

Mục lục

- 1 Tổng quan đề tài
- 2 Tổng quan các phương pháp
- 3 Thực nghiệm
- 4 Kết luận & Hướng phát triển

Mục lục

- 1 Tổng quan đề tài
- 2 Tổng quan các phương pháp
- 3 Thực nghiệm
- 4 Kết luận & Hướng phát triển

1.1. Giới thiệu bài toán NER

Định nghĩa

Nhận diện thực thể tên riêng (Named Entity Recognition - NER) là bài toán **gán nhãn cho từng từ/token trong câu** để xác định các thực thể như *tên người (PER)*, *tổ chức (ORG)*, *địa danh (LOC)*, v.v.

Ví dụ (BIO)

Hà_Nội/**B-LOC** là/O thủ_dô/O của/O Việt_Nam/**B-LOC** ./O

Ứng dụng

- Trích xuất thông tin trong tin tức, báo cáo.
- Hỗ trợ hệ thống hỏi đáp, tìm kiếm thực thể.
- Tiền xử lý cho dịch máy, tóm tắt văn bản.

Thách thức với tiếng Việt

- Tên riêng thường nhiều từ: *Thành phố Hồ Chí Minh*.
- Ngữ cảnh phụ thuộc mạnh, khó đoán chỉ từ một token.
- Chữ hoa không nhất quán trong corpora tiếng Việt.
- Nhiều tên tổ chức dùng địa danh \Rightarrow dễ nhầm ORG \leftrightarrow LOC.

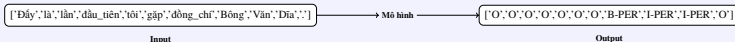
1.3. Phát biểu bài toán

Input

- **Tập dữ liệu huấn luyện** gồm N câu đã gán nhãn theo chuẩn BIO:
 - Mỗi câu: dãy token (w_1, w_2, \dots, w_t) (token đã được tách sẵn trong dataset).
 - Mỗi token có nhãn tương ứng (l_1, l_2, \dots, l_t) .
- **Câu mới cần dự đoán:** chuỗi token (w_1, w_2, \dots, w_t) .

Output

- Chuỗi nhãn dự đoán (l_1, l_2, \dots, l_t) ứng với từng token.
- Mỗi nhãn thuộc tập nhãn L theo chuẩn BIO:
 - **B-** và **I-**: bắt đầu và bên trong một thực thể (PER, ORG, LOC, MISC).
 - **O**: token không thuộc thực thể nào.



Mục lục

- 1 Tổng quan đề tài
- 2 Tổng quan các phương pháp
- 3 Thực nghiệm
- 4 Kết luận & Hướng phát triển

2.1. Mô hình Hidden Markov Model (HMM)

Tổng quan mô hình

- **Hidden Markov Model (HMM)** là mô hình xác suất dùng để gán nhãn chuỗi.
- Giả định tồn tại chuỗi trạng thái ẩn $Y = (y_1, \dots, y_T)$ sinh ra chuỗi quan sát: $X = (x_1, x_2, \dots, x_T)$
- Hai xác suất cốt lõi:
 - **Transition:** $P(y_t | y_{t-1})$
 - **Emission:** $P(x_t | y_t)$
- **Training:** ước lượng MLE;
- **Decoding:** Viterbi (tối ưu chuỗi nhãn).

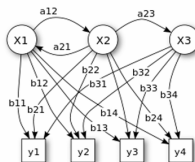
Ưu điểm

- Dễ hiểu, nền tảng xác suất rõ ràng.
- Huấn luyện nhanh, phù hợp dữ liệu nhỏ.
- Mô hình hóa tốt bài toán gán nhãn chuỗi cơ bản.

Nhược điểm

- Giả định Markov bậc 1 \Rightarrow chỉ phụ thuộc trạng thái trước.
- Không tận dụng được ngữ cảnh dài.
- Hiệu quả thấp với dữ liệu mất cân bằng nhãn (NER tiếng Việt).

Hidden Markov Model



Sơ đồ HMM: trạng thái ẩn (Y) sinh ra quan sát (X).

Quy trình dự đoán (Decoding Process)

- 1 **Khởi tạo:** Tính xác suất cho các nhãn có thể ở vị trí đầu tiên.
- 2 **Đệ quy (Forward):** Tại mỗi bước t , tính xác suất tích lũy cực đại để đến trạng thái y_t :

$$\delta_t(y) = \max_{y'} [\delta_{t-1}(y') \cdot P(y|y')] \cdot P(x_t|y)$$

- 3 **Truy vết (Backtracking):** Từ trạng thái cuối cùng có điểm cao nhất, đi ngược về đầu câu để khôi phục chuỗi nhãn tối ưu.

26/12/2025 9 / 26

10 / 26

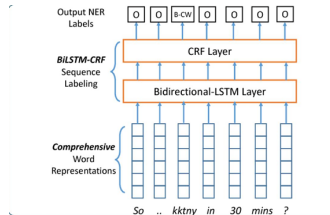
2.4. Mô hình BiLSTM + CRF

Kiến trúc tổng thể

- ① **Embedding Layer:** Ánh xạ token \rightarrow vector ngữ nghĩa.
- ② **BiLSTM Layer:** Học ngữ cảnh hai chiều (trái \leftrightarrow phải).
- ③ **CRF Layer:** Giải mã chuỗi nhãn tối ưu toàn cục.

Luồng xử lý dự đoán (Prediction Pipeline)

- ① **Input:** Chuỗi index token.
- ② **BiLSTM Forward:**
 - Qua LSTM 2 chiều \rightarrow vector ngữ cảnh h_t .
 - Tạo **Emission Scores** (điểm phát xạ).
- ③ **CRF Layer:**
 - Cung cấp **Transition Scores** (điểm chuyển nhãn).
- ④ **Viterbi Decoding:**
 - Tổng hợp Emission + Transition.
 - Tìm đường đi tối ưu.



Sơ đồ kiến trúc BiLSTM + CRF.

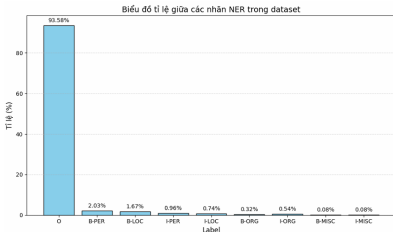
Cài đặt trong bài toán

- **Embedding:** Train from scratch (random init), không dùng pre-trained.
- **CRF:** Decode bằng Viterbi.
- **Optimizer:** Adam, $lr=0.001$.
- **Config:** epochs=12, batch_size=32, early_stop=3.

Mục lục

- 1 Tổng quan đề tài
- 2 Tổng quan các phương pháp
- 3 **Thực nghiệm**
- 4 Kết luận & Hướng phát triển

3.1. Dataset



Biểu đồ tỉ lệ giữa các nhãn NER trong dataset.

Thành phần nhận trong VLSP2016:

- **O (Outside):** 93.58%
- **B-PER:** 2.03%
- **B-LOC:** 1.67%
- **B-ORG:** 0.32%
- **Các nhãn I-/MISC khác:** $\sim 2.4\%$

Nhân xét dữ liệu

Dữ liệu **mất cân bằng nghiêm trọng**: nhãn O chiếm **93.58%** tổng token.

Thông tin dữ liệu

Nguồn: VLSP2016-NER-data (HuggingFace)

Dạng dữ liệu: file .parquet / token + nhãn IOB.

Cấu trúc: cột 1 = token, cột 2 = tag (B-, I-, O-).

Quy mô: $\sim 13.5k$ câu train, $\sim 3.4k$ câu test (300k token).

Nhận xét: gán nhãn thủ công, khá sạch nhưng có thể nhiều nhe.

Hệ thống nhãn BIO trong VLSP2016

Giá trị	Ý nghĩa	Loại thực thể
0	O – Outside	Không phải thực thể
1	B-PER (Begin of Person)	Bắt đầu tên người
2	I-PER (Inside Person)	Bên trong tên người
3	B-ORG (Begin of Org)	Bắt đầu tổ chức
4	I-ORG (Inside Org)	Bên trong tổ chức
5	B-LOC (Begin of Location)	Bắt đầu địa danh
6	I-LOC (Inside Location)	Bên trong địa danh
7	B-MISC (Begin of Misc)	Bắt đầu thực thể khác
8	I-MISC (Inside Misc)	Bên trong thực thể khác

14 / 26

15 / 26

16 / 26

3.5. Phân tích lỗi – HMM

Các lỗi NER phổ biến (Top errors)

```
>>> Top NER Errors:
O -> B-LOC: 280
B-LOC -> O: 168
O -> I-ORG: 164
O -> B-PER: 156
O -> B-ORG: 131
I-ORG -> O: 94
O -> I-LOC: 82
B-PER -> O: 79
O -> B-MISC: 69
O -> I-MISC: 60
```

Nhận xét:

- Lỗi phổ biến nhất là $O \rightarrow B\text{-}*/I\text{-}*$ và $B\text{-}*/I\text{-}* \rightarrow O$.
- Cho thấy mô hình vẫn **thiên lệch về nhãn O** dù đã tối ưu.
- Các thực thể dài (ORG, LOC) dễ bị **đứt chuỗi**.

Ví dụ câu bị gán nhãn sai

Sentence 8:		
WORD	TRUE	PRED
lại	0	0
một	0	0
lần	0	0
nữa	0	0
,	0	0
chính	0	0
bai	B-ORG	0
quản_lí	I-ORG	0
rừng_phòng_hộ	I-ORG	0
đầu	I-ORG	0
nguồn	I-ORG	0
Sêrêpôk	I-ORG	0
đó	0	0
*	0	0
thủ_của	0	0
*	0	0
cho	0	0
tự	0	0
hình_thành	0	0
một	0	0
khu	0	0
đầu_lịch	0	0
trở_lại	0	0
như_thế	0	0
...	0	0

Phân tích nguyên nhân:

- Cụm ban quản_lí rừng_phòng_hộ đầu nguồn Sêrêpôk là một thực thể **ORG** dài.
- HMM gán đúng B-ORG ở token đầu nhưng **mất toàn bộ chuỗi I-ORG** phía sau.
- Chỉ cần một token có **emission hoặc transition thấp** \Rightarrow toàn span bị gãy.
- Mô hình **không khai thác được ngữ cảnh dài** để duy trì chuỗi thực thể.

Kết luận

- Đây là **lỗi mang tính cấu trúc của HMM**. Do giả định **Markov bậc 1** và không có biểu diễn ngữ cảnh dài hạn, HMM dễ **lạm dụng các thực thể dài** như ORG.
- Các kỹ thuật tối ưu giúp cải thiện **Recall** cho nhãn hiếm, nhưng **không thể khắc phục triệt để** hạn chế này, là động lực để sử dụng các mô hình ngữ cảnh sâu hơn như **CRF** và **BiLSTM+CRF**.

3.6. Kết quả thực nghiệm – CRF (Test set)

Nhận xét tổng quan (Test set)

- **Accuracy rất cao: 0.9904.**
- **Weighted F1 (ALL): 0.9901** (bao gồm nhãn O).
- **Non-O Weighted F1: 0.9076** \Rightarrow nhận diện thực thể tốt.
- **Span-level F1: 0.9191** (đánh giá ở mức thực thể hoàn chỉnh).

Ý nghĩa:

- CRF học được **phụ thuộc nhãn toàn chuỗi**, giảm lỗi đứt span.
- Hiệu quả rõ rệt với thực thể dài (ORG, LOC).
- Vượt trội so với HMM trong bài toán NER tiếng Việt.

So sánh theo đặc trưng (Test set)

Metric	Base	+Lower	+Pre/Suf	+Shape
Accuracy	0.9563	0.9850	0.9898	0.9904
Precision	0.9505	0.9843	0.9894	0.9900
Recall	0.9563	0.9850	0.9898	0.9904
F1-score	0.9484	0.9841	0.9895	0.9901

Shape features: `is_upper`, `is_title`,
`is_digit`

Kết luận

CRF cho kết quả vượt trội nhờ mô hình hóa **phụ thuộc nhãn toàn chuỗi** và khai thác hiệu quả **đặc trưng hình thái + ngữ pháp**, đặc biệt phù hợp với NER tiếng Việt.

3.7. Phân tích lỗi – CRF

Top confusions (Test set)

Top confusions:

I-ORG → O : 26
 B-LOC → O : 21
 B-PER → O : 20
 B-ORG → O : 19
 O → I-ORG : 18
 B-LOC → I-LOC : 14
 O → B-LOC : 12
 I-LOC → B-LOC : 11
 O → B-ORG : 10
 B-PER → B-LOC : 10
 B-LOC → B-PER : 9
 I-PER → I-ORG : 8
 I-ORG → B-LOC : 8
 B-ORG → I-ORG : 7
 B-ORG → B-LOC : 7
 I-LOC → O : 7
 O → I-LOC : 7
 I-ORG → I-LOC : 7
 O → B-PER : 6
 I-PER → O : 6

Nhận xét nhanh:

- Lỗi nổi bật là **I-ORG** → **O** và **B-ORG** → **O** ⇒ **dứt span** **ORG** dài (mất nhân bên trong thực thể).
- Nhiều lỗi **B-LOC** → **O** / **I-LOC** ⇔ **B-LOC** ⇒ sai ranh giới **thực thể** (boundary) trong chuỗi LOC.
- Nhầm **PER** ⇔ **LOC** (vd: B-PER → B-LOC, B-LOC → B-PER) thường đến từ **viết hoa / token dạng tên riêng**.

Worst case (ví dụ tiêu biểu)

Worst cases:
 idx=899 nonO_errors=14

X Huyện_Duc	gold=B-LOC	pred=O
X bất	gold=B-LOC	pred=O
X An_Nhon_Tay	gold=I-LOC	pred=B-LOC
X Trung_Hoa	gold=B-LOC	pred=O
X bất	gold=I-LOC	pred=B-LOC
X Phú_Hoà_Dong	gold=I-LOC	pred=B-LOC
X bất	gold=B-LOC	pred=O
X cầu	gold=I-LOC	pred=B-LOC
✓ Bận_Khang	gold=I-LOC	pred=I-LOC
✓ huyện_Ly	gold=B-LOC	pred=B-LOC
✓ Cù_Chí	gold=I-LOC	pred=I-LOC
X căn_cứ	gold=B-LOC	pred=O
X Đông_Du	gold=I-LOC	pred=B-LOC
X sự_đoàn	gold=B-ORG	pred=O
X dù	gold=I-ORG	pred=O
X 25	gold=I-ORG	pred=O
✓ Ng	gold=B-LOC	pred=B-LOC

Giải thích nguyên nhân:

- Nhiều token **địa danh nhiều từ** bị gán B-LOC rồi các token sau lại rơi về O ⇒ **dứt chuỗi LOC** (boundary + transition yếu).
- Các token dạng **số / ký hiệu** (vd: 25, 101, 325) làm nhiễu đặc trưng, dễ kéo nhân I-ORG về O.
- Một số cụm có **từ khóa gợi ORG/LOC** (vd: UBND, tỉnh, Ủy_ban) nhưng ngữ cảnh cụ thể không đủ mạnh ⇒ dễ nhầm **ORG ⇔ LOC** hoặc **ORG → O**.

Thông điệp chính

Các đặc trưng hình thái & ngữ pháp giúp CRF nhận diện tốt hơn, nhưng vẫn khó với **thực thể dài** và **boundary LOC/ORG**.

Để giảm lỗi span dài & boundary, cần **ngữ cảnh mạnh hơn** (BiLSTM/Transformer) hoặc bổ sung **feature theo cụm từ / gazetteer** cho LOC/ORG.

3.8. Kết quả thực nghiệm – BiLSTM-CRF (Test set)

Kết quả trên Test set (BiLSTM-CRF)

[Token] Weighted F1 (ALL incl O): 0.9843

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

B-LOC	0.9092	0.7694	0.8335	1314
B-MISC	0.9259	0.9259	0.9259	54
B-ORG	0.8361	0.5752	0.6815	266
B-PER	0.9342	0.9268	0.9305	1502
I-LOC	0.8933	0.7189	0.7966	594
I-MISC	0.9091	0.8929	0.9009	56
I-ORG	0.8420	0.6143	0.7103	477
I-PER	0.9357	0.8829	0.9086	692
O	0.9896	0.9979	0.9937	69944

accuracy			0.9851	74899
macro avg	0.9083	0.8116	0.8535	74899
weighted avg	0.9843	0.9851	0.9843	74899

[Token] Weighted F1 (Non-O only): 0.8642

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

B-LOC	0.9405	0.7694	0.8464	1314
B-MISC	1.0000	0.9259	0.9615	54
B-ORG	0.9217	0.5752	0.7083	266
B-PER	0.9574	0.9268	0.9418	1502
I-LOC	0.9242	0.7189	0.8087	594
I-MISC	0.9259	0.8929	0.9091	56
I-ORG	0.9099	0.6143	0.7334	477
I-PER	0.9517	0.8829	0.9160	692

micro avg	0.9437	0.8046	0.8686	4955
macro avg	0.9414	0.7883	0.8532	4955
weighted avg	0.9417	0.8046	0.8642	4955

[Span] P=0.9253 R=0.8450 F1=0.8834 | TP=409 FP=33 FN=75

Tóm tắt chỉ số (Test set)

Nhận xét nhanh

- Token-level (ALL) cao do nhãn O chi phối \Rightarrow **Weighted F1 = 0.9843**.
- Khi chỉ xét **Non-O**, F1 giảm còn **0.8642** \Rightarrow khó ở thực thể hiếm/dài.
- Lỗi phổ biến: bỏ sót thực thể (dự đoán về O), đặc biệt **LOC/ORG**.

Token-level

- ALL incl O**: Weighted F1 = **0.9843**, Acc = **0.9851**, Macro F1 = **0.8535**
- Non-O only**: Weighted F1 = **0.8642**, Micro F1 = **0.8686**, Macro F1 = **0.8532**

Span-level (Entity-level)

P=**0.9253** | R=**0.8450** | F1=**0.8834**
TP=409, FP=33, FN=75

Kết luận

BiLSTM-CRF cải thiện biểu diễn ngữ cảnh so với HMM, nhưng trong đồ án này **CRF vẫn tốt hơn** (Non-O F1 và Span-F1 cao hơn) \Rightarrow đặc trưng thủ công + CRF phù hợp dữ liệu hiện tại.

3.9. Phân tích lỗi – BiLSTM-CRF

Top confusions (Test set)

```

Top confusions:
B-LOC -> O : 238
I-ORG -> O : 137
B-ORG -> O : 101
I-LOC -> O : 92
B-PER -> O : 91
I-PER -> O : 63
O -> B-LOC : 37
O -> B-PER : 36
B-LOC -> B-PER : 30
O -> I-ORG : 26
I-LOC -> B-LOC : 24
I-ORG -> B-LOC : 20
I-LOC -> I-PER : 20
O -> B-ORG : 17
O -> I-LOC : 16
B-LOC -> I-LOC : 16
I-LOC -> I-ORG : 15
I-ORG -> I-LOC : 14
I-LOC -> B-PER : 14
I-PER -> B-PER : 11

```

Nhận xét:

- Lỗi lớn nhất: **B-LOC** → **O** (238), **I-ORG** → **O** (137).
- Mô hình thường **bỏ sót thực thể** (FN tăng) ⇒ Recall span giảm.
- Có hiện tượng **nhầm ranh giới BIO**: I-LOC ↔ B-LOC, B-LOC → B-PER.

Ý nghĩa

Sai lệch chủ yếu đến từ **entity dài** và **tên riêng/viết tắt/ngoại ngữ** khiến emission không ổn định ⇒ dễ rơi về nhãn O.
⇒ **Kết luận**: BiLSTM-CRF học ngữ cảnh tốt hơn HMM, nhưng vẫn cần **feature/embedding mạnh** hơn hoặc **pretrained encoder** (PhoBERT/XLM-R) để xử lý entity dài và tên riêng đa dạng.

Ví dụ worst case (Test set)

```

Worst cases:
idx=2662 nonO_errors=14
✓ Jane                gold=B-PER      pred=B-PER
✓ Barton              gold=I-PER      pred=I-PER
X quản lý             gold=O          pred=B-ORG
X Bệnh viện           gold=B-LOC      pred=I-ORG
X Washington          gold=I-LOC      pred=I-ORG
X ;                   gold=O          pred=I-ORG
X Sophie              gold=B-PER      pred=I-ORG
X Quinn              gold=I-PER      pred=I-ORG
✓ VN                  gold=B-LOC      pred=B-LOC
✓ Trung tâm           gold=B-ORG      pred=B-ORG
✓ Nghiên cứu          gold=I-ORG      pred=I-ORG
✓ vẽ                  gold=I-ORG      pred=I-ORG
✓ triết học           gold=I-ORG      pred=I-ORG
X ,                   gold=I-ORG      pred=O
X văn hoá             gold=I-ORG      pred=O
X và                  gold=I-ORG      pred=O
X xã hội              gold=I-ORG      pred=O
X VN                  gold=I-ORG      pred=B-LOC
X Trường              gold=B-ORG      pred=B-LOC
X DH                  gold=I-ORG      pred=I-LOC
X Temple              gold=I-ORG      pred=O
✓ Nguyễn              gold=B-PER      pred=B-PER
✓ Thị                 gold=I-PER      pred=I-PER
✓ Mai                 gold=I-PER      pred=I-PER
✓ Quảng Nam           gold=B-LOC      pred=B-LOC

```

Giải thích nguyên nhân:

- Câu chứa nhiều thực thể dài/đan xen (ORG-LOC-PER) ⇒ khó giữ ranh giới span.
- Một vài token “nhiều” (dấu câu, viết tắt như DH, VN, tên ngoại như Temple) làm CRF layer khó nối chuỗi I-*
- Khi **một token bị gán O**, cả thực thể dài có thể bị **đứt chuỗi** ⇒ giảm F1 span.

3.10. So sánh tổng quan các mô hình

Nhận xét so sánh mô hình

- **CRF tốt nhất (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):** Macro-F1 **0.51** → **0.72**, nhưng vẫn kém do hạn chế Markov.
- **Token F1 (ALL)** cao vì 0 nhiều ⇒ ưu tiên **Non-O F1 & Span-F1**.

Kết luận ngắn

- **CRF phù hợp nhất** với dữ liệu hiện tại (feature thủ công + ràng buộc chuỗi).
- **BiLSTM+CRF** sẽ mạnh hơn nếu có **data** lớn hơn / **embedding pretrained**.

Bảng so sánh hiệu suất (Test set)

Chỉ số	HMM	CRF	BiLSTM
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

*Ghi chú: HMM có Macro/Weighted/Acc;
CRF/BiLSTM-CRF có thêm Non-O & Span-F1.*

Ý nghĩa chỉ số

- **Token F1 (ALL):** dễ “ảo” vì 0 nhiều.
- **Non-O F1:** chất lượng nhận diện thực thể (bỏ 0).
- **Span F1:** đúng thực thể hoàn chỉnh (khắc khe nhất).

24 / 26

Tài liệu tham khảo



VLSP 2016. *Vietnamese Named Entity Recognition Shared Task Dataset*. (HuggingFace).



Ngoc C. Le, Ngoc-Yen Nguyen, Anh-Duong Trinh. *On the Vietnamese Named Entity Recognition: A Deep Learning Method Approach*. In: IEEE RIVF, 2020.
<https://arxiv.org/pdf/1912.01109>



Daniel Jurafsky, James H. Martin. *Speech and Language Processing (3rd ed. draft)*, 2023.
https://web.stanford.edu/~jurafsky/slp3/ed3book_aug25.pdf

Cảm ơn!

Xin cảm ơn cô Nguyễn Thị Quý và các bạn!

Hỏi đáp — Demo