CSE: **Faculty of Computer Science and Engineering**
Thuyloi University

NHẬN DẠNG THỰC THỂ CÓ TÊN (Named Entity Recognition)

TS. Nguyễn Thị Kim Ngân
Email: ngannguyen@tlu.edu.vn

Nhận dạng thực thể có tên

- Nhận dạng thực thể có tên (Named Entity Recognition – NER) nhằm nhận biết các chuỗi từ trong văn bản là tên của một đối tượng nào đó, điển hình như tên người, tên tổ chức, tên địa danh, thời gian, …

- NER là nhiệm vụ đóng vai trò quan trọng trong các ứng dụng trích xuất thông tin

Ví dụ

- Anh [PER **Thanh**] là cán_bộ [ORG **Ủy_ban nhân_dân Thành_phố Hà_Nội**].
- [PER **Jane Villanueva**] of [ORG **United**] , a unit of [ORG **United Airlines Holding**], said the fare applies to the [LOC **Chicago**] route.

Một số kiểu thực thể

NE type	Examples
ORGANIZATION	*Georgia-Pacific Corp., WHO*
PERSON	*Eddy Bonte, President Obama*
LOCATION	*Murray River, Mount Everest*
DATE	*June, 2008-06-29*
TIME	*two fifty a m, 1:30 p.m.*
MONEY	*175 million Canadian Dollars, GBP 10.40*
PERCENT	*twenty pct, 18.75 %*
FACILITY	*Washington Monument, Stonehenge*
GPE	*South East Asia, Midlothian*

Nhãn thực thể theo cấu trúc BIO

- Nhãn thực thể được gán theo cấu trúc IO, BIO, BIOES như định dạng dữ liệu phân cụm CoNLL (tham khảo bộ dữ liệu CoNLL 2023 https://github.com/rahinic/BIO-to-BIOES-tagger):
  - B: Begin (bắt đầu), for the first token of a chunk phrase/NE
  - I: Inside (bên trong), for tokens inside chunk phrase/NE's
  - E: End (kết thúc), for the end tokens of chunk phrase/NE's
  - O: Outside (từ không cần nhận diện), for tokens outside/other any chunk phrase/NE
  - S: for unit/single length chunk phrase/NE’s

Ví dụ

- [PER **Jane Villanueva**] of [ORG **United**] , a unit of [ORG **United Airlines Holding**], said the fare applies to the [LOC **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

Tập dữ liệu huấn luyện cho tiếng Việt

- VLSP 2016 (https://vlsp.org.vn/vi/node/56): File dữ liệu huấn luyện chứa một văn bản đã tách từ và gán nhãn. Mỗi từ được đặt trên một dòng riêng biệt và mỗi câu được phân cách nhau bởi một dòng trống. Mỗi dòng bao gồm các cột được cách nhau bởi một khoảng trắng:
  - Cột đầu tiên là một từ
  - Cột thứ hai là từ loại của từ
  - Cột thứ 3 là nhãn phân cụm cú pháp
  - Cột thứ 4 là nhãn thực thể
  - Cột thứ 5 là nhãn thực thể lồng

Ví dụ

- Anh [PER **Thanh**] là cán_bộ [ORG **Ủy_ban nhân_dân Thành_phố Hà_Nội**].

Word	POS	Phrase	Nhãn thực thể	Nhãn thực thể lồng
Anh	N	B-NP	O	O
Thanh	NPP	B-NP	B-PER	O
là	V	B-VP	O	O
cán_bộ	N	B-NP	O	O
Uỷ_ban	N	B-NP	B-ORG	O
nhân_dân	N	I-NP	I-ORG	O
Thành_phố	N	I-NP	I-ORG	B-LOC
Hà_Nội	NPP	I-NP	I-ORG	I-LOC
.	CH	O	O	O

Phương pháp nhận dạng thực thể

- Conditional Random Fields (CRF)
- Mạng nơ-ron hồi tiếp