Introduction to NLP

Phạm Quang Nhật Minh

minhpham0902@gmail.com

December 13, 2025

Table of Contents

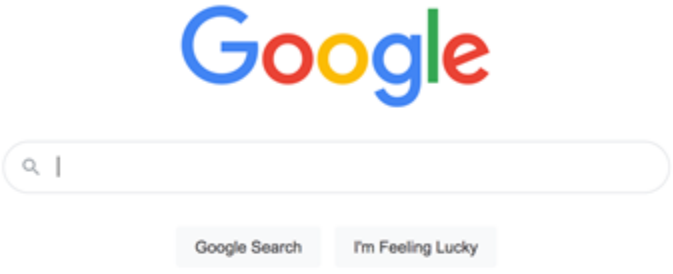
- What is Natural Language Processing
- Why NLP is hard?
- A Brief History of NLP
- NLP Tasks
  - Fundamental Problems in NLP
  - NLP Applications
- How to learn NLP?

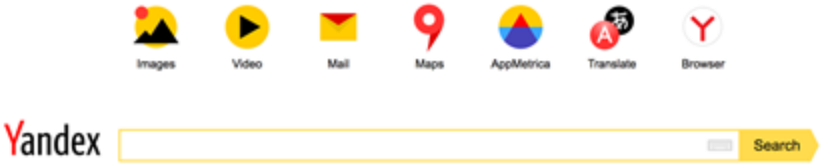
What is Natural Language Processing?

- Technology to handle human language (usually text) using computers

- To get computers to perform useful tasks involving human languages
    - Aid human-machine communication (e.g. question answering, dialog, code generation)
    - Aid human-human communication (e.g. machine translation, spell checking, assisted writing)
    - Analyze/understand language (e.g. syntactic analysis, text classification, entity/relation recognition/linking)

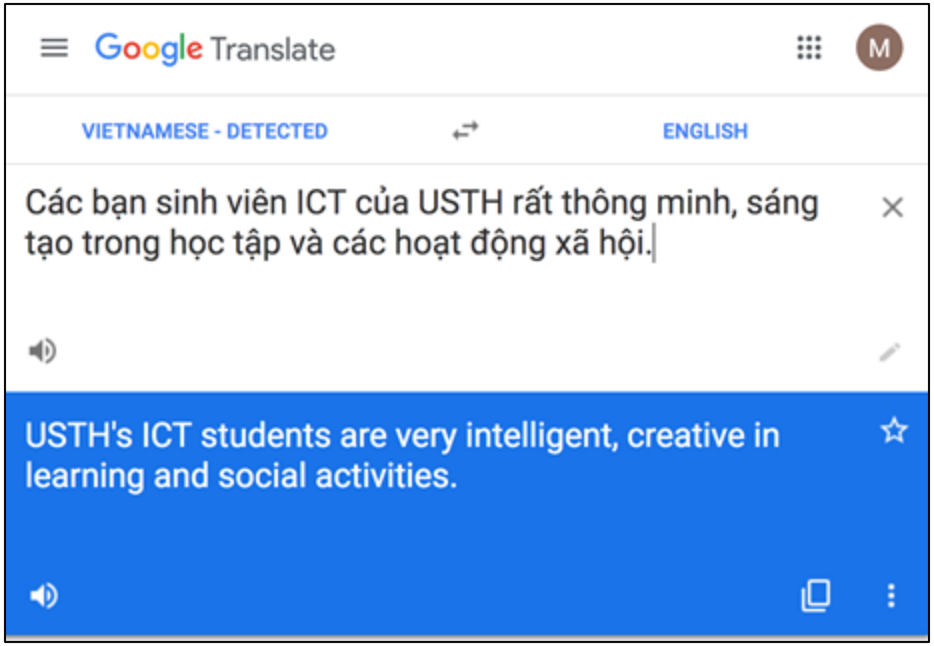
Search Engines





Machine Translation

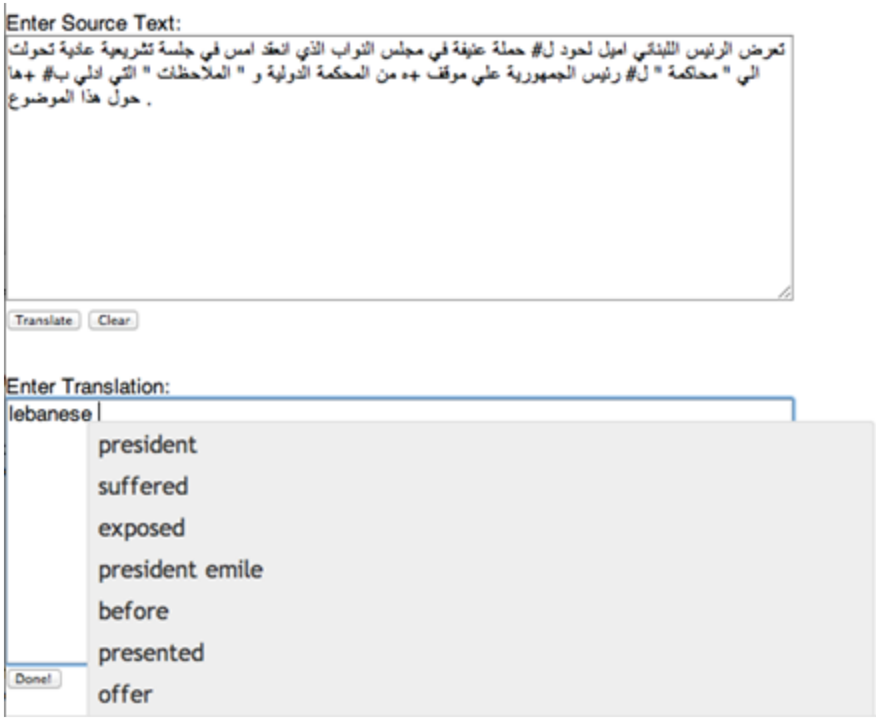
- Fully automatic



Các bạn sinh viên ICT của USTH rất thông minh, sáng tạo trong học tập và các hoạt động xã hội.

USTH's ICT students are very intelligent, creative in learning and social activities.

- Helping human translators



Enter Source Text:

تعرض الرئيس اللبناني اميل لحود ل# حملة عنيفة في مجلس النواب الذي انعقد امس في جلسة تشريعية عادية تحولت الي " محاكمة " ل# رئيس الجمهورية علي موقف +ه من المحكمة الدولية و " الملاحظات " التي ادلي ب# +ها حول هذا الموضوع .

Enter Translation:

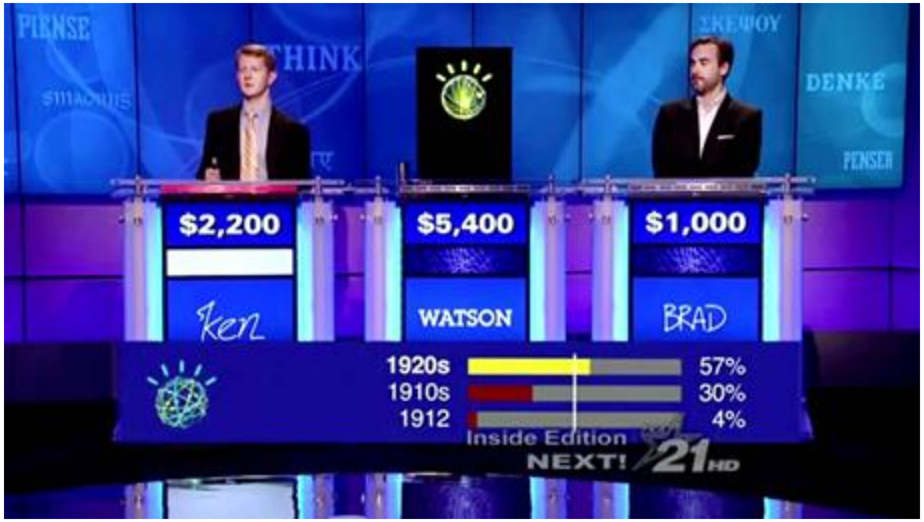
lebanese

Question Answering: IBM’s Watson

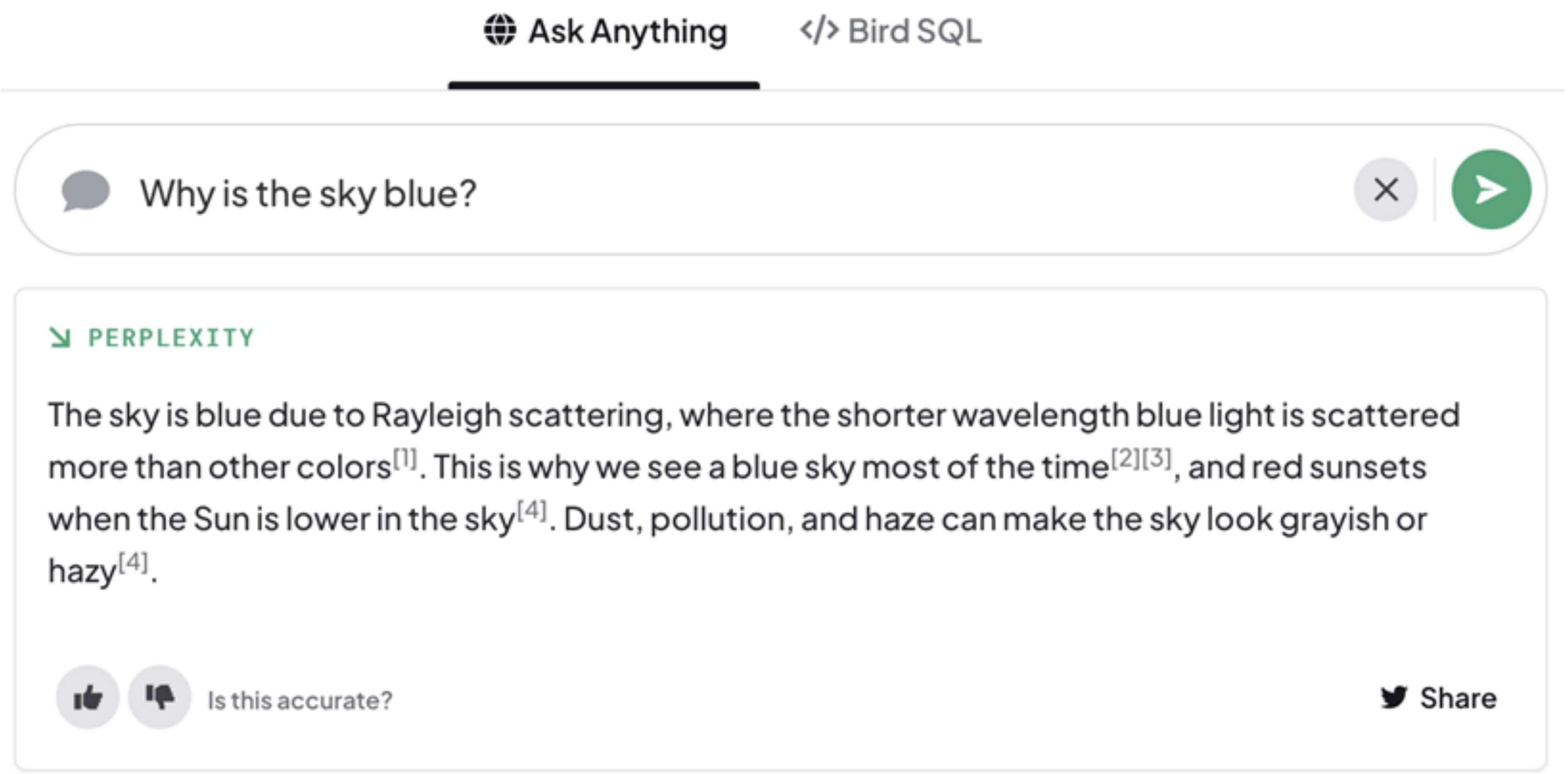
- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON’S
“AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA”
INSPIRED THIS AUTHOR’S
MOST FAMOUS NOVEL

Bram
Stoker



Question Answering: Perplexity.AI



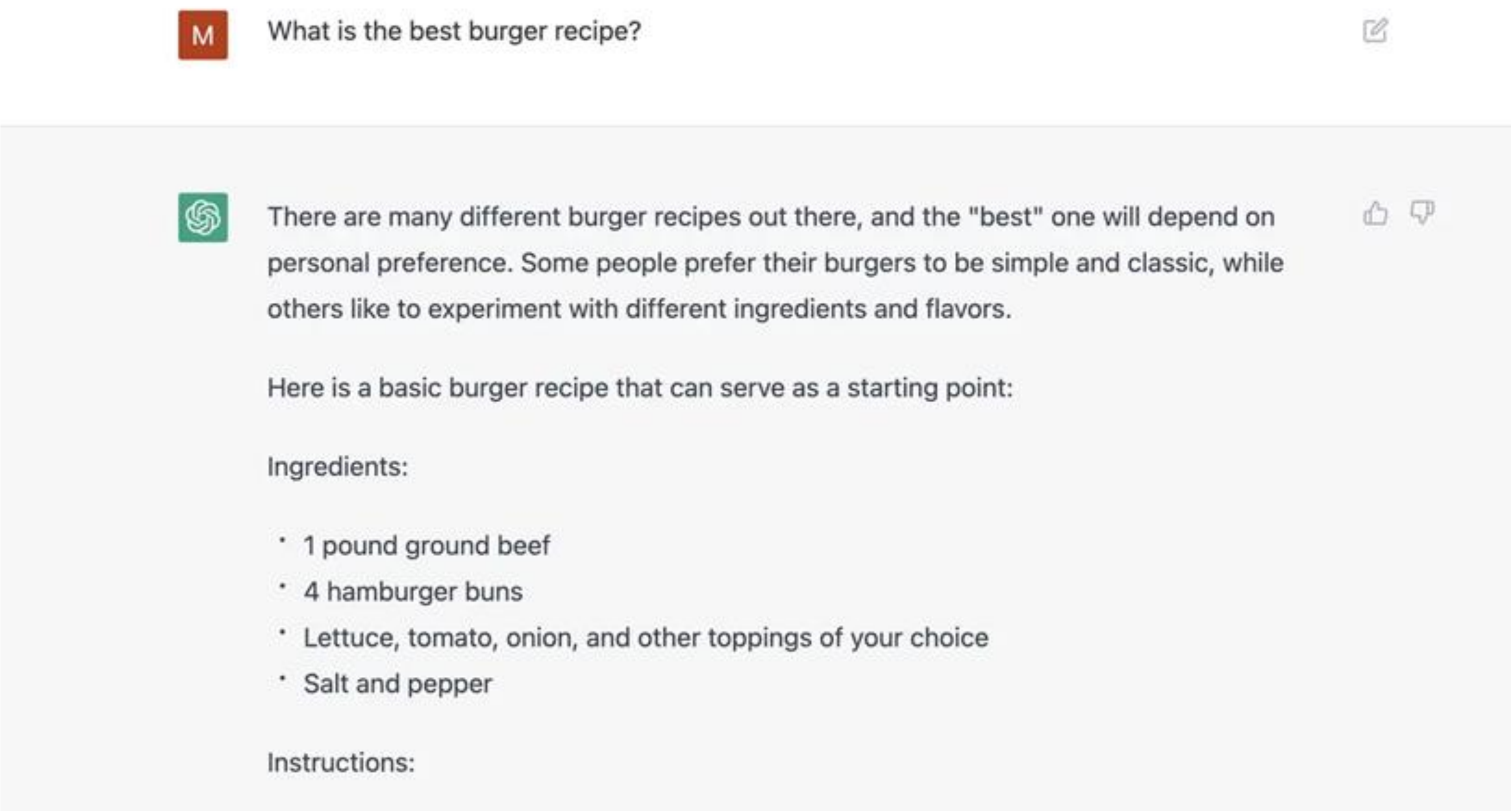
Why is the sky blue?

The sky is blue due to Rayleigh scattering, where the shorter wavelength blue light is scattered more than other colors[1]. This is why we see a blue sky most of the time[2][3], and red sunsets when the Sun is lower in the sky[4]. Dust, pollution, and haze can make the sky look grayish or hazy[4].

https://www.perplexity.ai

Question Answering: ChatGPT

- ChatGPT can answer questions, write code, peoms, suggest ideas for an essay



M

What is the best burger recipe?

There are many different burger recipes out there, and the "best" one will depend on personal preference. Some people prefer their burgers to be simple and classic, while others like to experiment with different ingredients and flavors.

Here is a basic burger recipe that can serve as a starting point:

Ingredients:

- 1 pound ground beef
- 4 hamburger buns
- Lettuce, tomato, onion, and other toppings of your choice
- Salt and pepper

Instructions:

Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we’ve now scheduled the curriculum meeting.
It will be in Gates 159 tomorrow from 10:00-11:30. ▼
-Chris

Create new Calendar entry

Information Extraction & Sentiment Analysis

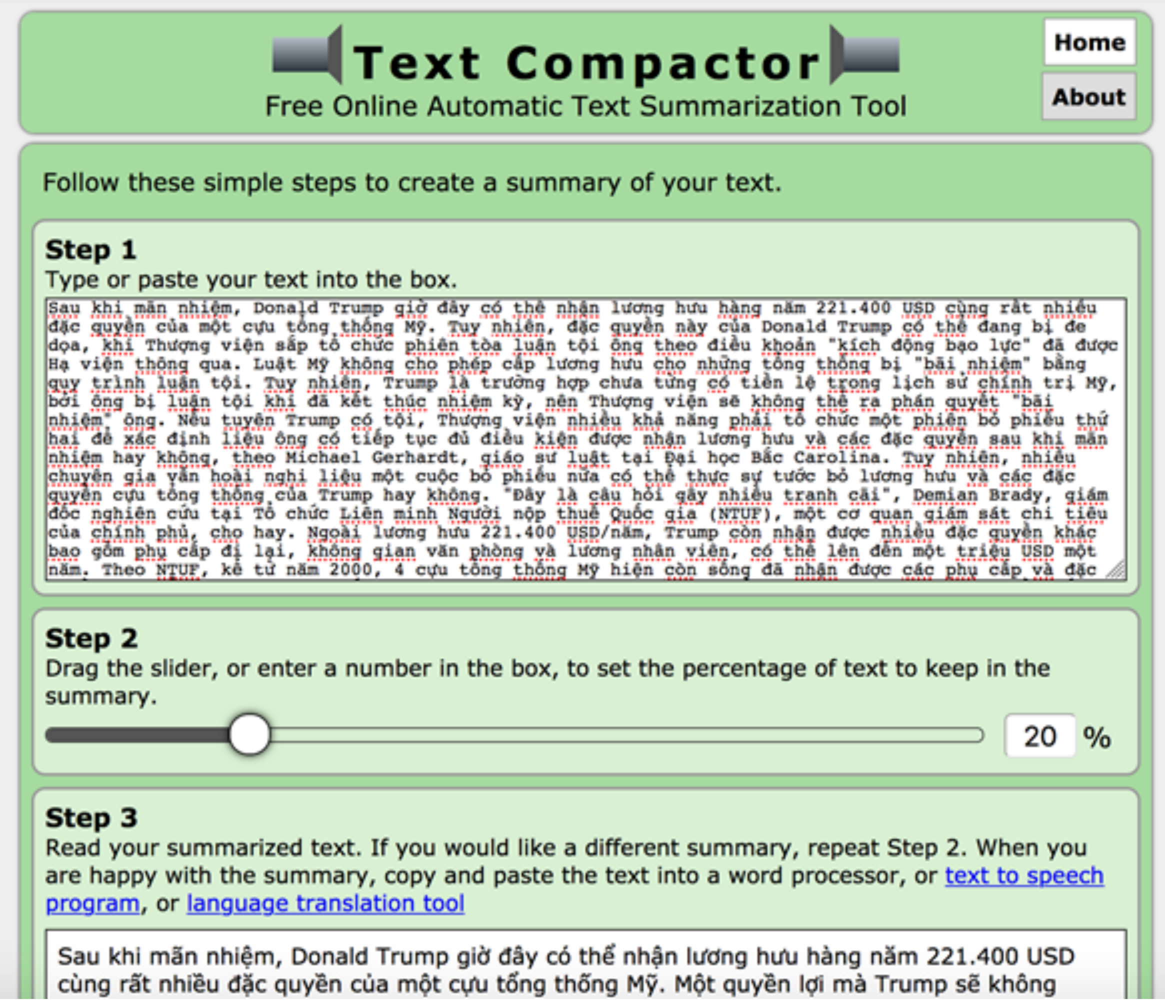
Attributes:
- zoom
- affordability
- size and weight
- flash
- ease of use

Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and ligh carry around those heavy, bulky profes her!
- ✗ the camera feels flimsy, is plastic eight you have to be very delicate in the handling of this camera

Text Summarization

https://www.textcompactor.com



Follow these simple steps to create a summary of your text.

Step 1
Type or paste your text into the box.

Sau khi mãn nhiệm, Donald Trump giờ đây có thể nhận lương hưu hàng năm 221.400 USD cùng rất nhiều đặc quyền của một cựu tổng thống Mỹ. Tuy nhiên, đặc quyền này của Donald Trump có thể đang bị đe dọa, khi Thượng viện sắp tổ chức phiên tòa luận tội ông theo điều khoản "kích động bạo lực" đã được Hạ viện thông qua. Luật Mỹ không cho phép cấp lương hưu cho những tổng thống bị "bãi nhiệm" bằng quy trình luận tội. Tuy nhiên, Trump là trường hợp chưa từng có tiền lệ trong lịch sử chính trị Mỹ, bởi ông bị luận tội khi đã kết thúc nhiệm kỳ, nên Thượng viện sẽ không thể ra phán quyết "bãi nhiệm" ông. Nếu tuyên Trump có tội, Thượng viện nhiều khả năng phải tổ chức một phiên bỏ phiếu thứ hai để xác định liệu ông có tiếp tục đủ điều kiện được nhận lương hưu và các đặc quyền sau khi mãn nhiệm hay không, theo Michael Gerhardt, giáo sư luật tại Đại học Bắc Carolina. Tuy nhiên, nhiều chuyên gia vẫn hoài nghi liệu một cuộc bỏ phiếu nữa có thể thực sự tước bỏ lương hưu và các đặc quyền cựu tổng thống của Trump hay không. "Đây là câu hỏi gây nhiều tranh cãi", Demian Brady, giám đốc nghiên cứu tại Tổ chức Liên minh Người nộp thuế Quốc gia (NTUF), một cơ quan giám sát chi tiêu của chính phủ, cho hay. Ngoài lương hưu 221.400 USD/năm, Trump còn nhận được nhiều đặc quyền khác bao gồm phụ cấp đi lại, không gian văn phòng và lương nhân viên, có thể lên đến một triệu USD một năm. Theo NTUF, kể từ năm 2000, 4 cựu tổng thống Mỹ hiện còn sống đã nhận được các phụ cấp và đặc

Step 2
Drag the slider, or enter a number in the box, to set the percentage of text to keep in the summary.

20 %

Step 3
Read your summarized text. If you would like a different summary, repeat Step 2. When you are happy with the summary, copy and paste the text into a word processor, or text to speech program, or language translation tool

Sau khi mãn nhiệm, Donald Trump giờ đây có thể nhận lương hưu hàng năm 221.400 USD cùng rất nhiều đặc quyền của một cựu tổng thống Mỹ. Một quyền lợi mà Trump sẽ không

Dialogue Systems

Apple Siri (2011)

Google Now (2012)
Google Assistant (2016)

Microsoft Cortana (2014)

Amazon
Alexa/Echo (2014)

Google Home (2016)

Apple HomePod (2017)

Why NLP?

- Languages involve many human activities
- Voice-based user interfaces
    - Remote controls, virtual assistants
- Mining big textual data
    - E.g., Biomedical texts

Ambiguity makes NLP hard!

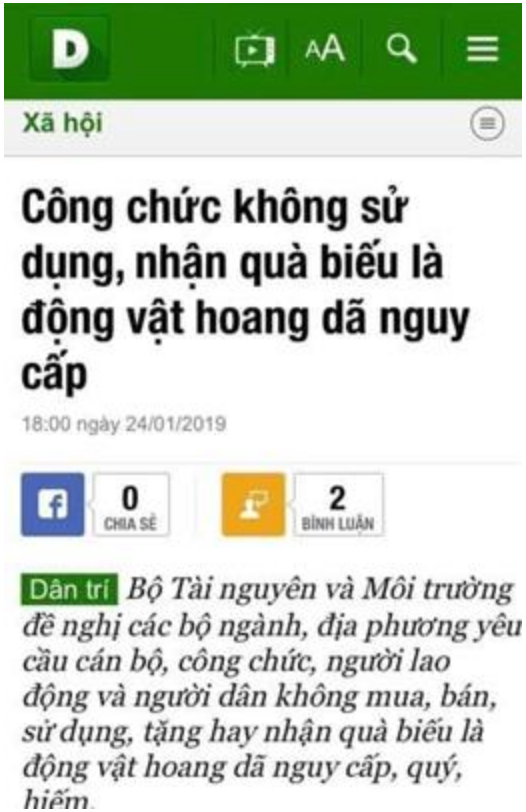
- Five different meanings of “I made her duck”
    1. I cooked waterfowl for her
    2. I cooked waterfowl belong to her
    3. I created the (plastic) duck she owns
    4. I caused her to quickly lower her head or body
    5. I waved my magic wand and turned her into undifferentiated waterfowl
- NLP is to resolve or disambiguate ambiguities

NLP is highly ambiguous (1)

- Word-level ambiguity
  - “duck” can be a noun or a verb (ambiguous POS)
  - “make” can mean “create” or “cook” (ambiguous sense)
- Syntax-level ambiguity
  - ”her” can be a direct object or indirect object of the verb “make”

NLP is highly ambiguous (2)

- Syntactic ambiguity
  - Natural language processing
  - I shot an elephant in my pajasma.



Công chức không sử dụng, nhận quà biếu là động vật hoang dã nguy cấp

18:00 ngày 24/01/2019

Dân trí *Bộ Tài nguyên và Môi trường đề nghị các bộ ngành, địa phương yêu cầu cán bộ, công chức, người lao động và người dân không mua, bán, sử dụng, tặng hay nhận quà biếu là động vật hoang dã nguy cấp, quý, hiếm.*

It is 100% real

NLP is highly ambiguous (3)

- Anaphora resolution
  - “John persuaded Bill to buy a TV for himself.” (himself = John or Bill?)
- Natural languages involve reasoning about the world
  - E.g., It is unlikely that an elephant wears a pajama

Why else is NLP difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

But that's what makes it fun!

Machine Translation Needs

- NLP emerged from the need of Machine Translation in the 1940s.
  - Russian – English language pair
- Lousy era during 1966 after a report of ALPAC
  - "we do not have useful machine translation and there is no immediate or predictable prospect of useful machine translation”
  - MT/NLP almost died

Better condition from 1980s

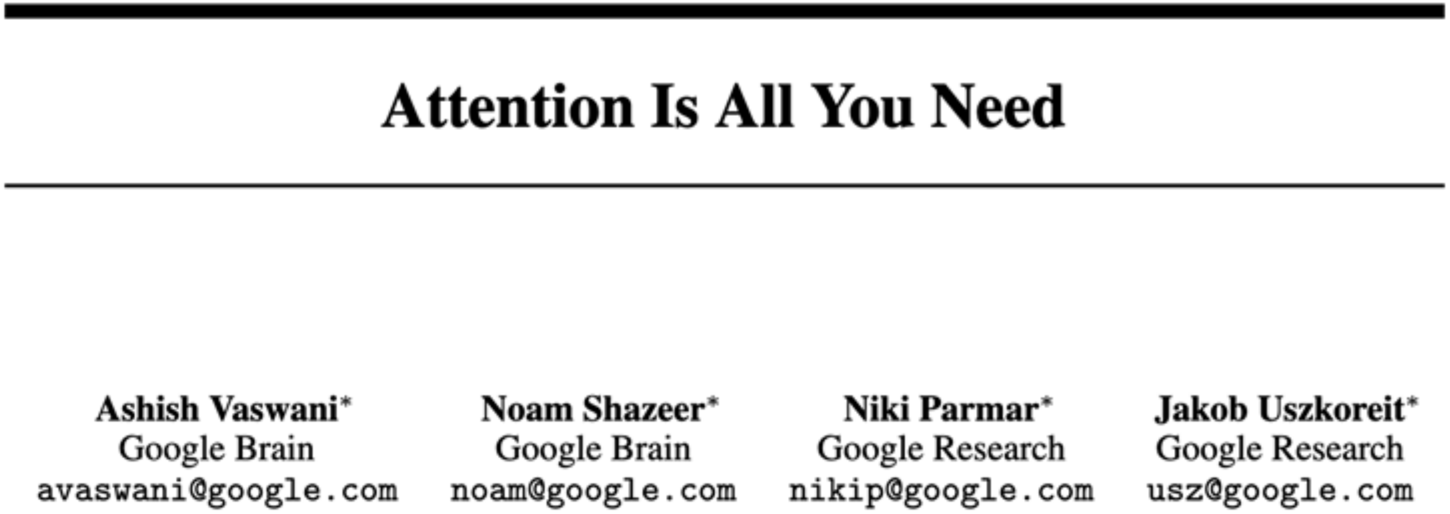
- MT/NLP products started providing some results
  - LUNAR (QA system) developed in 1978 by W.A woods
- Statistical Machine Translation (SMT) by IBM in late 1980s and early 1990s

The Rise of Machine Learning 2000 - 2007

- Large amount of spoken and written materials become widely available
  - More annotated NLP corpora
- Development of statistical machine learning models
  - Support vector machines (Vapnik, 1995)
  - Multinomial logistic regression (MaxEnt) (Berger et al., 1996)
  - Bayesian models (Pearl, 1988)

The Rise of Large Language Models (2018 ~)

- Transformers
- BERT
- GPT Models
- Open-source large language models
- Multi-modal foundation models



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

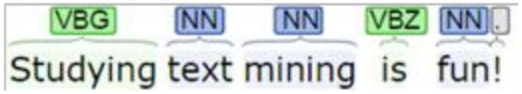
Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

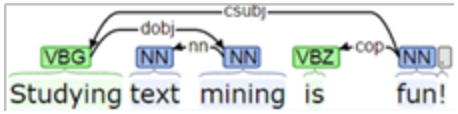
Jakob Uszkoreit*
Google Research
usz@google.com

Fundamental Problems in NLP

- Tokenization
  - ☐ "Studying text mining is fun" → "studying" + "text" + "mining" + "is" + "fun"
- Part-of-Speech tagging



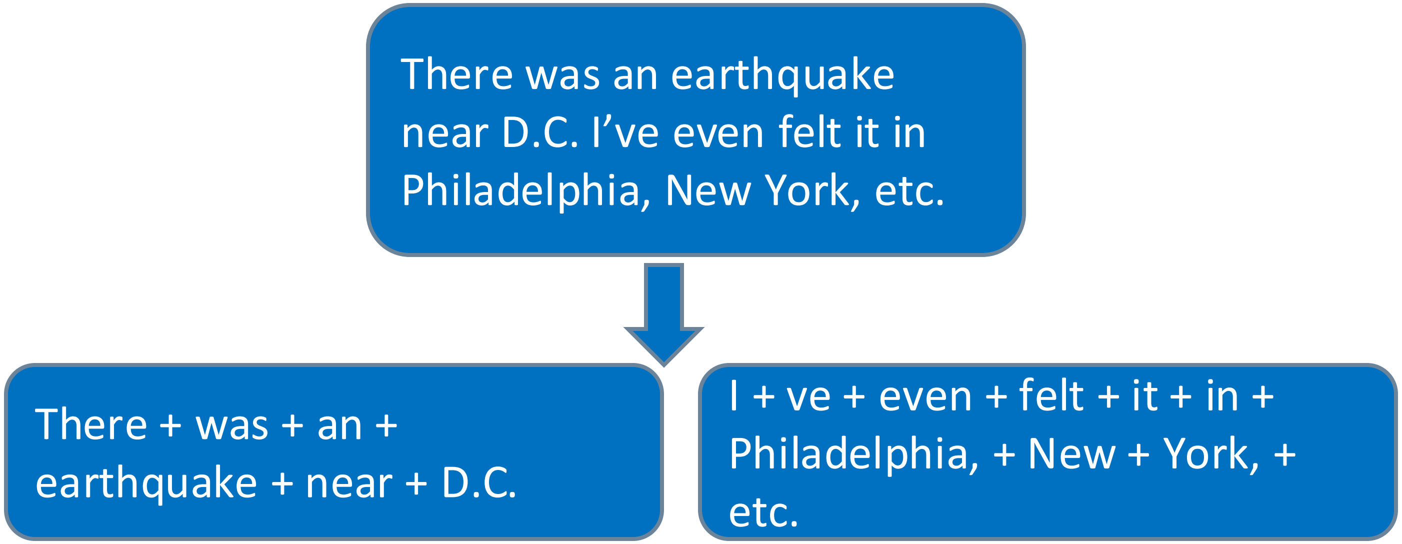
- Chunking
- Named entity recognition
- Syntactic parsing



- Semantic analysis

Tokenization

- Split text into words and sentences



There was an earthquake near D.C. I’ve even felt it in Philadelphia, New York, etc.

There + was + an + earthquake + near + D.C.

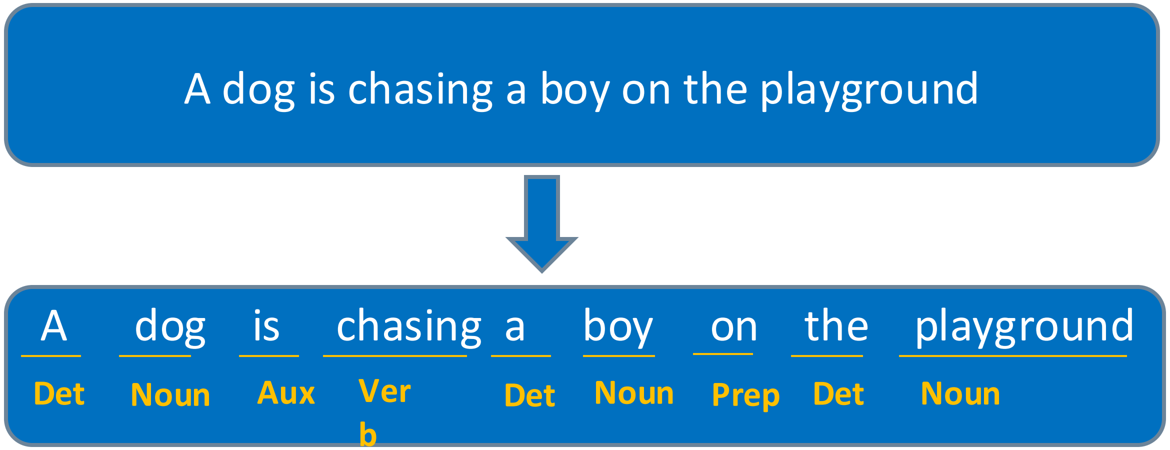
I + ve + even + felt + it + in + Philadelphia, + New + York, + etc.

Word Segmentation

- Sentences in Japanese or Chinese are written without space
  - Word segmentation adds spaces between words
    - 単語文割を行う → 単語 文割 を 行 う
- Vietnamese, a compound word may contain several syllables (smallest units in Vietnamese). There are only spaces between syllables.
  - E.g., Nhật Bản luôn là thị trường thương mại quan trọng của Việt Nam
  - Word segmentation determines contiguous syllables that make a word
    - Nhật_Bản luôn là thị_trường thương_mại quan_trọng của Việt_Nam

Part-of-speech tagging

- Marking up a word in a text (corpus) as corresponding to a part of speech



A dog is chasing a boy on the playground

A	dog	is	chasing	a	boy	on	the	playground
Det	**Noun**	**Aux**	**Verb**	**Det**	**Noun**	**Prep**	**Det**	**Noun**

Named-entity recognition

- Determine text mapping to proper names

Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

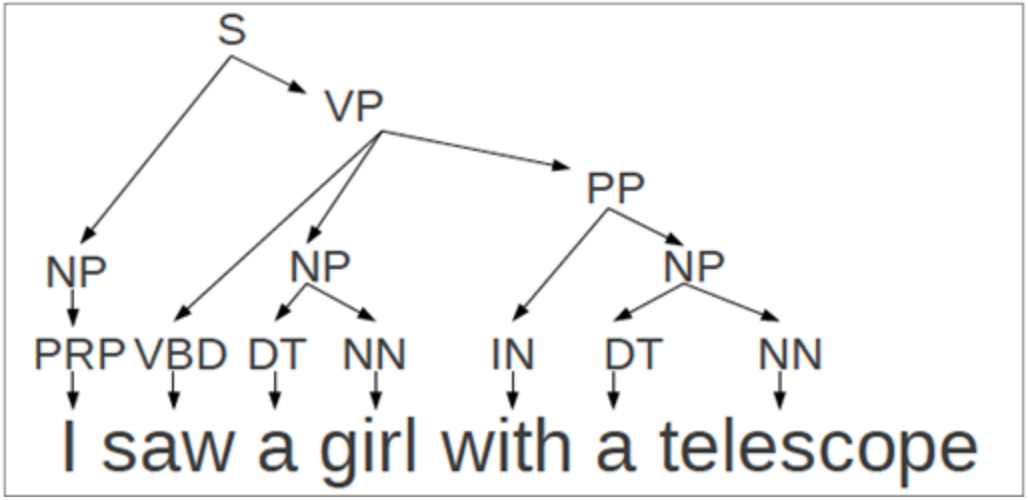
Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

Organization, Location, Person

Syntactic parsing

- Perform grammatical analysis for a given sentence and assign a syntactic structure to it
- An important task in NLP with many applications
  - ☐ Intermediate state of representation for semantic analysis

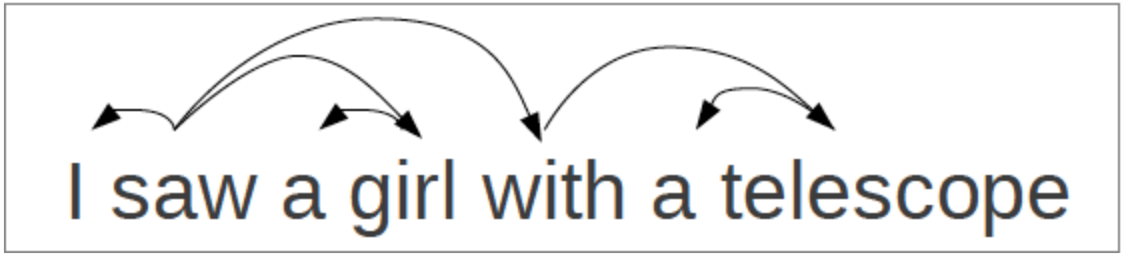
I saw a girl with a telescope



Dependency parsing

- Assign a dependency structure to a given sentence
  - Focuses on relations between words

I saw a girl with a telescope



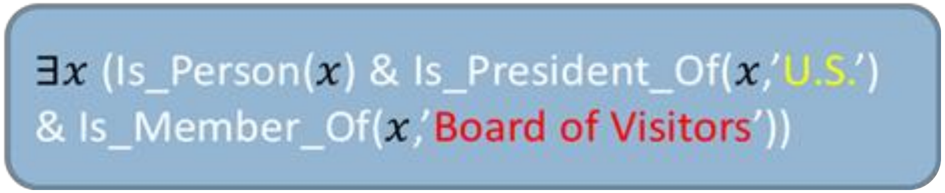
Semantic Analysis

- Syntax parsing trees gives no information about semantics
- Semantic considers:
  - Meaning Representation
  - Translation from syntax into the meaning representation
  - Word meaning disambiguation
  - Relations between words

Meaning Representations

- Convert chunks of text into more formal representations
  - Deep semantic analysis: e.g., first-order logic structures

Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

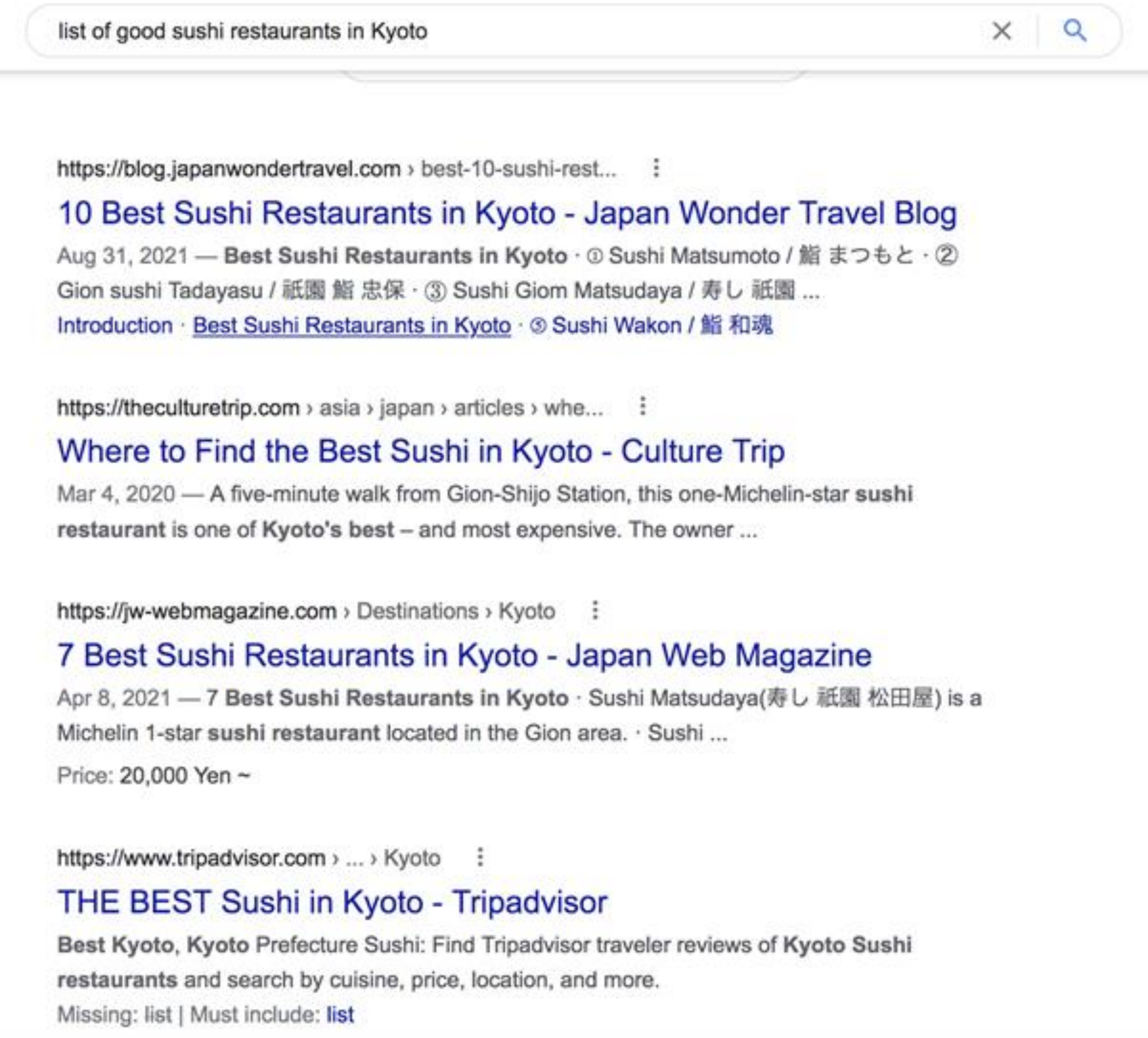


$\exists x$ (Is_Person(x) & Is_President_Of(x,'U.S.') & Is_Member_Of(x,'Board of Visitors'))

Application Tasks

- Information Retrieval
- Information Extraction
- Question Answering
- Text Summarization
- Machine Translation
- Chatbot & Dialogue Systems

Information Retrieval



list of good sushi restaurants in Kyoto

https://blog.japanwondertravel.com › best-10-sushi-rest...

10 Best Sushi Restaurants in Kyoto - Japan Wonder Travel Blog

Aug 31, 2021 — **Best Sushi Restaurants in Kyoto** · ① Sushi Matsumoto / 鮨 まつもと · ② Gion sushi Tadayasu / 祇園 鮨 忠保 · ③ Sushi Giom Matsudaya / 寿し 祇園 ...

Introduction · Best Sushi Restaurants in Kyoto · ⑤ Sushi Wakon / 鮨 和魂

https://theculturetrip.com › asia › japan › articles › whe...

Where to Find the Best Sushi in Kyoto - Culture Trip

Mar 4, 2020 — A five-minute walk from Gion-Shijo Station, this one-Michelin-star **sushi restaurant** is one of **Kyoto's best** – and most expensive. The owner ...

https://jw-webmagazine.com › Destinations › Kyoto

7 Best Sushi Restaurants in Kyoto - Japan Web Magazine

Apr 8, 2021 — **7 Best Sushi Restaurants in Kyoto** · Sushi Matsudaya(寿し 祇園 松田屋) is a Michelin 1-star **sushi restaurant** located in the Gion area. · Sushi ...

Price: 20,000 Yen ~

https://www.tripadvisor.com › ... › Kyoto

THE BEST Sushi in Kyoto - Tripadvisor

Best Kyoto, Kyoto Prefecture Sushi: Find Tripadvisor traveler reviews of **Kyoto Sushi restaurants** and search by cuisine, price, location, and more.

Missing: list | Must include: list

Question Answering

- A system that automatically return answers for an input question by retrieving information from a collected documents
- Differences from IR
  - QA system's goal is to respond exact answers instead of documents related to the question
  - QA system requires more complicated semantic analysis

Question Answering

- Factoid question answering
  - Who/What/Where/When
  - Answers are often short phrases
- Non-factoid question answering
  - Definition questions
  - How/Why
  - Answers may span multiple sentences (paragraph)

Text Summarization

- Process of distilling the most important information from a text to produce an abridged version of a particular task or user

- Useful in the era of information explosion

- Summarization types
  - Single-document/Multi-document summarization
  - Extractive/Abstractive summarization

Chatbot & Dialogue Systems

- NLP systems that can communicate with humans in natural languages
    - ChatGPT, Siri, Google Assistant
- Significantly advanced recently

How to learn NLP (1)

- Have background/knowledge about
  - Probabilistics and Statistics
  - Basic math (linear algebra, calculus)
  - Machine Learning
  - Programming
- Learn from textbooks or courses

How to learn NLP (2)

- Learning by doing!
  - Build up somethings: customize open-source codes, re-implement some models, etc
- Compete in Kaggle data science challenges
  - https://www.kaggle.com/search?q=NLP
- Read papers on ACL Anthology (for ones who want to do research on NLP)

NLP Engineer job

- Solid NLP/ML background
- Software development skills: backend, database, front-end
- English or Japanese