TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÀI TẬP LỚN**

HỌC PHẦN: XỬ LÝ NGÔN NGỮ TỰ NHIÊN

**ĐỀ TÀI: XÂY DỰNG MÔ HÌNH NHẬN DẠNG THỰC THỂ**

Giáo viên hướng dẫn:

Nguyễn Thị Kim Ngân

Học viên thực hiện:

Cao Thị Mai Phương - 2054015772

Nguyễn Văn Luyện - 2151264669

Nguyễn Đình Cường - 2151264648

Hoàng Hiểu Nhi - 2154020987

**Hà Nội, năm 2024**

**MỤC LỤC**

[Phần 1: Lý thuyết 1](#_Toc170508829)

[1. Lý thuyết nhận dạng thực thể 1](#_Toc170508830)

[2. Các loại nhãn thực thể 2](#_Toc170508831)

[3. Mô hình CRF 3](#_Toc170508832)

[Phần 2: Ứng dụng 4](#_Toc170508833)

[1. Mô tả tập dữ liệu 4](#_Toc170508834)

[2. Thiết kế mô hình CRF (áp dụng với Bài 2 của Bài tập lớn số 2) 4](#_Toc170508835)

[3. Kết luận 6](#_Toc170508836)

[Tài liệu tham khảo 8](#_Toc170508837)

# Phần 1: Lý thuyết

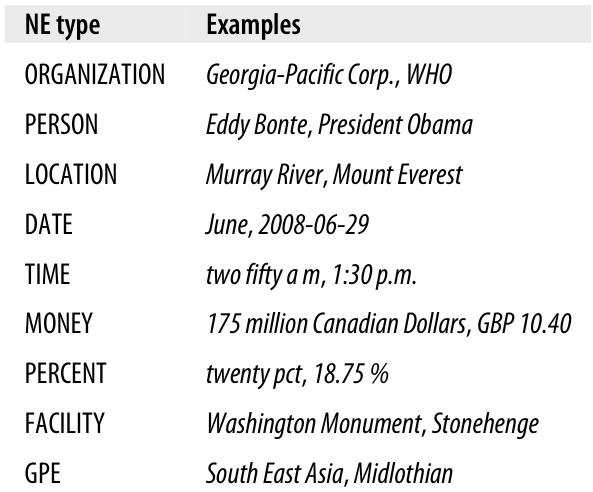
## Lý thuyết nhận dạng thực thể

Nhận dạng thực thể (Named Entity Recognition - NER): là tác vụ cơ bản trong lĩnh vực Xử lý ngôn ngữ tự nhiên. Vai trò chính của tác vụ này là nhận dạng các cụm từ trong văn bản và phân loại chúng vào trong các nhóm đã được định trước như tên người, tổ chức, địa điểm, thời gian, loại sản phẩm, nhãn hiệu, v.v…

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

Ví dụ: Thanh <B-PER> là <O> nhà <O> khoa học <O> . <O>

Một số kiểu thực thể:



* Tên tổ chức (Organization) bao gồm các loại tên sau:
* Các cơ quan chính phủ (các bộ ngành, uỷ ban nhân dân, hội đồng nhân dân, toà án, cơ quan báo chí, hội nghề nghiệp, đoàn thể chính trị, phòng ban, ...)
* Công ti (ngân hàng, thị trường chứng khoán, hãng phim, nhà sản xuất, hợp tác xã, phòng ban, ....)
* Các thương hiệu
* Các tổ chức chính trị (các đảng phái chính trị, ...)
* Các ấn phẩm (các tạp chí, báo)
* Các công ty âm nhạc (ban nhạc, dàn nhạc, đội hợp xướng ...)
* Các tổ chức công cộng (trường học, tổ chức từ thiện)
* Các tổ chức khác của con người (câu lạc bộ thể thao, các hiệp hội, nhà hát, công ti, tôn giáo, tổ chức thanh niên...)
* Tên người (Person) bao hàm các loại tên sau:
* Tên, tên đệm và họ của một người
* Tên động vật và các nhân vật hư cấu
* Các bí danh
* Tên địa lí (Địa danh - Location) bao gồm các thực thể có toạ độ địa lí nhất định,ghi lại được trên bản đồ:
* Tên gọi các hành tinh: Mặt Trăng, Mặt Trời, Trái Đất…
* Tên gọi các thực thể mang yếu tố địa lí tự nhiên và địa lí lịch sử (quốc gia, vùng lãnh thổ, châu lục), các vùng quần cư (làng, thị trấn, thành phố, tỉnh, giáo khu, giáo xứ), các điểm kinh tế (vùng nông nghiệp, khu công nghiệp)
* Tên gọi các thực thể tự nhiên (đèo, núi, dãy núi, rừng, sông, suối, hồ, biển, vịnh, vũng, eo biển, đại dương, thung lũng, cao nguyên, đồng bằng, khu bảo tồn thiên nhiên, bãi biển, khu sinh thái, v.v.)
* v.v.v…

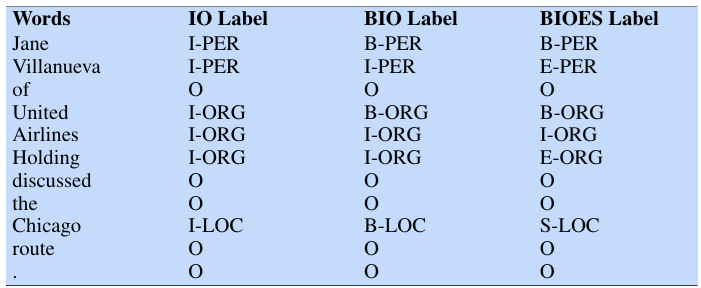
## Các loại nhãn thực thể

Nhãn thực thể được gán theo cấu trúc IO, BIO, BIOES như định dạng dữ liệu

phân cụm CoNLL (bộ dữ liệu CoNLL 2023)

* B: Begin (Bắt đầu) cho từ đầu tiên của một đoạn thực thể.
* I: Inside (Bên trong) cho các từ nằm bên trong đoạn thực thể.
* E: End (Kết thúc) cho từ cuối cùng của đoạn thực thể.
* O: Outside (Bên ngoài) cho các từ không thuộc đoạn thực thể hoặc các từ khác.
* S: Thực thể có độ dài đơn vị (Single), tức là chỉ gồm một từ.

Ví dụ: [PER Jane Villanueva ] of [ORG United] , a unit of [ORG United Airlines Holding], said the fare applies to the [LOC Chicago] route.



Các thực thể tùy thuộc vào loại cấu trúc sẽ chọn được các nhãn khác nhau.

## Mô hình CRF

CRF là mô hình học tập quan hệ. NER sử dụng CRF dựa trên mô hình đồ họa vô hướng của automata trạng thái hữu hạn xác suất được huấn luyện có điều kiện. CRFđược sử dụng để tính xác suất có điều kiện của các giá trị trên các nút đầu ra được chỉ định cho các giá trị đã cho trên các nút đầu vào được chỉ định khác. Nó kết hợp các tính năng phụ thuộc và học tập phụ thuộc vào ngữ cảnh. Nó cho phép biểu diễn sự phụ thuộc vào các phân loại trước đó trong một diễn ngôn. Ý tưởng cơ bản là bối cảnh xung quanh tên sẽ trở thành bằng chứng tốt trong việc gắn thẻ một lầnxuất hiện khác có cùng tên trong một bối cảnh không rõ ràng khác. Hầu hết các nghiên 3 cứu về NER sử dụng vectơ đặc trưng cần triển khai CRF bao gồmcác tính năng ngôn ngữ và thẻ POS, bộ phân tích hình thái, công báo và kho văn bản có chú thích NE. Đối với học máy, dữ liệu cần được chuyển đổi thành vectơ đặc trưng cho mỗi từ trong đó bản đồ đặc trưng chứa thông tin ngữ cảnh của từ đó. Hệ thống (Mner-CRF) đã triển khai chức năng tính năng lấy một câu, vị trí của từ hiện tại, nhãn của từ hiện tại và nhãn của từ trước đó làm tham số đầu vào cho hàm từ tập dữ liệu huấn luyện và dự đoán thẻ NE thích hợp nhất cho một từ trong câu dựa trên việc học. Việc đào tạo hệ thống CRF được thực hiện với tập dữ liệu huấn luyện trả về tệp mô hình làm đầu ra. Tệp mô hình được hệ thống CRF sử dụng trong quá trình thử nghiệm. Đào tạo CRF và tạo tệp mô hình là mô-đun cốt lõi trong quá trình phát triển hệ thống NER. Thuật toán trường ngẫu nhiên có điều kiện sử dụng câu x làm chuỗi đầu vào, trong đó x1, x2, . . . , xn , là các từ của một câu và chuỗi thẻ t , trong đó t1, t2, . . . , tn là chuỗi các trạng thái đầu ra, tức là các thẻ NE. Trong CRF, xác suất có điều kiện được mô hình hóa bằng cách xác định một bản đồ đặc trưng ánh xạ toàn bộ chuỗi đầu vào được ghép nối với toàn bộ chuỗi trạng thái P(t1, t2, . . . , tn|x1, x2, . . . , xn) tới một số vectơ đặc trưng d-chiều. Sau đó, chúng ta có thể mô hình hóa xác suất dưới dạng mô hình log-tuyến tính với vectơ tham số.

# Phần 2: Ứng dụng

## Mô tả tập dữ liệu

- Nguồn lấy dữ liệu thu thập từ github

<https://raw.githubusercontent.com/datnnt1997/bert_vn_ner/master/data/sample_data/raw_data/train.csv>

<https://raw.githubusercontent.com/datnnt1997/bert_vn_ner/master/data/sample_data/raw_data/test.csv>

- Tập dữ liệu có bao nhiêu câu đã được gán nhãn

2413 câu train

34 câu test

## 2. Thiết kế mô hình CRF (áp dụng với Bài 2 của Bài tập lớn số 2)

- Input: Dữ liệu đầu vào là các câu văn bản, trong đó mỗi câu được chia thành các từ. Các từ này sẽ được đại diện dưới dạng các đặc trưng (features) để mô hình có thể học.

- Output: Dữ liệu đầu ra là các nhãn thực thể tương ứng với từng từ trongcâu. Các nhãn này bao gồm các loại thực thể như tên người (B-PER, I-PER), tổ chức(B-ORG, I-ORG), địa điểm (B-LOC, I-LOC), và các nhãn khác

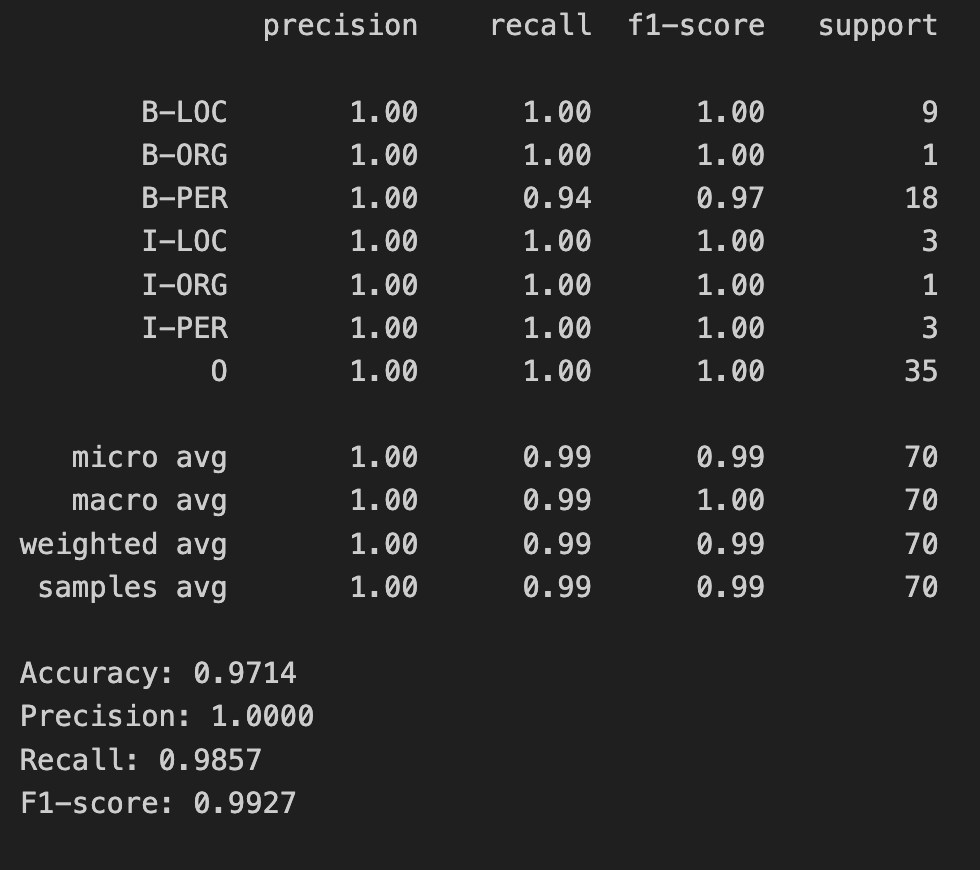
- Kiến trúc mô hình:

Mô hình CRF hoạt động dựa trên việc tối ưu hóa hàm mục tiêu thông qua thuật toánL-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno), với các tham số quan trọng:

- c1 và c2: Các hệ số điều chỉnh (regularization coefficients) giúp kiểm soát mức độ phức tạp của mô hình để tránh hiện tượng overfitting.

- max\_iterations: Số lượng vòng lặp tối đa mà thuật toán sẽ chạy để tìm nghiệm tối ưu.

Kết quả đánh giá mô hình như sau:



Kết quả đánh giá trên tập dữ liệu kiểm tra (test data) cho thấy mô hình CRF đạt được các chỉ số đo lường tốt trên tập dữ liệu kiểm tra, với các giá trị Precision, Recall và F1-score cao, chứng tỏ mô hình có khả năng nhận diện và phânloại cácthực thể một cách chính xác.

Biểu đồ so sánh các độ đo đánh giá chất lượng mô hình (Accuracy, Recall, Precision và F1-score) của Bài 1 và Bài 2.

A graph of blue and orange bars

Description automatically generated

Thư viện underthesea cho kết quả dự đoán thấp hơn so với mô hình CRF trên tất cả các độ đo. Tuy vậy, thư viện underthesea vẫn là một sự lựa chọn tốt trong việc nhận diện thực thể.

## 3. Kết luận

Trong bài tập lớn này, chúng tôi đã nghiên cứu và thực hiện việc nhận dạng thực thể (NER) sử dụng thư viện Underthesea và mô hình CRF.

**Sử dụng thư viện** Underthesea : Thư viện Underthesea cung cấp các công cụ mạnh mẽ để xử lý ngôn ngữ tự nhiên bằng tiếng Việt, bao gồm cả nhận dạng thực thể. Các chức năng của thư viện bao gồm phân đoạn từ, nhận dạng thực thể, và những công cụ khác hỗ trợ nghiên cứu và phát triển các ứng dụng xử lý ngôn ngữ tự nhiên.

**Áp dụng mô hình CRF**: Mô hình CRF là một mô hình học máy phổ biến trong bài toán nhận dạng thực thể, đặc biệt là trong ngữ cảnh xử lý ngôn ngữ tự nhiên. Chúng tôi đã huấn luyện mô hình CRF để nhận dạng các loại thực thể trong văn bản.

**Đánh giá và kết quả**: Chúng tôi đã sử dụng các tập dữ liệu thí nghiệm để đánh giá hiệu suất của hệ thống nhận dạng thực thể. Kết quả cho thấy mô hình CRF cung cấp độ chính xác cao trong việc nhận dạng các thực thể trong văn bản tiếng Việt. Độ đo Accuracy, Recall, Precision và F1-score của hệ thống đã đạt được mức đáng kể, cho thấy tính hiệu quả và khả năng áp dụng trong thực tế của mô hình.

**Hướng phát triển tiếp theo**: Để cải thiện hiệu suất, chúng tôi có thể xem xét việc tinh chỉnh các siêu tham số của mô hình CRF hoặc thử nghiệm các mô hình khác như BERT hay LSTM cho nhiệm vụ nhận dạng thực thể. Mở rộng ứng dụng của hệ thống để xử lý các loại văn bản phức tạp hơn và đa dạng hơn, đặc biệt là trong các ngữ cảnh thực tế.

Với kết quả đạt được, chúng tôi tin rằng nghiên cứu và triển khai các phương pháp này sẽ đóng góp ý nghĩa vào lĩnh vực xử lý ngôn ngữ tự nhiên và ứng dụng của nó trong thực tế.

# Tài liệu tham khảo

<https://vlsp.org.vn/sites/default/files/2019-06/Guidelines-NER-2018_0_0.pdf>

<https://viblo.asia/p/seri-nlp-nhan-dang-thuc-the-ner-phan-1-Ljy5VyWzlra>

<https://vi.shaip.com/blog/named-entity-recognition-and-its-types/>