



# Introduction to NLP

**Phạm Quang Nhật Minh**

Aimesoft JSC

[minhpham0902@gmail.com](mailto:minhpham0902@gmail.com)

January 06, 2024



# Table of Contents

2

- 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?

3

- A subfield of computer science, artificial intelligence, and computational linguistics
- To get computers to perform useful tasks involving human languages
  - Human-Machine communication
  - Improving human-human communication
  - Extracting information from texts



# Search Engines

4

# Google

Google Search

I'm Feeling Lucky



Images



Video



Mail



Maps



AppMetrica



Translate



Browser

# Yandex

 Search

# DuckDuckGo

Search the web without being tracked



# Baidu 百度

 百度一下

# NAVER

5인 이상 모임은 조금만 미뤄요

# COCOCOC



# Machine Translation

5

## ■ Fully automatic

Google Translate

VIETNAMESE - DETECTED ↔ ENGLISH

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 لـ حملة عنيفة في مجلس النواب الذي انعقد امس في جلسة تشريعية عادية تحولت الي " محاكمة " لـ رئيس الجمهورية علي موقف +ه من المحكمة الدولية و " الملاحظات " التي ادلي بـ# +ها حول هذا الموضوع .

Translate Clear

Enter Translation:

lebanese |

- president
- suffered
- exposed
- president emile
- before
- presented
- offer

Done!



# Question Answering: IBM's Watson

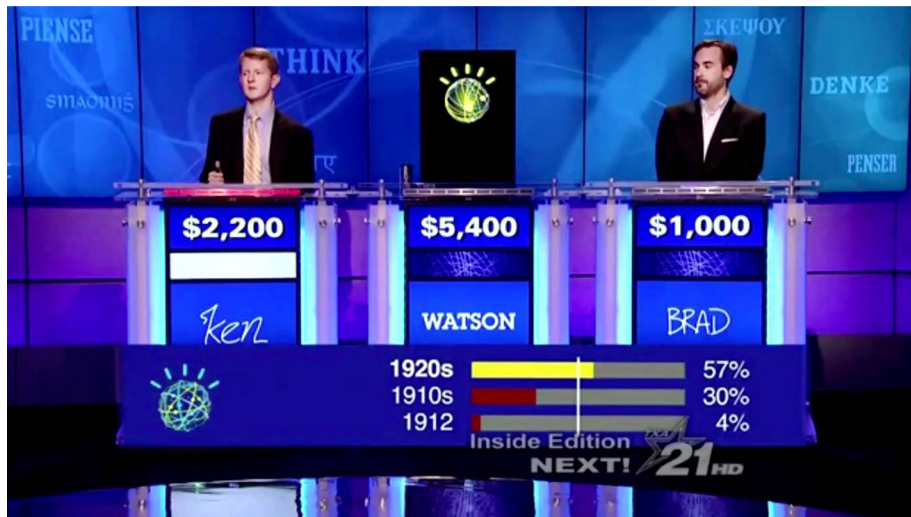
6

- 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

7

 Ask Anything

 Bird SQL

 Why is the sky blue?




 PERPLEXITY

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



Is this accurate?

 Share

<https://www.perplexity.ai>



# Question Answering: ChatGPT

8

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



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

9

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jurafsky

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 light around those heavy, bulky profes
- ✗ ■ the camera feels flimsy, is plastic you have to be very delicate in the handling of this camera



carry her!  
weight



# Text Summarization

11

<https://www.textcompactor.com>

## Text Compactor

Free Online Automatic Text Summarization Tool

Home

About

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 nửa 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

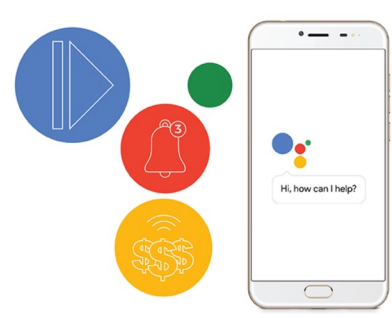


# Dialogue Systems

12



Apple Siri (2011)



Google Now (2012)  
Google Assistant (2016)



Microsoft Cortana (2014)



Amazon  
Alexa/Echo (2014)



Google Home (2016)



Apple HomePod (2017)



# Why NLP?

13

- 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!

14

- 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)

15

- 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”



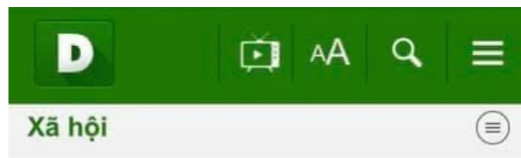
# NLP is highly ambiguous (2)

16

## ■ Syntactic ambiguity

Natural language processing

I shot an elephant in my pajama.

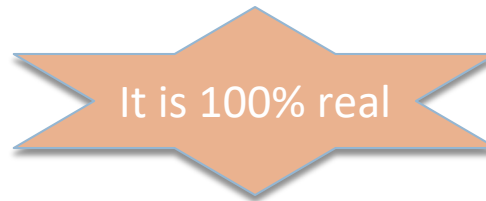


**Công chức không sử dụng, nhận quà biếu là động vật hoang dã nguy cấp**

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.







# NLP is highly ambiguous (3)

17

- 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?

18

## 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

20

- 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

21

- 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 ~)

22

- 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



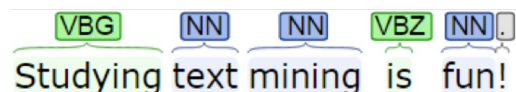
# Fundamental Problems in NLP

23

## ■ Tokenization

- “Studying text mining is fun” → “studying” + “text” + “mining” + “is” + “fun”

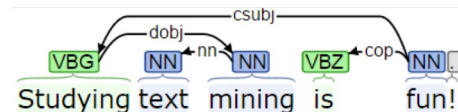
## ■ Part-of-Speech tagging



## ■ Chunking

## ■ Named entity recognition

## ■ Syntactic parsing



## ■ Semantic analysis



# Tokenization

24

- 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

25

- 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

26

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

A dog is chasing a boy on the playground



<u>A</u>	<u>dog</u>	<u>is</u>	<u>chasing</u>	<u>a</u>	<u>boy</u>	<u>on</u>	<u>the</u>	<u>playground</u>
Det	Noun	Aux	Ver	Det	Noun	Prep	Det	Noun
			b					



# Named-entity recognition

27

- 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**

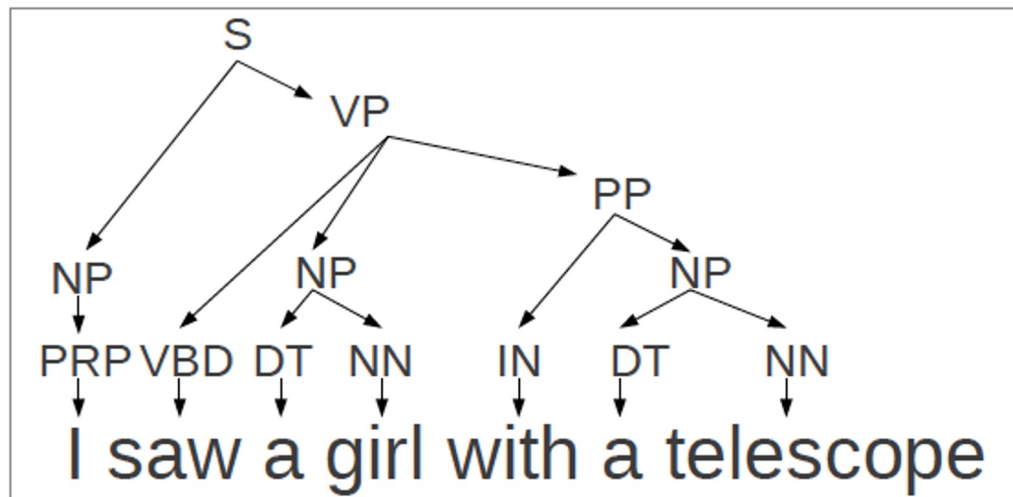


# Syntactic parsing

28

- 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



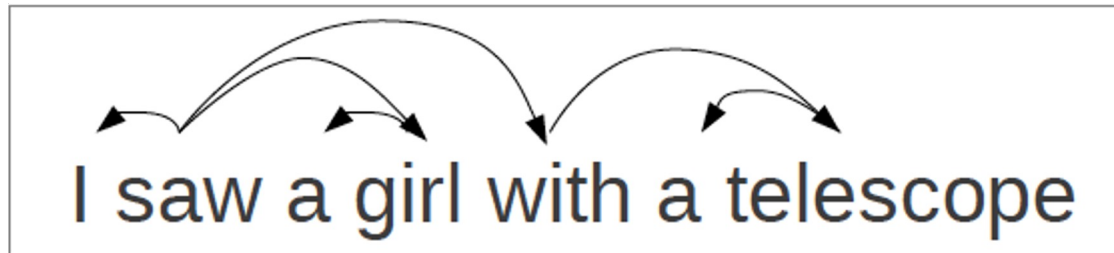


# Dependency parsing

29

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

I saw a girl with a telescope





# Semantic Analysis

30

- 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

31

- 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

32

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





# Information Retrieval

33

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

34

- 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

35

- 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

36

- 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

37

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



# How to learn NLP (1)

38

- 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)

39

- 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 Usecase: Law Update System

40

- Customer is a large law book publishing company
- Law books need to be updated when cited laws changed





# Development Content

41

Information about changes/additions in the related terms and conditions.

```
<SNLLQuestion>
所得税に関し、審査請求の対象となる処分は具体的にどのようなものがありますか。
</SNLLQuestion>
<answer>
  税務署長等が行う次の処分等です。
</answer>
<explanation>
(1) 税務署長が行う処分
  @ 更正・決定（通則法24条～26条）
  A 加算税の賦課決定（通則法65条～68条）
  B 更正の請求に対するその請求の一部を認めた更正又はその更正をすべき理由がない旨の通知（通則法23条）
  C 特別農業所得者の申請に対する却下（所法110条）
  D 予定納税額の減額承認申請に対する一部承認又は却下（所法113条）
  E 純損失の繰戻しによる還付請求の一部を認めた所得税の還付又はその還付をすべき理由がない旨の通知（所法142条）
  F 青色申告の承認申請の却下（所法145条）
  G 青色申告の承認の取消し（所法150条）
  H 棚卸資産・有価証券の評価方法の変更申請の却下（所令101条・107条）
  I 減価償却資産の償却方法の変更申請の却下（所令124条）
  J 所得税の申告等の期限延長申請の却下（通則令3条）
(2) 国税局長が行う処分
  納税地の指定（国税庁長官が行うものを除きます。）（所法18条）
```



Law Update System



```
<SNLLQuestion>
所得税に関し、審査請求の対象となる処分は具体的にどのようなものがありますか。
</SNLLQuestion>
<answer>
  税務署長又は税関長等が行う次の処分等です。
</answer>
<explanation>
(1) 税務署長又は税関長が行う処分
  @ 更正・決定（通則法24条～26条）
  A 加算税の賦課決定（通則法65条～68条）
  B 更正の請求に対するその請求の一部を認めた更正又はその更正をすべき理由がない旨の通知（通則法23条）
  C 特別農業所得者の申請に対する却下（所法110条）
  D 予定納税額の減額承認申請に対する一部承認又は却下（所法113条）
  E 純損失の繰戻しによる還付請求の一部を認めた所得税の還付又はその還付をすべき理由がない旨の通知（所法142条）
  F 青色申告の承認申請の却下（所法145条）
  G 青色申告の承認の取消し（所法150条）
  H 棚卸資産・有価証券の評価方法の変更申請の却下（所令101条・107条）
  I 減価償却資産の償却方法の変更申請の却下（所令124条）
  J 所得税の申告等の期限延長申請の却下（通則令3条）
(2) 国税局長が行う処分
  納税地の指定（国税庁長官が行うものを除きます。）（所法18条）
```

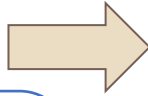


# Our Solution

42

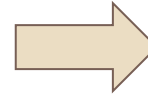
## Text Alignment

- Align the paragraphs in the response/explanation section with those in the related terms and conditions.



## Update Detection

- Identify the sections that need updating when there are changes in the related terms and conditions.
- Determine the type of update to be carried out.



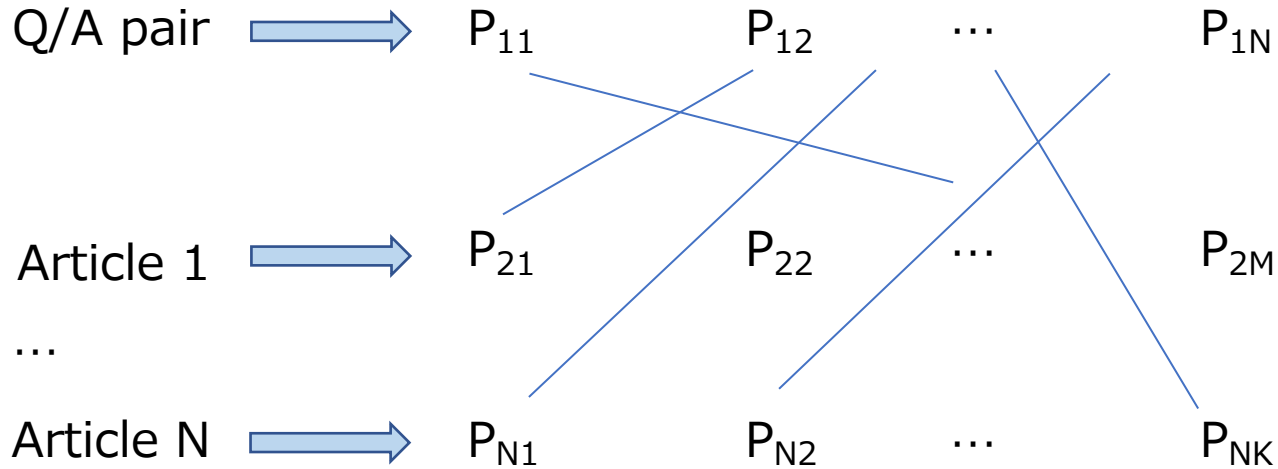
## Content Updating

- Identify the types of updates that can be automated.
- Construct rules for rewriting sentences.



# Text Alignment

43



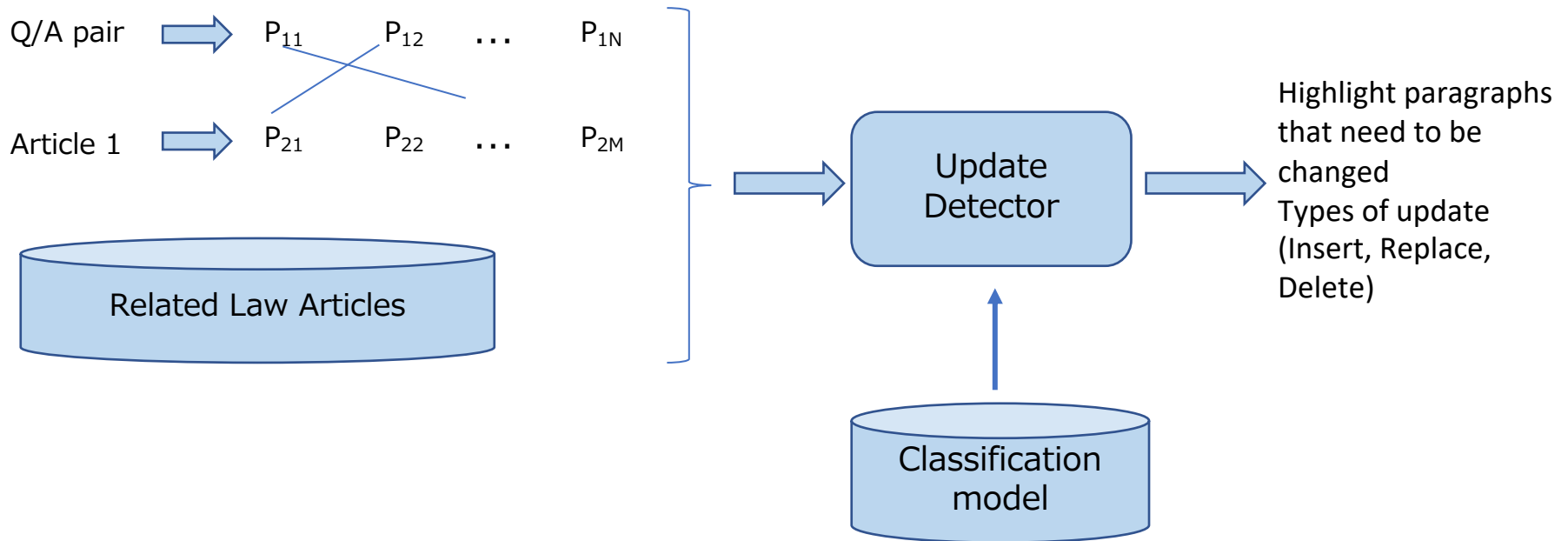
## Algorithm

- Text Similarity
- Construct a binary classifier to determine if a pair (paragraph 1, paragraph 2) is related or not.



# Update Detection

44



- From the results of Text Alignment and data about changes in the related terms and conditions, identify the sections that need updating and the type of update required.
- The determination of the update types is done by combining machine learning models and rules.



# Law Update System

45

前の設問

作業 619,586\_環境手引 - 追録番号 10, ページ 191

## 解析結果

191

次の設問

### 設問

### 改正法令情報

元に戻す

① 前年度における事業者全体の「エネルギー使用状況届出書」を提出し、その結果を5月末日までに、本社の所在地を管轄する経済産業局に「エネルギー使用状況届出書」を提出します。なお、個別の工場や事業場など事業所単位で1,500Kl/年以上のエネルギー使用量（原油換算量）があった場合には、当該工場・事業場のエネルギー使用量を事業者全体のエネルギー使用量の内訳として「エネルギー使用状況届出書」に記載します（省エネ法7③・\*1819①）。

② 把握したエネルギー使用量の合計が1,500Kl/年以上であった場合には、その結果を5月末日までに、本社の所在地を管轄する経済産業局に「エネルギー使用状況届出書」を提出します。なお、個別の工場や事業場など事業所単位で1,500Kl/年以上のエネルギー使用量（原油換算量）があった場合には、当該工場・事業場のエネルギー使用量を事業者全体のエネルギー使用量の内訳として「エネルギー使用状況届出書」に記載します（省エネ法7③・\*1819①）。

③ \*「エネルギー使用状況届出書」を届け出ると、国がその事業者を「特定事業者」又は「特定連鎖化事業者」と指定します（省エネ法7①）。また、3,000Kl/年以上のエネルギーを使用している工場・事業場は「第一種エネルギー管理指定工場等」、1,500Kl/年以上3,000Kl/年未満のエネルギーを使用している工場・事業場は「第二種エネルギー管理指定工場等」として指定されます（省エネ法10・13①）。

**\*4 エネルギーの使用の合理化等に関する基本方針**

経済産業大臣は、工場等、輸送、建築物、機械器具等に係るエネルギーの使用の合理化及び\*電気の需要の平準化非化石エネルギーへの転換並びに電気の需要の最適化を総合的に進める見地から、エネルギーの使用の\*合理化等合理化及び非化石エネルギーへの転換等に関する基本方針（基本方針）を定め、これを公表します（省エネ法3①）。

平成25年12月27日に公表された基本方針（平25・12・27経産告268）は、①エネルギーの使用の合理化のためにエネルギーを使用する者等が講ずべき措置に関する基本的な事項、②電気の需要の平準化を図るために電気を使用する者等が講ずべき措置に関する基本的な事項、③エネルギーの使用の合理化等の促進のための施策に関する基本的な事項から構成されています。このうち、工場等（当該者が連鎖化事業者である場合にあっては当該者が行う連鎖化事業者の加盟者が設置している当該連鎖化事業に係る工場等を含み、当該者が認定管理統括事業者である場合にあってはその管理関係事業者が設置している工場等（当該管理関係事業者が連鎖化事業者である場合にあっては、当該者が行う連鎖化事業者の加盟者が設置している当該連鎖化事業に係る工場等を含みます。）を含みます。）においてエネルギーを使用して事業を行う者について、次に示す措置を講じてエネルギー消費原単位又は電気需要平準化評価原単位の改善を図ることとされています。

① 工場等に係るエネルギーの使用の定率、エネルギーの使用の合理化に関する取組等を把握すること

確認済み  修正なし

プレビュー

前の設問

191

次の設問

法令	DL	条文	未使用	未施行
エネルギーの使用の合理化等に関...	<input type="checkbox"/>	第1条	<input type="checkbox"/>	<input type="checkbox"/>
エネルギーの使用の合理化等に関...	<input type="checkbox"/>	第2条	<input type="checkbox"/>	<input type="checkbox"/>
エネルギーの使用の合理化等に関...	<input type="checkbox"/>	第3条	<input type="checkbox"/>	<input type="checkbox"/>
エネルギーの使用の合理化等に関...	<input type="checkbox"/>	第4条	<input type="checkbox"/>	<input type="checkbox"/>
エネルギーの使用の合理化等に関...	<input type="checkbox"/>	第7条	<input type="checkbox"/>	<input type="checkbox"/>
エネルギーの使用の合理化等に関...	<input type="checkbox"/>	第18条	<input type="checkbox"/>	<input type="checkbox"/>
エネルギーの使用の合理化等に関...	<input type="checkbox"/>	第18条	<input checked="" type="checkbox"/>	<input type="checkbox"/>
エネルギーの使用の合理化等に関...	<input type="checkbox"/>	第129条	<input type="checkbox"/>	<input type="checkbox"/>



# NLP Engineer job

46

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