구조를 활용하여 하나의 동사가 여러 엔티티를 연결하는 경우를 반영한다. 이 방식은 관계를 문장 수준의 구문 정보에 기반하여 추정하는 전통적인 정보 추출 접근법에 해당한다.

LLM-only 방식은 도메인 특화 모델이나 규칙 기반 구문 분석을 사용하지 않고, LLM을 단일 단계(end-to-end) 정보 추출기로 활용하여 텍스트로부터 엔티티와 관계를 동시에 추출하는 방법이다. 입력된 텍스트에 대해 LLM이 문맥을 기반으로

주요 기술 개념을 엔티티로 식별하고, 이들 사이의 관계를 (subject, relation, object) 형태의 삼중항으로 직접 생성한다. 이 때 사용되는 구체적인 프롬프트는 부록 A의 <Figure 13>에 제시한다.

Proposed Hybrid 방식은 3장에서 서술한 바와 같이, Domain-only 방식에서 추출된 도메인 특화 엔티티를 입력 제약으로 활용하여 LLM을 통해 엔티티 정제 및 관계 추출을 수행하는 접근법이다. 이 방식은 도메인 기반 엔티티 식별을 통해 관계 추출 과정의 잡음을 완화하면서도, LLM의 언어적 추론 능력을 활용하여 관계 표현을 확장하는 것을 목표로 한다.

이와 같은 비교 실험 구성은 서로 다른 정보 추출 전략이 반도체 특허로부터 구축되는 지식 그래프의 특성에 어떠한 차이를 만들어내는지를 체계적으로 분석하기 위한 것이다.

4.3 평가 방법

본 연구는 레이블이 없는 도메인 텍스트로부터 지식 그래프를 구축하는 문제를 다루므로, 정답 삼중항을 기준으로 한 전통적인 정밀도(precision) 및 재현율(recall) 기반 평가를 적용하기 어렵다. 이러한 설정은 정답이 존재하지 않는 zero-resource 또는 약지도(weakly supervised) 환경에 해당한다.

Choi *et al.*(2025)은 이러한 환경에서 지식 그래프의 품질을 내적(intrinsic) 품질과 외적(extrinsic) 품질로 구분하여, 그래프의 구조적 일관성과 같은 내부적 특성과 실제 응용에서의 유용성을 함께 고려하는 다면적 평가의 필요성을 제시하였다. 이에 따라 정답 삼중항이 존재 하지 않는 환경에서는 단일 정량 지표에 의존하기보다, 관련 연구들에서 활용되어 온 다양한 평가 관점을 종합적으로 고려하는 접근이 주로 사용된다.

예를 들어, 원문 텍스트에 대한 충실성(faithfulness) 또는 사실성(factuality)은 생성 모델 평가 분야에서 널리 논의되어 왔으며 (Maynez *et al.*, 2020; Manakul *et al.*, 2023), 그래프 자체의 구조적 품질은 기존 지식 그래프 품질 분석 연구에서 주요 평가 기준으로 다루어져 왔다(Zaveri *et al.*, 2015; Paulheim, 2016; Seo *et al.*, 2022; Xue *et al.*, 2022). 또한, 구축된 지식 그래프의 유용성은 하위 응용 과제에서의 활용 가능성을 통해 간접적으로 평가될 수 있다(Heist *et al.*, 2023).

이에 따라 구축된 지식 그래프의 품질을 (1) 충실성(faithfulness), (2) 구조적 품질 (structural quality), 그리고 (3) 하위 과제 기반 외재적 성능(extrinsic performance)의 세 가지 관점에서 평가한다.

(1) 충실성(Faithfulness)

첫 번째 평가는 추출된 삼중항이 원문 텍스트에 의해 실제로 지지되는지를 측정하는 것을 목표로 한다. 생성된 출력이 원문에 의해 지지되는지(supported)를 평가하는 문제는 LLM의 환각과 사실성문제를 분석하기 위해 자연어 생성 및 정보 추출 연구에서 널리 다루어져 왔다(Maynez *et al.*, 2020; Manakul *et al.*, 2023). LLM-as-a-judge(Liu *et al.*, 2023) 전략을 적용하여, 하나의 원문 텍스트와 하나의 삼중항을 입력으로 받아 해당 삼중항이 텍스트에 의해 명시적으로 언급되거나 논리적으로 지지되는지를 판단한다. 평가 모델로 GPT-5-mini (*gpt-5-mini*)를 사용하였으며, 사용된 프롬프트는 부록 A의 <Figure 14>에 제시한다.

평가 모델은 추출된 각 삼중항 $t \in T$을 SUPPORTED, NOT SUPPORTED, UNCLEAR의 세 가지 범주 중 하나로 분류한다. SUPPORTED는 삼중항의 내용이 텍스트에 의해 직접적으로 표현되거나 명확하게 추론 가능한 경우를 의미하며, NOT SUPPORTED는 텍스트에 의해 뒷받침되지 않는 정보가 포함된 경우를 의미한다. UNCLEAR는 해당 삼중항이 원문 표현과의 관계를 명확히 판단하기 어려운 경우에 한정하여 부여된다. 이러한 판정 결과를 전체 삼중항 집합에 대해 집계하여, SUPPORTED로 분류된 삼중항의 비율을 Supported Triplet Rate(STR), NOT SUPPORTED로 분류된 삼중항의 비율을 Hallucination Triplet Rate(HTR) 로 정의한다(수식 (8)).

$$\mathrm{STR} = \frac{|\mathrm{t} \in T \mid t \text{ is SUPPORTED}|}{|T|}, \quad \mathrm{HTR} = \frac{|\mathrm{t} \in T \mid t \text{ is NOT SUPPORTED}|}{|T|} \tag{8}$$

(2) 구조적 품질(Structural quality)

두 번째 평가는 생성된 지식 그래프의 내부 구조적 특성을 분석하는 것을 목표로 한다. 기존의 지식 그래프 품질 평가 연구에서는 완전성(completeness), 일관성(consistency), 간결성(conciseness) 등이 지식 그래프의 핵심 구조적 특성으로 논의되어 왔으며(Zaveri *et al.*, 2016; Paulheim, 2017), 단순한 삼중항수나 엔티티 수와 같은 규모 기반 지표만으로는 지식 그래프의 품질을 충분히 설명하기 어렵다는 한계가 지적되어 왔다(Seo *et al.*, 2022). 이에 따라 중복성, 비정상적 구조, 엔티티 활용도와 같은 구조적 특성을 반영하여 내적 품질을 정량화하기 위한 평가 지표의 필요성이 제기되었고, 실제로 Seo *et al.* (2022)은 온톨로지 활용도, 구조적 완전성, 관계 표현의 다양성을 반영하는 구조 기반 지표를 제안하여 지식 그래프의 품질이 단순 규모뿐만 아니라 내부 구조가 얼마나 합리적으로 조직되어 있는지에 의해 좌우됨을 보였다.

이러한 논의를 바탕으로, 기존 지식 그래프 품질 평가 문헌에서 논의되어 온 내적 품질 개념들을 본 연구의 설정에 맞게 정량화한다. 구체적으로, 지식 그래프 G와 엔티티 집합 E에 대해 세 가지 구조적 지표, 즉 Entity Completeness, Redundancy Rate, Structural Inconsistency Rate을 정의한다. 각 지표의 정의는 <Table 2>에 제시되어 있다. 이 지표들은 그래프가 추출된 엔티티를 관계 구조에 얼마나 활용하는지, 중복되거나 비정상적인 삼중항이 어느 정도 포함되는지를 내부 구조의 관점에서

정량적으로 분석하기 위한 기준으로 사용된다.

Table 2. Structural Knowledge Graph Evaluation Metrics

Metric	Definition
Entity Completeness	The proportion of extracted entities that appear in at least one knowledge graph triple. $\lvert\{e \in E \mid \exists (s,r,o) \in G,\ e = s \vee e = o\}\rvert / \lvert E \rvert$
Redundancy Rate	The proportion of duplicated (subject, relation, object) triples in the extracted graph. $1 - \lvert \text{unique}(G) \rvert / \lvert G \rvert$
Structural Inconsistency Rate	The proportion of triples in which the subject and the object are identical. $\lvert\{(s,r,o) \in G \mid s = o\}\rvert / \lvert G \rvert$

(3) 하위 과제 기반 외재적 성능(Extrinsic performance)

마지막으로, 구축된 지식 그래프의 실제 활용 가능성을 평가하기 위해 특허 메타데이터인 CPC 코드를 활용한 외재적(extrinsic) 평가를 수행한다. 최근 연구들은 정답 삼중항이 없는 경우에도, 지식 그래프를 이용해 검색이나 분류와 같은 하위 응용 과제(downstream tasks)를 수행하고 그 성능을 측정함으로써 그래프의 실질적 유용성을 평가할 수 있음을 보여주었다(Heist *et al.*, 2023; Choi *et al.*, 2025).

CPC 코드는 지식 그래프 구축 과정과 독립적인 외부 분류 기준이므로, 그래프 기반 문서 표현이 특허의 주제적 구조를 어느 정도 반영하는지를 간접적으로 평가하는데 활용할 수 있다. 또한 반도체 특허는 기술 요소가 분산 서술되는 경우가 많아 단순 키워드 기반 유사도만으로는 문서 간 관련성을 충분히 포착하기 어렵다. 이러한 맥락에서 CPC 기반 검색 성능은 지식 그래프가 특허 문서에 내재된 기술 의미 구조를 얼마나 효과적으로 통합하여 표현하고 있는지를 평가하는 실질적인 외재적 평가 지표로 활용될 수 있다.

구체적으로, 각 문서 d_i로부터 구축된 지식 그래프는 그래프 임베딩 기법인 node2vec(Grover *et al.*, 2016)을 통해 고정 차원의 임베딩 벡터 z_e로 변환된다. 문서 단위 표현 d_i는 수식 (9)와 같이 해당 문서의 지식 그래프에 포함된 엔티티 임베딩의 평균으로 정의한다.

$$d_i = \frac{1}{|E_i|} \sum_{e \in E_i} z_e \tag{9}$$

문서 벡터 간의 코사인 유사도를 기반으로 동일한 CPC 코드를 공유하는 문서들을 관련 문서로 간주하여 Recall@k, MAP@k, nDCG@k를 계산한다. 이러한 지표들은 정보 검색 분야에서 널리 사용되는 표준 랭킹 품질 지표이며, 기존 문헌에서 제시된 표준 정의를 따른다(Koren *et al.*, 2009; Schütze *et al.*, 2008; Järvelin and Kekäläinen, 2002).

충실성 평가는 문장 단위, 외재적 평가는 문서 단위에서 수행되며, 이와 같이 서로 다른 분석 단위를 갖는 세 가지 평가 축을 통해 Domain-only, LLM-only, Proposed Hybrid 방식의 특성을 종합적으로 비교한다.

(4) 구조적 특성 분석

위의 세 가지 평가는 정량적 지표를 기반으로 한 비교를 제공하지만, 이러한 결과만으로는 서로 다른 구축 방식에 따라 형성되는 그래프 구조의 차이를 충분히 설명하기 어렵다. 이에 정량적 평가 결과의 해석을 보완하기 위해, 관계 분포와 경로 구조를 중심으로 한 구조 분석을 추가로 수행한다. 이는 지식 그래프의 품질을 직접 평가하기 위한 지표로 사용되지 않으며, 서로 다른 구축 전략이 관계 스키마의 분포와 엔티티 간 연결 구조에 어떠한 차이를 만들어내는 지를 설명하기 위한 보조적 분석으로 활용된다.

5. 실험 결과 및 분석

본 장에서는 제안한 하이브리드 지식 그래프 구축 프레임워크의 성능을 정량적 비교와 정성적 분석을 통해 종합적으로 평가한다. 4.3절에서 정의한 세 가지 축을 기준으로 Domain-only, LLM-only, Proposed Hybrid 방법을 비교한다. 5.1절에서는 세 방법의 전반적 성능을 비교하고, 5.2절에서는 관계 분포 및 경로 구조를 중심으로 그래프의 구조적 · 의미적 특성을 분석한다. 마지막으로 5.3절에서는 세부 주제별 하위 그래프 사례를 통해 제안 방법이 기술 문서의 구조적 논리와 설계 맥락을 어느 정도까지 반영하는지 정성적으로 논의한다.

5.1 성능 비교

3.4절 관계 추출에서 서술한 바와 같이, 제안한 방법이 각 세부 주제별로 수행한 관계의 군집화 및 병합 결과는 <Table 3>에 제시하였다. LLM을 통해 추출된 초기 관계는 세부 주제별 평균 187개의 관계가 생성되었으나, 군집화 및 병합 과정을 거친 후 평균 25개 수준으로 감소하여, 다수의 중복되거나 유사한 관계가 소수의 의미적으로 일관된 관계 그룹으로 효과적으로 통합되었음을 확인할 수 있다.

Table 3. Results of Clustering and Merging Relations Per Subtopic

Subtopic		1	2	3	4	5	6	7	8
#	Triplets	595	511	570	635	564	677	566	593
#	Raw relations	186	184	187	190	200	181	182	183
#	Merged relations	16	22	26	24	25	25	28	33

이러한 관계 정제 과정을 거쳐 구축된 최종 지식 그래프의 기본 통계는 <Table 4>에 제시되어 있다. 기본 통계는 제안한 관계 정제 과정이 그래프의 규모와 구조를 효과적으로 정리했음을 보여주며, 이후 평가는 이 최종 그래프를 대상으로 수행한다.

Table 4. Overview of Final Triplet Statistics

# of triples	# of unique entities	# of relations
4,551	2,899	30

(1) 충실성 (Faithfulness)

<Table 5>는 LLM-as-a-judge 기반 충실성 평가 결과를 보여준다. Domain-only 방식은 환각 삼중항이 상대적으로 많이 발생하여 전반적인 텍스트 충실성이 가장 낮게 나타난 반면, LLM-only 방식은 이를 크게 개선하여 지원된 삼중항 비율이 현저히 증가하였다. 제안한 Hybrid 방식은 두 기준 방식 대비 가장 안정적인 성능을 보이며, 지원된 삼중항 비율이 가장 높고 환각 및 불확실 삼중항의 비율이 가장 낮게 관측되었다. 이는 도메인 지식과 LLM 추론을 결합한 접근이 텍스트에 대한 충실성을 효과적으로 향상시킴을 시사한다. 다만, 해당 평가는 LLM 평가자에 기반한 내적 평가이므로, 평가 모델의 판정 편향 가능성을 완전히 배제할 수는 없으며, 본 결과는 동일한 평가 방식 하에서의 상대 비교로 고려할 수 있다.

Table 5. Faithfulness Evaluation Results Based on LLM-as-a-judge

Method	Supported Triplet Rate(%) ↑	Hallucination Triplet Rate(%) ↓	Unclear Triplet Rate(%)
Domain-only	60.23	35.78	3.98
LLM-only	89.47	8.42	2.11
Hybrid	97.35	2.21	0.44

(2) 구조적 품질(Structural quality)

<Table 6>은 세 방법의 구조적 품질을 비교한 결과를 제시한다. 전반적으로 LLM-only 방식은 추출된 엔티티가 삼중항 구조에 포함되는 비율이 가장 높아, 관계 구조가 상대적으로 풍부하게 형성되는 경향을 보인다. 반면 Hybrid 방식은 중복 삼중항이나 주어 · 목적어 동일과 같은 비정상적 구조가 가장 낮게 관측되어, 구조적 잡음이 상대적으로 적은 그래프를 형성한 것으로 나타났다. 이러한 양상은 규칙 기반 추출이 구조적으로 보수적인 그래프를 생성하는 반면, LLM 기반 추출은 표현 범위를 넓히는 과정에서 구조적 변동성이 함께 증가할 수 있다는 점을 시사한다. 따라서 구조적 품질 지표는 본 논문에서 의미적 유용성을 단독으로 설명하는 지표로 사용하지 않고, 외재적 성능 지표와 병행하여 분석한다.

Table 6. Intrinsic Structural Quality Evaluation Results

Method	Entity Completeness ↑	Redundancy Rate ↓	Structural Inconsistency ↓
Domain-only	0.8813	0.0022	0.0027
LLM-only	**0.9408**	0.0019	0.0047
Hybrid	0.8875	**0.0003**	**0.0009**

(3) 하위 과제 기반 외재적 성능(Extrinsic performance)

<Table 7>은 CPC subclass 기준의 문서 검색 성능을 비교한 결과를 보여준다. Hybrid 방식은 MAP 및 nDCG의 랭킹 품질 지표에서 전반적으로 가장 우수한 값을 기록하여, 동일 CPC 범주의 특허가 상위 순위에 더 안정적으로 배치되는 경향을 보였다. 다만 재현율인 Recall은 k 값에 따라 방법 간 차이가 일정하지 않게 나타나므로, 본 결과는 Hybrid 방식이 특히 상위 랭크 품질 측면에서 상대적으로 유리함을 보여주는 것으로 파악할 수 있다.

또한 CPC 코드는 기술 구성 요소와 기능의 존재 여부를 중심으로 분류되며, 공정 단계의 순서나 인과 구조와 같은 세부 관계 의미를 직접 반영하지 않는다. 따라서 CPC 기반 검색 성능은 지식 그래프가 포착한 의미 구조 중 문서 수준 유사도의 일부 측면을 간접적으로 보여주는 지표이며, 관계 의미의 정교함이나 경로 수준 추론 가능성을 직접 측정하는 결과로 해석하기는 어렵다.

Table 7. CPC-based Extrinsic Retrieval Performance (CPC-subclass)

Method	Recall @10	Recall @20	MAP @10	MAP @20	nDCG @10	nDCG @20
Domain-only	0.0262	0.0515	0.9258	0.9226	0.9515	0.9511
LLM-only	0.0276	0.0538	0.9427	0.9333	0.9603	0.9562
Hybrid	**0.0284**	**0.0546**	**0.9632**	**0.9568**	**0.9728**	**0.9686**

5.2 지식 그래프의 구조적 특성 분석

본 절에서는 Domain-only, LLM-only, Hybrid 방식으로 구축된 지식 그래프의 차이를 관계 구조의 분포 특성과 그래프 응집성의 관점에서 분석한다. 각 방법이 생성한 지식 그래프가 관계 표현을 어떠한 구조로 조직하는지를 비교하는 것을 목표로 하여, 그래프의 구조적 특성을 기술적으로 비교하는데 초점을 둔다. 이를 위해 관계 유형 분포에 대한 정량적 분석과 대표 특허 문서를 대상으로 한 경로 수준의 구조 분석을 수행하였다.

(1) 관계 분포 분석

먼저, 각 방법으로 생성된 지식 그래프에서 관계 유형이 분포되는 양상을 비교하여 <Table 8>에 제시 한다. Domain-only 방식은 규칙 기반 의존 구문 분석을 통해 문장 단위의 다양한

관계 표면형을 그대 로 유지하는 특성을 가지며, 그 결과 동일하거나 유사한 의미를 갖는 관계가 서로 다른 관계 유형으로 분리되어 나타나는 경향이 관찰된다. 이러한 구조에서는 관계 유형의 수가 크게 증가하고, 관계 표현이 그래프 전반에 걸쳐 분산되는 양상을 보인다. 이는 그래프가 포함하는 정보의 양과는 무관하게, 관계 스키마가 상대적으로 비응집적인 형태로 구성될 수 있음을 시사한다.

LLM-only 방식은 추론을 통해 관계를 생성함으로써, 규칙 기반 방식에 비해 관계 표현의 변이가 일부 감소하는 경향을 보인다. 그러나 명시적인 관계 정규화나 병합 절차가 포함되지 않기 때문에, 유사한 의미를 갖는 관계들이 여전히 서로 다른 표현으로 생성되는 경우가 발생하며, 관계 스키마가 완전히 정제된 형태로 수렴하지는 않는다. 이로 인해 관계 표현의 다양성과 구조적 변동성이 함께 나타난다.

반면, Hybrid 방식은 관계 병합 및 정규화 과정을 통해 의미적으로 유사한 관계 표현을 대표 관계로 통합한다. 그 결과 관계 유형의 수가 상대적으로 제한되며, 관계 스키마가 보다 응집된 형태로 구성된다. 이러한 구조적 응집성은 관계 표현의 표면적 다양성을 축소하는 대신, 관계 유형 간의 정합성을 높이는 방향으로 작용한다. 따라서 Hybrid 방식의 관계 분포 특성은 정보의 축소라기보다는, 관계 스키마가 보다 일관되게 조직된 결과로 파악할 수 있다.

Table 8. Relation Type Distribution Statistics Across Different Baseline Methods

Method	# Triplets	# Relation Types	Distribution Score	Normalized Distribution Score
Domain-only	7,070	3,471	10.05	0.85
LLM-only	7,818	1,597	7.66	0.72
Hybrid	4,551	30	3.61	0.65

(2) 경로 기반 의미 분석

관계 분포의 차이가 실제 그래프 구조에 어떠한 영향을 미치는지를 살펴보기 위해, 동일한 특허 문서를 대상으로 경로 수준의 구조를 비교한 결과를 <Figure 7>에 제시하였다. 사례로 사용한 특허 문서(US10847349)의 경우, Domain-only 그래프에서는 주요 기술 엔티티들이 문장 단위 관계를 통해 개별적으로 연결되어 있으나, 그래프 차원에서 엔티티 간 연결이 연속적인 구조로 확장되는 양상은 제한적으로 나타났다. 이는 관계 표현이 다수 존재하더라도, 관계 유형 간의 정합성이 낮을 경우 경로 구조가 명확히 형성되지 않을 수 있음을 보여준다.

이에 비해 Hybrid 방식으로 구축된 그래프에서는 동일한 엔티티 집합을 기반으로 관계가 정제되고 통합되면서, 일부 엔티티 쌍 사이에서 보다 해석 가능한 다단계 연결이 형성되는 양상이 관찰되었다. 이러한 차이는 관계 스키마의 응집성이 경로 구조의 해석 가능성에 영향을 미칠 수 있음을 보여준다.

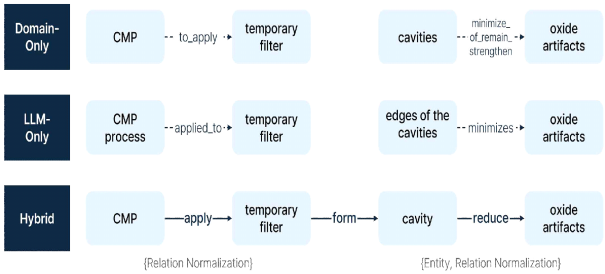


Figure 7. Path-level Semantic Connectivity Comparison Across the Baseline Methods

종합하면, Domain-only 방식은 관계 표면형의 다양성이 높아 관계 구조가 분산되는 경향을 보이는 반면, Hybrid 방식은 관계 병합을 통해 관계 유형수를 축소하고 그래프 구조를 보다 응집된 형태로 구성한다. 본 절의 분석은 이러한 구조적 차이를 관계 분포와 경로 구조의 관점에서 기술적으로 비교한 것으로, 의미적 품질이나 응용 성능에 대한 평가는 5.1절의 외재적 성능 결과와 함께 해석될 필요가 있다.

5.3 정성 분석 및 사례 연구

정량 지표는 방법 간 성능 비교에 유용하지만, 구축된 지식그래프가 특허 기술의 구조적 논리와 설 계 맥락을 어느 정도까지 반영하는지 직접적으로 설명하기에는 한계가 있다. 특히 공정 단계의 위계 구조나 설계 논리, 그리고 조건에서 성능으로 이어지는 연결은 문장 단위의 단순 관계 집계만으로 충분히 포착되기 어렵다. 이에 본 절에서는 대표 특허 사례를 선정하여 하위 그래프 수준에서 제안 방법의 해석 가능성과 활용 가능성을 논의한다.

먼저 구축된 관계 그룹의 대표 관계명(main relation)과 이에 대응되는 원시 관계 표현의 예시는 <Table 9>에 제시하였다. 세부 주제별 관계 그룹을 모두 통합 · 정제하여 최종 그래프 구축에 활용한 관계는 총 30개이며, 이를 유사한 의미 군집으로 범주화한 최종 결과는 <Table 10>에 정리하였다. 이러한 범주화는 관계 체계를 단일 스키마로 강제하기보다는, 유사 관계를 묶어 해석 가능성을 높이는 방식으로 이해될 수 있다.

Table 9. Examples of Main Relations and Their Corresponding Expressions

Main Relation	Corresponding Relations
Form	form, form in, formed by, used to form, fill, etc.
Connect	adhere to, attach, bond to, connected to, contact, etc.
Process	process, performed, performed in, result in, etc.
Effect	effect, accommodate, contribute to, strengthen, etc.
Chemical	dissolve, react, polished by, heated by, etc.

Table 10. Summary of 30 Relation Categories Grouped Into 5 Semantic Clusters

Group	Relations
Structure/ Fabrication	form, cover, etch, apply, produce, modify, divide
Material/Component Manipulation	include, remove, supply, connect, position, protect, move, extend
Operation/Function	operate, use, activate, process, relate
Sensing/ Characterization	measure, detect, comparison, property, effect, increase, reduce, oppose
Interactions	chemical, emit

사례 분석의 첫 번째 대상으로, 실온에서 수행되는 웨이퍼 건조 공정을 다루는 특허 문서 (US11417511)를 분석하였으며, 그 결과는 <Figure 8>에 제시하였다. 하위 그래프에서는 'wafer drying'이 'cleaning step', 'reacting step', 'pressure releasing step'과 포함 관계로 연결되며, 공정 단계들이 상위 공정 개념 아래 구조적으로 조직되는 양상을 확인할 수 있었다. 또한 'reaction chamber'가 'cleaning solvent', 'supercritical fluid', 'pressure'와 연결되고, 'supercritical fluid'가 'critical temperature'와 연결되는 등 공정 조건과 물성 개념이 함께 구조화되는 관계가 나타났다. 이는 특허 텍스트 전반에 산재된 공정 요소들이 지식 그래프 상에서 결합되어 표현되는 양상을 보여주며, 공정 단계와 조건 및 물성 개념 간의 연결 구조가 어떻게 조직되는지를 드러낸다.

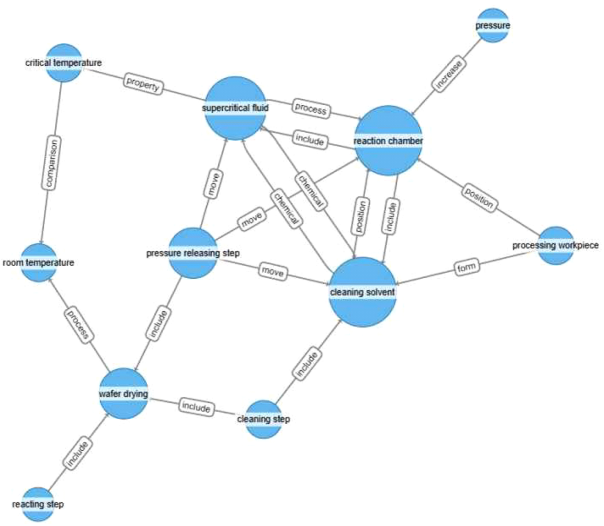


Figure 8. Case Study Subgraph Extracted from Publication No. US11417511 (room-temperature wafer drying process).

이어서 AlGaN 성장과 자외선 LED 응용을 다루는 특허 문서 (US10535801)를 분석하고, 그 결과를 <Figure 9>에 제시하였다. 하위 그래프에서는 excess Ga, 성장 방법, AlGaN 물질 관련 엔티티들이 연결되어 있고, 이후 물리적 과정 및 광 발생 관련 개념과 응용 분야로 이어지는 연결이 관측되었다. 이를 통해 특정 공정이나 재료 요소가 성능 및 응용 기술로 연결되는 서술 구조가 지식 그래프 상에서 하나의 연결된 구성 요소로 조직되는 양상을 확인할 수 있었다. 이러한 구조는 문서 내에서 함께 기술된 공정 조건, 재료 특성, 성능 요소들이 관계 형태로 통합 되어 표현되는 방식을 보여준다.

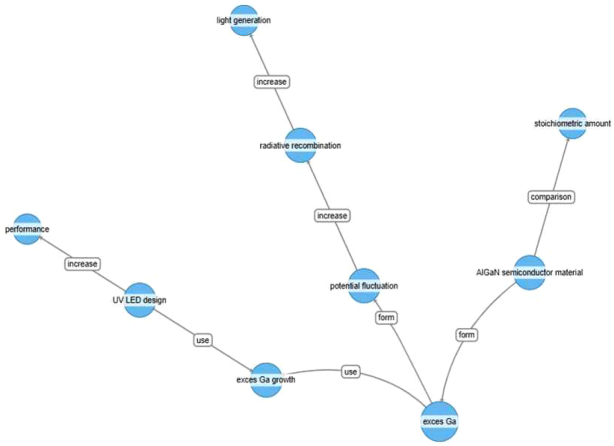


Figure 9. Case Study Subgraph Extracted from Publication No. US10535801 (AlGaN growth using excess Ga and UV-LED applications)

이상의 사례 분석은 제안한 하이브리드 구축 절차가 특허 기술 요소들을 하위 그래프 형태로 구조화하여 해석 가능성을 제공할 수 있음을 보여준다. 이러한 분석은 5.1절에서 보고한 정량적 성능 결과를 구조적 관점에서 보완하며, 지식 그래프가 기술 문서의 구조적 맥락을 어떻게 조직하는지를 구체적으로 보여준다.

6. 결 론

본 연구에서는 반도체 도메인의 레이블이 제공되지 않은 특허 데이터를 기반으로, 도메인 특화 정보 추출 기법과 대규모 언어 모델(LLM)을 결합한 하이브리드 지식 그래프 구축 방법론을 제안하였다. 제안한 접근은 사전 정의된 온톨로지나 대규모 레이블링 데이터에 의존하지 않고도, 도메인 중심의 엔티티 식별과 관계 추출을 단계적으로 수행함으로써 복합 기술 도메인에서의 지식 구조화를 가능하게 한다는 점에 의의가 있다.

이를 위해 도메인 특화 모델을 활용하여 엔티티 후보를 사전에 식별하고, 이를 LLM 추론의 제약 조건으로 활용함으로써 LLM 단독 방식에서 발생할 수 있는 도메인 비정합 엔티티 생성과 관계 표현의 불안정성을 완화하였다. 또한 관계 표현을 반복적으로 정제 및 병합하는 절차를 도입하여, 세부 주제별로 생성된 관계 표현을 통합하고 일관된 관계 사전을 자동으로 유도하였다. 이러한 절차는 반도체 특허의 CPC 분류 체

계를 기준으로 적용되어, 주제별 하위 그래프와 이를 통합한 도메인 특화 지식 그래프를 구축하는 방식으로 수행되었다.

평가 결과, 제안한 하이브리드 방식은 LLM-only 방식에 비해 원문에 대한 충실성이 높고, Domain-only 방식에 비해 관계 표현과 연결 구조가 보다 유연하게 형성되는 경향을 보였다. 또한 구조적 품질 분석에서는 중복성이나 비정상적 구조가 상대적으로 억제된 안정적인 삼중항 집합이 관찰되었다. 외재적 평가로 수행한 CPC 기반 문서 검색 실험에서는 하이브리드 방식이 상위 랭크 품질 측면에서 상대적으로 우수한 성능을 보였으며, 이는 문서 수준의 기술 주제 유사성을 표현하는 데 일정 수준 기여할 수 있음을 시사한다. 다만 이러한 결과는 CPC 분류 체계의 특성상 관계 의미의 정교함이나 경로 수준 추론 능력을 직접 반영하지는 않는다.

종합하면, 본 연구는 레이블이 없는 도메인 텍스트 환경에서도 도메인 중심의 지식 그래프를 자동으로 구축할 수 있음을 실험적으로 확인하였다. 제안한 하이브리드 접근은 관계 표현의 유연성과 구조적 정합성 간의 균형을 유지하면서, 반도체 특허와 같은 복합 기술 문서에서 기술 요소 간의 연결 구조를 비교적 안정적으로 조직할 수 있음을 보여준다. 이러한 결과는 LLM이 관계 추출의 전 과정을 대체하기보다, 관계 체계의 정제와 통합 단계에서 보조적이면서 구조화된 역할로 효과적으로 활용될 수 있음을 시사한다.

한편 반도체 특허라는 단일 도메인을 대상으로 실험을 수행하였다는 한계를 가지며, 향후에는 다른 기술 도메인에 대한 적용을 통해 방법론의 일반화 가능성을 추가로 검증할 필요가 있다. 또한 CPC 기반 외재적 평가는 문서 수준의 주제 유사성을 간접적으로 반영하는데 그치므로, 관계 의미의 정교함이나 경로 수준 추론 가능성을 직접 평가하기에는 한계가 있다. 더불어 관계 정제 및 통합 과정에서 LLM의 의미적 판단에 의존하는 특성상, 프롬프트 설계나 사용 모델에 따라 결과의 안정성이 달라질 수 있다는 점 역시 고려되어야 한다.

이에 따라 향후 연구에서는 지식 그래프 기반 질의응답이나 기술 요소 간 구조적 의존성 분석과 같은 보다 구조 중심의 하위 과제를 통해 그래프의 활용 가능성을 추가로 검증할 필요가 있다. 나아가 LLM 기반 평가가 내포할 수 있는 잠재적 편향을 완화하기 위한 평가 전략과, 시간에 따라 축적되는 신규 문서를 반영할 수 있는 지식 그래프 구축 방향 역시 향후 연구 과제로 남아 있다.

참고문헌

Abu-Salih, B. (2021), Domain-specific knowledge graphs: A survey, *Journal of Network and Computer Ap lications*, **185**, 103076.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... McGrew, B. (2023), Gpt-4 technical report, ArXiv, abs/2303.08774.

Andriopoulos, K. and Pouwelse, J. (2023), Augmenting LLMs with Knowledge: A survey on hallucination prevention, ArXiv, abs/2309.16459.

Bian, H. (2025), LLM-empowered knowledge graph construction: A survey, ArXiv, abs/2510.20345.

Bodenreider, O. (2004), The Unified Medical Language System (UMLS): Integrating biomedical terminology, *Nucleic Acids Research*, **32**, D267-D270.

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., and Yang, G. (2020), A deep learning based method for extracting semantic information from patent documents, *Scientometrics*, **125**(1), 289-312.

Chen, C., Shi, S. S., and Peng, S. L. (2024), A Construction of Knowledge Graph for Semiconductor Industry Chain Based on Lattice-LSTM and PCNN Models, *Journal of Internet Technology*, **25**(2), 313-329.

Choi, S. and Jung, Y. (2025), Knowledge Graph Construction: Extraction, Learning, and Evaluation, *Applied Sciences*, **15**(7), 3727.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., ... and Jain, A. (2024), Structured information extraction from scientific text with large language models, *Nature communications*, **15**(1), 1418.

Ding, L., Zhou, S., Xiao, J., and Han, J. (2024), Automated construction of theme-specific knowledge graphs, ArXiv, abs/2404.19146.

Friedrich, A., Adel, H., Tomazic, F., Hingerl, J., Benteau, R., Marusczyk, A., and Lange, L. (2020), The SOFC-exp corpus and neural approaches to information extraction in the materials science domain, In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 1255-1268.

Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. (2017), Creating training corpora for nlg micro-planning, In *55th Annual Meeting of the Association for Computational Linguistics*, ACL 2017, Association for Computational Linguistics (ACL), 179-188.

Grover, A. and Leskovec, J. (2016), node2vec: Scalable feature learning for networks, In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855-864.

Guo, Q., Jin, Z., Qiu, X., Zhang, W., Wipf, D., and Zhang, Z. (2020), CycleGT: Unsupervised Graph-to-Text and Text-to-Graph Generation via Cycle Training, In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, 77-88.

Hajikhani, A. and Cole, C. (2024), A critical review of large language models: Sensitivity, bias, and the path toward specialized ai, *Quantitative Science Studies*, **5**(3), 736-756.

Heist, N., Hertling, S., and Paulheim, H. (2023), KGrEaT: A framework to evaluate knowledge graphs via downstream tasks, In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3938-3942.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... and Zimmermann, A. (2021), Knowledge graphs, *ACM Computing Surveys (Csur)*, **54**(4), 1-37.

Hulth, A. (2003), Improved automatic keyword extraction given more linguistic knowledge, In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 216-223.

Jarvelin, K. and Kekalainen, J. (2002), Cumulated gain-based evaluation of IR techniques, *ACM Transactions on Information Systems (TOIS)*, **20**(4), 422-446.

Ji, S., Pan, S., Cambria, E., Marttinen, P., and Yu, P. S. (2021), A survey on knowledge graphs: Representation, acquisition, and applications, *IEEE Transactions on Neural Networks and Learning Systems*, **33**(2), 494-514.

Kaur, S., Smiley, C., Gupta, A., Sain, J., Wang, D., Siddagangappa, S., ... and Shah, S. (2023), REFinD: Relation extraction financial dataset, In *Proceedings of the 46th international ACM SIGIR Conference on*

Research and Development in Information Retrieval, 3054-3063.

Khorashadizadeh, H., Mihindukulasooriya, N., Tiwari, S., Groppe, J., and Groppe, S. (2023), Exploring In-Context Learning Capabilities of Foundation Models for Generating Knowledge Graphs from Text, In *TEXT2KG/BiKE@ ESWC*, 132-153.

Kim, H., Park, S., Lee, S., Cho, S., and Jeon, S. (2023), Named entity recognition in materials science using a bi-encoder and contrastive learning, *Proceedings of the Fall Conference of the Korean Institute of Industrial Engineers*, 2972-2983.

Kim, M., Seol, Y., Lee, S., and Yoon, J. (2025), Problem-Solution based Patent Analysis for Identifying Semiconductor Technology Trends, *Journal of Intellectual Property*, **20**(1).

Koren, Y., Bell, R., and Volinsky, C. (2009), Matrix factorization techniques for recommender systems, *Computer*, **42**(8), 30-37.

Kulkarni, M., Mahata, D., Arora, R., and Bhowmik, R. (2022), Learning rich representation of keyphrases from text, In *Findings of the Association for Computational Linguistics: NAACL 2022*, 891-906.

Lairgi, Y., Moncla, L., Cazabet, R., Benabdeslem, K., and Cleau, P. (2024), itext2kg: Incremental knowledge graphs construction using large language models, In *International Conference on Web Information Systems Engineering*, Singapore: Springer Nature Singapore, 214-229.

Lavrinovics, E., Biswas, R., Bjerva, J., and Hose, K. (2025), Knowledge graphs, large language models, and hallucinations: An nlp perspective, *Journal of Web Semantics*, **85**, 100844.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020), BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, **36**(4), 1234-1240.

Levenshtein, V. I. (1965), Binary codes capable of correcting deletions, insertions, and reversals. In *Doklady Akademii Nauk*, **163**(4), 875-848, Russian Academy of Sciences.

Li, J., Tang, T., Zhao, W. X., Wei, Z., Yuan, N. J., and Wen, J. R. (2021), Few-shot Knowledge Graph-to-Text Generation with Pretrained Language Models, In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1558-1568.

Lin, C. Y. (2004), Rouge: A package for automatic evaluation of summaries, In *Text summarization branches out*, 74-81.

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., ... Zhao, L. (2025), Domain specialization as the key to make large language models disruptive: A comprehensive survey, *ACM Computing Surveys*, **58**(3), 1-39.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023), G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment, In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511-2522.

Luan, Y., He, L., Ostendorf, M., and Hajishirzi, H. (2018), Multi-Task Identification of Entities, Relations, and Conference for Scientific Knowledge Graph Construction, In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3219-3232.

Manakul, P., Liusie, A., and Gales, M. (2023), Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 9004-9017.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020), On Faithfulness and Factuality in Abstractive Summarization, In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906-1919.

Navarro, G. (2001), A guided tour to approximate string matching, *ACM computing surveys (CSUR)*, **33**(1), 31-88.

Orlando, R., Cabot, P. L. H., Barba, E., and Navigli, R. (2024), ReLiK: Retrieve and LinK, Fast and Accurate Entity Linking and Relation Extraction on an Academic Budget, In *Findings of the Association for Computational Linguistics ACL 2024*, 14114-14132.

Pai, L., Gao, W., Dong, W., Ai, L., Gong, Z., Huang, S., ... and Zhang, Y. (2024), A survey on open information extraction from rule-based model to large language model, *Findings of the association for computational linguistics: EMNLP 2024*, 9586-9608.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002), Bleu: a method for automatic evaluation of machine translation, In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311-318.

Paulheim, H. (2017), Knowledge graph refinement: A survey of approaches and evaluation methods, *Semantic Web*, **8**(3), 489-508.

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., ... and Tang, S. (2024), Graph retrieval-augmented generation: A survey, *ACM Transactions on Information Systems*.

Peng, C., Xia, F., Naseriparsa, M., and Osborne, F. (2023), Knowledge graphs: Opportunities and challenges, *Artificial Intelligence Review*, **56**(11), 13071-13102.

Popovic, M. (2015), chrF: character n-gram F-score for automatic MT evaluation, In *Proceedings of the tenth workshop on statistical machine translation*, 392-395.

Reimers, N., and Gurevych, I. (2019), Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982-3992.

Ribeiro, L. F., Schmitt, M., Schütze, H., and Gurevych, I. (2021), Investigating pretrained language models for graph-to-text generation, In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, 211-227.

Sarica, S., Luo, J., and Wood, K. L. (2020), TechNet: Technology semantic network based on patent data, *Expert Systems with Applications*, **142**, 112995.

Schneider, P., Schopf, T., Vladika, J., Galkin, M., Simperl, E., and Matthes, F. (2022), A Decade of Knowledge Graphs in Natural Language Processing: A Survey, In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 601-614.

Schutze, H., Manning, C. D., and Raghavan, P. (2008), Introduction to information retrieval (Vol. 39, pp. 234-265), Cambridge: Cambridge University Press.

Seo, S., Cheon, H., Kim, H., and Hyun, D. (2022), Structural quality metrics to evaluate knowledge graph quality, Semantic Web.

Sharma, S., Nayak, T., Bose, A., Meena, A. K., Dasgupta, K., Ganguly, N., and Goyal, P. (2022), FinRED: A dataset for relation extraction in financial domain, In *Companion Proceedings of the Web Conference 2022*, 595-597.

Siddharth, L., Blessing, L. T., Wood, K. L., and Luo, J. (2022), Engineering knowledge graph from patent database, *Journal of Computing and Information Science in Engineering*, **22**(2), 021008.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., ... and Jain, A. (2019), Unsupervised word embeddings capture latent knowledge from materials science literature, *Nature*, **571**(7763), 95-98.

Venugopal, V., and Olivetti, E. (2024), MatKG: An autonomously

generated knowledge graph in Material Science, *Scientific Data*, **11**(1), 217.

Wang, B. and Guo, L. (2017), Knowledge graph embedding: A survey of approaches and applications, *IEEE Transactions on Knowledge and Data Engineering*, **29**(12), 2724-2743.

Xue, B. and Zou, L. (2022), Knowledge graph quality management: A comprehensive survey, *IEEE Transactions on Knowledge and Data Engineering*, **35**(5), 4969-4988.

Yamaguchi, K., Asahi, R., and Sasaki, Y. (2020), SC-CoMIcs: A superconductivity corpus for materials informatics, In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6753-6760.

Yang, R., Yang, B., Ouyang, S., She, T., Feng, A., Jiang, Y., ... and Li, I. (2024), Graphusion: Leveraging large language models for scientific knowledge graph fusion and construction in nlp education, ArXiv, abs/2407.10794.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016), Quality assessment for Linked Data: A Survey, *Semantic Web*, **7**(1), 63-93.

Zhang, Y., Zhong, V., Chen, D., Angeli, G., and Manning, C. D. (2017), Position-aware attention and supervised data improve slot filling, In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 35-45.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., and Lu, Z. (2019), BioWordVec, improving biomedical word embeddings with subword information and MeSH, *Scientific Data*, **6**(1), 52.

Zhang, B. and Soh, H. (2024), Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction, In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9820-9836.

Zhong, L., Wu, J., Li, Q., Peng, H., and Wu, X. (2023), A comprehensive survey on automatic knowledge graph construction, *ACM Computing Surveys*, **56**(4), 1-62.

Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., ... and Zhang, N. (2024), Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities, *World Wide Web*, **27**(5), 58.

Zou, X. (2020), A survey on application of knowledge graph, *In Journal of Physics: Conference Series*, **1487**(1), 012016, IOP Publishing.

<부록>

A. 프롬프트 템플릿

본 부록에서는 사용된 LLM 프롬프트 템플릿의 구체적인 설계와 예시를 제시한다. <Figure 10>~<Figure 12>는 제안하는 방법론에서 지식 그래프를 구축하기 위해 사용된 프롬프트들이다. 구체적으로, <Figure 10>은 도메인 모델을 활용해 추출된 엔티티를 정제하고 관계를 추출하기 위한 프롬프트를, <Figure 11>은 세부 주제 내에서 관계 표현들을 단계적으로 군집화하고 병합하기 위한 프롬프트를 보여준다. <Figure 12>는 서로 다른 세부 주제에서 생성된 관계 표현들을 통합하고 표준화하기 위한 단계별 관계 통합 프롬프트를 제시한다.

또한, 비교 실험을 위해 사용된 LLM-only 기준 방법의 통합 삼중항 추출 프롬프트는 <Figure 13>에 제시하며, <Figure 14>는 생성된 지식 그래프의 충실도를 평가하기 위해 LLM을 판별자로 활용한 평가 프롬프트를 보여준다.

```
You are an expert in information extraction and knowledge graph construction for the semiconductor industry. Given a patent abstract and a pre-extracted list of technology-centered entities, your task is to extract meaningful (subject, predicate, object) triplets that represent core technological relationships. Follow these instructions strictly:

1. Use only entities from the provided list as subjects or objects in the triplets. You may reword or introduce new entities only if you are 80% certain they refer to or improve clarity of items in the entity list.
3. Simplify entity names to concise and clear KG node labels by removing redundant descriptors and focusing on the core technical concept (e.g., “adhesive composition for semiconductor” → “semiconductor adhesive”).
4. Focus on technology-centered entities such as materials, components, operations, and properties.
5. Do not use vague or procedural phrases (e.g., “method for dicing …”) as entities. Convert them into technology-centered concepts or actions (e.g., “semiconductor wafer dicing”).
6. Extract only clear logical, physical or functional relations such as “include”, “form”, “contain”, “used in”, etc. Do not use ambiguous or purely spatial relations like “is”, “for”, “of”, “on”, “by”, etc.

Output only in the format (subject, predicate, object). Use parentheses, separate elements with commas, and include no extra text.

Text: {text}
Entity List: {entity_list}
```

Figure 10. Prompt for Entity Refinement and Relation Extraction

```
You are an expert in the semiconductor domain. Given a list of relations extracted from the abstract of a patent document in the semiconductor field, your task is to systematically analyze and integrate these relations. Ensure technical accuracy, coherence, and relevance to the semiconductor industry.

Follow these steps:

1. Clustering: Group the relations based on semantic and contextual similarity, especially their functional roles in semiconductor descriptions.
2. Merging: Within each group, consolidate semantically similar or equivalent relations. Define a representative main relation that best describes each group.
3. Dictionary Formatting: Return a dictionary where:
  - Each key is the selected representative relation (e.g., “form”).
  - Each value is a list of merged or semantically equivalent relations that map to that representative.

Output Format: Return only the final relation dictionary, with no additional commentary or explanation.

Input Relations:
{relations}
```

Figure 11. Prompt for step-by-step Clustering and Merging Relations Within Each Subtopic

```
You are an expert in the semiconductor domain. The following list of relations consists of representative relations refined from multiple subtopics.
Although these relations were refined separately, semantically equivalent or highly similar relations may still exist across subtopics.

Your task is to systematically analyze and integrate these relations using a structured Chain-of-Thought approach:

(1) Alignment and Clustering:
    Identify relations that express the same or very similar meanings across subtopics and group them based on semantic and contextual similarity.

(2) Merging and Canonicalization:
    Merge similar relations within each group and define a single representative (canonical) relation that can be consistently applied across the entire document set.

Ensure that the integration resolves naming inconsistencies across subtopics while preserving semantic distinctions when necessary.
Maintain technical accuracy and coherence within the semiconductor domain.

[Relations]
{relations}
```

Figure 12. Prompt for step-by-step integration and canonicalization of relations across subtopics

```
You are an information extraction system for semiconductor patent text.

Task:
1) Extract a list of domain-relevant entities as short noun phrases (e.g., "reacting step", "room temperature", "wafer").
2) Extract relationships as (subject, relation, object) triplets.
   - subject/object MUST be chosen from the extracted entities list (string match).
   - relation should be a short verb or verb+preposition phrase (e.g., "includes", "performed_at", "uses", "forms").

Rules:
- Output MUST be valid JSON that matches the provided schema.
- Do not invent entities not present in the text.
- Prefer meaningful relations that are explicitly supported by the sentence.
- If the sentence enumerates steps, also add "precedes" relations following the textual order.

Text:
"{sentence}"
```

Figure 13. Prompt for end-to-end Triplet Extraction Used in the LLM-only Baseline Method

```
You are evaluating a knowledge graph extraction from a patent sentence.

Sentence:
"{sentence}"

Triplet:
Subject: {subj}
Relation: {rel}
Object: {obj}

Question:
Is this triplet explicitly supported by the sentence?

Answer with only one of:
SUPPORTED
NOT SUPPORTED
UNCLEAR
```

Figure 14. Prompt for LLM-as-a-judge faithfulness evaluation.

저자소개

이수연: 서울대학교 경영학과에서 2017년 학사, 서울대학교 산업공학과에서 2023년 석사학위를 취득하고 산업공학과 박사과정에 재학 중이다. 연구분야는 Natural Language Processing, Dialogue System, Large Language Models이다.

박소형: 서울대학교 산업공학과에서 2021년 학사학위를 취득하고 서울대학교 산업공학과 박사과정에 재학 중이다. 연구분야는 데이터마이닝, 인공지능, 자연어처리이다.

김현종: 포항공과대학교 산업경영공학과에서 2021년 학사학위

를 취득하고 서울대학교 산업공학과 박사과정에 재학 중이다. 연구분야는 Natural Language Processing, Large Language Models, Sentiment Analysis이다.

조성준: 서울대학교 산업공학과에서 학사, 석사학위를 취득하고 미국 워싱턴대학교에서 컴퓨터 사이언스학과에서 인공지능 석사학위 및 메릴랜드대학교 컴퓨터사이언스 학과에서 뉴럴네트워크, 머신러닝 분야로 박사학위를 취득하였다. 이후 공공데이터전략위원장, 정부3.0추진위원회 빅데이터전문위원장과 한국BI데 이터마이닝 회장 등을 역임했다. 현재 서울대학교 산업공학과 교수 및 빅데이터 AI 센터장으로 재직하고 있다. 연구분야는 딥러닝, 텍스트마이닝 등 빅데이터 및 AI, 산업 응용이다.