

**COLE.VN**  
connecting knowledge

**Chủ đề:**

**Tổng quan về xử lý ngôn ngữ tự nhiên**

# Mục đích buổi học

- Học viên tiếp cận được các khái niệm cơ bản trong xử lý ngôn ngữ tự nhiên
- Nắm được các nhóm bài toán trong lĩnh vực xử lý ngôn ngữ tự nhiên

# Nội dung chính

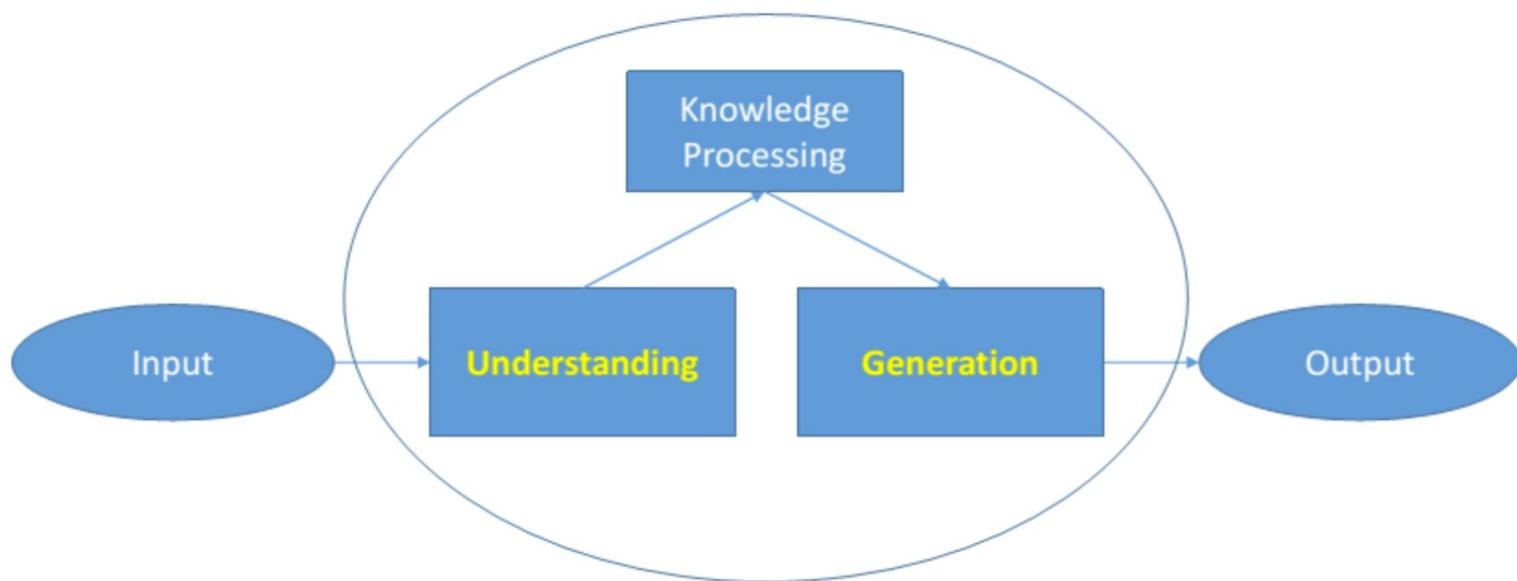
**NLP là gì?**

**Các bài toán cơ bản trong NLP**

# Giới thiệu về NLP

# Mục tiêu của NLP

NLP là lĩnh vực nghiên cứu và ứng dụng nhằm giúp máy tính có thể hiểu và sinh ngôn ngữ tự nhiên.



# Mục tiêu của NLP

Hiểu và sinh ngôn ngữ là các nhiệm vụ khó

Cái phòng của khách sạn tôi đặt cho ông chú hôm qua nó hơi nhỏ

Ông già đi nhanh quá

Sinh ra các văn bản có nghĩa như tiểu thuyết, trả lời câu hỏi,...

# Mục tiêu của NLP

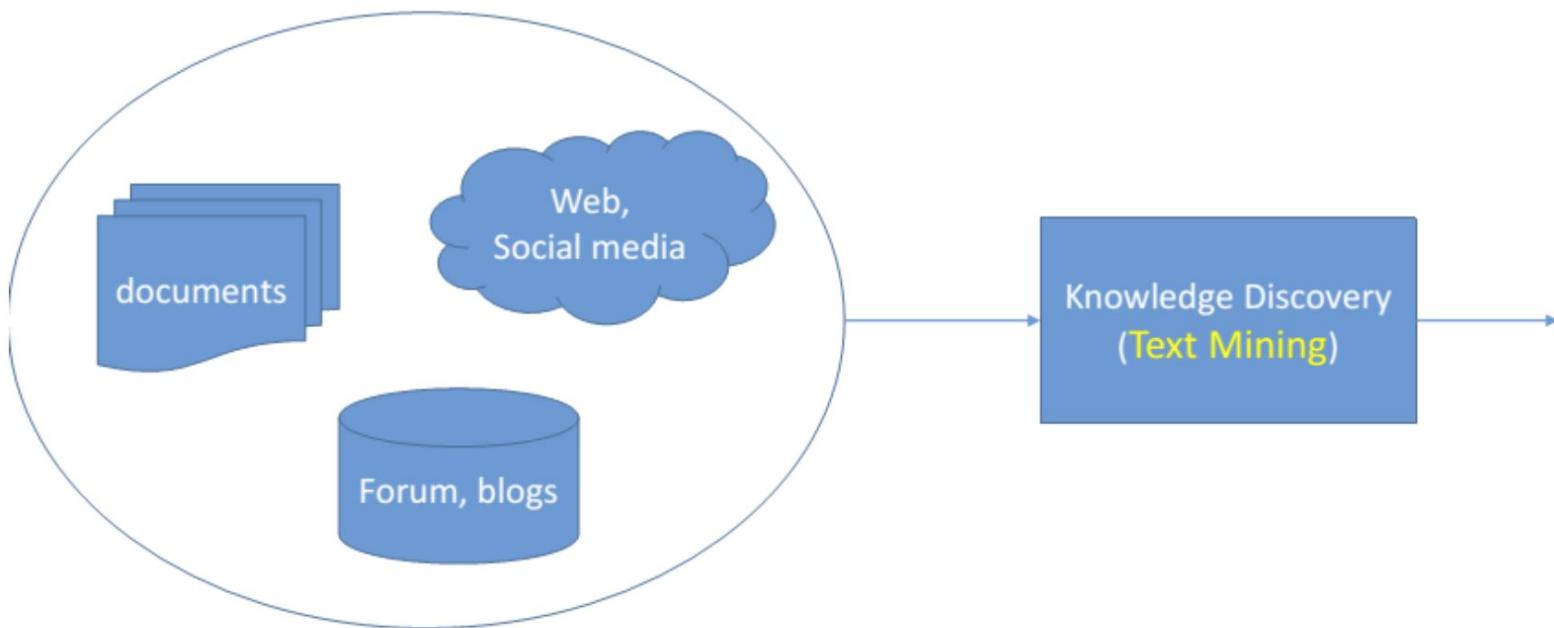
Tại sao xử lý ngôn ngữ tự nhiên khó?

- Nhập nhằng (ambiguity).
- Phụ thuộc ngữ cảnh.
- Phụ thuộc văn hóa, vùng miền.
- Phụ thuộc vào background knowledge.

# Mục tiêu của NLP

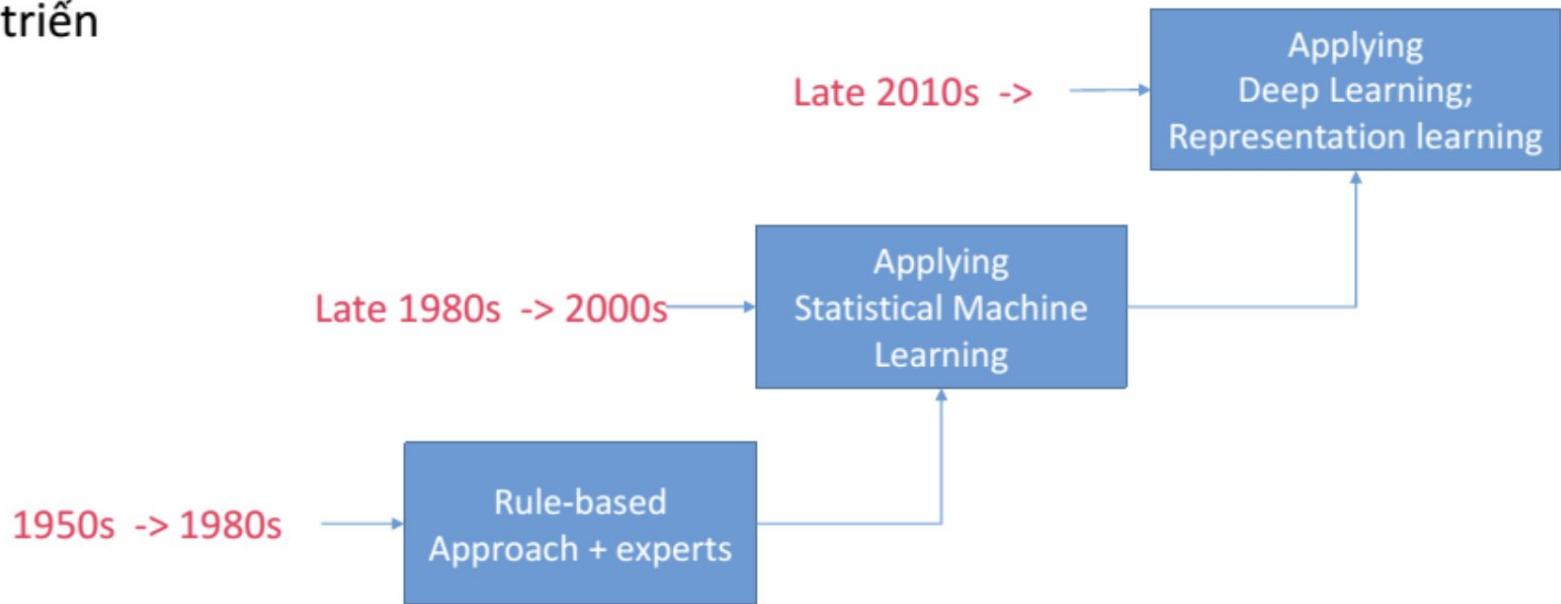
Tại sao xử lý ngôn ngữ tự nhiên khó?

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



# Lược sử phát triển của NLP

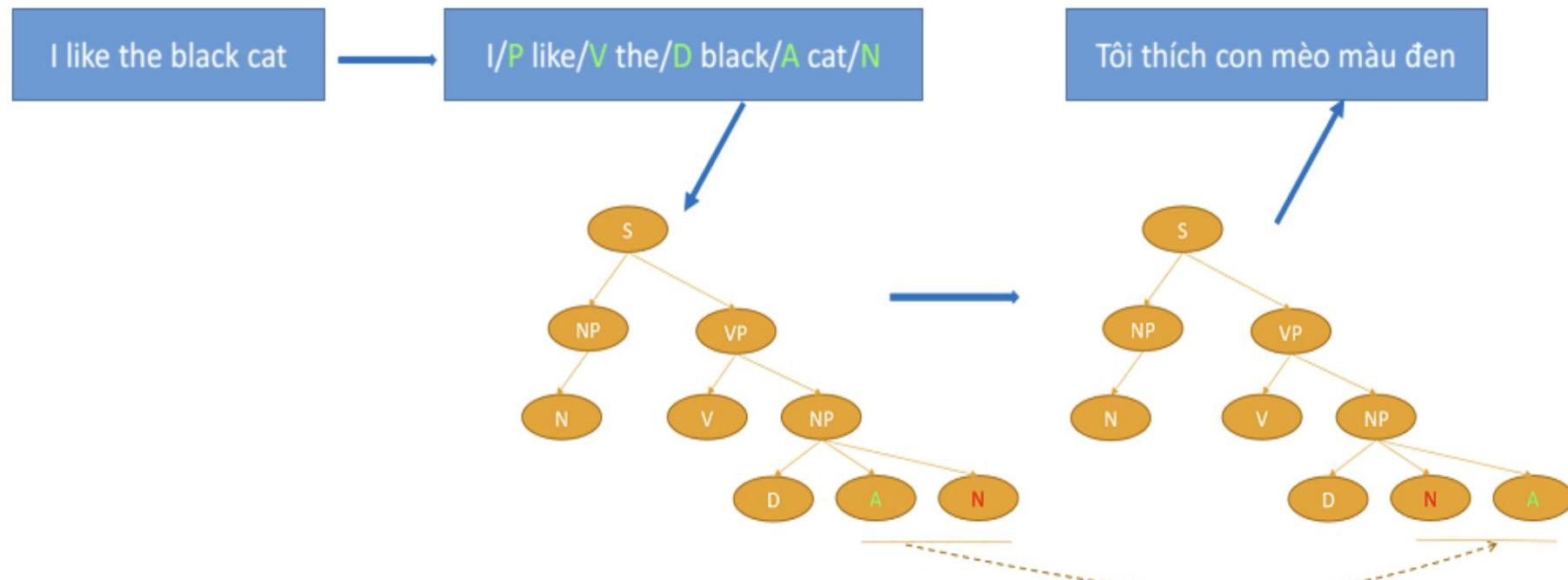
- Có 3 giai đoạn phát triển



# Lược sử phát triển của NLP

Ví dụ về dịch máy (Machine Translation)

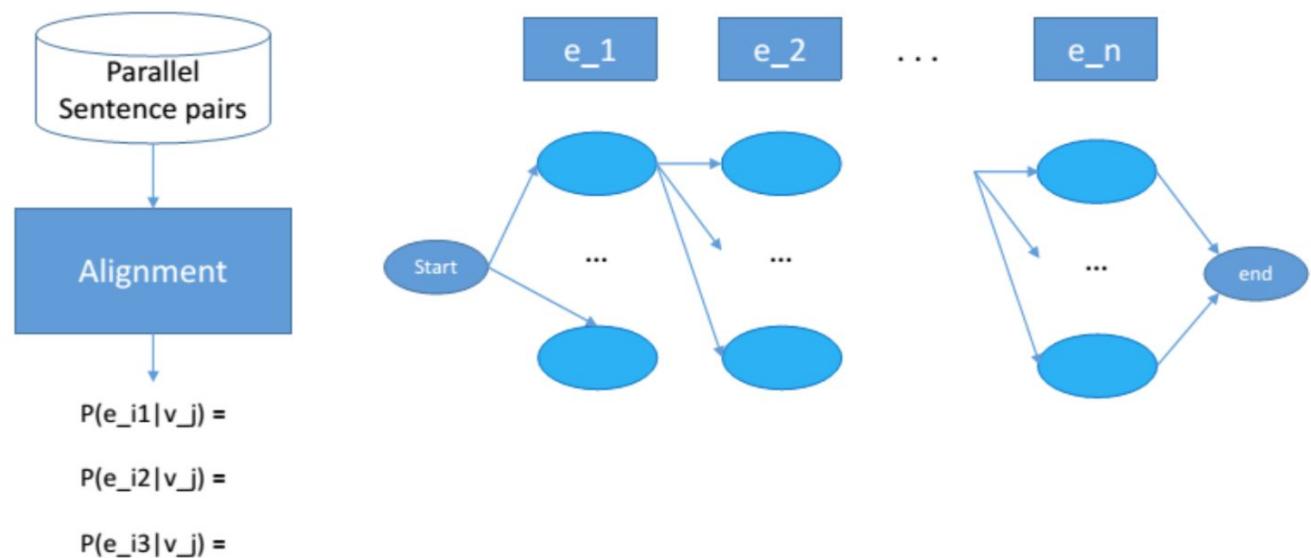
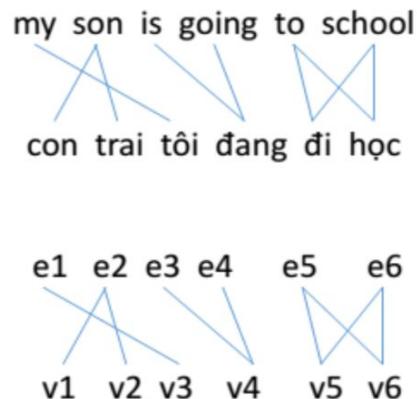
Giai đoạn 1: Rule Based Machine Translation (RBMT)



# Lược sử phát triển của NLP

Ví dụ về dịch máy (Machine Translation)

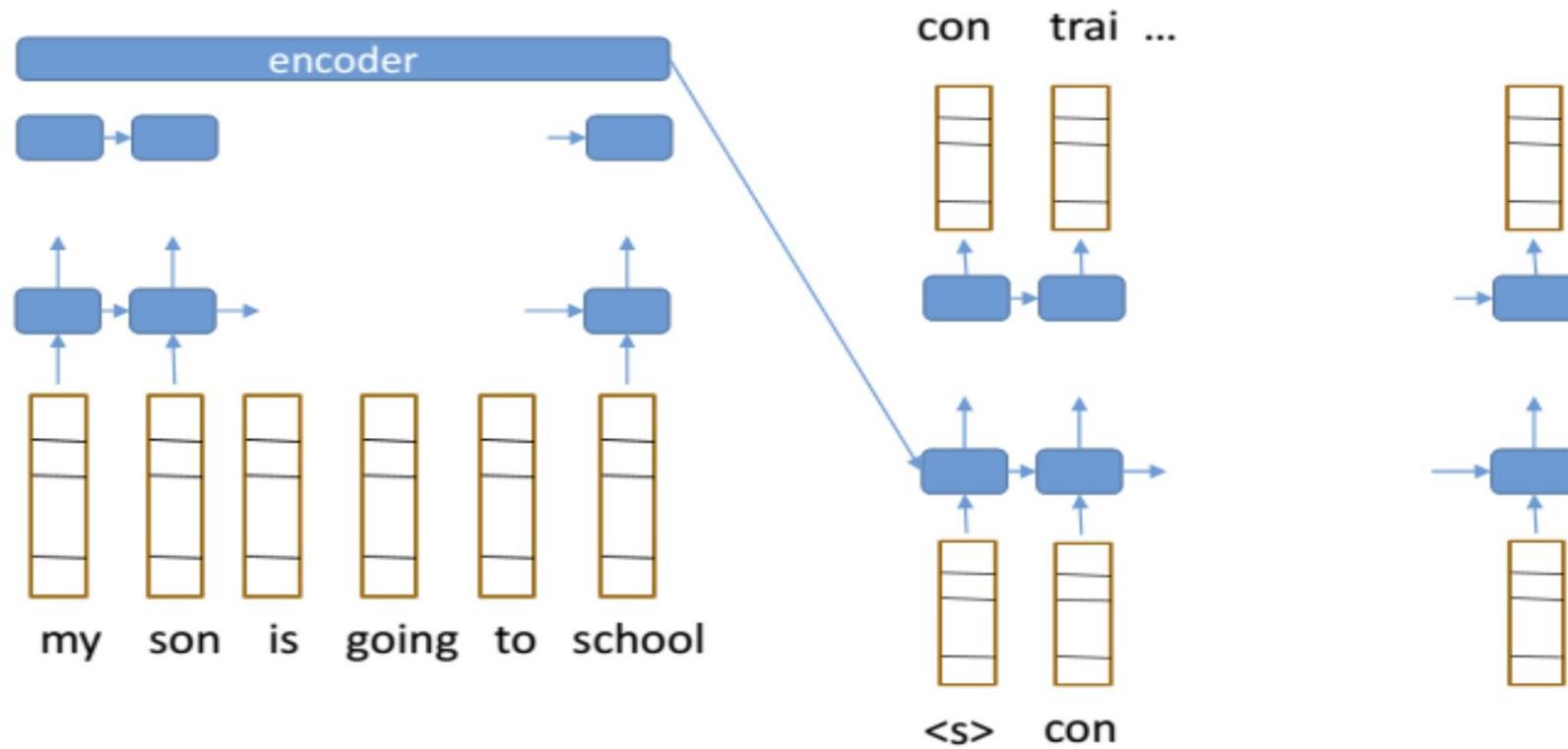
Giai đoạn 2: Statistical Machine Translation (SMT)



# Lược sử phát triển của NLP

Ví dụ về dịch máy (Machine Translation)

Giai đoạn 3: Neural Machine Translation (NMT)



# Các bài toán cơ bản của NLP

# Các mức xử lý ngôn ngữ

Việc xử lý và nghiên cứu ngôn ngữ có thể được thực hiện theo các mức:

- Từ vựng (Lexical).
- Cú pháp (Syntactic).
- Ngữ nghĩa (Semantic).
- Ngữ dụng (Pragmatic).

# Các bài toán cơ bản cho NLP

## Xử lý từ (Word Processing)

- Tokenizing.
- Word segmentation.

## 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 cho 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

# Các bài toán cơ bản cho NLP

Ví dụ về Core NLP problems:

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



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



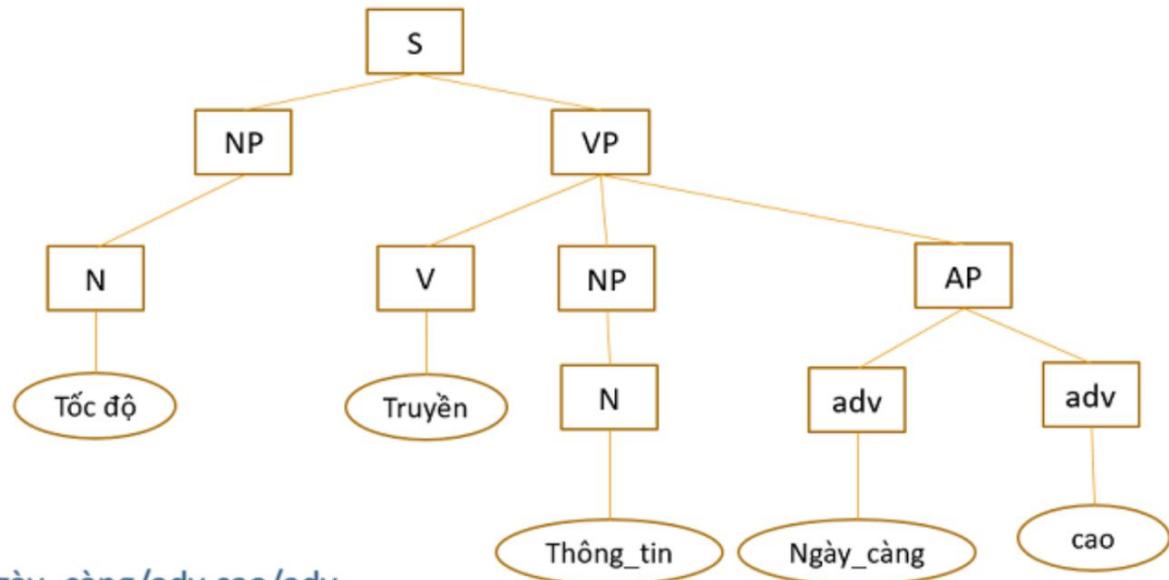
- Ông\_ấy nói : “ tốc\_độ truyền \_thông\_tin ngày\_càng cao ” .



- Ông\_ấy/N nói/V :/sym “/sym tốc\_độ/N truyền/V thông\_tin/N ngày\_càng/adv cao/Adv ” .

# Các bài toán cơ bản cho NLP

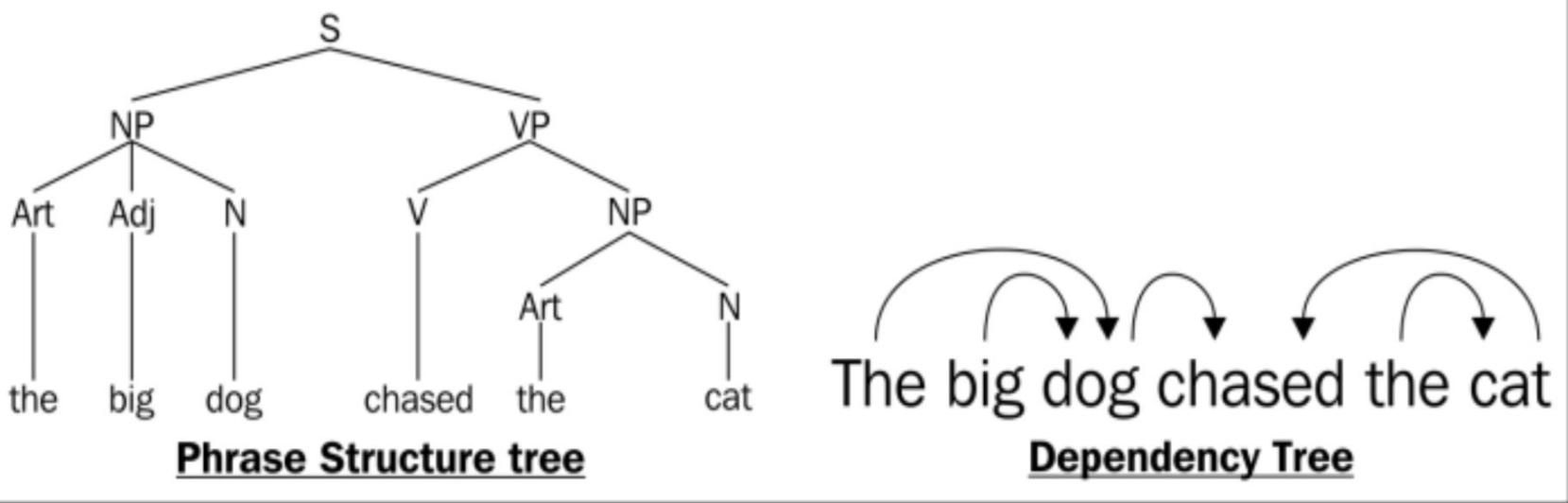
Ví dụ về Core NLP problems:



Tốc \_độ/N truyền/V thông \_tin/N ngày \_càng/adv cao/adv

# Các bài toán cơ bản cho NLP

Ví dụ về Core NLP problems:



# Các bài toán cơ bản cho NLP

Ví dụ về Core NLP problems:

## Word Sense Disambiguation

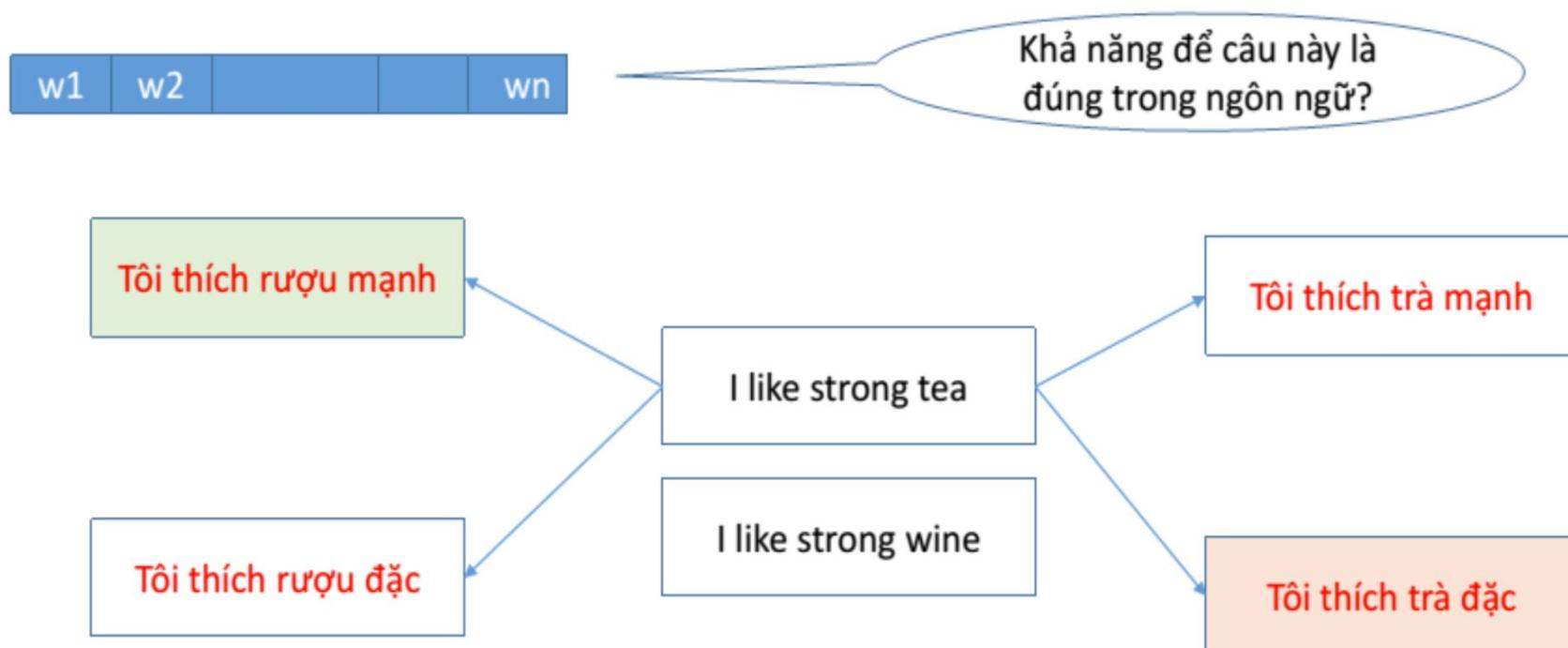
- Sau khi bị **bồ** **đá** nó đã trở thành một người khác hẳn
- Tôi để quên chìa khóa trong phòng mà giờ **nó** bị khóa cửa rồi.



## Co-reference

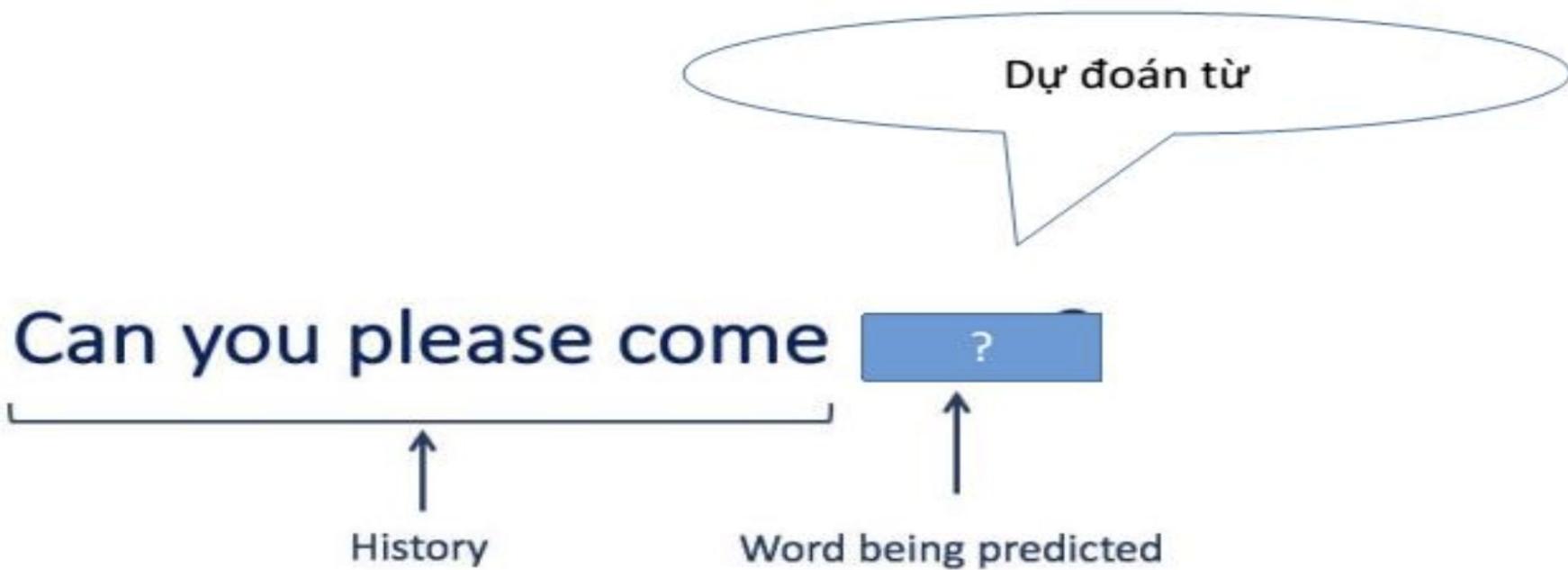
# Các bài toán cơ bản cho NLP

Ví dụ về Core NLP problems: Mô hình ngôn ngữ



# Các bài toán cơ bản cho NLP

Ví dụ về Core NLP problems: Mô hình ngôn ngữ



# Các bài toán cơ bản cho NLP

Ví dụ về Core NLP problems: Mô hình ngôn ngữ

5-gram



# Các bài toán cơ bản cho NLP

Ví dụ về Core NLP problems: Mô hình ngôn ngữ

Tôi thích khoa

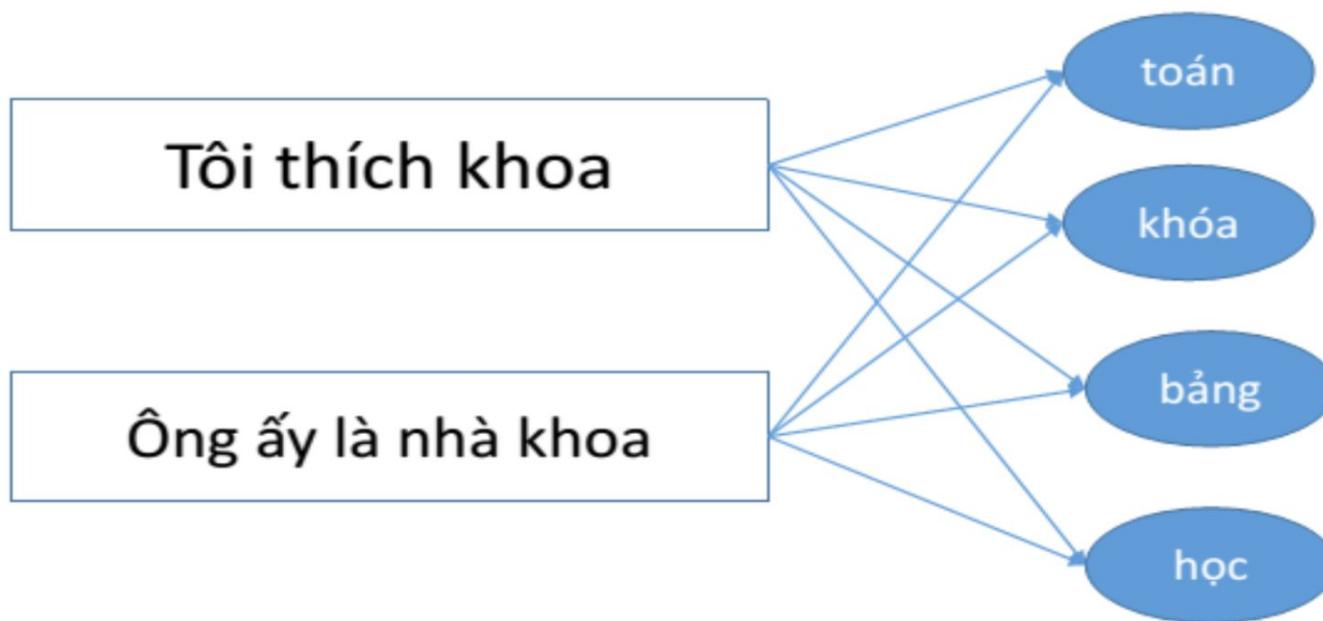
???

Ông ấy là nhà khoa

???

# Các bài toán cơ bản cho NLP

Ví dụ về Core NLP problems: Mô hình ngôn ngữ



# Các bài toán cơ bản cho NLP

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 PEN Tree banks).
- Core NLP (e.g Stanford Core NLP).
- Tools, Libraries for text and document processing/mining (NLTK, Spacy, ...).

# Các ứng dụng của NLP

# Các bài toán ứng dụng

## 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

# Các bài toán ứng dụng

- **Text Clustering.**
  - Topic Modeling.
- **Discourse Analysis.**

## Các ứng dụng: (ví dụ)

### Applications

- Spell checking
- Grammar checking
- Plagiarism checking

- Related Applications
- OCR (optical character recognition)
- Speech recognition
- Information Retrieval

# Các bài toán ứng dụng

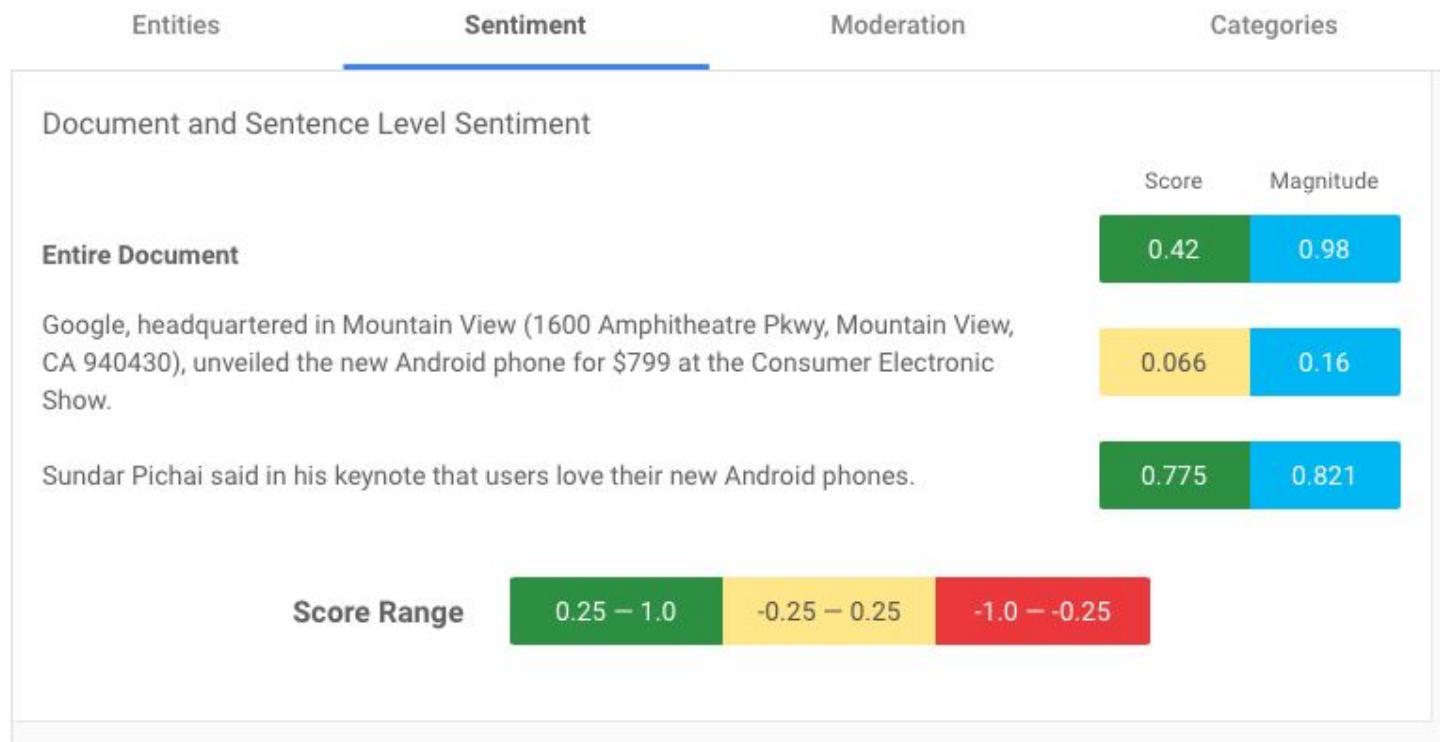
Ví dụ:

1. **Sentiment Analysis** (phân tích cảm xúc).
2. **Recommender Systems** (hệ tư vấn).
3. **Question Answering and Chatbot** (hỏi đáp và hội thoại tự động)
4. **Information Extraction** (trích chọn thông tin).
5. **Document Classification** (phân loại văn bản).
6. **Topic Modeling** (mô hình chủ đề).
7. **Information Retrieval** (tìm kiếm thông tin).
8. **OCR** (nhận dạng chữ in).

# Các bài toán ứng dụng

Ví dụ: Phân tích/Phân loại cảm xúc (Sentiment Analysis)

- Google Cloud (<https://cloud.google.com/natural-language>): for document and sentence level
- Document level



# Các bài toán ứng dụng

Ví dụ: Phân tích/Phân loại cảm xúc (Sentiment Analysis)

- Sentence level

Entities	Sentiment	Moderation	Categories
<Google> <sub>10</sub> , headquartered in <Mountain View> <sub>11</sub> ((<1600> <sub>2</sub> <1600 Amphitheatre Pkwy, Mountain View, CA> <sub>5</sub> <1600 Amphitheatre Pkwy> <sub>6</sub> , <Mountain View> <sub>11</sub> , <CA> <sub>8</sub> <940430> <sub>3</sub> ), unveiled the new <Android> <sub>7</sub> <phone> <sub>14</sub> for <\$799> <sub>1</sub> <799> <sub>4</sub> at the <Consumer Electronic Show> <sub>9</sub> . <Sundar Pichai> <sub>12</sub> said in his <keynote> <sub>13</sub> that <users> <sub>16</sub> love their new <Android> <sub>7</sub> <phones> <sub>15</sub> .			
1. \$799	PRICE	2. 1600	NUMBER
3. 940430	NUMBER	4. 799	NUMBER
5. 1600 Amphitheatre ...	ADDRESS	6. 1600 Amphitheatre ...	LOCATION
7. Android	CONSUMER GOOD	8. CA	LOCATION
9. Consumer Electroni...	EVENT	10. Google	ORGANIZATION

# Các bài toán ứng dụng

Ví dụ: Phân tích/Phân loại cảm xúc (Sentiment Analysis)

- 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 C

# Các bài toán ứng dụng

Ví dụ: **Hỏi đáp tự động** (Question/ Answering Systems)

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

# Các bài toán ứng dụng

Ví dụ: **Hỏi đáp tự động** (Question/ Answering Systems)  
IR-based question answering.

- IR: Information Retrieval (like Google, Bing)
- Input: question
- Output:  
related documents

**IR-based Question Answering**

Google Where is the Louvre Museum located?

Search About 904,000 results (0.00 seconds)

Everything Best guess for Louvre Museum Location is Paris, France  
Mentioned on at least 7 websites including wikipedia.org, answers.com and ask.com - Show sources - Feedback

Images

Maps

Videos

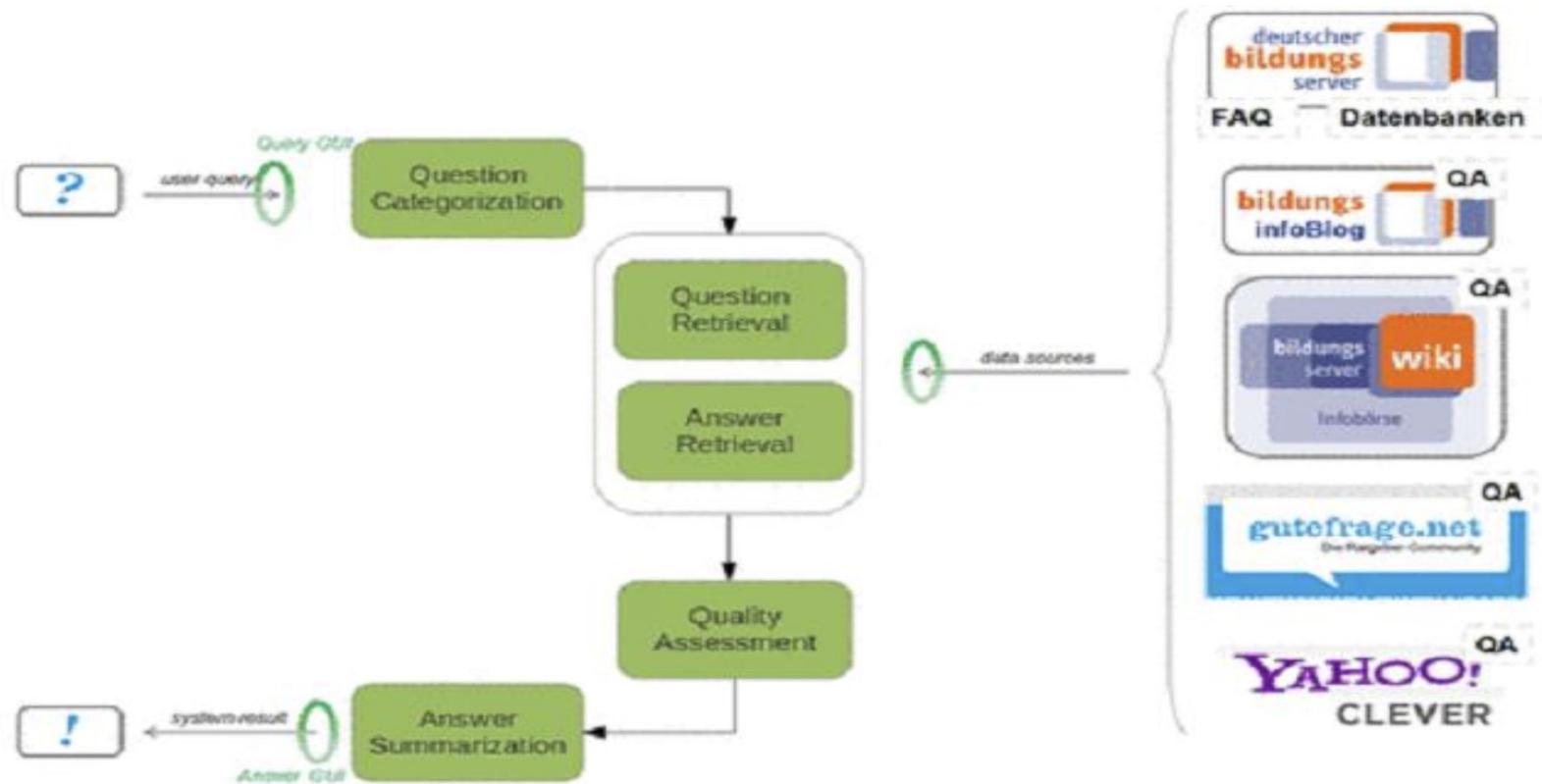
News

Musée du Louvre - Wikipedia, The free encyclopedia  
[en.wikipedia.org/wiki/Mus%C3%A9e\\_du\\_Louvre](https://en.wikipedia.org/wiki/Mus%C3%A9e_du_Louvre)  
Musée du Louvre is located in Paris. Location: within Paris. Established: 1793. Location: Palais Royal; Musée du Louvre, 75001 Paris, France. Type: Art museum ...  
Louvre Palace - List of works in the Louvre - Category:Musée du Louvre

# Các bài toán ứng dụng

Ví dụ: **Hỏi đáp tự động** (Question/ Answering Systems)

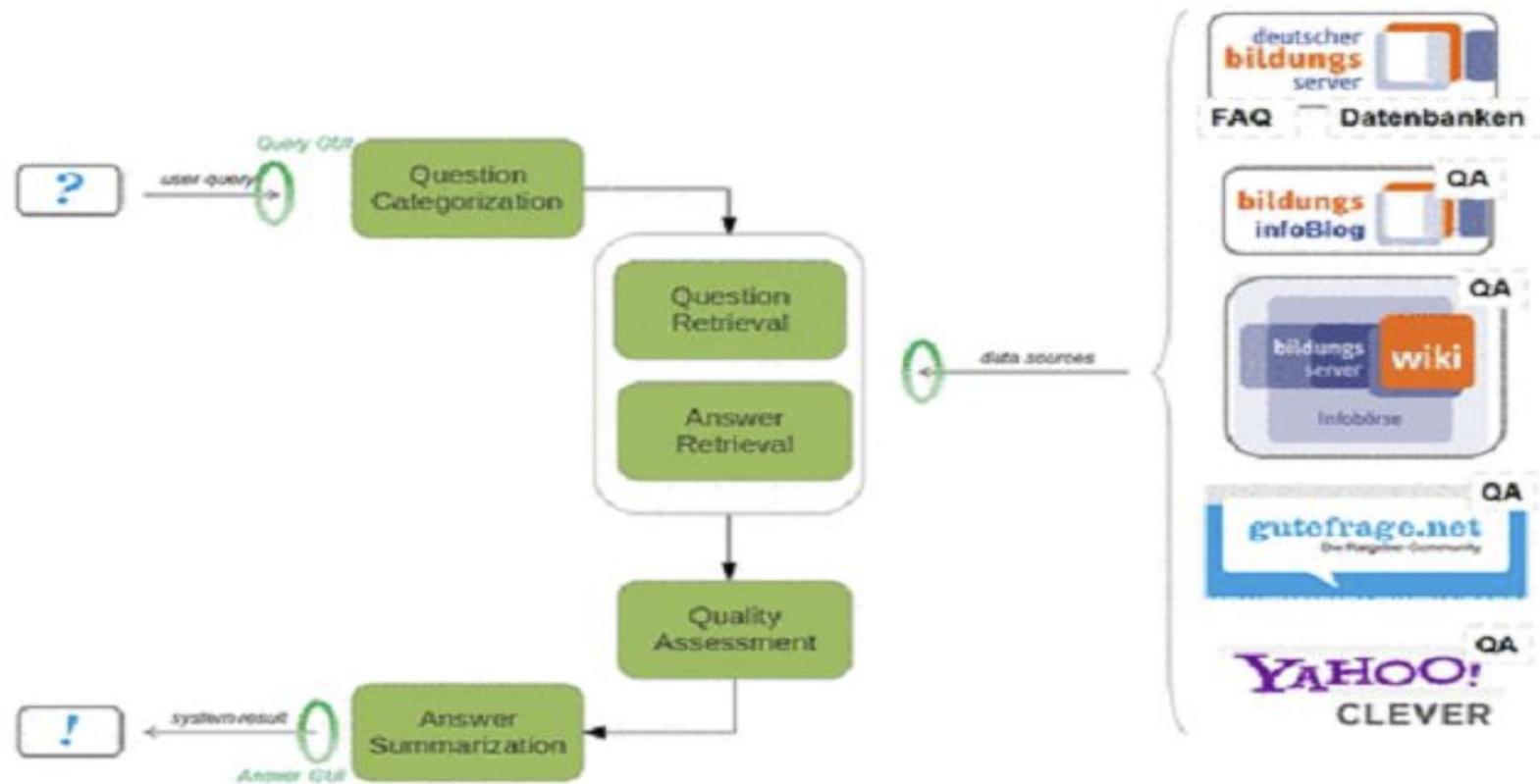
Community based question answering ([www.quora.com](http://www.quora.com))



# Các bài toán ứng dụng

Ví dụ: **Hỏi đáp tự động** (Question/ Answering Systems)

Community based question answering ([www.quora.com](http://www.quora.com))



# Các bài toán ứng dụng

Ví dụ: **Hội thoại, Chatbot, Personal Assistant**

(Alexa, Google Assistant, Google DialogFlow, MS Bot Framework).



# Các bài toán ứng dụng

## Ví dụ: Trích rút thông tin (Information Extraction - IE)

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

---

### TIE-UP-1:

Relationship:	TIE-UP
Entities:	"Bridgestone Sports Co." "a local concern" "a Japanese trading house"
Joint Venture Company:	"Bridgestone Sports Taiwan Co."
Activity:	ACTIVITY-1
Amount:	NT\$20000000

---

---

### ACTIVITY-1:

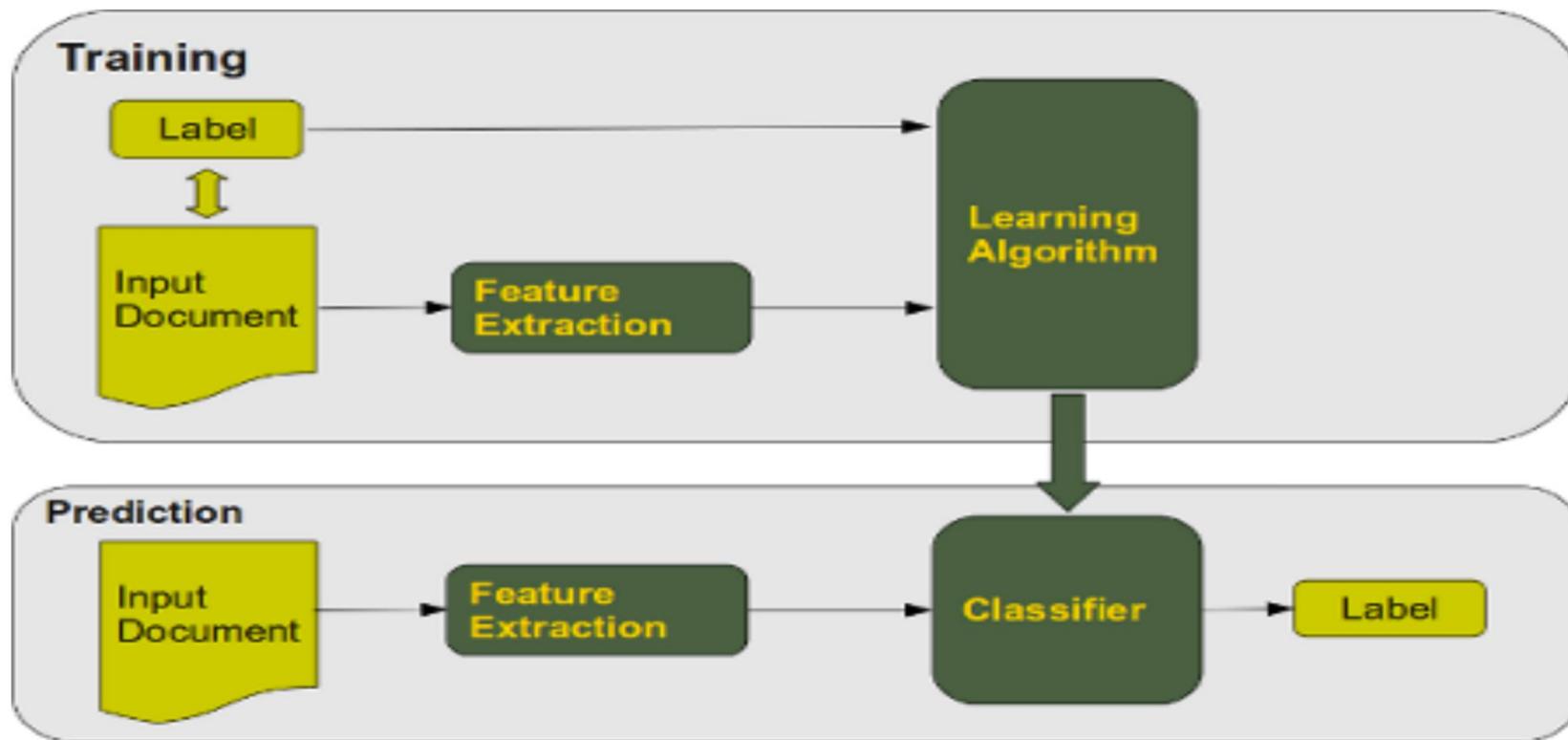
Activity:	PRODUCTION
Company:	"Bridgestone Sports Taiwan Co."
Product:	"iron and 'metal wood' clubs"
Start Date:	DURING: January 1990

---

# Các bài toán ứng dụng

Ví dụ: **Phân lớp văn bản** ((Text Classification)

(eg, phân loại văn bản theo chủ đề, phát hiện spam mail, ...)



# Các bài toán ứng dụng

Ví dụ: **Topic modeling** (Mô hình chủ đề)

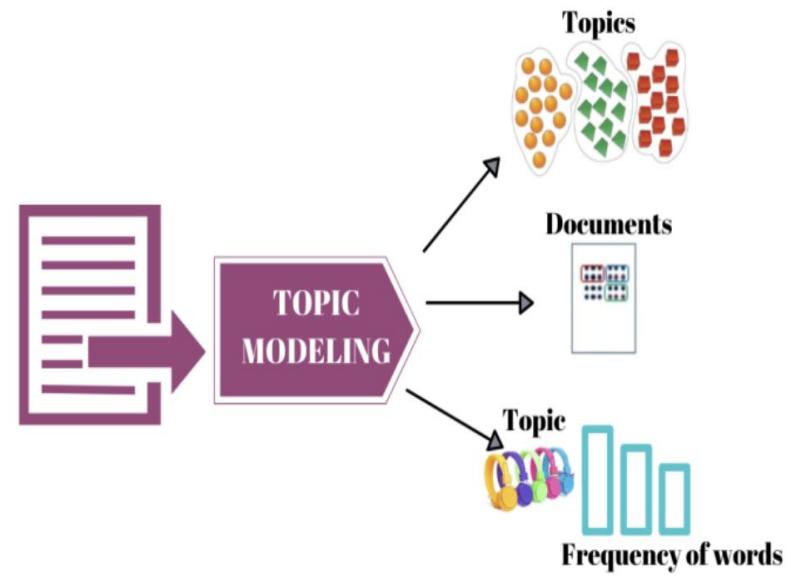
Bài toán đặt ra là đối với 1 tập các văn bản (documents) cần xác định được:

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 suất.

Các kỹ thuật phổ biến:

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



# Các bài toán ứng dụng

Ví dụ: **Nhận dạng chữ in** (Optical Character Recognition – OCR)

Image

5. Bảo trì công trình hạ tầng thu nhận dữ liệu ảnh viễn thám là tập hợp các

Tesseract-ocr

5. Bảo trì công trình hạ tầng thu nhận dữ liệu ảnh viễn thám là tập hợp các

Spelling using  
Language Model

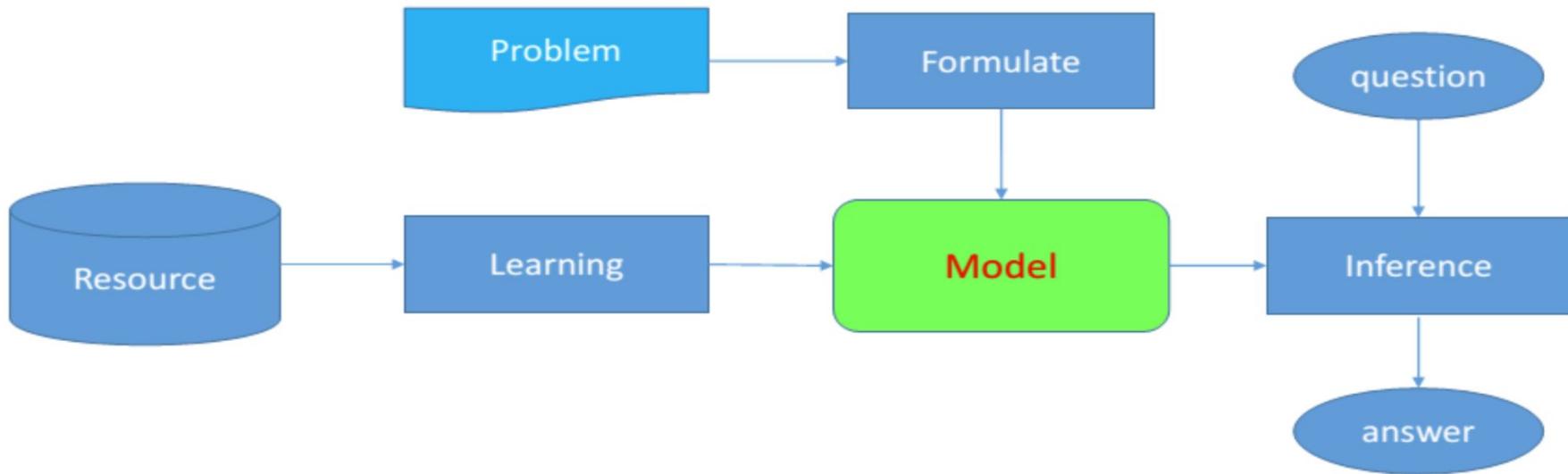
5. Bảo trì công trình hạ tầng thu nhận dữ liệu ảnh viễn thám là tập hợp các

# Mô hình Ngôn ngữ và Kỹ thuật xử lý

# Tiếp cận chung giải bài toán NLP

Các bước giải quyết bài toán NLP:

- Bước 1: Mô hình hóa bài toán (problem formulation).
- Bước 2: Xác định mô hình (ngôn ngữ- model formulation).
- Bước 3: Huấn luyện mô hình (Model training/learning).
- Bước 4: Suy diễn dựa trên mô hình (Model-based inference).

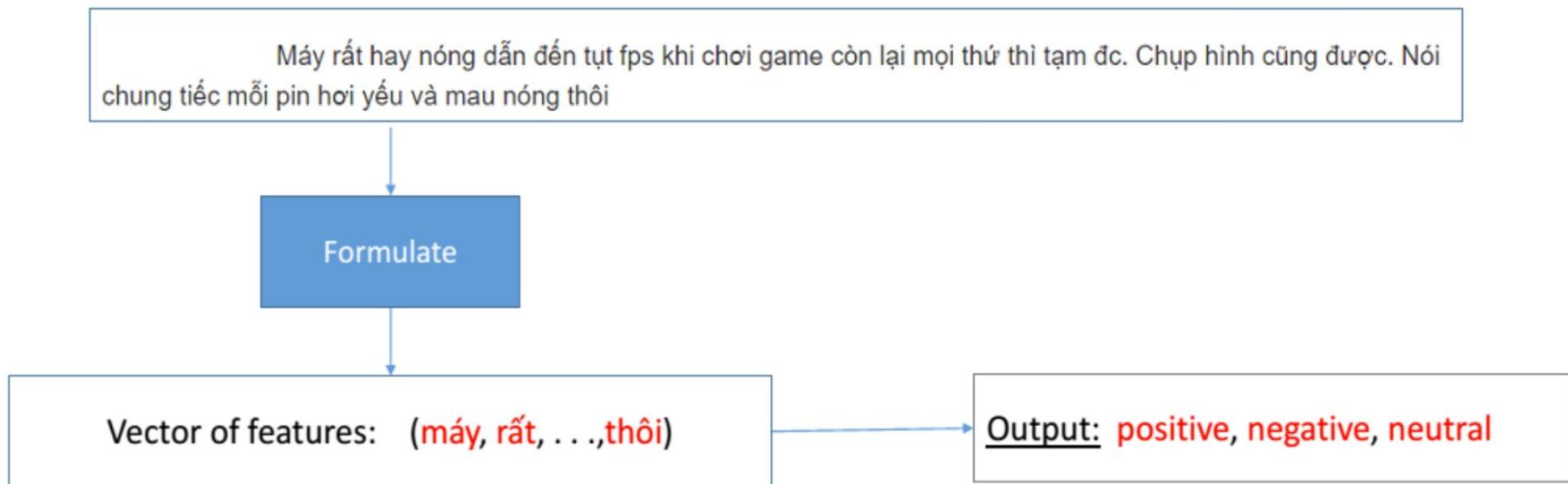


# Tiếp cận chung giải bài toán NLP

## Bài toán Phân lớp văn bản (Text Classification)

Là các bài toán học có giám sát (Supervised learning) trên texts.

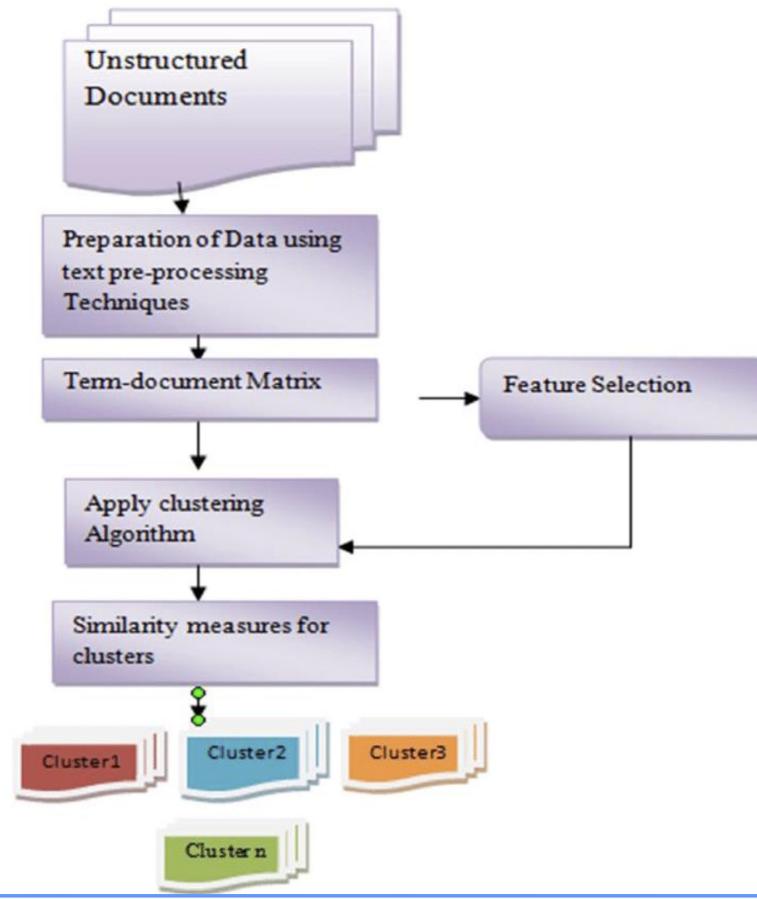
- Input: a textual document or a sentence



# Tiếp cận chung giải bài toán NLP

## Bài toán Phân cụm văn bản (Document Clustering)

Là các bài toán học có không giám sát (upervised learning) trên texts. Ví dụ: Topic modeling.



# Tiếp cận chung giải bài toán NLP

Bài toán gán nhãn (ngữ nghĩa) chuỗi (Semantic Sequence Labelling)

Ứng dụng: Name Entity, Information Extraction, Speech Recognition, ...

Thủ\_tướng Nguyễn\_Xuân\_Phúc đến thăm Trường Đại\_Học Quốc\_Gia Hà\_Nội  
B-PER I-PER O O B-ORG I-ORG I-ORG I-ORG

- **Information extraction problems:**

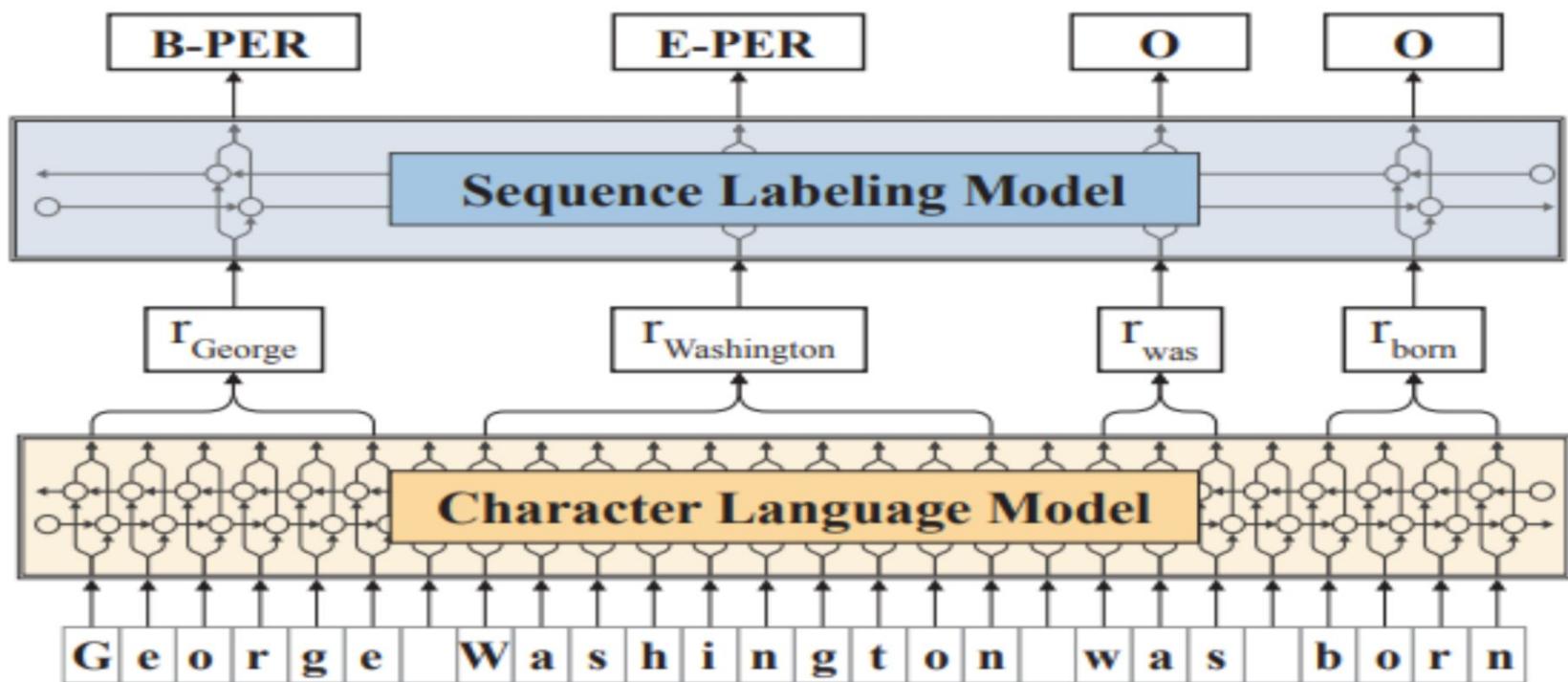
- Named Entity Recognition
- Meta Data Extraction

Số/Ký hiệu	35/2019/NĐ-CP
Ngày ban hành	25/04/2019
Ngày có hiệu lực	10/06/2019
Người ký	Nguyễn Xuân Phúc
Trích yếu	Quy định xử phạt vi phạm hành chính trong lĩnh vực Lâm Nghiệp
Cơ quan ban hành	Chính phủ
Phân loại	Nghị định

# Tiếp cận chung giải bài toán NLP

Bài toán gán nhãn (ngữ nghĩa) chuỗi (Semantic Sequence Labelling)

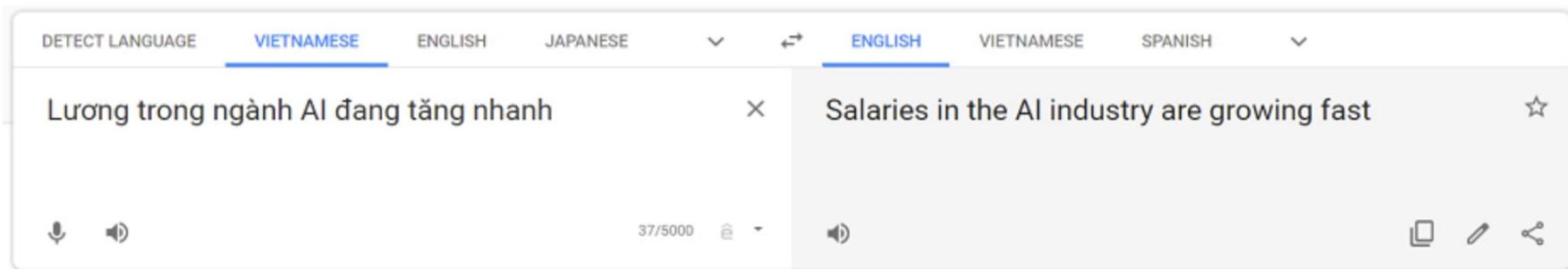
Ứng dụng: Name Entity, Information Extraction, Speech Recognition, ...



# Tiếp cận chung giải bài toán NLP

## Bài toán sinh văn bản (Text Generation)

Ứng dụng: Machine translation, news generation, chatbot, QA systems...



The screenshot shows a machine translation interface. At the top, there are language selection tabs: DETECT LANGUAGE, VIETNAMESE (underlined in blue), ENGLISH, JAPANESE, and a double-headed arrow icon. Below these are two text boxes. The left text box contains the Vietnamese sentence "Lương trong ngành AI đang tăng nhanh". The right text box contains the English translation "Salaries in the AI industry are growing fast". Below the text boxes are various icons for audio playback, file download, and sharing.

**Duplex AI:** "Hi. I'm calling to book a woman's haircut for a client. Um, I'm looking for something on May 3?"

**Human receptionist:** "Sure. Give me onnne second...."

**AI:** "Mm-hmm."

**Human:** "Sure, what time are you looking for, around?"

**AI:** "At 12 P.M."

**Human:** "We don't have 12 available. The closest we have to that is a 1:15."

**AI:** "Do you have anything between 10 A.M. and, uh, 12 P.M.?"

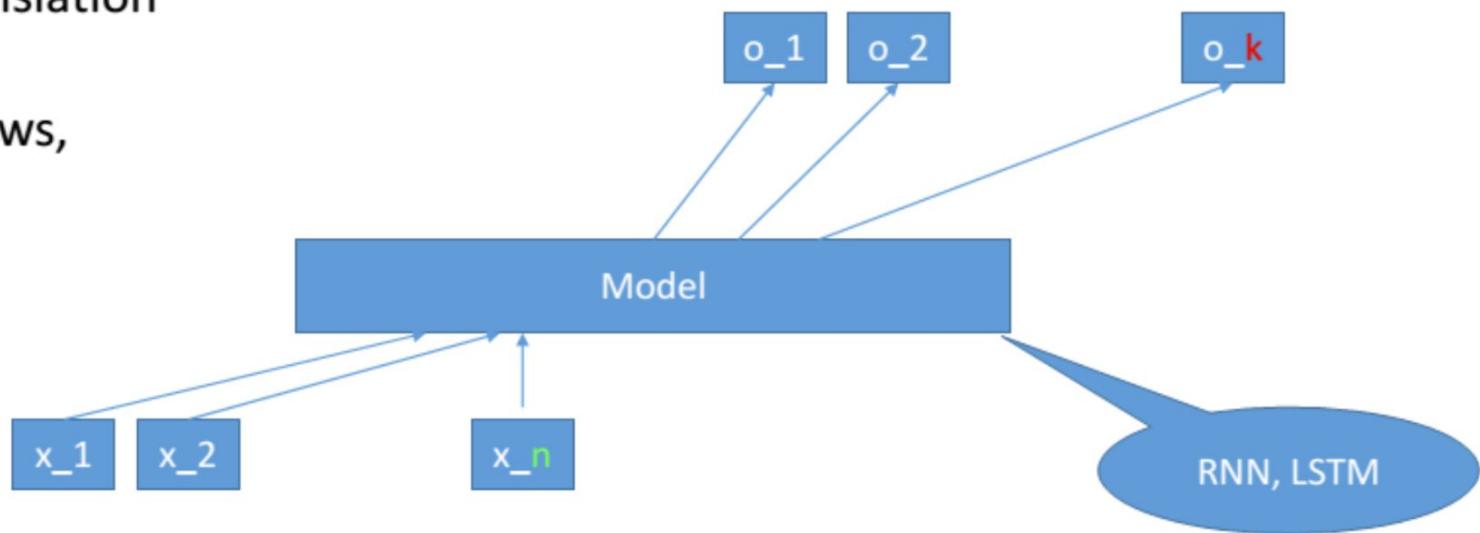
**Human:** "Okay, we have a 10 o'clock."

# Tiếp cận chung giải bài toán NLP

## Bài toán sinh văn bản (Text Generation)

Ứng dụng: Machine translation, news generation, chatbot, QA systems...

- Machine Translation
- Chatbot
- AI writing (news, advertisement, comments)



# Mô hình ngôn ngữ và kỹ thuật

## Knowledge/Rule based and Heuristics

Sử dụng trong giai đoạn đầu (50-90) với symbolic AI

- IF Condition
- THEN Action

**Given input:** “that”

**If**

(+1 A/ADV/QUANT)  
(+2 SENT-LIM)  
(NOT -1 SVOC/A)

**Then** eliminate non-ADV tags

**Else** eliminate ADV

# Mô hình ngôn ngữ và kỹ thuật

## Knowledge/Rule based and Heuristics

Sử dụng trong giai đoạn đầu (50-90) với symbolic AI

### RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger

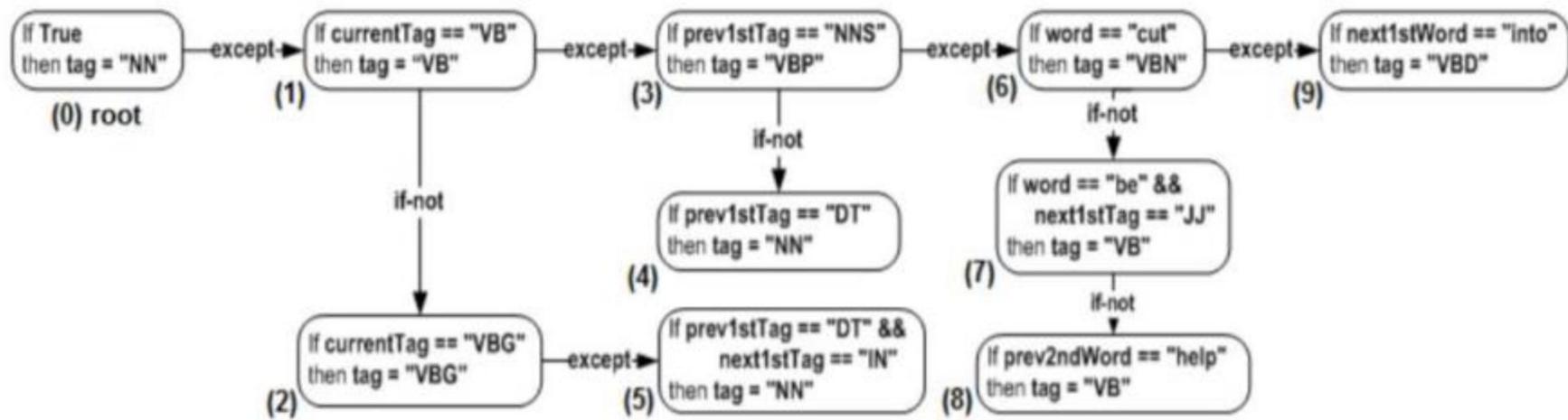


Figure 1: A part of our SCRDR tree for English POS tagging.

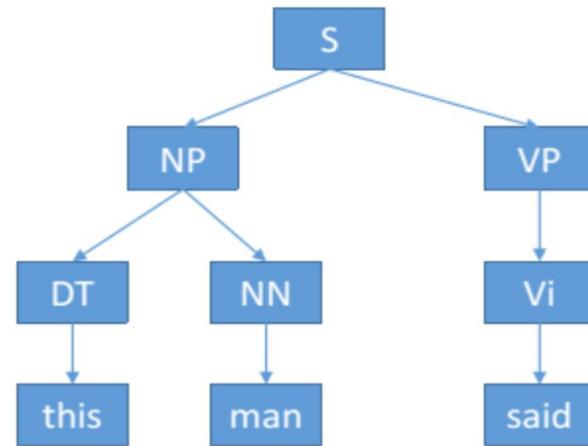
# Mô hình ngôn ngữ và kỹ thuật

## Knowledge/Rule based and Heuristics

Sử dụng trong giai đoạn đầu (50-90) với symbolic AI

## Rules and Knowledge Bases

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	P	NP



# Mô hình ngôn ngữ và kỹ thuật

## Knowledge/Rule based and Heuristics

Sử dụng trong giai đoạn đầu (50-90) với symbolic AI

- Sử dụng luật để biểu diễn thông tin ngôn ngữ (using template or rule for representing linguistics information)
- Sử dụng tri thức chuyên gia (using experts' knowledge)
- Đưa ra quyết định cứng (derive Hard decision)

# Mô hình ngôn ngữ và kỹ thuật

Knowledge/Rule based and Heuristics

Sử dụng trong giai đoạn đầu (50-90) với symbolic AI

## Hạn chế của tiếp cận Rule

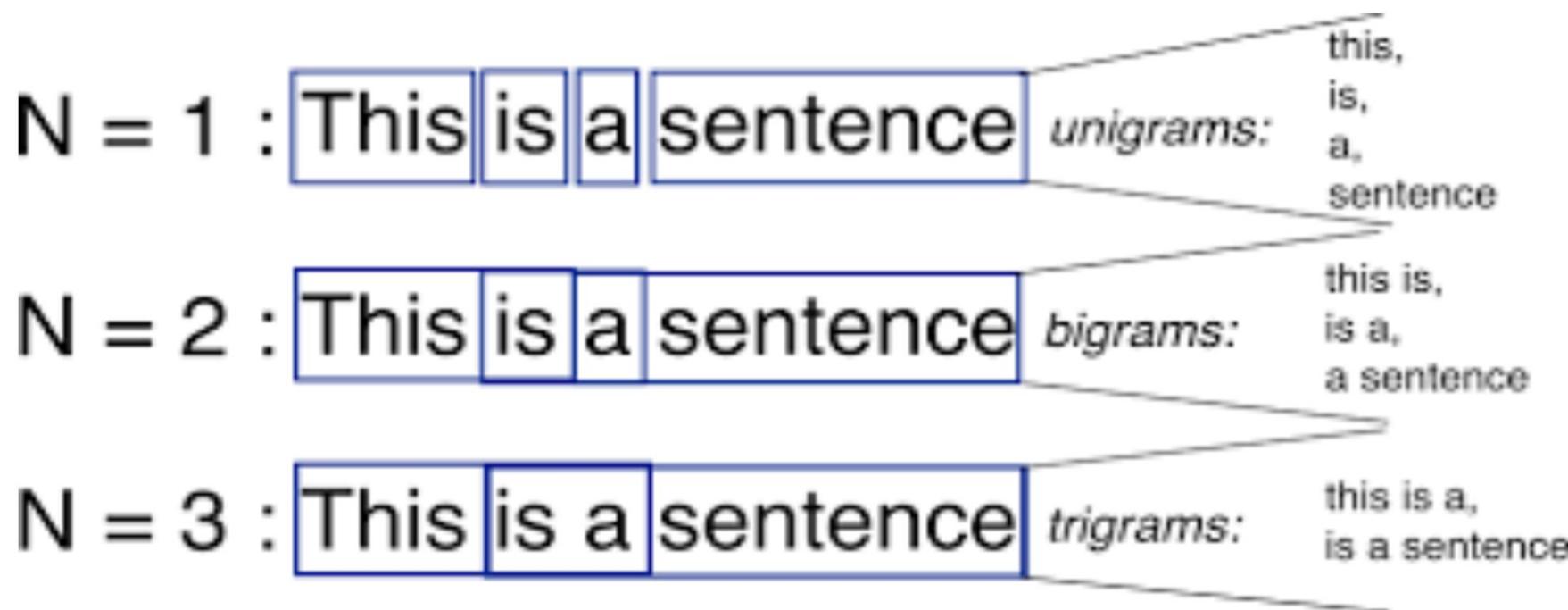
- Many ambiguities: nhiều nhập nhằng
- Many conflict of Rules: luật mâu thuẫn
- Scope of rule coverage: luật chỉ bao phủ một phạm vi nhất định
- How to trust the rules: độ tin tưởng của luật?

# Mô hình ngôn ngữ và kỹ thuật

Mô hình ngôn ngữ thống kê (Statistical Language Models)

Từ đầu những năm 90 và phổ biến cho đến tận gần đây

Mô hình n-grams



# Mô hình ngôn ngữ và kỹ thuật

Mô hình ngôn ngữ thống kê (Statistical Language Models)

Từ đầu những năm 90 và phổ biến cho đến tận gần đây

Văn phạm xác suất

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	P	NP



S	$\Rightarrow$	NP	VP	1.0
VP	$\Rightarrow$	Vi		0.4
VP	$\Rightarrow$	Vt	NP	0.4
VP	$\Rightarrow$	VP	PP	0.2
NP	$\Rightarrow$	DT	NN	0.3
NP	$\Rightarrow$	NP	PP	0.7
PP	$\Rightarrow$	P	NP	1.0

# Mô hình ngôn ngữ và kỹ thuật

Mô hình ngôn ngữ thống kê (Statistical Language Models)

Từ đầu những năm 90 và phổ biến cho đến tận gần đây

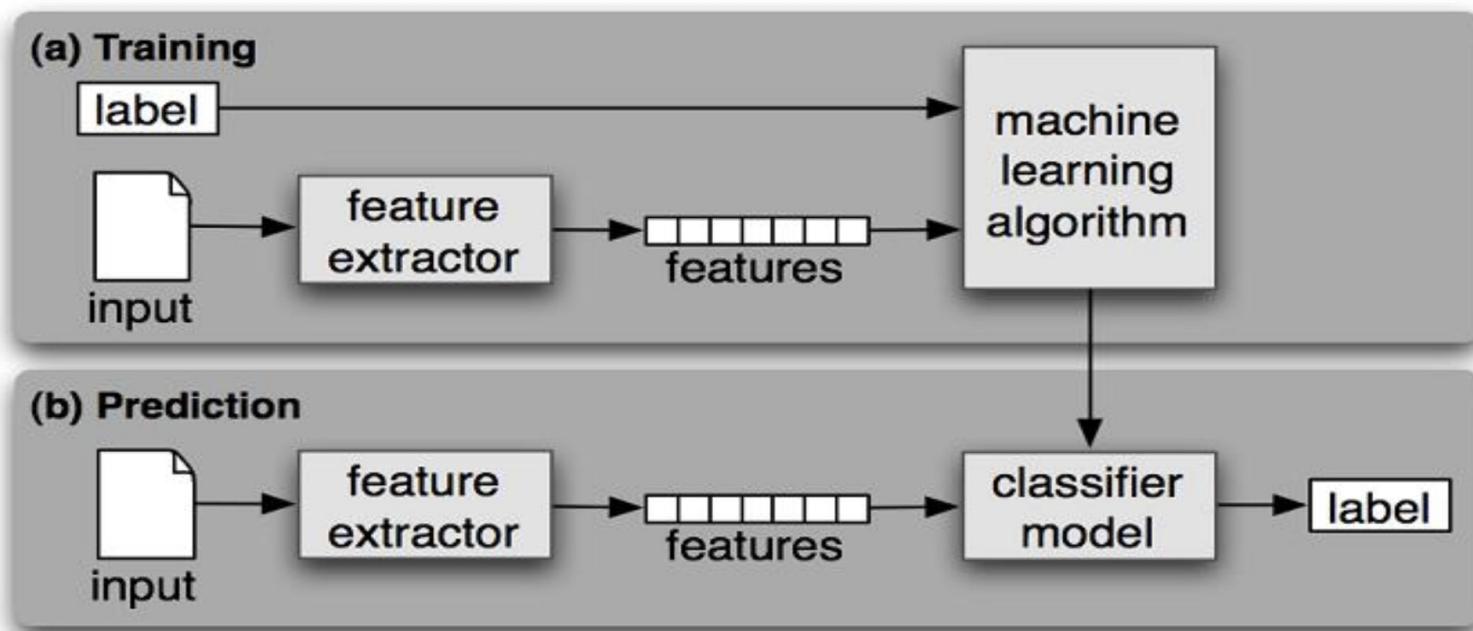
- Coi NLP như một ứng dụng của học máy (Consider NLP problems as statistical machine learning problems)
- Giải quyết nhập nhằng (Solving limitations of hard decision)
- Học từ dữ liệu (Learning knowledge from data (i.e. data-driven) not from experts)

# Mô hình ngôn ngữ và kỹ thuật

Mô hình ngôn ngữ thống kê (Statistical Language Models)

Từ đầu những năm 90 và phổ biến cho đến tận gần đây

Bài toán phân lớp văn bản



# Mô hình ngôn ngữ và kỹ thuật

Mô hình ngôn ngữ thống kê (Statistical Language Models)

Từ đầu những năm 90 và phổ biến cho đến tận gần đây

Bài toán phân cụm văn bản

Input: Biểu diễn của văn bản dưới dạng vector các đặc trưng.

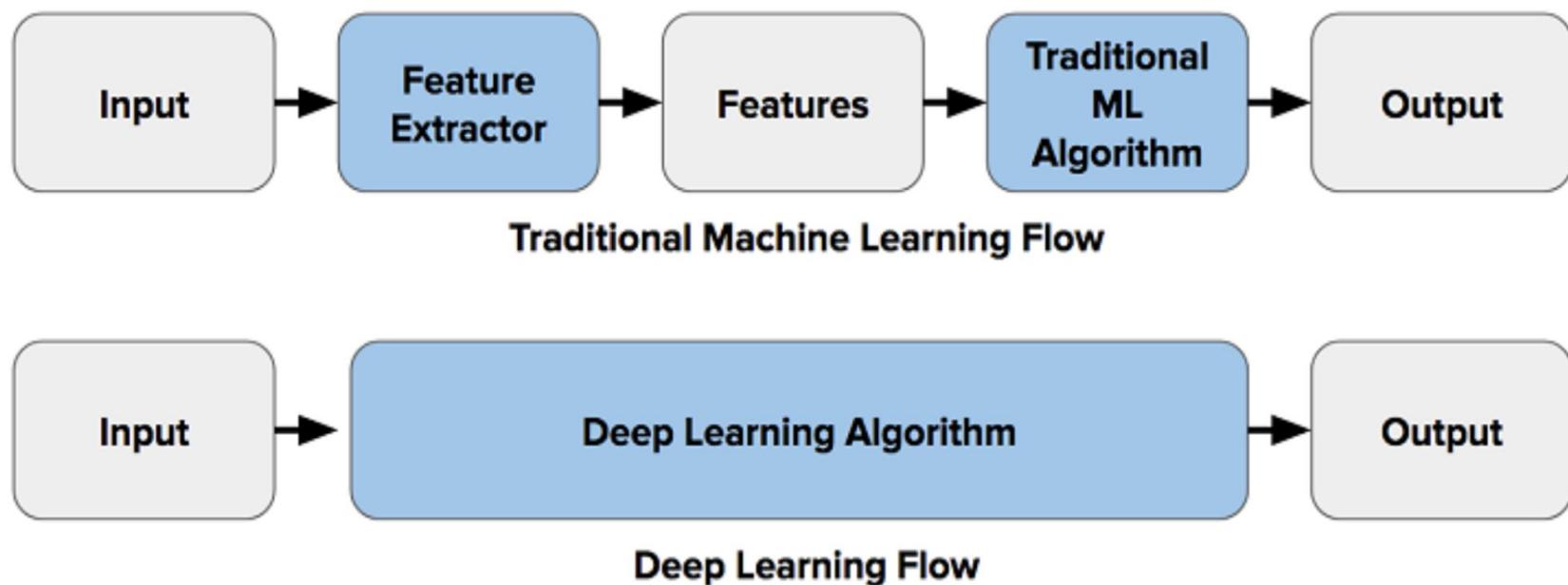
Output: Các cụm văn bản

Các phương pháp: K-means, K-median, Spectral clustering, Hierarchical clustering, matrix factorization, ...

# Mô hình ngôn ngữ và kỹ thuật

Mô hình ngôn ngữ Neural (Neural Language Models)

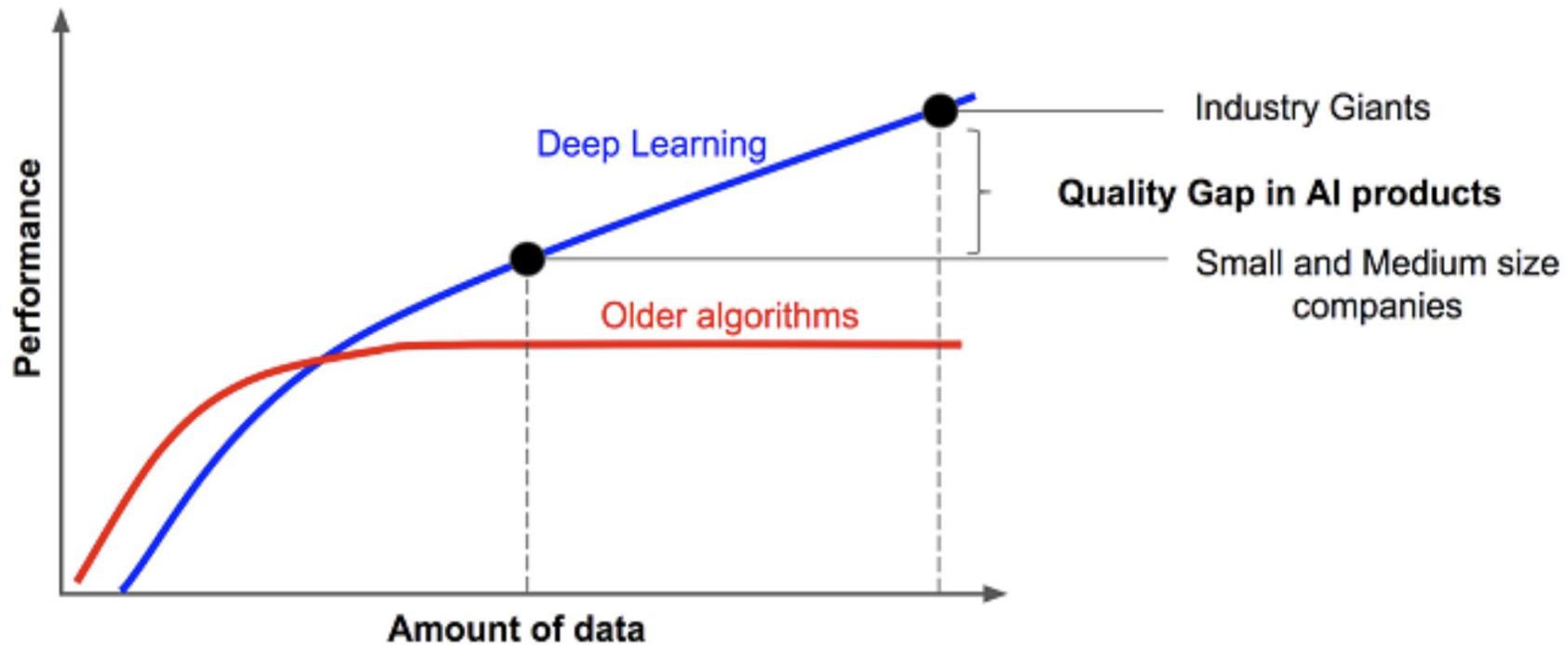
Deep Learning vs. Traditional Machine Learning



# Mô hình ngôn ngữ và kỹ thuật

Mô hình ngôn ngữ Neural (Neural Language Models)

Deep Learning vs. Traditional Machine Learning



# Mô hình ngôn ngữ và kỹ thuật

Mô hình ngôn ngữ Neural (Neural Language Models)

Deep Learning vs. Traditional Machine Learning

## Biểu diễn dữ liệu

- Input is a text, for example:

“I will return to this restaurant because of this food.”

- Represented as a vector of words:

(I, will, return, to, this, restaurant, because, of, this, food, .)

- Put it into the model

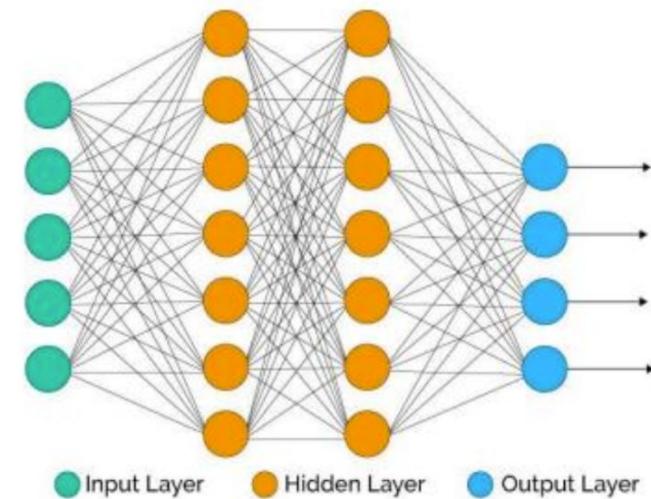
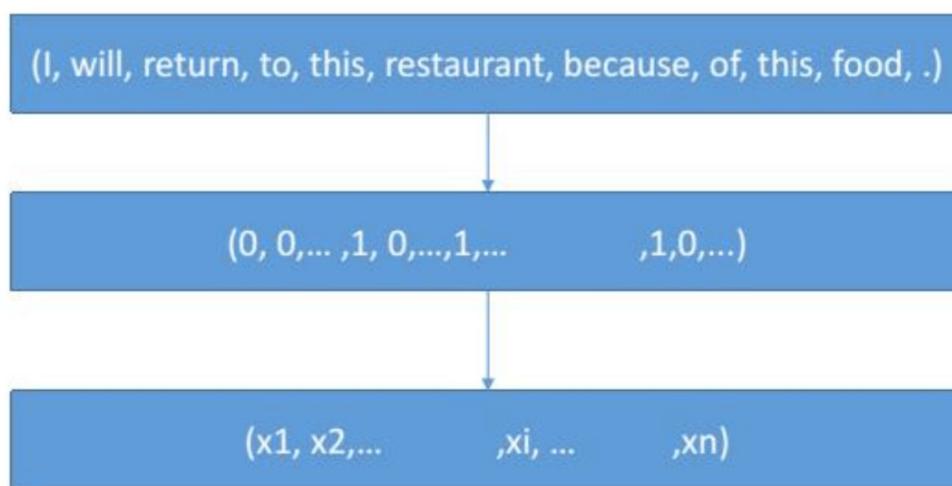
Model

# Mô hình ngôn ngữ và kỹ thuật

## Mô hình ngôn ngữ Neural (Neural Language Models)

### Deep Learning vs. Traditional Machine Learning

- There is a problem:



- Size of the vector is very large: vocabulary size, n is about ~50k -> 100k

# Mô hình ngôn ngữ và kỹ thuật

## Mô hình ngôn ngữ Neural (Neural Language Models)

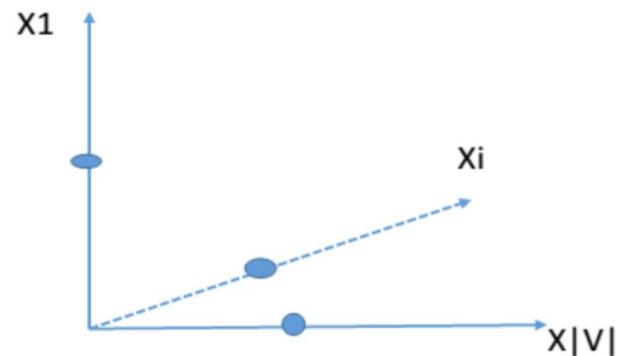
### Word Representation (Biểu diễn từ)

➤ The words: “bố”, “mẹ”, “con cái”,  
“bàn”, “ghé”, “sách”, “vở”

have no relationships in one-hot  
vector representations.

➤ The words are placed in a space of

Rome      Paris  
Rome = [1, 0, 0, 0, 0, 0, ..., 0]  
Paris = [0, 1, 0, 0, 0, 0, ..., 0]  
Italy = [0, 0, 1, 0, 0, 0, ..., 0]  
France = [0, 0, 0, 1, 0, 0, ..., 0]



# Mô hình ngôn ngữ và kỹ thuật

## Mô hình ngôn ngữ Neural (Neural Language Models)

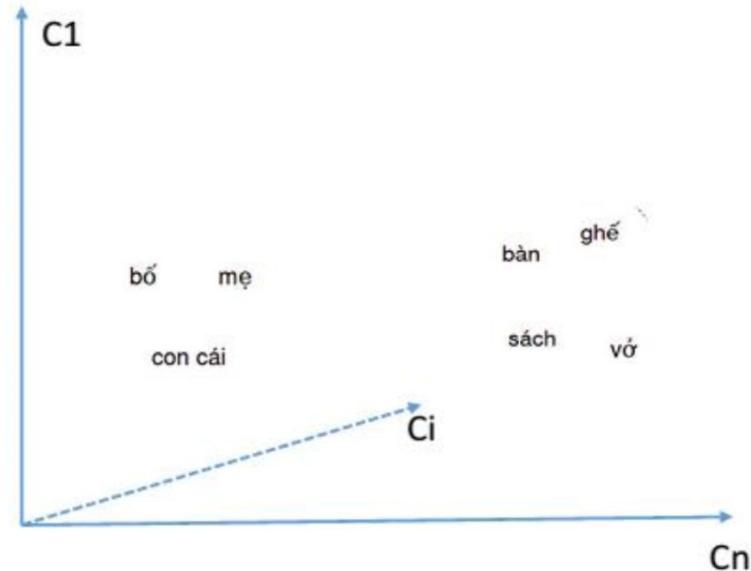
### Word Representation (Biểu diễn từ)

- The words: “bố”, “mẹ”, “con cái”, “bàn”, “ghé”, “sách”, “võ”

Embed the words into a space of abstract concepts

- The words are placed in a space of  $n \ll |V|$  dimensions.

Word vectors	Dimensions			
	animal	domesticated	pet	fluffy
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37



# Mô hình ngôn ngữ và kỹ thuật

## Mô hình ngôn ngữ Neural (Neural Language Models)

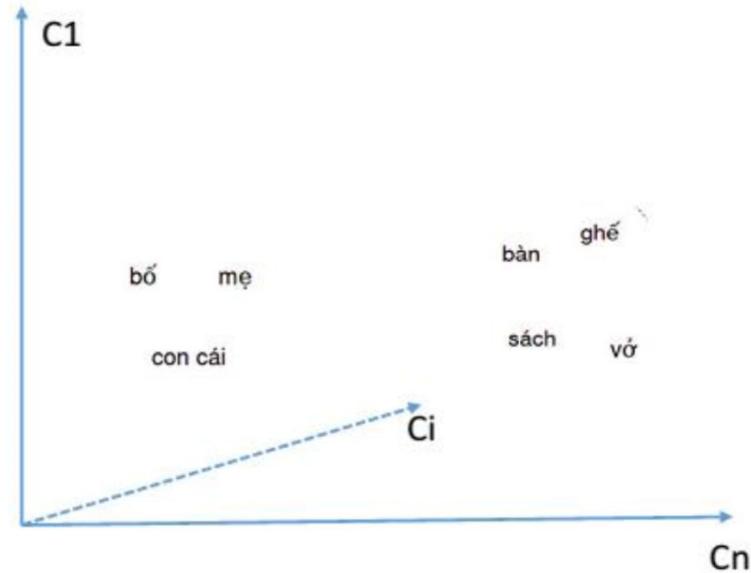
### Word Representation (Biểu diễn từ)

- The words: “bố”, “mẹ”, “con cái”, “bàn”, “ghé”, “sách”, “võ”

Embed the words into a space of abstract concepts

- The words are placed in a space of  $n \ll |V|$  dimensions.

Word vectors	Dimensions			
	animal	domesticated	pet	fluffy
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37



# Tách từ tiếng Việt

- Mục đích: xác định ranh giới của các từ trong câu.
- Là bước xử lý quan trọng đối với các hệ thống XLNNTN, đặc biệt là đối với các ngôn ngữ đơn lập, ví dụ: âm tiết Trung Quốc, âm tiết Nhật, âm tiết Thái, và tiếng Việt.
- Với các ngôn ngữ đơn lập, một từ có thể có một hoặc nhiều âm tiết.

Vấn đề của bài toán tách từ là khử được sự nhập nhằng trong ranh giới từ.

# Tách từ tiếng Việt

## Từ vựng

- Tiếng Việt là ngôn ngữ không biến hình
- Từ điển từ tiếng Việt (Vietlex): >40.000 từ, trong đó:
  - 81.55% âm tiết là từ : từ đơn
  - 15.69% các từ trong từ điển là từ đơn
  - 70.72% từ ghép có 2 âm tiết
  - 13.59% từ ghép  $\geq 3$  âm tiết
  - 1.04% từ ghép  $\geq 4$  âm tiết

# Tách từ tiếng Việt

## Từ vựng

- Tiếng Việt là ngôn ngữ không biến hình
- Từ điển từ tiếng Việt (Vietlex): >40.000 từ

Độ dài	# từ	%
1	6,303	15.69
2	28,416	70.72
3	2,259	5.62
4	2,784	6.93
5	419	1.04
Tổng	40,181	100

Bảng 1. Độ dài của từ tính theo âm tiết

## Qui tắc cấu tạo từ tiếng Việt

- Từ đơn: dùng một âm tiết làm một từ.
  - Ví dụ: tôi, bác, người, cây, hoa, đi, chạy, vì, đã, à, nhỉ, nhé...
- Từ ghép: tổ hợp (ghép) các âm tiết lại, giữa các âm tiết đó có quan hệ về nghĩa với nhau.
  - Từ ghép đẳng lập. các thành tố cấu tạo có quan hệ bình đẳng với nhau về nghĩa.
    - Ví dụ: chợ búa, bếp núc
  - Từ ghép chính phụ. các thành tố cấu tạo này phụ thuộc vào thành tố cấu tạo kia. Thành tố phụ có vai trò phân loại, chuyên biệt hoá và sắc thái hoá cho thành tố chính.
    - Ví dụ: tàu hỏa, đường sắt, xấu bụng, tốt mã, ngay đơ, thẳng tắp, sưng vù...

## Qui tắc cấu tạo từ tiếng Việt

- **Từ láy:** các yếu tố cấu tạo có thành phần ngũ âm được lặp lại; nhưng vừa lặp vừa biến đổi. Một từ được lặp lại cũng cho ta từ láy.
- **Biến thể của từ:** được coi là dạng lâm thời biến động hoặc dạng "lời nói" của từ.
  - Rút gọn một từ dài thành từ ngắn hơn
    - ki-lô-gam → ki lô/ kí lô
  - Lâm thời phá vỡ cấu trúc của từ, phân bố lại yếu tố tạo từ với những yếu tố khác ngoài từ chen vào. Ví dụ:
    - khổ sở → lo khổ lo sở
    - ngọt nghẽo → cười ngọt cười nghẽo
    - danh lợi + ham chuộng → ham danh chuộng lợi

## Qui tắc cấu tạo từ tiếng Việt

- Các diễn tả gồm nhiều từ (vd, “bởi vì”) cũng được coi là 1 từ
- Tên riêng: tên người và vị trí được coi là 1 đơn vị từ vựng
- Các mẫu thường xuyên: số, thời gian

## Các hướng tiếp cận

- Tiếp cận dựa trên từ điển
- Tiếp cận dựa trên học máy
- Kết hợp hai phương pháp trên.

## Tách từ dựa trên từ điển

- Thuật toán so khớp từ dài nhất
- Yêu cầu:
  - Từ điển
  - Chuỗi đầu vào đã tách các dấu câu và âm tiết
- Tư tưởng: thuật toán tham lam
  - Đi từ trái sang phải hoặc từ phải sang trái, lấy các từ dài nhất có thể, dừng lại khi duyệt hết
  - Độ phức tạp tính toán:  $O(n \cdot V)$ 
    - n: Số âm tiết trong chuỗi
    - V: Số từ trong từ điển

## Tách từ dựa trên từ điển

- **Ưu điểm:**
  - Cài đặt đơn giản
  - Độ phức tạp tính toán hợp lý
  - Không yêu cầu dữ liệu huấn luyện
- **Nhược điểm:**
  - Phụ thuộc vào từ điển
  - Chưa giải quyết được vấn đề nhập nhằng

## Cách tách từ đơn giản

- Phát hiện các mẫu thông thường như tên riêng, chữ viết tắt, số, ngày tháng, địa chỉ email, URL,... sử dụng biểu thức chính quy
- Chọn chuỗi âm tiết dài nhất từ vị trí hiện tại và có trong từ điển, chọn cách tách có ít từ nhất
- Hạn chế: có thể đưa ra cách phân tích không đúng.
- Giải quyết: liệt kê tất, có 1 chiến lược để chọn cách tách tốt nhất.

## Tách từ sử dụng biểu thức chính quy

- là một khuôn mẫu được so sánh với một chuỗi
- Các ký tự đặc biệt:
  - \* - bất cứ chuỗi ký tự nào, kể cả không có gì
  - x – ít nhất 1 ký tự
  - + - chuỗi trong ngoặc xuất hiện ít nhất 1 lần
- Ví dụ:
  - Email: x@x(.x)+
  - dir \*.txt
  - ‘\*John’ -> ‘John’, ‘Ajohn’, “Decker John”
- Biểu thức chính quy được sử dụng đặc biệt nhiều trong:
  - Phân tích cú pháp
  - Xác nhận tính hợp lệ của dữ liệu
  - Xử lý chuỗi
  - Trích rút thông tin

## Một số công cụ tách từ

- JvnSegmenter (Nguyễn Cẩm Tú) : CRF
  - <http://jvnsegmenter.sourceforge.net>
- VnTokenizer (Lê Hồng Phương)
  - <https://github.com/phuonglh/vn.vitk>
- Dongdu (Lưu Anh Tuấn): SVM
  - <http://viet.jnlp.org/dongdu>
- Pyvi (Trần Việt Trung) : <https://github.com/trungtv/pyvi>
- Từ điển từ:
  - <http://tratu.coviet.vn/tu-dien-lac-viet.aspx>
  - <http://tratu.soha.vn/>
  - <https://www.informatik.uni-leipzig.de/~duc/Dict/>
  -

# THANK YOU !

**COLE.VN**  
Connecting knowledge



[www.cole.vn](http://www.cole.vn)