

Xử lý ngôn ngữ tự nhiên Natural Language Processing (NLP)

Giảng viên: Lê Thanh Hương
Trường Công nghệ Thông tin và Truyền thông, ĐHBKHN

Nội dung khóa học

- Khóa học được chia thành 19 buổi, bao gồm:
 - 15 bài giảng x 3 giờ (02 giờ lý thuyết + 01 giờ thực hành)
 - 01 buổi kiểm tra giữa kỳ (02 giờ kiểm tra + 01 giờ chữa bài)
 - 01 buổi kiểm tra cuối kỳ (02 giờ kiểm tra + 01 giờ chữa bài)
 - 02 buổi thuyết trình dự án (04 giờ/buổi)
- Nằm trong chương trình đào tạo kỹ sư AI Vingroup, với các môn học:
 - Học máy, Thị giác máy tính, Xử lý ngôn ngữ tự nhiên, Data engineering
- Với cách tiếp cận dựa trên bài toán (Problem-based), dự án (project-based), người học có được các kỹ năng:
 - Sử dụng các kỹ thuật và công cụ của NLP.
 - Thực hành và giải quyết các vấn đề thực tế

Nội dung khóa học

- Học viên được tiếp cận các khái niệm, kỹ thuật, và mô hình NLP cơ bản và phổ dụng
 - Mô hình ngôn ngữ thống kê
 - Mô hình ngôn ngữ neural
- Học viên được tiếp cận với các bài toán ứng dụng NLP với nhiều ví dụ minh họa, dữ liệu, và dự án từ các bài toán thực tế.
 - Các kỹ thuật xử lý văn bản.
 - Các bài toán ứng dụng như (phân loại văn bản, phân cụm, phân tích quan điểm, tóm tắt tự động ...)

Bài 1

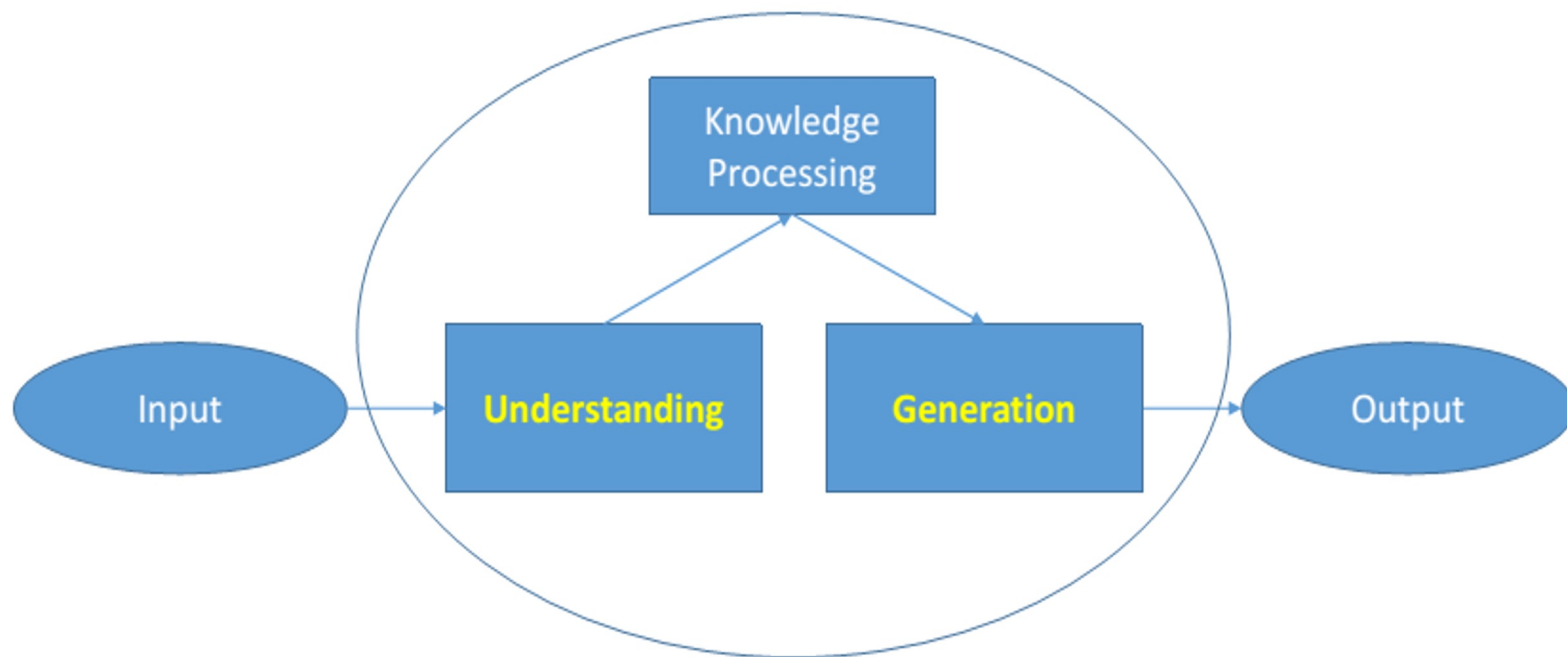
Introduction

Nội dung

- Giới thiệu
- Các mức phân tích trong NLP
- Các ứng dụng của NLP
- Lược sử phát triển của NLP

NLP là gì?

- NLP là lĩnh vực con của trí tuệ nhân tạo, nghiên cứu và tạo ra các ứng dụng giúp máy tính có thể hiểu được và sinh được ngôn ngữ như con người.



Tại sao NLP khó?

- Nhập nhằng (ambiguity).
- Phụ thuộc ngữ cảnh.
- Phụ thuộc văn hóa, vùng miền.
- Phụ thuộc vào các tri thức nền tảng của cá nhân

Nhập nhằng mức từ

- Một câu có thể có n khả năng tách từ, nhưng chỉ 1 trong chúng là đúng
- Vấn đề: chồng chéo từ
 - Học_sinh học_sinh học.
 - Học_sinh học sinh_học.

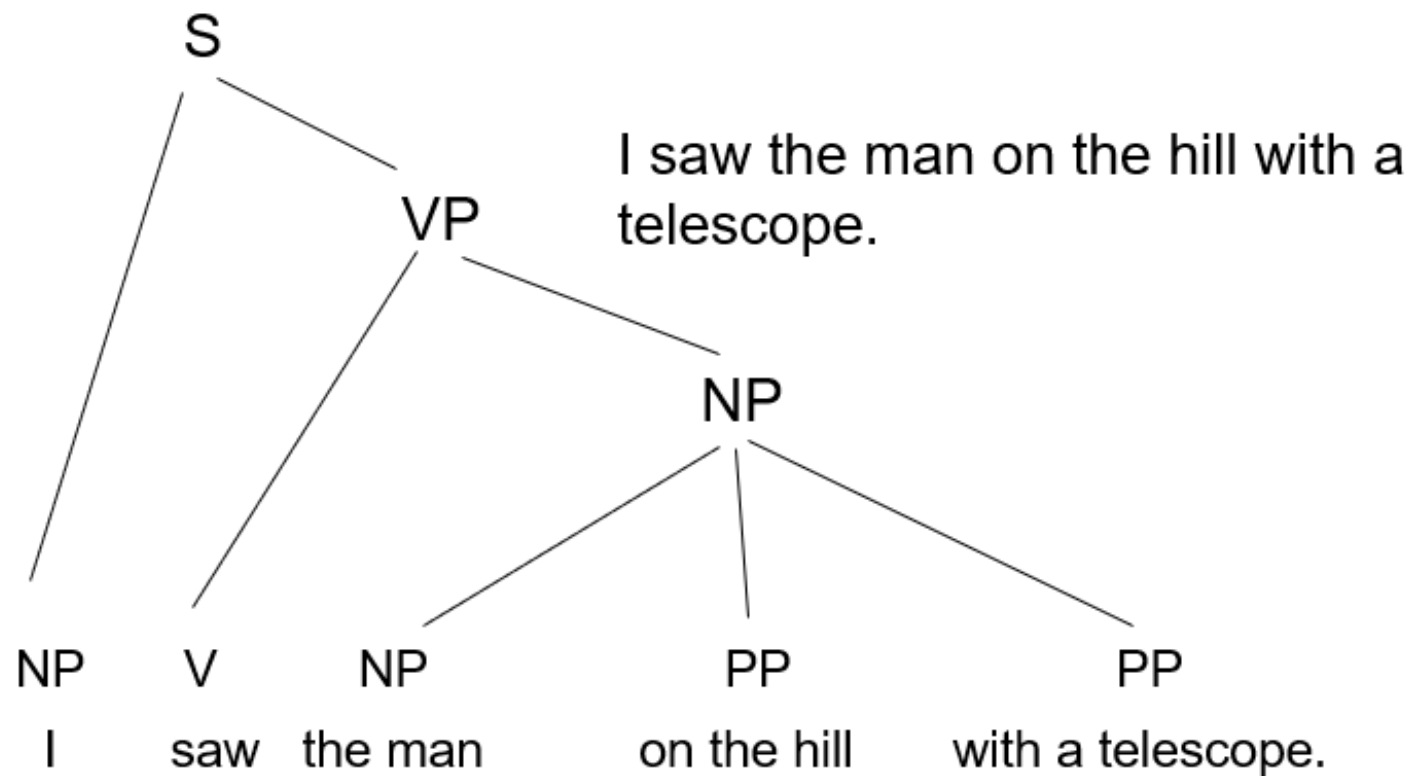
Nhập nhằng trong gán nhãn từ loại

Con ngựa đá con ngựa đá.

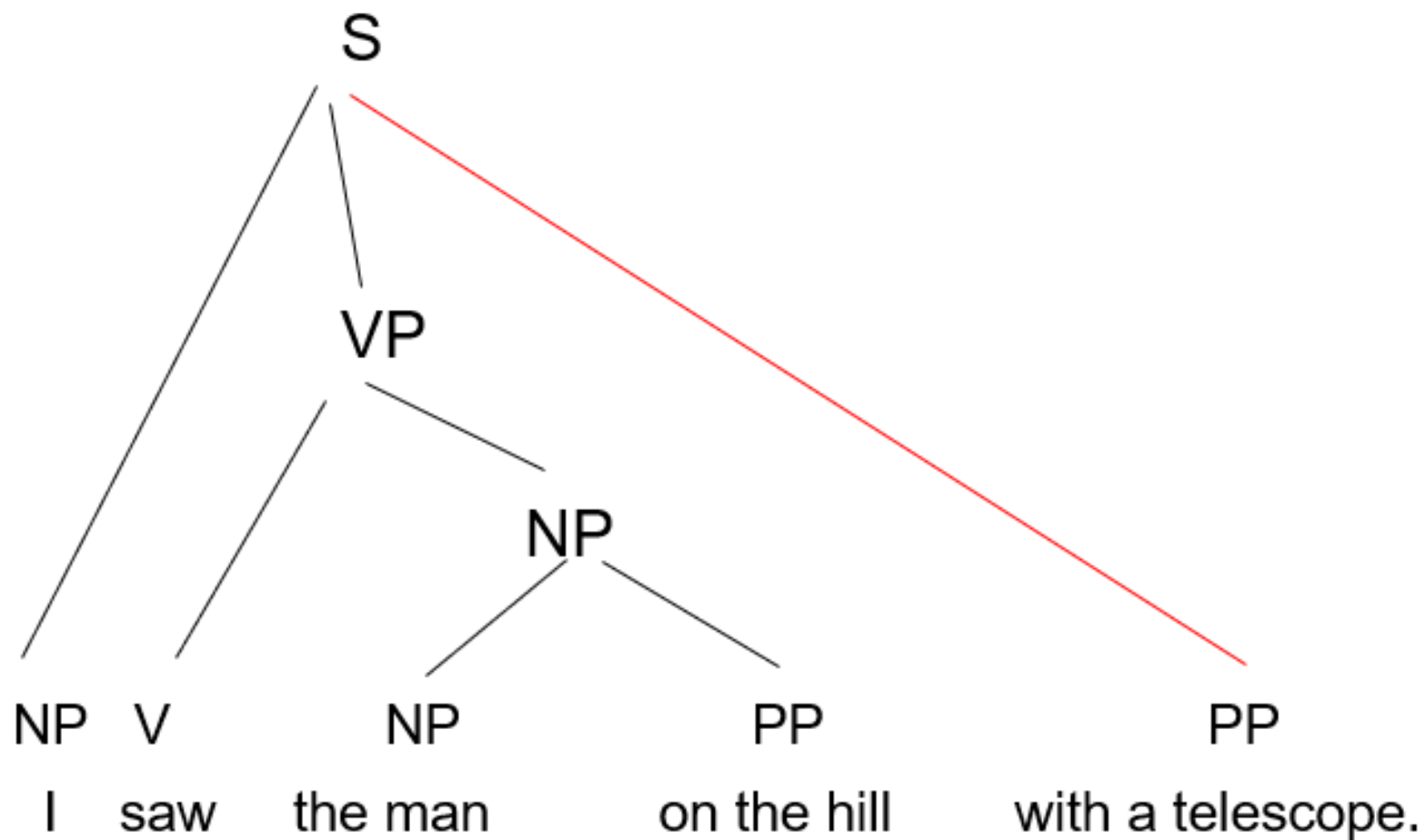
- Con_ngựa/DT đá/ĐgT con_ngựa/DT đá/DT.
- Ông/ĐaT già/TT đi/Phó_từ nhanh/TT quá/trạng_từ.
- Ông_già/DT đi/ĐgT nhanh/TT quá/trạng_từ.

Nhập nhằng trong PTCP

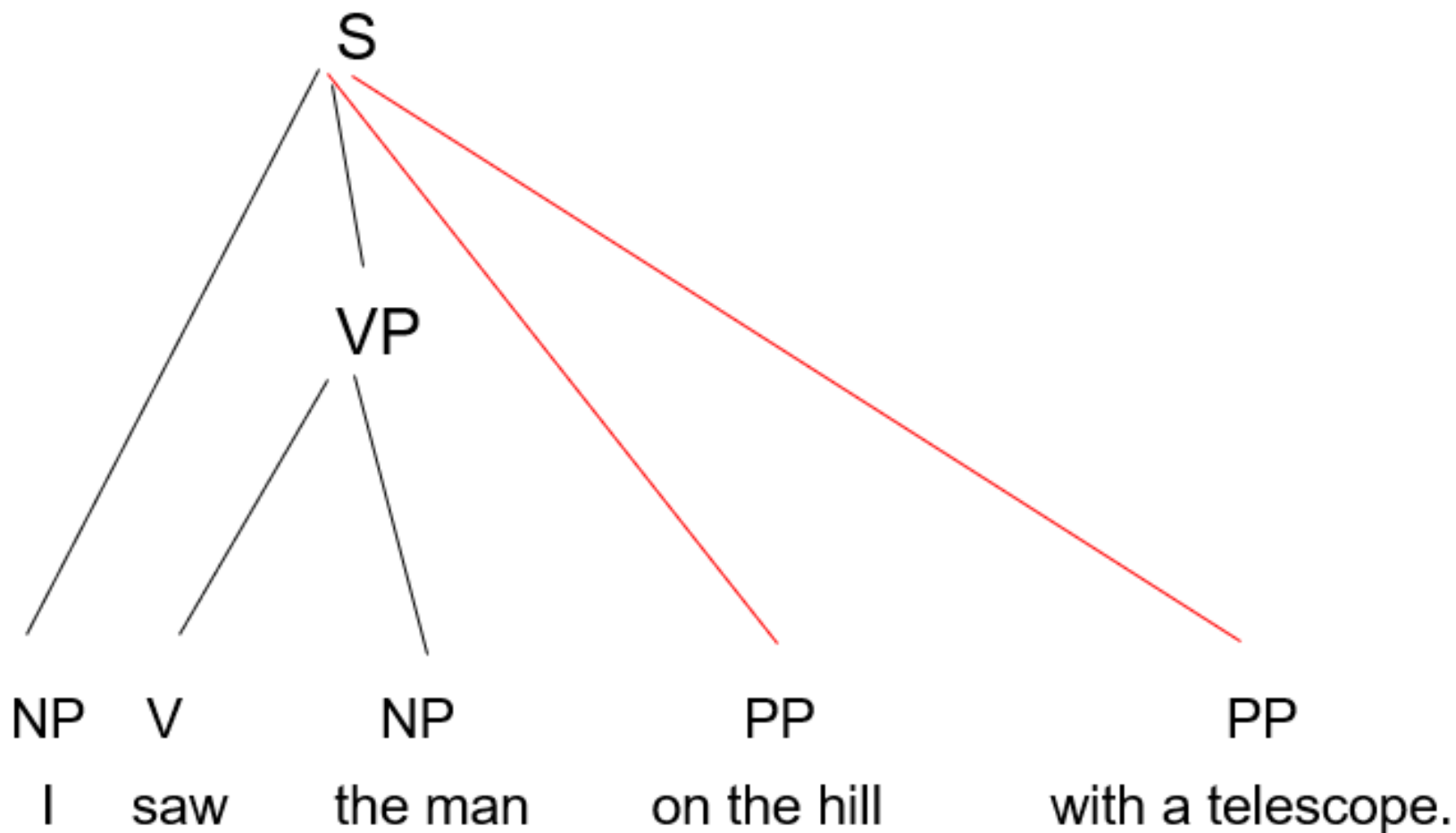
Cây cú pháp theo văn phạm phi ngữ cảnh



Ngữ pháp: nhập nhằng cấu trúc (liên kết)



Ngữ pháp: nhập nhằng cấu trúc (liên kết)



Ngữ pháp: nhập nhằng cấu trúc (từ loại)

Time flies like an arrow.

Time // flies like an arrow.
 VBZ IN (giới từ so sánh)

Time flies // like an arrow.
 NNS VBP

Ngữ pháp: nhập nhằng cấu trúc (từ loại)

Ông_già // đi nhanh quá.

Ông // già đi nhanh quá.

Ngữ nghĩa: nhập nhằng mức từ vựng

- I walked to the bank ...

of the river.

to get money.

- The bug in the room ...

was planted by spies.

flew out the window.

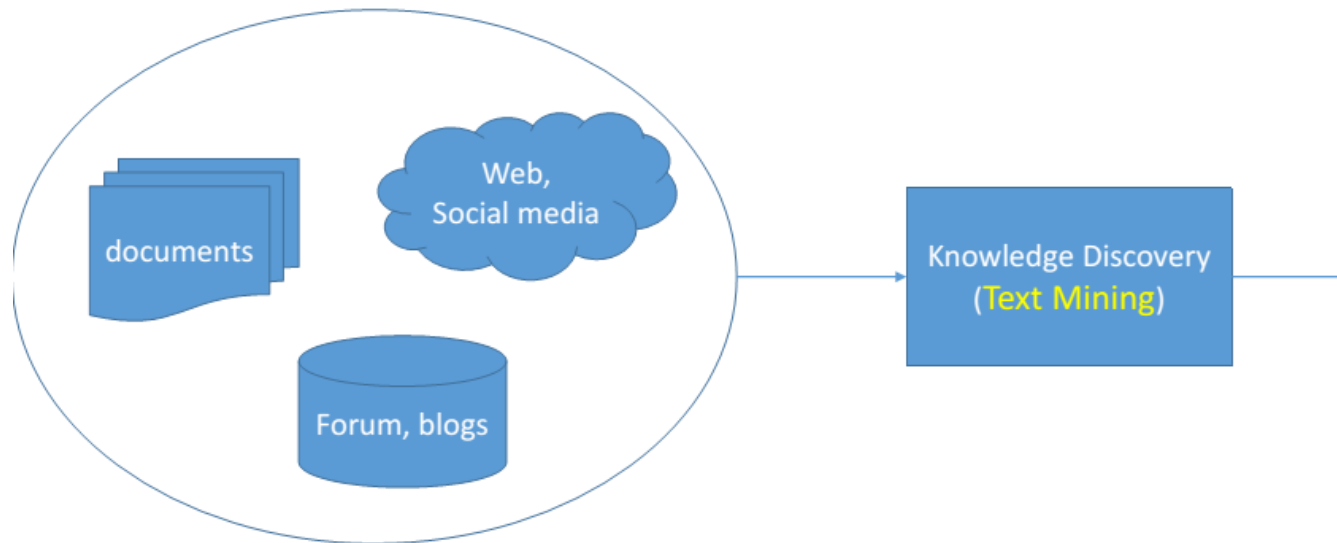
- I work for John Hancock ...

and he is a good boss.

which is a good company.

Tại sao NLP khó?

- Khối lượng dữ liệu khổng lồ phi cấu trúc (unstructured) và bán cấu trúc (semi structure) với những thách thức mới



Nội dung

- Giới thiệu
- Các mức phân tích trong NLP
- Các ứng dụng của NLP
- Lược sử phát triển của NLP

Các mức phân tích trong NLP

- **Lexical (từ vựng)**: cách từ được xây dựng, các tiền tố và hậu tố của từ
- **Syntax (cú pháp)**: mối liên hệ về cấu trúc ngữ pháp giữa các từ và ngữ
- **Semantics (ngữ nghĩa)**: nghĩa của từ, cụm từ, và cách diễn đạt
- **Pragmatic (ngữ dụng)**: mục đích phát ngôn, cách sử dụng ngôn ngữ trong giao tiếp

Các bài toán cơ bản trong NLP

Xử lý từ (word processing)

- Tokenization
- Word segmentation

Các bài toán cơ bản trong NLP

Xử lý cú pháp (Syntactic Processing)

- Morphology analysis
 - Word stemming.
 - Stop word processing.
- Part of Speech (POS) Tagging.
- Syntactic Parsing
 - Syntactic tree generation.
 - Dependency parsing

Các bài toán cơ bản trong NLP

Xử lý ngữ nghĩa (Semantic Processing)

- Ngữ nghĩa từ
 - Word sense disambiguation
 - Anaphora resolution (co-reference).
- Ontological word semantics (Wordnet)
 - Semantic based terms.
 - Semantic relations.
- (semantic) Language Modeling
 - Language generation.
 - Word prediction.

Các bài toán cơ bản trong NLP

Xử lý mức độ ngữ dụng (Pragmatic Processing)

Chưa được xử lý nhiều, bài toán liên quan có:

- Xử lý sắc thái ngôn ngữ
 - Language affection

Tài nguyên ngôn ngữ

Tài nguyên ngôn ngữ (Language Resources) là các tài nguyên phục vụ cho việc xử lý các bài toán NLP trên máy tính:

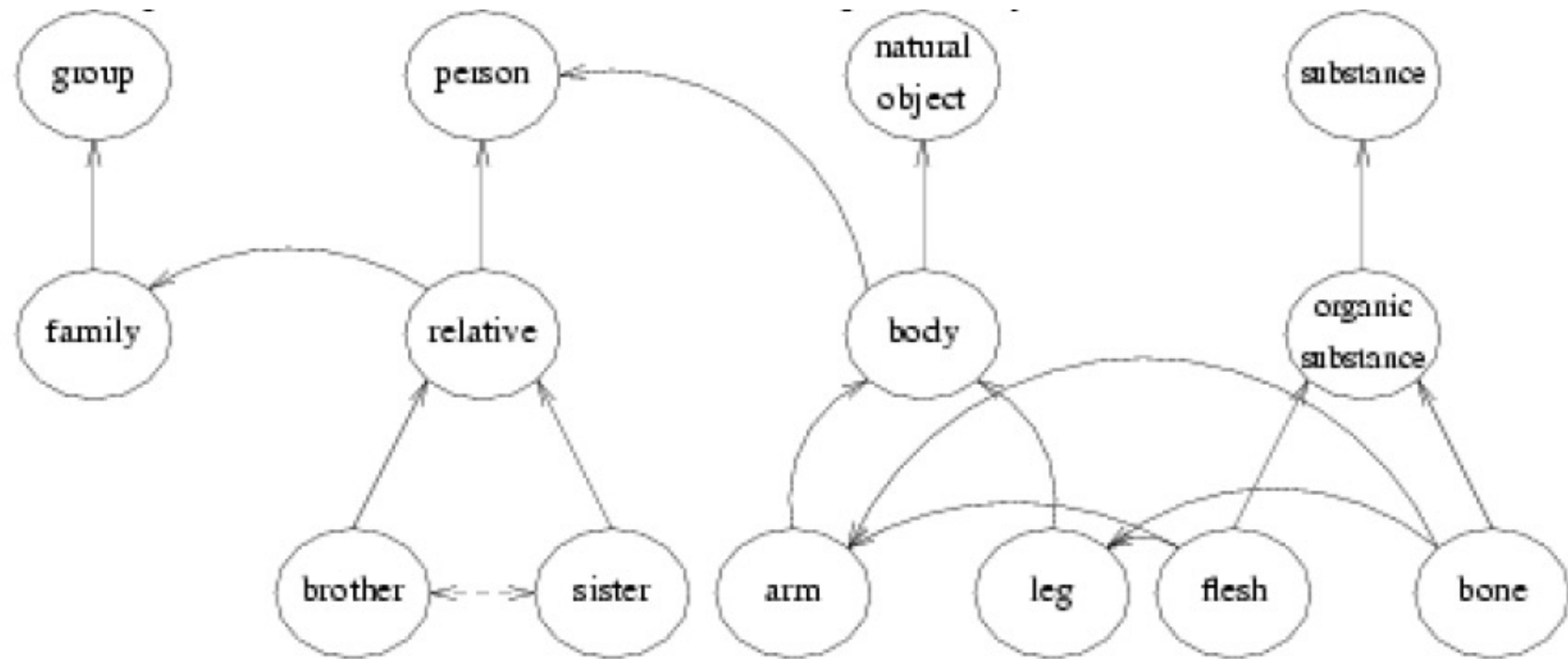
- Các corpora (đơn hoặc đa ngữ).
- Dictionaries, Thesaurus, WordNet
- TreeBank (e.g PENN Tree banks).
- Các thư viện xử lý văn bản (NLTK, Spacy, CoreNLP,...).

PENN Tree banks

```
( (S
  (NP-SBJ
    (NP (NNP Pierre) (NNP Vinken) )
    (, ,)
    (ADJP
      (NP (CD 61) (NNS years) )
      (JJ old) )
    (, ,) )
  (VP (MD will)
    (VP (VB join)
      (NP (DT the) (NN board) )
      (PP-CLR (IN as)
        (NP (DT a) (JJ nonexecutive) (NN director) ))
      (NP-TMP (NNP Nov.) (CD 29) )))
  (. .) ))
```


WordNet

- Các từ nối theo chiều dọc biểu diễn quan hệ rộng (holonymy) - hẹp (hyponymy), theo chiều ngang biểu diễn quan hệ bộ phận meronymy (part_of) và holonymy (has_part) .
- Mỗi nghĩa của từ được biểu diễn bằng 1 số synset



<http://wordnet.princeton.edu/>

→ hyponymy - - - -> antonymy ⤿ meronymy

Các thư viện xử lý văn bản

- **Natural Language Toolkit (NLTK)** (Python)
 - Các tác vụ: chuyển chữ hoa về chữ thường, tách câu, tách từ, lấy gốc từ , loại bỏ ký tự đặc biệt, từ dừng, từ hiếm, emoji, URL,... tìm từ đồng nghĩa, trái nghĩa
- **SpaCy** (Python và Cython)
 - Nhanh hơn NLTK
 - Các tác vụ: tách câu, tách từ, loại bỏ từ dừng, lấy gốc từ , gán nhãn từ loại, phân tích cú pháp phụ thuộc, nhận dạng thực thể có tên
- **CoreNLP** (java)
 - tách câu, tách từ, lấy gốc từ , gán nhãn từ loại, phân tích cú pháp phụ thuộc, nhận dạng thực thể có tên, phân giải đồng tham chiếu

Các bài toán trong CoreNLP

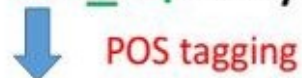
- Ông ấy nói: “tốc độ truyền thông tin ngày càng cao”.



- Ông ấy nói : “ tốc độ truyền thông tin ngày càng cao ” .



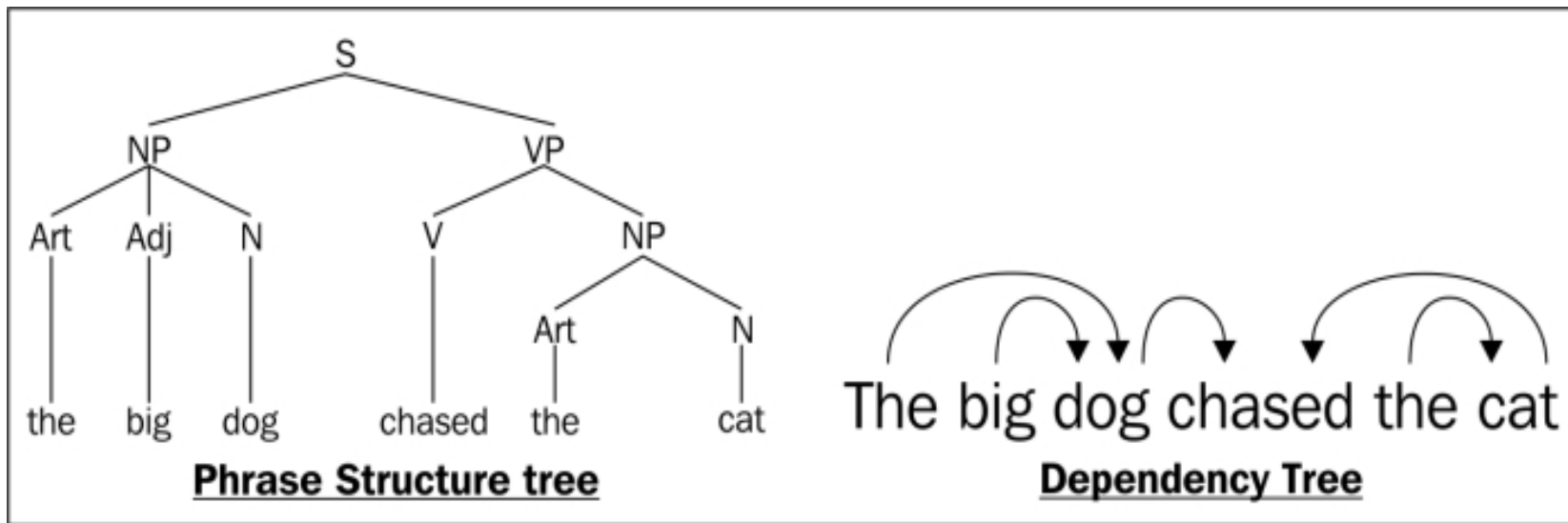
- Ông_ấy nói : “ tốc_độ truyền thông_tin ngày_càng cao ” .



- Ông_ấy/N nói/V :/sym “/sym tốc_độ/N truyền/V thông_tin/N ngày_càng/adv cao/Adv ” .

Các bài toán trong CoreNLP

Phân tích cú pháp



Các bài toán trong CoreNLP

Word Sense Disambiguation

Sau khi bị **bồ đá**, nó đã trở thành một người khác hẳn.

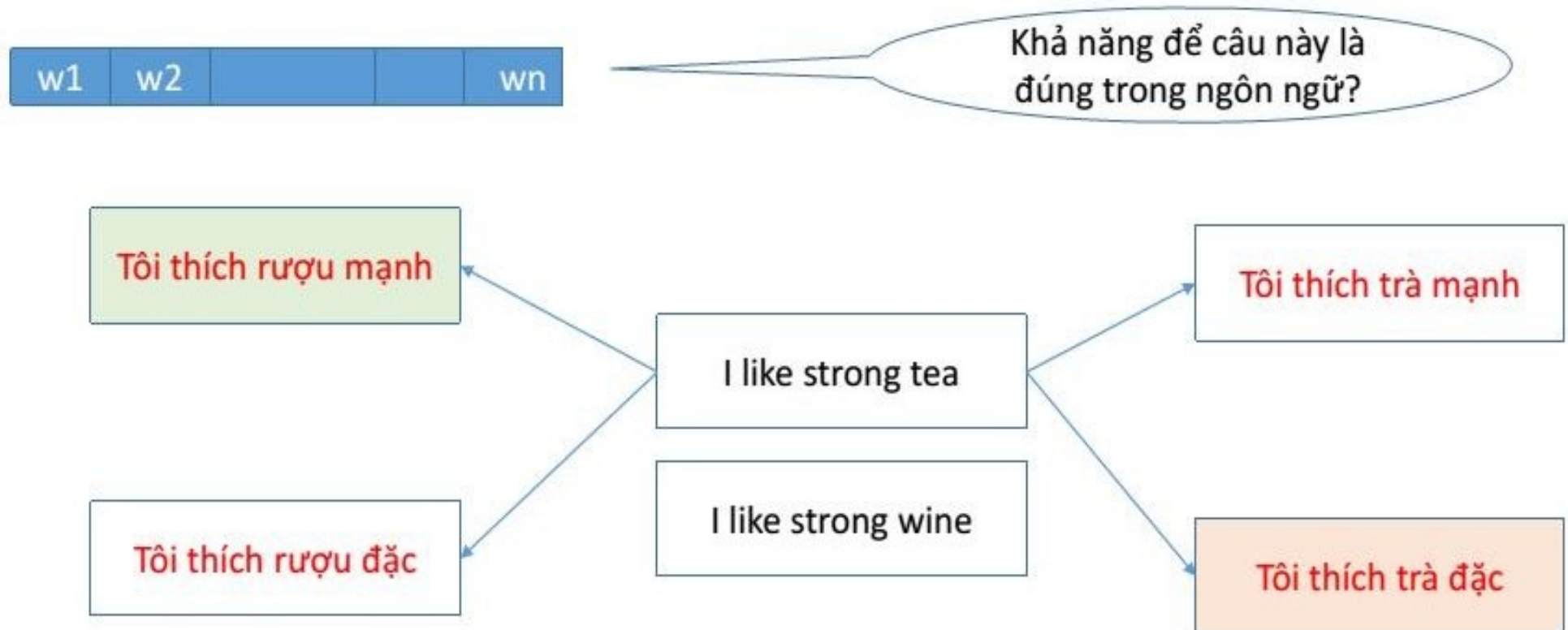
Co-reference

Tôi dễ quên chìa khóa trong phòng mà giờ nó bị khóa cửa rồi.



Các bài toán trong CoreNLP

Mô hình ngôn ngữ



Các bài toán trong CoreNLP

Mô hình ngôn ngữ



Các bài toán trong CoreNLP

Mô hình ngôn ngữ



Các bài toán trong CoreNLP

Mô hình ngôn ngữ

Tôi thích khoa

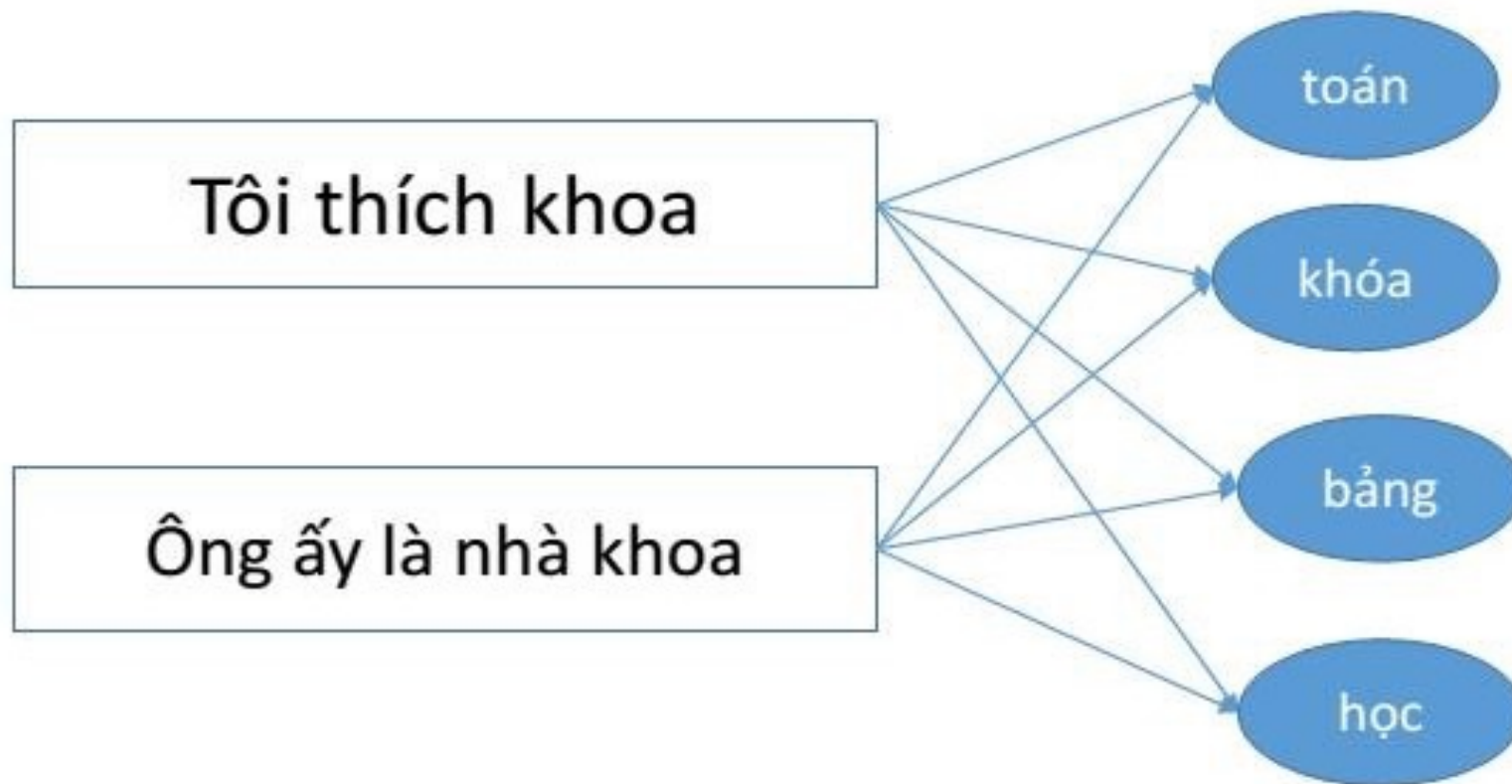
???

Ông ấy là nhà khoa

???

Các bài toán trong CoreNLP

Mô hình ngôn ngữ



Các thư viện cho các bài toán cơ bản cho tiếng Việt

- **Underthesea** (Python): tách từ, gán nhãn từ loại, phân tích cú pháp phụ thuộc, phân loại văn bản, phân tích quan điểm, nhận dạng thực thể có tên
- **VnCoreNLP** (java): tách từ, gán nhãn từ loại, phân tích cú pháp phụ thuộc, nhận dạng thực thể có tên
- **vn.vitk** (java): tách từ, gán nhãn từ loại, phân tích cú pháp phụ thuộc
- **pyVi** (python): tách từ, gán nhãn từ loại, loại bỏ dấu (accents removal), thêm dấu (accents adding)

Đầu ra của VnCoreNLP

Index	Word	POS	NER	head index	dependency relation
1	Ông	Nc	O	4	sub
2	Nguyễn_Khắc_Chúc	Np	B-PER	1	nmod
3	đang	R	O	4	adv
4	làm_việc	V	O	0	root
5	tại	E	O	4	loc
6	Đại_học	N	B-ORG	5	pob

Nội dung

- Giới thiệu
- Các mức phân tích trong NLP
- Các ứng dụng của NLP
- Lược sử phát triển của NLP

Phân nhóm ứng dụng NLP

Information Extraction

- Name Entity Recognition
- Job information extraction
- Sentiment extraction
- Keyword extraction

Text Generation

- Writing suggestion
- News generation
- Summarization
- Chatbot
- Question answering
- Machine translation

Text Classification

- Spam filtering
- Document classification
- Sentiment classification (Social listening)
- Recommendation

Phân nhóm ứng dụng NLP

■ Discourse Analysis

- Coreference Resolution
- Clause-level Discourse Analysis

■ Document checking

- Spell checking
- Grammar checking
- Plagiarism checking

■ Other applications

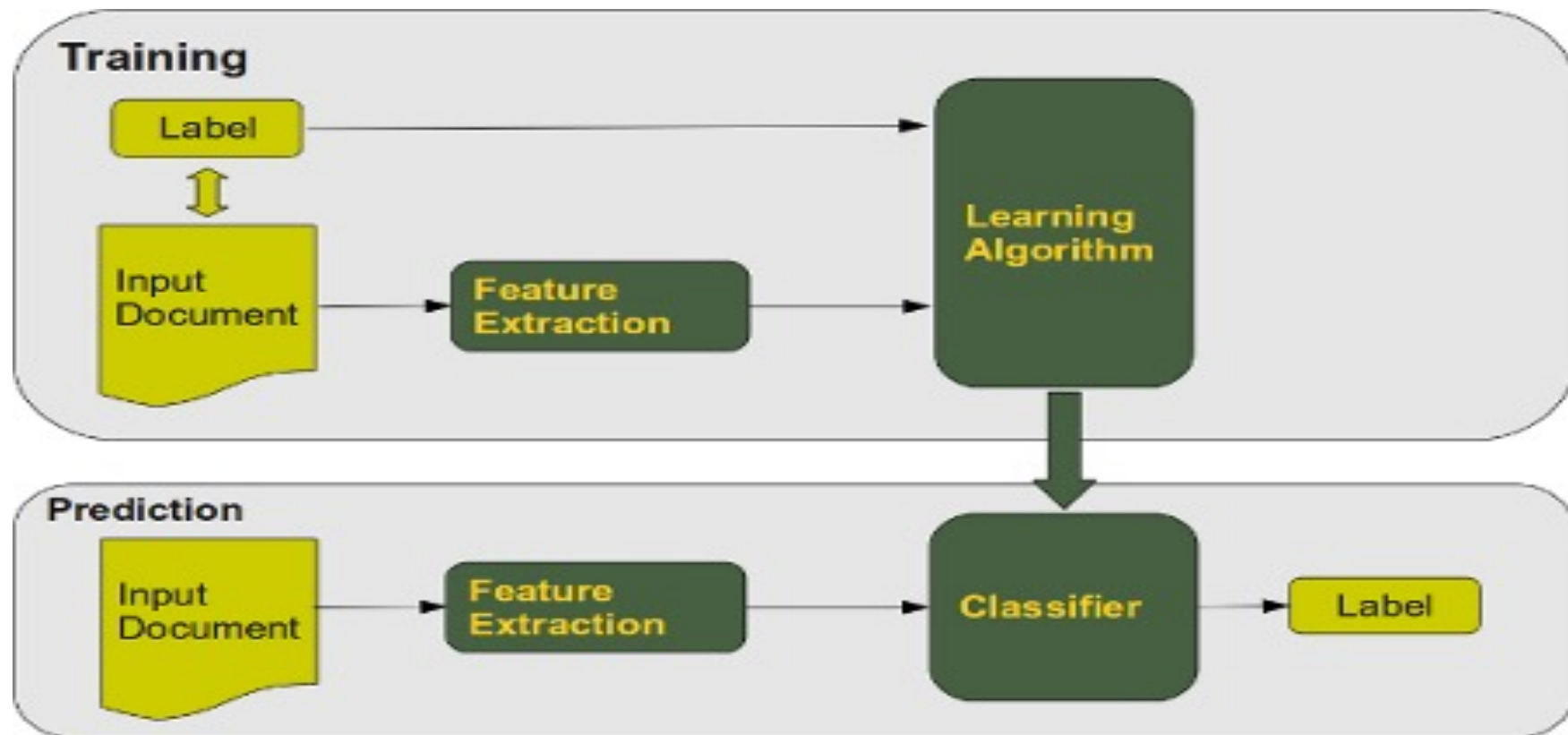
- OCR (optical character recognition)
- Speech recognition
- Text Clustering
- Information Retrieval

Các ứng dụng của NLP

1. Phân loại văn bản (document classification)
2. Tìm kiếm thông tin (information retrieval)
3. Trích rút thông tin (information extraction)
4. Phân tích cảm xúc (sentiment analysis)
5. Tóm tắt văn bản (text summarization)
6. Dịch máy (machine translation)
7. Hỏi đáp (question answering/chatbot)
8. Hệ gợi ý (recommender system)
9. Nhận dạng chữ viết (OCR)
10. ...

Phân loại văn bản

Ví dụ: phân loại văn bản theo chủ đề, phát hiện spam mail, ...



Phân cụm văn bản

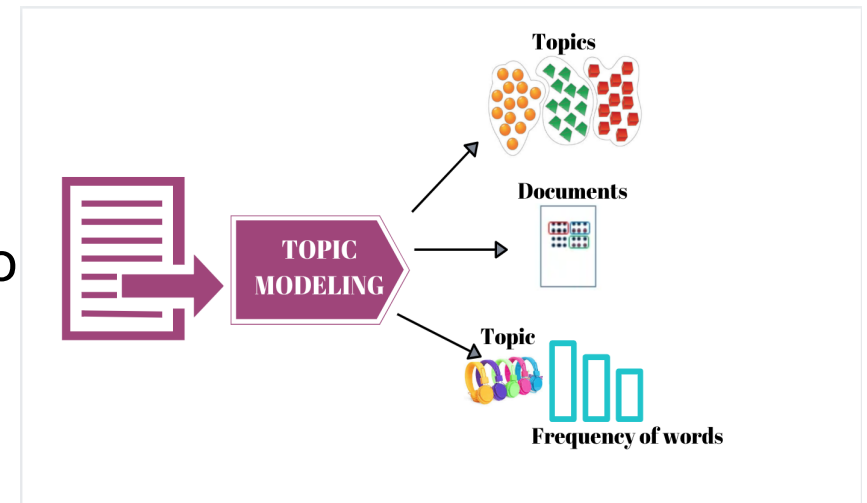
Bài toán đặt ra là đối với 1 tập các văn bản (documents) cần xác định được:

1. Các topic (chủ đề) ẩn trong tập các văn bản.
2. Phân cụm các văn bản theo các chủ đề nêu trên.

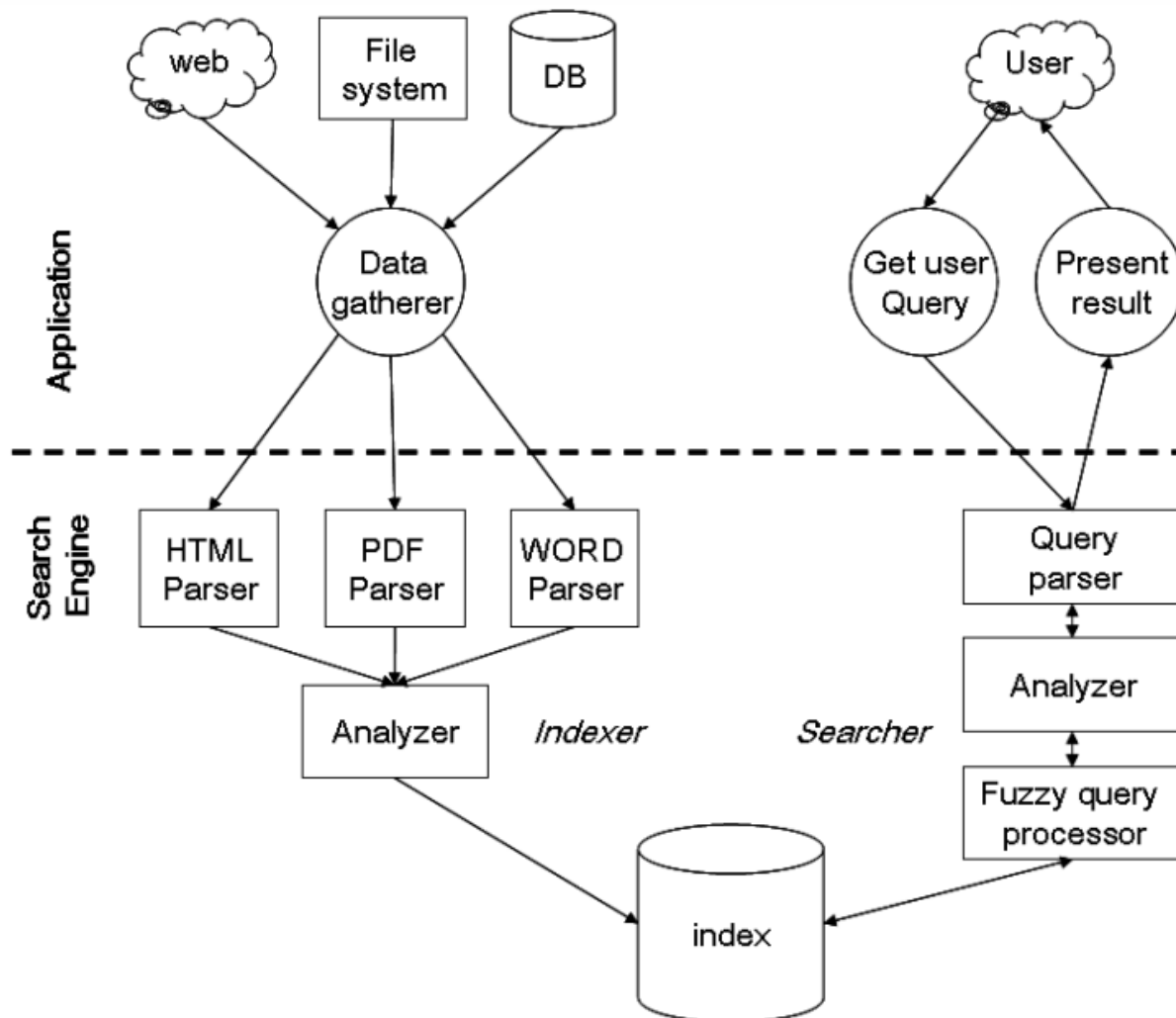
Thường chủ đề được xác định bởi 1 tập các từ khóa và tần xuất.

Các kỹ thuật phổ biến:

- LSA (Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation).
- NMF (Non-negative Matrix Factorization)



Tìm kiếm thông tin



Trích rút thông tin

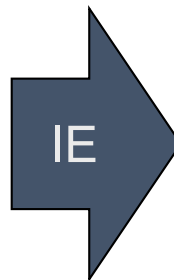
October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation CEO Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Phân tích cảm xúc

- Phân loại: **positive**, **negative** and **neutral**

- Ứng dụng:

Social listening; Reputation management; Advertisement campaign evaluation;...



Phân tích cảm xúc

- Stanford CoreNLP (<https://corenlp.run>)

— Text to annotate —

President Trump said Thursday that the United States would raise tariffs on \$200 billion worth of Chinese goods at 12:01 a.m. Friday and was taking steps to tax nearly all of China's imports as he accused Beijing of backtracking on a trade deal.

— Annotations —

sentiment ✕

— Language —

English ▼

Submit

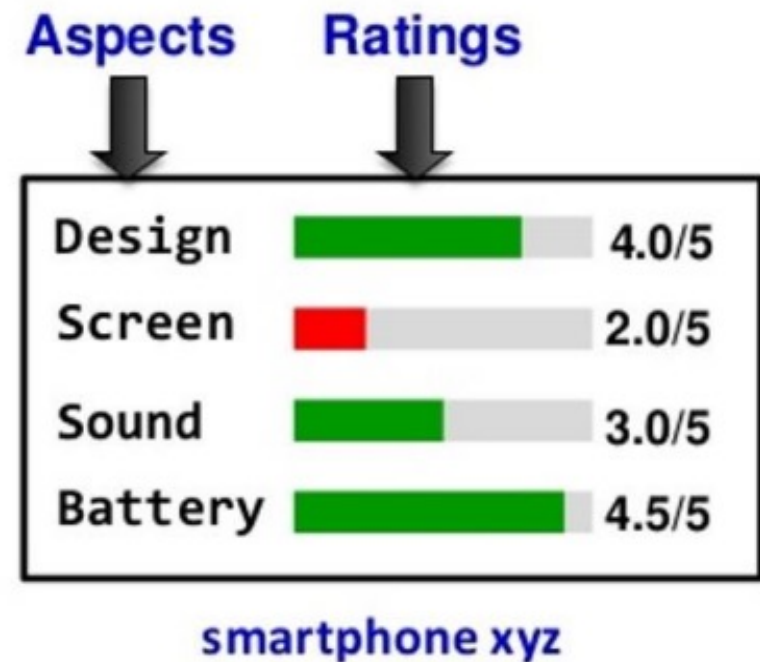
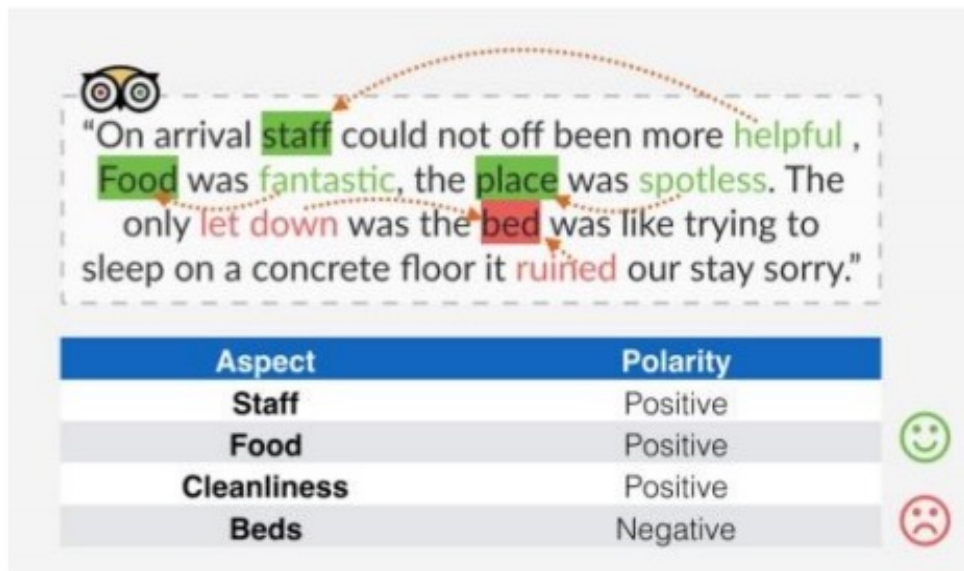
Sentiment:

1

NEGATIVE

President Trump said Thursday that the United States would raise tariffs on \$ 200 billion worth of Chinese goods at 12:01 a.m. Friday and was taking steps to tax nearly all of CI

Phân tích cảm xúc



Newsinessence [Radev & al. 01]

NewsInEssence: Web-based News Summarization - Microsoft Internet Explorer provided by AT&T WorldNet Service

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print Copy Paste

Links Customize Links Free Hotmail Windows Media Windows Like Music - Try AOL! WorldNet Customer Care WorldNet Home Page WorldNet Member Services

Address http://www.newsinessence.com/nie.cgi Go

...www...NewsInEssence...com...

Interactive Multi-source News Summarization

Home
[Current Clusters](#)
[Create Cluster](#)
[Summarize Cluster](#)
[Track Cluster](#)
[User Cluster Archive](#)
[CIDR Cluster Archive](#)
[Google Cluster Archive](#)

Help
[About NewsInEssence](#)
[Contact Us](#)

CLAIM
MEAD
[summarization.com](#)

4 Killed In Florida Fireworks Blast July 2, 2003 19:10:48

4 Killed In Florida Fireworks Blast July 2, 2003 19:10:48. BONITA SPRINGS, Fla., July 2, 2003 Investigators and firefighters gather at the scene of a tractor-trailer that exploded as workers were unloading fireworks in Bonita Springs, Fla., Wednesday, July 2, 2003. Kevin McKenzie was mowing a strip of grass at Lover's Key about 300 feet from the tractor trailer when the explosion happened at 2:10 p.m., shooting flames and fireworks from the truck.

[\[8 Articles from 7 Sources\]](#) [\[4 Summaries\]](#)

Recent User Clusters (more)

- ['Liberia's Taylor bans church radio station'](#)
11 articles, 3 summaries: 07/02, 9:57 PM
- ['Knesset backs Sharon on roadmap'](#)
7 articles, 3 summaries: 07/01, 11:48 AM
- ['Israel pulls out of Bethlehem'](#)
5 articles, 4 summaries: 07/01, 11:25 AM

Recent CIDR Clusters (more)

- ['Bush challenge to Iraq attackers: Bring them on'](#)
25 articles, 4 summaries: 07/02, 7:40 PM
- ['Bill sparks massive Hong Kong protest'](#)
14 articles, 4 summaries: 07/02, 7:40 PM
- ['Edinburgh Evening News - Top Stories - Palestinian police back in Bethlehem'](#)
13 articles, 4 summaries: 07/02, 7:40 PM

NIE Headlines
[Build your own cluster of articles.](#)

NewsTroll from URL:
URL must be from [CNN](#), [Yahoo!](#), [MSNBC](#), [BBC](#), or [USA Today](#).

NewsTroll from query:

[Advanced Options](#)

User Clusters (Archive)

- ['Liberia's Taylor bans church radio station'](#)
11 articles, 3 summaries: 07/02, 9:57 PM
- ['Knesset backs Sharon on roadmap'](#)
7 articles, 3 summaries: 07/01, 11:48 AM
- ['Israel pulls out of Bethlehem'](#)
5 articles, 4 summaries: 07/01, 11:25 AM
- ['India cool on Pakistan offer'](#)
1 article, 3 summaries: 06/25, 10:33 AM

4 Killed In Florida Fireworks Blast July 2, 2003 19:10:48

produced on 07/02, 7:40 PM

2% Summary

4 Killed In Florida Fireworks Blast July 2, 2003 19:10:48 (4:1) BONITA SPRINGS, Fla., July 2, 2003 Investigators and firefighters gather at the scene of a tractor-trailer that exploded as workers were unloading fireworks in Bonita Springs, Fla., Wednesday, July 2, 2003. (4:2)

Done Internet

NewsInEssence: Web-based News Summarization - Microsoft Internet Explorer provided by AT&T WorldNet Service

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Dell Home Real.com Messenger


Address <http://www.newsinessence.com/nie.cgi?CID=20020830135218> Go

Birmingham, England, according to police in Vaesteraas, 60 miles northwest of the capital, Stockholm. (Z:6) Security officers at Vaesteraas airport found the weapon in a toiletries bag when they scanned the man's hand luggage on Thursday, police spokesman Ulf Palm said. (Z:7)


Summaries of all documents: [\[10%\]](#) [\[20%\]](#)

Cluster Documents

Included	Index	Title	Source	Publication Date
<input checked="" type="checkbox"/>	1	Hijack suspect 'denies having gun' [Use As Seed] http://news.bbc.co.uk/1/hi/world/europe/2224395.stm	news.bbc.co.uk	08/30, 5:23 PM
<input checked="" type="checkbox"/>	2	Swedish airport security praised [Use As Seed] http://news.bbc.co.uk/1/hi/world/europe/2225741.stm	news.bbc.co.uk	08/30, 12:34 PM
<input checked="" type="checkbox"/>	3	'It can't get more scary than this' [Use As Seed] http://news.bbc.co.uk/1/hi/world/europe/2225342.stm	news.bbc.co.uk	08/30, 11:10 AM
<input checked="" type="checkbox"/>	4	Hijack suspect 'not attending conference' [Use As Seed] http://news.bbc.co.uk/1/hi/england/2225318.stm	news.bbc.co.uk	08/30, 8:54 AM
<input checked="" type="checkbox"/>	5	Terror experts quiz hijack suspect [Use As Seed] http://www.cnn.com/2002/WORLD/europe/08/30/stockholm.gun/index.html	www.cnn.com	08/30, 5:57 AM
<input checked="" type="checkbox"/>	6	Swede charged with plans to hijack plane [Use As Seed] http://www.msnbc.com/news/801304.asp?cp1=1	www.msnbc.com	08/30, 12:00 AM
<input checked="" type="checkbox"/>	7	Swede faces attempt hijack charge [Use As Seed] http://www.msnbc.com/news/801297.asp	www.msnbc.com	08/29, 12:00 AM

 Redraw Reset Compression: [10%](#) Summarize

Track This Topic: Receive an update on this topic via email

 Email: Time: [Now!](#) Go

Internet 49

Dịch máy

→ ↺ translate.google.com/?hl=vi&sl=en&tl=vi&text=Natural%20language%20prompting%20is%20the%20method%20of%20reformatting%20NLP%... 🔍 ↗ ☆ 🏠 📺

Google Dịch

📄 Văn bản

🖼️ Hình ảnh

📄 Tài liệu

🌐 Trang web

Phát hiện ngôn ngữ Anh Pháp Việt ▼

↔️ Việt Pháp Anh ▼

Natural language prompting is the method of reformatting NLP tasks in the format of a natural language response to natural language input. The development of text-to-text pretrained models such as T5 (Raffel et al., 2020) makes prompts a particularly useful method for multitask learning. For example, Khashabi et al. (2020a) reformat 20 question-answering datasets into a single prompt of question: ... (A)... (B)... (C)... context: ..., while later work such as Zhong et al. (2021) and Wang et al. (2021) cast a range of datasets into a single boolean QA prompt or a single NLI prompt, respectively. Although effective, these single-prompt methods typically do not generalize to new prompts or new tasks inexpressible in their fixed format. ✕

Nhắc ngôn ngữ tự nhiên là phương pháp định dạng lại các nhiệm vụ NLP theo định dạng của phản hồi ngôn ngữ tự nhiên đối với đầu vào ngôn ngữ tự nhiên. Sự phát triển của các mô hình huấn luyện trước chuyển văn bản thành văn bản như T5 (Raffel và cộng sự, 2020) làm cho lời nhắc trở thành một phương pháp đặc biệt hữu ích cho việc học đa nhiệm. Ví dụ, Khashabi et al. (2020a) định dạng lại 20 bộ dữ liệu trả lời câu hỏi thành một câu hỏi duy nhất: ... (A)... (B)... (C)... ngữ cảnh: ..., trong khi công việc sau này chẳng hạn như Zhong et al. (2021) và Wang et al. (2021) truyền một loạt bộ dữ liệu vào một dấu nhắc QA boolean duy nhất hoặc một dấu nhắc NLI tương ứng. Mặc dù hiệu quả, nhưng các phương pháp nhắc đơn này thường không khái quát hóa thành các lời nhắc mới hoặc tác vụ mới không thể diễn đạt được ở định dạng cố định của chúng. ☆

Dịch máy

vnexpress-net.translate.google/ha-noi-chap-nhan-dap-an-lam-nham-de-thi-toan-in-mo-4616597.html?_x_tr_sl=vi&_x_tr_tl=en&_x_tr_hl=vi&_x_tr_pto...

Google Dịch Việt → Anh Bản dịch ▾

VNEXPRESS Monday, 12/6/2023 Hanoi 30° Local weather × Latest News by region International

News View World Video Podcasts Business Real estate Science Entertainment Sport Law Education Health Life Tourism

Education > Admissions > Common Monday, 12/6/2023, 18:27 (GMT+7)

Hanoi accepts the wrong answer for the math exam

The Hanoi Department of Education and Training accepts and calculates points for the candidate's wrong answer because the grade 10 math exam is blurry, on the afternoon of June 12.

Mr. Ha Xuan Nham, Head of Secondary Education Department, Hanoi Department of Education and Training, said that after reviewing the entire process, comparing the original and the copy, the Department evaluated the math exam questions in grade 10 content errors. However, in the process of copying more than 100,000 copies

BEST SELLING
- Our Most Popular

View more

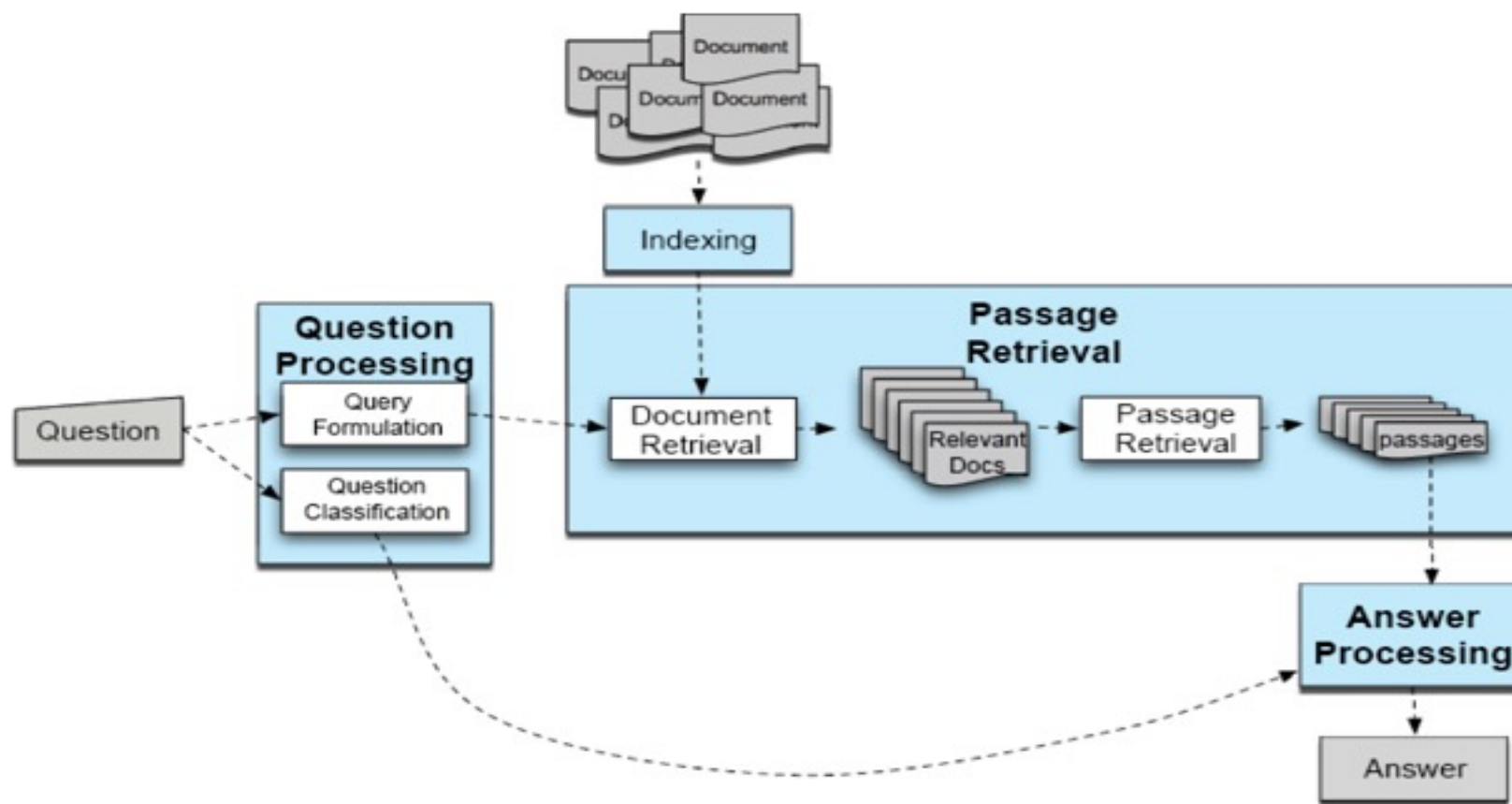
Answers to the 10th grade English exam in Hanoi

Hỏi đáp tự động

- Information Retrieval based Question Answering
(Hỏi đáp dựa trên tìm kiếm)
- Community based Question Answering
(Hỏi đáp dựa trên cộng đồng)
- Conversation/Chatbot
(Hội thoại)

Hỏi đáp tự động

IR-based question answering (SOTA – entity oriented)



ChatGPT

H

Cho đoạn văn sau: "Ngày 20/9, một tuần sau hỏa hoạn tại chung cư mini 10 tầng ở phố Khương Hạ, quận Thanh Xuân, Cơ quan Cảnh sát điều tra Công an Hà Nội thông báo nguyên nhân gây cháy do "chập mạch điện đường dây dẫn điện tại khu vực của bình ắc quy" nằm ở phần đầu của chiếc xe máy tay ga."

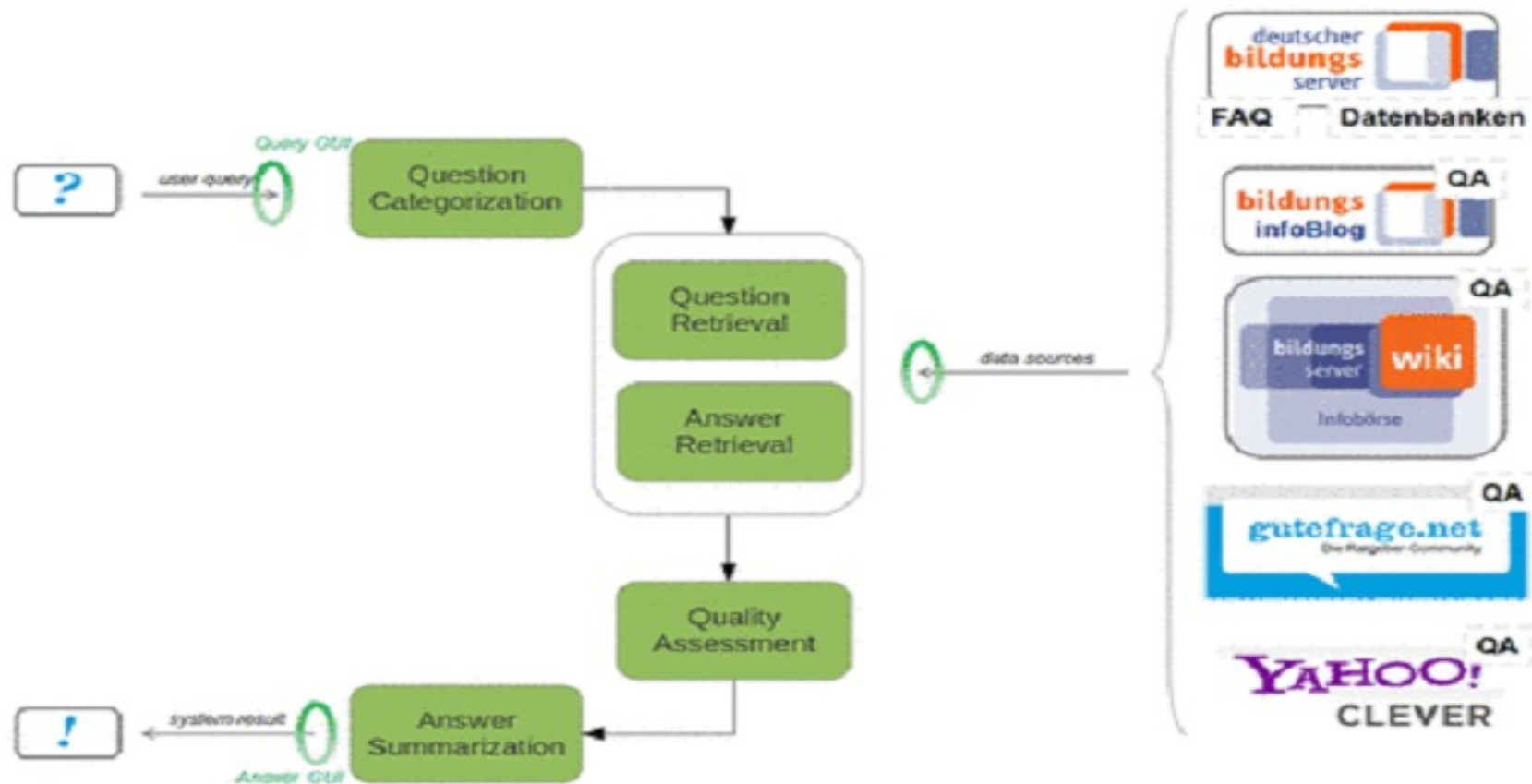
Hãy trả lời câu hỏi: " Nguyên nhân hỏa hoạn do đâu?"



Nguyên nhân hỏa hoạn là do "chập mạch điện đường dây dẫn điện tại khu vực của bình ắc quy" nằm ở phần đầu của chiếc xe máy tay ga.

Hỏi đáp tự động

Community based question answering (www.quora.com)

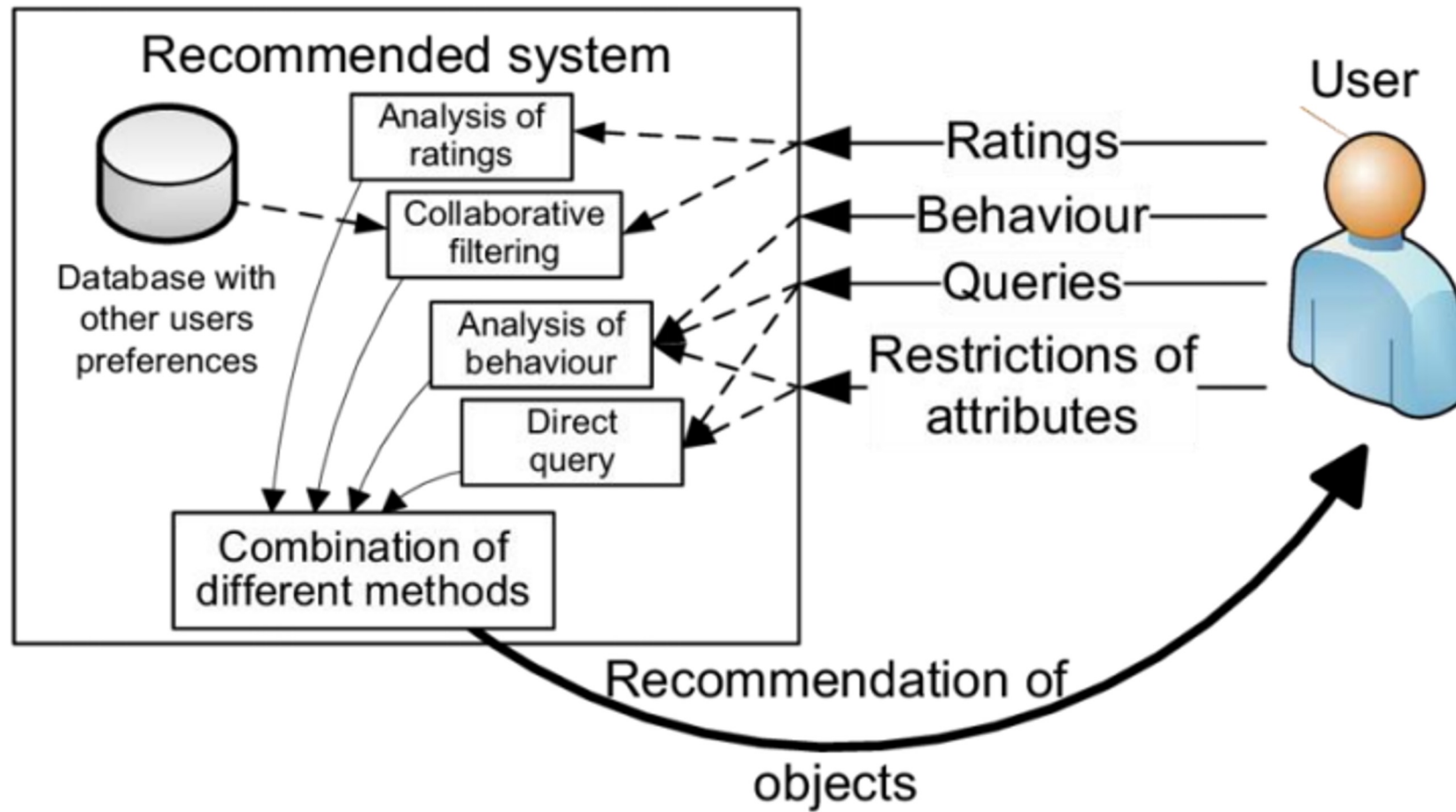


Hội thoại (Chatbot, Personal Assistant)

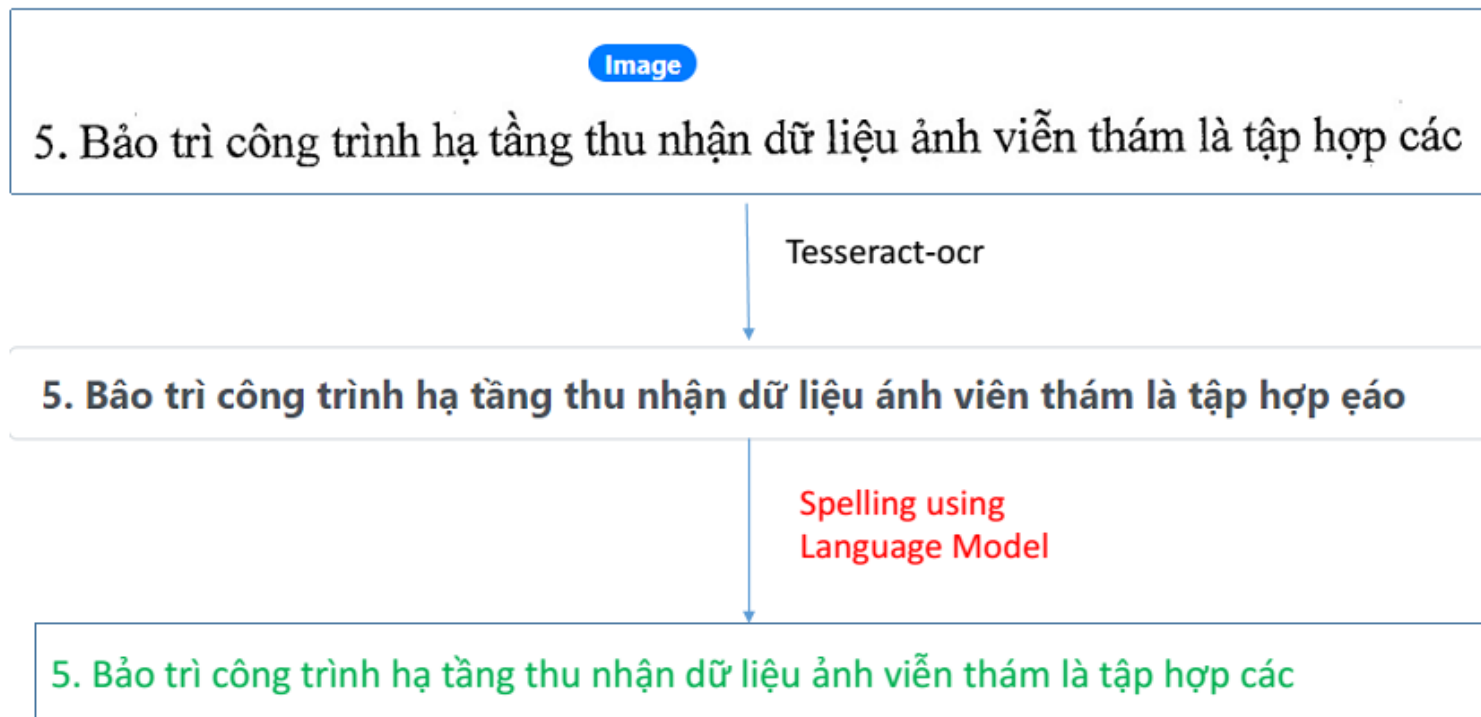
Ví dụ: Alexa, Google Assistant, Google DialogFlow, MS Bot Framework.



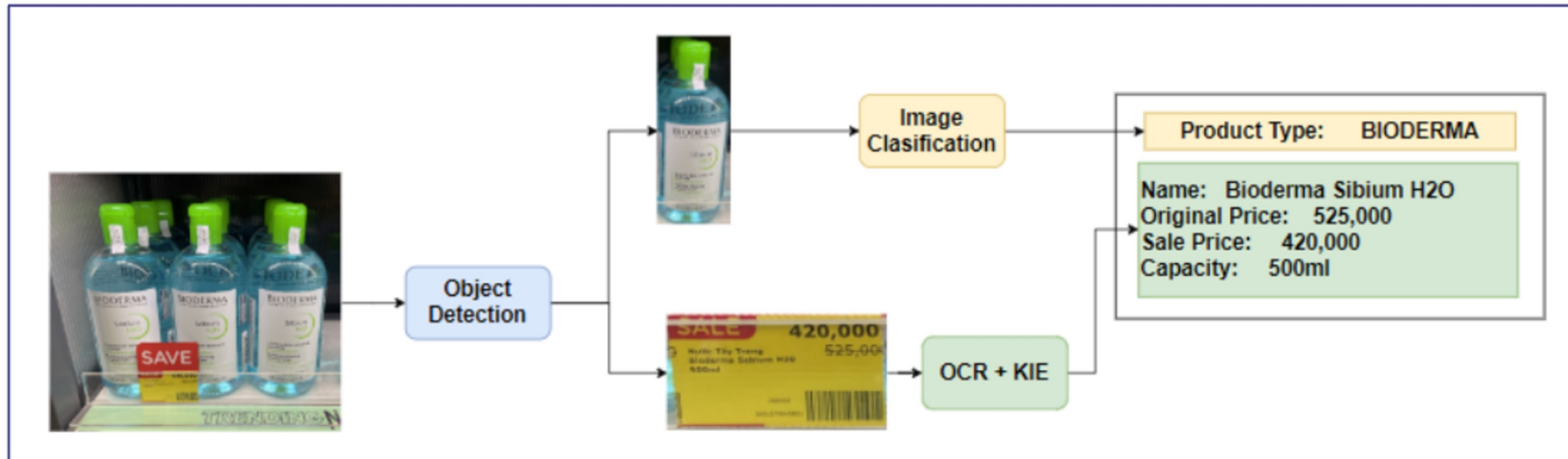
Hệ gợi ý



Nhận dạng hình ảnh + chữ viết



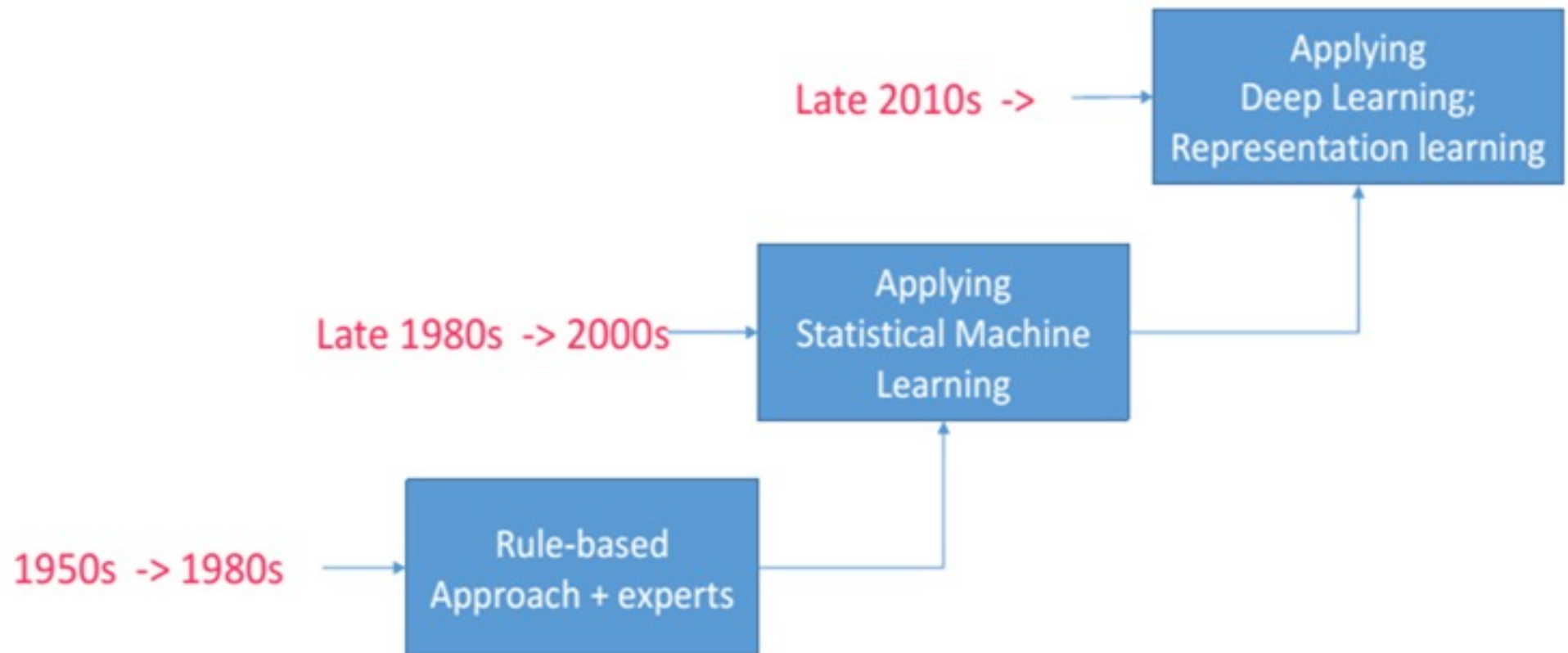
Nhận dạng hình ảnh + chữ viết



Nội dung

- Giới thiệu
- Các mức phân tích trong NLP
- Các ứng dụng của NLP
- **Lược sử phát triển của NLP**

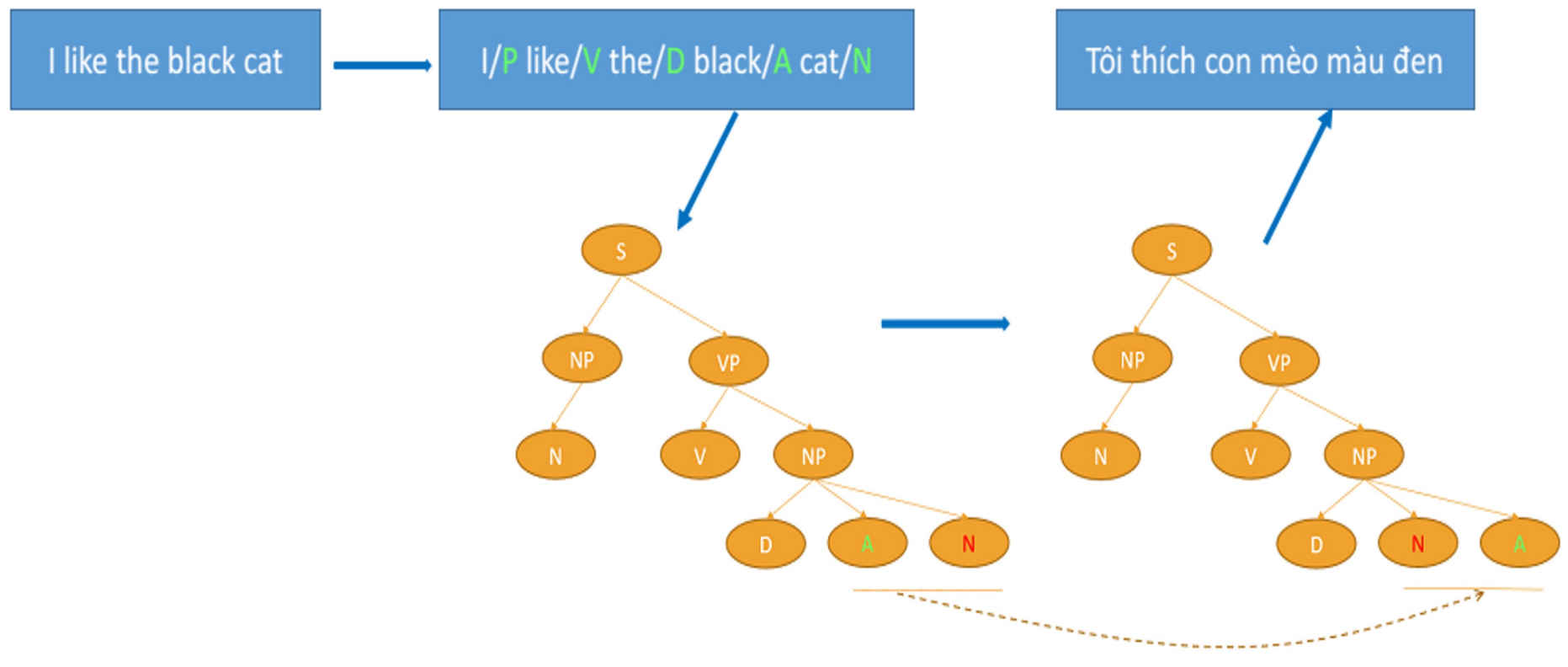
Lược sử phát triển của NLP



Lược sử phát triển của NLP

Ví dụ về dịch máy (Machine Translation)

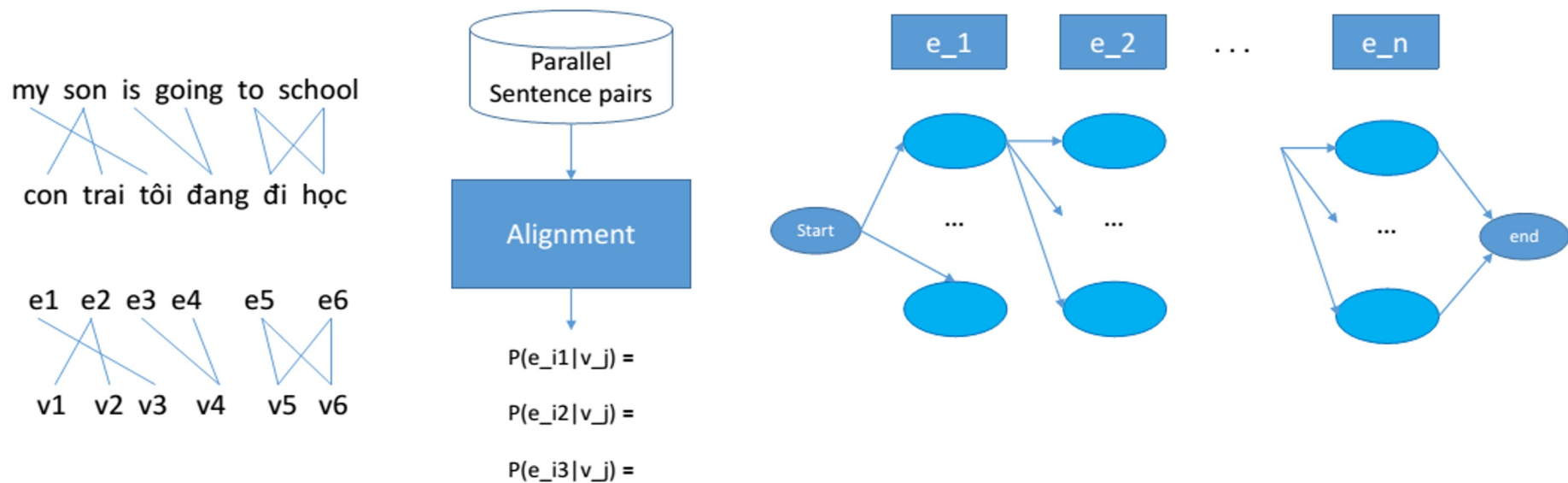
Giai đoạn 1: Rule Based Machine Translation (RBMT)



Lược sử phát triển của NLP

Ví dụ về dịch máy (Machine Translation)

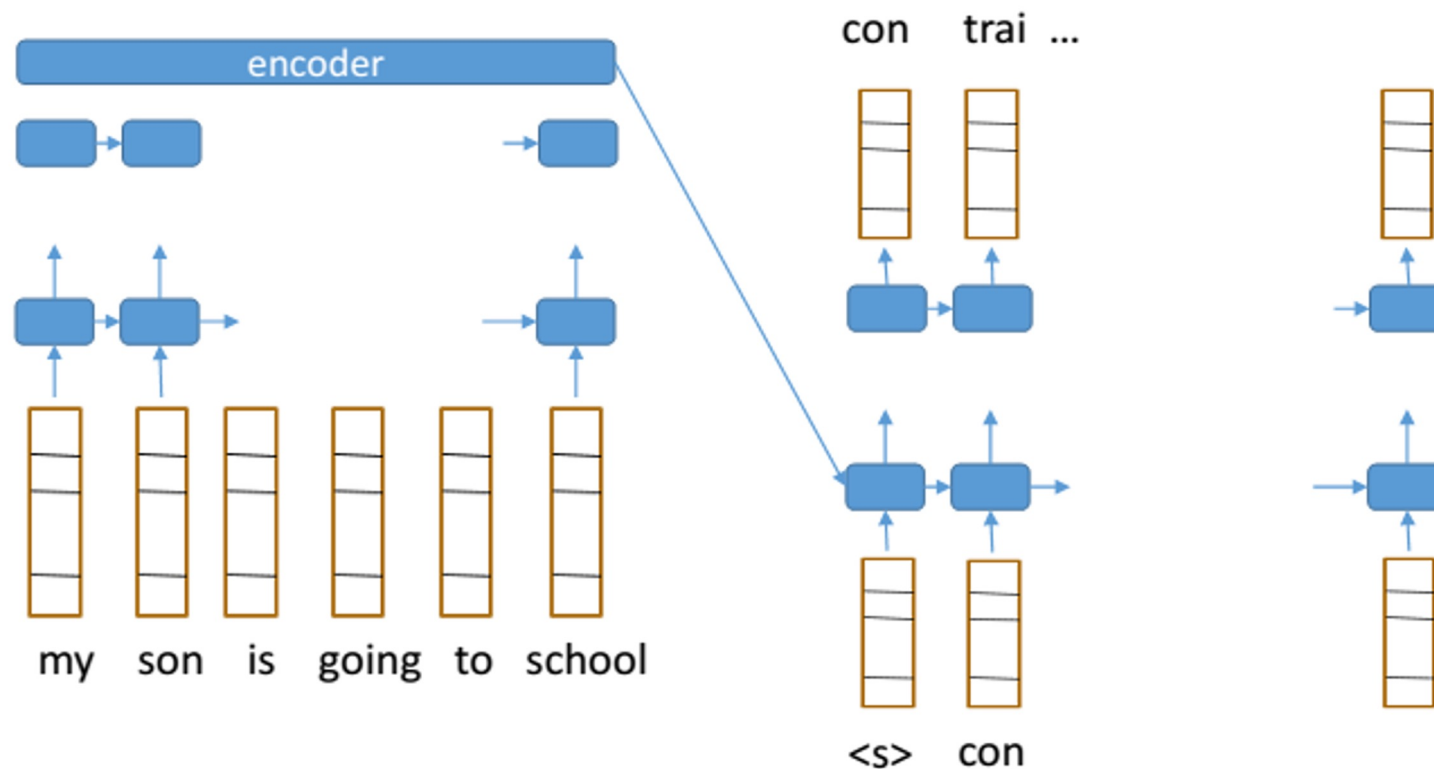
Giai đoạn 2: Statistical Machine Translation (SMT)



Lược sử phát triển của NLP

Ví dụ về dịch máy (Machine Translation)

Giai đoạn 3: Neural Machine Translation (NMT)



Các hệ thống dựa trên luật

- Dựa vào các luật thủ công, hướng miền ứng dụng
- Thường được sử dụng trong các bài toán đơn giản như trích rút dữ liệu có cấu trúc (vd, tên, ngày tháng,...) từ các dữ liệu phi cấu trúc (webpages, emails).
- Do sự phức tạp của ngôn ngữ, các tiếp cận này không tổng quát, dễ bị sai khi gặp các tình huống mới.

Các hệ thống dựa trên học máy truyền thống

- Có thể giải quyết các bài toán khó hơn (như phân loại thư rác).
- Sử dụng tập ngữ liệu huấn luyện (đã được gán nhãn) và các đặc trưng (vd, túi từ, nhãn từ loại) để xây dựng các mô hình học
- Các mô hình này khám phá các mẫu từ dữ liệu huấn luyện để dự đoán các dữ liệu mới

Các hệ thống dựa trên học sâu

- Có tính tổng quát hơn
- Không cần các đặc trưng thủ công vì tự động sinh đặc trưng và mô hình hệ thống.
- Cần tập ngữ liệu huấn luyện lớn
- Thường cho kết quả tốt hơn, có thể giải quyết các bài toán khó hơn trong NLP (vd dịch máy)

Q&A

Thank you!