제품 소개서

HanBert_Base . HanBert_MRC . HanBert_NER . HanBert_Sentiment . Moran . Moran_Post OCR

2021.09

(주) 투블럭 Ai

TWOBLOCK Ai

Deep NLP & Conversational Ai

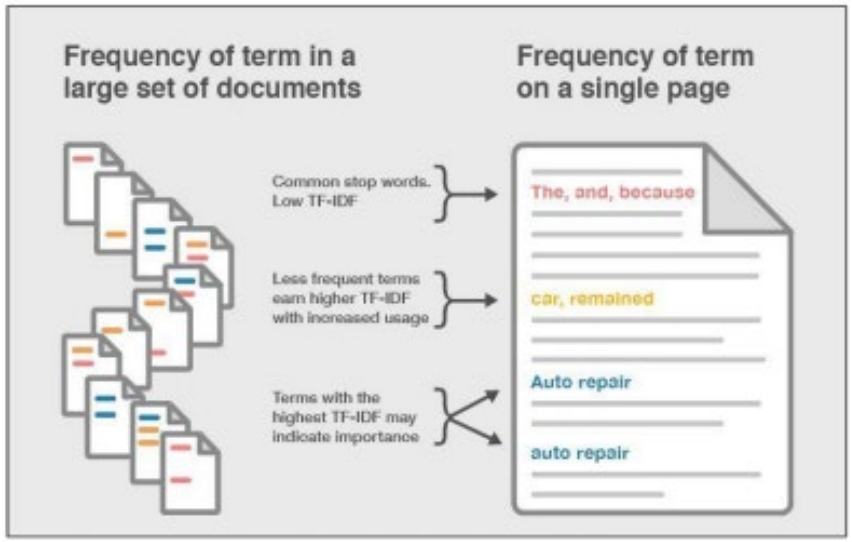
Deep NLP 필요성 | 70점짜리 키워드 분석 수준의 한계

정보량의 법칙 (tf/idf) 의 한계

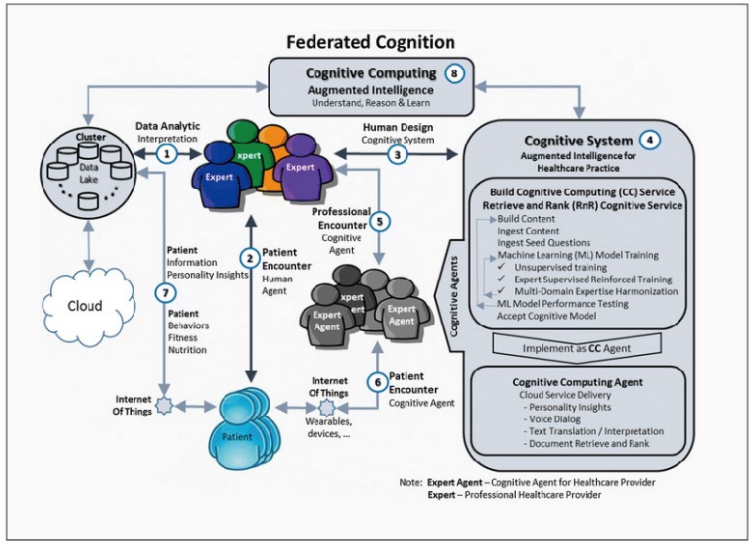
- 문서에서 중요한 단어 추출하여 문서의 의미를 대신하는 방법
- 문서를 자동으로 요약하고 싶은데, 지금 수준으로는 못 쓰겠어요.
- 게시판 댓글을 분석하려는데, 시스템 구축하는 비용이 부담스러워요.

Classical NLP 구축 절차와 비용

- 사전, 문법, 동의어 사전, 온톨로지, 의미 분류 등의 지식 베이스 구축이 필요
- 패턴 수집과 규칙 작성의 일관성와 상호 간섭을 고려하기가 어려움
- 시장에서 기대하는 성능을 만족시키지 못하고 있음



NLP 성능
리비히의 법칙
>



Deep NLP 필요성 | BERT의 놀라운 성능

과거 기술을 통한 언어이해 성능을 혁신적로 추월.

BERT 는 구글이 공개한 모델이며 업체마다 자신들만의 Corpus와 다양한 전처리 방법을 사용하여 자신들만의 모델로 학습을 시켜야 함. (자체 한국어 BERT 모델 보유 기업은 많지 않음. ETRI가 공개한 KoBERT 가 대표적)

(예) 네이버 영화평 DB에서 긍정과 부정의 판별

적용기술	성능	속도	학습비용
임의로 긍/부정 판별	50점	0ms	0원
Symbolic NLP 기술	75점 수준	50ms~100ms	1억원 수준
범용 딥러닝 기술을 활용	85점 수준	10ms~20ms	100만원
BERT에 추가학습	91점	15ms~20ms	10만원

(예)논문에 발표된 BERT와의 성능 비교

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

TBAi Deep NLP Package의 구성

자체 기술로 개발한 HanBert_Base 모델을 고객의 다양한 요구사항에 대응 가능한 Task block으로 구성.
자체 형태소 분석기를 통해 산업별로 다양한 언어모델을 적용하여 Fine Tuning의 자유도가 매우 높은 것이 특징 .
Task block 간의 유기적인 Pipe Line구성으로 RPA , Text Analysis 등 다양한 목적으로 구축 활용 가능.

BUTTERBLOCK

TBAi ButterBlock™ Suite

HanBert_Base (90k / 54k)
TBAi Deep NLP Base Model

Moran
형태소 분석기

General
NER, MRC 등 Task 가능

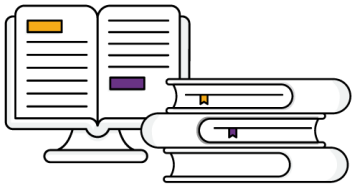
+

HanBERT_NER 개체명 인식	Domain 특화 개체명 인식 Fine Tuning
HanBERT_MRC 기계독해	Domain 특화 기계독해 Fine Tuning
HanBERT_Sentiment 감성분석 TASK	댓글 , 감성분석 Fine Tuning
OCR + NLP	자동 오류수정 , 중요단어 보정 등

* 54k의 경우 연구목적으로 Open Source 공개 중

HanBERT_Base

자체 기술로 정제한 대규모 코퍼스를 활용하고 특허 출원한 한국어 표기법을 Deep NLP에 적용하여 BERT 모델 학습 .
용도에 따라 2가지의 Vocab과 3가지 크기의 HanBERT 모델을 개발하여 활용 용도에 맞추어 적용.
최고 성능의 한국어 범용 심층 언어 모델을 기계독해, 감성분석, 개체명인식 등에 활용.



BERT + Corpus

HanBERT_Base

- 다양한 형식과 내용의 110G 한국어 코퍼스
- MorAn을 활용, 한국어 코퍼스 정제, 선별
- 54k, 90k Vocab, 3가지 크기의 모델
- 건당 30 ms 수준의 처리 속도, 초당 50K 문서 처리

TBAi 의 심층언어모델 HanBERT를 다양한 응용에 적용 합니다.

모델코드	설명			
54k / 90k	**Vocab의 크기**	54k : 54000 표제어 90k : 90000 표제어		
S/N/ML	**모델의 크기와 학습량**	S 6층	히든 768	3.8M 학습
		N 12층	히든 768	5 M 학습
		ML 18층	히든 1024	10M 학습

HanBERT_Base Features

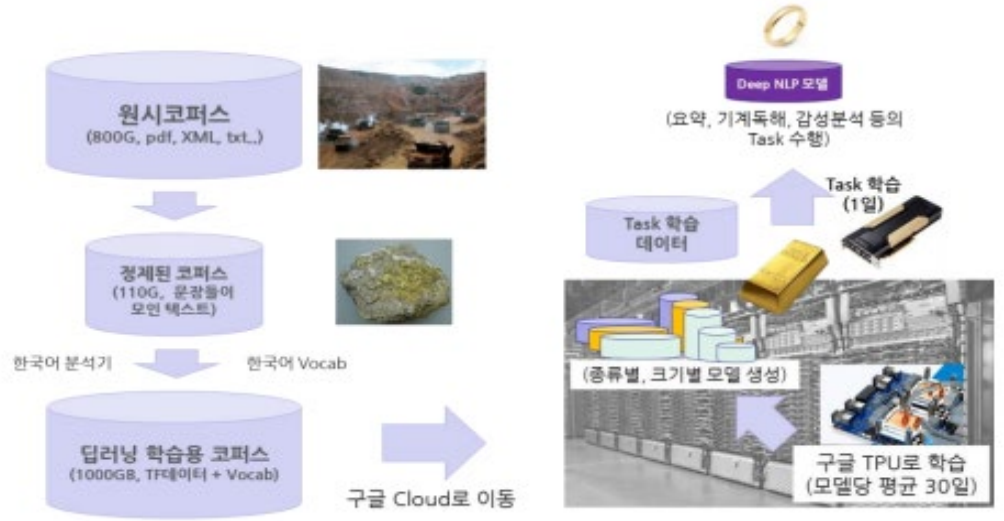
Basic

Basic 한국어 분석기 Moran 기본 탑재
모델 사용 라이센스
용도에 맞게 모델 선택
빠른 처리 속도 (< 30 ms)
다양한 크기의 모델 (54k ,90k)
Vocab에 충분한 여유공간

Premium

고객 데이터 학습 지원
Task별 추가 학습을 위한 기술지원

TBAi HanBERT_Base를 만드는 과정



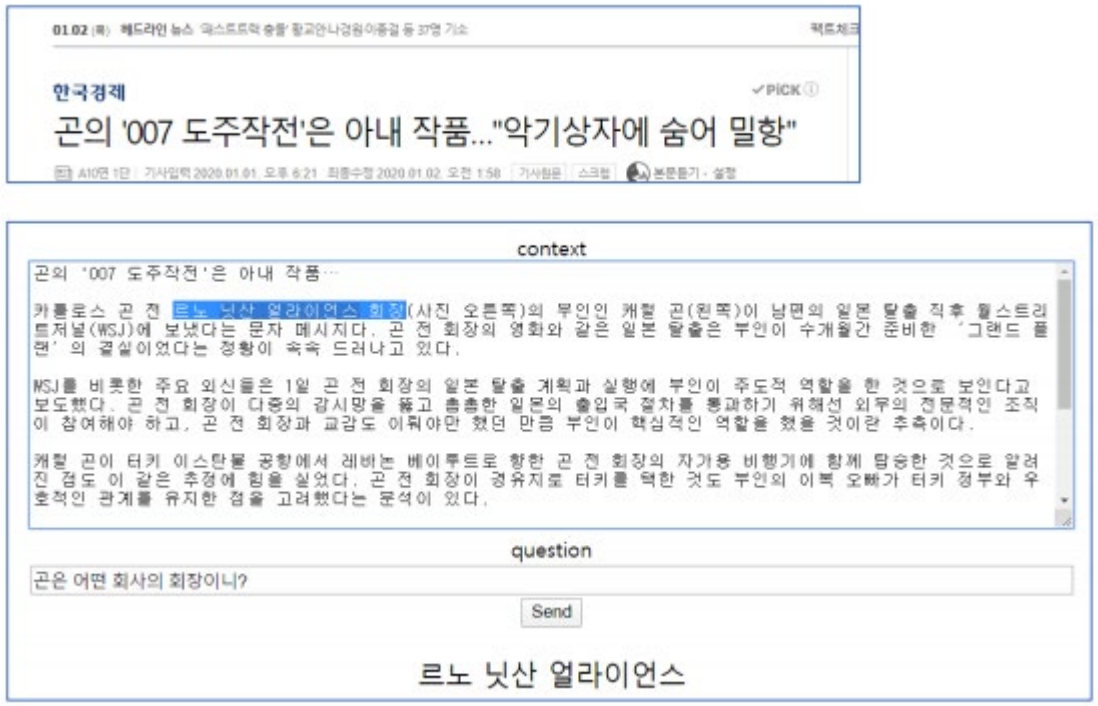
코퍼스 수집 [기존 + 1MM]	코퍼스 정제 [2MM]	형태소 분석 [0.1MM]	Vocab 생성 Size 0.1MM
TF 레코드 생성 Duo, Vocab [0.4MM]	BERT 모델 학습 Loss관찰 [1MM]	모델 성능평가 [2MM]	

> 7개월 <

HanBERT MRC 기계독해

- HanBERT에 기계독해 학습용 데이터를 추가 학습하여 HanBERT_MRC 모델 생성
- 기업 문서에 많이 포함되어 있는 테이블을 처리하기 위해서 HTML 문서에 대한 독해 기능을 제공하며 OCR 엔진과 연동 가능
- 한국어 분석을 위한 Moran과 연동되어 CPU/GPU 서버에서 수행되며, Cloud 서비스를 위한 API도 제공

HanBERT_MRC는 다양한 분야의 문서와 질문에 우수한 성능을 발휘



곤은 어떤 회사의 회장이지 ? ➔ 르노 닛산 얼라이언스

어디에서 탈출했다는 거지 ? ➔ 일본

어디에 숨어서 도망간거지 ? ➔ 대형 악기상자

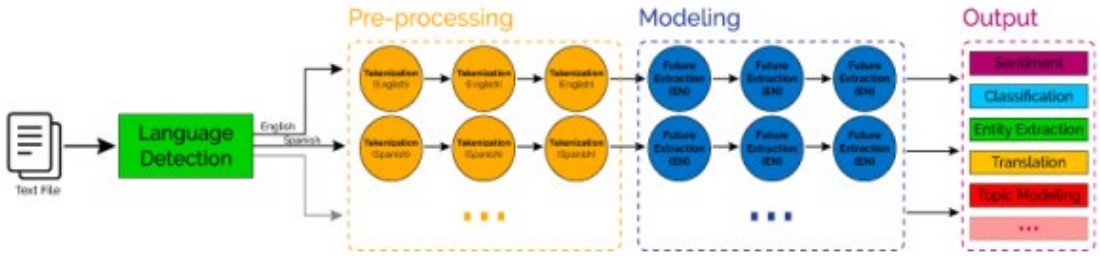
누구 도와준 사람 있어 ? ➔ 부인

어디에서 자가용 비행기를 탔어 ? ➔ 간사이 국제공항

HanBERT MRC Features

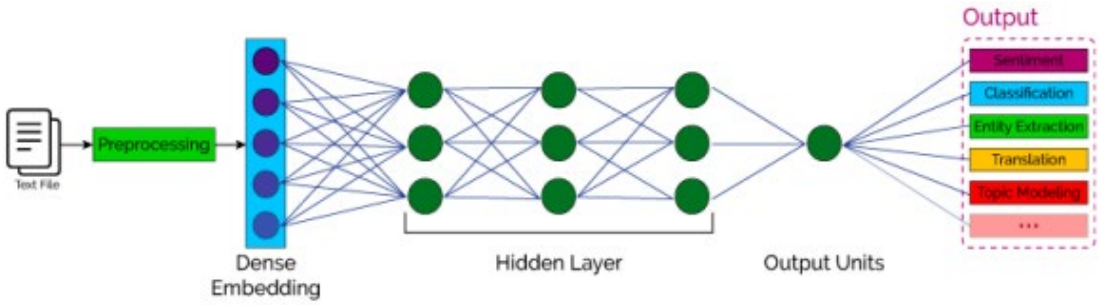
Classical NLP

사전, 규칙, 통계 : 정확하지만 범위 제한



Deep Learning

데이터, 학습장비 : 빠르고 만들기 쉬움

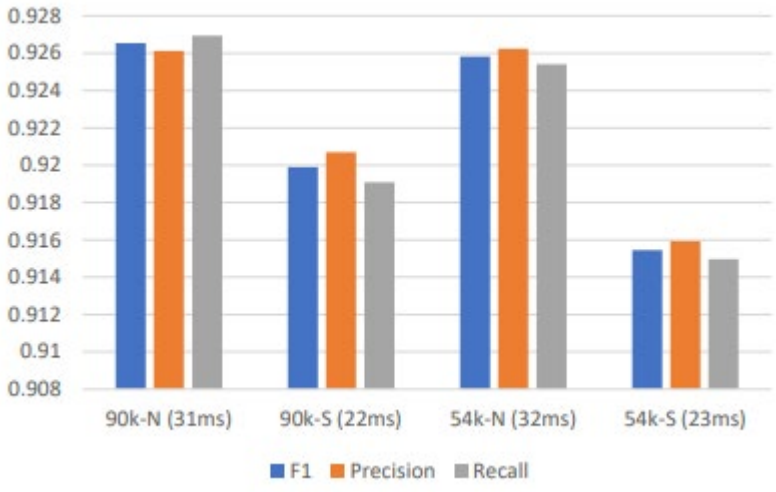


- 2019 Ai Starthone, 기계독해 분야 1위
- 2020 인공지능 온라인경진대회, 기계독해 1위 / 문자인식 1위로 종합 1위
- KoQuAD 1.0, 2.0 리더보드 리스팅
- 검증된 MRC 분야 최고 수준의 성능
- 띄어쓰기에 강한 언어모델 적용
- 딥러닝을 위한 한국어 표현(자사 특허) 모델 성능 향상
- RPAi 에 활용 가능 하도록 중요단어 사전 지원
- OCR 솔루션과 연동하여 OCR 성능 향상 (+% 3~ 5%)
- 고객 데이터 학습 지원
- Q&A 세트 추가 학습 지원

HanBERT_NER 개체명 인식

- 최신 대규모 개체명 인식 학습 데이터를 HanBERT에 적용하여, HanBERT 90k에서 최상의 성능을 발휘
- 500만 개체명 DB를 동시에 활용하여 Deep NLP에서 학습되지 않은 명칭어의 경우에도 추출
- 한국어 분석을 위한 MorAn과 연동되어 CPU/GPU 서버에서 수행되며, Cloud 서비스를 위한 API도 제공

HanBERT_NER은 최신 학습 데이터와 대규모 명칭어 DB를 활용 Domain 별 Vertical NER 구축 가능



HanBert_NER

Domain 별 NE 데이터베이스 Customizing 가능

개체명 인식

HanBERT_Sentiment 감성분석

- SNS Text 분석 , 기업브랜드 평판 조사 , 제품 후기 등 다양한 채널에서 수집되는 제품에 관한 고객의 소리를 분석
- Domain 별로 특화된 컨텐츠로 BERT 의 학습량이 많으면 성능이 점진적으로 향상
- 감성분석 외에도 다양한 Text 의 Classification, Opinion Mining 에 활용가능

(예) 네이버 영화평 DB에서 긍정과 부정의 판별

적용기술	성능	속도	학습비용
임의로 긍/부정 판별	50점	0ms	0원
Symbolic NLP 기술	75점 수준	50ms~100ms	1억원 수준
범용 딥러닝 기술을 활용	85점 수준	10ms~20ms	100만원
BERT에 추가학습	91점	15ms~20ms	10만원

Moran 형태소 분석기

C 언어로 작성되어 20년 넘게 다양한 레퍼런스를 보유하고 있으며, 표준 태그와도 호환되는 쉬운 튜닝.
한국어 분석의 기본이 되는 형태소 분석의 원천 기술, 자체 TRIE 구조, 융통성 있는 분석 결과 등을 보유.
초당 0.5~1.0MB의 처리 속도, Python, JAVA와 연동되며, 딥러닝을 위한 한국어 표현 방식 제공.

CJKE Text mining R&D Company
moransoft

Reference

MyScript, 네이버, 다음커뮤니케이션, SKT, LG CNS,
LG U+, 삼성전자 등

Technology

통합 TRIE 사전 구조 한글 전용 내부 코드
Weighted 문법 다단계 프로세싱

Flexibility

50만 기분석 사전 100만 형태소 사전 400만
명칭어 사전 분야별 사전

Specification

100~200M 메모리 초당 0.5~1MB 처리속도
Java, Pathon 연동 다양한 API 제공

TWOBLOCK Ai

Deep Learning

딥러닝을 위한 한국어 표기 방법 제공 (특허출원)

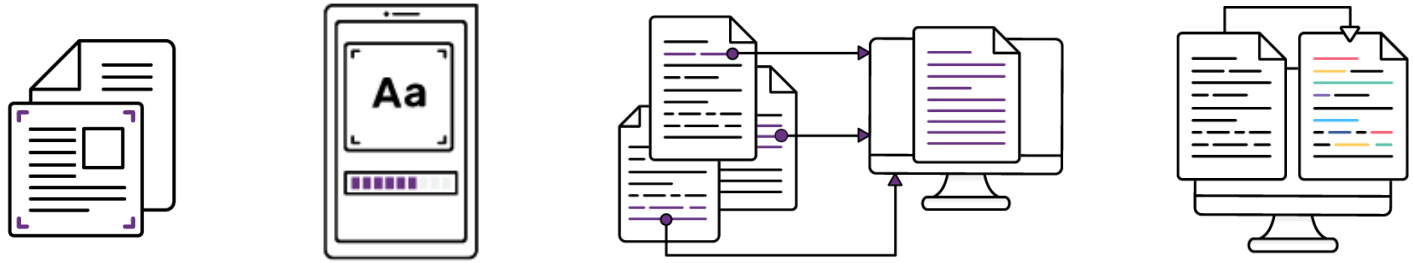
조사/어미 등의 기능어 구별이 가능
복합 명사를 구성하는 단어들의 위치 구별이 가능
형태소를 글자 단위로 분리해서 토큰화가 가능
분석 후 원문으로 복원하는 것이 가능

연세대학교학생이 농민 폭력 시위를 주도한 혐의로 지명수배 된 날은 ?

연세대 ~학교 ~학생 ~~이 농민 폭력 시위 ~~를 주도 ~~한 혐의 ~~로
지명 ~수배 된 날 ~~은

Moran_Post OCR, NLP Package

OCR 엔진과 결합하여 전체적인 문서의 인식율 제고.
언어모델 및 NLP 후처리 기술의 적용으로 인식오류 (띄어쓰기, 오탈자 등) 자동 제거 가능.
문서에 포함되어 있는 단어들과 언어모델을 참조하여 낱글자로 인식된 결과를 단어로 조합 수행.



Scanned Image > OCR 엔진 > NLP Post Processing > NLP Task

기재된 < 기재 된	손해배상을 < 손 해배상 을	용역계약서 < 용 역 계 약 서
변화하는 만큼 < 변화하는 만중	한국야쿠르트는 < 한국앙국르트는	주원료로 한 < 주원금로 한

NLP 적용시 OCR 인식율 +3%~+5% 향상가능 , 자동화 적용가능 수준으로 인식율 제고

OCR 엔진은 고객사 보유엔진을 사용합니다. 원하실 경우 제휴사의 엔진을 추천해드립니다

Post OCR NLP Package Features

대규모 코퍼스에서 기본 언어모델을 자료화, 분야별 코퍼스에서 해당 분야 언어모델을 자료화하여 활용.
문서에 포함되어 있는 단어들과 언어모델을 참조하여 낱글자로 인식된 결과를 단어로 조합하여 수행.
최종적으로 Moran의 기능을 활용해서 문장 단위 분리 작업, 철자 교정 작업을 수행.

| 4단계의 처리를 거쳐서 문자 인식된 문서의 오류를 자동 보정

대상 문서 언어모델	스캔된 문서에서 단어를 추출해서 대상 문서에 1차 적용
분야별 코퍼스 언어모델	법률, 경제, 특허 등의 분야 코퍼스에서 추출한 각 10M Patterns
대규모 일반 코퍼스 언어모델	총 110G 코퍼스에서 추출한 16M Patterns
Moran 문장 후처리	숫자, 날짜, 요일 등의 규칙형 토큰과 문장 분리, 철자 교정 수행

ButterBlock Suite 가격 정책

BUTTERBLOCK

심층 NLP Solotion

HanBERT

HanBERT Base 언어모델 라이센스

설치되는 서버의 GPU수,
기업규모,서비스 목적, 규모에 따라 최적화

Domain 전용 언어모델 개발

1

필요한 Deep NLP 추가모델
NER, MRC,Sentiment 등

2

Task를 위한 학습

3

학습용 데이터 정제

4

형태소 분석기 튜닝

5

고객 특화 Deep NLP 모델,
서비스용 API 등

6

고객 컨설팅 기획,
전략 등의 서비스

감사합니다.

주 소 서울특별시 서초구 남부순환로 350길 54 (양재동, 브이타워 1층) 전 화 02.6677.1111 이메일 info@twoblockai.com 홈페이지 www. twoblockai.com

(주)투블럭에이아이

말을 알아 듣는 Ai가 만들어 가는 편리한 세상, 사람에 대한 호기심이 세상을 바꿉니다.