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 :
Nguyễn Công Phát - 23521143
Phạm Trần Khánh Duy - 23520384
Nguyễn Lê Phong - 23521168

GVHD :
TS. Nguyễn Thị Quý

Ngày 24 tháng 12 năm 2025

Mục lục

LỜI NÓI ĐẦU

Nhận diện thực thể tên (Named Entity Recognition – NER) là một bài toán quan trọng trong xử lý ngôn ngữ tự nhiên, nhằm tìm và gán nhãn cho các cụm từ biểu diễn tên người, địa điểm, tổ chức hoặc các loại thực thể khác trong câu. Đối với tiếng Việt, cấu trúc từ nhiều âm tiết, cách viết có dấu gạch dưới hoặc tên riêng dài khiến bài toán NER có nhiều thách thức hơn so với một số ngôn ngữ khác.

Trong báo cáo này, nhóm tiến hành cài đặt và thực nghiệm ba mô hình gán nhãn chuỗi phổ biến: mô hình thống kê HMM, mô hình CRF và mô hình học sâu BiLSTM kết hợp CRF. Tất cả các mô hình đều được huấn luyện và đánh giá trên bộ dữ liệu NER tiếng Việt VLSP 2016 đã được nhóm chuyển về định dạng dòng (CoNLL) trong các tệp `train.txt` và `test.txt`. Nội dung báo cáo được xây dựng dựa trên các notebook thực nghiệm của nhóm (`HMM.ipynb`, `CRF + BiLSTM-CRF.ipynb`) và bộ slide trình bày nội bộ.

Nhóm xin chân thành cảm ơn giảng viên đã hướng dẫn, hỗ trợ và cung cấp kiến thức nền tảng để nhóm hoàn thành đề tài.

TÓM TẮT ĐỒ ÁN

Đề tài thực hiện bài toán Nhận diện thực thể tên (Named Entity Recognition – NER) cho tiếng Việt trên bộ dữ liệu VLSP 2016 (định dạng CoNLL). Nhóm cài đặt, huấn luyện và so sánh ba mô hình gán nhãn chuỗi: HMM, CRF và BiLSTM-CRF. Ngoài phần thực nghiệm, nhóm triển khai demo Flask cho phép thử câu tiếng Việt và chuyển đổi giữa CRF và BiLSTM-CRF.

Repository: `https://github.com/paht2005/CS221.Q12-Vietnamese-Named-Entity-Recognition`

Chương 1

GIỚI THIỆU & DATASET

1.1 Bài toán

NER là bài toán gán nhãn cho từng token trong câu theo hệ nhãn BIO để nhận biết các cụm từ biểu diễn thực thể. Đối với tiếng Việt, đặc thù từ nhiều âm tiết (thường được nối bằng dấu gạch dưới trong dữ liệu), cùng với sự mất cân bằng nhãn (nhãn `O` chiếm đa số) khiến việc đánh giá cần chú ý các chỉ số chất lượng thực thể (Non-O F1, Span F1), không chỉ Accuracy/Token F1.

1.2 Dataset

Nhóm sử dụng VLSP 2016 Vietnamese NER, được chuẩn hoá về hai tệp: `dataset/train.txt` và `dataset/test.txt`. Mỗi dòng gồm `token` và `tag`, câu được ngăn cách bởi dòng trống.
Hệ nhãn: BIO với các loại thực thể chính: `PER`, `ORG`, `LOC`, `MISC` và `O` (không phải thực thể).
Ví dụ:

```
Hị_Ni B-LOC
lị O
th_ũ O
ca O
Vit_Nam B-LOC
```

Chương 2

PHƯƠNG PHÁP

2.1 HMM (Baseline)

Hidden Markov Model mô hình hoá chuỗi nhãn là trạng thái ẩn và chuỗi token là quan sát, ước lượng xác suất chuyển trạng thái $P(y_t|y_{t-1})$ và xác suất phát xạ $P(x_t|y_t)$. Dùng Viterbi để suy diễn chuỗi nhãn tối ưu.

2.2 CRF

Conditional Random Fields là mô hình xác suất có điều kiện cho toàn bộ chuỗi nhãn, cho phép dùng nhiều đặc trưng ngữ cảnh/hình thái (token hiện tại, token lân cận, chữ hoa/thường, có chữ số, dấu gạch dưới, prefix/suffix, BOS/EOS). Nhóm huấn luyện CRF bằng L-BFGS với regularization.

2.3 BiLSTM-CRF

BiLSTM-CRF kết hợp BiLSTM để học biểu diễn ngữ cảnh hai chiều và lớp CRF đầu ra để ràng buộc chuỗi nhãn hợp lệ, tối ưu theo toàn câu. Mô hình được huấn luyện bằng PyTorch; CRF layer dùng `pytorch-crf`.

Chương 3

THỰC NGHIỆM & KẾT QUẢ

3.1 Thiết lập đánh giá

Nhóm huấn luyện trên `train.txt` và đánh giá trên `test.txt`. Do nhãn `O` chiếm tỷ lệ lớn, nhóm ưu tiên:

- **Non-O F1**: Token-level F1 bỏ nhãn `O`.
- **Span F1**: Entity-level F1 (khắt khe nhất), yêu cầu đúng biên thực thể.

3.2 Bảng so sánh tổng quan (Test set)

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

Bảng 3.1: So sánh hiệu suất ba mô hình trên tập Test.

3.3 Nhận xét

- **CRF tốt nhất** trên Test: **Acc=0.9904**, **Span-F1=0.9191**, **Non-O F1=0.9076**.
- **BiLSTM-CRF** đứng sau: **Acc=0.9851**, **Span-F1=0.8834**, **Non-O F1=0.8642**.
- **HMM** (tối ưu) cải thiện Macro-F1 **0.51** $\rightarrow$ **0.72**, nhưng vẫn kém do giả định Markov.
- Token-level F1 (ALL) có thể cao do `O` nhiều $\Rightarrow$ Non-O F1 và Span F1 phản ánh chất lượng NER tốt hơn.

Chương 4

DEMO & TRIỂN KHAI

Nhóm xây dựng demo Flask để thử mô hình trực quan:

- Chuyển đổi giữa **CRF** và **BiLSTM-CRF**.
- Highlight thực thể dự đoán theo nhãn.
- Hiển thị bảng token/tag và thông số mô hình.

Chạy demo:

```
pip install -r requirements.txt
python app.py
# http://127.0.0.1:5000
```

Ảnh demo: `outputs/demo.png`.

Chương 5

KẾT LUẬN & HƯỚNG PHÁT TRIỂN

5.1 Kết luận

Nhóm đã hoàn thiện pipeline NER tiếng Việt trên VLSP 2016 và so sánh ba mô hình. Với dữ liệu hiện tại, **CRF cho kết quả tốt nhất** nhờ đặc trưng thủ công phù hợp tiếng Việt và ràng buộc chuỗi nhãn. **BiLSTM-CRF** cạnh tranh nhưng cần dữ liệu lớn hơn hoặc embedding pretrained để phát huy tối đa.

5.2 Hướng phát triển

- Tích hợp embedding pretrained (fastText/PhoBERT) để cải thiện BiLSTM-CRF.
- Thử các mô hình Transformer cho NER tiếng Việt.
- Mở rộng đánh giá trên nhiều miền dữ liệu (tin tức, pháp luật, y tế).

Tài liệu tham khảo

[1] VLSP 2016, *Vietnamese Named Entity Recognition Shared Task*, 2016.

[2] J. Lafferty, A. McCallum, F. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, ICML, 2001.

[3] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, *Neural Architectures for Named Entity Recognition*, NAACL, 2016.