해외사례를 중심으로

데이터 라벨링으로 만드는 혁신

NIA 지능정보원
한국지능정보사회진흥원

해외사례를 중심으로 데이터 라벨링으로 만드는 혁신

글 | **이지현** IT전문기자 j.lee.reporter@gmail.com / **우창완** 책임연구원 woo@nia.or.kr

최근 몇 년간 데이터 라벨링 기업들이 크게 성장하고 있다. 스케일AI[1], 라벨박스[2], 하이브[3], 클라우드팩토리[4]같은 데이터 라벨링 기업들은 꾸준히 투자를 유치하고 있으며,[5] 그 시장은 점점 커지는 추세다. 한 조사기관은 2020년 기준 데이터 라벨링 시장 가치가 13억 달러 규모를 넘어섰고 2028년까지 연간 25%씩 성장할 거라는 분석을 내놓았다.[6] 이렇게 데이터 라벨링 산업이 성장하는 데는 인공지능 기술의 역할이 컸다. 인공지능 프로젝트 작업 중 25%[7]가 데이터 라벨링 작업이라고 하니, 인공지능 연구가 많아질수록 데이터 라벨링 시장에 고객이 몰리는 셈이다.

1 https://scale.com/

2 https://labelbox.com/

3 https://thehive.ai/about-us

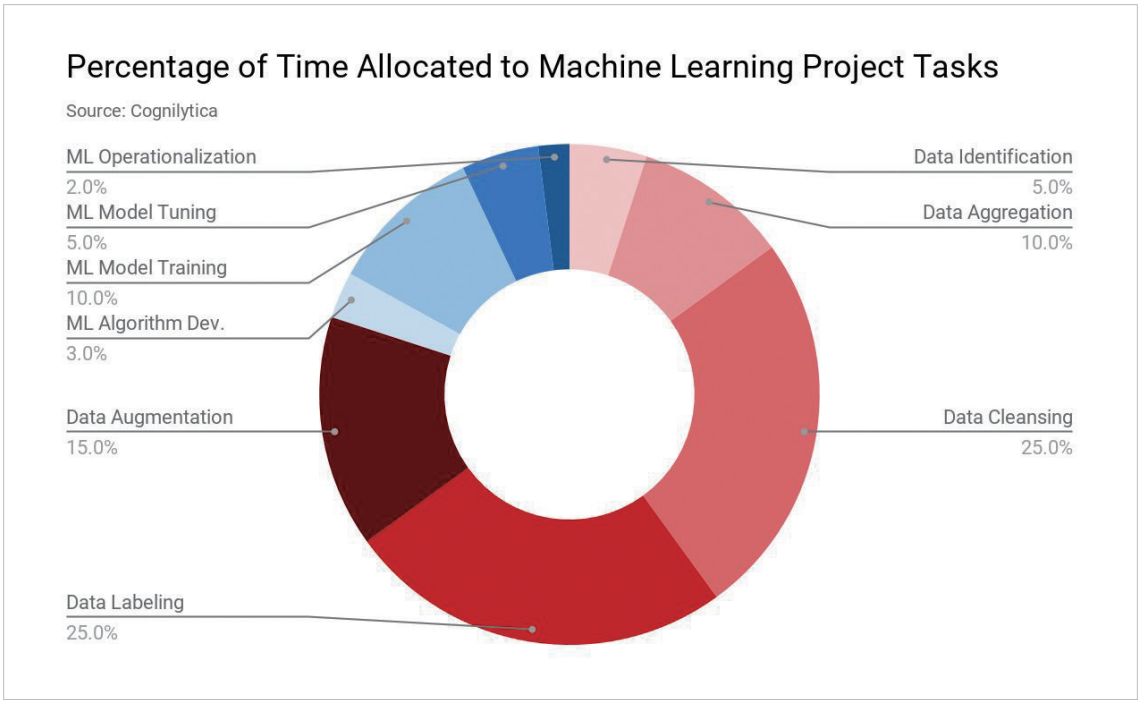
4 https://www.cloudfactory.com/

5 스케일AI는 1억5,500만 달러, 라벨박스는7,900만 달러, 하이브는 1,670만 달러, 클라우드팩토리는 7,800만 달러 규모의 투자를 유치했다.

6 https://www.grandviewresearch.com/industry-analysis/data-collection-labeling-market

7 https://www.forbes.com/sites/cognitiveworld/2020/02/02/the-human-powered-companies-that-make-ai-work/

[머신러닝 프로젝트에서 필요한 작업 종류와 소요 시간 비율]



www.forbes.com/sites/cognitiveworld/2020/02/02/the-human-powered-companies-that-make-ai-work/

데이터 라벨링 작업이란 인공지능 혹은 기계가 해당 데이터가 무엇인지 이해할 수 있도록 추가 정보를 달아주는 일이다. 사람은 고양이와 호랑이 사진을 보고 구분할 수 있지만 기계는 이를 보고 헷갈려한다. 그래서 고양이인지 호랑이인지 이미지에 이름표를 달아주는 라벨링 작업이 필요하며, 라벨링 할 대상은 이미지 외에도 오디오, 텍스트, 음성, 영상 등 다양하다. 이때 라벨링 작업을 사람이 직접 수동으로 할 수도 있고, 프로그램을 통해 자동화할 수 있다.

글로벌 마켓 인사이트가 펴낸 보고서에 따르면 2019년 기준 전체 데이터 라벨링에서 75%가 사람이 직접 참여하는 구조였고 2026년까지 자동 라벨링보단 사람이 개입한 라벨링 작업이 더 많은 거라고 예상했다.[8] 사람이 직접 데이터 라벨링 하는 경우는 다시 두 가지로 나뉜다.

8 https://www.gminsights.com/industry-analysis/data-annotation-tools-market

내부 직원이 참여하는 경우와 외부 사람에게 아웃소싱하는 경우로, 후자를 '크라우드소싱(Crowd sourcing)' 방식이라 부른다. 이런 데이터 라벨링 작업은 어찌 보면 지루한 단순 작업으로 보일 수 있으나 실제론 그 작은 작업이 쌓이고 쌓여 새로운 혁신의 중요한 밑거름이 된다. 아래에서 살펴볼 프로젝트들이 바로 그 예이다.

의학 분야와 집단 지성과의 만남

의학 분야에는 진료기록이나 엑스레이 사진 등으로 구성된 다양하고 방대한 데이터가 쌓여 있다. 최근 헬스케어 산업의 성장으로 의료 데이터를 활용하는 사례가 늘고 있는 가운데, 연구 과정에서 필요한 데이터 정제 작업을 외부에 맡기려는 시도도 함께 많아지고 있다. 누군가는 전문성이 필요한 의학 분야에 일반인이 나서 데이터를 정제해도 되는지 의문이 들 수 있다. 아래 프로젝트는 그런 통념과 반대로 일반인에게 데이터 라벨링 작업을 의도적으로 맡긴 사례다. 대부분 데이터 량이 너무 많아서 내부 인력으로는 이를 감당하지 못하거나 의료인들이 가진 사전 지식을 데이터 품질에 방해가 돼서 크라우드소싱 방식을 선택하고 있는 것을 볼 수 있다.

자폐아 연구를 위한 모바일 앱 '게스왓'

주요내용

스탠퍼드 바이오인포매틱스부 연구진은 자폐 아동들의 증상을 진단하고 치료를 돕는 '게스왓(Guess What)'[9]을 2018년 개발했다.[10] 게스왓은 기본적으로 자폐증을 가진 아이들을 위한 모바일 게임 앱이나 핵심은 자폐증 아이들의 얼굴 데이터 수집에 있다. 기존 얼굴 인식 기술이 비장애인 얼굴 사진을 중심으로 개발됐다는 점을 주목해 자폐증 연구를 위한 얼굴 이미지 데이터를 별도로 수집하고 있는 것이다.

게임 내용은 자폐증 아이와 부모가 퀴즈를 내고 정답을 맞히는 구조다. 모바일 화면에는 특정 감정이나 단어를 묘사하는 사진과 자막이 나오고 아이는 그에 맞는 표정을 지으면 된다. 부모가 보기에 아이가 제대로 된 표정을 지었다고 생각이 들면 휴대폰을 아래로 기울이면

9 https://guesswhat.stanford.edu/

10 https://news.stanford.edu/2019/03/06/smartphone-app-treat-track-autism/

된다. 앱은 이를 인식해 정답처리하며 다시 휴대폰을 위로 젖히면 다음 문제로 넘어간다. 사전에 녹화 영상을 공유하기로 동의했다면 퀴즈에 참여하는 자폐아의 얼굴은 비디오 파일 형식으로 저장되고 스탠포드 연구진에게 전송된다. 그렇게 얻은 비디오 데이터는 다시 프레임별로 나눠 이미지로 저장된다. 이미지가 저장될 때는 퀴즈 사진과 이에 대한 자막 내용이 라벨링 정보로 자동 추가된다. 연구진은 그렇게 수집한 데이터를 머신러닝 분석에 활용해 궁극적으로 자폐아가 짓는 표정, 행동 움직임의 특성을 알아내고 있다. 게스왓은 향후 자폐증 진단 도구나 감정 표현 및 교류 연습을 할 수 있는 치료 앱으로 발전시킬 예정이라고 한다.

[게스왓 게임 예시]

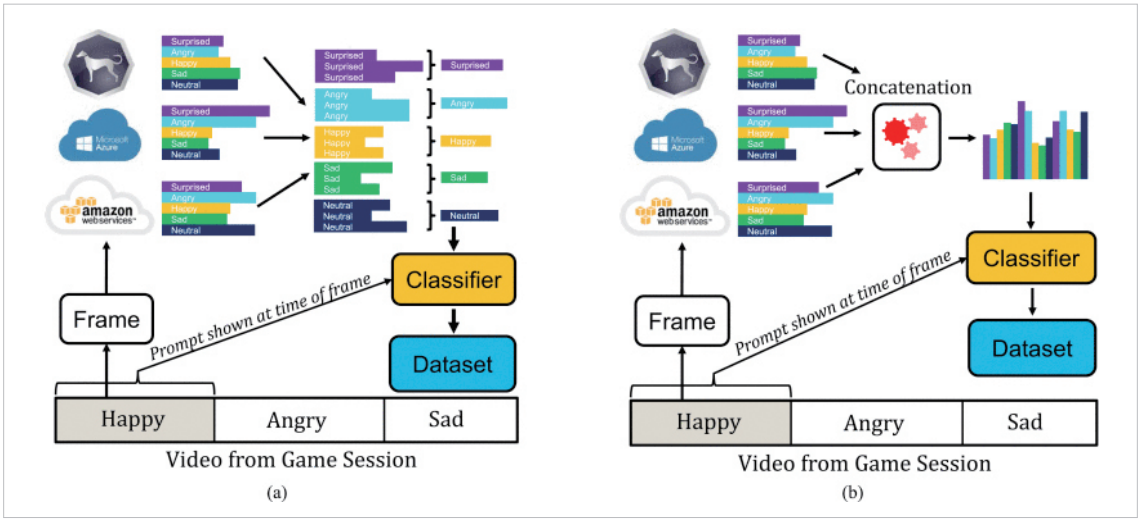
guesswhat.stanford.edu

라벨링 참여방법

스탠퍼드 연구진은 일종의 정답 데이터(ground truth)를 만들기 위해 진행했던 라벨링 작업을 논문에 공개하기도 했다. 해당 라벨링 작업에는 자폐아 관련 치료나 연구 경험이 없는 비전문가를 투입했다. 라벨링 데이터를 자폐아 연구 전문가만 이해할 수 있는 데이터가 아닌, 비전문가도 이해할 수 있는 수준으로 만들려고 했기 때문이다. 이미 같은 연구실에서

진행한 연구에 따르면 임상 지식이 없는 평가자가 라벨링을 했을 때 데이터 분석 신뢰도가 더 좋다는 결과를 얻은 적이 있었다고 한다.[11] 최종적으로 연구진은 사진 1,350개에 감정 표현 정보를 추가해 라벨링 작업을 마쳤다고 한다.

[게스왓 데이터 라벨링 과정]



ieeexplore.ieee.org/abstract/document/8501577

11 https://www.sciencedirect.com/science/article/pii/S0933365718302598 5.2. Data processing 참고

트위터로 보건 동향을 알아내는 '크라우드 브레이크'

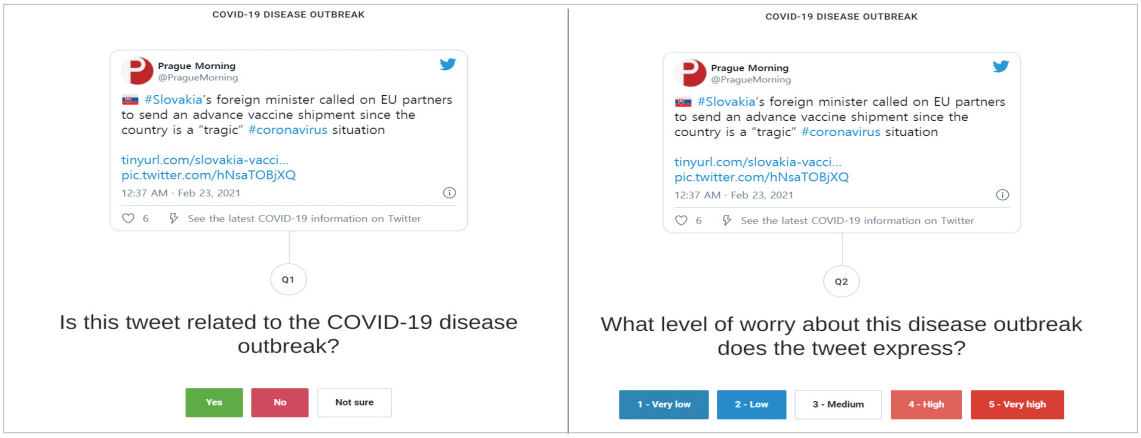
주요내용

소셜 미디어 데이터는 사람들의 생각과 생활모습을 알아내기 좋은 도구다. 이미 많은 데이터 분석 기관들이 소셜 미디어 자료로 문화나 여론을 분석하고 있으며, 의학계에서도 비슷한 실험을 진행 중이다. 스위스 연방 과학대학(EPFL) 디지털 전염병 연구실에서는 트위터 데이터로 현재 전 세계에서 일어나는 보건 및 질병 동향을 파악하고 있다. 연구진은 기존 의학계에서 활용하는 소셜미디어 데이터가 정제하고 실시간 형태로 수집하기 어렵다는 점을 주목해 '크라우드 브레이크'[12]를 기획했다.

라벨링 참여방법

크라우드 브레이크는 트위터 API를 이용해 특정 단어를 포함한 트윗을 불러오고, 해당 트윗의 내용을 라벨링 해준다. 라벨링 과정은 18세 이상이라면 전 세계 누구나 무료로 참여할 수 있다. 스위스 연방과학대학은 현재 두 가지 주제로 나눠 데이터를 모으고 있으며, 하나는 '백신에 대한 사람들의 감정'이며 다른 하나는 '코로나19 사태에 대한 사람들의 감정'과 관련이 있다.

[크라우드브레이크 데이터 라벨링 과정. 아래 버튼 클릭해 사용자가 데이터 라벨링을 할 수 있다]



* 출처 : 공식 홈페이지

12 https://www.crowdbreaks.org/?locale=en

가령 크라우드 브레이크에선 트윗을 보여주고, '해당 트윗이 코로나와 연관이 있습니까?' 라고 물어본다. 라벨링 참여자는 '예/아니오' 중 하나를 클릭할 수 있고 그렇게 입력된 답이 라벨링 정보로 추가된다. 많은 시민이 이 프로젝트에 자발적으로 참여하고 있으며 현재까지 백신과 관련한 데이터는 15만 개, 코로나19 관련 데이터는 8만 개가 분류되고, 35만 개가 넘는 트윗에 라벨 정보가 입력되었다.

스위스 연방과학대 연구진은 해당 데이터로 어떤 분석 결과를 얻을 수 있는지도 논문을 통해 공개했다. 이들은 2018년과 2019년에 작성된 트윗을 종합한 결과 백신에 대한 사람들의 반응은 중립 혹은 긍정적었다는 점을 알아냈다.[13] 연구진은 크라우드 브레이크로 고품질 데이터를 모아 인공지능 연구에 활용하면 향후 독감이나 점염병 예측을 미리 할 수 있을 거라고 기대하고 있다.

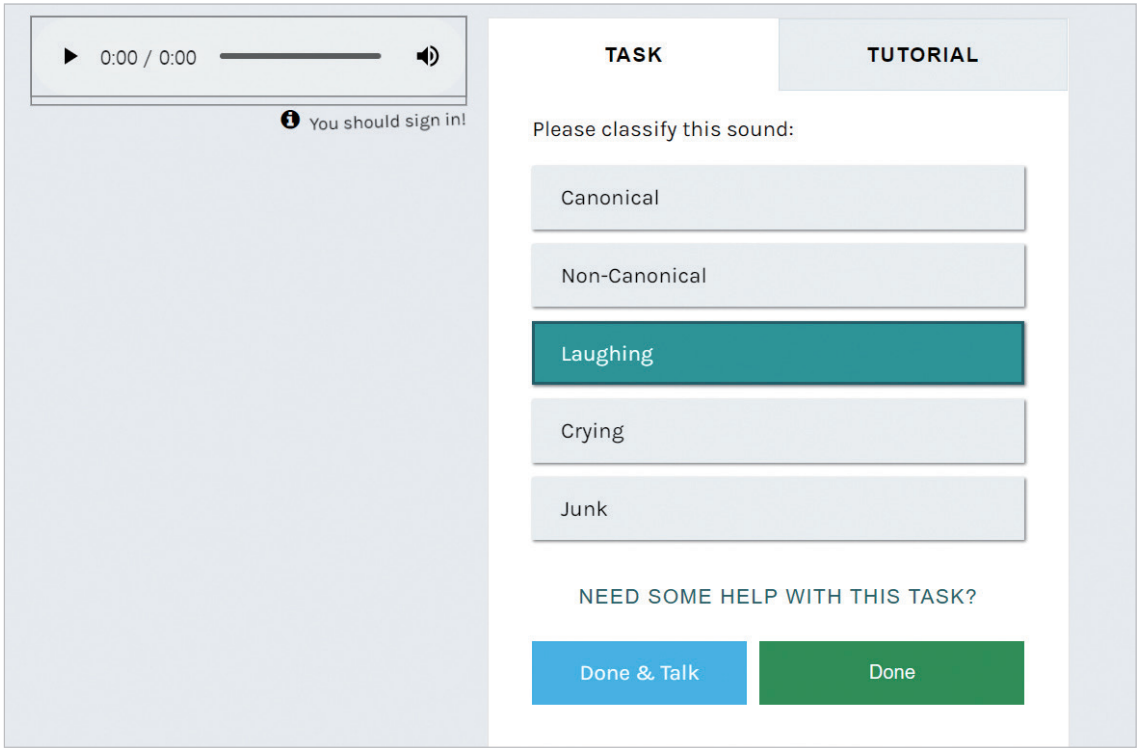
13 https://www.frontiersin.org/articles/10.3389/fpubh.2019.00081/full

유아 언어 분석을 위한 '머츄어티 오브 베이비 사운드'

주요내용

유럽 내 한 심리언어학 연구기관[14]은 2020년 어린아이의 옹알이를 분석해 장애 아이들과 비장애 아이의 언어 차이점을 알아내는 연구를 진행했다. 이때 '머츄어티 오브 베이비 사운드(Maturity of Baby Sounds)'라는 데이터 라벨링 프로젝트를 통해 음성 데이터를 정교하게 만들었다.

[머츄어티 오브 베이비 사운드에서 요청한 음성 데이터 라벨링 작업]



www.zooniverse.org/projects/chiarasemenzin/maturity-of-baby-sounds

14 프랑스 연구기관인 The Laboratoire de Sciences Cognitives et Psycholinguistique (LSCP)와 네덜란드 연구기관인 The Max Planck Institute이 주관했다.

라벨링 참여방법

연구진은 아프리카, 미국, 호주, 아시아, 남미 등에 거주하고 있는 아이들의 옹알이를 녹음하고 분석했는데, 음성 파일에서 어른 목소리나 외부 소음이 섞여 있는 문제를 발견했다. 음성 파일은 10-16시간 정도로 길었기 때문에 연구자가 일일이 들으며 해당 소음을 삭제하기는 힘들었다. '머츄어티 오브 베이비 사운드'는 그런 소음이 있는 음성파일을 시민들의 참여로 걸러내는 프로젝트다. 여기서 연구진은 1초 미만[15]으로 쪼개진 음성 파일을 만들었고, 참여자들은 그 음성파일을 들으며 정상 옹알이 소리인지, 웃는 소리인지, 우는 소리인지 정보를 입력하면 됐다. 2020년 12월 기준 6천 명의 자원봉사자가 이 프로젝트에 참여했고 라벨링 데이터 10만 개를 만들어냈다. 이를 통해 연구에 필요한 데이터 정제 작업을 효율적으로 할 수 있었다.

15 데이터 라벨링 참여자들이 들어어야할 음성 분량을 줄이면서 음성 파일 내용 노출을 최소화해 개인정보 보호하고자 분량을 짧게 만들었다.

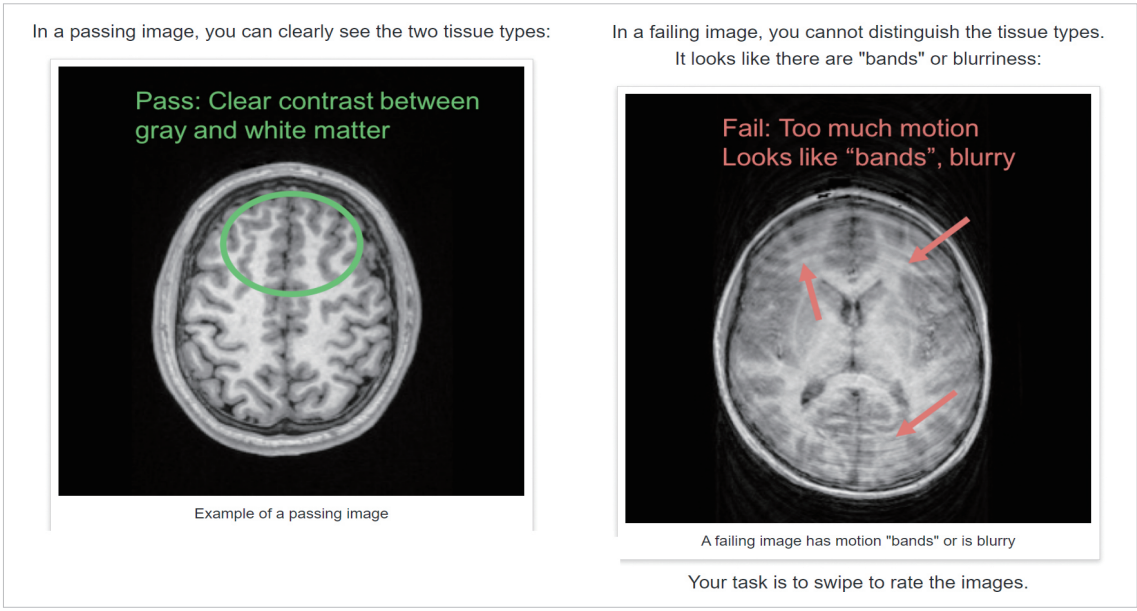
의료계 연구를 도와주는 '브레인닥터/스와이프포사이언스'

주요내용

의료 업계에서 MRI나 CT 사진이 많아 이미지 데이터를 라벨링 하는 작업이 많다. 워싱턴 대학의 한 연구진은 해당 라벨링 작업을 일반인에게 맡길 수 있는 프로그램 '브레인 닥터'를 개발했다.

2019년 워싱턴대학은 청소년기 뇌의 발달 정도와 정신질환 환자들의 뇌 활동을 연구하던 중에 오픈소스로 공개된 뇌 MRI 데이터를 활용하고 있었다.[16] 연구진은 MRI를 보고 의학적 판단하기 전에 사진이 흐리거나 잘못 촬영되지 않았는지 확인해야 했는데, 연구 인력이 모자라서 사진 수백 장을 직접 검수하기 힘들었다. 이를 해결하고자 만든 것이 '브레인 닥터'다.

[브레인닥터 예시 - MRI 사진에서 하얀색과 회색 영역이 분명하게 보이면 정상 사진이고, 주름이 보이거나 흐릿한 사진이면 데이터 분석에 활용할 수 없어 데이터 분석에 활용할 수 없다]



braindr.us/#/

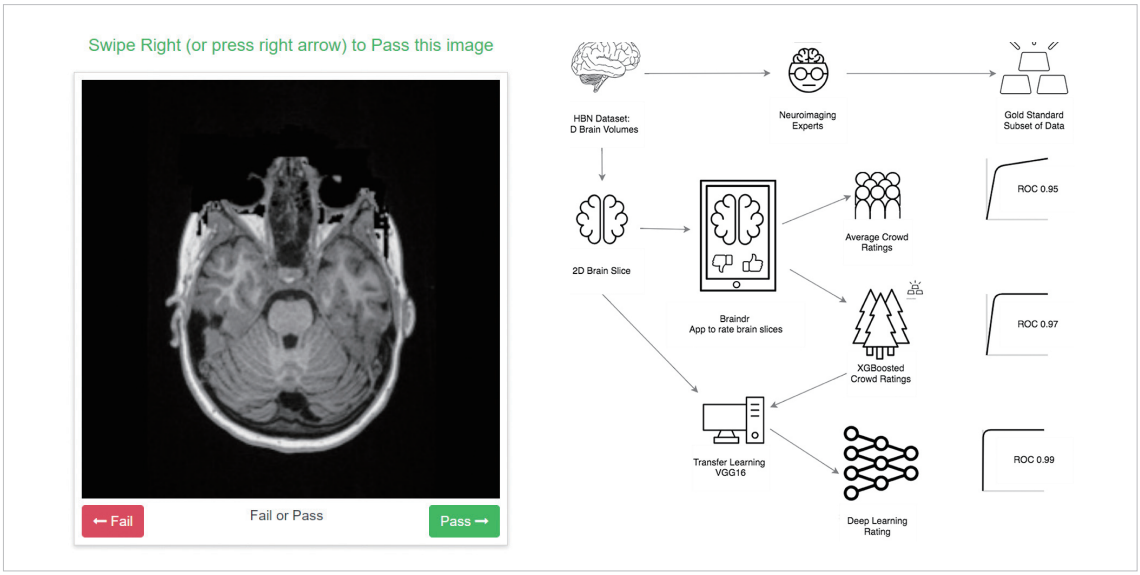
16 https://escience.washington.edu/combining-citizen-science-and-deep-learning-to-amplify-expertise-in-neuroimaging/

브레인 닥터는 웹 기반 게임으로 MRI 사진을 보여주고 참여자에게 '정상(Pass)', '사용불가(Fail)' 버튼을 누르도록 유도한다.[17] 그렇게 입력된 답은 MRI 사진에 자동으로 추가돼 고품질 MRI 사진을 골라내준다. MRI를 해석하는 게 아니라 그림의 이상한 점을 찾는 것과 유사하기 때문에 일반인도 쉽게 참여할 수 있는 게임이다.

라벨링 참여방법

워싱턴 대학이 한 달 동안 해당 서비스를 운영한 결과, 외부 사용자 261명이 브레인닥터로 라벨링 작업에 참여했으며 총 8만 개의 라벨링 정보를 얻을 수 있었다.[18] 연구진은 브레인 닥터로 얻은 1차 데이터를 기반으로 다시 가장 높은 품질의 사진을 연구에 활용했고 결과적으로 연구 시간을 단축시킬 수 있었다.

[브레인 닥터 예시와 구조]



* 사진출처 : 논문

17 이는 틴더라는 데이팅 앱 UI에 영향을 받았다고 한다. 틴더에선 자신이 선호하는 외모 사진을 보고 좋아요 싫어요 형태로 입력하는 디자인으로 구성됐다.

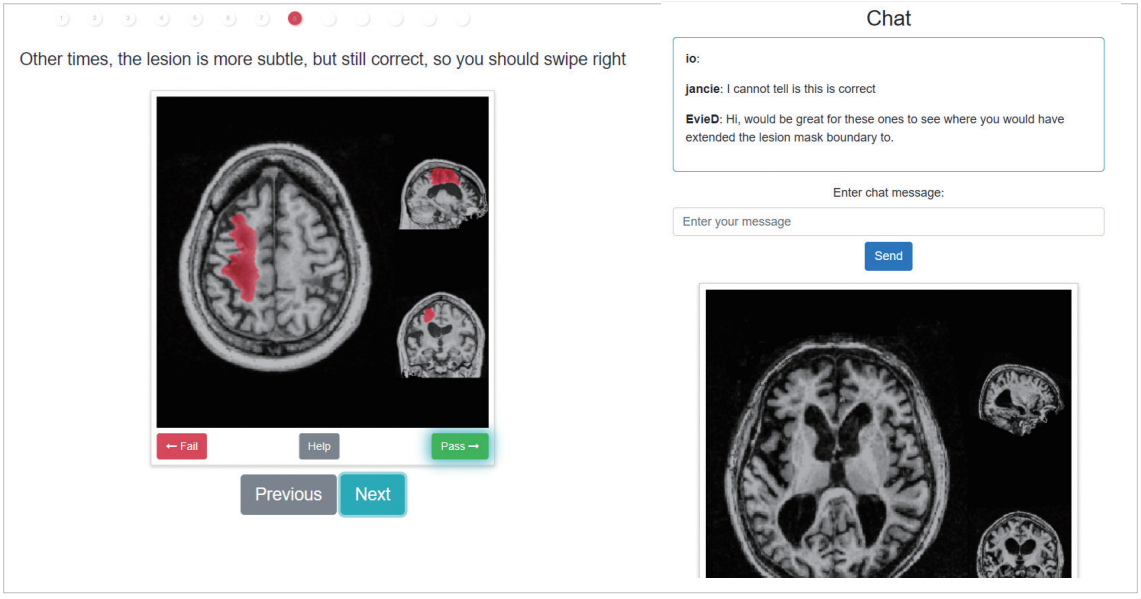
18 https://www.frontiersin.org/articles/10.3389/fninf.2019.00029/full

브레인닥터의 진화

브레인닥터는 이후 '스와이프 포 사이언스(Swipes for Science)'[19]로 발전된다. 이는 브레인닥터처럼 두 가지 정답을 주고 하나를 선택하는 게임이다. 오픈소스 기술[20]이기 때문에 내용을 자유자재로 바꿀 수 있다. 현재까지 이 기술로 제작된 서비스를 살펴보자면 일단 '웨일닥터'[21]가 있다.

웨일닥터는 바다 속 생물의 음성 데이터를 연구하는데 사용되며, 외부 누구나 소리를 듣고 조건에 맞는 소리인지 아닌지 입력할 수 있다. 뇌질환 이미지 데이터 라벨링 도구인 '브레인닥터레스'[22]나 '브레인스팟'[23]은 기존 브레인닥터보다 좀 더 진화한 형태다. 두 게임

[브레인닥스리스 서비스]



braindrles.us/#/tutorial

19 https://slides.com/anishakeshavan/swipes-for-science#/16

20 https://github.com/SwipesForScience/SwipesForScience

21 https://whale-dr.firebaseapp.com/

22 https://braindrles.us/#/play

23 http://brainspot.braindr.us/#/

모두 뇌졸중 증상이 보이는 MRI 사진 예시를 보여주고 조건에 맞아 '통과하기(Pass)'를 입력하면 자동으로 해당 값이 사진에 추가된다. 답을 입력하는 과정에서 잘 모르는 내용은 '대화(Chat)'하기 기능을 이용해 연구진에게 직접 질문할 수 있다. 대화 내용은 웹상에서 계속 노출되기 때문에 같은 라벨링 작업을 하는 다른 사용자들에게도 유용한 자료가 된다. 이 밖에도 '앱스트랙트'[24]는 텍스트 분석 라벨링 도구로 특정 단락에서 포함하고 있는 단어나 숫자를 외부에서 입력할 수 있다.

24 https://appstract.pub/#/

공공성이 높은 데이터 라벨링 프로젝트

크라우드소싱 기반의 데이터 라벨링 작업에서 참여자는 금전적 보상을 받는다. 그런데 금전적 보상이 전혀 없는 프로젝트도 있다. 주로 위키백과처럼 공공성이 높은 프로젝트들에서 그렇다. 학계에서 진행하는 프로젝트에서 특히 크라우드소싱 방식으로 데이터 라벨링을 하는데, 전 세계 많은 시민이 자신의 시간과 노동력을 연구진을 위해 기여하고 있다. 일부 연구진은 데이터 라벨링 과정에서 비용을 발생했더라도 그 결과를 시민에게 공유하기도 한다. 다음은 그런 공공성이 돋보이는 데이터 라벨링 프로젝트다.

화상 탐사 로봇이 더 똑똑해질 수 있도록 돕자, 'AI포마스'

주요내용

화성 탐사 중인 로봇들은 위험한 지형에 들어가 움직이는 못하곤 한다. 나사(NASA)는 이를 방지하기 위해 탐사 로봇이 마치 자율 주행을 하듯 스스로 화성 내 위험한 지형을 학습하고 판단하는 기술을 만들고 있으며, 여기에 필요한 데이터를 시민들에게 받고 있다. 바로 AI포마스(AI For Mars)[25]라는 프로젝트다.

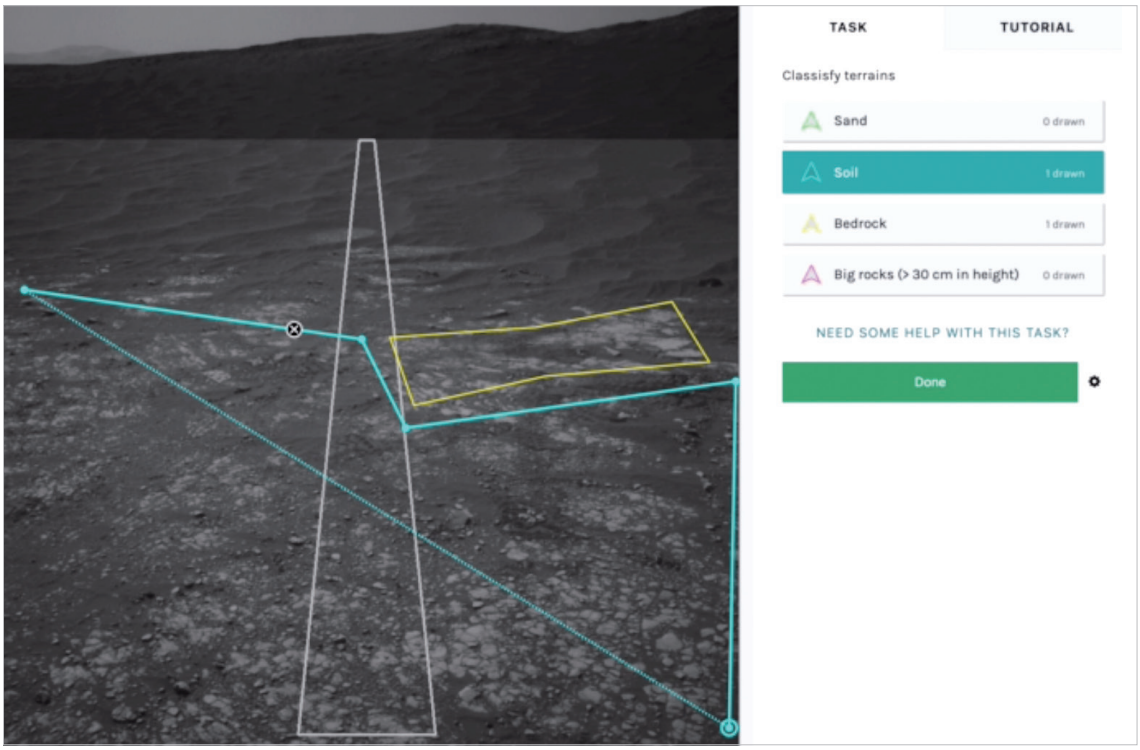
라벨링 참여방법

해당 프로젝트에선 화성 탐사 로봇이 찍은 사진을 보여주고, 참가자들에게 지형 정보를 추가해달라고 요청한다. 어디가 모래이고. 무엇이 돌인지 그림으로 표시해주는 형식이다. 2020년 이 프로젝트가 시행된 이후 현재까지 9,900명이 넘는 자원봉사자가 참여했고, 37만 개 분류 데이터, 4만9천 개의 물체 정보가 입력됐다.[26]

25 https://www.zooniverse.org/projects/hiro-ono/ai4mars/about/research

26 https://www.zooniverse.org/projects/hiro-ono/ai4mars

[AI포마스에서 이미지 라벨링을 하는 과정]



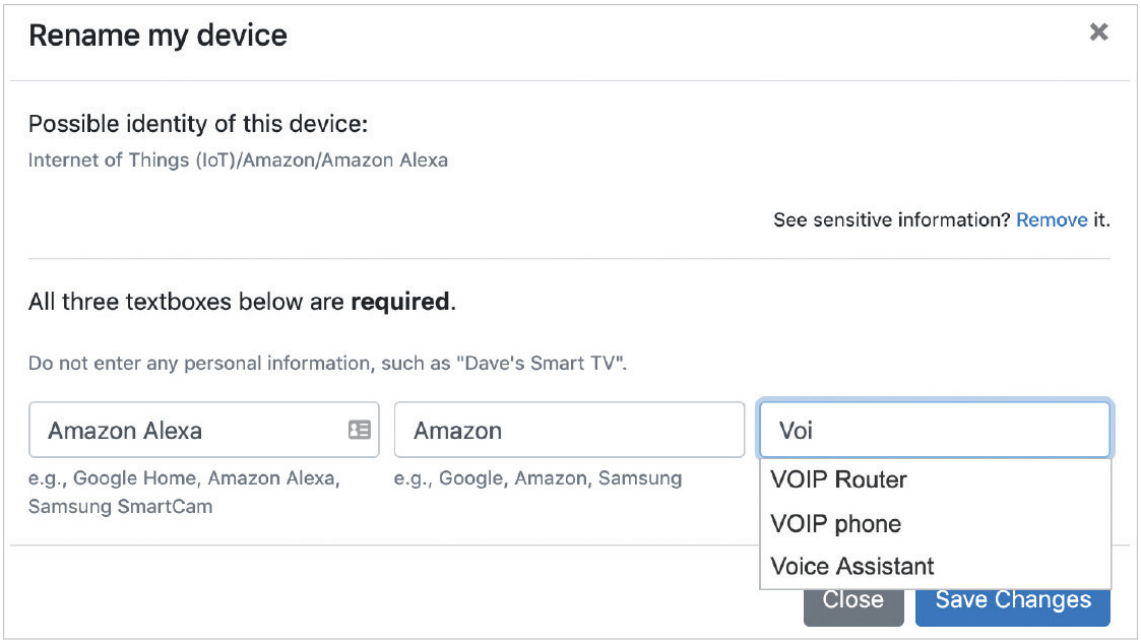
www.zooniverse.org/projects/hiro-ono/ai4mars/classify

스마트홈 기기들의 보안성을 감시하는 'IoT 인스펙터'

주요내용

프린스턴대학, 조지아 공과대학, 시카고 대학 등으로 구성된 연구진은 스마트홈 기기의 사용현황과 보안성 분석에 필요한 데이터를 얻고자 'IoT 인스펙터'[27]를 개발했다. 비슷한 기존 연구들에서는 데이터들의 상세 정보가 부족해 연구를 확장성 있게 하지 못했으나 IoT 인스펙터는 사용자가 실시간 형태로 데이터 상세 정보를 입력하게 유도해 연구자들에게 고품질 데이터를 전달하고 있다. 사용자가 IoT 인스펙터에서 제공하는 소프트웨어를 컴퓨터에 설치하고 스마트홈 기기들의 정보를 입력하면, 네트워크 트래픽 정보와 기기 정보가 연구진에게 전송되는 식이다. 이때 데이터 속에 있는 민감한 정보는 삭제되고 익명 처리된다.

[스마트홈 기기에 데이터 라벨링을 입력하는 과정]



* 사진출처 : 논문[28]

27 https://iotinspector.org/

28 https://arxiv.org/pdf/1909.09848.pdf

라벨링 참여방법

사용자가 데이터 라벨링에 참여하는 과정은 간단하다. IoT 인스펙터 프로그램 내에서 카테고리나 제품명, 제조사명을 입력하면 데이터 라벨링이 자동으로 완성된다. 예를 들어 집에 스마트 스피커가 있다면 제품명은 '아마존 알렉사', 제조사명은 '아마존', 카테고리는 '음성 비서'로 입력하면 된다.

연구진은 IoT 인스펙터를 2019년 출시하고 5개월간 사용자 8,488명을 모으는데 성공했다. 여기서 얻은 데이터 정보는 15만 개였다. 네트워크 트래픽 데이터는 전체 이용자의 절반 가량인 4,322명이 공유했다고 한다.[29] 연구진은 해당 데이터로 스마트홈 기기 제품들의 인기도나, 어떤 기기들이 보안성을 높은지, 어떤 광고회사가 스마트TV 정보를 추적하고 있는지 알아냈다.[30] 관련 데이터들은 외부 연구에서에서도 쓰일 수 있도록 공개했다.

29 https://arxiv.org/pdf/1909.09848.pdf

30 https://docs.google.com/presentation/d/10ke_xQYsDE0S9SMbhGrD4bRlH5miRQBUw-PkhKJW6J8/edit#slide=id.g504dab15f3_0_378

장애인이 다니기 쉬운 도로를 만들자 '프로젝트 사이드워크'

주요내용

도시를 계획할 때 정부 당국은 시각장애인이나 휠체어 등이 필요한 교통 약자를 고려하기 마련이다. 문제는 도로 모양은 시간이 지나면서 조금씩 변하게 되고 이는 교통 약자들에게 여러 불편을 준다. 공사로 인해 도보가 갑자기 끊기거나, 보도블록이 망가지거나 횡단보도와 도로를 잇는 턱이 너무 높은 경우가 대표적이다. 미국에서는 이런 문제를 파악하기 위해 직접 교통약자들이 도로를 돌아다니며 문제점을 점검하고 시 당국에 보고하곤 했으나, 해당 방식은 시간이 너무 많이 소요되고 관련 인력도 너무 부족하다는 단점이 있었다. 워싱턴 대학은 프로젝트 사이드워크[31]라는 아이디어를 내고 집단 지성을 통해 장애인에게 문제가 되는 도로 정보 데이터를 수집하고 업데이트하고 있다.

라벨링 참여방법

프로젝트 사이드워크는 일종의 틀린 그린 찾기 게임하고 비슷하다. 참여자가 사이드워크에서 제공하는 도구에 접속하면 일단 지도를 볼 수 있다. 얼핏 보면 구글 스트리트뷰와 비슷하나 도로 문제점을 표시하는 라벨링 기능이 제공된다. 화면에 보이는 도로가 파손되거나 턱이 너무 높이가 장애물이 있다면 관련 라벨을 클릭해 선택하면 된다. 문제의 심각도도 표시하는 기능도 있다. 같은 장소에 여러 명이 라벨링 했다면 총 몇 명이 입력했는지 관련 통계도 옆에 나온다.

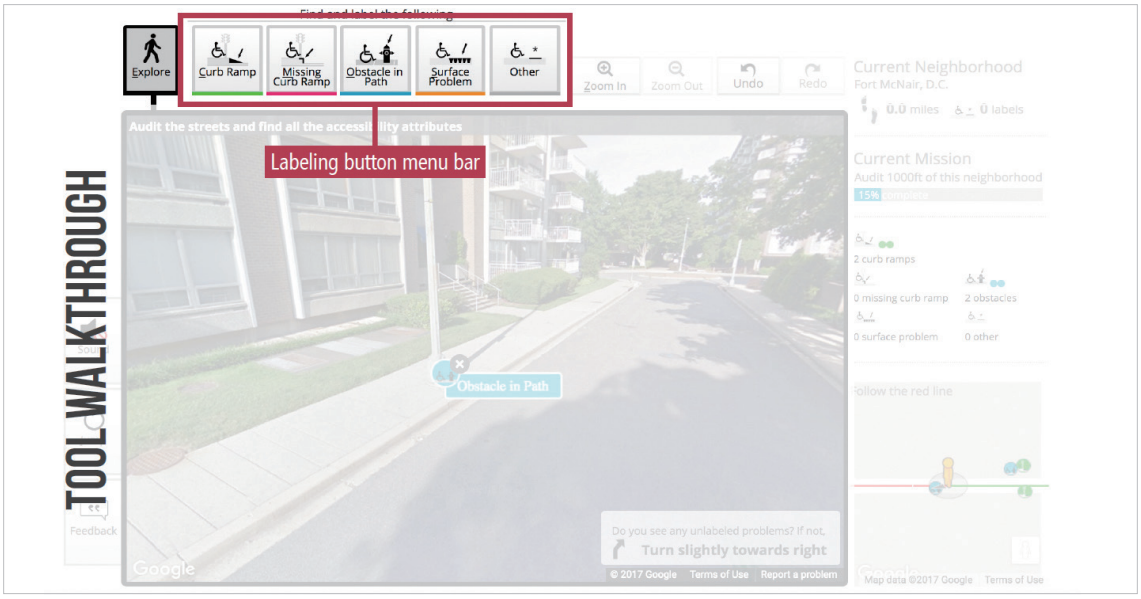
[도로 턱이 너무 높아 휠체어 이동이 힘든 구간(왼쪽), 휠체어 이동 구간에 장애물이 있는 경우(가운데), 도로가 망가진 경우(오른쪽). 이런 환경은 모두 시각 장애인 및 교통약자들이 이동하는데 방해가 된다.]

* 사진출처 : 프로젝트 사이드워크 발표자료[32]

31 https://sidewalk-sea.cs.washington.edu/

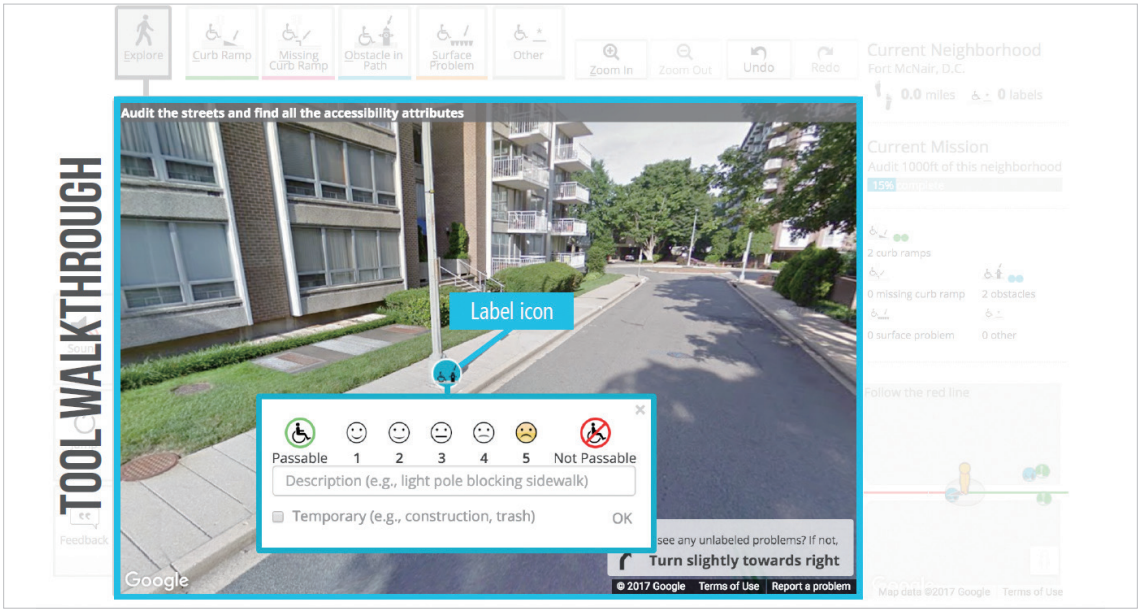
32 https://makeabilitylab.cs.washington.edu/media/talks/Saha_ProjectSidewalk_GoogleGeo2020_web.pdf

[라벨링 할 수 있는 종류는 5가지로, 지도를 보다가 문제가 되는 도로를 발견하면 적절한 라벨을 추가하면 된다]



* 사진출처 : 프로젝트 사이드워크 발표자료[33]

[문제 정도가 얼마나 심각한지 표시할 수 있다]

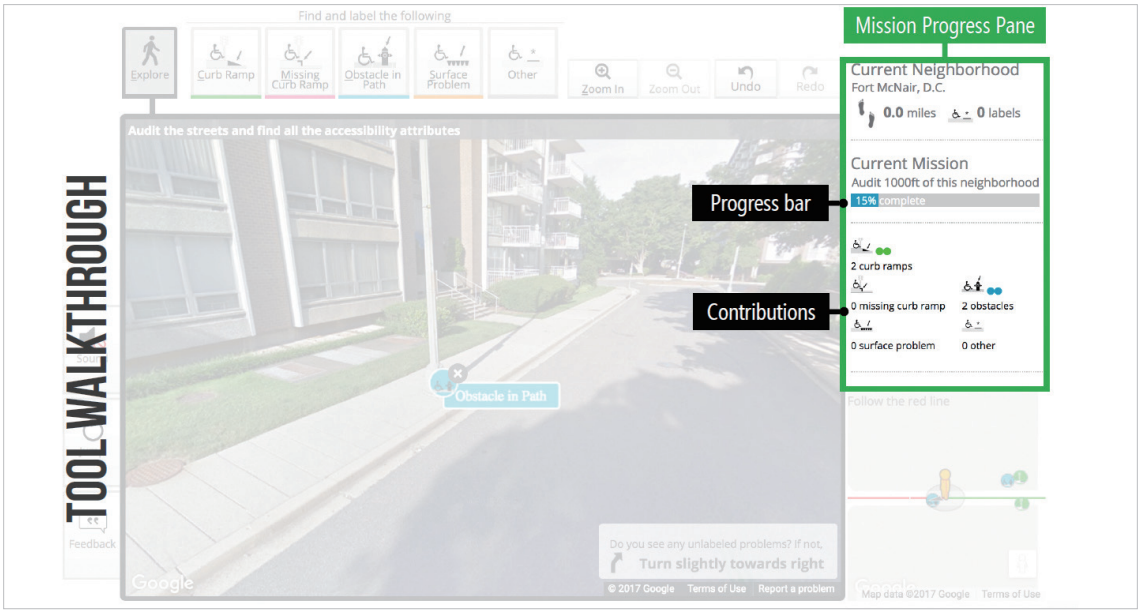


* 사진출처 : 프로젝트 사이드워크 발표자료[34]

33 https://makeabilitylab.cs.washington.edu/media/talks/Saha_ProjectSidewalk_GoogleGeo2020_web.pdf

34 https://makeabilitylab.cs.washington.edu/media/talks/Saha_ProjectSidewalk_GoogleGeo2020_web.pdf

[거리를 얼마나 검토했는지, 참여자는 몇 명인지 통계를 볼 수 있다]

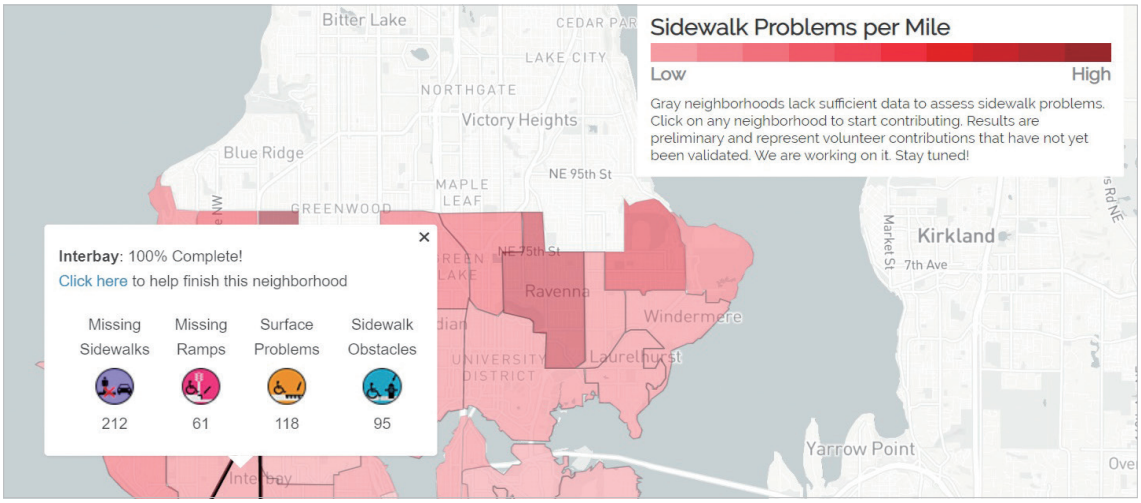


* 사진출처 : 프로젝트 사이드워크 발표자료[35]

프로젝트 사이드워크는 튜토리얼 형태로 참여방법을 자세히 소개해 사전 지식 없이 누구나 관련 도로를 점검할 수 있도록 도와준다. 이렇게 해서 저장된 데이터들은 자동으로 실제 위치와 결합해서 도로의 접근성 현황을 파악하는 시각화 자료로 활용되고 있다. 또한 접근성이 높은 위치를 파악할 수 있어 장애인 및 교통 약자를 위한 별도의 지도 앱이나 길찾기 서비스를 만드는 데 활용되고 있다. 워싱턴 대학은 프로젝트 사이드워크에서 직접 수집하고 라벨링 한 데이터를 API 형태로 외부에 공개하고 관련 소프트웨어도 오픈소스로 개방 중이다.

35 https://makeabilitylab.cs.washington.edu/media/talks/Saha_ProjectSidewalk_GoogleGeo2020_web.pdf

[프로젝트 사이드워크에서 수집한 데이터 라벨링 값들로 도로 접근성을 시각화한 자료]



* 사진출처 : 공식홈페이지[36]

워싱턴 대학은 연구 초기 관련 데이터를 많이 수집하기 위해 자원봉사자들을 모집하는 동시에 아마존 메커니컬 터크로 외부 참여자를 고용하기도 했다. 한 가지 흥미로운 건 워싱턴 대학은 별도로 고용한 데이터 라벨링 참여자에게 법적 최저 임금을 보장하려고 노력했다는 점이다. 당시 법적 최저 임금은 시간당 7.25달러였는데, 연구진은 참여자들이 온라인으로 검토한 도로 길이를 기준으로 임금을 책정했고, 최종적으로 시간당 평균임금 8.21달러[37]를 제공했다.

워싱턴 대학은 첫 프로젝트를 2016년 가을부터 18개월 동안 진행했으며 워싱턴 DC 거리 위주로 데이터를 수집했다. 총 자원봉사자 627명과 아마존 메커니컬 터크로 고용한 데이터 라벨러 170명이 데이터 라벨링 작업을 수행했다. 이들이 점검한 워싱턴 DC 거리는 4,733km 분량이었으며, 총 20만 개 넘는 라벨링 데이터를 생성했다. 프로젝트 사이드워크는 꾸준히 관심을 받으며 지금은 워싱턴 DC 외에도 미국의 3개 도시와 멕시코의 2개 도시의 지도 데이터를 만들어 접근성 관련한 데이터를 수집하고 라벨링하고 있다.

36 https://sidewalk-sea.cs.washington.edu/results

37 https://makeabilitylab.cs.washington.edu/media/publications/Saha_ProjectSidewalkAWebBasedCrowdsourcingToolForCollectingSidewalkAccessibilityDataAtScale_CHI2019.pdf

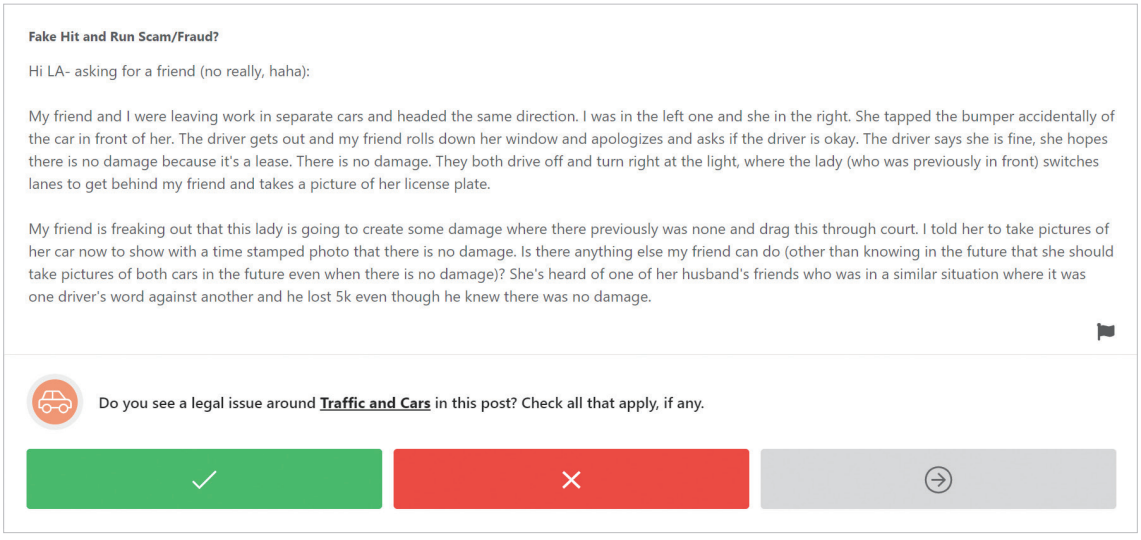
법률 분야 인공지능 연구를 돕는 '러니드 핸드'와 '애티커스 프로젝트'

주요내용 및 라벨링 참여방법

스탠포드 로스쿨과 서퍽 대학 법률 혁신 기술 랩은 '러니드핸드'[38]라는 프로젝트를 시작해 법률계 인공지능 발전을 위한 데이터를 모으고 있다. 러니드핸드의 구조는 간단하다.

이 서비스는 '레딧'[39]이나 소셜미디어에서 법률 관련 콘텐츠를 가져오고 해당 내용이 법률 내용에 속하는지 사용자에게 묻는다. 참여자는 해당 질문이 법률 문제를 다루고 있는지 파악해서 '예/아니오'로 답을 입력하고 세부적으로 노동, 주택, 세금 등 어느 범주에 속하는 내용인지 선택할 수 있다. 이렇게 얻어진 데이터로 스탠포드 대학은 사람들이 최근 어떤 법률 문제를 고민하는지 흐름을 파악하고 수요를 예측하고 있다. 데이터 라벨링 작업은 일반인 누구나 할 수 있으나 로스쿨 학생이나 법률 전문가도 직접 참여하고 있다. 미국 변호사라면 러니드 핸드에서 데이터를 입력하는 과정을 무료 법률 활동(프로 보노)으로

[러니드핸드에서 제공하는 질문. 해당 콘텐츠가 교통법과 관련된 내용인지 묻고 있다]



* 사진출처 : 공식홈페이지

38 https://betterinternet.law.stanford.edu/about-the-project/learned-hands/

39 한국의 네이버 지식인, 다음 카페와 유사한 서비스. 영미권 유명 커뮤니티다.

보고 미국 변호사 협회에서 그 시간을 공식 인정해주고 있다. 이 프로젝트는 2018년에 시작해 지금까지 3,400여 개 법률 데이터를 분석해 3만5천 개가 넘는 데이터 라벨값을 생성했다. 러니드 핸드에서 만든 데이터는 누구나 다운로드할 수 있다.[40]

애티커스 프로젝트[41]에서도 법률 분야 AI 연구에 필요한 데이터를 제작하고 라벨링 작업을 하고 있다. 2020년 비영리 기관으로 출범한 이곳은 세일즈포스에 속한 기업 변호사 웨이 첸[42]이 설립했다.

애티커스 프로젝트의 라벨링 작업은 주로 변호사나 로스쿨 학생들이 하고 있다. 2021년 첫 번째로 공개한 데이터[43]는 금융과 관련된 계약서 510개를 분석해서 1만3천 개 라벨 정보를 만들었다. 여기에 UC 버클리대 AI 연구팀과 변호사 16명과 로스쿨 학생 22명 기술 고문 4명이 참여했다.[44] 데이터는 누구나 무료로 사용할 수 있게 CC BY 4.0 라이선스 하에 공개됐다.

애티커스 프로젝트는 데이터 라벨링 관련 전문가를 체계적으로 영입하기 위해 '오픈 데이터 펠로우십 프로그램'[45]도 운영하고 있다. 변호사와 로스쿨 학생을 학생으로 대상으로 진행되는 이 프로그램에선 리걸테크(Legal Tech)를 발전을 도모하며 이와 연관된 프로젝트를 구상하고 데이터를 생성하는 작업을 주도하고 있다.

40 https://learned-hands.github.io/project-hub/data.html

41 https://www.atticusprojectai.org/

42 https://www.linkedin.com/in/weichen221/

43 https://www.atticusprojectai.org/cuad-v1-performance-announcements

44 https://www.atticusprojectai.org/cuad

45 https://www.atticusprojectai.org/open-data-fellow-programs

음성 인식 연구를 도와주는 '아이히어유'&'ML커먼즈'

주요내용 및 라벨링 참여방법

'아이히어유'[46]는 음성인식 연구에 필요한 데이터를 크라우드 소싱 방식으로 모으는 프로젝트다. 독일 대학 연구진들이 이끌고 있으며, 비영어권 국가의 음성이나 감정을 분석하는데 필요한 음성 데이터를 게임 형태로 얻고 있다. 게임에선 짧은 음성을 들려주고 해당 목소리의 성별이나 나이대, 감정을 사용자에게 선택해달라고 요청한다. 입력된 답안이 음성 데이터 라벨값으로 입력되는 식이다. 직접 지문을 보여주고 해당 문장을 읽어 녹음해 올리는 과제도 있다. 음식을 먹으면서 말을 하고 있는 음성을 올리고 '해당 음성에서 어떤 음식을 먹고 있는가'라든지, '피곤했을 때 목소리를 녹음해서 올려달라' 같은 흥미로운 과제도 있다.

아이히어유에서 모인 데이터는 익명처리해서 저장하며, 자체적으로 발표하는 연구[47]에서는 참여자의 나이대나 성별 등을 추가로 입력받아 음성 인식 기술에 활용했다. 현재까지 아이히어유에 모인 데이터는 1,600여 개[48]로 주로 독일어나 영어 중심 데이터가 많다.

ML커먼즈[49]은 2018년부터 시작한 프로젝트로 여러 IT 기업과 학계가 모여 미래 사회에 필요한 데이터를 만들고 있다. '모두를 위한 기계학습'이란 목표를 두고 있어 특히 인공지능 개발에 적합한 데이터 개발에 집중하고 있다. 알리바바, 구글, 시스코, 페이스북, 바이두, 삼성 등 대형 IT 기업들이 설립 멤버로 참여했다.

46 https://www.ihearu-play.eu/

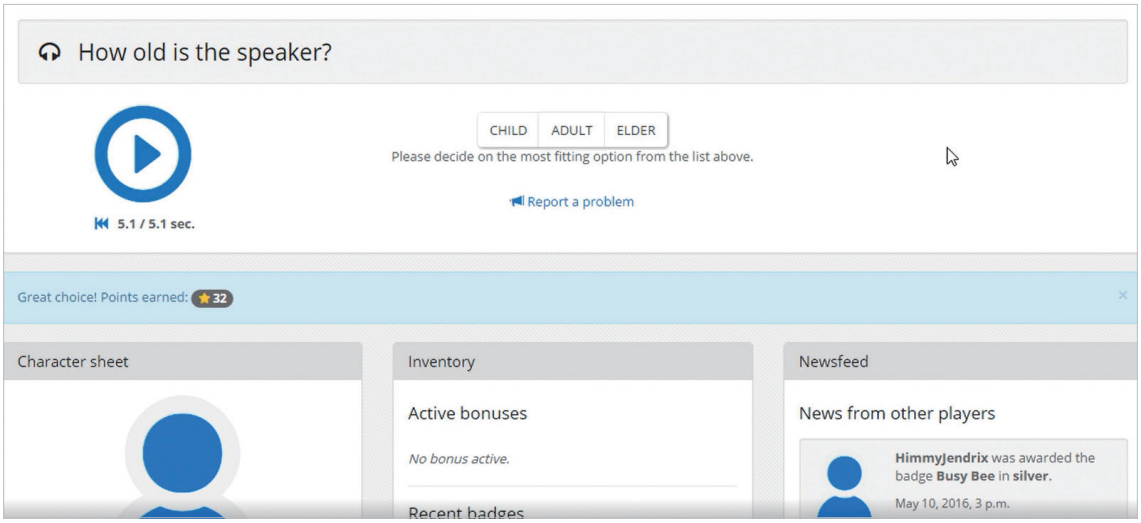
47 http://www.ijac.net/en/article/doi/10.1007/s11633-019-1180-0

48 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154486#abstract0

49 https://mlcommons.org/ko/

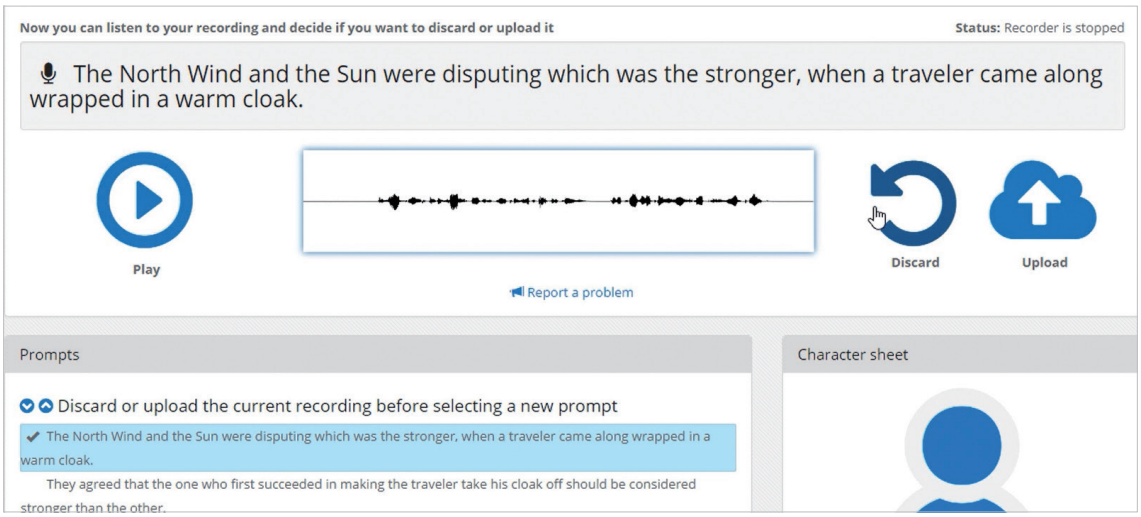
최근 진행하는 만들고 있는 데이터는 음성 인식 관련 데이터[50]다. 아직 파일을 공개하진 않았지만 해당 데이터는 56개 언어로 제작되고 분량은 8만7,000시간 규모가 될 예정이다. 라벨링 데이터와 대본도 지원된다. ML커먼즈는 컴퓨터 비전 분야의 발전을 촉진시켰던 '이미지넷'에서 영감을 받아 해당 데이터를 구축했다고 밝혔다.

[아이히어유 과제 예시. 누구나 해당 데이터를 듣고 답을 입력할 수 있다]



* 사진출처 : 공식홈페이지

[아이히어유 녹음 과제 예시]



* 사진출처 : 공식홈페이지

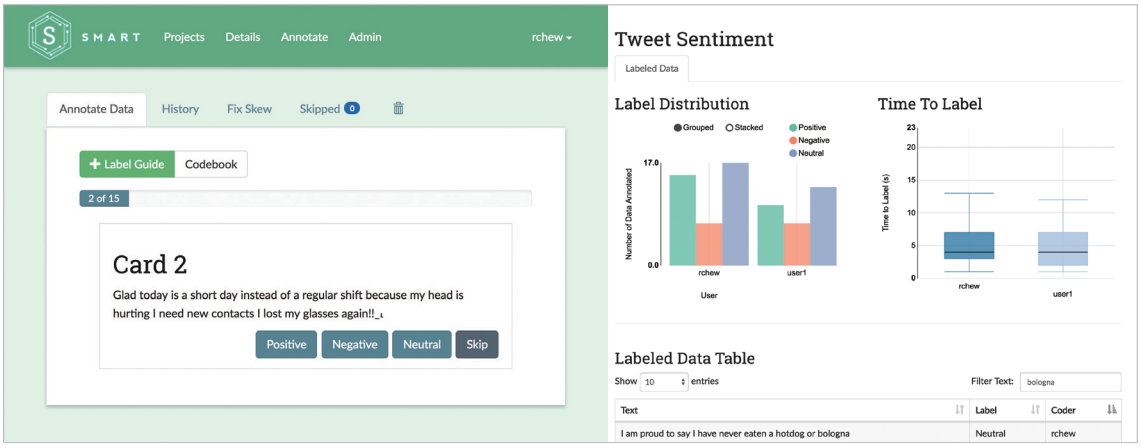
50 https://mlcommons.org/en/peoples-speech/

오픈소스 데이터 라벨링 도구 '스마트' & '라벨스튜디오'

주요내용 및 라벨링 참여방법

최근엔 데이터 라벨링 기업이 많아지자 오픈소스 기반 데이터 라벨링 도구들도 관심을 받고 있다. 스마트[51]는 데이터 라벨링을 도와주는 웹 도구로 특히 크라우드 소싱 방식의 라벨링 작업을 할 때 유용하다. 현재는 텍스트 데이터 라벨링만 가능하고, 라벨링 과정에서 필요한 과제를 객관식 형태로 지원하거나 도구 안에서 데이터를 바로 학습시키거나(Active Learning) 또는 데이터 대한 신뢰성도 직접 확인 할 수 있다.[52] 이 기술은 비영리 학술 단체인 RTI 인터내셔널이 개발했으며 데이터 과학 국가 컨소시엄(The National Consortium for Data Science (NCDS))의 지원을 받고 있다. NCDS은 IBM, GE, 딜로이트, 노스캐롤라이나 대학교, 미국 환경보호국 등 재계, 학계, 정부 기관이 모여 만든 단체다.

[스마트로 만든 데이터 라벨링 프로젝트 예시. 논문에서는 트윗 데이터의 감정을 라벨링하는 작업을 예시로 보여줬다]



arxiv.org/abs/1812.06591

51 https://rtiinternational.github.io/SMART/

52 https://smart-app.readthedocs.io/en/latest/

라벨 스튜디오[53]도 최근 오픈소스 라벨링 도구로 각광받고 있다. 하텍스라는 데이터 라벨링 기업이 만든 이 기술은 오픈소스 기반으로 사용자가 맞춤화한 데이터 라벨링을 할 수 있도록 돕는다. 라벨링 할 수 있는 데이터는 이미지, 텍스트, 오디오, 영상 등 다양하며 라벨링 방식에 대한 템플릿을 제공하거나 외부 인프라기술과 쉽게 결합할 수 있게 만들었다. 현재 수백만 개 데이터가 라벨 스튜디오로 라벨링 작업이 이뤄지고 있고 IBM, 엔비디아, 인텔 등도 이 서비스의 고객으로 등록돼있다.

53 https://labelstud.io/

기업들의 데이터 라벨링 활용 사례

야후, 애플 등에서 엔지니어 생활을 하다가 데이터 라벨링 기업인 데이터소어를 설립한 이반 리 CEO는 한 인터뷰[54]에서 자신이 이전에 엔지니어로 일할 당시 수백만 달러 예산을 투자하며 데이터에 라벨링 작업을 하는 것을 보고 관련 기업을 설립했다고 밝혔다. 중국 AI 전문 미디어인 싱크는 데이터 분석을 많이 하는 자율 주행차 기업은 한 달에 데이터 라벨링 작업에 수백만 달러를 투자한다고 설명했다.[55] 이런 기사를 보았을 때 IT 기업들이 방대한 데이터를 라벨링 하는데 많은 예산을 투입하고 있다는 것을 유추할 수 있다.

분야로 따지면 헬스케어, 자동차, 소프트웨어 중심 기업들이 주로 데이터 라벨링 기업들의 핵심 고객이다. 기업들은 자신이 어떤 데이터 라벨링 서비스를 이용하는지 정도는 공개하나 그 과정은 구체적으로 잘 밝히지 않는다. 아래는 연구 논문, 개발 블로그 등으로 공개된 기업들의 데이터 라벨링 사례들이다. 이를 통해 기업들이 어떤 목적을 가지고 어떤 과정을 거쳐 데이터 라벨링 작업을 하는지 엿볼 수 있다.

IBM의 스포츠 게임 영상 분석 사례

IBM은 스포츠 경기의 하이라이트 영상으로 자동으로 추출하는 기술을 개발하는 과정에서 데이터 라벨링 기술을 활용했으며 이 내용을 논문[56]으로 공개했다. 해당 기술을 만들기 위해 IBM은 지난 13년 동안 열린 국제 테니스 경기 영상 454개를 활용했다. 기존 영상들은

54 https://techcrunch.com/2020/09/29/datasaur-snags-3-9m-investment-to-build-intelligent-machine-learning-labeling-platform/

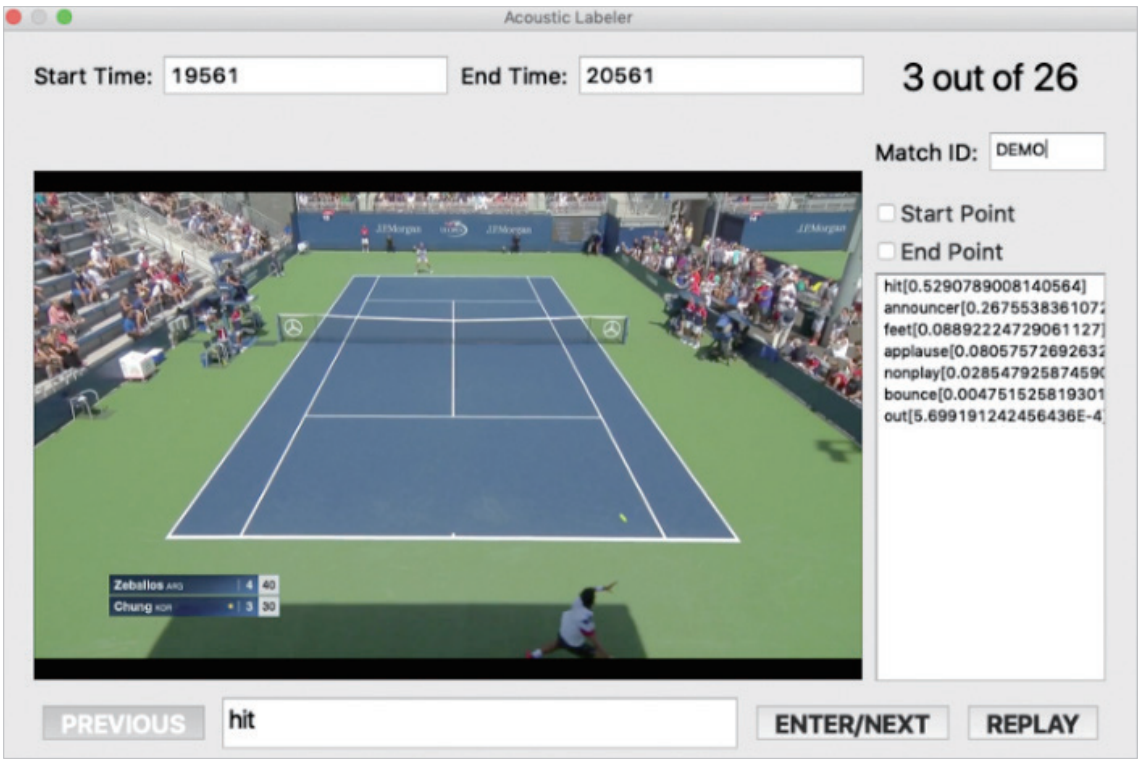
55 https://medium.com/syncedreview/data-annotation-the-billion-dollar-business-behind-ai-breakthroughs-d929b0a50d23

56 https://researcher.watson.ibm.com/researcher/files/us-wangshiq/AB_MMSports2019.pdf?s_tact=C43202QW

최종점수, 경기장소, 선수 이름, 날짜, 시간, 선수가 움직인 거리 등의 간단한 정보가 있었는데, 이런 데이터로는 고도화된 영상 분석 서비스를 만들 수 없었으며 결국 IBM은 데이터를 새로 만드는 작업에 착수했다.

IBM 연구팀은 먼저 영상에서 소리가 가장 높아지거나 낮아지는 구간을 중심으로 영상을 잘랐다. 분석하려는 영상은 900시간 분량이었으며, 구간별로 자른 영상에 중계진 음성, 박수 소리, 공 소리, 발소리 등으로 정보를 추가하는 절차를 거쳤다. IBM은 이를 위해 별도의 라벨링 도구를 개발했고 사람이 직접 라벨링 정보를 추가하도록 지원했다. 그 결과 음성 데이터 2만 개의 라벨링 작업을 5일 안에 완성할 수 있었다. IBM은 이 데이터를 영상 관련 머신러닝 기술 개발에 활용했다고 밝혔다.

[IBM이 직접 만든 데이터 라벨링 도구]



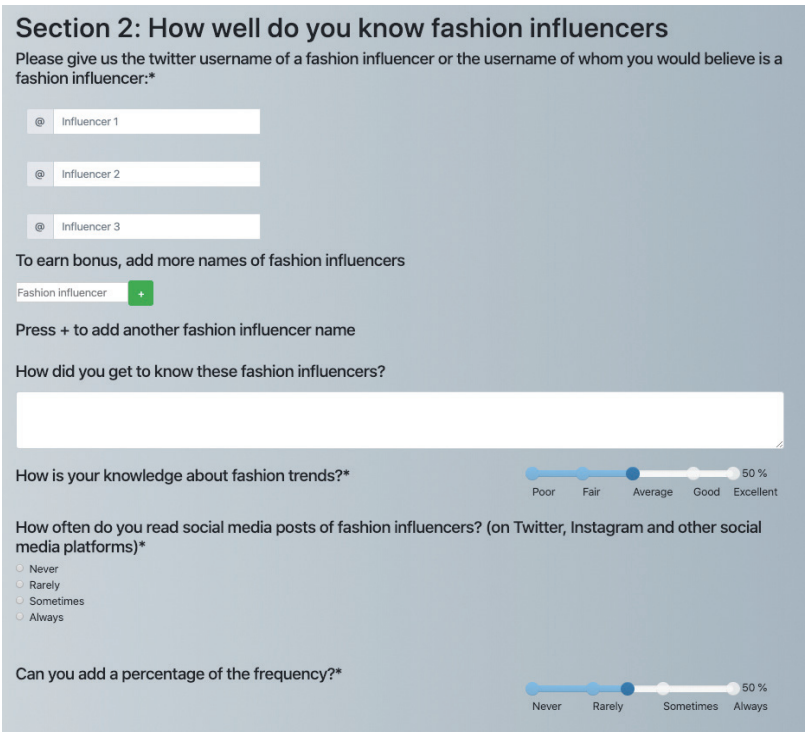
* 사진출처 : 논문[57]

57 https://researcher.watson.ibm.com/researcher/files/us-wangshiq/AB_MMSports2019.pdf?s_tact=C43202QW

유럽의 패션 데이터 통합 프로젝트 '패션 브레인'

패션브레인[58]은 유럽 내 기업과 대학들이 모여 패션업계 데이터를 통합해 분석하려는 프로젝트다. 독일의 온라인 쇼핑몰 잘란도, 독일 베를린보이트 기술대학, 영국 셰필드대학교, 스위스 프리부르 대학교, 스위스의 이미지 분석 기업 패쉬웰,[59] 네달란드에서 만든 오픈소스 DB 모넷DB 개발자들이 함께 진행했으며 이들은 다양한 종류의 데이터로 패션 트렌트 예측 기술이나 패션 업계에 맞춤화된 추천 서비스나 검색 기술을 개발하고 있다.

[패션브레인에서 적용한 데이터 라벨링 과정. 데이터 라벨링 참여자에게 알고 있는 패션 인플루언서 계정 정보를 입력해달라고 요청했다]



* 사진출처 : 관련 논문[60]

58 https://fashionbrain-project.eu/

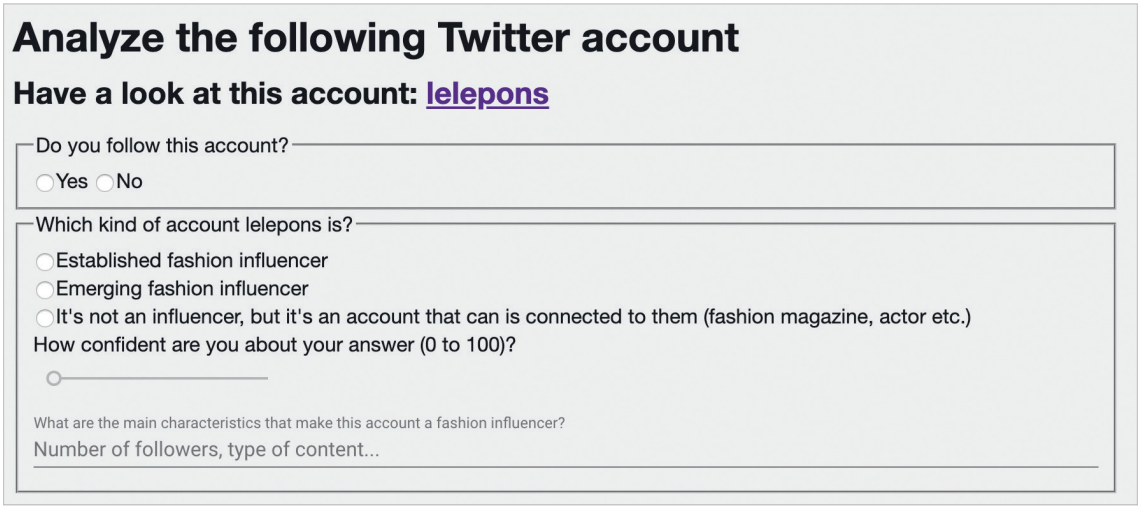
59 현재는 애플에 인수된 상태

60 https://fashionbrain-project.eu/wp-content/uploads/2021/01/D3.3_v2_Dec19.pdf

패션브레인은 기술 개발 과정에서 패션업계에 영향력이 높은 인물 정보를 웹에서 수집하는 과정을 거쳤는데 이 과정에서 피규어 에이트[61]와 아마존 메커니컬 터크를 이용했다. 먼저 패션업계 영향력이 높은 인물 정보를 수집하기 위해 피규어 에이트로 외부 참여자 250명을 고용했고 그들이 알고 있는 패션계 인플루언서 이름과 트위터 계정 주소를 입력하도록 요청했다. 이를 통해 890개 데이터가 모였으며 연구진은 이 중 34%가 연구에서 쓰기 좋은 의미 있는 데이터였다고 밝혔다.

또한 유명 패션 인플루언서 계정에서 연결된 해쉬태그나 리트윗 글을 통해 새로운 패션 업계 인물을 발굴하기도 했다. 이를 위해 아마존 메커니컬 터크로 참여자 175명을 고용했고 여기서 375개 라벨 데이터를 얻어냈다.

[패션브레인에서 적용한 데이터 라벨링 과정. 특정 소셜 계정에 대한 정보를 파악하기 위해 질문들이다]

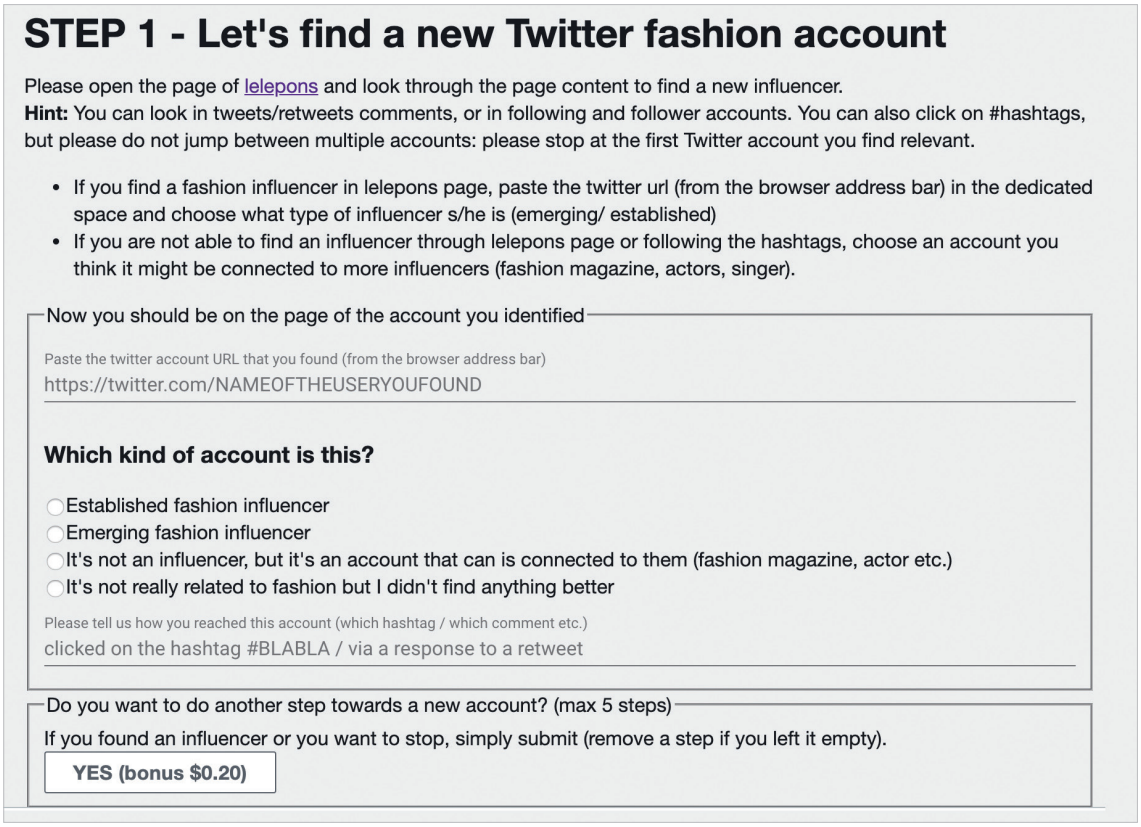


* 사진출처 : 관련 논문[62]

61 Figure Eight. 현재는 애픈(Appen)에 인수된 상태

62 https://fashionbrain-project.eu/wp-content/uploads/2021/01/D3.3_v2_Dec19.pdf

[패션브레인에서 적용한 데이터 라벨링 과정]



STEP 1 - Let's find a new Twitter fashion account

Please open the page of lelepons and look through the page content to find a new influencer.
Hint: You can look in tweets/retweets comments, or in following and follower accounts. You can also click on #hashtags, but please do not jump between multiple accounts: please stop at the first Twitter account you find relevant.

- If you find a fashion influencer in lelepons page, paste the twitter url (from the browser address bar) in the dedicated space and choose what type of influencer s/he is (emerging/ established)
- If you are not able to find an influencer through lelepons page or following the hashtags, choose an account you think it might be connected to more influencers (fashion magazine, actors, singer).

Now you should be on the page of the account you identified

Paste the twitter account URL that you found (from the browser address bar)
https://twitter.com/NAMEOFTHEUSERYOUFOUND

Which kind of account is this?

- Established fashion influencer
- Emerging fashion influencer
- It's not an influencer, but it's an account that can is connected to them (fashion magazine, actor etc.)
- It's not really related to fashion but I didn't find anything better

Please tell us how you reached this account (which hashtag / which comment etc.)
clicked on the hashtag #BLABLA / via a response to a retweet

Do you want to do another step towards a new account? (max 5 steps)

If you found an influencer or you want to stop, simply submit (remove a step if you left it empty).

YES (bonus $0.20)

* 사진출처 : 관련 논문[63]

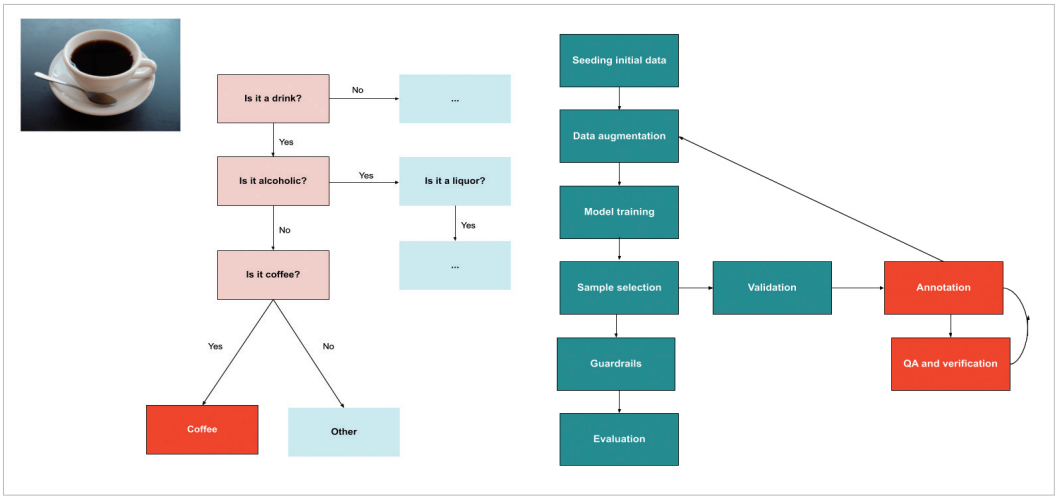
63 https://fashionbrain-project.eu/wp-content/uploads/2021/01/D3.3_v2_Dec19.pdf

도어대쉬 이미지 데이터 라벨링 사례

도어대쉬(doordash)는 미국의 '배달의 민족'이라고 불리는 업계 1위 음식배달 중개업체다. 이들은 플랫폼 기업으로 음식 정보, 후기, 식당 이름 같은 데이터 수천만 개를 보유하고 있었는데, 해당 데이터로 어떻게 검색 기능을 개선했는지 기술 블로그에 공개했다. 여기서 기계학습에 필요한 데이터를 어떻게 구축하고 라벨링 했는지 살펴볼 수 있다.[64]

일단 도어대쉬는 사용자가 '디저트'라는 메뉴를 검색할 때 보다 적절한 검색 결과를 내보내고 싶었다. 이를 위해선 각 데이터 안에 속성값이 라벨링 됐어야했고 아마존 메커니컬 터크를 포함해 여러 라벨링 기업 서비스를 통해 라벨링 작업을 진행했다. 이때 도어대쉬는 라벨값이 구체적으로 나올 수 있도록 객관식으로 문제를 내고, 너무 주관적인 평가값은 입력하지 않도록 유도했다. '유명한', '편리한' 같은 주관적인 라벨값은 이후 사람이나 시기에 따라 변화할 수 있고, 데이터 학습 결과에 안 좋은 영향을 줄 것이라고 본 것이다.

[도어대쉬 라벨링 작업]



* 출처 : 공식 기술 블로그[65]

64 https://doordash.engineering/2020/08/28/overcome-the-cold-start-problem-in-menu-item-tagging/

65 https://doordash.engineering/2020/08/28/overcome-the-cold-start-problem-in-menu-item-tagging/

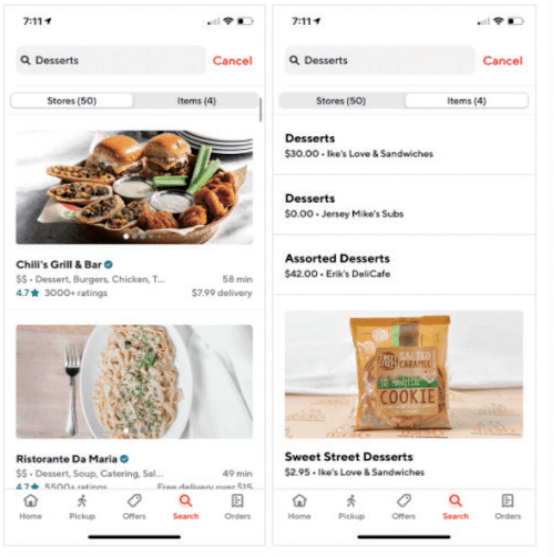
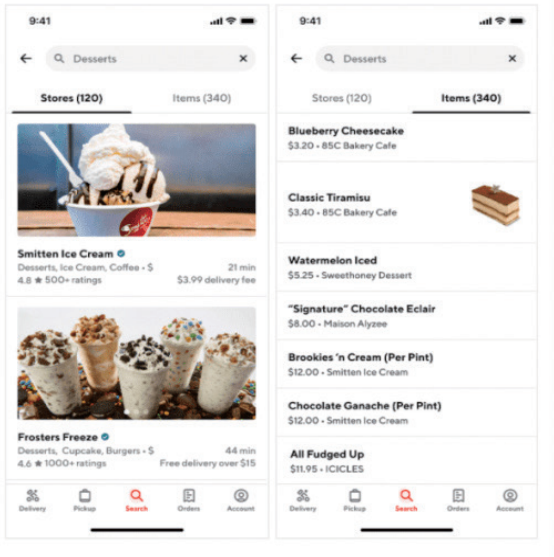
이런 과정을 거쳐 결과적으로 도어대쉬는 데이터 검색 기능을 따로 수정하지 않고 데이터 학습만으로 검색 수준을 개선할 수 있었다고 설명한다.

[데이터 라벨링 질문과 답변 예시]

Item Name	Picture	Task Question	Annotator Response
Fried Chicken Sub		Is this: **sandwich** **burger** **other** **not sure**	burger
Fried Chicken Sub		Is this: **vegan** **other** **not sure**	other

* 출처 : 공식 기술 블로그[66]

[데이터 라벨링 및 기계학습 후 서비스 변화(오른쪽). 같은 '디저트' 단어를 입력했을 때 결과 값이 더 많이 나올 수 있었다]

Current state of item search results	*Future state of item search results*
	

* 출처 : 공식 기술 블로그[67]

66 https://doordash.engineering/2020/08/28/overcome-the-cold-start-problem-in-menu-item-tagging/

67 https://doordash.engineering/2020/08/28/overcome-the-cold-start-problem-in-menu-item-tagging/

디파인드크라우드 음성인식 및 얼굴인식 모델 개발

디파인드크라우드[68]는 니보(Neevo)라는 개방형 데이터 라벨링 서비스를 운영하는 기업이다. 누구나 니보를 통해 데이터 라벨링 작업에 참여할 수 있으며, 기업이 데이터 라벨링 작업을 요청했을때 니보는 가장 적절한 참여자를 기업에 연결해준다. 현재 니보에 등록한 사용자는 21만 명이 넘은 상태다.

한 오디오 하드웨어[69] 기업은 디파인드크라우드 기술로 음성인식 기능을 개선했고 이에 대한 과정을 자세히 공유했다.[70] 해당 기업은 '재생', '노래 반복' 같은 기본적인 명령어부터 '근처 식당 좀 찾아줘' 같은 복합적인 명령어를 인식하는 기술을 구현하고자 했으며, 동시에 집 안, 야외 등 소음 유무나 사용자의 억양에 상관없이 인식이 잘 되는 음성 기술을 계획하고 있었다.

해당 기업은 니보를 통해 데이터 라벨링 참여자를 뽑고 네 그룹으로 나눠 각자 다른 과제를 부여했다. 먼저 230명으로 구성된 첫 번째 그룹에게 목소리를 녹음해 올리는 과제를 부여했다. 라벨링 작업에 참여하는 사람에게 '자신이 듣고 싶은 음악을 요청하기' 같은 상황에 맞은 표현을 말하고 녹음해 공유했다. 그 결과 총 녹음 파일 9,434개가 수집됐고 이 데이터는 음성 인식 알고리즘을 훈련시키는데 활용됐다.

68 https://www.definedcrowd.com/

69 정확한 고객명은 밝히지 않음

70 https://www.definedcrowd.com/success-stories/building-a-voice-assistant-model/

두 번째 그룹은 143명으로 구성됐으며, 이들은 녹음파일 내용을 문서로 기록하고 작성된 문서 내용이 실제 오디오 녹음파일 내용과 일치하는지 검수하는 일을 수행했다. 이를 통해 1만8,415개 문서가 만들어졌다. 세 번째 그룹은 주석을 다는 작업을 맡았다. 여기에 75명이 투입됐으며, 이들은 해당 프로젝트로 만들어진 문서와 고객이 제공한 기존 문서를 결합해 각 문장들이 지시하는 핵심 의미를 주석(Data Annotation)으로 작성했다. 예를 들어 '볼륨 조절(adjusting volume)', '외부 소음 차단(noise cancellation)', '식당 찾기(finding a restaurant)' 같은 정보를 추가하는 식이었다. 이 외에도 특정 가수, 음악 장르, 가수 이름 같은 정보를 따로 분류하는 작업을 수행해, 그 결과 2만2,526개 주석을 만들어냈다.

이렇게 만들어진 주석은 다시 태깅 작업으로 연결된다. 90명으로 구성된 네 번째 그룹은 제공된 데이터를 보고 연관성 있는 것끼리 연결하는 작업을 했다 '레이디 가가'라는 데이터에 '가수'라는 태그를 붙여넣는 식이다. 해당 기업은 원래 기존에 있는 모든 데이터에 태그를 했어야 했지만 데이터 수집 및 라벨링 작업을 거치면서, 태그 해야 할 데이터 범위를 좁혀 결과적으로 개발 비용을 줄일 수 있었다고 한다.

한 글로벌 전자제조업체에선 얼굴 인식 기능을 개발하기 위해 디파인드크라우드를 이용했다. 해당 기업은 특히 아시아인으로 구성된 가족 얼굴을 정교하게 인식할 수 있는 기술을 개발 중이었다.[71] 궁극적으로 사진 안에 몇 명의 구성원이 있는지, 누가 부모고 딸인지 아들인지 파악할 수 있는 기술을 개발하고 싶었는데 여기에는 사진 해상도가 640×640픽셀 이상이면서 실내조명이 다양하게 반영된 이미지 데이터가 필요했다.

71 https://www.definedcrowd.com/success-stories/expanding-facial-recognition-models/

디파인드크라우드는 니보에서 가족 50팀을 선정했고 각 팀마다 사진 20장을 찍도록 요청해 총 이미지 1,000장을 확보했다. 또한 사진 속 인물의 나이, 출생지, 가족 관계 정보를 추가하는 과정이 진행됐다. 이 과정을 마치기까지 총 6주가 걸렸으며 기업은 빠른 시간 안에 맞춤화된 데이터를 구축했다고 설명한다.

포르투갈 전력공사인 EDP는 송전 기술 과정을 예측하는 기술을 만들기 위해 디파인드크라우드를 이용했다.[72] EDP에선 과거 송전 시설이나 전봇대 유지 보수할 때는 헬리콥터를 이용해 직접 사진을 찍고 전력 설비들을 검수했으나 현재는 인공지능 기술로 시설물 고장을 미리 예측하는 기술을 개발하고 있다. 여기에 필요한 데이터를 크라우드소싱 방식으로 수집했다.

EDP는 일단 1만2,500개 사진을 찍어 모으고, 디파인드크라우드를 통해 고용한 참여자에게 이미지 데이터 라벨링 작업을 요청했다. 참여자들은 약 7,000개 사진에 전봇대나 송전 시설 위치와 모양을 표시하는 라벨링 작업을 했고 그중 3,500개가 실제 알고리즘 개발 과정에서 활용됐다. 또 사진 900개를 검토하도록 요청해 문제가 있는 전력 설비를 표시하는 이미지 라벨링을 진행했고, 해당 데이터를 기반으로 시설 고장 여부를 파악하는 기술을 구축했다.

72 https://www.definedcrowd.com/success-stories/automating-utilities-inspection/

해외사례를 중심으로

데이터 라벨링으로 만드는 혁신

작성 IT전문기자 **이지현**
한국지능정보사회진흥원 정책본부 정책기획팀 **우창완** 책임

기획 한국지능정보사회진흥원 정책본부 **박원재** 본부장
한국지능정보사회진흥원 정책본부 정책기획팀 **이규엽** 팀장

문의 j.lee.reporter@gmail.com / woo@nia.or.kr

주소 대구 광역시 동구 첨단로 53(우 41068) 한국지능정보사회진흥원
T. 053 230 1114 F. 053 230 1907 www.nia.or.kr

● 이 보고서는 방송통신발전기금으로 수행한 과학기술정보통신부 정보통신·방송연구개발사업 (ICT진흥 및 혁신기반 조성-지식정보사회의 국가발전전략연구 사업)의 결과입니다.