

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

**Đề án môn Ngôn ngữ học ngữ liệu:**

**Tìm hiểu quy trình phát triển dữ liệu cho bài toán  
“Nhận diện thực thể tên riêng tiếng Việt”**

**Sinh viên thực hiện:** Nhóm 3 – CS321.M11.KHCL

1. Trần Tuấn Vỹ - 18520406
2. Huỳnh Trọng Khoa - 18520918
3. Nguyễn Tấn Phúc - 18521259

# NỘI DUNG TRÌNH BÀY

---

1. Giới thiệu đề tài
2. Giai đoạn chuẩn bị
3. Giai đoạn thực hiện
4. Đánh giá





# GIỚI THIỆU ĐỀ TÀI

# Giới thiệu đề tài

## I. PHÁT BIỂU BÀI TOÁN

## ĐỊNH NGHĨA

### Nhận diện thực thể tên riêng (Named entity Recognition)

**Nhận diện** các từ, cụm từ trong văn bản và **phân loại** chúng vào trong các **nhóm** đã được định trước.

### Ví dụ

Nhận dạng thực thể cho câu "Bệnh nhân 44 là bác sĩ" với các loại thực thể "PATIENT\_ID" (mã số bệnh nhân), "JOB" (nghề nghiệp).



# Giới thiệu đề tài

## I. PHÁT BIỂU BÀI TOÁN

## ĐỊNH NGHĨA

### Nhận diện thực thể tên riêng (Named entity Recognition)

**Nhận diện** các từ, cụm từ trong văn bản và **phân loại** chúng vào trong các **nhóm** đã được định trước.

### Mục tiêu

Một trong những tác vụ tiền đề trong việc xây dựng các ứng dụng Truy xuất thông tin, Hệ thống hỏi-đáp, Chatbox, ...



# Giới thiệu đề tài

## II. THÔNG TIN VỀ TẬP DỮ LIỆU

## NGUỒN DỮ LIỆU

### Bài báo tham khảo

**“COVID-19 Named Entity Recognition for Vietnamese”** - Nhận dạng thực thể tên riêng Tiếng Việt cho nội dung liên quan đến COVID-19.

### Tác giả

Nhóm nghiên cứu thuộc VinAI Research, Ha Noi, Vietnam



# Giới thiệu đề tài

## II. THÔNG TIN VỀ TẬP DỮ LIỆU

Số lượng dữ liệu

Bao gồm: 10,037 câu

Tỉ lệ train/dev/test

5/2/3

Entity Type	Train	Valid	Test	All
PATIENT_ID	3,240	1,276	2,005	6,521
NAME	349	188	318	855
AGE	682	361	582	1,625
GENDER	542	277	462	1,281
JOB	205	132	173	510
LOCATION	5,398	2,737	4,441	12,576
ORGANIZATION	1,137	551	771	2,459
SYMPTOM&DISEASE	1,439	766	1,136	3,341
TRANSPORTATION	226	87	193	506
DATE	2,549	1,103	1,654	5,306
#Entities in total	15,767	7,478	11,735	34,984
#Sentences in total	5,027	2,000	3,000	10,027

*Bảng 1. Bảng thống kê bộ dữ liệu*



# Giới thiệu đề tài

## II. THÔNG TIN VỀ TẬP DỮ LIỆU

#COVID-19 #COVID

THANH NIÊN

BÁO MƠI.com

VNEXPRESS  
TIN NHANH VIETNAM

FEB 2020 – AUG 2020

Entity Type	Train	Valid	Test	All
PATIENT_ID	3,240	1,276	2,005	6,521
NAME	349	188	318	855
AGE	682	361	582	1,625
GENDER	542	277	462	1,281
JOB	205	132	173	510
LOCATION	5,398	2,737	4,441	12,576
ORGANIZATION	1,137	551	771	2,459
SYMPTOM&DISEASE	1,439	766	1,136	3,341
TRANSPORTATION	226	87	193	506
DATE	2,549	1,103	1,654	5,306
#Entities in total	15,767	7,478	11,735	34,984
#Sentences in total	5,027	2,000	3,000	10,027

*Bảng 1. Bảng thống kê bộ dữ liệu*





# Giới thiệu đề tài

## II. THÔNG TIN VỀ TẬP DỮ LIỆU

## CÁC LOẠI THỰC THỂ

STT	Nhãn thực thể	Định nghĩa
1	PATIENT_ID	Mỗi bệnh nhân COVID-19 ở Việt Nam có một mã định danh "X" (bệnh nhân thứ X)
2	NAME	Tên bệnh nhân hoặc người có liên quan đến bệnh nhân.
3	AGE	Tuổi của bệnh nhân hoặc người có liên quan đến bệnh nhân.
4	GENDER	Giới tính của bệnh nhân hoặc người có liên quan đến bệnh nhân.
5	JOB	Nghề nghiệp của bệnh nhân hoặc người có liên quan đến bệnh nhân.

**Bảng 2. Định nghĩa nhãn thực thể (1)**



# Giới thiệu đề tài

## II. THÔNG TIN VỀ TẬP DỮ LIỆU

## CÁC LOẠI THỰC THỂ

STT	Nhãn thực thể	Định nghĩa
6	LOCATION	Địa điểm/nơi ở của bệnh nhân hoặc từng đến.
7	ORGANIZATION	Các tổ chức liên quan đến bệnh nhân: Công ty, Tổ chức Chính phủ, ...
8	TRANSPORTATION	Phương tiện giao thông mà bệnh nhân sử dụng (Chỉ gán số hiệu, biển số, .. )
9	DATE	Tất cả các ngày có trong câu
10	SYMPTOM&DISEASE	Triệu chứng và bệnh mà bệnh nhân gặp phải

**Bảng 3. Định nghĩa nhãn thực thể (2)**



# Giới thiệu đề tài

## II. THÔNG TIN VỀ TẬP DỮ LIỆU

## ĐỊNH DẠNG DỮ LIỆU

### Định dạng CoNLL

**“Conference on Natural Language Learning”**: định dạng dữ liệu phân cụm.

### Mô tả tập dữ liệu

- 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 hai cột, các cột được cách nhau bởi một khoảng trắng.



# Giới thiệu đề tài

## II. THÔNG TIN VỀ TẬP DỮ LIỆU

## ĐỊNH DẠNG DỮ LIỆU

### Ví dụ

Bệnh_nhân	O
44	B-PATIENT-ID
Là	O
bác_sĩ	B-JOB

### Cấu trúc BIO

Với mỗi thực thể bao gồm nhiều từ tạo thành (T):

- gán nhãn từ bắt đầu của (T) là **B-T**
- các từ tiếp theo là **I-T**.

Không có nhãn: gán 'O'





**GIAI ĐOẠN  
CHUẨN BỊ!**

# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC CHUNG

**1** Thực thể phải được định danh rõ ràng.

“Khoa Khám Bệnh, Bệnh viện Bạch Mai”, “Số 127 đường Trần Hưng Đạo”

**2** Gán nhãn các câu liên quan đến bệnh nhân COVID-19 và liên quan đến tình hình dịch ở Việt Nam

**Các trường hợp như:** công bố ca bệnh mới, lịch trình di chuyển, triệu chứng, bệnh nền, công bố tử vong, ...



# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN **PATIENT\_ID**

**Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự (1, 2, 3, 100, 200, ...).**

- Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X  
→ chỉ gán "X" với nhãn **PATIENT\_ID**
- BNX → gán nhãn cho "BNX" là **PATIENT\_ID**



# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN **AGE**

1

Chỉ gán nhãn giá trị tuổi của bệnh nhân và những người liên quan (tiếp xúc)

2

Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên hoặc có mã bệnh nhân)

"Bệnh nhân 100, nam, **55** tuổi, địa chỉ ở quận 8, TP HCM." => gán nhãn "55" là thực thể kiểu AGE





# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN **NAME**

1

### Tên bệnh nhân

"N.H.N", "N.H.N.", "T.", "T" (chú ý phân biệt dấu chấm trong tên bệnh nhân và các dấu chấm cuối câu)

2

### Tên những người có liên quan hay tiếp xúc với bệnh nhân

"Ngày 21/8, bệnh nhân có tiếp xúc với cô A., là nhân viên siêu thị Điện máy Xanh Đà Nẵng". → gán nhãn "A."



# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN **GENDER**

1

Chỉ gán nhãn giới tính của bệnh nhân và những người liên quan trực tiếp (tiếp xúc) với bệnh nhân.

2

Trong một câu, giới tính cần được gán với một đối tượng được định danh (có tên hoặc có mã bệnh nhân)

"Bệnh nhân 125 là **nữ**, quốc tịch Nam Phi, 22 tuổi, trú tại quận 7, TP.HCM." => gán nhãn "nữ" là thực thể nhãn GENDER.



# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN **JOB**

1

**Chỉ gán nhãn nghề nghiệp của bệnh nhân và các cá nhân có liên quan trực tiếp (tiếp xúc, gặp mặt, ở gần).**

2

**Ngoài ra, những từ chỉ nghề nghiệp cần phải được gán với 1 cá nhân nhất định trong câu (có tên hoặc có mã bệnh nhân).**

""Bệnh nhân 35" là nhân viên bán hàng tại Siêu thị Điện máy Xanh ở quận Hải Châu, Đà Nẵng."  
→ gán nhãn "nhân viên bán hàng" là thực thể kiểu JOB do là nghề nghiệp của bệnh nhân 35.



# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN SYMPTOM\_AND\_DISEASE

1

### Triệu chứng liên quan tới bệnh nhân COVID-19

"Khoảng 4-5 ngày sau tôi sốt, ho, đau họng dữ dội, xét nghiệm dương tính", bệnh nhân chia sẻ."

2

### Các loại bệnh khác mà bệnh nhân COVID-19 mắc phải

"Bệnh nhân 737" ... tiền sử suy thận mạn giai đoạn cuối đã chạy thận nhân tạo chu kỳ và đặt stent, suy tim, tăng huyết áp."



# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN **LOCATION**

1

**Tên Châu lục, tên Quốc gia**

2

**Tên đơn vị hành chính của quốc gia.**

"tỉnh Hải Dương", "thành phố Hà Nội", "quận Hoàn Kiếm" (gán nhãn cả các từ chỉ đơn vị hành chính: tỉnh, thành phố, quận, huyện, đường)

3

**Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng**



# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN ORGANIZATION

1

**Tên các cơ quan chính phủ: bộ ngành, uỷ ban nhân dân**

"UBND Hà Nội", "Sở Y Tế Quảng Ninh".

2

**Tên các cơ quan liên quan tới việc xử lý dịch tễ**

"Cơ quan kiểm soát dịch bệnh Hà Nội", "HCDC", "Viện Pasteur TP Hồ Chí Minh". .

3

**Tên các công ty, tổ chức nơi bệnh nhân làm việc**



# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN **TRANSPORTATION**

**Chỉ gán nhãn biển số, số hiệu của loại phương tiện di chuyển, không gán nhãn loại phương tiện di chuyển.**

"Bệnh nhân 110", nữ, 19 tuổi, ở Đống Đa, Hà Nội, là du học sinh tại Mỹ, quá cảnh tại Nhật Bản, về Hà Nội trên chuyến bay **JL751**, số ghế 1A ngày 19/3.

Sau khi xuống sân bay, 15h30 cùng ngày, T. đi xe khách Công Tạo mang biển số **51B-142.48** về huyện Bình Đại



# Giai đoạn chuẩn bị

## I. THÔNG TIN VỀ GUIDELINE

## QUY TẮC GÁN NHÃN **DATE**

1

**Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)  
(Gán nhãn cả năm nếu có)**

Sau khi trở về địa phương, bệnh nhân 61 từng làm chủ hôn hai đám cưới và dự thánh lễ tại xã Phước Nam trước khi nhập viện điều trị vào ngày **15-3**.

2

**Các trường hợp chỉ khoảng thời gian: gán nhãn riêng biệt ngày bắt đầu và kết thúc.**

“Từ ngày **1-6/8**, bệnh nhân được chuyển lên cách ly và ... ”.  
→ 2 entities kiểu **DATE** là : "1" và "6/8"





# Giai đoạn chuẩn bị

## II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

## MÔ HÌNH MÁY HỌC

### CONDITIONAL RANDOM FIELDS (CRFs)

CRFs là một mô hình phân lớp sử dụng ngữ cảnh tham gia vào quá trình gán nhãn. Thường được sử dụng trong các bài toán gán nhãn từ loại, nhận diện thực thể tên riêng trong lĩnh vực Xử lý ngôn ngữ tự nhiên.

### Thư viện hỗ trợ

```
import sklearn_crfsuite  
crf = sklearn_crfsuite.CRF()
```



# Quy trình chuẩn bị

## II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

## CHUẨN BỊ DỮ LIỆU

### Clone dữ liệu từ github

```
!git clone https://github.com/VinAIResearch/PhoNER\_COVID19/
```

### Sử dụng tập Train, Test

```
dir_train_data = 'PhoNER_COVID19/data/word/train_word.conll'  
dir_test_data = 'PhoNER_COVID19/data/word/test_word.conll'
```



# Quy trình chuẩn bị

## II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

## XỬ LÝ DỮ LIỆU

### Input format (X)

```
X_format = [  
    ['Đồng_thời', ',', 'bệnh_viện', 'tiếp_tục', 'thực_hiện', 'các', ...],  
    ['Bác_sĩ', 'căn_cứ', 'vào', 'triệu_chứng', 'lâm_sàng', ... ], ...  
]
```

### Output format (Y)

```
y_format = [  
    ['O', 'O', 'O', 'O', 'O', 'O', ...],  
    ['O', 'O', 'O', 'O', 'O', ...], ...  
]
```



# Quy trình chuẩn bị

## II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

## HUẤN LUYỆN MÔ HÌNH VÀ ĐÁNH GIÁ

### Huấn luyện mô hình

```
crf = CRF()  
crf.fit(X_train,y_train)  
y_pred = crf.predict(X_test)
```

### Đánh giá mô hình dựa trên F1-SCORE

	PAT.	NAM	AGE	GEN.	JOB.	LOC.	ORG.	SYM.	TRA.	DAT.	Mic-F1	Mac-F1
F1-Score	0.76	0.18	0.06	0.56	0.29	0.67	0.61	0.23	0.77	0.92	<b>0.68</b>	<b>0.64</b>



# Quy trình chuẩn bị

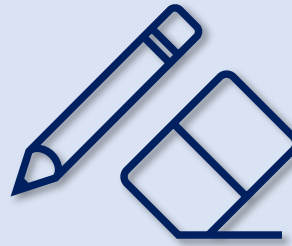
## II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

## GIAO DIỆN CÔNG CỤ

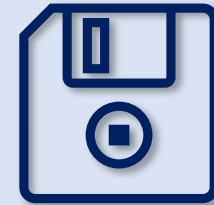
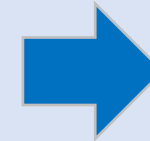
### Các tính năng cơ bản



Chọn câu



Chỉnh sửa



Lưu kết quả



Google Colaboratory



Google Drive



# Quy trình chuẩn bị

II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

GIAO DIỆN CÔNG CỤ

Chọn câu để gán nhãn

sentence: " 1

Show code

Khoảng 17h30 ngày 16 - 5 , ông V. đi ngang qua khu cách_ly bệnh_viện thì ông H.H.T. ( 33 tuổi , ngụ tỉnh Hải_Dương ) là		
	word	ner
0	Khoảng	O
1	17h30	B-PATIENT_ID
2	ngày	O
3	16	B-DATE
4	-	I-DATE
5	5	I-DATE
6	,	O



# Quy trình chuẩn bị

II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

GIAO DIỆN CÔNG CỤ

### Điều chỉnh nhãn

Index: " 14

Pre: I-

Ner: NAME

O\_label: ☐

Show code

Đã gán từ Hà\_Tĩnh (vị trí 14) với nhãn I-LOCATION

	word	ner
0	Hiện	<input type="radio"/>
1	bệnh_nhân	<input type="radio"/>
2	đang	<input type="radio"/>
3	được	<input type="radio"/>



# Quy trình chuẩn bị

II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

GIAO DIỆN CÔNG CỤ



Lưu kết quả

Round: 1

Annotator\_ID: 1

Show code

Saved

	annotator_1_sentence_47.txt 	me	2:45 PM	88 bytes
---	---	----	---------	----------






# Giai đoạn chuẩn bị


II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ


GIAO DIỆN CÔNG CỤ


Folders


Name ↑


 Data


 Report


 SET 1


 SET 2


 SET 3


 SET 4


 SET 5


 SET 6


 SET 7

 SET 8

 SET 9

 SET 10

 SET 10 (Bán tự động)

 Set 10 (Thủ công)



# Giai đoạn chuẩn bị

## II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

## ĐÁNH GIÁ CÔNG CỤ

### So sánh quy trình gán nhãn thủ công và bán tự động:

- Sử dụng 5 câu của set 10 để thực hiện gán nhãn bằng 2 phương pháp.
- Đo thời gian thực hiện gán nhãn.
- So sánh thời gian thực hiện trung bình của 2 phương pháp.



# Giai đoạn chuẩn bị

II. XÂY DỰNG CÔNG CỤ GÁN NHÃN THỰC THỂ

ĐÁNH GIÁ CÔNG CỤ

**Bảng so sánh thời gian gán nhãn giữa hai phương pháp:**

	thủ công		Bán tự động	
Câu	Annotator 1	Annotator 2	Annotator 1	Annotator 2
1	64	191	41	31
2	148	171	36	41
3	81	133	48	46
4	42	173	52	85
5	38	81	31	24
Tổng Thời gian	6'13s	12'29s	3'28s	3'47s





# **GIẢI ĐOẠN THỰC HIỆN**

# Quy trình thực hiện

## I. TỔ CHỨC THỰC HIỆN GÁN NHÃN

### Dùng 50 câu trong tập Test

- Annotator chỉ biết được nội dung của câu
- Nhãn của câu sẽ được dùng làm **nhãn gold** để đo **độ chính xác**.

### Quá trình huấn luyện anntotator

- Đọc guideline
- Thực hiện gán nhãn bằng tool (2 annotator)
- Đo độ chính xác, đồng thuận
- Thảo luận



# Quy trình thực hiện

## I. TỔ CHỨC THỰC HIỆN GÁN NHÃN

## ĐỘ CHÍNH XÁC, ĐỘ ĐỒNG THUẬN

### ĐỘ ĐO F1-SCORE

$$F1 - Score = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

### Thư viện hỗ trợ

```
from segeval.metrics import f1_score  
a = f1_score(y_gold,y_annotator_1)  
b = f1_score(y_gold,y_annotator_2)  
c = f1_score(y_annotator_1,y_annotator_2)
```

### Độ chính xác, độ đồng thuận

**Độ chính xác** :tính thông qua nhãn gold và nhãn của từng annotator gán

**Độ đồng thuận**: tính thông qua nhãn do 2 annotator gán.



# Quy trình thực hiện

## I. TỔ CHỨC THỰC HIỆN GÁN NHÃN

## THẢO LUẬN

### Bước 1

Khi 2 annotator hoàn thành trong mỗi vòng , tiến hành tính độ chính xác, độ đồng thuận.

### Bước 2

Tìm hiểu nguyên nhân gán sai của từng annotator.

### Bước 3

Đưa ra hướng giải quyết cho từng trường hợp cụ thể: Đọc lại guideline, sửa nhãn gold cho đúng, thêm ghi chú vào guideline...





**ĐÁNH GIÁ**



# Đánh giá

## ĐỘ CHÍNH XÁC, ĐỘ ĐỒNG THUẬN

SET		1	2	3	4	5
Độ chính xác	Annotator 1	93.48%	97.14%	87.5%	100%	84.85%
	Annotator 2	91.3%	97.14%	91.43%	100%	94.12%
Độ đồng thuận		97.87%	100%	85.71%	100%	91.42%
SET		6	7	8	9	10
Độ chính xác	Annotator 1	85.71%	90.47%	93.1%	91.42%	91.43%
	Annotator 2	96.55%	100%	91.52%	100%	91.43%
Độ đồng thuận		88.88%	90.47%	91.22%	91.42%	100%

Bảng kết quả đo độ chính xác, độ đồng thuận trong 10 vòng



# Đánh giá

## PHÂN TÍCH MỘT SỐ KẾT QUẢ TRONG QUÁ TRÌNH THỰC HIỆN GÁN NHÃN

SET 1

### Câu gán gán sai:

“Tính đến ngày 30 - 7 , Việt\_Nam có thêm một\_số ca bệnh nặng đang điều\_trị tại các bệnh\_viện Đà\_Nẵng , ...

Nhãn gold: (Đà\_nẵng, LOCATION)

Nhãn 2 annotator gán: LOCATION

Nguyên nhân: Hiểu nhầm guideline (*Trong trường hợp này vì có từ “các” nên không xác định được đối tượng cụ thể*).



# Đánh giá

PHÂN TÍCH MỘT SỐ KẾT QUẢ TRONG QUÁ TRÌNH THỰC HIỆN GÁN NHÃN

SET 2

## Câu gán gán sai:

“Tuy\_nhiên , bệnh\_nhân chỉ được phép duy\_trì tư\_thể này trong vòng 16 tiếng mỗi ngày , nếu không sẽ bị **loét điểm tì** . ”

Nhãn gold: ○

Nhãn 2 annotator gán: SYMPTOM\_AND\_DISEASE

Nguyên nhân: Hiểu nhầm nghĩa của từ (do bệnh này không liên quan đến triệu chứng, di chứng hay hậu di chứng liên quan đến COVID-19).



# Đánh giá

## PHÂN TÍCH MỘT SỐ KẾT QUẢ TRONG QUÁ TRÌNH THỰC HIỆN GÁN NHÃN

SET 3

### Câu gán gán sai:

Lần thứ 4 hiến máu tại bệnh\_viện , **nữ hộ\_sinh** Nguyễn\_Việt\_Dung , khoa Phụ ngoại , cho biết lần này có khác\_biệt khi chị và các đồng\_nghiep đều phải đeo khẩu\_trang đi hiến máu . ”

Nhãn gold: ○ ○

Nhãn 2 annotator gán: Annotator 1 (B-GENDER B-JOB), Annotator 2 (B-JOB I-JOB)

### Nguyên nhân:

- Chưa đọc kỹ toàn bộ câu (*Chủ thể không liên quan đến bệnh nhân COVID-19*).



# Đánh giá

PHÂN TÍCH MỘT SỐ KẾT QUẢ TRONG QUÁ TRÌNH THỰC HIỆN GÁN NHÃN

SET 8

## Câu gán gán sai:

**UBND TP Biên\_Hoà , Đồng\_Nai** dỡ bỏ cách\_ly tuyến đường Hồ\_Văn\_Đại , nơi cư\_trú của hai bệnh\_nhân 595 và 669 sau 14 ngày phong\_toả .

Nhãn gold: ORGANIZATION, LOCATION

Nhãn annotator gán: ORGANIZATION

Nguyên nhân: Guideline chưa đề cập.



# Đánh giá

## PHÂN TÍCH MỘT SỐ KẾT QUẢ TRONG QUÁ TRÌNH THỰC HIỆN GÁN NHÃN

## NHẬN XÉT CHUNG

### 1. Cần đọc kỹ câu để nắm rõ ngữ cảnh.

**Ví dụ:** Bốn người "ngoài **VN0054**" gồm N.T.T, nữ 24 tuổi từ Anh về nước ngày 9-3 trên máy bay thuê riêng, người nhà và lái xe của bệnh nhân số 17, nam bệnh nhân 27 tuổi đi từ Hàn Quốc về trên chuyến bay VJ981 ngày 4-3."

"VN0054" là số hiệu chuyến bay nhưng đi kèm với từ phủ định "ngoài" nên trong câu này, "VN0054" không phải là phương tiện di chuyển của bệnh nhân

→ không gán nhãn "**VN0054**" là **TRANSPORTATION**



# Đánh giá

PHÂN TÍCH MỘT SỐ KẾT QUẢ TRONG QUÁ TRÌNH THỰC HIỆN GÁN NHÃN

NHẬN XÉT CHUNG

## 2. Xác định rõ thực thể LOCATION và ORGANIZATION:

**Ví dụ:** Trước đó, vào sáng 19-9, **Bệnh viện Phổi Đà Nẵng** đã chuyển hai bệnh nhân COVID-19 đã được công bố khỏi bệnh viện trước đó nhiều ngày là bệnh nhân số 416 và BN 478 về **Bệnh viện Đà Nẵng** để tiếp tục điều trị bệnh nền.

"Bệnh viện Phổi Đà Nẵng" đóng vai trò là chủ ngữ của câu, thực hiện hành động là chuyển hai bệnh nhân => gán nhãn **ORGANIZATION**.

"Bệnh viện Đà Nẵng" là nơi tiếp nhận bệnh nhân, được dùng như một địa điểm => gán nhãn **LOCATION**



# Đánh giá

PHÂN TÍCH MỘT SỐ KẾT QUẢ TRONG QUÁ TRÌNH THỰC HIỆN GÁN NHÃN

NHẬN XÉT CHUNG

## 3. Thực thể định danh rõ ràng

**Ví dụ:** **Lãnh đạo huyện Chương Mỹ** cho biết trên địa bàn huyện có một người tiếp xúc với bệnh nhân thứ 17, đó là chị D.T.T, đã tiếp xúc với bệnh nhân ngày 2 và ngày 4-3.

"Lãnh đạo huyện Chương Mỹ" chưa được định danh rõ ràng => chỉ gán "huyện Chương Mỹ" là **LOCATION**







**THANK YOU!**