



# XỬ LÝ NGÔN NGỮ TỰ NHIÊN

## CHƯƠNG 5: Sequential labeling

ThS. Lưu Thanh Sơn



## NỘI DUNG

1. Định nghĩa bài toán.
2. Part of speech tagging
- 3. Named entity recognition.**
4. Markov chains
5. HMM Tagger
6. CRF
7. Độ đo đánh giá



### 3. Named Entities Recognition



# Named entity recognition

- **Named entity** (tạm dịch: thực thể tên) là các thực thể đề cập đến một danh từ riêng cụ thể (proper name). Các danh từ riêng này có thể là: người (person - PER), địa điểm (location - LOC), tổ chức (organization - ORG), hoặc là thực thể địa chính trị (geo-political entity - GPE)
- Bài toán nhận diện các thực thể tên trong văn bản được gọi là bài toán **named-entity recognition** (NER).



# Một số thực thể tên trong tiếng Anh

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	Turing is a giant of computer science.
Organization	ORG	companies, sports teams	The IPCC warned about the cyclone.
Location	LOC	regions, mountains, seas	Mt. Sanitas is in Sunshine Canyon.
Geo-Political Entity	GPE	countries, states	Palo Alto is raising the fees for parking.

Ví dụ về nhận dạng thực thể tên trong văn bản

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY \$6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].



# Vai trò của NER

- **Trong sentiment analysis:** nhận diện thực thể tên giúp cho việc phân tích cảm xúc của một người dùng đối với một đối tượng cụ thể (VD: sản phẩm, hàng hoá, dịch vụ, nhân viên, ...)
- **Trong question-answering:** nhận diện thực thể tên giúp cho việc trích xuất các thông tin về các sự kiện (event) được đề cập trong câu hỏi hoặc câu trả lời và mối liên kết giữa các sự kiện đó với thông tin cần truy xuất để tìm ra câu trả lời chính xác.
- .....



# Natural language understanding





# Khó khăn đối với NER

- Vấn đề nhận diện thực thể cho một **spans of text** → phụ thuộc vào vấn đề segmentation trong text.
  - Spans of text: một hợp các từ trong một khoảng nhất định trong văn bản.
  - VD: [John F. Kennedy] → spans gồm các từ: “John”, “F.”, “Kennedy”
- Sự mơ hồ về ngữ nghĩa của từ (type ambiguity).

VD: từ **JFK** có thể là:

- Tên một người (PER): John F. Kennedy (tổng thống thứ 35 của Mỹ).
- Tên một địa điểm (LOC): sân bay JFK ở TP New York.
- Tên một trường trung học (ORG).
- Tên một con đường (ORG).
- ....



# Ví dụ về type ambiguity trong bài toán NER

[PER Washington] was born into slavery on the farm of James Burroughs.

[ORG Washington] went up 2 games to 1 in the four-game series.

Blair arrived in [LOC Washington] for what may well be his last state visit.

In June, [GPE Washington] passed a primary seatbelt law.



# BIO trong NER

- Cách tiếp cận theo BIO trong NER: một thực thể tên riêng sẽ bao gồm 2 thành phần:
  - *Loại tên thực thể (named entity types).*
  - *Phạm vi (boundary).*
- Cách đánh nhãn tên thực thể theo cách tiếp cận BIO tagging:
  - *Token bắt đầu spans sẽ được ký hiệu là B.*
  - *Token nằm trong một spans được ký hiệu là I.*
  - *Token nằm ngoài spans được ký hiệu là O.*
- Ứng với mỗi spans sẽ luôn có 2 nhãn B và I cho một thực thể tên.
- Như vậy, với n nhãn thực thể tên ban đầu, theo cách tiếp cận của BIO tagging chúng ta sẽ có tổng cộng **2n+1** nhãn thực thể tên.



## Các dạng khác theo cách tiếp cận BIO tagging cho spans of text

- IO: giống như BIO tagging, nhưng bỏ đi ký hiệu B (boundary).
- BIOES: Giống như BIO tagging, nhưng thêm vào ký hiệu E để ký hiệu cho token kết thúc spans, và ký hiệu S để chỉ spans chỉ có một ký tự.



# Ví dụ về gán nhãn thực thể tên

Text: *Jane Villanueva of United Airlines Holding discussed the Chicago route.*

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O



# Một số dataset dùng cho NER

- Tiếng Anh:
  - CoNLL-2002 and CoNLL-2003 (British newswire).
  - ACE.
  - MUC-6 and MUC-7 (American newswire)
- Tiếng Việt:
  - VLSP 2016 (<https://vlsp.org.vn/resources-vlsp2016>).



# Phương pháp dùng cho POS-Tagging và NER

- HMM – Hidden Markov chains.
- CRF – Conditional Random Fields.