

Part of speech:

NP NNP RB VBD IN NNP NNP CC PRP VBZ RB VBG PRP IN PRP .
Mrs. Clinton previously worked for Mr. Obama, but she is now distancing herself from him .

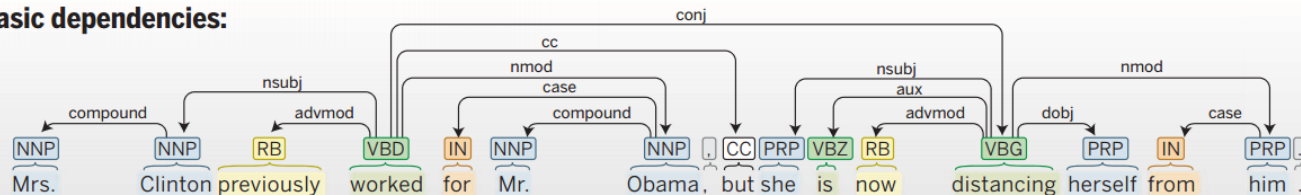
Named entity recognition:

Person Date Person Date
Mrs. Clinton previously worked for Mr. Obama, but she is now distancing herself from him.

Co-reference:

Mention Ment M Mention M
Mrs. Clinton previously worked for Mr. Obama, but she is now distancing herself from him.

Basic dependencies:



Lecture 2: Text Preprocessing

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

01 Introduction to NLP

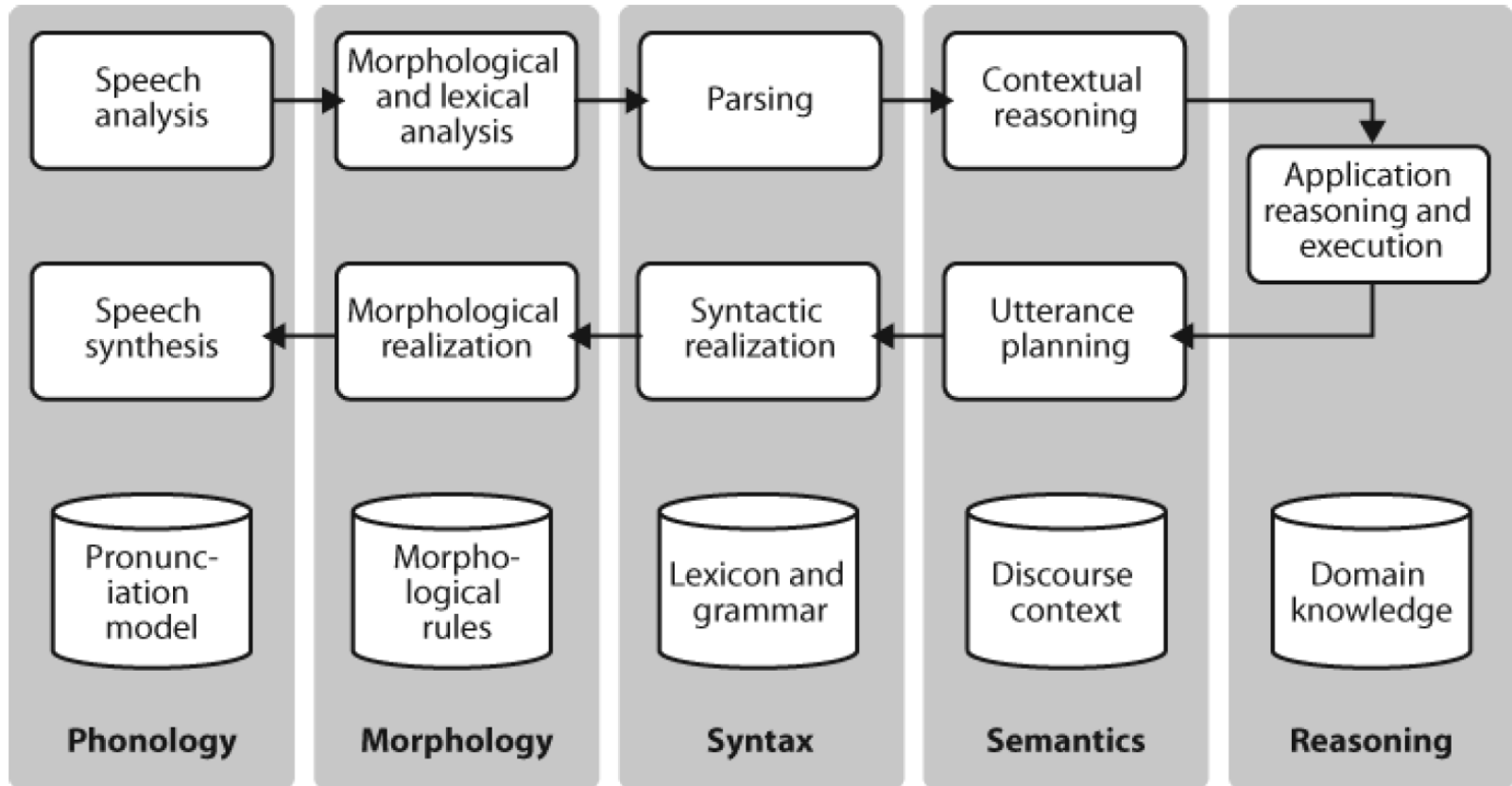
02 Lexical Analysis

03 Syntax Analysis

04 Other Topics in NLP

Natural Language Processing

- Natural language processing sequence



Natural Language Processing

Witte (2006)

- Classical categorization of NLP

Classical Categorization

To deal with the complexity of natural language, it is typically regarded on several levels (cf. Jurafsky & Martin):

Phonology the study of linguistic sounds

Morphology the study of meaningful components of words

Syntax the study of structural relationships between words

Semantics the study of meaning

Pragmatics the study of how language is used to accomplish goals






Discourse the study of larger linguistic units

Importance for Text Mining

- *Phonology* only concerns spoken language
- *Discourse*, *Pragmatics*, and even *Semantics* is still rarely used

Natural Language Processing

- Phonology is the first gate of AI solutions

 <p>음성 인식 기술 비교</p>	 <p>무엇을 도와드릴까요?</p> <p>애플 시리(Siri)</p>	 <p>Google</p> <p>"Ok Google" 말하기</p> <p>구글 구글 나우(Now)</p>	 <p>Here are today's top headlines.</p> <p>SpaceX to ferry astronauts to space station USA Today · 8 hours ago</p> <p>News from the web</p> <p>See all</p> <p>마이크로소프트 코타나(Cortana)</p>	 <p>아마존 알렉사(Alexa)</p>
명칭	출시일	출시일	출시일	출시일
사용 가능 언어	영어·중국어·한국어 등 13개 언어	영어·일본어·스페인어·한국어 등 9개 언어	영어·중국어·스페인어 등 7개 언어(※한국어 사용 불가)	영어(※한국어 사용 불가)
사용처	아이폰·아이패드에 탑재	구글 안드로이드 스마트폰에 탑재	윈도10 운영체제(OS)에 탑재	음성 인식용 스피커 에코에 탑재
주요 기능	-음성으로 스마트폰 조작 -자주 사용하는 앱 추천	-음성으로 스마트폰 조작 -구글의 검색·지도 등 다른 서비스와 연동	-평온·당황 등의 감정을 이모티콘으로 표현 -윈도10 OS를 쓰는 기기에 모두 기본 탑재	-가정용 제품으로 설계 -생활 소음 속에서도 목소리 잘 인식 -아마존 서비스와 연동해 물건 주문 가능

Natural Language Processing

- Speech to Text (STT)



build passing Kaldi Speech Recognition Toolkit

To build the toolkit: see `./INSTALL`. These instructions are valid for UNIX systems including various flavors of Linux; Darwin; and Cygwin (has not been tested on more "exotic" varieties of UNIX). For Windows installation instructions (excluding Cygwin), see `windows/INSTALL`.

To run the example system builds, see `egs/README.txt`

If you encounter problems (and you probably will), please do not hesitate to contact the developers (see below). In addition to specific questions, please let us know if there are specific aspects of the project that you feel could be improved, that you find confusing, etc., and which missing features you most wish it had.

Kaldi information channels

For HOT news about Kaldi see [the project site](#).

Documentation of Kaldi:

- Info about the project, description of techniques, tutorial for C++ coding.
- Doxygen reference of the C++ code.

Kaldi forums and mailing lists:

We have two different lists

- User list [kaldi-help](#)
- Developer list [kaldi-developers](#):

To sign up to any of those mailing lists, go to <http://kaldi-asr.org/forums.html>:

<https://github.com/kaldi-asr/kaldi>



Natural Language Processing

- Top 16 Speech Recognition Startups (2020.02.06)

1 Mobvoi



Country: China | **Funding:** \$252.7M
Mobvoi is an AI company that developed Chinese voice recognition, natural language processing, and vertical search technology in-house.

2 SoundHound



Country: **USA** | Funding: **\$115M**
SoundHound develops voice-enabled AI and conversational intelligence technologies. It provides Speech-to-Meaning engine and Deep Meaning Understanding technology that can be built in other services and devices. It also develops app for music recognition and voice assistant for search.

3 Liulishuo



Country: **China** | Funding: **\$100M**
Through cutting-edge AI technology & innovative product design, Liulishuo helps users learn English more efficiently and communicate with the world.

4 Invoca



Country: **USA** | Funding: **\$60.8M**
 Invoca provides complete call intelligence for business. Its machine learning algorithms analyze live phone conversations to understand caller intent and outcomes. Marketers can utilize these insights to make smarter decisions on everything from PPC bidding strategy to digital retargeting audiences.

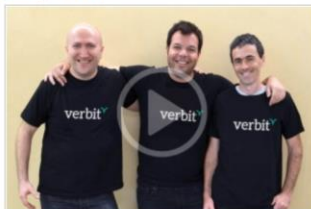
5 Dialogflow



Country: USA | **Funding:** \$8.6M

Dialogflow is a conversational user experience platform enabling natural language interactions for devices, applications and services. Developers can use Dialogflow services for speech recognition, natural language processing (intent recognition and context awareness), and conversation management to quickly and easily differentiate their business, increase customer satisfaction and improve business processes.

6 Verbit



Country: **Israel** | Funding: **\$34M**
Verbit is using smart AI technology to disrupt transcription and captioning with automation and speed.

7 Speechmatics



Country: **UK**
Speechmatics provides automatic speech recognition technologies that can be used anywhere, by anyone, in any language.

8 Notable



Country: **USA** | Funding: **\$19.2M**
 Notable uses AI to automate and digitize every physician-patient interaction. It automates recording of doctor's visits and updating of electronic health records. The company has developed a technology that uses natural language processing and voice recognition to automatically record doctor-patient interactions and structure the data for inclusion in a patient's medical records.



Natural Language Processing

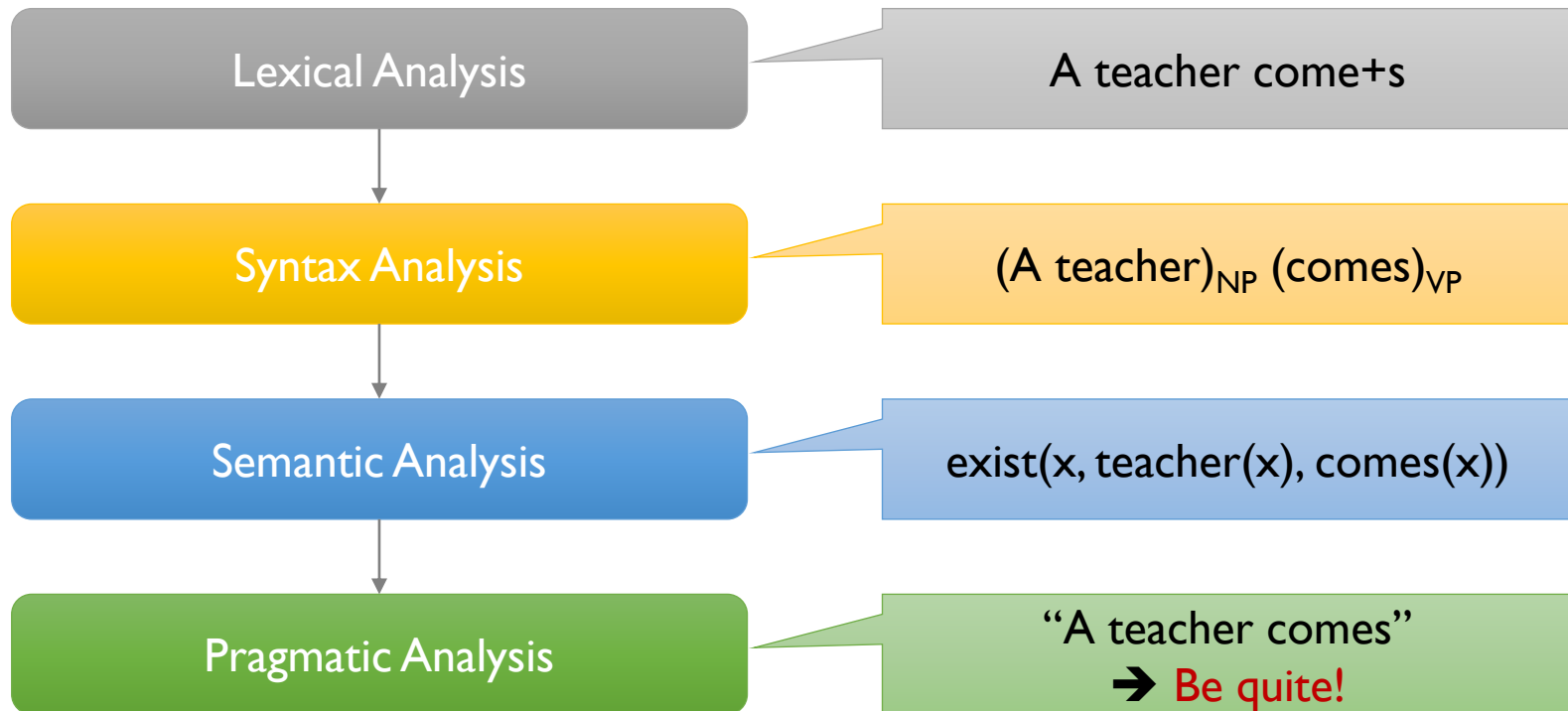
- Text to Speech (TTS) Example



Natural Language Processing

Witte (2006)

- An example of NLP

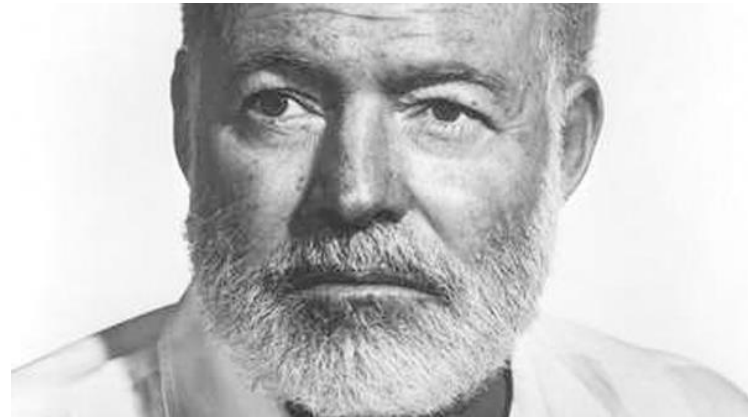


Natural Language Processing

- Is Pragmatic Analysis Possible?



“여섯 단어로 우리를 울릴만한
소설을 써 보시지?”



"For sale:

Baby shoes. Never worn."

E. Hemingway

Is NLP Easy? No!

Witte (2006)

- Why is NLP hard?

Difference to other areas in Computer Science

Computer scientist are used to dealing with precise, closed, artificial structures

- e.g., we build a “mini-world” for a database rather than attempting to model every aspect of the real world
- programming languages have a simple syntax (around 100 words) and a precise semantic

This approach does not work for natural language:

- tens of thousands of languages, with more than 100 000 words each
- complex syntax, many ambiguities, constantly changing and evolving

A corollary is that a TM system will never get it “100% right”

Is NLP Easy? No!


- Programming Language

```
17 from __future__ import absolute_import
18 from __future__ import division
19 from __future__ import print_function
20
21 import re
22 import tensorflow as tf
23
24
25 def create_optimizer(loss, init_lr, num_train_steps, num_warmup_steps, use_tpu):
26     """Creates an optimizer training op."""
27     global_step = tf.train.get_or_create_global_step()
28
29     learning_rate = tf.constant(value=init_lr, shape=[], dtype=tf.float32)
30
31     # Implements linear decay of the learning rate.
32     learning_rate = tf.train.polynomial_decay(
33         learning_rate,
34         global_step,
35         num_train_steps,
36         end_learning_rate=0.0,
37         power=1.0,
38         cycle=False)
```

<https://github.com/google-research/bert/blob/master/optimization.py>

Is NLP Easy? No!

- How to annoy graduate students with four lines of Python code



```
import tensorflow as plt  
import pandas as np  
import numpy as tf  
import matplotlib.pyplot as pd
```

Is NLP Easy? No!

Witte (2006)

- Ambiguity of a natural language

Ambiguity appears on every analysis level

The classical examples:

- *He saw the man with the telescope.*
- *Time flies like an arrow. Fruit flies like a banana.*

And those are simple...

This does not get better with real-world sentences:

- *The board approved [its acquisition] [by Royal Trustco. Ltd.] [of Toronto] [for \$27 a share] [at its monthly meeting].*

(cf. Manning & Schütze)

Is NLP Easy? No!

Witte (2006)

- Complex and subtle relationship between concepts in texts
 - ✓ “AOL merges with Time-Warner”
 - ✓ “Time-Warner is bought by AOL”
- Ambiguity and context sensitivity
 - ✓ automobile = car = vehicle = Hyundai



vs.



Research Trends in NLP

Witte (2006)

- From rule-based approaches to statistical approaches

The classical way: until late 1980's

Rule-based approaches:

- are too rigid for natural language
- suffer from the *knowledge acquisition bottleneck*
- cannot keep up with changing/evolving language
ex. "to google"

The statistical way: since early 1990's

"Statistical NLP" refers to all quantitative approaches, including Bayes' models, Hidden Markov Models (HMMs), Support Vector Machines (SVMs), Clustering, ...

- more robust & more flexible
- need a *Corpus* for (supervised or unsupervised) learning

But real-world systems typically combine both.

Research Trends in NLP

Collobert et al. (2011)

- From statistical approaches to machine-learning (deep-learning) approaches

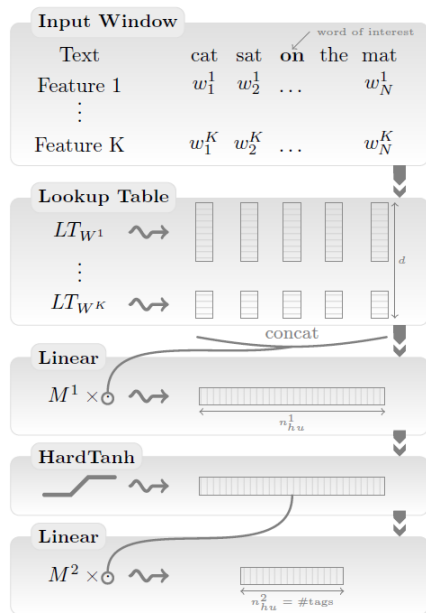


Figure 1: Window approach network.

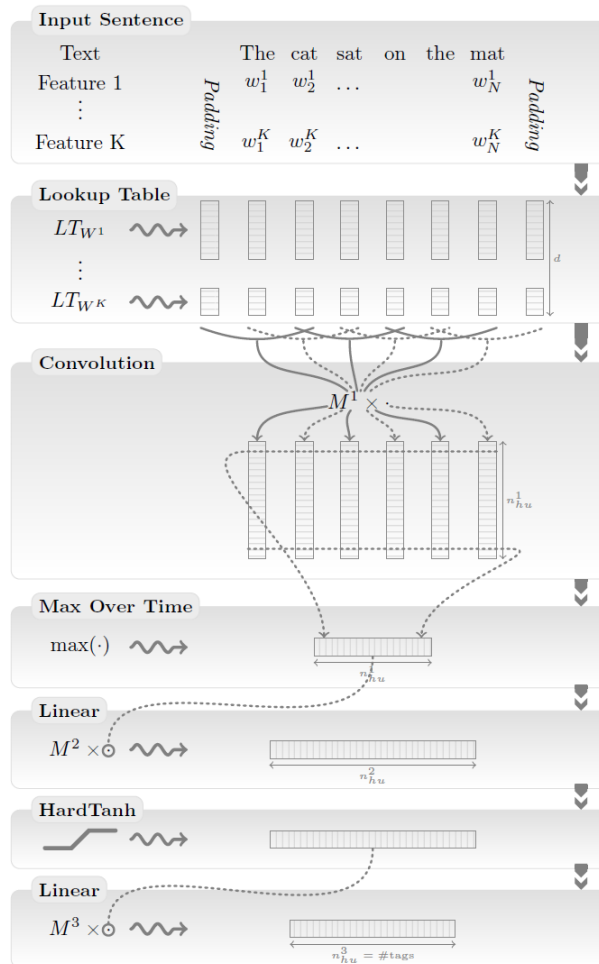


Figure 2: Sentence approach network.

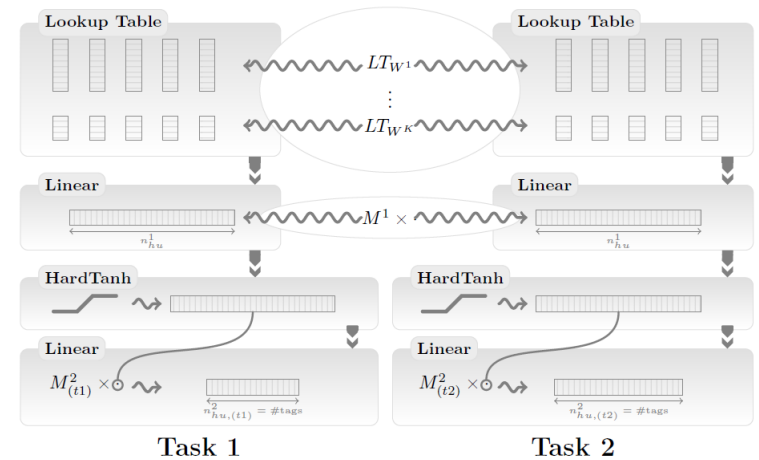


Figure 5: Example of multitasking with NN. Task 1 and Task 2 are two tasks trained with the window approach architecture presented in Figure 1. Lookup tables as well as the first hidden layer are shared. The last layer is task specific. The principle is the same with more than two tasks.

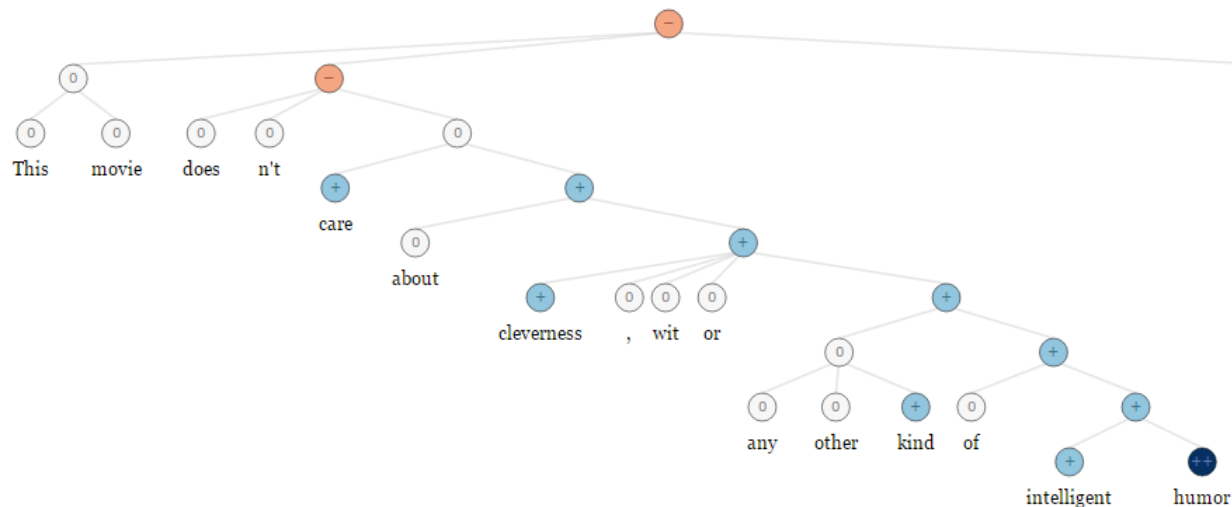
Research Trends in NLP

Socher et al. (2013)

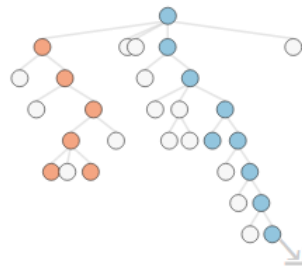
- From statistical approaches to machine-learning (deep-learning) approaches

Sentiment Trees

You can double-click on each tree figure to see its expanded version with greater details. There are 5 classes of sentiment classification: **very negative**, **negative**, **neutral**, **positive**, and **very positive**.



All labels are now correct



<http://nlp.stanford.edu:8080/sentiment/rntnDemo.html>

Research Trends in NLP

Wu et al. (2016)

- From statistical approaches to machine-learning (deep-learning) approaches

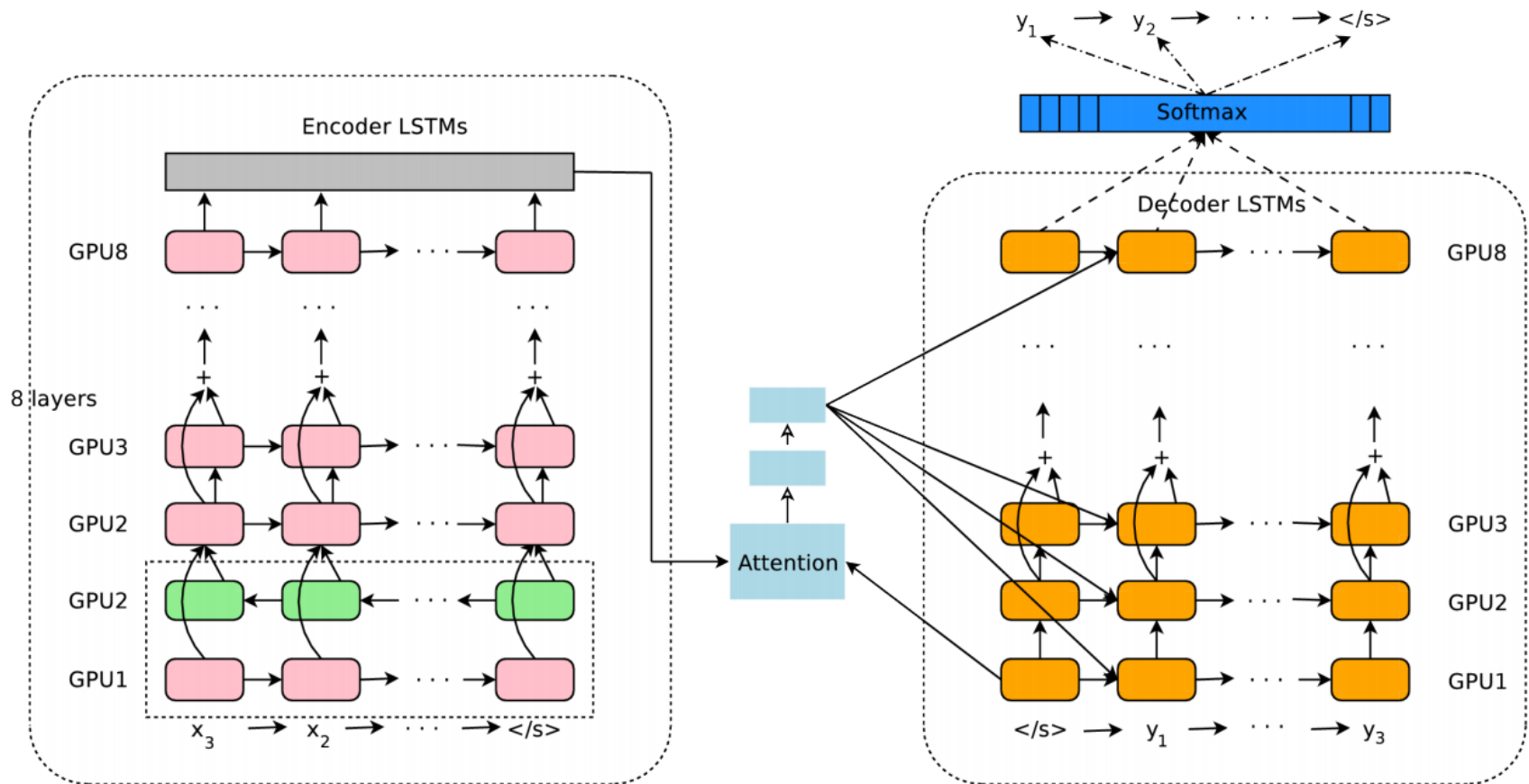
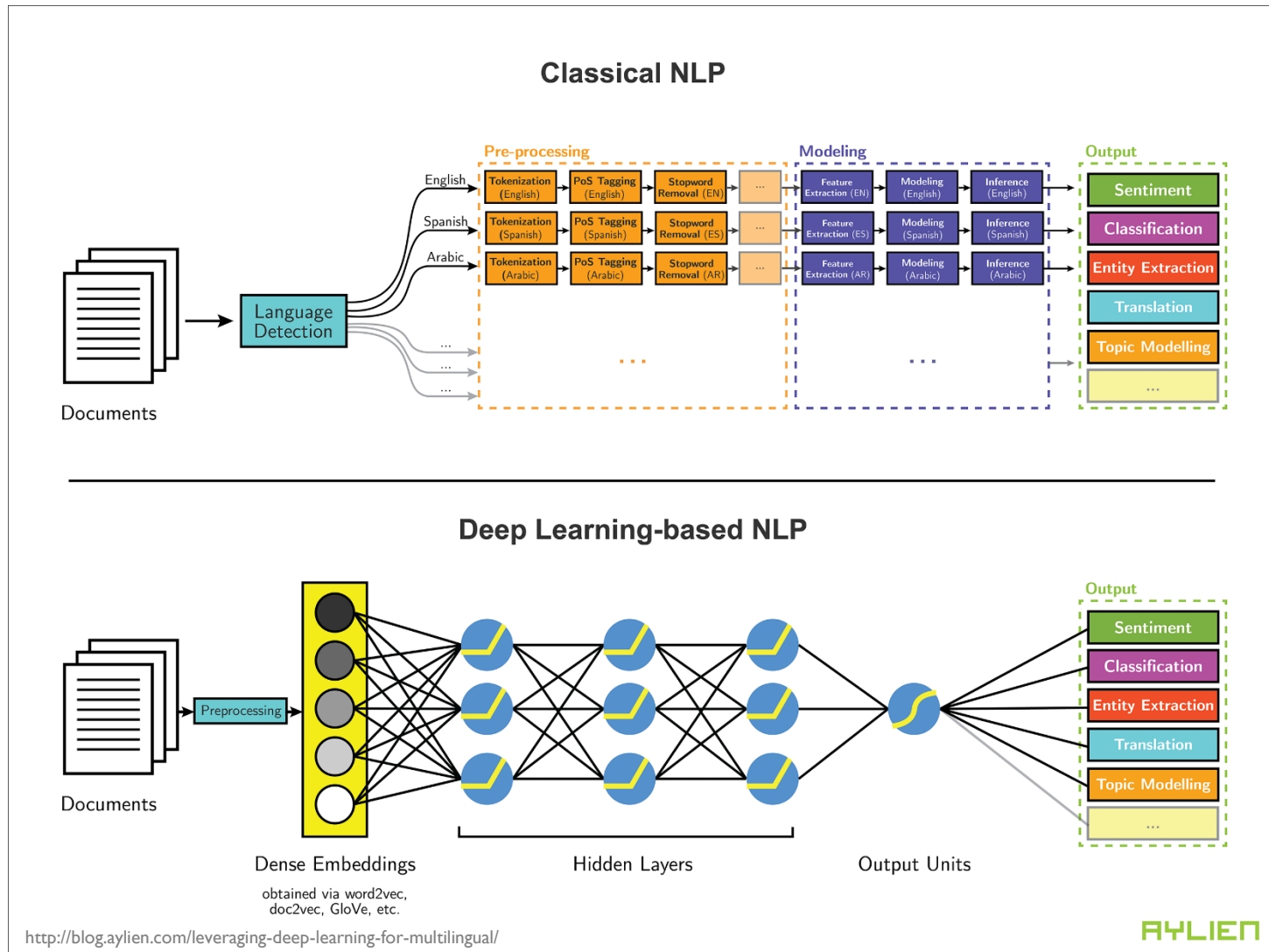


Figure 1: The model architecture of GNMT, Google's Neural Machine Translation system. On the left

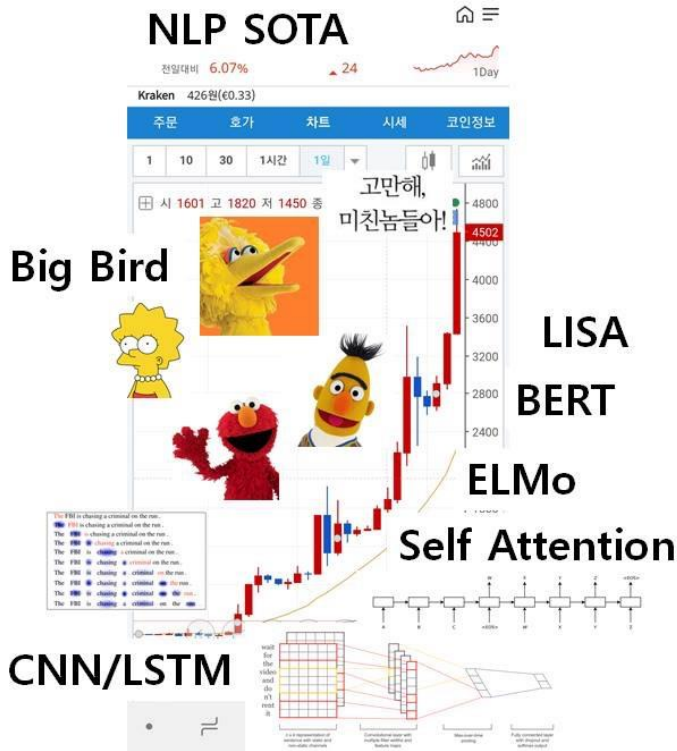
Research Trends in NLP

- End-to-End Multi-Task Learning



Research Trends in NLP

- Performance Improvements



Browse > Natural Language Processing

Natural Language Processing

434 leaderboards • 232 tasks • 100 datasets • 3563 papers with code

Representation Learning



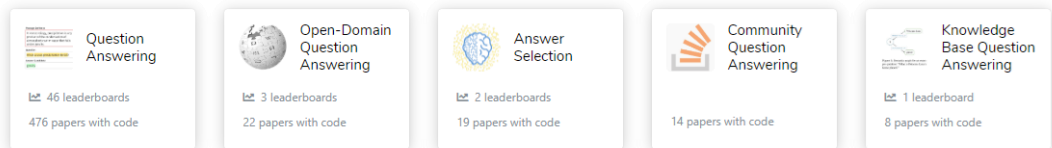
See all 17 tasks

Machine Translation

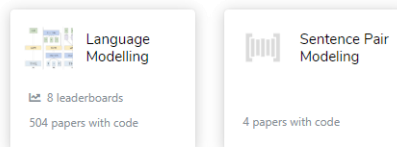


See all 6 tasks

Question Answering

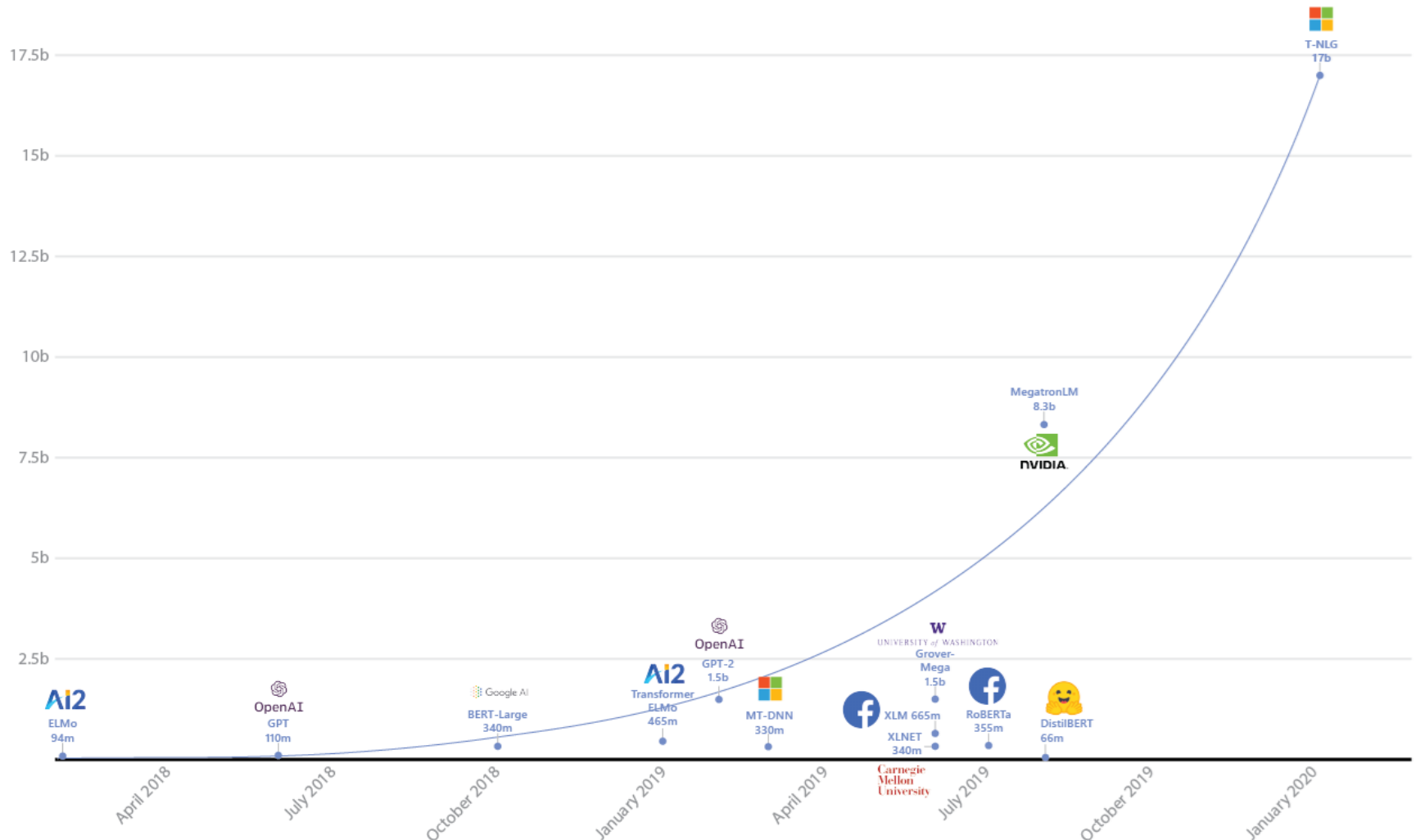


Language Modelling



Research Trends in NLP

- Performance Improvements with a huge model



Research Trends in NLP

• An Era of optimization???

수학적 최적화(Mathematical Optimization·MO)
수퍼 컴퓨터가 고차원의 수학 방정식을 이용해 현재 주어진 상황에서 최적의 답을 찾아내는 기술

최적해:
최대 매출에 필요한 생산량

현재해

매출

생산량

수학적 최적화(MO)와 인공지능(AI) 비교

수학적 최적화	구분	인공지능
제조·물류·에너지·교통 등 최적의 해결책이 필요한 분야	적용 분야	얼굴·이미지·음성 분석 등 빅데이터 분석 필요한 분야
수학 방정식으로 최적의 답 계산 (연역 추론)	방식	과거 데이터 분석해 유사성 또는 향후 추세 분석(귀납 추론)
-시간·비용이 상대적으로 적게 들 -새 변수에 바로 대응 가능	장점	-데이터 양 많을수록 정교해짐 -과거에 못 풀던 문제 해결 가능
수학 공식화 어려운 분야에 적용 힘들어	단점	분석 결과가 100% 맞지 않음

수학적 최적화 적용 사례

OLED(유기 발광 다이오드) 생산 라인

기존 방식

수학적 최적화 방식

유기물

디스플레이 글라스(유리)

글라스(유리) 위 수백만 개 소자에 유기물을 붙이는 증착 과정이 중요. 글라스 투입 위치를 수학으로 계산해 수율 높임

MO는 고성능 컴퓨터에 기반해 복잡한 연산을 한다는 점에서 AI와 비슷하다. 하지만 과거의 빅데이터를 분석해 최선의 답을 도출하는 AI와 달리 MO는 수학 공식으로 현재 주어진 한정된 조건에서 가장 이상적인 해결책을 찾아준다. AI가 경험을 바탕으로 가장 나은 방안을 제시한다면, MO는 수학 이론을 동원해 최적 답을 알려주는 것이다. MO는 AI처럼 빅데이터를 분석하는 과정이 필요 없기 때문에 시간·비용이 상대적으로 적게 든다.

최신순 찬성순 반대순

김병인 (betteren****)

좋은 기사 감사합니다. 하지만
업공학에서 경영과학(Operat
어 왔고, 산업체에 효과도 많
와 같은 기업에서도 많이 활

댓글



김병인

박사학위 : Rensselaer Polytechnic Ins.
전공분야 : 물리최적화, OR 응용
연구실 : 물류 연구실

가

신고

동의하기 힘들습니다. MO는 산
는 이름으로 수십년간 강의되
시스템틀일 것입니다. 포스코

15

0

Research Trends in NLP

- 10 Exiting ideas of 2018 in NLP (<http://runder.io/10-exciting-ideas-of-2018-in-nlp/>)
 - ✓ Unsupervised Machine Translation
 - ✓ Pretrained language models
 - ✓ Common sense inference datasets
 - ✓ Meta-learning
 - ✓ Robust unsupervised methods
 - ✓ Understanding representations
 - ✓ Clever auxiliary tasks
 - ✓ Combining semi-supervised learning with transfer learning
 - ✓ QA and reasoning with large documents
 - ✓ Inductive bias

Research Trends in NLP

- Major NLP Achievements & Papers from 2019
 - ✓ Language Models Are Unsupervised Multitask Learners
 - ✓ XLNet: Generalized Autoregressive Pretraining for Language Understanding
 - ✓ RoBERTa: A Robustly Optimized BERT Pretraining Approach
 - ✓ Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts
 - ✓ Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems
 - ✓ Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks
 - ✓ Probing the Need for Visual Context in Multimodal Machine Translation
 - ✓ Bridging the Gap between Training and Inference for Neural Machine Translation
 - ✓ On Extractive and Abstractive Neural Document Summarization with Transformer Language Models
 - ✓ CTRL: A Conditional Transformer Language Model For Controllable Generation
 - ✓ ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

Research Trends in NLP

- 14 NLP research breakthrough you can apply to your business
 - ✓ BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
 - ✓ Sequence Classification with Human Attention
 - ✓ Phrase-Based & Neural Unsupervised Machine Translation
 - ✓ What you can cram into a single vector: Probing sentence embeddings for linguistic properties
 - ✓ SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference
 - ✓ Deep contextualized word representations
 - ✓ Meta-Learning for Low-Resource Neural Machine Translation
 - ✓ Linguistically-Informed Self-Attention for Semantic Role Labeling
 - ✓ A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks
 - ✓ Know What You Don't Know: Unanswerable Questions for SQuAD
 - ✓ An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling
 - ✓ Universal Language Model Fine-tuning for Text Classification
 - ✓ Improving Language Understanding by Generative Pre-Training
 - ✓ Dissecting Contextual Word Embeddings: Architecture and Representation

Research Trends in NLP

- Statistical translation vs. deep learning-based translation

Translate Text

Input

Enter or paste text from a passage.

Korean



Output

Copy output from this field to clipboard.

English



Text

Rest API

흔히 기업·가계 등 경제 주체와 금융시장이 가장 싫어하는 건 '불확실성'이라고 한다. 주요국 주식시장은 트럼프 미국 대통령 당선에 따른 경제회복 기대감으로 아직 상승하고 있지만, 정책당국과 경제연구기관·학계 모두 우리 경제 안팎에 '불확실성 충격'이 엄습하고 있다고 한목소리를 낸다. 트럼프발 정책 불확실성과 혼선, 미·중 통상마찰 심화 등 정치경제적 혼돈과 급변이 '불확실성 시대'의 실체로 지목된다. 전문가들은 "지금의 불확실성은 통제하기 어려운 요인에서 비롯되고 있다"며 "사전에 결과를 예측하기도 어렵고, 시장의 예측과 상반되는 결과가 나타날 가능성이 크므로 다양한 관점에서의 정보 획득이 필요한 때"라고 말한다.



Text

JSON

Often, companies and households as the economic and financial market hates the most is uncertainty it is called " X-inefficiency. Brazilian stocks fell as hopes for an economic recovery with Trump in the US presidential election is still rising, but both government policy-makers and economic research institutes and universities have shelters the 'uncertainty shock' and our economy. With TeuReomPeuBal, the China-US trade policy uncertainty and confusion entity of class variation " era of uncertainty with a deep political and economic chaos. Experts said the current uncertainty is difficult to control factors arising from competing, and

Research Trends in NLP

- Statistical translation vs. deep learning-based translation

[로그인](#)

번역

즉석 번역 사용 안함



한국어 영어 독일어 한국어 - 감지됨

↔

영어 한국어 일본어





번역하기


흔히 기업·가계 등 경제 주체와 금융시장이 가장 싫어하는 건 '불확실성'이라고 한다. 주요국 주식시장은 트럼프 미국 대통령 당선에 따른 경제회복 기대감으로 아직 상승하고 있지만, 정책당국과 경제연구기관·학계 모두 우리 경제 안팎에 '불확실성 충격'이 엄습하고 있다고 한목소리를 낸다. 트럼프발 정책 불확실성과 혼선, 미·중 통상마찰 심화 등 정치경제적 혼돈과 급변이 '불확실성 시대'의 실체로 지목된다. 전문가들은 "지금의 불확실성은 통제하기 어려운 요인에서 비롯되고 있다"며 "사전에 결과를 예측하기도 어렵고, 시장의 예측과 상반되는 결과가 나타날 가능성이 크므로 다양한 관점에서의 정보 획득이 필요한 때"라고 말한다.



348/5000

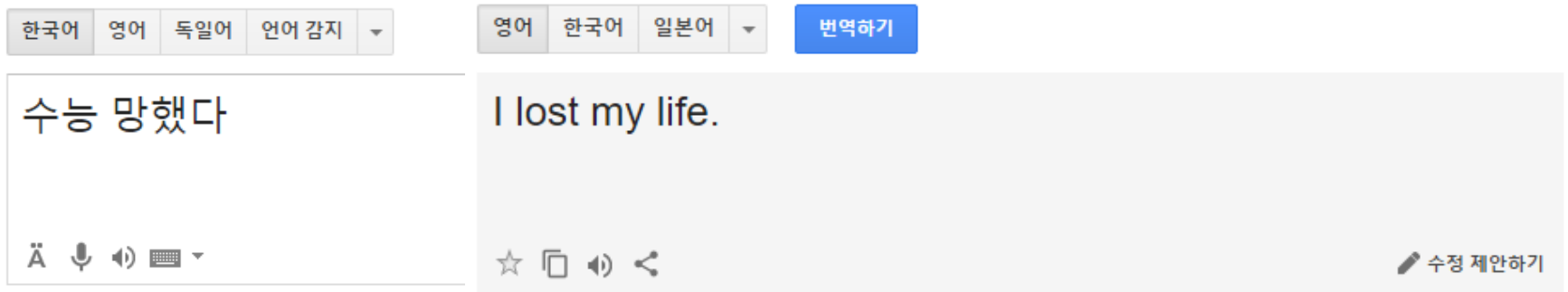
It is often said that economic entities such as corporations and households and financial markets hate the 'uncertainty'. Although the stock market of major countries is still rising due to the expectation of economic recovery following the election of US President George W. Bush, policy makers, economic research institutes and academics both call for a "shock of uncertainty" inside and outside of our economy. Uncertainty, confusion, deepening frictions between the US and China, and political and economic chaos and sudden changes are pointing to the reality of the 'uncertainty age'. Experts say that "uncertainty now comes from factors that are difficult to control", "it is difficult to predict the outcome in advance, and there is a high possibility that the results will be in conflict with the market forecast. It says.



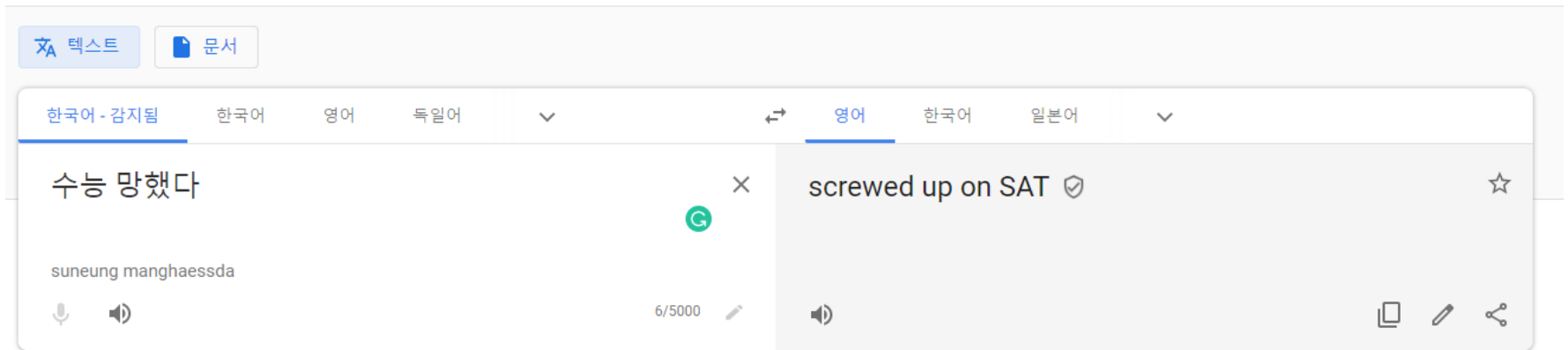
 수정 제안하기

Research Trends in NLP

- Provide your inputs to improve the machine translator!

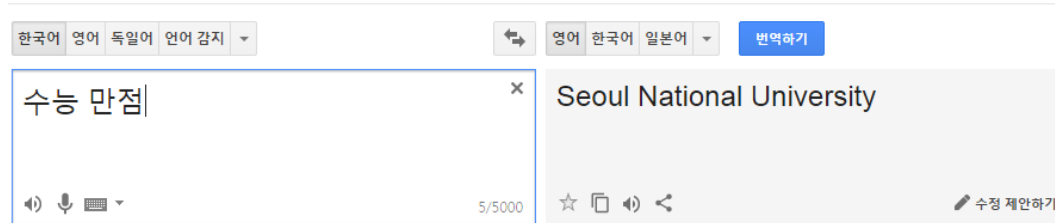


- 2018.03.06 & 2019.03.02 & 2020.03.02

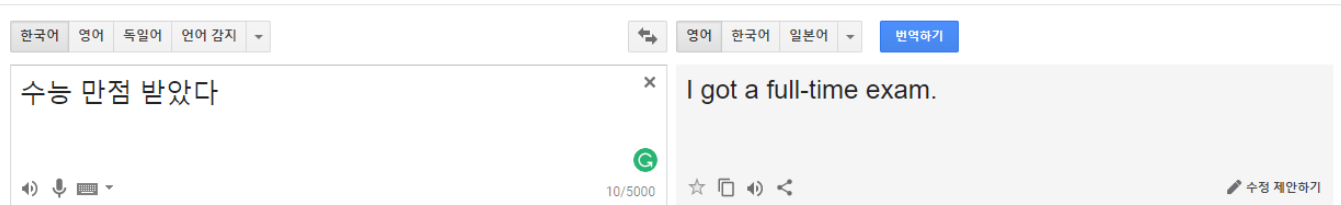


Research Trends in NLP

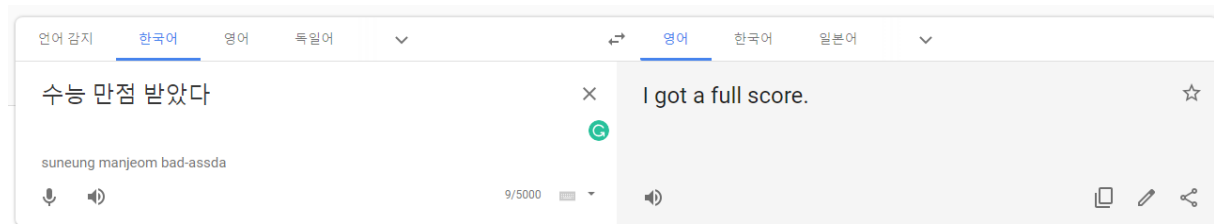
- Provide your inputs to improve the machine translator!



- 2018.03.06

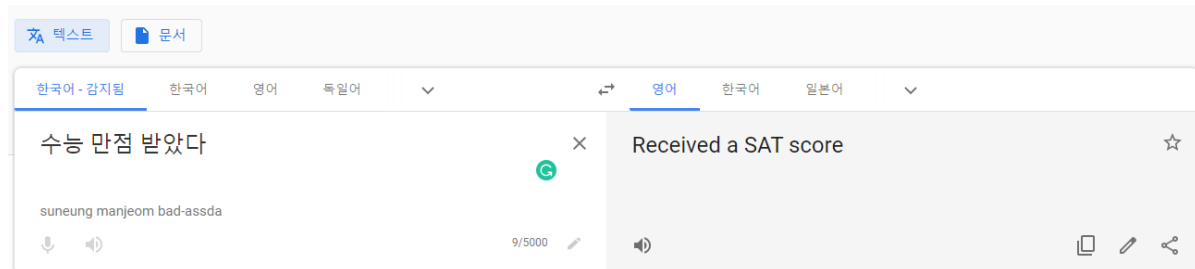


- 2019.03.02



의견 보내기

- 2020.03.02



의견 보내기

Data Quality in NLP

학습 모델 및 데이터 목록

- ExoBrain Project

한국어 BERT 언어모델	언어처리 학습데이터	음성 학습데이터	객체검출 학습데이터
---------------	------------	----------	------------

언어처리 학습데이터

한국어 분석 및 질의응답 기술을 개발하기 위한 과학기술정보통신부 소프트웨어 분야 R&D인 엑소브레인 과제에서는 다양한 지식산업 환경에서 전문가 수준의 질의응답 서비스의 제공을 위하여 ETRI, 울산대, KAIST, 충북대, 강원대 등 국내 여러 연구기관이 힘을 합쳐 연구하고 있습니다.

엑소브레인 과제를 수행하면서 ETRI와 함께 각 연구기관에서 구축한 언어처리 학습데이터(엑소브레인 말뭉치 v4.0)를 공개하여 유사 분야 연구에 도움이 되고자 합니다. 공개하는 엑소브레인 말뭉치 v4.0은 아래와 같이 구성되어 있습니다.

엑소브레인 QA Datasets (ETRI)	퀴즈 QA Datasets <ul style="list-style-type: none">퀴즈 분야 질의응답을 위한 4개 유형 (객관식/주관식/가부형/연상형)의 퀴즈 QA datasets (569개)
	SQuAD 한국어 QA Dataset <ul style="list-style-type: none">SQuAD 질문의 위키피디아 한국어 번역 QA datasets (표준태깅, 3397개) MRC 한국어 QA Dataset <ul style="list-style-type: none">한국어 위키피디아를 대상으로 구축한 MRC(Machine Reading Comprehension) QA datasets(10,000개)
	위키피디아 단문질문 QA Datasets <ul style="list-style-type: none">상/중/하 난이도 별 패러프레이즈 QA datasets(표준태깅, 3007개)일반상식 분야 QA dataset(기본 태깅, 1,776개)
엑소브레인 언어분석 말뭉치 (ETRI/강원대)	언어분석 통합 말뭉치 <ul style="list-style-type: none">언어분석 6개 기술(형태소분석, 다의어 어휘의미분석, 세분류 개체명인식, 의존구문분석, 의미역인식, 상호참조해결)의 태깅 가이드라인과 자연어 질의응답을 위한 질문/정답 포맷의 뉴스기사 대상 태깅 말뭉치 (2,593문장, 33,131어절)
	세부기술 별 말뭉치 <ul style="list-style-type: none">개체명 인식 태깅 가이드라인 및 말뭉치 (인명/장소/조직/날짜/시간 5개 태그, 10,000 문장)의미역 인식 태깅 가이드라인 및 말뭉치 (625문장, 7,436어절)의존구문분석 태깅 가이드라인 및 말뭉치 (2,225문장, 27,317어절) TTA 공인인증 말뭉치 (추후제공예정) <ul style="list-style-type: none">엑소브레인 과제 결과물의 객관적인 성능 측정을 위해 2018년 한국정보통신기술협회(TTA)에서 시행한 공인인증 평가에 사용된 평가셋(개체명 인식: 459문장, 의미역 인식(필수역 대상): 450문장)
UCorpus-HG 말뭉치 (울산대학교)	울산대학교 형태/의미 말뭉치(UCorpus-HG) <ul style="list-style-type: none">표준국어대사전 기반 모든 동형이의어 대상으로 어개번호를 부착한 말뭉치원문: 세종 형태의미 말뭉치, 신문, 초등학교 국어교과서, 법률, 사전뜻풀이/용례전체 1,909,840 문장, 18,869,517 어절 (학습 말뭉치 90%, 평가 말뭉치 10%로 분리 제공)
엑소브레인 Korean TimeBank 및 SpaceBank (KAIST/충북대)	한국어 시간 정보 주석 말뭉치: Korean TimeBank (KAIST) <ul style="list-style-type: none">한국어 시간 정보 자동 추출을 위한 한국어 시간 정보 주석 가이드라인 및 말뭉치말뭉치 구성: 812 문서, 5,467 문장 태깅 (시간개체 4,509개, 시간관계 5,182개)
	한국어 공간 정보 주석 말뭉치: Korean SpaceBank (충북대학교) <ul style="list-style-type: none">한국어 공간 정보 자동 추출을 위한 한국어 공간 정보 주석 가이드라인 및 말뭉치말뭉치 구성 : 개체 태그 7종류, 관계 태그 4종류, 2,264 문장 태깅
엑소브레인 패러프레이즈 말뭉치 (KAIST)	한국어 패러프레이즈 말뭉치: Korean Paraphrase Corpus(KAIST) <ul style="list-style-type: none">한국어 패러프레이즈 인식 및 평가를 위한 주석 가이드라인 및 말뭉치말뭉치 구성: 패러프레이즈 관계 2,000문장 쌍과 출처, 유사도(0-5)/난이도(상/중/하) 표준 태깅, 의미(실질) 형태소 정보 태깅

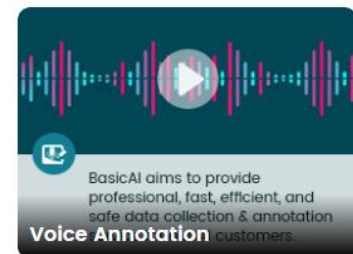
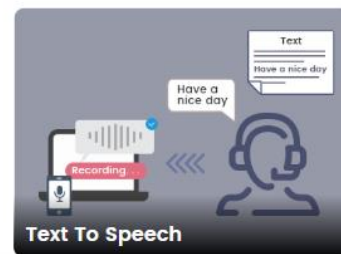
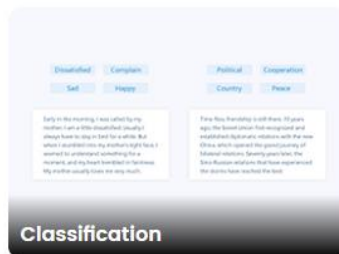
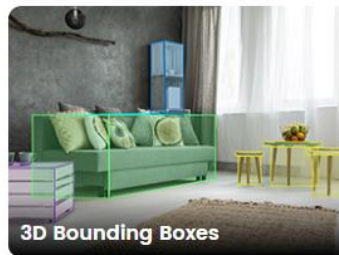
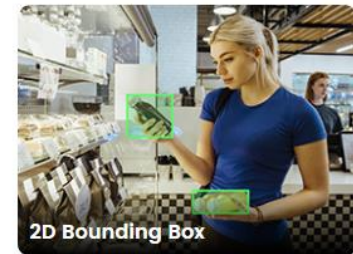
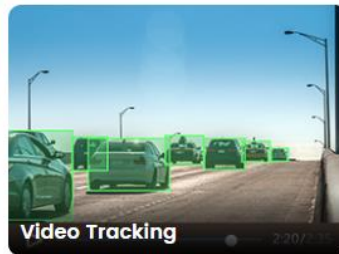
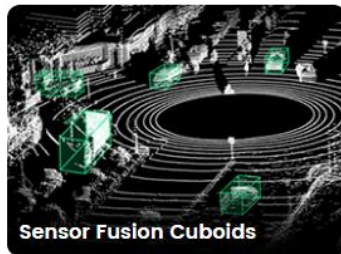
Data Quality in NLP

- Data Annotation as a Business Model

- ✓ Scale AI: <https://scale.com/>

- ✓ Basic AI: <https://www.basic.ai/>

Data Labeling Services



Data Quality in NLP

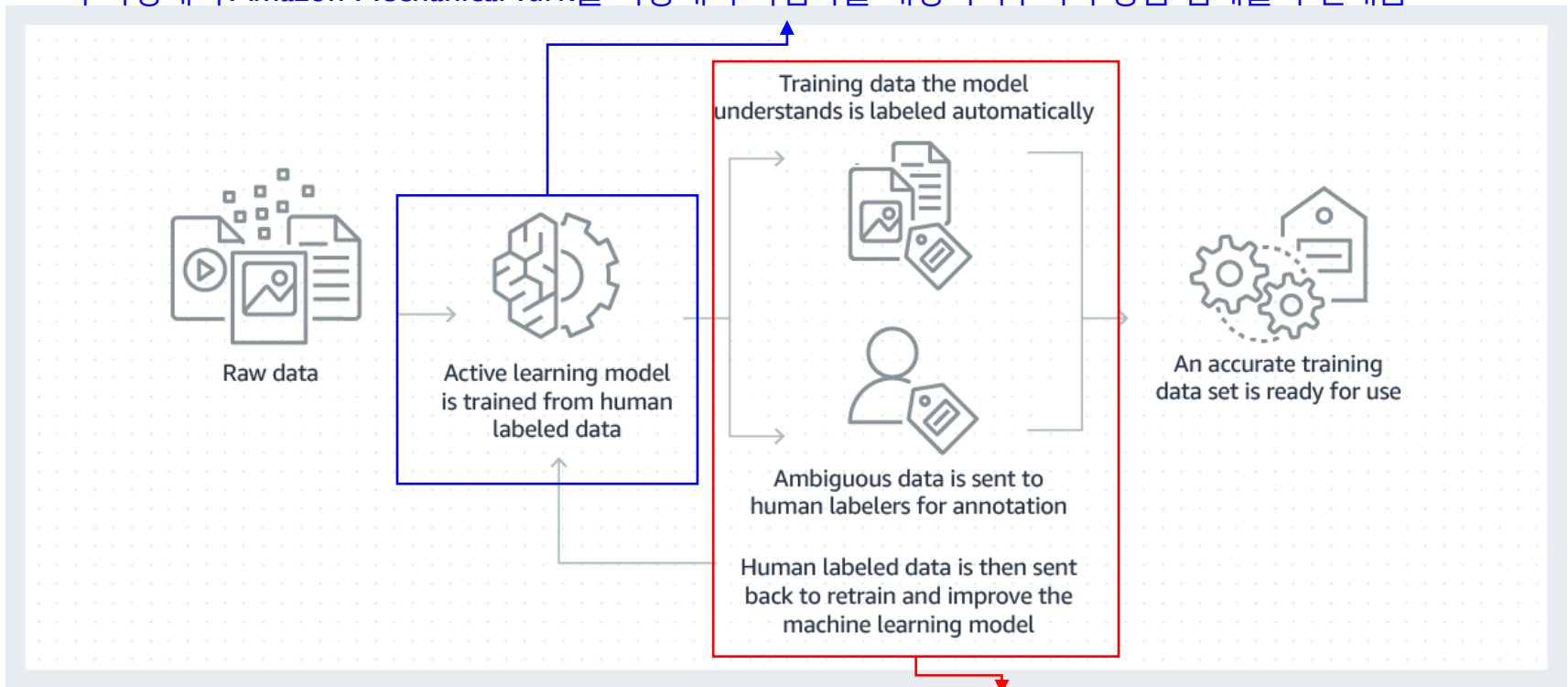
- Data Annotation as a Business Model

- ✓ Amazon SageMaker Ground Truth: <https://aws.amazon.com/ko/sagemaker/groundtruth/>

- Data labeling Platform

처음에는 사람에 의해 labeling 작업 수행

이 과정에서 Amazon Mechanical Turk를 사용해서 작업자를 매칭시켜주거나 공급 업체를 추천해줌



1차 레이블링된 데이터를 이용해서 AI 모델을 학습시킨 후,
모델의 신뢰도가 낮을 경우에 사람에게 확인 요청을 하는 feedback loop를 거침

Data Quality in NLP

• Data Annotation as a Business Model (Social Enterprise)

✓ DataMaker: <https://www.rdproject.kr/#section-service>

✓ 테스트웍스: <http://www.testworks.co.kr/>



플랫폼 특징점 3

이중 전수 검수를 통한 고품질의 데이터 보장

저희는 아프리카 가나 데이터 랩에서 1차 작업 및 전수 검수, 대한민국 데이터 랩에서 전담 검수자들이 2차 전수 검수를 시행합니다. 모든 작업자들은 보상금을 지급받기 위해서 데이터메이커의 엄격한 통과 기준을 준수해야 합니다. 저희는 2가지 방식의 검수 시스템을 보유하고 있으며, 의뢰자님이 검수 옵션을 선택하실 수 있습니다.

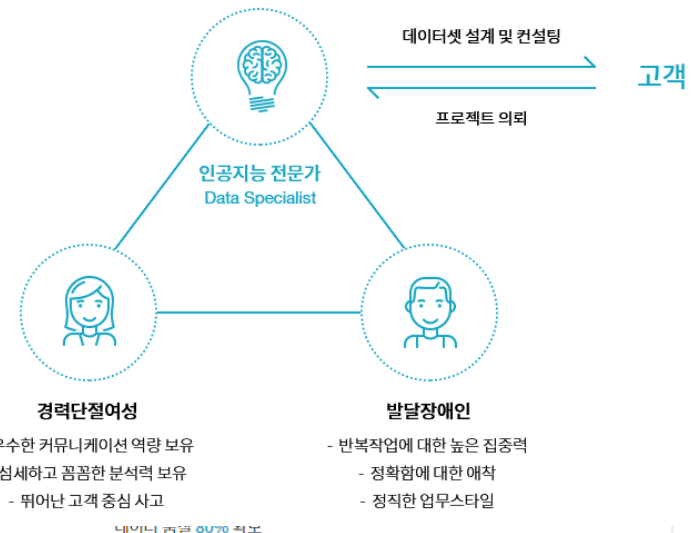


데이터 가공

인공지능 알고리즘 개발을 위해 학습/검증용 데이터셋 구축 서비스를 제공

인공지능 데이터셋 구축 전문 인력

인공지능 전문가의 리드 하에, 데이터 가공 작업에 풍부한 경험과 특장점을 가진 인원들이 프로젝트에 투입됩니다.



A person, likely a woman, is holding a white rectangular sign in front of her face. The sign has the text "ANY questions?" written on it in a black, handwritten-style font. The person is wearing a dark blue blazer over a light blue and white striped shirt. The background is slightly blurred, showing some orange and white elements, possibly a wall or a display.

ANY
questions?