정부혁신 보다나은 정부

보 도 자 료

한국과학기술정보연구원 Korea Institute of Science and Technology Information

배포 즉시 보도 가능합니다.

대전(본원): 대외협력실 이종성 042-869-0976 / 이해준 0676 / 손영주 0997
문의: 콘텐츠큐레이션센터 황혜경 센터장(042-869-1785)

배포번호 : 2020-75 배포일자 : 2020.10.26.(월)	매수 : 보도자료 6매 (첨부자료 포함)	배포처 : 대외협력실

기계학습 데이터 현황과 이슈

- 과학기술 분야를 중심으로 -

한국과학기술정보연구원(원장 최희윤, 이하 KISTI)은 코로나19 사태로 인한 극심한 경기침체의 극복과 경제의 구조적 대전환을 위해 추진되는 정부의 핵심 사업인 데이터댐 구축에 과학기술분야 핵심정보를 기계학습용 데이터로 구축하여 외연과 깊이를 확장해 가고 있다. 『KISTI 이슈브리프 제26호』를 통해 디지털 뉴딜·데이터댐의 핵심인 인공지능(AI) 개념과 국내외 기계학습 데이터 현황을 살펴보고 KISTI에서 추진하고 있는 과학기술 기계학습 데이터 구축 및 활용방안을 소개한다.

* KISTI 이슈브리프 : KISTI는 국가과학기술정보 분야 대표 연구기관으로서, 최근의 국가·사회 이슈에 대해 폭넓은 조사와 정보/데이터 기반 분석 기법을 통해 문제 해결을 위한 지식과 시사점, 대응 방안을 제공하고자 "KISTI 이슈브리프"를 발간함 (https://www.kisti.re.kr/promote/post/issuebrief)

□ 인공지능(Artificial Intelligence, AI)과 기계학습

○ **(기계학습)** 기계학습은 AI의 특정 연구 분야로 "명시적으로 프로그램 되지 않고 학습할 수 있는 능력을 컴퓨터에게 주는 AI 연구 분야"로 정의함.

○ **(기계학습 데이터)** 특정목적(문서분류, 문서요약, 영상이해, AI 비서 등)을 위하여 기계가 학습할 수 있도록 용도에 맞게 잘 정제된 데이터를 뜻하며 AI 성공은 지속적이며 신뢰성 있는 양질의 학습 데이터 확보가 필수적임.

○ **(기계학습 데이터 가공의 필요성)** 기계학습에서 데이터는 자동차의 연료와 같은 역할을 하며, 높은 수준의 기계학습을 구현하기 위해서는 많은 양의 양질의 데이터가 반드시 필요함.

○ **(활용 분야)** AI 기계학습은 방대한 학습 데이터로부터 패턴을 감지하고 시스템을 모델링해 의사결정의 효율성 향상, 비용절감 및 효율적인 자원 할당 등 전 산업분야에서 활용함.

□ 국내·외 기계학습 데이터 구축 현황

○ **(해외 기계학습 데이터 현황)** 이미지, 텍스트, 음성, 영상 등 다양한 분야에 대해 기계학습 데이터를 구축·공개하고 있음. 공개된 데이터는 전 세계에서 AI 기술 성능평가기준 데이터로 활용되고 있으며 AI 연구 촉진과 산업 발전에 기여하고 있음.

○ **(국내 기계학습 데이터 현황)** 주로 해외에서 공개된 기계학습 데이터를 활용하여 AI 기술개발 및 연구를 수행하고 있음.

-**(국내 특화 데이터 구축)** 독자적인 AI 활용을 위하여 한국형 이미지, 한국어 텍스트, 한국어 음성, 한국형 영상 데이터가 필요하기 때문에 우리나라에 특화된 기계학습 데이터를 구축하여 공개하고 있음.

□ 과학기술 기계학습 데이터 구축사업 추진 배경

○ **(한국판 뉴딜 종합 계획)** 정부는 코로나19 사태로 인해 유발된 경기침체 극복 및 구조적 대전환 대응을 위하여 『한국판 뉴딜 종합계획』을 발표함.

-디지털·그린 뉴딜을 강력 추진하고 사회적 합의를 바탕으로 고용·사회 안전망을 강화함.

○ **(기계 가독형 데이터의 개방)** 데이터 기반 연구 활동이 보편화되면서 기계 가독형 데이터가 중요해지고 있으며 오픈 데이터로 개방되고 있음.

○ **(과학기술정보의 지속 증가)** 매년 과학기술정보의 양이 급속하게 증가함에 따라 선행문헌조사를 도울 수 있는 연구지원 도구가 필요함.

○ **(의미기반 정보서비스 부재)** 연구자들이 최신 연구 동향을 손쉽게 파악하고 따라가기 위해서는 학술 결과물에 대한 빠른 파악이 필수적이며, 이를 위해서는 의미기반 정보서비스 제공이 선행되어야 함.

○**(과학기술분야 기계학습 데이터의 부족)** AI기반 R&D혁신을 위해서는 많은 양의 과학기술분야 학습데이터가 필요함에도 불구하고, 우리나라는 공공·일반 분야 기계학습 데이터 중심으로 구축하고 있어 과학기술분야 기계학습 데이터가 충분하지 못함.

○**(기계학습 데이터 확보 역량부족)** 국내 산·학·연 연구자들은 과학기술분야 연구를 AI 기반으로 수행하고자 하여도 연구에 필요한 학습데이터를 자체적 확보할 수 있는 역량이 부족하고 구축에 따른 많은 시간과 비용이 소요됨.

□ 과학기술 기계학습 데이터 구축사업 추진 내용

○**(사업 개요)** KISTI는 과학기술분야 기계학습 데이터를 대규모로 구축하여 AI 기반의 기술혁신으로 데이터 경제 견인에 이바지함.

-코로나19 사태 발생 이후, 일자리 축소 등 열악한 경제 상황에 대응하여 대규모 공공인프라 사업 추진을 통한 비대면 일자리를 창출함(목표 고용인력수: 2,000명).

○과학기술분야 국내논문, 국가 R&D 연구보고서 대상으로 하는 다수의 작업공정(직접입력, 레이블링, 태깅, 품질검수 등)으로 AI 학습 데이터 구축 활용확산 등을 추진함.

-과학기술분야 국내논문, 국가 R&D연구보고서 대상으로 AI 학습 데이터 구축에 필요한 교육·훈련, 기계학습 데이터 구축 및 운영 시스템 등을 개발함.

○**(과학기술 기계학습 데이터 구축)** 국내논문, 국가 R&D 연구보고서를 대상으로 과학기술분야 AI 학습데이터를 구축함.

-**(데이터선정)** AI 산업계 수요조사결과를 반영하여 AI 학습용 데이터 425.7만 건을 선정함

-**(품질관리)** 기계학습 데이터 구축 과정에 크라우드소싱(Crowdsourcing)[1] 방법을 적용하되, 다단계 검수를 통해 품질을 검증함.

□ 과학기술 기계학습 데이터 활용 및 기대효과

○**(연구분야)** 과학기술 분야 지식자원의 AI 연계 및 융합 연구 지원

-사례 분석, 연구방법론, 최근 연구 트렌드 분석 등 선행연구조사를 통하여 소요시간 단축이 가능함.

1) 크라우드소싱: 대중(Crowd)과 아웃소싱(Outsourcing)이라는 두 단어의 합성어로, 일반 대중에게 참여를 유도하여 상품 및 서비스의 개발 과정에 지식 및 의견을 반영하여 결과물을 이끌어 내는 것임.

-기계학습을 위한 원천 데이터의 원활한 활용으로 AI 기술성능 향상이 가능함.

-연구자가 관련 지식이 없는 학문분야를 초월한 새로운 지식 창출이 가능함.

○**(정책분야)** 국가 R&D 정책 수립 시 의사결정 지원

-R&D투자 효율성과 연구생산성 측면에서 기 수행된 연구주제를 AI 기반으로 분석·활용하여 新 성장 연구분야 및 공백기술 연구분야에 대한 국가 R&D 정책 수립이 가능함.

-과학기술분야 기계학습 데이터 활용 확산 관련 국가적 관리 체계를 마련함.

○**(산업분야)** 산업 기술 혁신을 위한 의사결정 및 비즈니스 활용

-AI기반 기업애로기술 솔루션 탐색, AI기반 유망기술분석, 데이터 전문기업, 비대면 전문교육 콘텐츠 제작 및 AI 과학교사 운영 등 데이터분야, 교육분야 비즈니스에 활용 가능함.

-선행기술조사 기간 및 범위를 축소시켜 기업의 新 사업 아이템 발굴 등에 소요되는 시간·비용을 절감함.

-데이터 가공·추출에 소요되는 시간·비용을 감소시켜 시뮬레이션 기반 기술 개발을 지원함.

□ 디지털 뉴딜의 성공조건

○**(데이터 표준화)** 디지털 뉴딜 정책 추진 시, 데이터를 수집·가공하고 다른 데이터화 결합하는 과정에서 표준화 정도가 성공의 중요 잣대임

-다양한 데이터 형태와 포맷에 대한 처리 방안에 과학기술 데이터의 표준화 방안을 확대 적용할 수 있음.

○**(기업 데이터와의 연계)** AI 기반으로 제공되는 과학기술 데이터를 활용하여 제품·서비스 경쟁력을 높이기 위해서는 기업의 데이터를 데이터댐과 연계·공유할 수 있는 파이프라인이 필요함.

-기업 스스로 데이터 분석 기반으로 제품·서비스에 대한 평가를 수행할 뿐만 아니라 의사결정에서도 데이터 기반으로 하는 문화가 선행되어야 함.

○**(데이터 활용 프로세스)** 기 구축된 데이터를 이용하여 새로운 제품서비스를 할 수 있게 하는 일련의 사업 프로세스 구축이 필요함.

-주요 산업 분야에 대한 사업전체과정에서 데이터 활용에 대한 매뉴얼 제

작이나 사업 프로세스별 전문가의 지원 체계 마련 등으로 해결 가능함.

○ **(데이터댐 활용 인프라 확충)** 막대한 예산으로 구축한 데이터댐을 효율적이고 효과적으로 활용하기 위해서는 클라우드 컴퓨팅과 AI 등의 분야에 대한 기술·인력이 필요함.

-우리나라는 미국, 중국 등과 비교하여 이 분야 경쟁력이 저조한 상황이기 때문에 정부의 지속적인 관심과 투자가 요구됨.

○ **(디지털 격차 해소)** 코로나19로 촉발한 비대면 중심의 디지털 대전환 속에서 다양한 계층 간 디지털 격차 해소 방안 마련이 필요함.

-디지털 역량 강화에 대한 교육훈련 사업 강화를 추진함.

□ KISTI 최희윤 원장은 "코로나19의 위기를 계기로 데이터의 중요성이 크게 부각되었다. 집단지성을 통해 대규모로 구축된 과학기술분야 기계학습 데이터는 과학기술연구 전주기를 지원하는 AI서비스에 활용되어, 학제 간 융합연구의 촉매제가 되고, 기술 혁신과 새로운 비즈니스 창출에 기여할 것" 이라고 밝혔다.

첨부 KISTI 이슈브리프 제26호 설명 이미지 자료