ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

XỬ LÝ NGÔN NGỮ TỰ NHIÊN - CS221.Q12

BÁO CÁO ĐỒ ÁN
NHẬN DIỆN THỰC THỂ TÊN RIÊNG (NER) TIẾNG VIỆT

Nhóm sinh viên (Nhóm 4):
Nguyễn Công Phát - 23521143
Phạm Trần Khánh Duy - 23520384
Nguyễn Lê Phong - 23521168

GVHD :
TS. Nguyễn Thị Quý

Ngày 26 tháng 12 năm 2025

Mục lục

1 Tổng quan đồ án

Đồ án được thực hiện trong khuôn khổ môn học **CS221.Q12 – Xử lý Ngôn ngữ Tự nhiên**, tập trung vào bài toán **Nhận diện thực thể tên (Named Entity Recognition – NER)** cho tiếng Việt.

Mục tiêu chính của đồ án là xây dựng một hệ thống NER hoàn chỉnh ở mức học thuật, bao gồm:

- Cài đặt ba mô hình NER: HMM, CRF và BiLSTM-CRF trên cùng định dạng dữ liệu.
- Huấn luyện và đánh giá mô hình thông qua các Jupyter Notebook trong `src/`.
- Đánh giá bằng các chỉ số quan trọng: Macro F1, Non-O F1 và Span F1.
- Lưu mô hình tốt nhất phục vụ suy diễn (`.joblib`, `.pt`).
- Demo Flask cho phép chuyển mô hình, trực quan hóa thực thể và hiển thị token/nhãn dự đoán.

Mã nguồn đồ án: `https://github.com/paht2005/CS221.Q12-Vietnamese-Named-Entity-Recognition`

2 Phương pháp

Đồ án triển khai ba mô hình gán nhãn chuỗi tiêu biểu cho bài toán NER, từ mô hình thống kê truyền thống đến mô hình học sâu, nhằm so sánh hiệu quả trên cùng bộ dữ liệu.

- **Hidden Markov Model (HMM – baseline):** Mô hình xác suất thống kê, xem chuỗi nhãn là trạng thái ẩn và chuỗi từ là quan sát. Chuỗi nhãn tối ưu được suy diễn bằng thuật toán Viterbi.
- **Conditional Random Fields (CRF):** Mô hình phân biệt cho toàn bộ chuỗi nhãn, khai thác đặc trưng ngữ cảnh và hình thái của token, đồng thời ràng buộc tính nhất quán của chuỗi nhãn.
- **BiLSTM-CRF:** Kết hợp BiLSTM để học ngữ cảnh hai chiều và lớp CRF ở đầu ra nhằm đảm bảo chuỗi nhãn hợp lệ, có khả năng học đặc trưng tự động từ dữ liệu.

Trong bài toán NER, các độ đo được ưu tiên sử dụng là **Precision**, **Recall** và đặc biệt là **F1-score**. Do nhãn `O` chiếm tỷ lệ lớn, đồ án tập trung phân tích **Non-O F1** (bỏ nhãn `O`) và **Span F1** ở mức thực thể, thay vì chỉ dựa vào Accuracy hoặc Token F1 tổng thể.

3 Dataset

- **Nguồn dữ liệu:** VLSP 2016 Vietnamese NER Shared Task, công bố công khai trên HuggingFace.
- **Định dạng:** Chuẩn CoNLL/BIO; mỗi dòng gồm `<token> <tag>`, các câu cách nhau bởi dòng trống.

- **Cấu trúc:** Hai tập dữ liệu `train.txt` (huấn luyện) và `test.txt` (đánh giá), theo tỉ lệ xấp xỉ 80/20.
- **Hệ thống nhãn:** `PER`, `ORG`, `LOC`, `MISC` và `O`, gán theo chuẩn BIO.
- **Đặc điểm:** Dữ liệu mất cân bằng mạnh (nhãn `O` chiếm đa số); thực thể tiếng Việt thường dài và nhiều token.

4 Cài đặt và triển khai

4.1 Cài đặt

```
1. git clone
↪  https://github.com/paht2005/CS221.Q12-Vietnamese-Named-Entity-Recognition.git
2. cd CS221.Q12-Vietnamese-Named-Entity-Recognition
3. pip install -r requirements.txt
```

4.2 Chạy demo

```
python app.py
# Open browser at: http://127.0.0.1:5000
```

5 Kết quả

Bảng dưới tổng hợp kết quả trên **test set**. Do nhãn `O` chiếm đa số, các chỉ số **Non-O F1** và **Span F1** phản ánh chất lượng NER rõ hơn so với Accuracy/Token F1 (ALL).

Chỉ số	HMM	CRF	BiLSTM-CRF
Accuracy	0.97	**0.9904**	0.9851
Token F1 (ALL)	0.98	**0.9901**	0.9843
Token F1 (Non-O)	–	**0.9076**	0.8642
Macro F1	**0.72**	0.8875	0.8535
Span F1	–	**0.9191**	0.8834

Ghi chú: Mô hình HMM chỉ đánh giá ở mức token (Accuracy, Macro F1); các chỉ số **Non-O F1** và **Span F1** không áp dụng do HMM không mô hình hoá trực tiếp thực thể (span).

Nhận xét ngắn: CRF cho kết quả tốt nhất trên tập test; BiLSTM-CRF đạt hiệu năng cạnh tranh nhưng còn hạn chế do dữ liệu không lớn và chưa sử dụng embedding tiền huấn luyện; HMM đóng vai trò mô hình baseline.

6 Kết luận

Nhóm đã hoàn thiện pipeline NER tiếng Việt từ dữ liệu → huấn luyện → đánh giá → demo. Trong điều kiện hiện tại, **CRF cho hiệu quả cao nhất**. Hướng phát triển: dùng embedding pretrained (fastText/PhoBERT) hoặc mô hình Transformer để cải thiện BiLSTM-CRF/NER.