# TADA: 과학기술 분야의 텍스트 분석 지원을 위한 데이터와 인프라

김성찬[1,2], 이승우[1,2], 최명석[1]

[1]한국과학기술정보연구원 기계학습데이터연구단, [2]과학기술연합대학원대학교 응용 AI

sckim@kisti.re.kr, swlee@kisti.re.kr, mschoi@kisti.re.kr

# An Introduction to TADA: Text Analysis with S&T Data/InfrAstructure

Seongchan Kim[1,2], Seungwoo Lee[1], Myung-seok Choi[1]

[1]Dept. of Machine Learning Data Research, Korea Institute of Science and Technology Information

[2]Applied AI, UST-KISTI

요 약

본 논문은 과학기술 분야의 텍스트 분석 지원을 위한 한국과학기술정보연구원 기계학습 데이터 공유·활용 서비스(AIDA)에서 제공하는 데이터와 인프라를 소개한다. 현재 과학기술 문헌분석을 위해 개발하고 있는 기술 및 공개된 데이터셋과 사전학습 언어모델, OpenAPI를 소개한다.

## Ⅰ. 서론

본 논문은 한국과학기술정보연구원(KISTI) 기계학습 데이터 공유·활용 서비스(AI Data Archive, AIDA)[1]가 제공하는 과학기술(S&T) 분야의 텍스트 분석 지원을 위한 데이터와 인프라를 소개한다. TADA(Text Analysis with S&T Data/Infrastructure)는 텍스트 데이터 세트 및 인프라 서비스를 위한 AIDA의 한 컴포넌트이다. TADA의 목적은 과학기술 문헌으로 데이터 기반의 AI와 기계학습 연구를 수행하는 연구원 및 개발자들의 연구 및 분석 프로세스를 효율적으로 지원하는 것이다. TADA는 데이터/AI 기반 문제해결 지원 시스템 상에서 R&D 혁신 및 AI 기술 개발을 지원하는데 그 목적이 있다.

## Ⅱ. 과학기술 문헌분석 기술개발 및 지원

그림 1은 KISTI 기계학습데이터 공유·활용 서비스가 목표로 하는 과학기술 문헌분석을 위한 지원기술에 관한 개요이다. 논문의 용어, 문장, 단락, 문서 4가지의 단위로 레벨을 나누고 이에 관련한 각 지원 기술을 개발한다. 먼저 용어(단어)레벨에서는 과학기술분야의 전문용어가 도메인내에서 어떠한 상세개체(Fine-grained Entity)를 갖게 되는지 인식하는 기술[2]을 개발하여 관련 데이터셋, 모델, OpenAPI등을 제공한다. 문장 레벨에서는 문장 단위의 의미(역할) 분류 딥러닝 기술을 개발하여 데이터셋, 모델, OpenAPI 등을 제공한다. 단락 모델에서도 각 단락이 어떤 역할을 하는지, 문서 레벨에서는 논문의 연구주제가 과학기술 표준분류 체계에서


그림 1 과학기술 논문 풀텍스트 분석(TADA) 지원을 위한 개요도

| 구분 | 데이터셋 명칭 | 건수 | 비고 |
|---|---|---|---|
| 데이터셋 | 국내 논문 텍스트 데이터셋 | 481,578건 | |
| | 국내 논문 QA 데이터셋 | 279,143건 | |
| | 국내 문장 의미 태깅 데이터 | 155,740건 | |
| | 보고서 테이블/그림 데이터셋 | 3,546,095건 | |
| | 식별된 조직 데이터 | 245,692건 | |
| | 메타데이터 추출용 코퍼스 | 3,815,987건 | |
| 사전학습 언어모델 | KorSciBERT | | 국내논문 및 특허 97GB 학습 |
| | KorSciELECTRA | | 국내논문, NTIS 연구과제, 특허, 뉴스, 한국어 위키 코퍼스 141GB 학습 |
| Open API (제공예정) | 문장의미 태깅 API | | 문장의미 태깅 카테고리 (‘문제정의’, ‘가설설정’, ‘이론/모형’,등 ) |
| | 상세개체 인식 API | | 보건의료분야 상세개체 유형 태그( Body Regions [A01], Musculoskeletal System [A02] 등) |
| | 연구주제 분류 API | | 과학기술표준분류체계의 중분류 카테고리 중 Top 3 (대수학NA01, 해석학 NA02, 위상수학 NA03 등) |

표 1 AIDA(https://aida.kisti.re.kr)에서 제공하는 과학기술 텍스트 자료(2022.10 기준)

어떠한 연구분야에 속하는지 분류하는 모델을 개발하여 관련 리소스를 제공한다. 과학기술 분야의 텍스트 자원을 가지고 개발된 사전학습 언어모델(KorSciBERT, KorSciELECTRA) 등을 기반으로 전이학습 하는 방식으로 각 모델들은 개발된다[3]. 이러한 모델들은 S&T 논문 전문 기반의 요약 및 군집 모델 개발에 활용되며 요약 및 군집모델은 ScienceON[4]과 같은 논문조사분석 시스템에 요약서비스, 이슈 트렌드 분석 서비스 등에 활용되는 것을 최종 목표로 한다.

AIDA(https://aida.kisti.re.kr)를 통하여 일반인에게 공개되어 모든 사람들이 사용할 수 있는 과학기술 텍스트 자원은 7개의 데이터셋과 2개의 사전 학습 언어 모델이며 표 1에 상세하게 기술되어 있다. 추가로 구축 및 확보된 데이터는 지속적으로 공개될 예정이다.

## Ⅲ. 결론

본 논문에서는 한국과학기술정보연구원이 제공하는 과학기술 문헌 분석을 위한 텍스트 및 인프라(TADA)를 소개하였다. 제공되는 텍스트 리소스와 인프라를 활용한다면 클러스터링, 요약, QA 등 과 같은 과학기술 텍스트 분석 모델의 다양한 애플리케이션의 성능 향상을 추구할 수 있다. 또한 과학기술 논문 및 문서에 대한 다양한 전문 분석 모델 API의 지원으로 관련 연구자 및 개발자의 AI 기술 개발 효율성이 향상될 것이다.

## ACKNOWLEDGMENT

이 연구는 2022년도 한국과학기술정보연구원의 기본사업 “Data/AI 기반 문제해결체계 구축”(K-22-L04-C05-S01)의 지원으로 수행되었습니다.

## 참 고 문 헌

[1] KISTI 기계학습 데이터 공유·활용 서비스, 한국과학기술정보연구원 (https//aida.kisti.re.kr).

[2] Xuan et. al., ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-Guided Distant Supervision. EMNLP 2021.

[3] Beltagy et al., “SciBERT: A Pretrained Language Model for Scientific Text,” EMNLP 2019.

[4] KISTI 과학기술 지식 인프라, 한국과학기술정보연구원 (https://scienceon.kisti.re.kr)