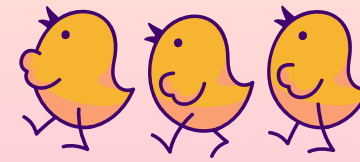


LangCon 2020



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

000 연구 / 개발을 해보자 💡



코퍼스 확보부터 해야지 📄



열심히 검색 / 수소문 🔍



구축해야겠네... 😓



어느새 고생길 진입 😊

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..

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 = **노가다가** 심한 분야

로 인식되기도 함

목적에 맞는 고품질 코퍼스 확보

대규모 인력을 쉽고 빠르게 동원

수천-수만명의 맴파워를
동원할 수 있는 클라우드소싱
플랫폼

언어 전문가 팀

비전문가가 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.

HQ Training Corpus as a Service

코퍼스 구축은 저희에게 맡기고 자연어처리 연구에 매진하세요.

클라우드소싱 플랫폼

AI Developer
Company, Organization, Laboratory



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

대기업, 연구소, 학교, 스타트업 등



Request



DeepNatural AI

클라우드 소싱 플랫폼

Export



Micro Task



Crowd Workforce
H.I. = Human Intelligence



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

학생, 직장인, 취업준비생, 주부 등

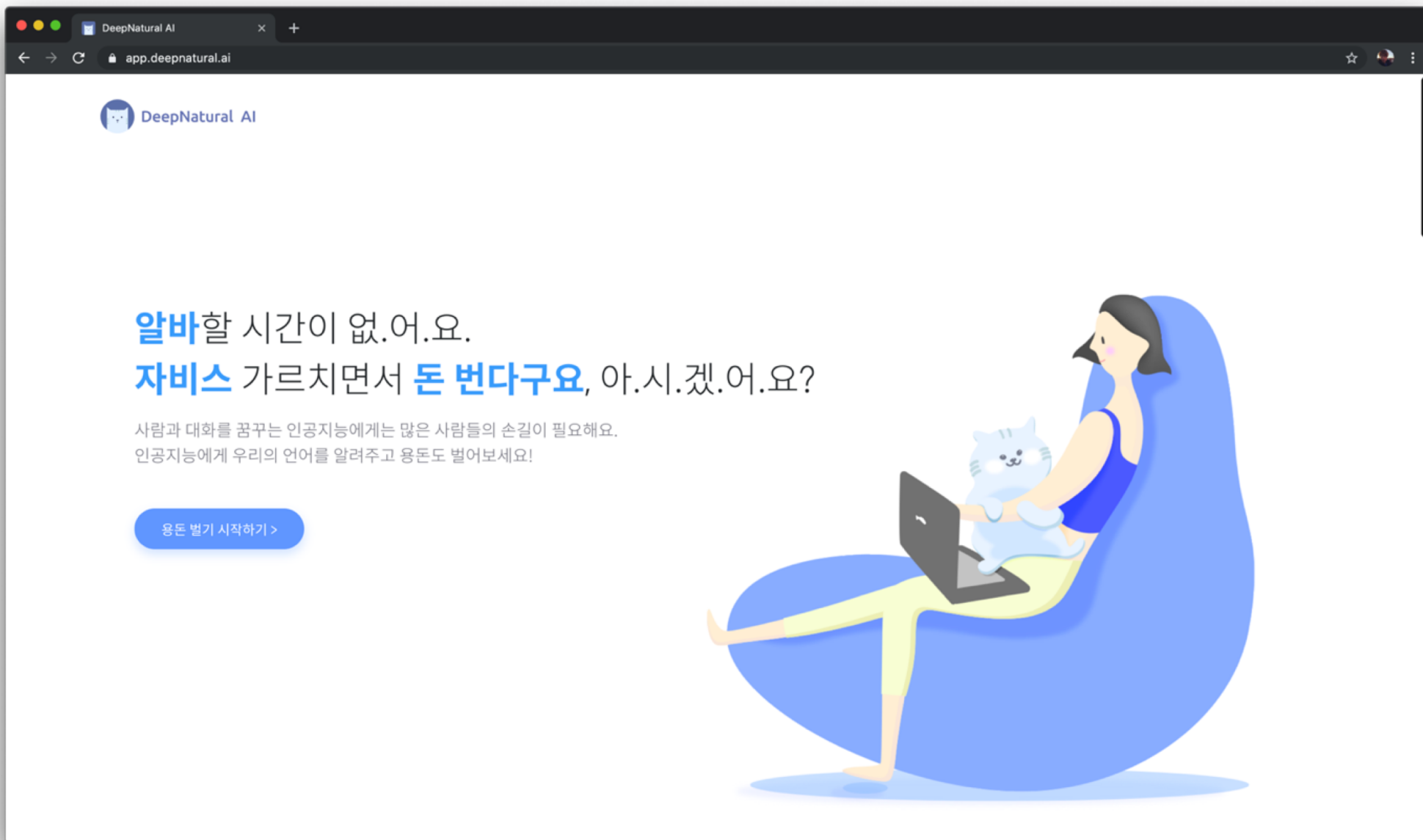
Training
Data



AI
Artificial Intelligence

Assist






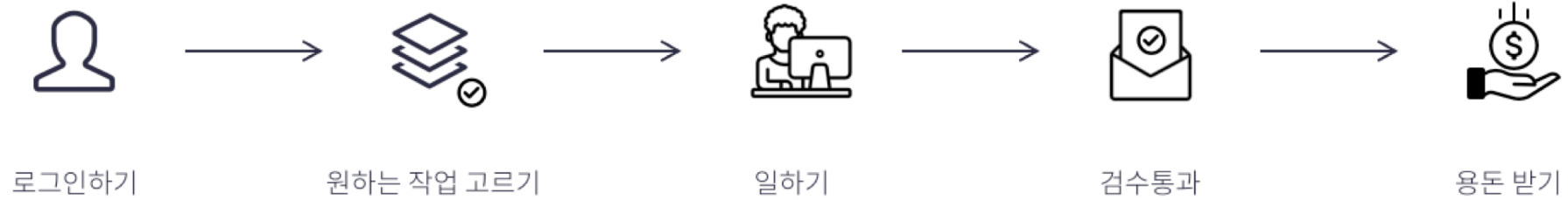
DeepNatural AI

알바할 시간이 없.어.요. 자비스 가르치면서 돈 번다구요, 아.시.겠.어.요?

사람과 대화를 꿈꾸는 인공지능에게는 많은 사람들의 손길이 필요해요.
인공지능에게 우리의 언어를 알려주고 용돈도 벌어보세요!

[용돈 벌기 시작하기 >](#)





app.deepnatural.ai/projects

DeepNatural AI 92,500

현재 보유 크레딧 **92,500**

이번 주 크레딧 **23,000 (+8,000)**

지난 주 크레딧 **15,000**

총 누적 크레딧 **589,000**

딤네추럴에 오신 것을 환영해요!

추천 프로젝트

<p>상품 정보 입력하기 (교육/문화용품)</p> <p>AI에게 상품 정보를 알려주세요! 다양한 각도에서 찍은 교육/문화용품 카테고리 제품 사진을 보고, 화면에 나온대로 제품의 정보를 정확히 입력하는 프로젝트입니다.</p> <p>100%</p> <p>600 198 완료 시까지</p> <p>자세히 보기 ></p>	<p>구문 / 의미역 분석 말뭉치 구축 (pilot)</p> <p>AI가 한국어를 더 독특하게 이해할 수 있도록 도와주세요. 주어진 지침에 따라 구문 / 의미역 분석 결과에 오류가 없는지 확인하고 바로잡아 주시면 됩니다.</p> <p>0% 0 0 완료 시까지</p> <p>자세히 보기 ></p>	<p>Frame이 올바른지 O/X 체크하기</p> <p>AI가 보다 독특하게 한국어를 이해할 수 있도록 도와주세요. 주어진 문장을 확인하고 서술어에 부착된 Frame이 올바른지 판단하여 O/X 체크해주시면 됩니다.</p> <p>60% 200 3 완료 시까지</p> <p>자세히 보기 ></p>	<p>영어 문장 듣고 평가하기</p> <p>문장별로 잘린 영어 문장을 듣고 자신감을 평가해 보세요!</p> <p>35% 0 9 완료 시까지</p> <p>자세히 보기 ></p>	<p>AI에게 음악듣는법 가르치기 _Cubase_Rock</p> <p>AI에게 음악 듣는 법 가르치기 Cubase의 Rock 장르입니다. 주어진 음원을 듣고 MIDI 데이터를 생성해 주시면 인공지능이 음악을 이해하는데 도움이 됩니다.</p> <p>0% 0 0 완료 시까지</p> <p>자세히 보기 ></p>
--	--	---	---	---

모든 프로젝트

AI에게 음악듣는법 가르치기_Logic-Hiphop	AI에게 음악 듣는 법 가르치기 Logic의 Hiphop 장르입니다.	4% 2 완료 시까지
구문 / 의미역 분석 말뭉치 구축 (교육)	AI가 한국어를 더 독특하게 이해할 수 있도록 도와주세요.	100% 1 완료 시까지
상품 정보 입력하기 (의약품/의료기기)	AI에게 상품 정보를 알려주세요!	100% 114 완료 시까지

app.deepnatural.ai/credit

DeepNatural AI

92,500

크레딧

크레딧 현황 지급 내역

보유 크레딧 92,500

실명인증이 완료된 작업자만 지급 신청이 가능합니다.
5,000 크레딧 이상부터 주 1회 신청 가능하며, 세금(3.3%) 공제 후 **등록된 계좌**로 지급됩니다.

전액	+ 5,000	+ 10,000	+ 10만
----	---------	----------	-------

90,000

예상 지급일은 2019년 11월 08일 입니다.

현재 크레딧	92,500
출금 크레딧	- 90,000
예상 잔액 크레딧	2,500

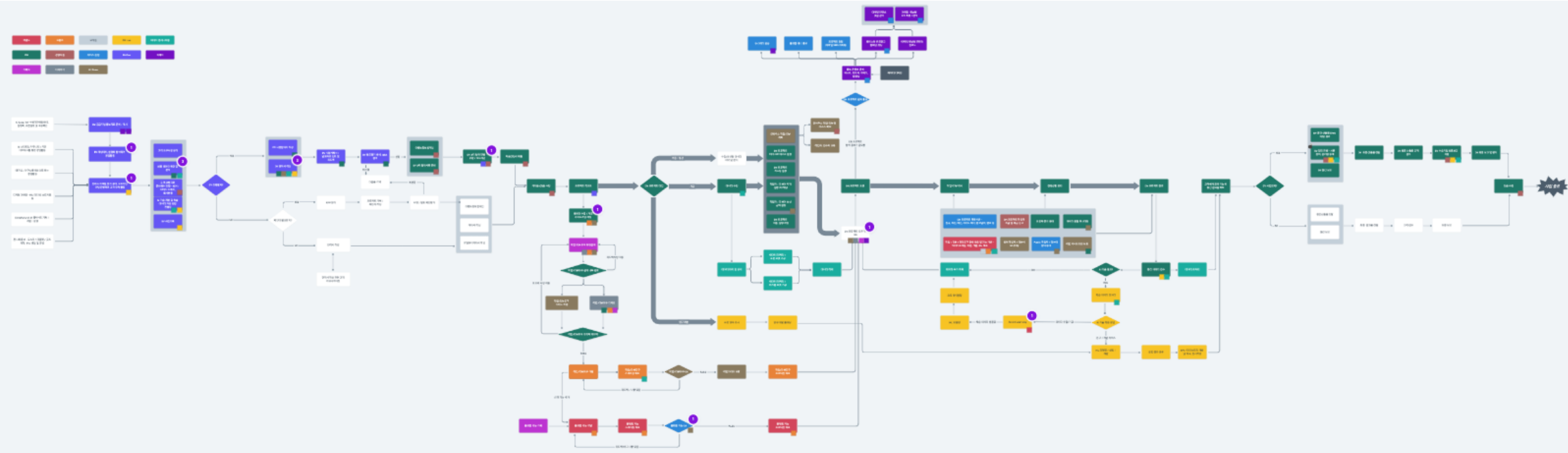
출금 계좌등록

은행

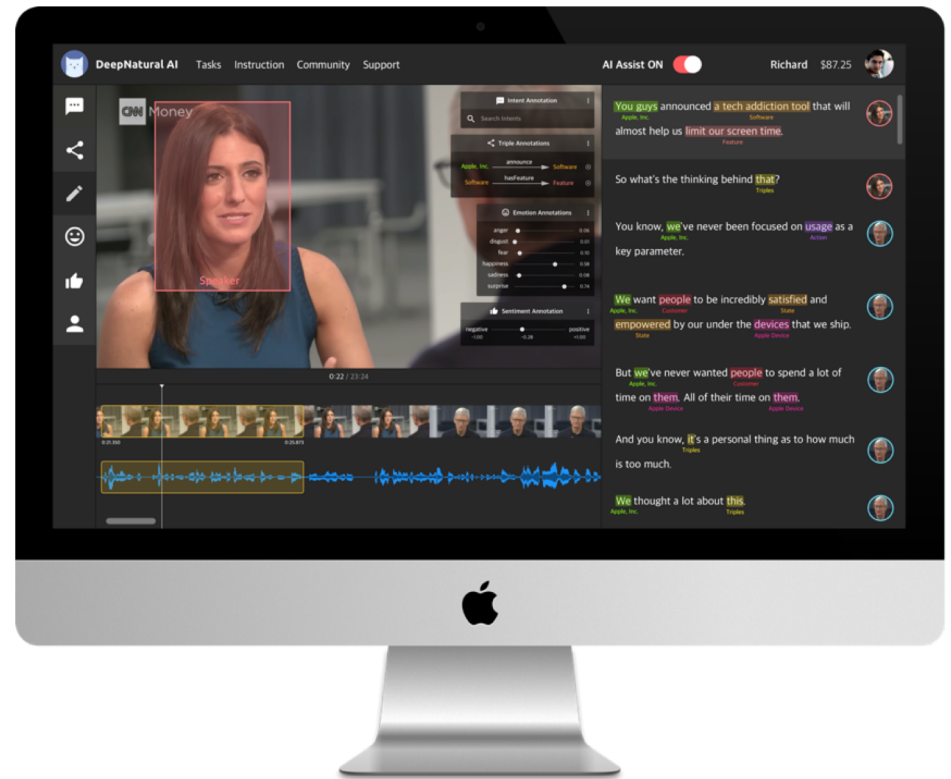
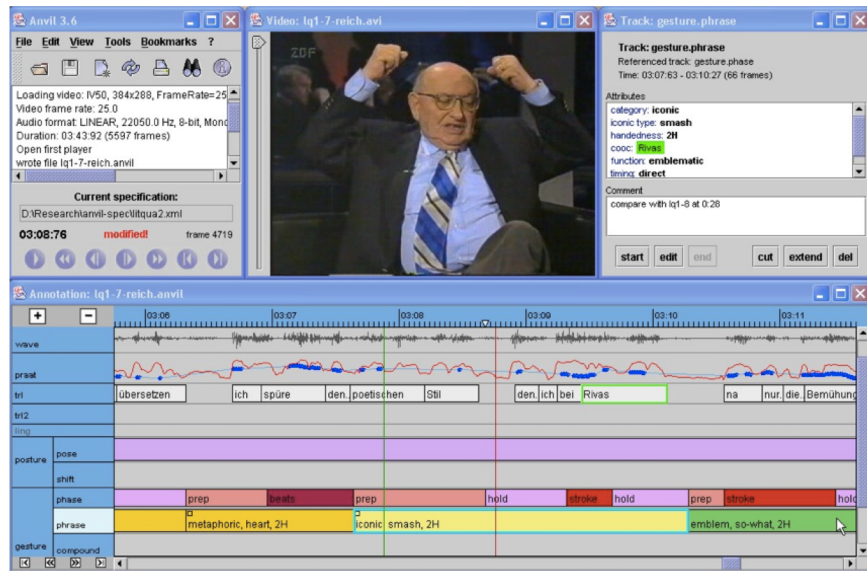
카카오뱅크

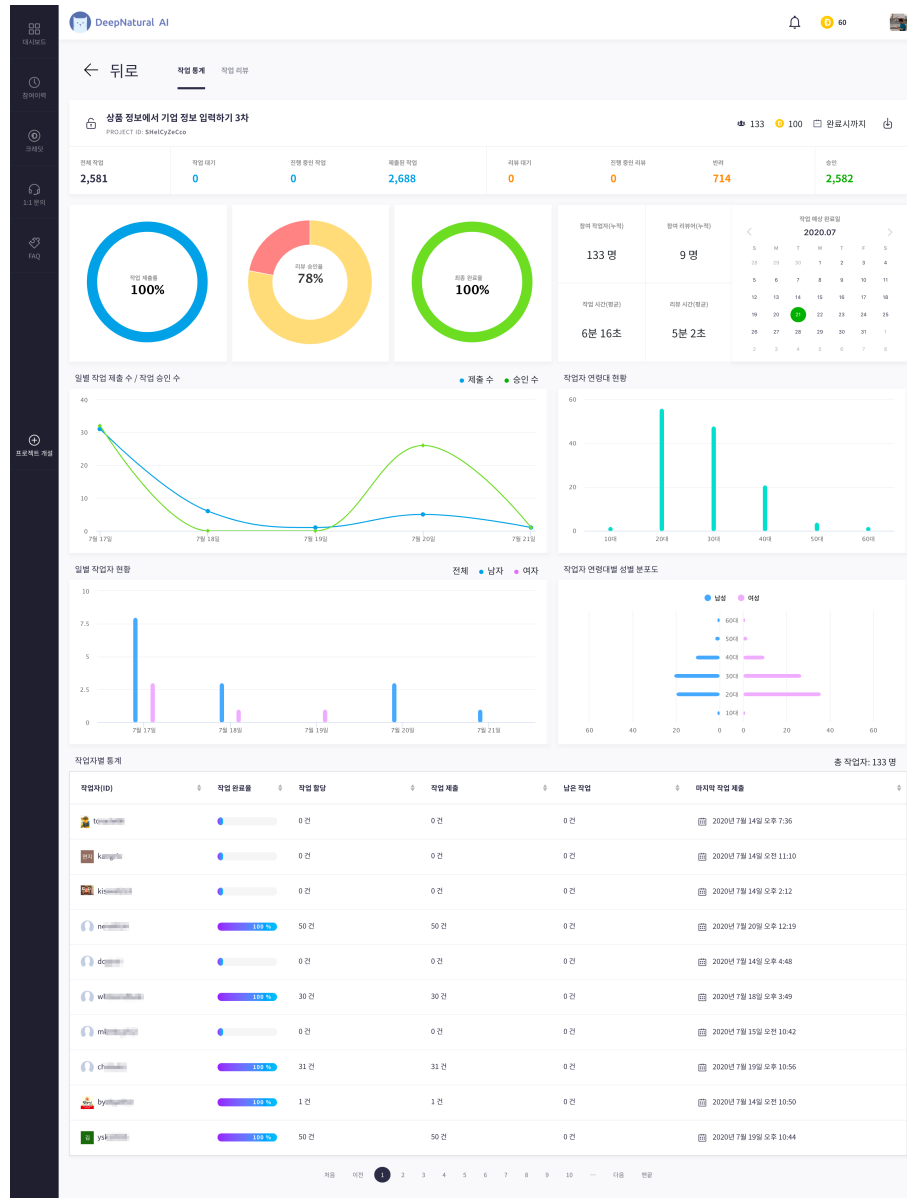
프로젝트 개설

코퍼스 구축 서비스 플로우



향상된 어노테이션 도구





프로세스 모니터링 및 검수 시스템



코퍼스 구축 사례

2019년 베타 서비스 시작부터 지금까지

국립국어원, KB금융지주, KAIST 연구실, 스폰라디오, 땡스플로우, YTN 등과
40+개 이상의 코퍼스 및 학습 데이터 구축 프로젝트 진행

대기업 / 기관 / 스타트업 / 학교 / 연구소

We are trusted by



👤
🔍
SKIP
나가기
제출하기

개요 ×

분석 결과 작성방법
가이드 라인을 작성해주세요.
아래 "자세히 보기" 버튼을 눌러서 확인해주세요.

자세히 보기 >

Q 아래 원 문장을 올바르게 수정해 주세요.

원문장 또 기업별로는 포함제철이 이번 상반기에 1천2백15억원으로 물류비지출이가장 컸으며 그 다음으로 유공(6백70억원)쌍용양회(5백17억원)금성사(3백68억원)등의 차례였다.

ID	FORM	LEMMA	XPOS
1	또	또	MAG
2	기업별로는	기업별 로 는	NNG + JKB + JX
3	포함제철이	포함제철 이	NNP + JKS
		◀ 포함제철	NNP ↓
		◀ 이	JKS ↑ 삭제
		◀ 형태소를 입력하세요	품사를 선택하세요
4	이번	이번	NNG
5	상반기에	상반기 에	NNG + JKB
6	1천2백15억원으로	1 천 2 백15억 원 으로	SN + NR + SN + NR + NNB + JKB
7	물류비지출이가장	물류비지출 이 가장	NNG + JKS + MAG
8	컸으며	크 었 으며	VA + EP + EC
9	그	그	MM
10	다음으로	다음 으로	NNG + JKB
11	유공(6백70억원)쌍용양회(5백17억원)금성사(3백68억원)등의	유공 (6백70억 원) 쌍용양회(5 백17억 원)금성사 (3 백68억 원)등 의	NNG + NR + NNB + SS + NNP + SN + NR + NNB + NNG + SS + SN + NR + NNB + JKG
12	차례였다.	차례 이 었 다 .	NNG + VCP + EP + EF + SF

☰
개요 닫기

개요

KB 개체명 인식
가이드 라인을 작성해주세요.
아래 "자세히 보기" 버튼을 눌러서 확인해주세요.

자세히 보기 >

Q 아래 문장을 가이드라인에 따라 분석해주세요

COMPANY_AND_BRAND 현대대상 < CCO PLO_PER 전세영 <

상무는 "아이들이 주어진 미션에 열심히 참여하면서 한층 밝아진 모습을 보니 뿌듯하고 보람을 느낀다"며 "청소년들의 건강한 성장을 돕는 다양한 활동들을 지속적으로 펼쳐 나갈 것"이라고 밝혔다.

PLO 태그

PLO_PER P PLO_LOC L PLO_ORG O

CATEGORIES

PERSON P AREA L COUNTRY O

GOVERNMENT P GOVERNMENT_OFFICE L FINANCIAL_INSTITUTION L RESEARCH_INSTITUTE L ECONOMIC_FEDERATION L ARMY L COMPANY_AND_BRAND L

SPORTS_TEAM L MEDIA L OTHER_ORGANIZATION L

CONTINENT L NATURE L ROAD L PUBLIC_SQUARE L HOSPITAL L SCHOOL L STORE L STADIUM_AND_MUSEUM L COMPLEX_BUILDING L

OTHER_LOCATION L

PRODUCT L FINANCIAL_PRODUCT L OBJECT L SOFTWARE L THEORY_AND_PHENOMENON L VALUE L DISEASE L TITLE L EVENT L

PROJECT L PRICE L DATETIME L PHONE_NUMBER L AMOUNT L OTHER L POLICY L ABSTRACT L

개요 닫기

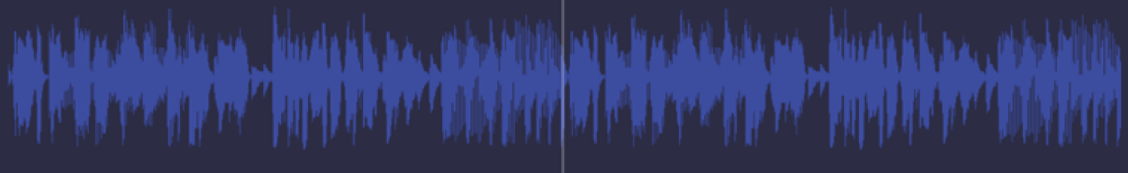
DeepNatural AI 작업설명 튜토리얼 1:1 문의하기 인공지능 비서 OFF 21,000

영어 스피킹 - 점수 알려주기

인공지능 영어 선생님이 더 똑똑해 질 수 있도록 도와주세요!
학생들이 녹음한 영어 스피킹 오디오를 듣고 Anxiety, Demotivated, Competent 3가지 점수를 측정해주시면 됩니다 🙌

Pitch, Loudness, Sound Pressure, Timbre, Tone 요소들을 고려해주세요 🧑

00:06 / 00:12



초보 AI 선생님에게 이 학생의 영어 스피킹 점수를 알려주세요 🙌

Anxiety 불안감, 긴장감
매우 불안하거나 심각하게 긴장함 전혀 불안하지 않고 긴장하지도 않음

Demotivated 관심/의욕 없음
관심이 없어보이며 의욕도 전혀 느껴지지 않음 매우 관심이 많고 의욕 또한 매우 충만함

Competent 자신감
자신감이 심각하게 부족함 자신감이 충분함

영어 말하기 듣기

반복 듣기 x 2

반복 듣기 x 5

선택 그리기 처음 이전 다음 끝 삭제 실행취소 다시실행 전체취소 화면모드 나가기 작업완료

DeepNatural

인공지능 비서 OFF

21,000

TASK

Frame 선택 및 논항 의미역 선택

25% (1/4)

AI가 보다 똑똑하게 한국어를 이해할 수 있도록 도와주세요. 주어진 문장 내 서술어에 대한 가장 적절한 Frame을 선택하고, 각 논항들에 대한 의미역을 알려주시면 됩니다 🐱

Sentence #15214

한편 16일 국제 유가는 OPEC 석유장관들이 오는 27일 빈 회의에서 증산에 합의할 것이라는 기대감에 힘입어 하락세를 지속했다.


Frame 선택

위 문장에서 붉은색으로 표시된 합의하다.v 의미 중 가장 비슷한 의미를 선택해주세요.

- Agree_or_refuse_to_act
 - 클린턴 행정부는 북한의 미사일 판매를 끝내기 위해 일년에 1십억 달러를 보내기로 합의하였나 ?
Speaker Purpose Proposed_action
 - 이라크 정부는 미국 대표 Tony Hall이 1990년 걸프전 이후 심해진 인도주의적 위기를 평가하기 위해 다음 주 금요일을 방문하도록 합의하였다고, 월요일 Hall의 사무실이 언급하였다.
Speaker Proposed_action
 - 한일 양국은 뉴라운드 농산물협상 등 국제 농업교섭에서 공동 보조를 취하기로 공식 합의했다.
Interlocutor Manner
 - 그녀는 결국 그녀와 이혼 서류에 도장을 찍기로 합의했다.
Speaker
- Be_in_agreement_on_action
 - 2004년 1월 말, 북한과 나이지리아는 미사일 거래에 합의한 것으로 전해지지만, 나이지리아는 미국의 압력 하에 2월 초에 합의를 취소하였다.
- Make_agreement_on_action
 - 부시 대통령은 마약 문제를 논의하기 위해 Barco, 페루 대통령 Alan Garcia 및 볼리비아 대통령 Jaime Paz Zamora와 90일 이내에 만나기로 합의하였다.
Party_1 Obligation

Frame 선택하기

각 Frame에 대한 예제 문장들을 정확히 확인하고 선택하셨나요? 그렇다면 [Frame 선택 완료] 버튼을 클릭하여 논항 의미역 선택을 진행해주세요.



Frame 선택 완료

선택 그리기 이전 다음 삭제 실행취소 다시실행 전체취소 화면모드 나가기 작업완료

DeepNatural AI

🔔
₩ 21,500

NWRW180000021-0001-0001 그리고 그 모든 필요 요소들을 결합시키는 데는 합치된 인식 전환과 공감이 절대적으로 전제가 된다. ✔ 규칙 검사 통과

ARG1	결합시키는	합치된	전환과	공감이	절대적으로	전제가	된다

의미역 태그 선택

ARG0 동작주, 행위자

ARG1 피동작주, 대상

ARG2 시작점, 수혜자

ARG3 도착점

ARGM LOC(장소) MNR(방법) PRD(보조서술) DIS(담화연결) INS(도구)

DIR(방향) TMP(시간) PRP(목적) ADV(부사적어구)

CND(조건) EXT(범위) CAU(발생이유) NEG(부정)

ETC ARGA ARGD rel aux

ETRI ARG1 (피동작주, 대상)

합치된 ———— 전환과

의존소				지배소			
ID	FORM	LEMMA	XPOS	ID	FORM	DEPREL	
1	그리고	그리고	MAJ	15	된다.	AP	
2	그	그	MM	5	요소들을	DP	
3	모든	모든	MM	5	요소들을	DP	
4	필요	필요	NNG	5	요소들을	NP	100%
5	요소들을	요소 들 을	NNG+XSN+JKO	6	결합시키는	NP_OBJ	100%
6	결합시키는	결합 시키 는	NNG+XSV+ETM	7	데는	VP_MOD	100%
7	데는	데 는	NNB+JX	15	된다.	NP_AJT	100%
8	각계의	각계 의	NNG+JKG	12	공감이	NP_MOD	75%
9	합치된	합치 되 ㄴ	NNG+XSV+ETM	11	전환과	VP_MOD	75%
10	인식	인식	NNG	11	전환과	NP	100%
11	전환과	전환 과	NNG+JC	12	공감이	NP_CNJ	100%
12	공감이	공감 이	NNG+JKS	15	된다.	NP_SBJ	100%
13	절대적으로	절대 적 으로	NNG+XSN+JKB	15	된다.	NP_AJT	100%
14	전제가	전제 가	NNG+JKC	15	된다.	NP_CMP	100%
15	된다.	되 ㄴ 다.	NNG+JKC	0	ROOT	VP	100%

준형: -게

날말 뜻: 일이 잘 이루어지다.

의미역 프레임: X:대상 Z:방법-게

용법: '-게' 대신에 '잘, 적당히, 원하는 대로' 따위의 부사나 부사어가 사용되기도 한다

예문: -일이 깔끔하게 되다.
-요즘은 사업이 그려져려 되고 있다.
-일이 되면 만 원 더 준답니다.

되다

- 되다 010101 [동사]
- 되다 010102 [동사]
- 되다 010103 [동사]
- 되다 010104 [동사]
- 되다 010105 [동사]
- 되다 010106 [동사]
- 되다 010107 [동사]
- 되다 010108 [동사]
- 되다 010201 [동사]
- 되다 010202 [동사]
- 되다 010203 [동사]
- 되다 010301 [동사]
- 되다 010302 [동사]
- 되다 010303 [동사]
- 되다 010304 [동사]
- 되다 010401 [동사]
- 되다 010402 [동사]
- 되다 010501 [동사]
- 되다 010502 [동사]
- 되다 010503 [동사]
- 되다 010504 [동사]
- 되다 010601 [동사]

📍 선택
📈 그리기
⏪ 처음
⏩ 이전
⏪ 다음
⏩ 끝
🗑️ 삭제
↶ 실행취소
↷ 다시실행
🔄 전체취소
💡 화면모드
📄 나가기
🔄 작업완료

DeepNatural

인공지능 비서 OFF 🔔 💰 21,000

TASK
1:1 실시간 상담 채팅 생성하기

75% (15/20)

인공지능 상담원이 능숙하게 응대할 수 있도록 다양한 대화 예제를 알려줍니다. 아래 주어진 상황에 몰입하여, 실제 고객과 상담원이 나눔법한 대화를 생성해주세요 🐱

현재 대화 장소 / 채널

숙박업 > 온라인문의 > 홈페이지
호텔 / 콘도 / 리조트 등 숙박시설 온라인 홈페이지, 결제수단, 예약현황

현재 대화주제

일반예약 / 당일예약 / 단체예약 관련한 예약, 취소, 환불, 할인

고객의 대화 목적

불만 토로하기

고객의 대화 분위기

부정적 중립 긍정적

상담원: 고객님 체크인 하시는 건가요?

고객: 네. 체크인하려고요. 저희가 조금 빨리왔는데 지금 바로 가능할까요?

상담원: 원래 4시부터 가능한데 전날 손님이 없어서 지금 바로 가능할게 같습니다.

고객: 잘됐네요. 감사합니다. 여기까지 온다고 피곤해서 지금 바로 쉬고 싶었거든요.

상담원: 저희도 기쁘네요. 지금 바로 체크인 도와 드리겠습니다. 또 다른 필요하신건 없으세요?

고객: 아 그럼 혹시 저희 원래 더블침대 1개로 예약을 한거 같은데 싱글 2개로 교체 해주실 수 있어요?

상담원: 지금 객실에서 바로 교체는 어려울거 같고 같은 타입의 객실으로 바꾸면 가능하신데 괜찮나요?

고객: 네 상관없습니다. 그렇게 해주세요.

상담원: 네 그렇게 진행하겠습니다. 그럼 잠시만 기다려 주세요. 다른거 필요하신건 없으신가요?

고객: 혹시 방 키는 몇개가 나오나요?

상담원: 각 방마다 한개씩 지급되는데 요청시에는 한개더 가능합니다 고객님.

고객: 그럼 방 키 한개만 더 주세요. 그리고 객실안에 물이 몇병정도 있나요?

상담원: 기본으로 두병 제공됩니다 고객님.

메시지 작성하기

아래 가이드라인과 규칙에 맞게 상담원과 고객의 대화를 실감나게 작성해주세요 🐱

가이드라인

- 대화 본문부터 시작 (반반하게 등장하는 일반적인 인사는 생략)
- 대화 끝에는 내용이 종결되도록 함 (문의 및 불만접수 완료)
- 상담원과 고객은 1회씩 번갈아가며 메시지 전송가능
- 단답형의 질문과 대답은 최대한 자제
- 기업명, 사람이름 등의 고유명사 사용 가능
- 표준어 사용, 맞춤법 준수

메시지 작성 규칙

- ✅ 메시지 평균 길이 (띄어쓰기 포함) : 31글자 ≥ 30글자
- ❌ 대화에 포함된 메시지 수 : 15 메시지 ≥ 20 메시지
- ❌ 문장의 끝 규칙 준수 : 온점 / 물음표 사용 (상담원=필수, 고객=권장)

상담원: 항상 중립의 감성 유지
이모티콘 사용 X

고객: 주어진 상황에 따라 감성 변화
과도한 이모티콘 사용 X

메시지를 작성 후 엔터키를 입력해주세요.

선택 그리기 이전 다음 삭제 실행취소 다시실행 전체취소 화면모드 나가기 작업완료

DeepNatural AI

인공지능 비서 OFF

대화 내용 텍스트로 받아쓰기

대화 단위로 영상 구간을 선택하고 음성 내용을 텍스트로 옮겨주세요. 받아쓰기 하듯이 최대한 동일하면서 맞춤법에 맞게 적어주시면 됩니다 🐱

김어준의 뉴스공장 8.20(월) 방송 박원순 서울시장 출연

tbs tv 시민의 방송

대화 내용

발화자

00:07 지난 한달간 옥담방 생활하셨죠?

00:11 그리고 강북 우선 투자발표를 하셨습니다.

00:16 어~ TBS가 이분 밑에 있습니다.

00:22 사실은 그리고 그래서 오히려 역차별을 받아 1년에 두번밖에 출연할 수 없는 분을 모셨습니다.

00:31 박원순 서울시장님을 모셨습니다. 안녕하십니까

00:33 네. 반갑습니다.

00:34 1년에 두번만 나오실수 있죠.

00:36 아니그런데 TBS가 내 밑에 있다고요?

00:39 그런걸로 알고있습니다.

00:40 어 그렇게 별로 생각해보본적이 없는데

00:42 그렇게 별로 행동한적이 없기때문에

00:44 아니 내 밑에 있으면 내 마음대로 시켜도 되는거예요?

00:47 아니 총수를 내 밑에 두고 있는거예요?

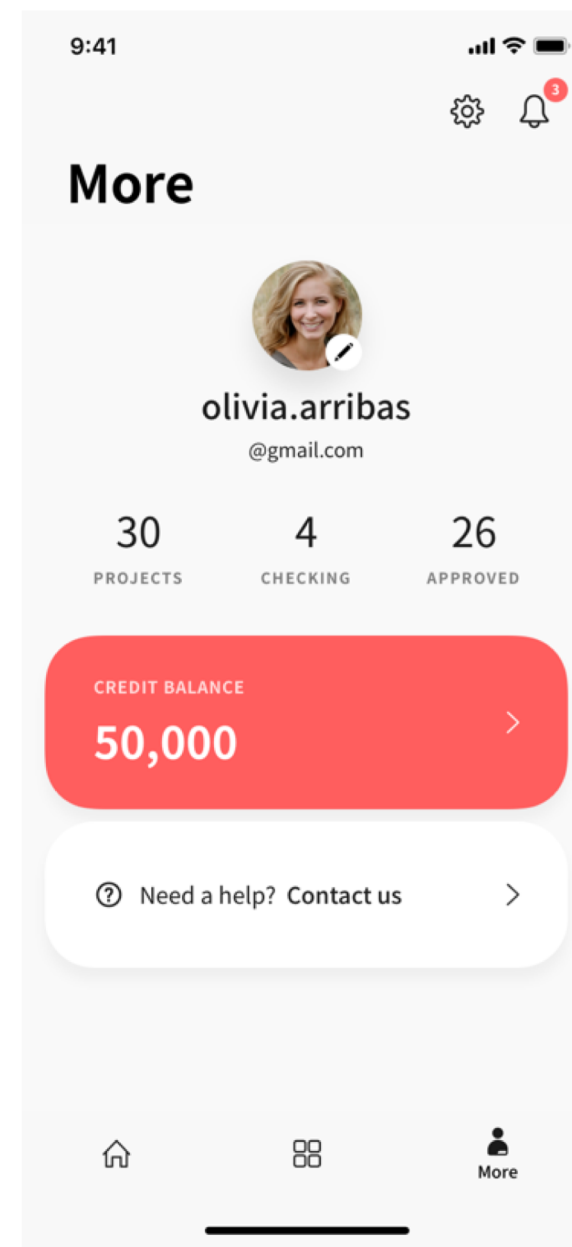
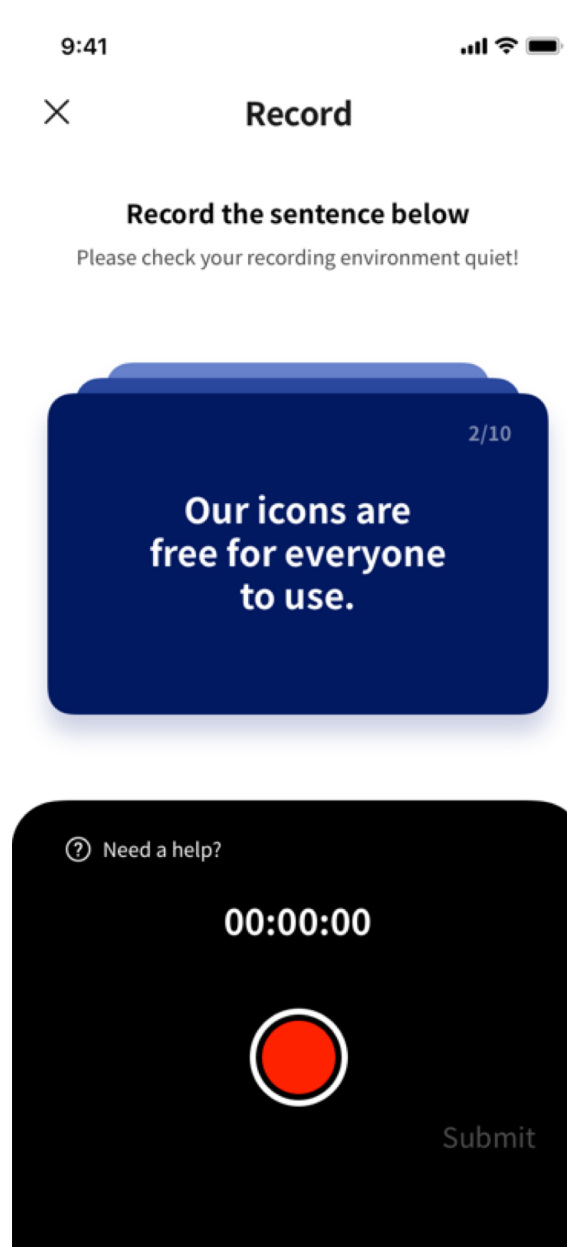
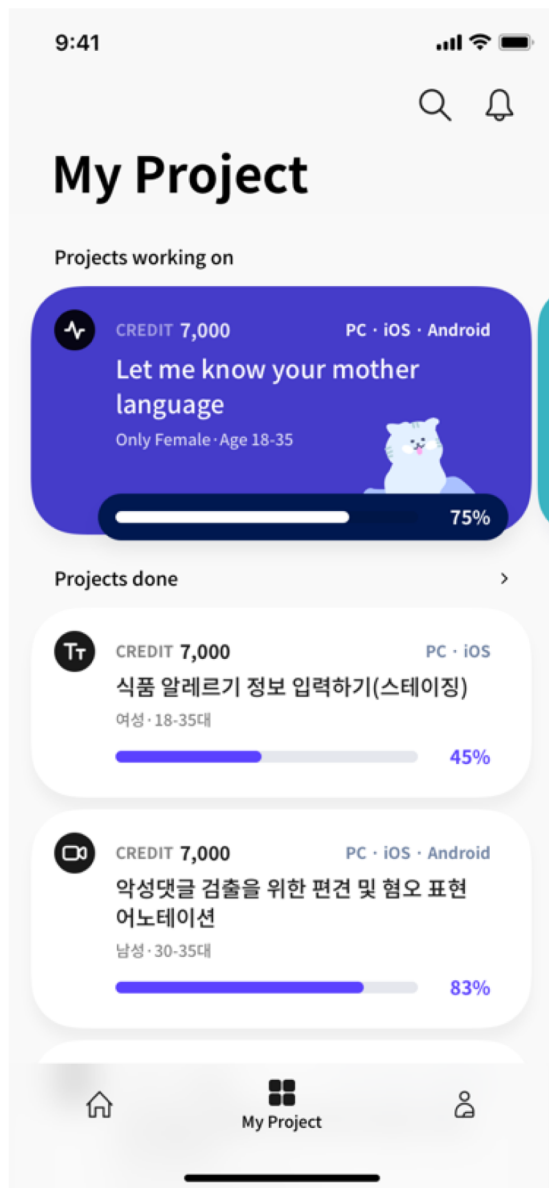
00:49 과거 어 교통방송은 그랬다고 하더군요

tbs TV 시청방법 : alleh tv 214번 | Btv 272번 | U+TV 176번 | 케이블TV

00:50 / 17:25

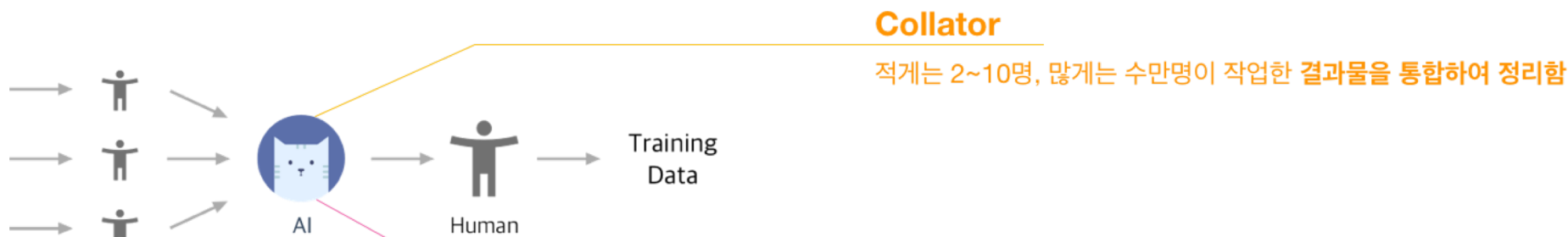
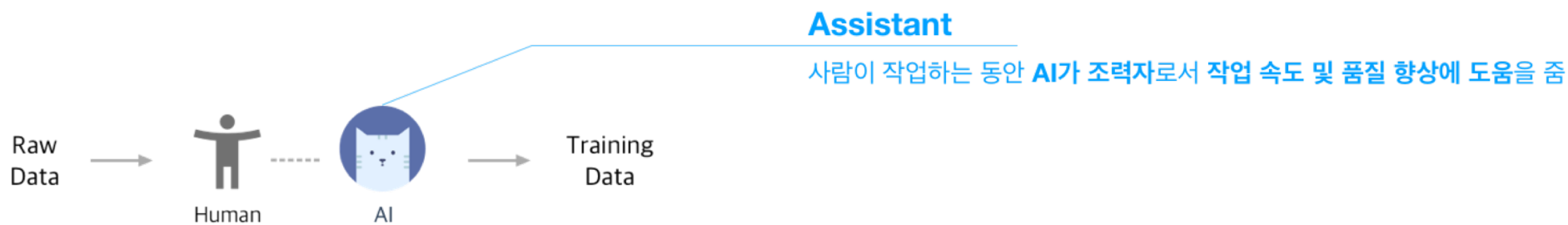
00:50 00:53

선택 그리기 처음 이전 다음 끝 삭제 실행취소 다시실행 전체취소 화면모드 나가기 제출

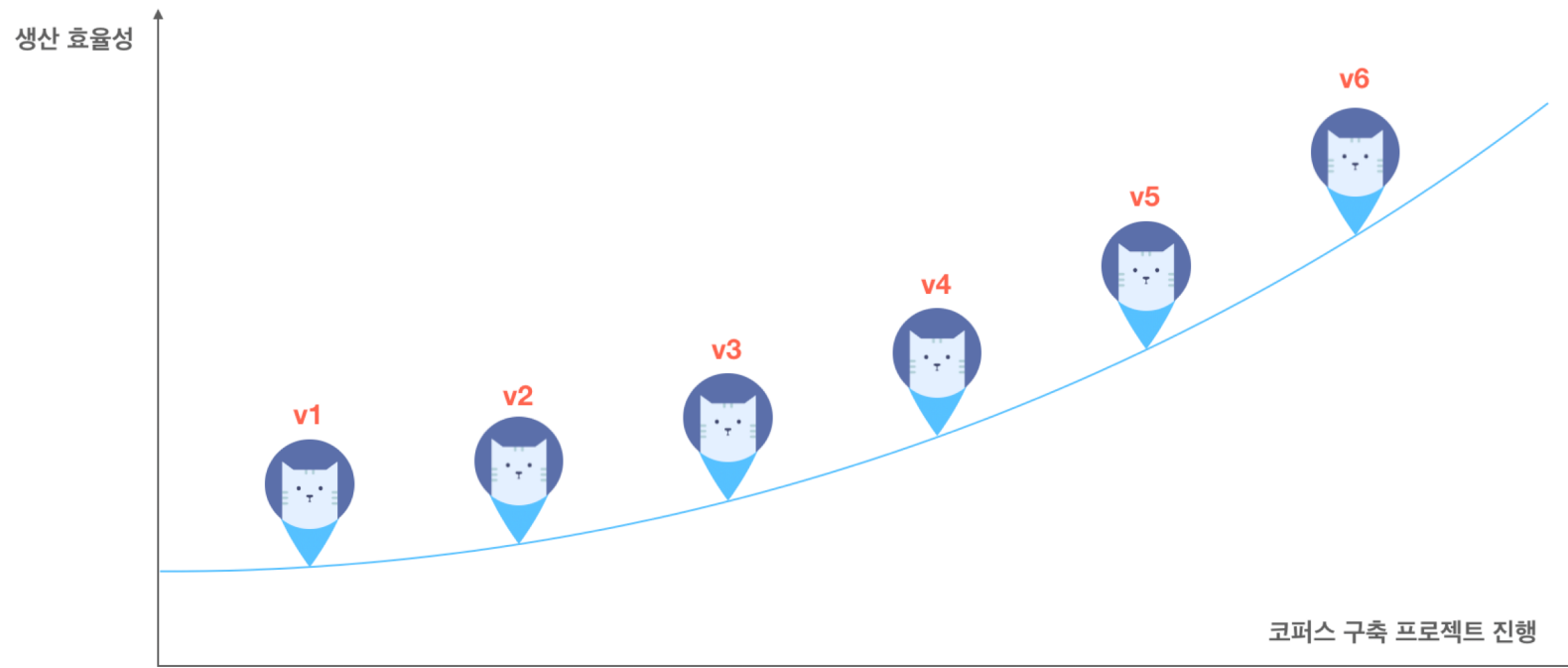




코퍼스 스타트업이 주목하는 기술

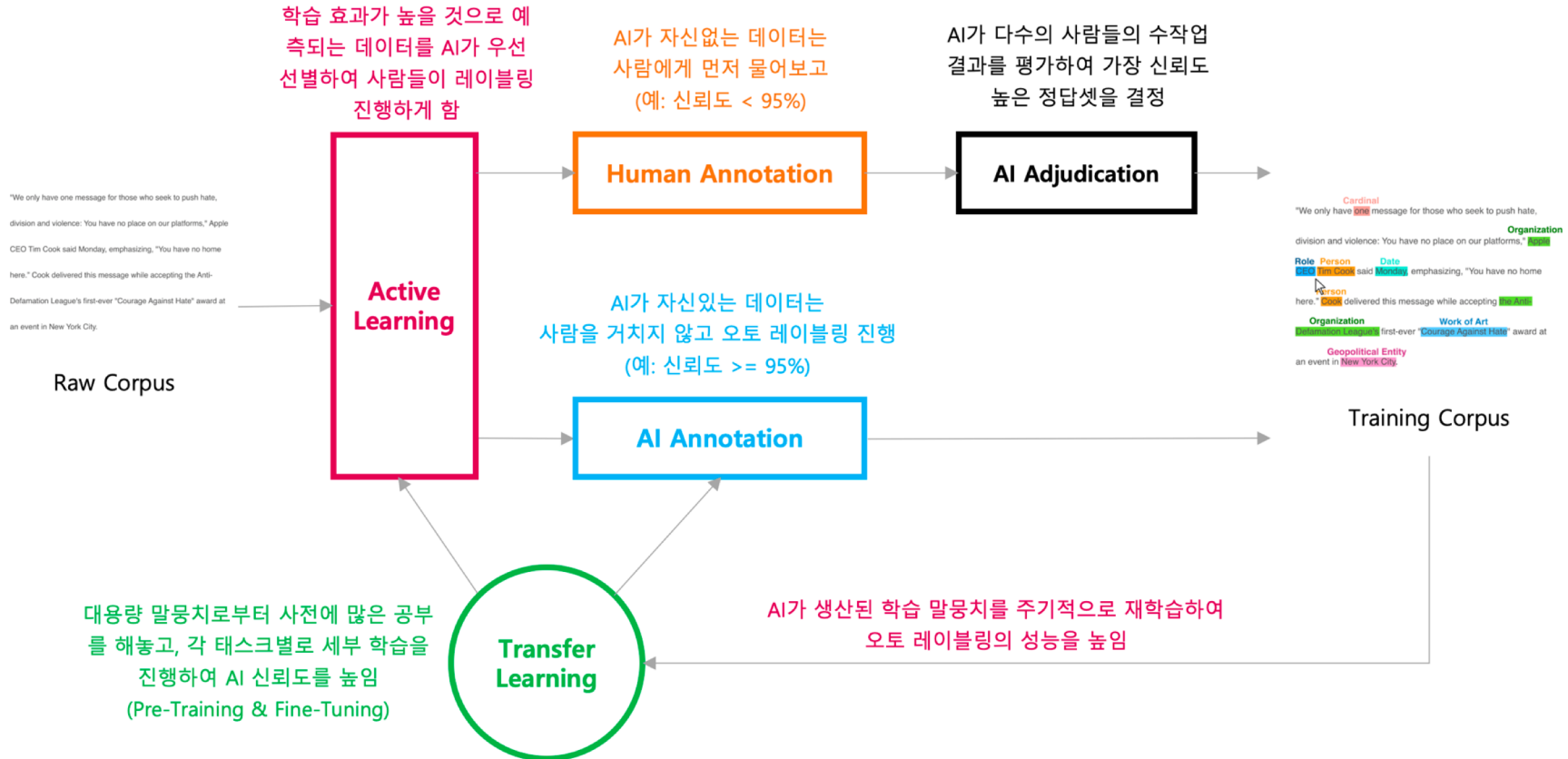


코퍼스 구축 프로젝트가 진행되면서
생산되는 학습 데이터를 주기적으로 학습하며
점점 더 똑똑해지는 AI Assistant

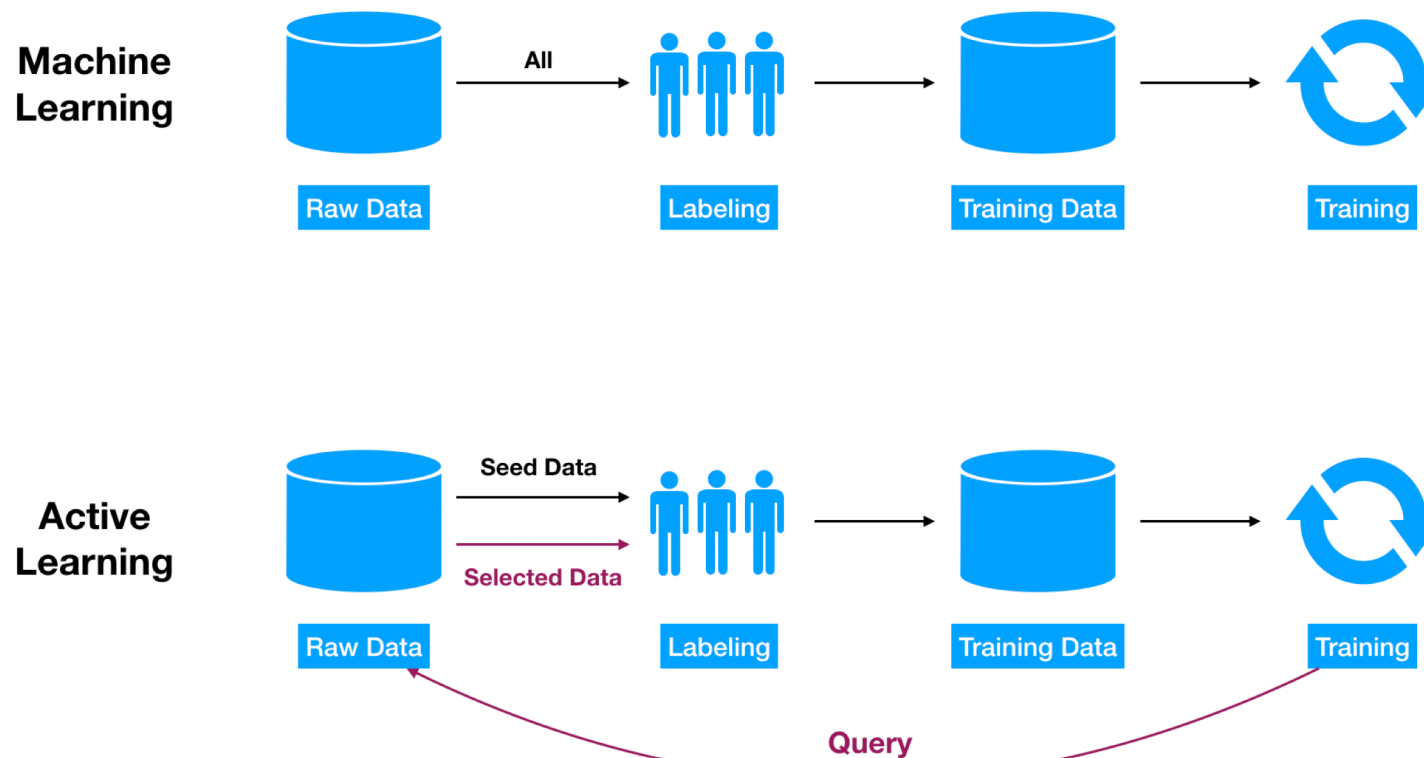


어떻게하면 AI Assistant의 성능을 더 빠르게 높일 수 있을까?

Transfer Learning + Active Learning

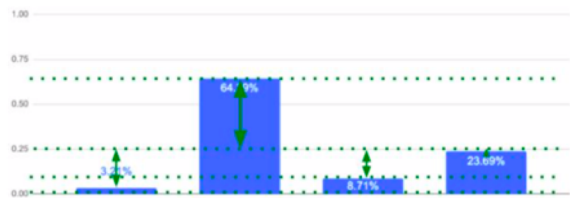


일반적인 기계학습과 액티브러닝의 차이는 학습 과정에서 인터랙티브하게 쿼리한다는 것



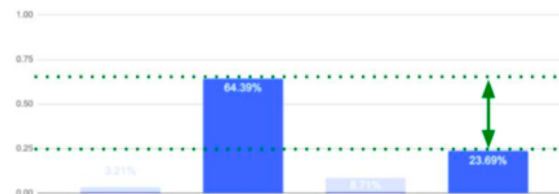
불확실성이 큰 데이터들을 먼저 학습

예측 결과의 불확실성을 판단하는 방법의 예



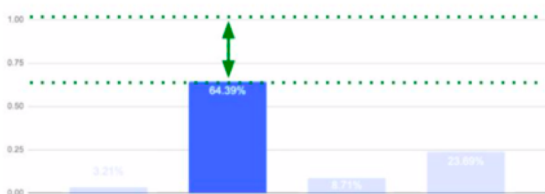
Entropy

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



Ratio of Confidence

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



Least Confidence

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

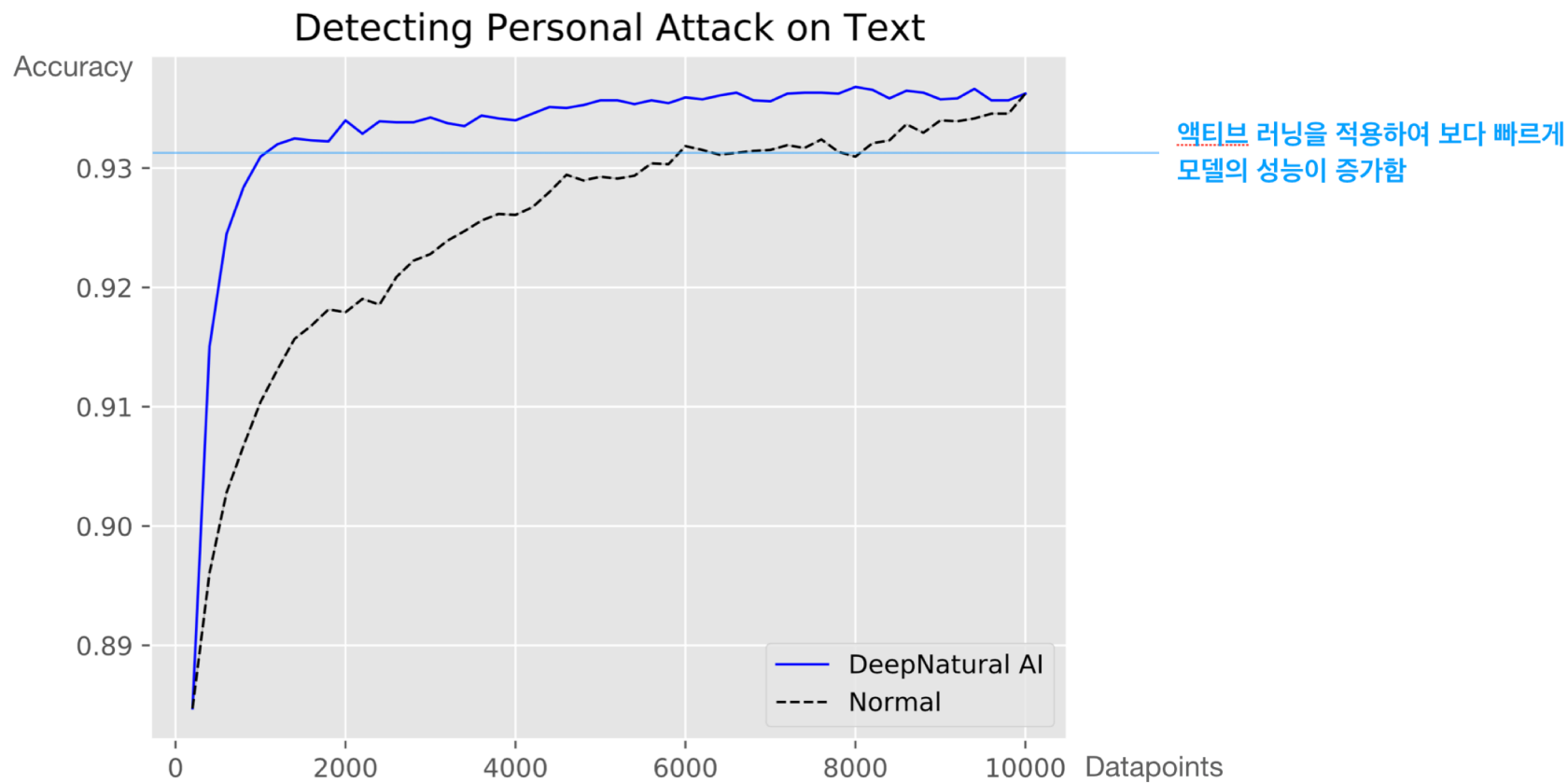


Margin of Confidence

$$1 - (P_{\theta}(y^*_1|x) - P_{\theta}(y^*_2|x))$$

Wikipedia Talk Labels: Personal Attacks 데이터셋에서

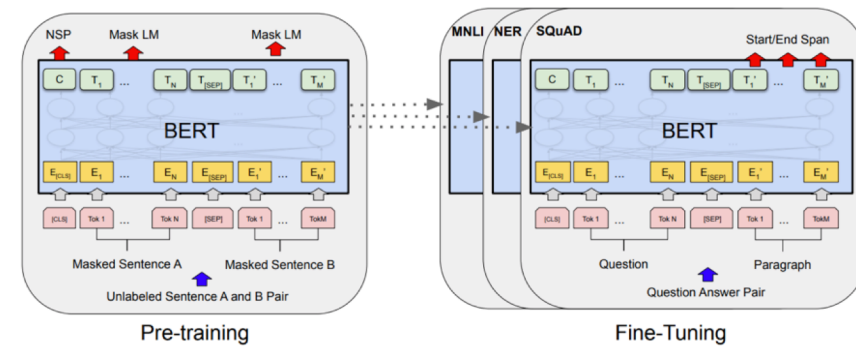
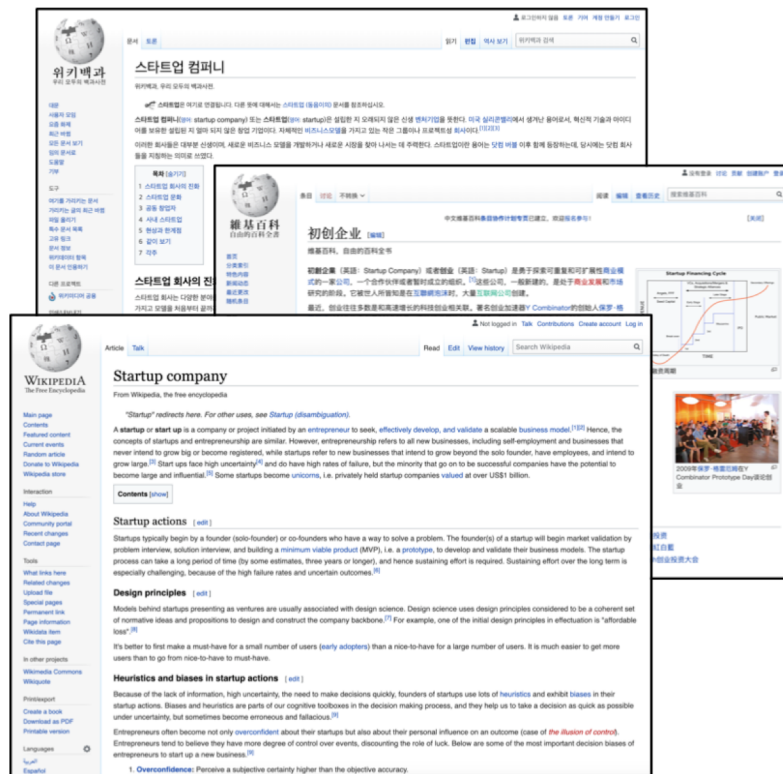
액티브러닝을 통한 빠른 성능향상 실험 결과



학습 데이터 구축 프로세스에 AI 기술을 적용하기 위해서는

학습 데이터가 전혀 없는 상태에서 시작하여, 소량의 학습 데이터를 이용하여 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 상태에서도 AI Assistant의 성능을 높일 수 있음

자연어처리가 점차 비언어 커뮤니케이션 영역으로 확대되지 않을까?

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





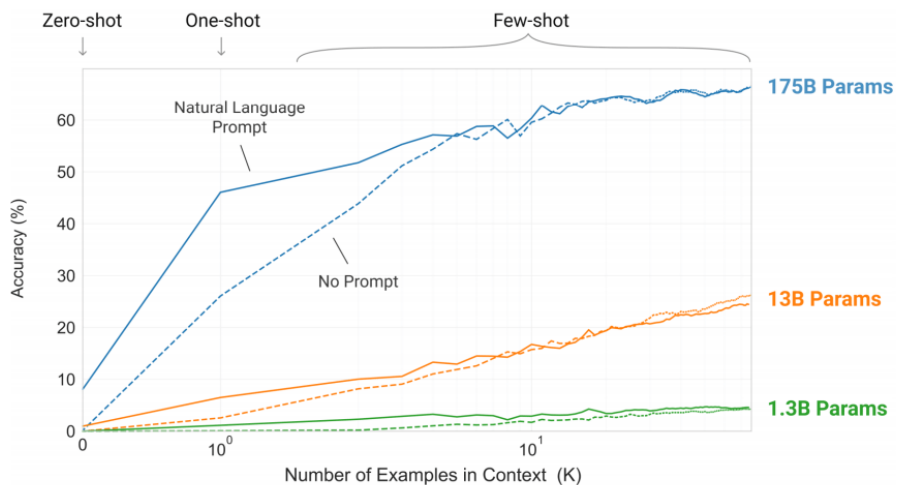
코퍼스 스타트업이 바라본 트렌드

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

국립국어원 모두의 말뭉치

국립국어원 모두의 말뭉치

모두의 말뭉치
미래를 준비하는 소중한 우리말 자원

말뭉치 신청

신문 말뭉치 신문지, 전문지, 연례연구 기관 신문 등 제작 기사와 구성된 말뭉치입니다. 신청하기	문어 말뭉치 텍, 잡지, 보고서 등으로 구성된 말뭉치입니다. 신청하기	구어 말뭉치 방송 자료, 일상 대화, 대본 등으로 구성된 말뭉치입니다. 신청하기	메신저 말뭉치 메신저와 채팅으로 구성된 말뭉치입니다. 신청하기
웹 말뭉치 블로그, 저널, 뉴스, 소셜망의 웹 자료로 구성된 말뭉치입니다. 신청하기	문서 요약 말뭉치 문서에서 추출된 주제문과 문서를 요약한 글로 구성된 말뭉치입니다. 신청하기	형태 분석 말뭉치 어휘를 분석하여 형태 표지를 부착한 말뭉치입니다. 신청하기	어휘 의미 분석 말뭉치 다의어를 구별하여 ~유지말뭉치에 의미 번호를 부착한 말뭉치입니다. 신청하기
개체명 분석 말뭉치 문장이나 단락을 단위로 하여 개체명 분석을 표시하고 분석 결과를 부착한 말뭉치입니다. 신청하기	구문 분석 말뭉치 문장의 구문 구조를 분석하여 구문 분석 결과를 부착한 말뭉치입니다. 신청하기	문법성 판단 말뭉치 한국어 예문 문법성(수용성)을 평가하여 문법성 판단 결과를 부착한 말뭉치입니다. 신청하기	유사 문장 말뭉치 원문이 아닌 유사 문장과 사람이 작성한 유사 문장으로 구성된 말뭉치입니다. 신청하기
어휘 관계 자료: NIKLex 비슷한 의미, 반대말, 동의어 등 어휘 관계를 언어 사용자가 평가한 자료입니다. 신청하기			

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

K-Data 데이터바우처

데이터구입비용 지원합니다!
데이터 구입과 가공서비스 비용 지원해드립니다!

데이터바우처 2차 수요기업 모집!

비용지원

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

AI Hub

AI Hub 소개, AI 데이터, AI 소프트웨어, AI 경진대회, AI 컴퓨팅, AI 인재개발, AI 서비스 사례, AI 개시판

AI Hub
AI 기술 및 제품/서비스 개발에 필수적인 데이터, AI SW/알고리즘, 컴퓨팅 자원 등 의사 인프라를 통합 제공하는 플랫폼입니다.

AI 학습용 데이터
AI 학습에 필요한 다양한 양질의 대용량 데이터를 마음껏 다운로드하세요.

자언어, 이미지, 영상, 지식베이스

한영 번역 말뭉치, 한국어 대화, 한국어 음성

AI 학습용 데이터, 고성능 컴퓨팅 자원, AI 경진대회, AISW API, AI 인재개발, AI 서비스 사례, AI 개시판, AI 허브 소스

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

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	상황별 음성 (4)	9	한국어 강의 데이터	
			10	회의 음성 데이터	
			11	고객 응대 데이터	
	4	열악환경 음성 (3)	12	상담 음성 데이터	
			13	비대면 환경 대화 데이터	
			14	열악한 환경 음성 데이터	
	5	한국어 방언 (5)	15	대화자 음성 합성 데이터	
			16	한국어 방언 발화 데이터(강원도)	
			17	한국어 방언 발화 데이터(경상도)	
			18	한국어 방언 발화 데이터(전라도)	
	6	요약 데이터 (3)	19	한국어 방언 발화 데이터(제주도)	
			20	한국어 방언 발화 데이터(충청도)	
			21	논문자료 요약 데이터	
	7	한국어 텍스트 (4)	22	도서자료 요약 데이터	
			23	한국어 대화 요약 데이터	
			24	도서자료 기계독해	
			25	민원(콜센터) 질의-응답 데이터	
	8	영어번역말뭉치 (2)	26	전문분야 말뭉치 데이터(분야별 개체명 인식 포함)	
			27	한국어 SNS 데이터	
	9	중국어-일본어 번역 말뭉치 (3)	28	한국어-영어 1 번역 말뭉치	
			29	한국어-영어 2 번역 말뭉치	
			30	한국어-중국어 번역 말뭉치 1	
	10	OCR (3)	31	한국어-중국어 번역 말뭉치 2	
			32	한국어-일본어 번역 말뭉치	
			33	아외 실제 촬영 한글 이미지	
				34	다양한 형태의 한글 문자 이미지 인식 데이터
				35	공공행정문서 OCR

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

플랫폼 노동 1탄
플랫폼 노동으로 1주일 살아보기

MBC is a South Korean public broadcast service. [Wikipedia](#)

#플랫폼노동 #쿠리어 #배민커넥터
[MBC 탐사기획 스트레이트 95회] 플랫폼 노동으로 1주일 살아보기 (2020.07.12)

딥네추럴은

- 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

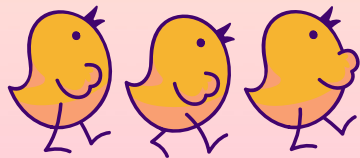


편하게 연락주세요

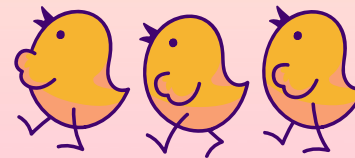
DeepNatural 박상원 대표 / anson@deepnatural.ai

Q&A





LangCon 2020



자연어처리, 데이터에게 길을 묻다

'고품질 코퍼스를 만들어 드립니다'

감사합니다!



DeepNatural

박상원 / CEO / anson@deepnatural.ai