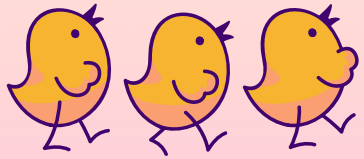
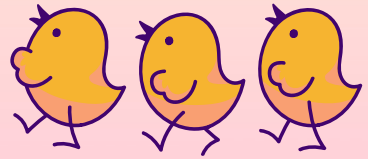


LangCon 2020



자연어처리, 데이터에게 길을 묻다

고품질 코퍼스를 만들어 드립니다

DeepNatural

박상원

코퍼스 구축 스타트업 이야기

1. 자연어처리에서 힘든 점
2. 코퍼스 구축을 쉽게 할 수는 없을까?
3. 고품질 코퍼스를 만들어 드립니다.
4. 코퍼스 구축 사례
5. 코퍼스 스타트업이 주목하는 기술
6. 코퍼스 스타트업이 바라본 트렌드
7. Q&A

자연어처리에서 힘든 점

LangCon 2020

NLP의 흔한 패턴

OOO 연구 / 개발을 해보자 💡

↓

코퍼스 확보부터 해야지 📰

↓

열심히 검색 / 수소문 🔍

↓

구축해야겠네... 😥

↓

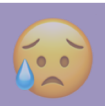
어느새 고생길 진입 🌝

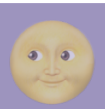
흔한 코퍼스 탐색 과정

ooo 연구 / 개발을 해보자 

코퍼스 확보부터 해야지 

열심히 검색 / 수소문 

구축해야겠네... 

어느새 고생길 진입 

공개된 코퍼스 확보
- 내 연구에 적합한 경우 그대로 활용
- 내 연구에 딱 맞지는 않지만 일단 아쉬운대로 활용

데이터를 보유한 기업 / 기관 찾음
- 비상업적 용도 활용 조건으로 제공 받음
- 유료 구매
- 비용이 비싸서 포기
- 외부 공개가 안됨

가용한 데이터가 없음
- 한국어 코퍼스는 특히 부족..

흔한 코퍼스 구축 고생길

Language Conference

ooo 연구 / 개발을 해보자 💡

↓

코퍼스 확보부터 해야지 📄

↓

열심히 검색 / 수소문 🔍

↓

구축해야겠네... 😥

↓

어느새 고생길 진입 🌕

크롤링 시작
- 잘 만들어진 크롤러가 있는 경우 쉽게 풀림
- 생각보다 크롤러 구현 까다로움

데이터 정제
- 스크립트 짜면 금방 될 줄 알았는데..
- 생각보다 노이즈 패턴이 다양함

어노테이션 도구
- 도구 없이 레이블링은 너무 어려움
- 오픈소스 어노테이션 도구 검색
- 도구 만들기 시작

본격 어노테이션 시작
- 작업자 동원 및 관리의 어려움
- 열심히 만든 레이블셋 / 작업 가이드는 계속 보완하게 됨
- 검수하다보면 말문이 막힐 때가 많음

기타 예상하지 못했던 이슈 발생

자연어처리 = **노가다**가 심한 분야

로 인식되기도 함

목적에 맞는 **고품질 코퍼스** 확보

코퍼스 구축을 쉽게 할 수는 없을까?

Language Conference

대규모 인력을 쉽고 빠르게 동원

수천-수만명의 맨파워를 동원할 수 있는 크라우드소싱 플랫폼

언어 전문가 팀

비전공자가 진행하기 어려운 코퍼스 구축은 인하우스 언어 전문가 팀이 담당

어노테이션 / 검수 도구

작업 유형에 따라 제공된 템플릿을 수정하여 내 프로젝트에 맞는 어노테이션 도구 손쉽게 제작

작업 배분, 취합, 모니터링

프로세스 모니터링 및 관리 시스템

AI Assistant

데이터 구축 공정에 AI 기술을 적용하여 생산성 극대화

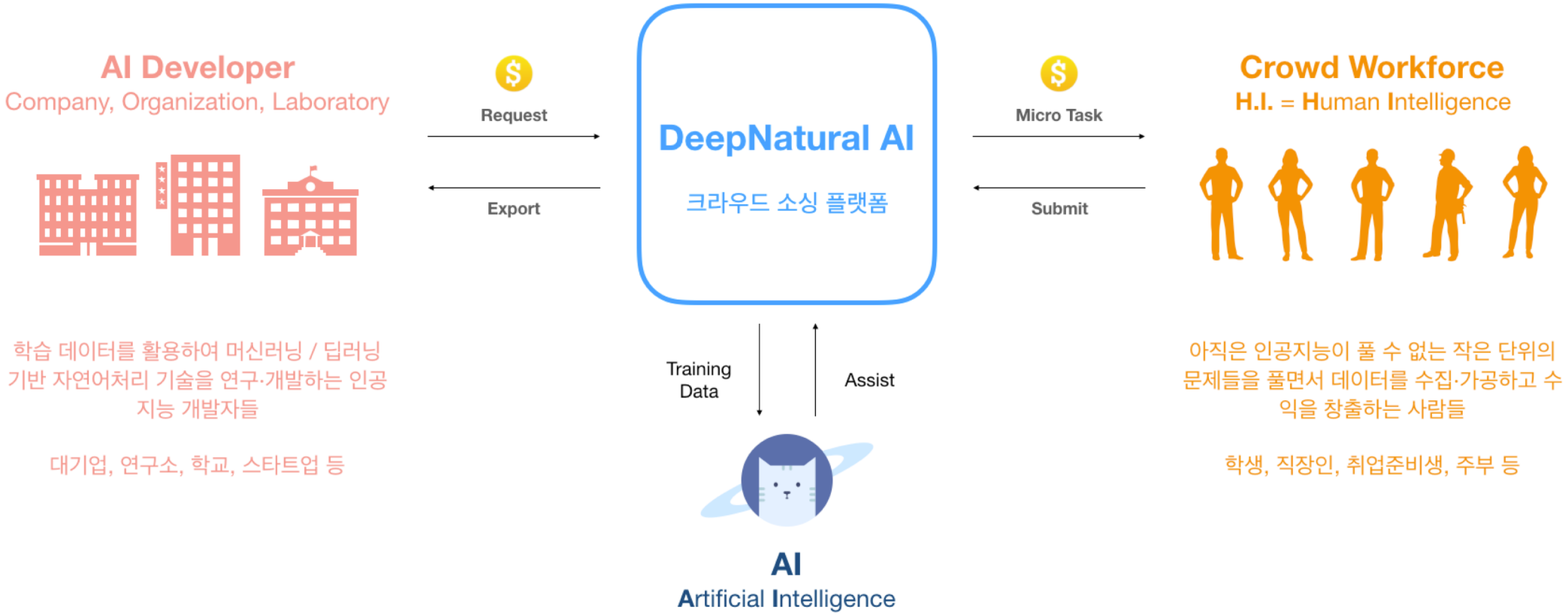
고품질 코퍼스를 만들어 드립니다.

LangCon 2020

HQ Training Corpus as a Service

코퍼스 구축은 저희에게 맡기고 자연어처리 연구에 매진하세요.

크라우드소싱 플랫폼



학습 데이터를 활용하여 머신러닝 / 딥러닝 기반 자연어처리 기술을 연구·개발하는 인공지능 개발자들

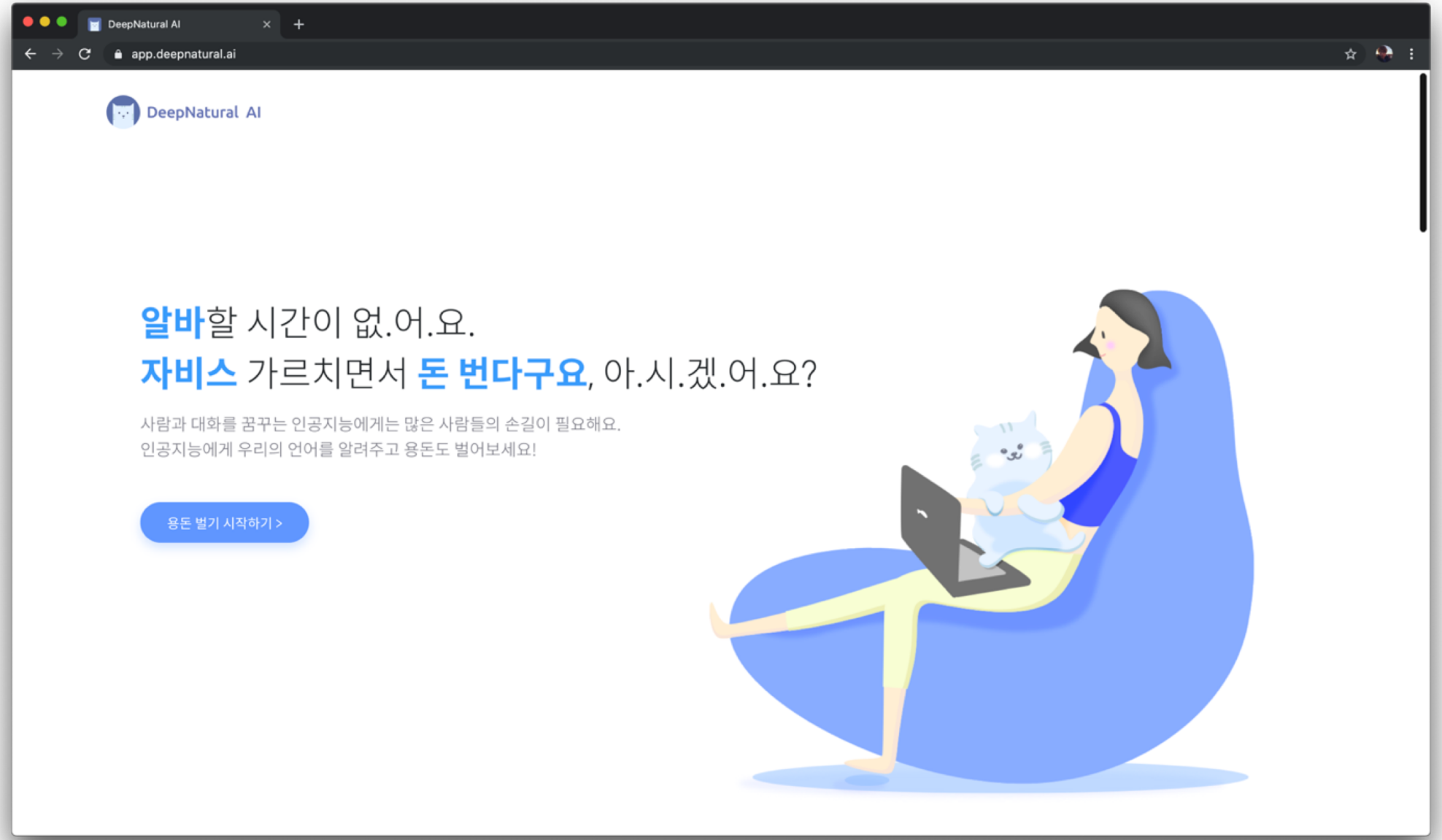
대기업, 연구소, 학교, 스타트업 등

아직은 인공지능이 풀 수 없는 작은 단위의 문제들을 풀면서 데이터를 수집·가공하고 수익을 창출하는 사람들

학생, 직장인, 취업준비생, 주부 등

인공지능에게 우리의 언어를 알려주고 용돈도 벌어보세요!

Language Conference

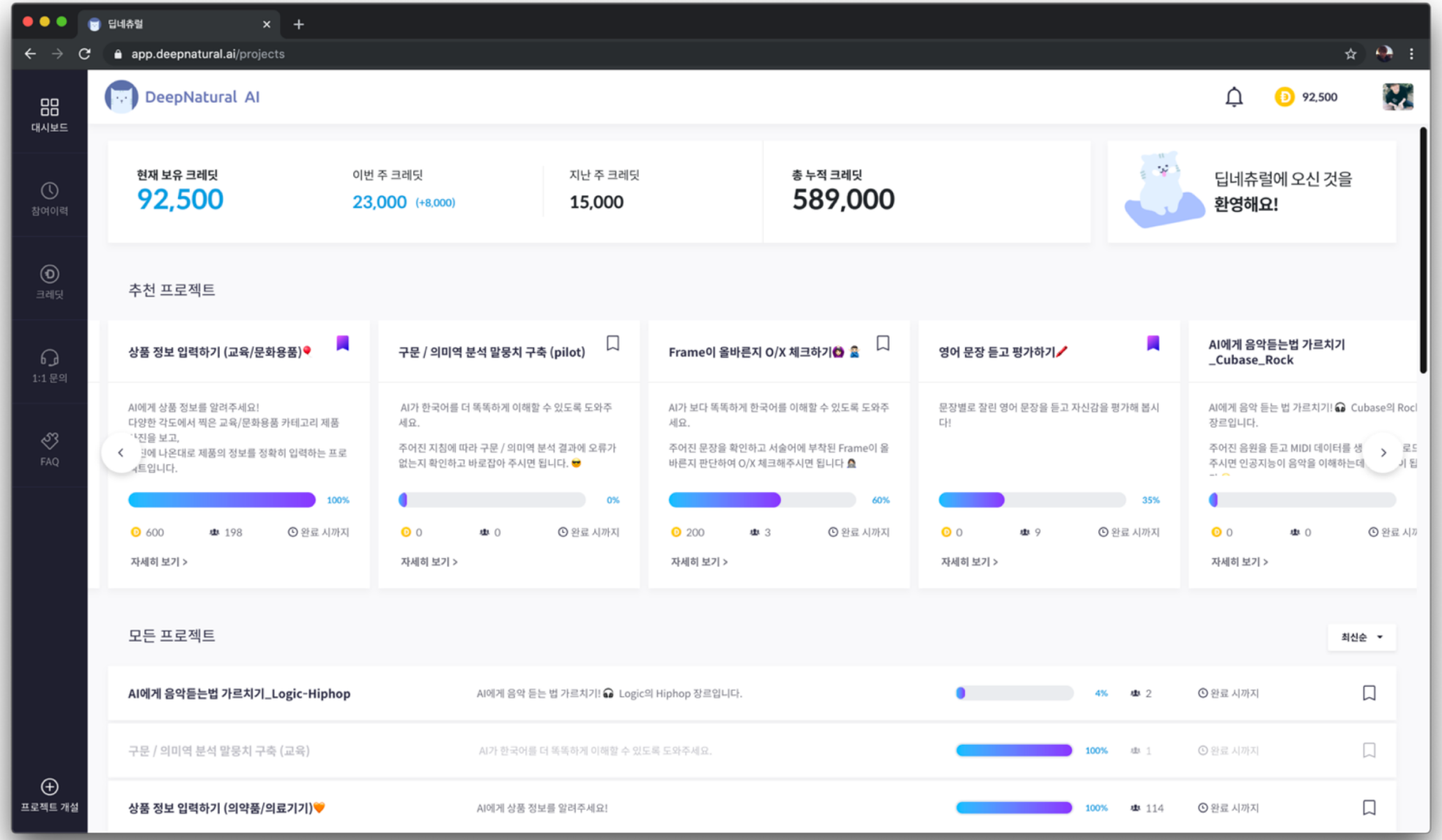


인공지능에게 우리의 언어를 알려주고 용돈도 벌어보세요!

Language Conference

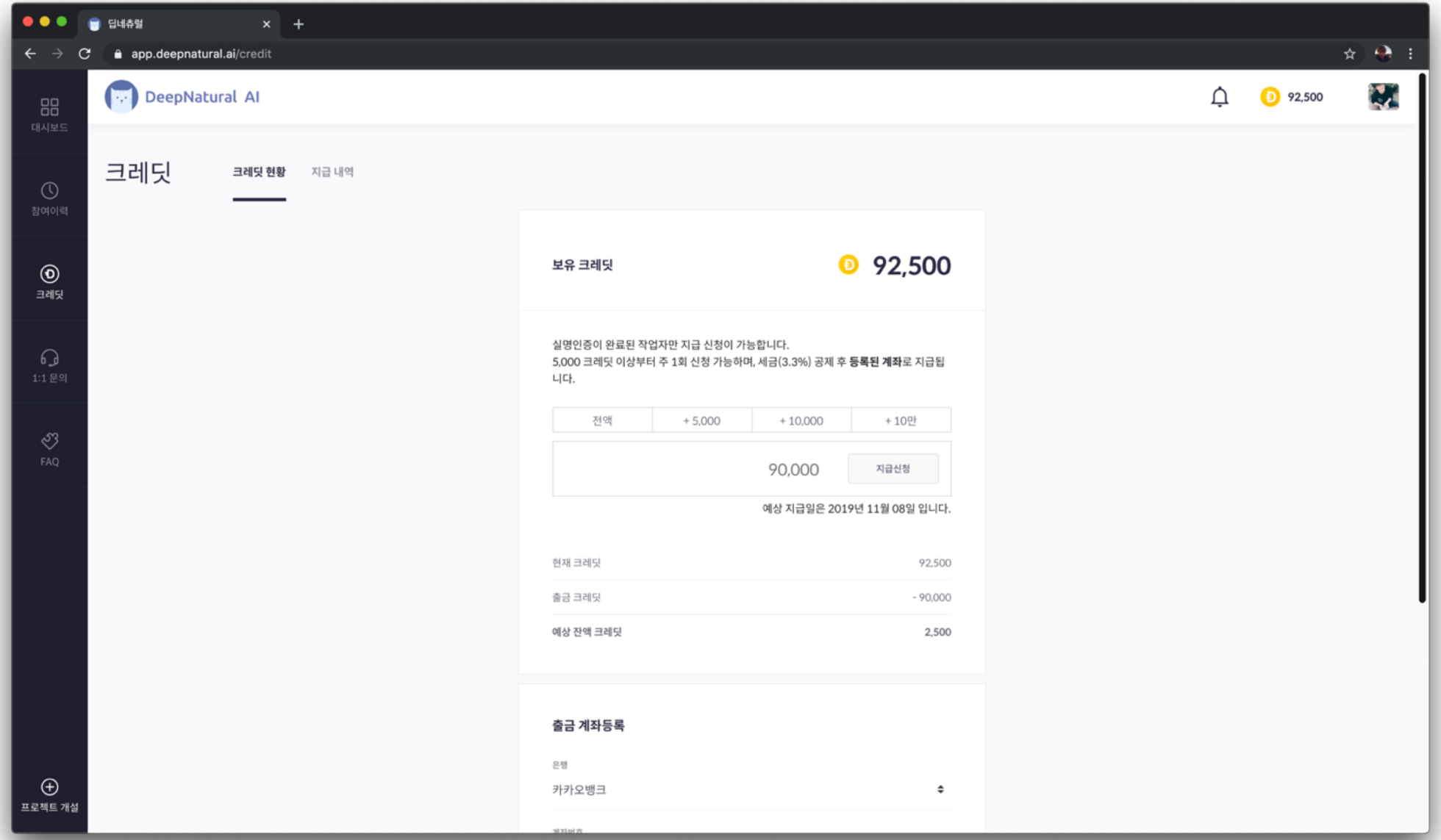
프로젝트 목록

Language Conference

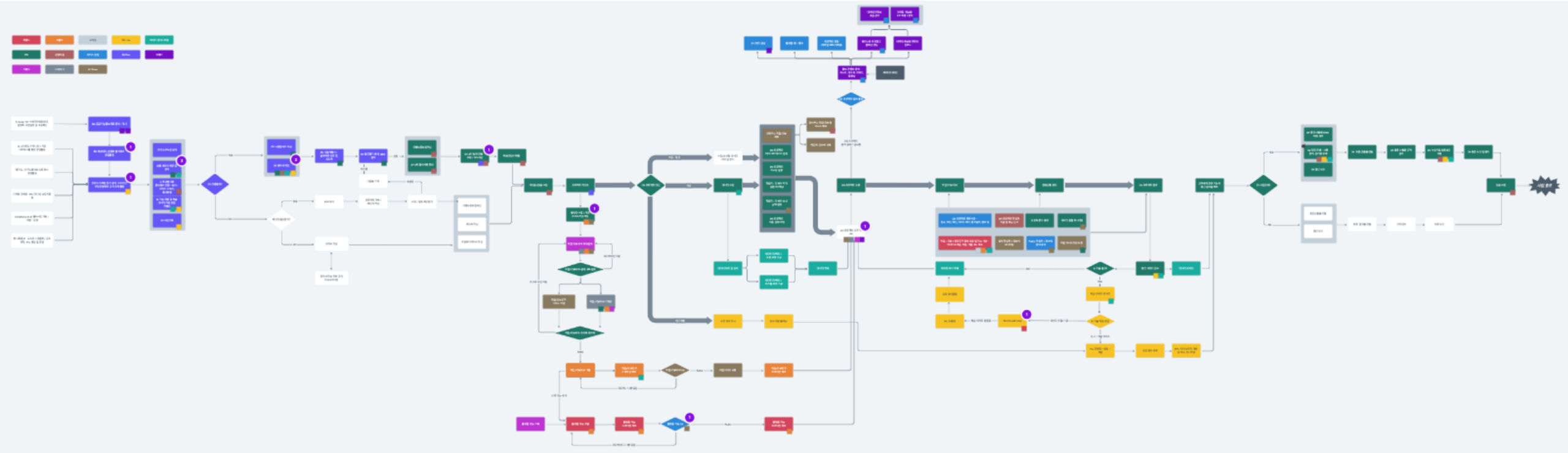


크레딧 출금

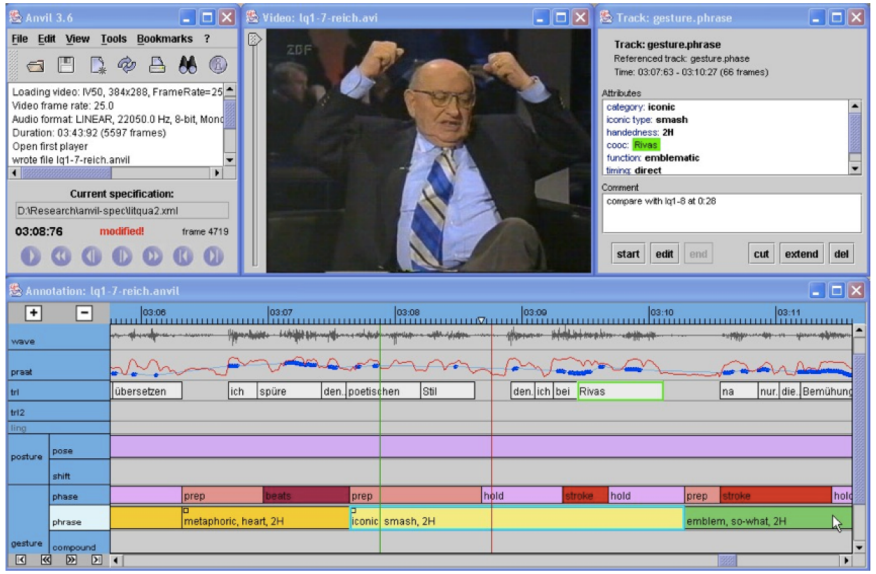
Language Conference



코퍼스 구축 서비스 플로우



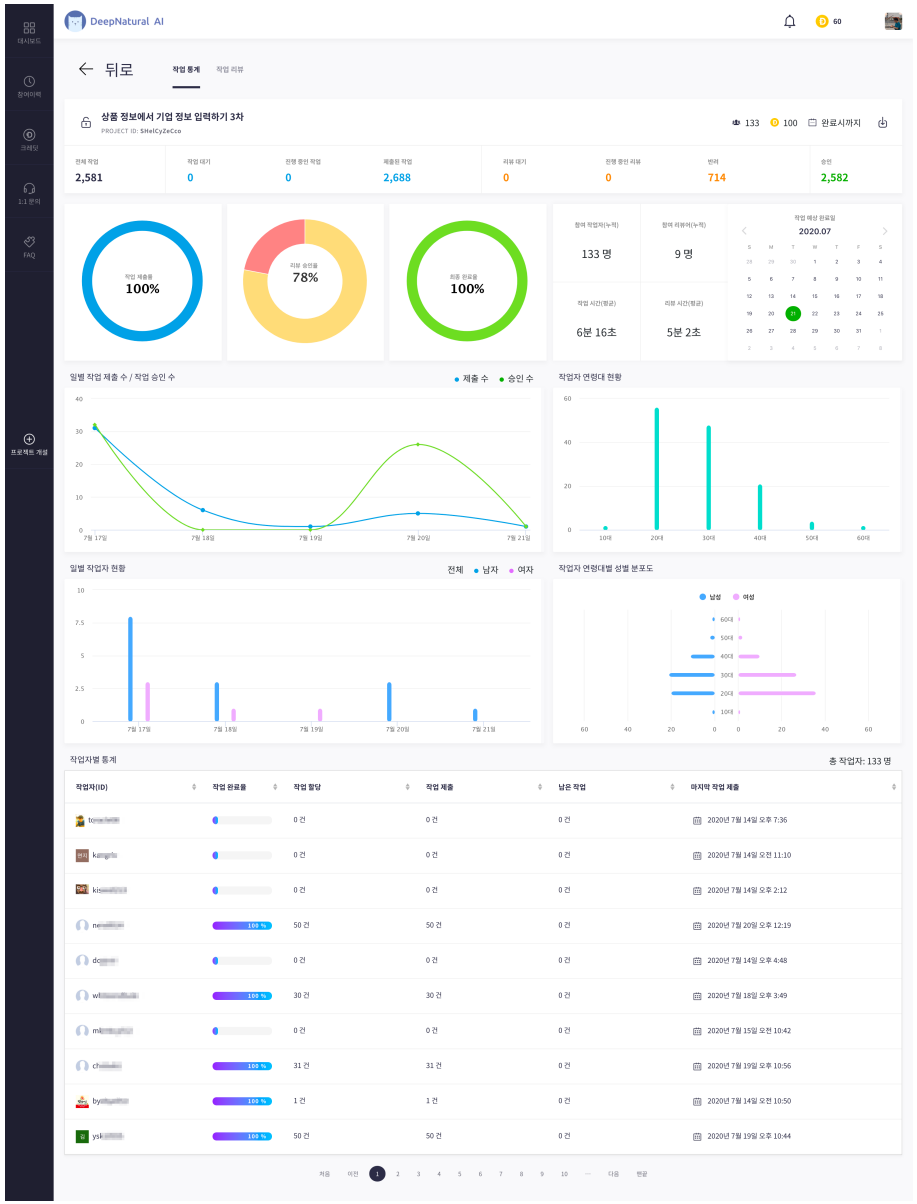
향상된 어노테이션 도구


Anvil 3.6
File Edit View Tools Bookmarks ?
Loading video: lv50, 384x288, FrameRate=25
Video frame rate: 25.0
Audio format: LINEAR, 22050.0 Hz, 8-bit, Mono
Duration: 03:43:92 (5597 frames)
Open first player
wrote file lq1-7-reich.anvil
Current specification:
D:\Research\anvil-spec\litqua2.xml
03:08:76 modified! frame 4719
Video: lq1-7-reich.avi
Track: gesture.phrase
Track: gesture.phrase
Referenced track: gesture.phrase
Time: 03:07:63 - 03:10:27 (66 frames)
Attributes
category: iconic
iconic type: smash
handedness: 2H
cooc: Rivas
function: emblematic
timing: direct
Comment
compare with lq1-8 at 0:28
start edit end cut extend del
Annotation: lq1-7-reich.anvil
wave
praat
tri
übersetzen ich spüre den. poetischen Stil den. ich bei Rivas na nur. die. Bemühung
posture
pose
shift
phase
prep beats prep hold stroke hold prep stroke hold
phrase
metaphoric, heart, 2H
iconic smash, 2H
emblem, so-what, 2H
gesture
compound


DeepNatural AI
Tasks Instruction Community Support
AI Assist ON
Richard $87.25
You guys announced a tech addiction tool that will almost help us limit our screen time.
So what's the thinking behind that?
You know, we've never been focused on usage as a key parameter.
We want people to be incredibly satisfied and empowered by our under the devices that we ship.
But we've never wanted people to spend a lot of time on them. All of their time on them.
And you know, it's a personal thing as to how much is too much.
We thought a lot about this.

고품질 코퍼스를 만들어 드립니다

Language Conference



프로세스 모니터링
및
검수 시스템

코퍼스 구축 사례

LangCon 2020

코퍼스 구축 사례

Language Conference

2019년 베타 서비스 시작부터 지금까지

국립국어원, KB금융지주, KAIST 연구실, 스푼라디오, 띵스플로우, YTN 등과

40+개 이상의 코퍼스 및 학습 데이터 구축 프로젝트 진행

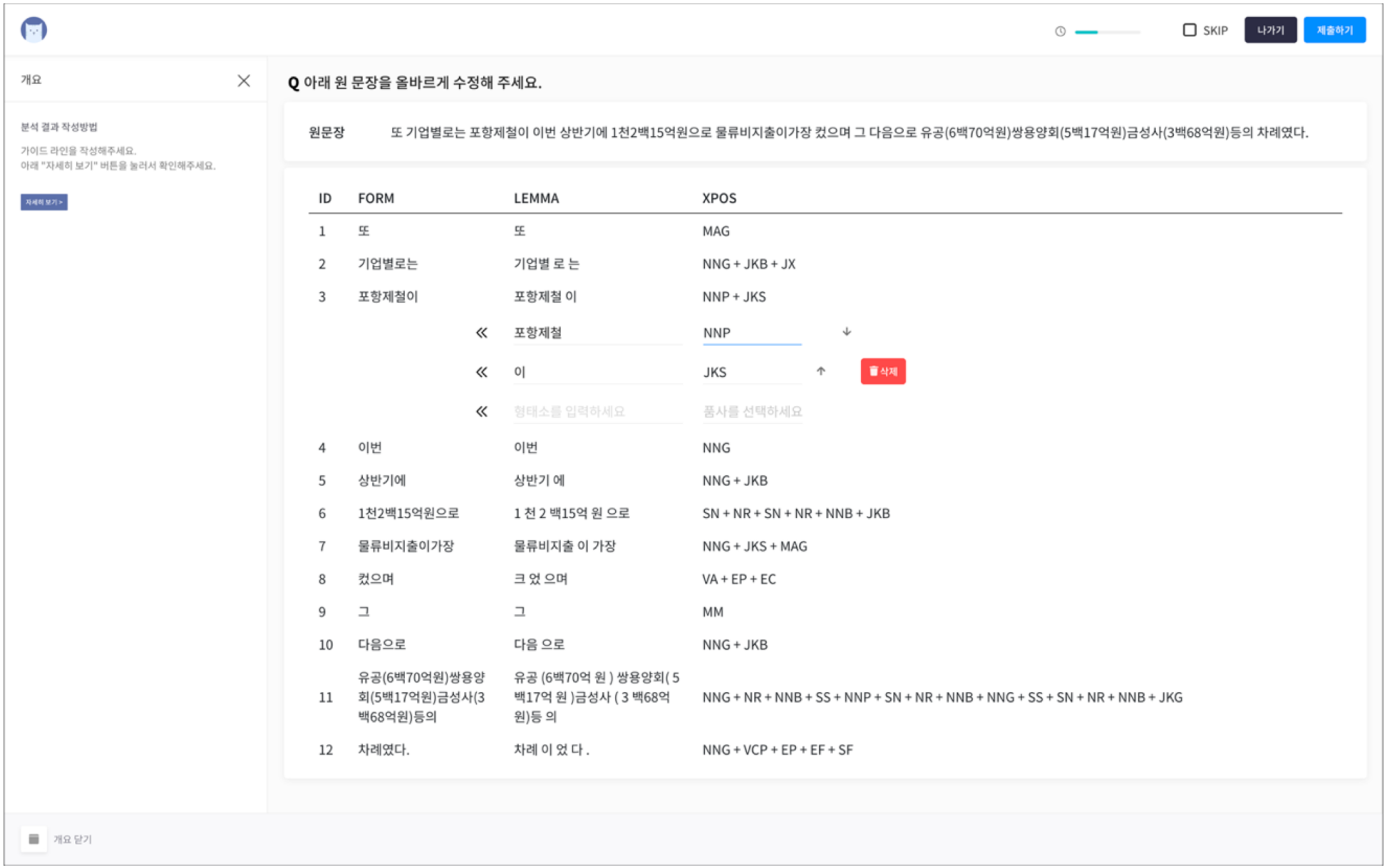
대기업 / 기관 / 스타트업 / 학교 / 연구소

We are trusted by

Bon Angels · KB금융지주 · 국립국어원 · DAESANG · YTN · LABS · Spoon · KAIST · MINDs Lab · 국립암센터 · DMQA · 건양대학교병원 · 연세대학교 산학협력단 · pactera EDGE · IIR TECH · MEGAZONE · acryl · things flow · Kdata · POZA Labs · NIA 한국정보화진흥원 · EVERYOUNG KOREA · EVERYOUNG PEOPLE · KOSA · Vinetech · KAIT · HCLab · Spirink · KOMACA · BLUESIS · Toddlewing · CultiLabs

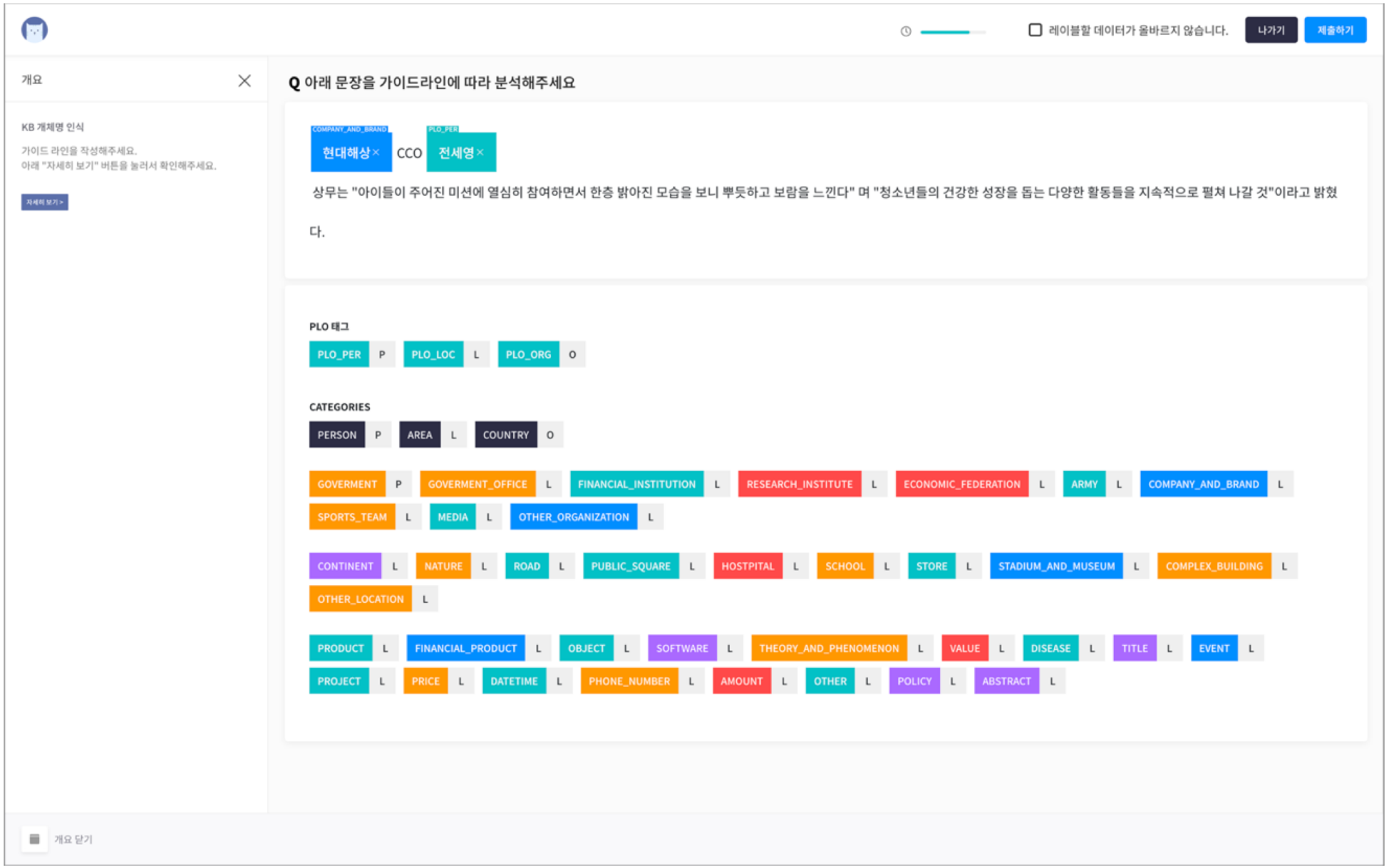
형태소 분석

Language Conference



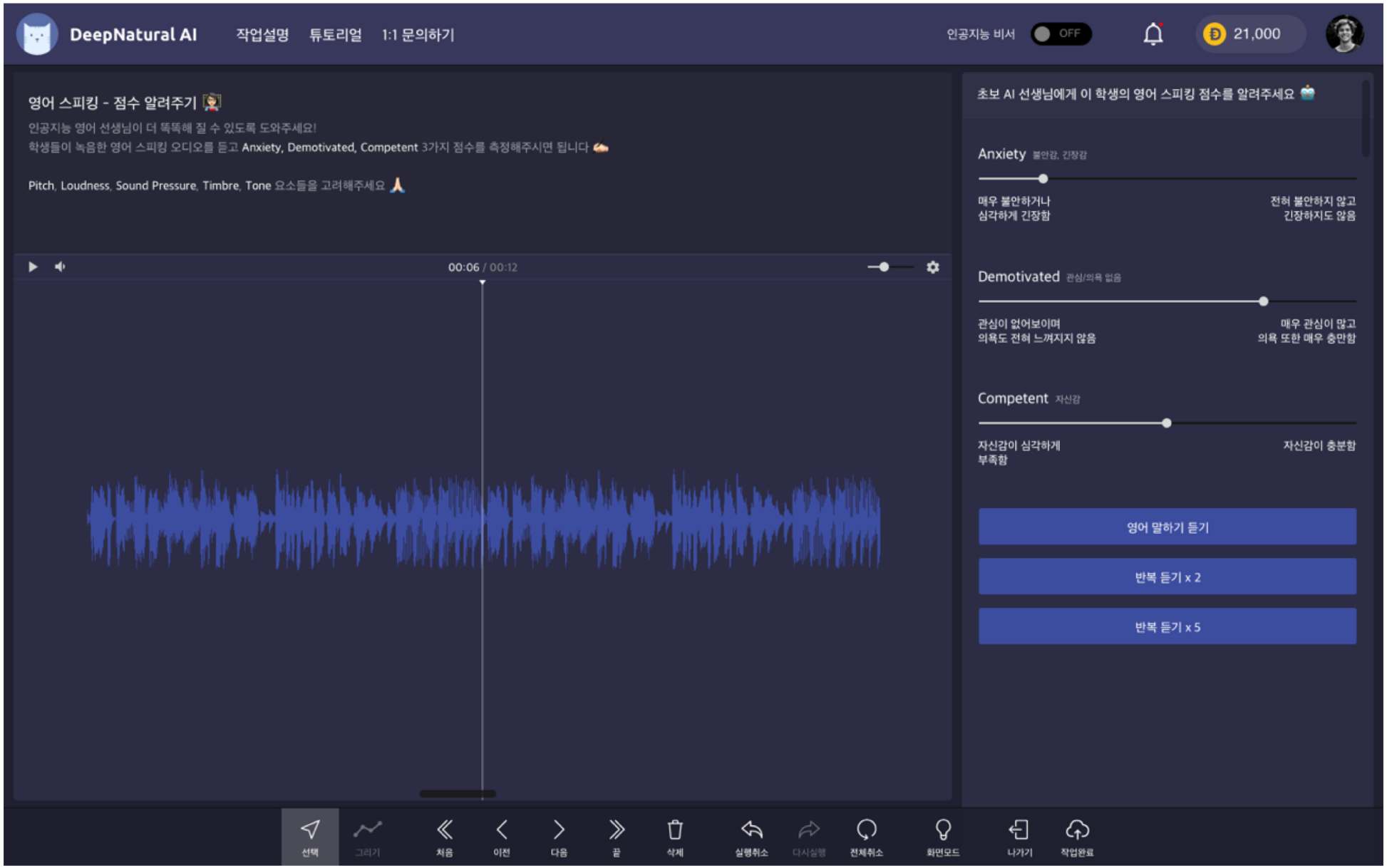
개체명 인식

Language Conference



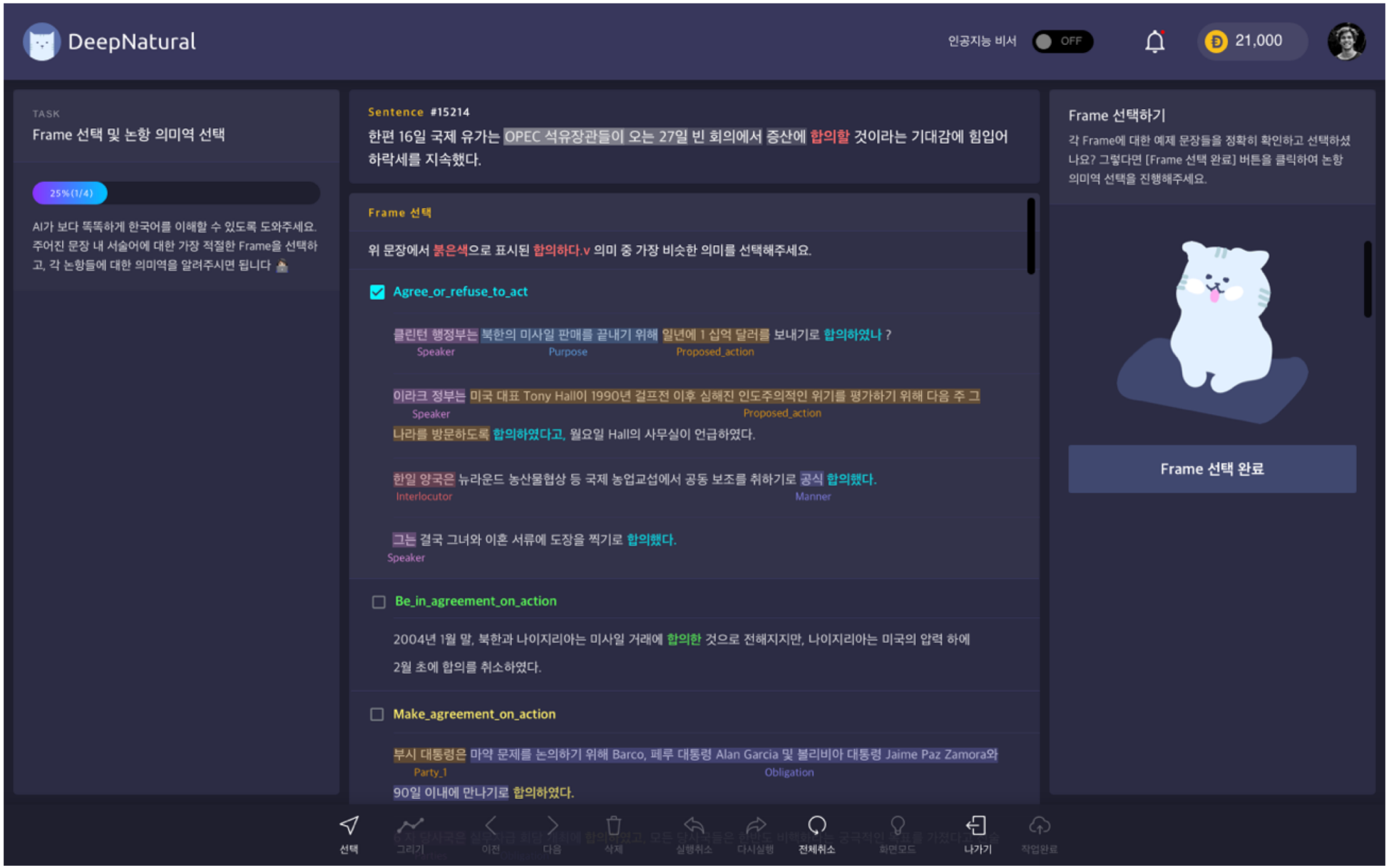
음성 점수 레이블링

Language Conference



서술어 의미 레이블링

Language Conference

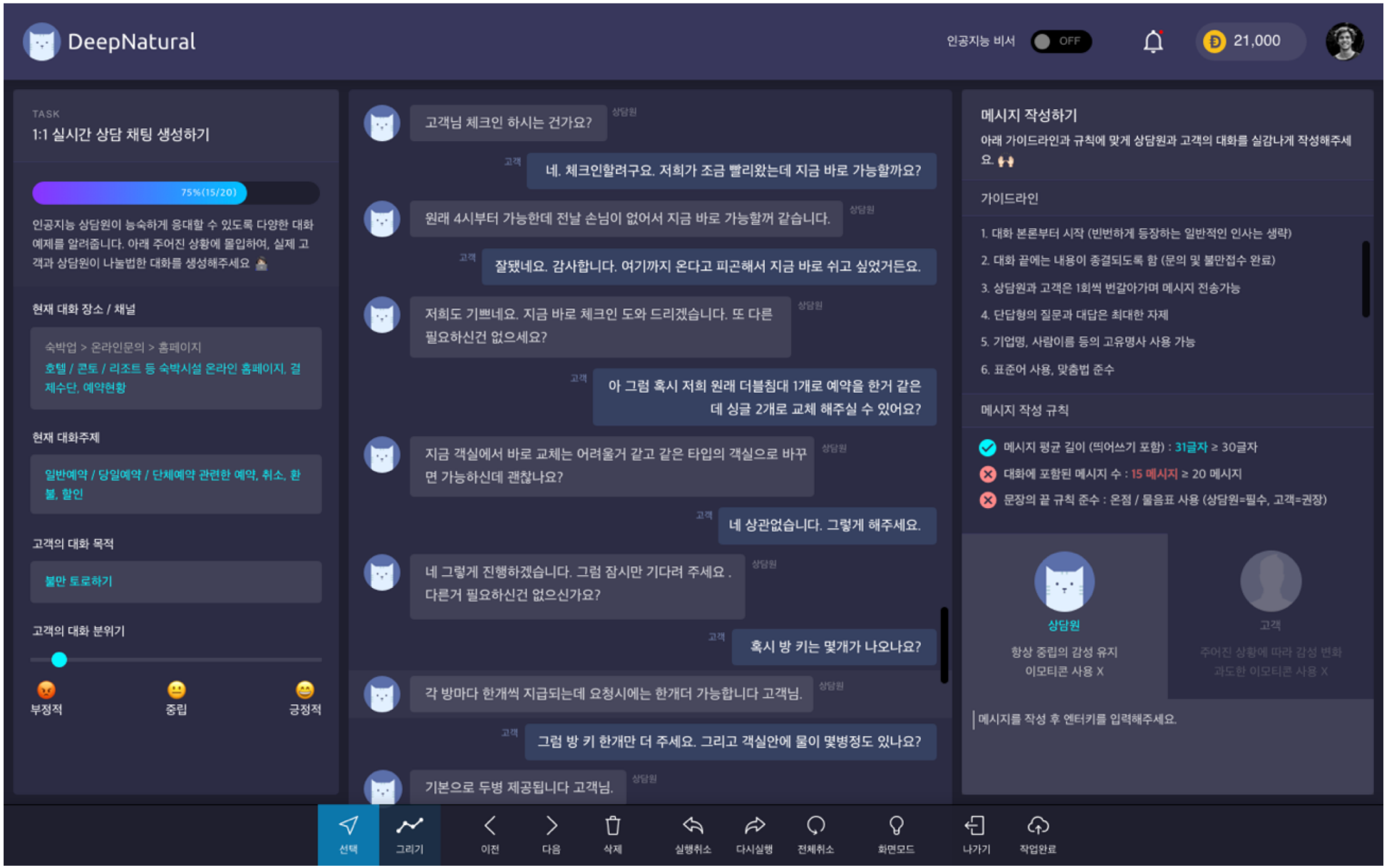


구문 / 의미역 분석

Language Conference

텍스트 대화

Language Conference

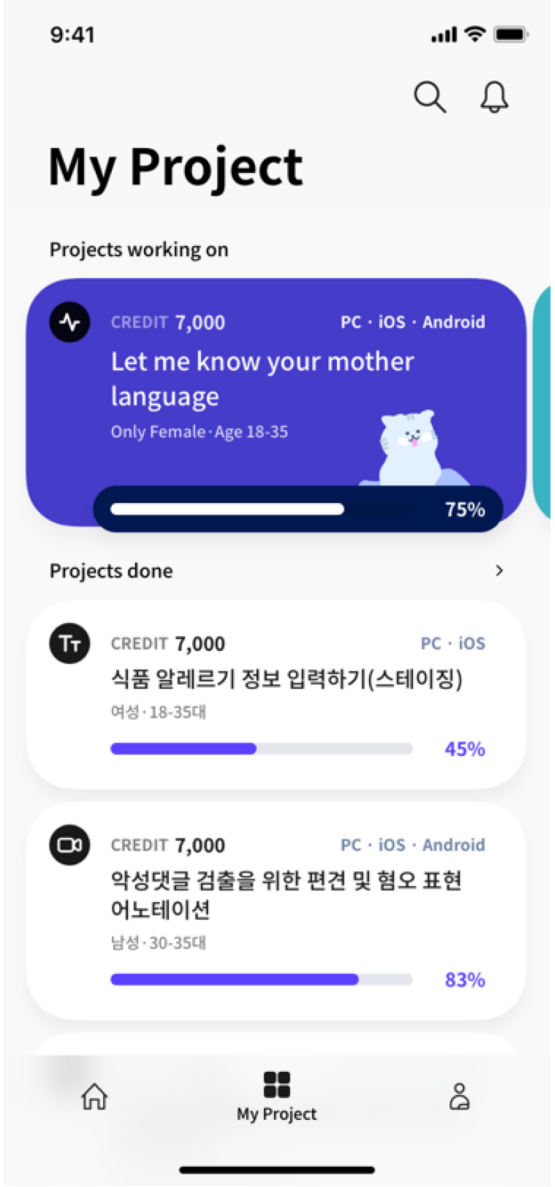


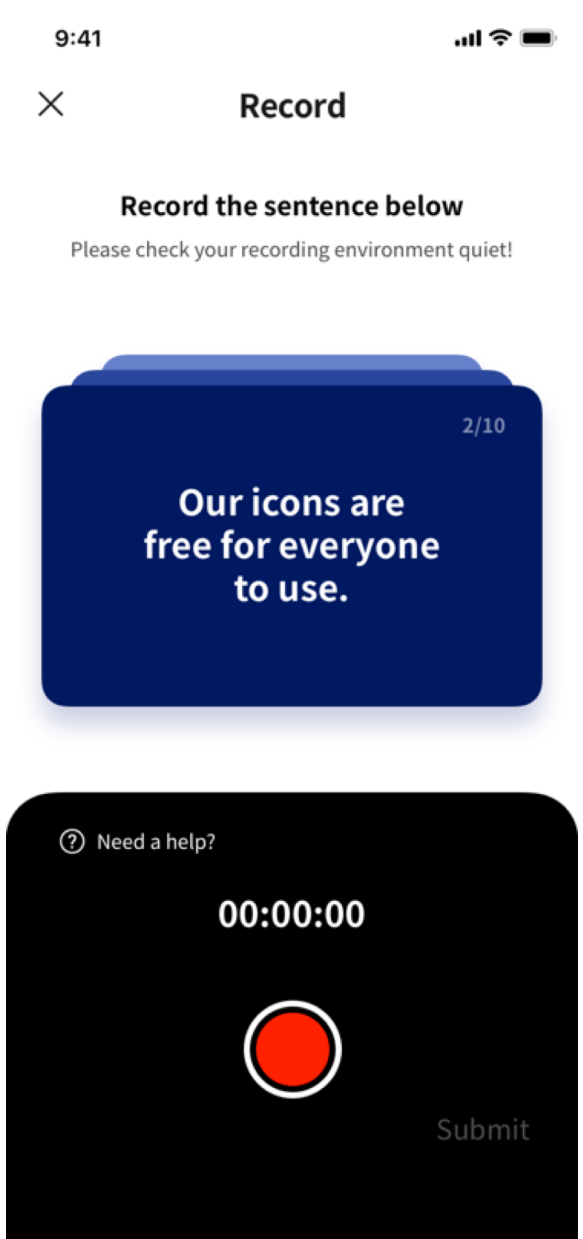
음성 대화

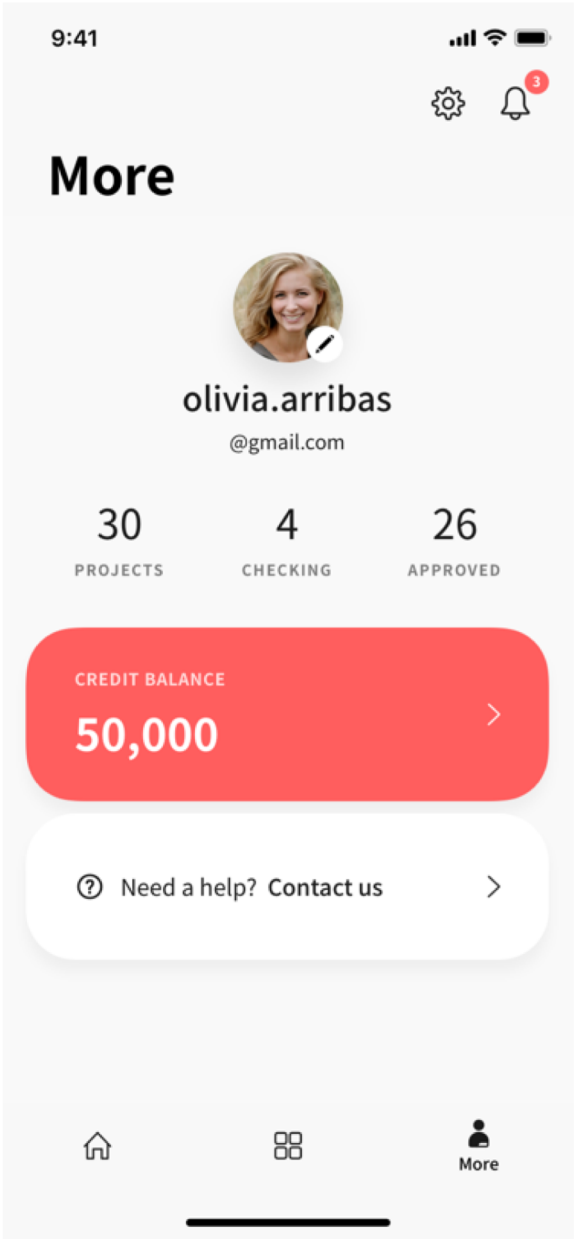
Language Conference

모바일 앱을 통한 데이터 수집 / 가공 (2020년 가을 출시 예정)

Language Conference





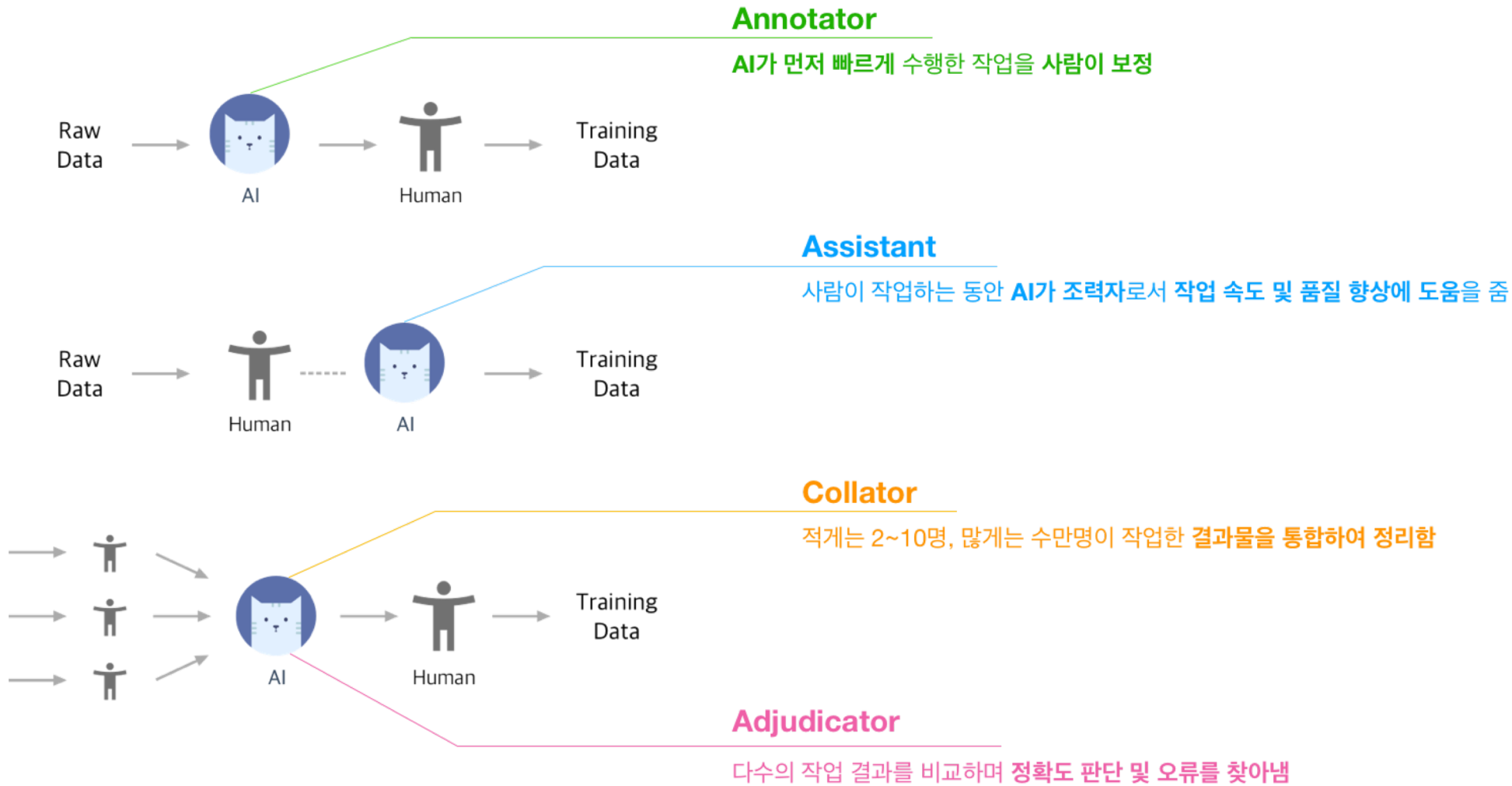


코퍼스 스타트업이 주목하는 기술

LangCon 2020

사람과 인공지능이 협력적으로 코퍼스 구축

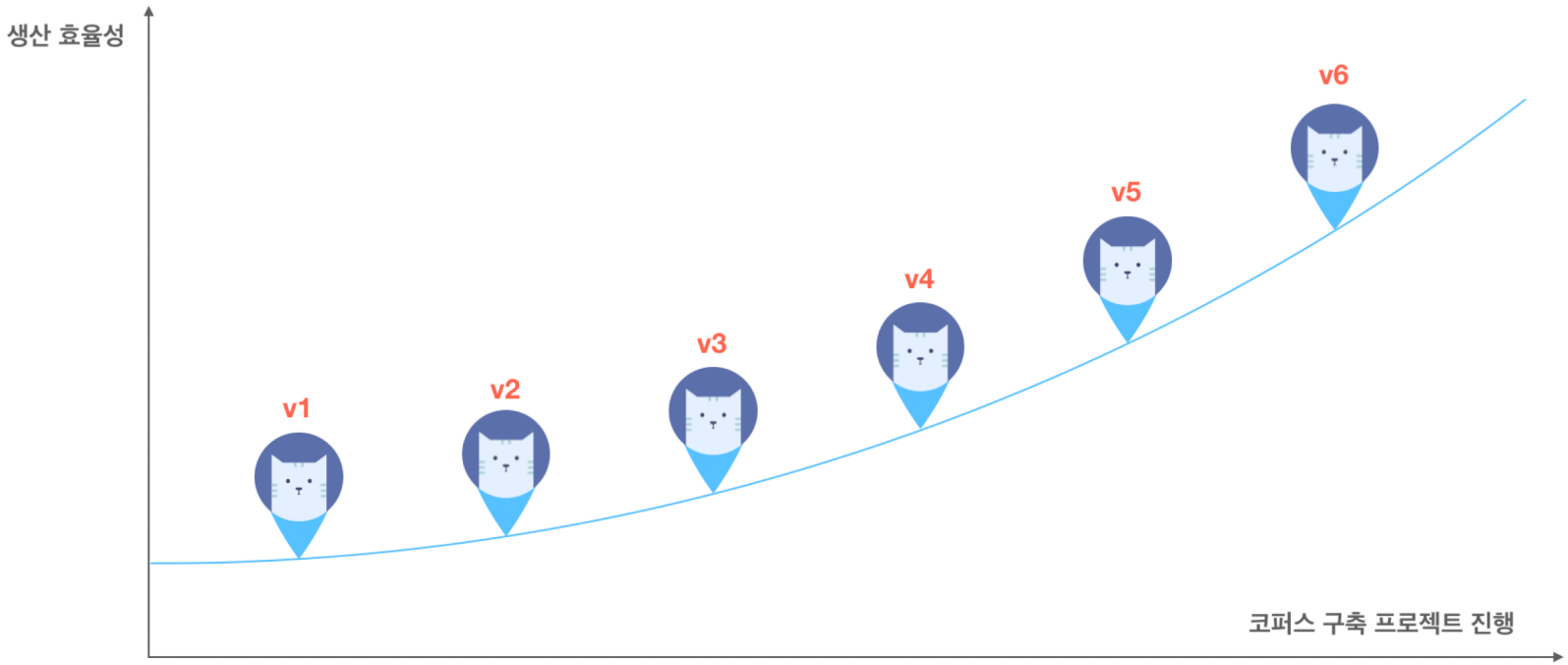
Language Conference



코퍼스 구축 프로젝트가 진행되면서

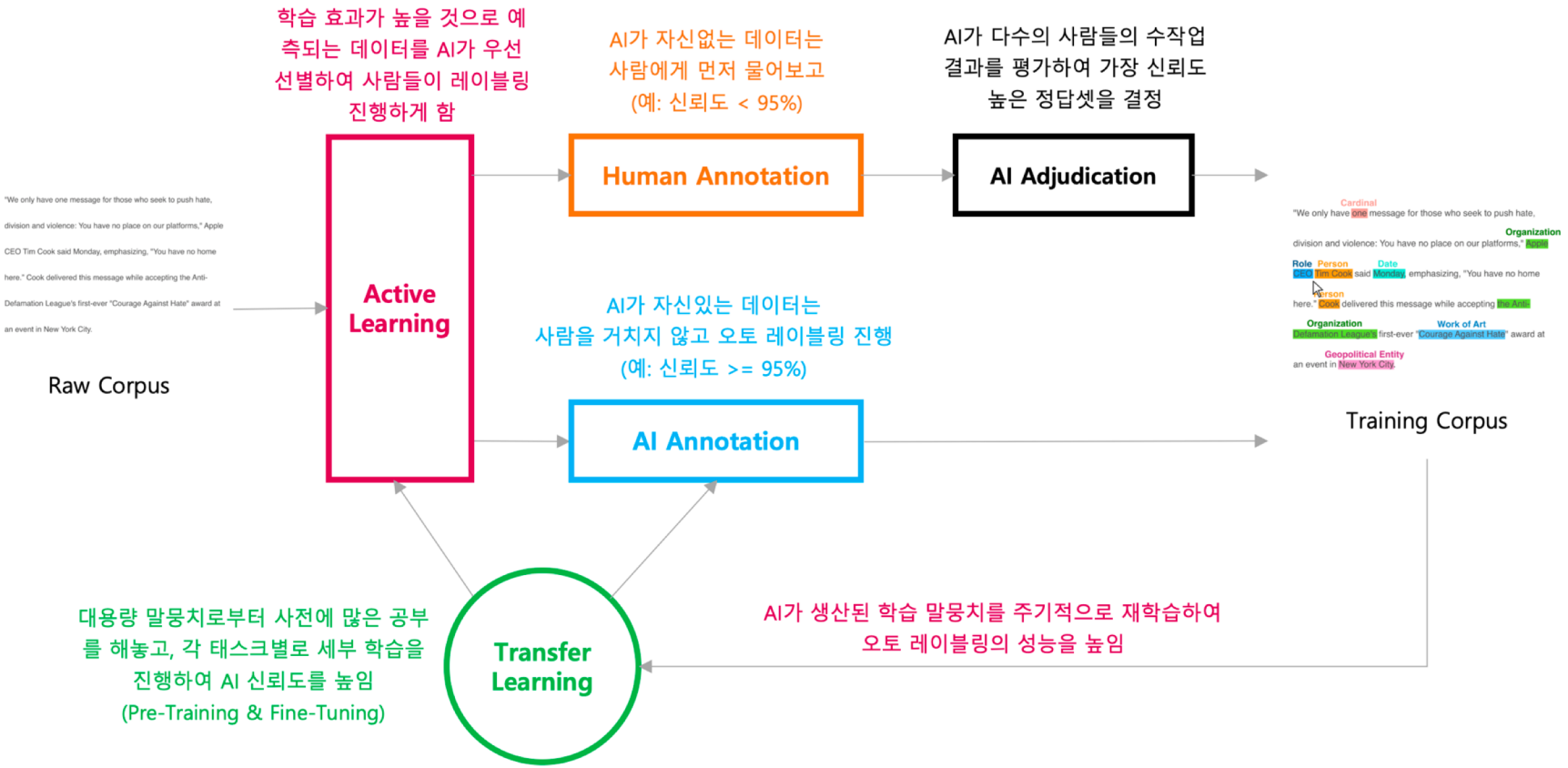
생산되는 학습 데이터를 주기적으로 학습하며

점점 더 똑똑해지는 AI Assistant

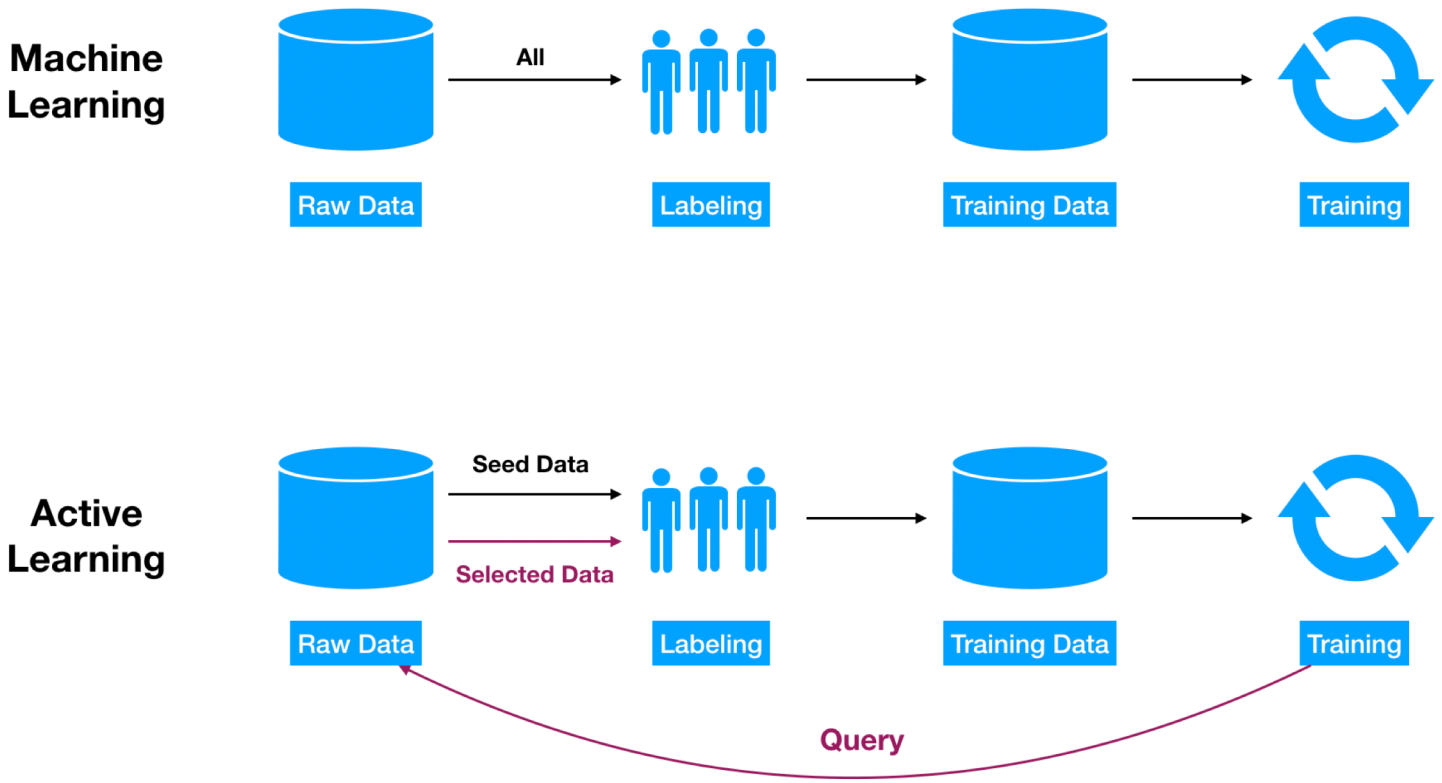


어떻게하면 AI Assistant 의 성능을 **더 빠르게** 높일 수 있을까?

Transfer Learning + Active Learning

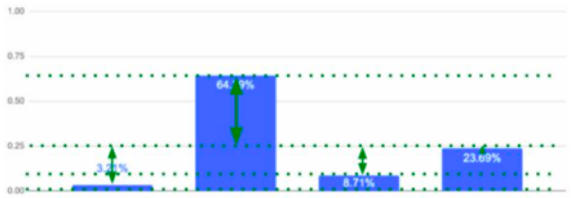


일반적인 기계학습과 **액티브러닝**의 차이는 **학습 과정**에서 **인터렉티브**하게 **쿼리**한다는 것



불확실성이 큰 데이터들을 먼저 학습

예측 결과의 불확실성을 판단하는 방법의 예



Entropy

$$\frac{-\sum_y P_\theta(y \mid x)\, \log_2 P_\theta(y \mid x))}{\log_2(n)}$$

Ratio of Confidence

$$\frac{P_\theta(y^*_2 \mid x)}{P_\theta(y^*_1 \mid x)}$$

Least Confidence

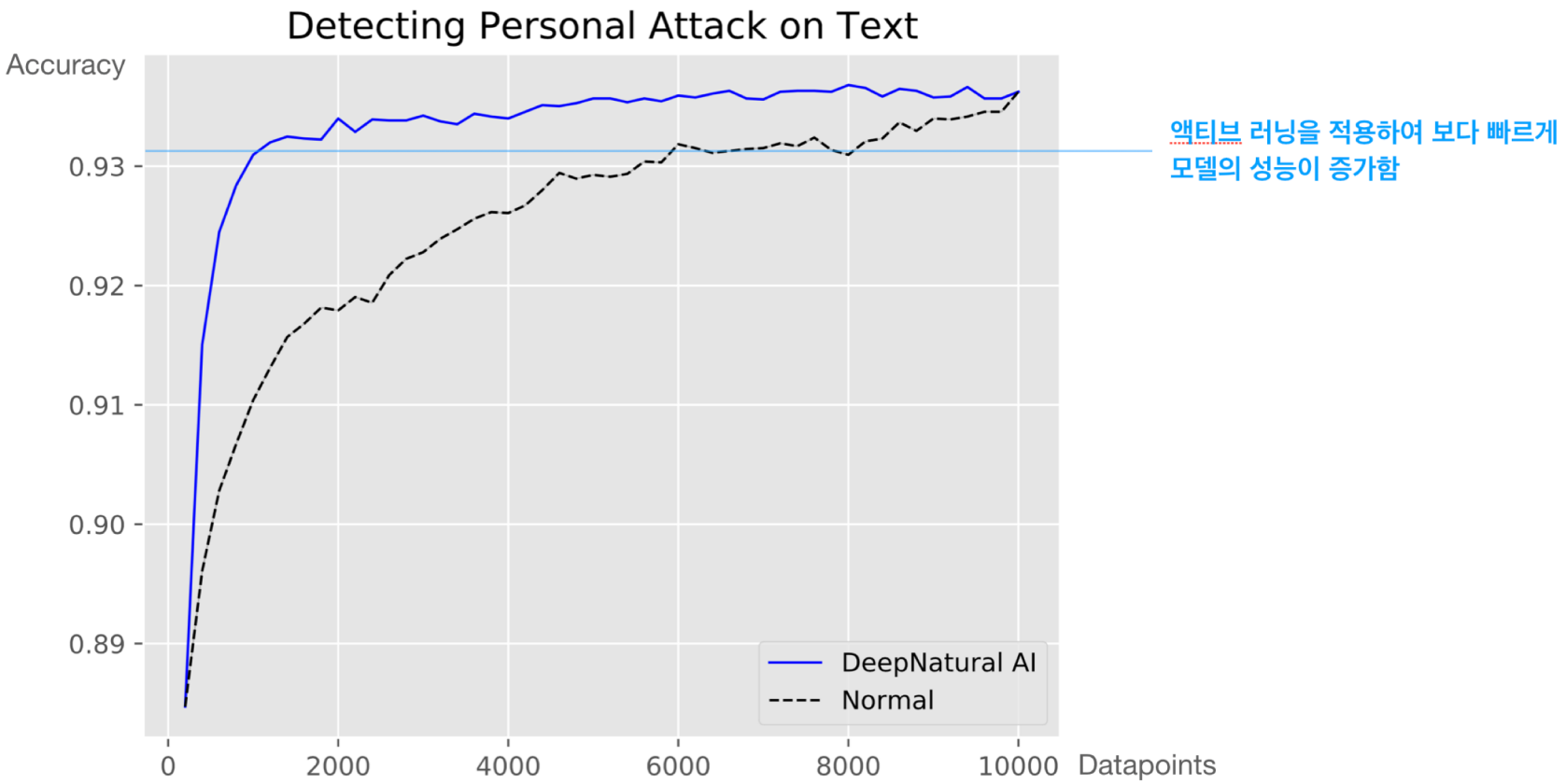
$$\frac{n\,(1 - P_\theta(y^*_1 \mid x))}{n - 1}$$

Margin of Confidence

$$1 - (P_\theta(y^*_1 \mid x) - P_\theta(y^*_2 \mid x))$$

Wikipedia Talk Labels: Personal Attacks 데이터셋에서

액티브러닝을 통한 빠른 성능향상 실험 결과



액티브 러닝을 적용하여 보다 빠르게
모델의 성능이 증가함

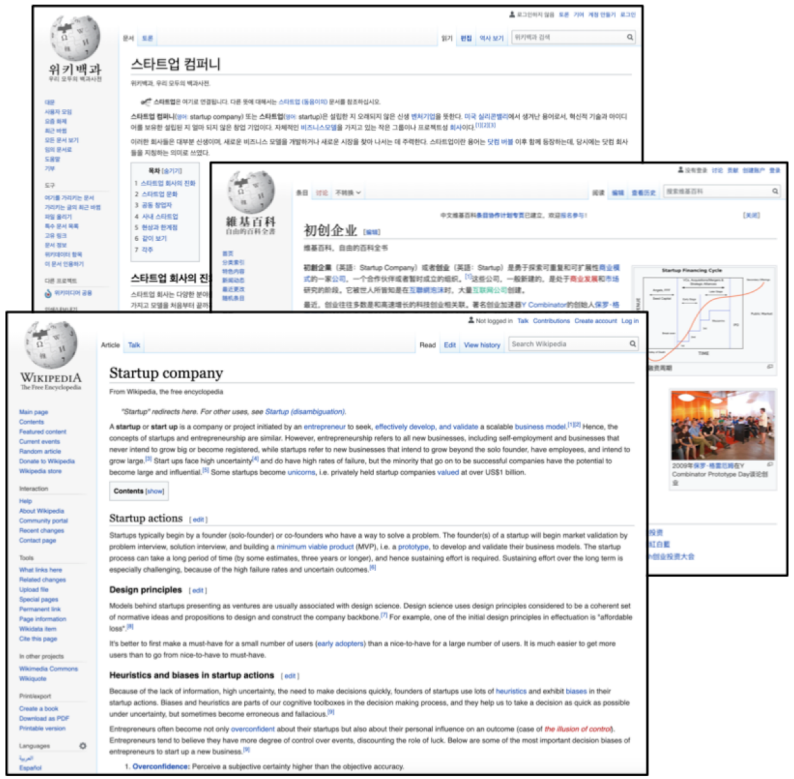
Transfer Learning :: BERT

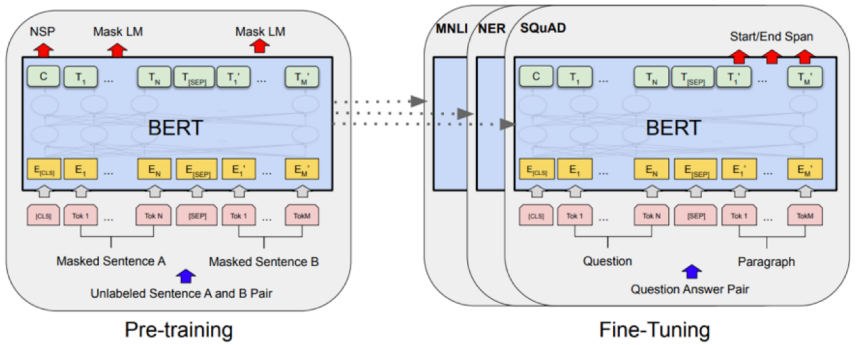
Language Conference

학습 데이터 구축 프로세스에 AI 기술을 적용하기 위해서는

학습 데이터가 전혀 없는 상태에서 시작하여, **소량의 학습 데이터를 이용**하여 AI Assistant의 성능을 빨리 끌어 올리는 것이 필요함

Pre-trained Model을 이용한 Transfer Learning이 중요한 역할

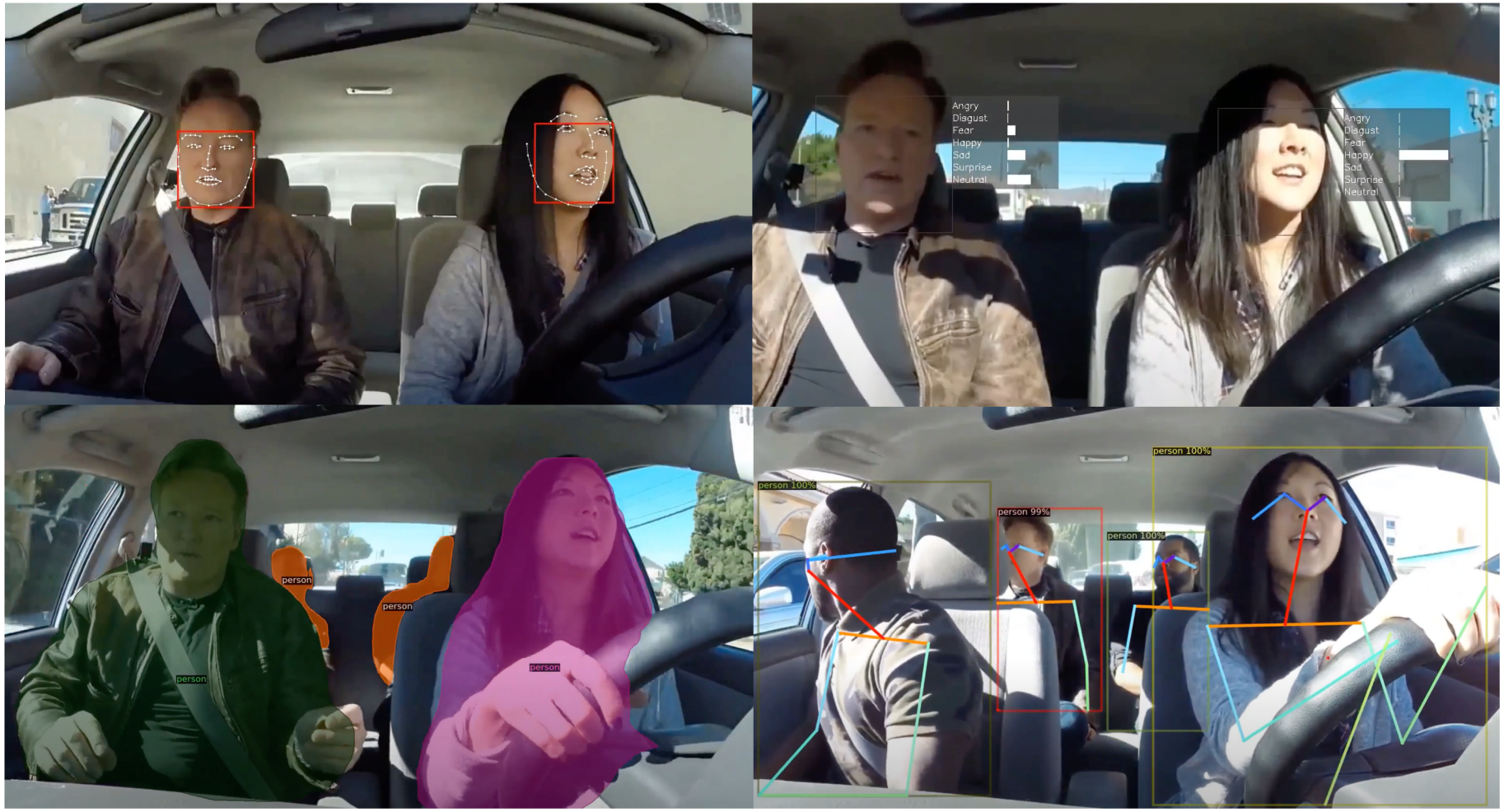




대량의 언어 데이터를 사전에 학습해놓은 PLM을 활용하면
학습 데이터 구축 프로젝트 초기에 **레이블된 데이터가 부족한 상태**에서도
AI Assistant의 성능을 높일 수 있음

자연어처리가 점차 **비언어 커뮤니케이션 영역**으로 확대되지 않을까?

화자 인식, 얼굴 표정, 제스처 등의 신체 움직임, 목소리 볼륨-속도-피치-음색, 공간 등을 포함한 코퍼스 구축



코퍼스 스타트업이 바라본 트렌드

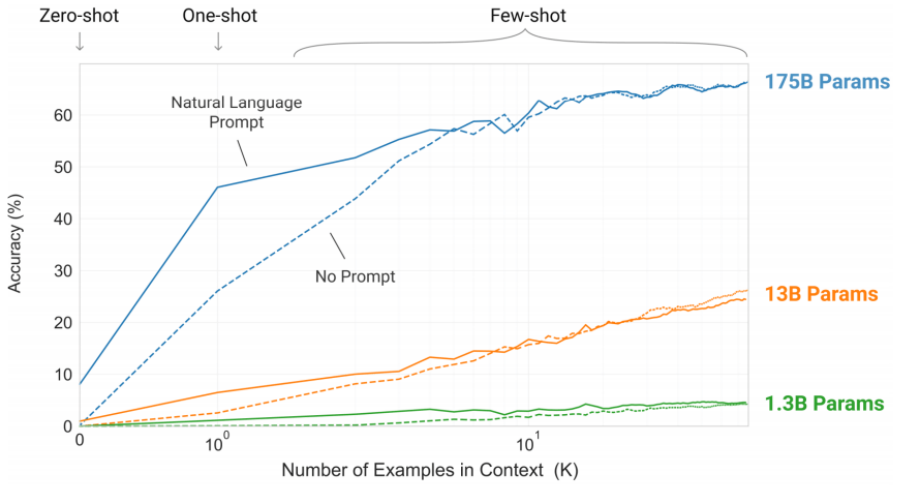
LangCon 2020

BERT

1. 공개된 BERT를 Fine-Tuning 하여 활용
2. 특정 도메인에 최적화된 코퍼스로 BERT를 구움

GPT-3

뜨거운 관심



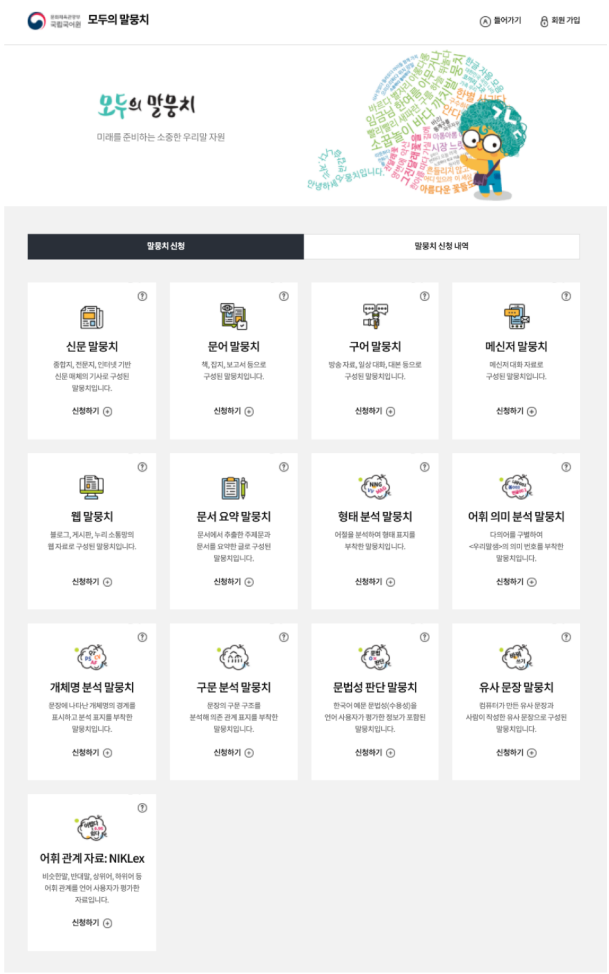
`BERT-Large, Uncased (Whole Word Masking)` : 24-layer, 1024-hidden, 16-heads, 340M parameters

`BERT-Base, Multilingual Cased` : 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters

정부 주도 데이터 구축 사업

Language Conference

국립국어원 모두의 말뭉치



https://corpus.korean.go.kr/

K-Data 데이터바우처

학습 데이터 구축하는 중소기업에게

7천만원 지원

NIPA AI 바우처

AI 솔루션 구입 / 활용, 학습 데이터 구축에

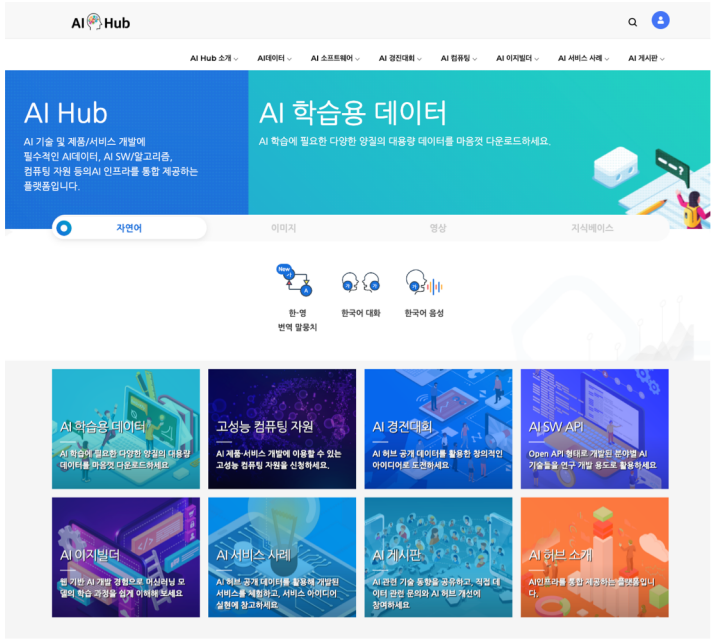
최대 3억원 지원

데이터 공개 x
데이터 소유권은 수요기업에게 있음

NIA 인공지능 학습용 데이터 구축 사업

2020년 상반기 400억 규모 데이터셋 구축 시작

2020년 하반기 3,000억 규모 데이터셋 구축 시작



http://www.aihub.or.kr/

정부 주도 데이터 구축 사업

NIA 인공지능 학습용 데이터 구축사업(2020년 하반기 추경)으로 내년에 공개될 자연어 분야 학습데이터 (지정과제)

- 총 132개 세부 데이터셋 (약 2,500억 규모)
- 자연어 분야는 35개 세부 데이터셋 (약 665억 규모)

영역	분야	소분류	번호	세부 데이터명
자연어 (35)	1	자유대화 (4)	1	자유대화(일반남여)
			2	자유대화(노인남여)
			3	자유대화(소아남여, 유아 등 혼합)
			4	한국인 외래어 발화
	2	명령어 (4)	5	명령어 데이터(일반남여/정형·비정형 포함)
			6	명령어 데이터(노인남여/정형·비정형 포함)
			7	명령어 데이터(소아남여,유아/정형·비정형 포함)
			8	차량 내 대화 및 명령어 데이터
	3	상황별음성 (4)	9	한국어 강의 데이터
			10	회의 음성 데이터
			11	고객 응대 데이터
			12	상담 음성 데이터
	4	열악환경 음성 (3)	13	비대면 환경 대화 데이터
			14	열악한 환경 음성 데이터
			15	다화자 음성 합성 데이터
	5	한국어 방언 (5)	16	한국어 방언 발화 데이터(강원도)
			17	한국어 방언 발화 데이터(경상도)
			18	한국어 방언 발화 데이터(전라도)
			19	한국어 방언 발화 데이터(제주도)
			20	한국어 방언 발화 데이터(충청도)
	6	요약 데이터 (3)	21	논문자료 요약 데이터
			22	도서자료 요약 데이터
			23	한국어 대화 요약 데이터
	7	한국어 텍스트 (4)	24	도서자료 기계독해
			25	민원(콜센터) 질의-응답 데이터
			26	전문분야 말뭉치 데이터(분야별 개체명 인식 포함)
			27	한국어 SNS 데이터
	8	영어번역말뭉치 (2)	28	한국어-영어 1 번역 말뭉치
			29	한국어-영어 2 번역 말뭉치
	9	중국어-일본어 번역 말뭉치 (3)	30	한국어-중국어 번역 말뭉치 1
			31	한국어-중국어 번역 말뭉치 2
			32	한국어-일본어 번역 말뭉치
	10	OCR (3)	33	야외 실제 촬영 한글 이미지
			34	다양한 형태의 한글 문자 이미지 인식 데이터
			35	공공행정문서 OCR

AI 시대 일자리의 변화

Language Conference

플랫폼 노동 ::

AI를 가르치는 데이터 구축 일자리 확대

컴퓨터와 인터넷만 있으면
누구나 언제 어디서나 데이터 구축에
참여하고 돈을 벌 수 있음

일자리 패러다임이 변화하고 있음

Join Us

Language Conference

딥네츄럴은

- 자연어처리 연구자들의 든든한 파트너로
- 사람의 지식을 인공지능에게 전달하는 일을 하며
- AI 밸류 체인에서 언어 데이터 구축에 포지셔닝하여 빠르게 성장하고 있습니다.

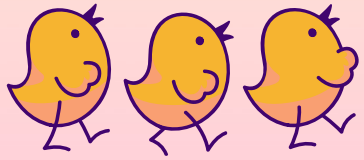
함께 성장할 동료를 찾고 있어요:

- 자연어처리 / 머신러닝 엔지니어 (AI Assistant)
- Linguist 언어 전문가
  - 국어 전공자. 형태, 구문, 의미 분석 말뭉치 구축 경험 등
- 데이터 엔지니어
  - 데이터 전/후처리
- 기타
  - 백엔드, 웹프론트엔드, 모바일, 프로덕트 디자이너, 데이터 구축 프로젝트 매니저, PO, ETC

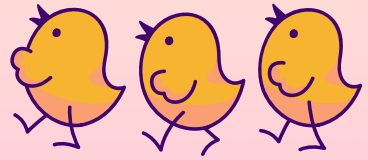
편하게 연락주세요

DeepNatural 박상원 대표 / anson@deepnatural.ai

Q&A



LangCon 2020



자연어처리, 데이터에게 길을 묻다

'고품질 코퍼스를 만들어 드립니다'

감사합니다!

DeepNatural

박상원 / CEO / anson@deepnatural.ai