ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

ĐOÀN NHẬT SANG	**21522542**
TRƯƠNG VĂN KHẢI	**21520274**
LÊ NGÔ MINH ĐỨC	**21520195**

TÊN ĐỀ TÀI

NHẬN DẠNG THỰC THỂ COVID-19 CHO TIẾNG VIỆT

MÔN HỌC

CS221: XỬ LÝ NGÔN NGỮ TỰ NHIÊN

TP. Hồ Chí Minh, 11/2023

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

TÊN ĐỀ TÀI

NHẬN DẠNG THỰC THỂ COVID-19 CHO TIẾNG VIỆT

MÔN HỌC

CS221: XỬ LÝ NGÔN NGỮ TỰ NHIÊN

GIẢNG VIÊN HƯỚNG DẪN

GV LT: Nguyễn Trọng Chỉnh

GV HDTH: Đặng Văn Thìn

GV HDTH: Nguyễn Đức Vũ

TP. Hồ Chí Minh, 11/2023

`

LỜI CẢM ƠN

Chúng tôi xin gửi lời cảm ơn tới các thầy cô trường đại học Công Nghệ Thông Tin, Đại học Quốc Gia TP.Hồ Chí Minh đã tận tình giảng dạy và truyền đạt kiến thức trong suốt khóa học vừa qua. Chúng tôi cũng xin được gửi lời cảm ơn đến các thầy cô trong khoa Khoa Học Máy Tính đã mang lại cho tôi những kiến thức vô cùng quý báu và bổ ích trong quá trình học tập tại trường.

Đặc biệt xin chân thành cảm ơn thầy giáo, TS. Nguyễn Trọng Chỉnh, thầy Nguyễn Đức Vũ, thầy Đặng Văn Thìn. Các thầy đã là người đã định hướng, giúp đỡ, trực tiếp hướng dẫn và tận tình chỉ bảo trong suốt môn học Xử Lý Ngôn Ngữ Tự Nhiên để tụi em có đủ kiến thức thực hiện đồ án cuối kỳ này.

Hồ Chí Minh, ngày 27 tháng 12 năm 2023

`

MỤC LỤC

`

`

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

Từ viết tắt	Từ chuẩn	Diễn giải
NER	Named Entity Recognition	Nhận dạng thực thể
VLSP	Vietnamese Language and Speech Processing	Tổ chức Xử lý Ngôn ngữ và Tiếng nói tiếng Việt
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
MHA	Multi-head Attention	Kĩ thuật attention nhiều head
MHSA	Multi-head Self Attention	Kỹ thuật self-attention nhiều head
FFN	Feed Foward Network	Mạng truyền thẳng
GELU	Gaussian Error Linear Unit	Đơn vị Tuyến tính với độ lỗi Gaussian
CLS	Classification	Token phân loại
SEP	Separator	Token ngăn cách
GPU	Graphics Processing Unit	Đơn vị xử lý đồ họa
RAM	Random Access Memory	Bộ nhớ truy xuất ngẫu nhiên

`

DANH MỤC BẢNG

`

DANH MỤC HÌNH VẼ VÀ ĐỒ THỊ

`

MỞ ĐẦU

Những năm gần đây, chứng kiến sự tăng trưởng vượt trội của mạng Internet, cũng như các mạng xã hội như Facebook, Zalo, Instagram,.... hay các công cụ tìm kiếm như Google đã có sự gia tăng khủng kiếp về số lượng người dùng. Điều này dẫn đến lượng thông tin được tạo ra trên mạng Internet từng giây ngày càng nhiều.

Các dạng dữ liệu này thường nằm ở dạng phi cấu trúc. Để chúng ta có thể sử dụng chúng có ý nghĩa và hiệu quả hơn, cần phải chuyển đổi chúng từ dạng phi cấu trúc thành dạng có cấu trúc đã định sẵn. Đây chính là mục tiêu của bài toán Nhận Dạng Thực Thể hay còn được gọi là NER (Named Entity Recognition).

Đối với dữ liệu Tiếng Việt, bài toán NER cũng được cộng đồng Xử Lý Ngôn Ngữ rất quan tâm và nghiên cứu thể hiện qua các bộ dữ liệu như VLSP 2016, VLSP 2018. Tuy nhiên, các bộ dữ liệu Tiếng Việt trên chủ đề này vẫn còn rất nhỏ và vẫn đang trong quá trình phát triển, vì vậy đây cũng là một phần lý do mà kết quả của các công trình nghiên cứu trước đó vẫn chưa được tốt.

Đó đều là những thách thức và cũng như là động lực để chúng tôi xây dựng một mô hình NER trên bộ dữ liệu mới nhất của VINAI là PHONER COVID19 được công bố tại hội nghị NAACL 2021.

Bài toán chính trong đồ án của chúng tôi có thể được phát biểu như sau:

1. **Đầu vào:** là một văn bản Tiếng Việt
2. **Đầu ra:** là câu văn bản Tiếng Việt đã được thêm các thực thể định danh cùng với loại định danh tương ứng

`

CHƯƠNG 1: GIỚI THIỆU BÀI TOÁN NHẬN DẠNG THỰC THỂ

Thực thể định danh, nói một cách đơn giản nó có thể chỉ về một thứ gì bất kì mà được gán với một cái tên thích hợp: Con người, Địa điểm, Tuổi, … **Bài Toán Nhận Dạng Thực Thể** đã được định danh là đi nhận dạng và phân loại đã được xác định theo một bộ các quy tắc xác định từ trước và quy tắc gán nhãn rõ ràng. Ví dụ cụ thể trong bộ dữ liệu NER VLSP 2016 có định nghĩa ra 4 thực thể xác định trước là: tên người (**PERSON - PER**), tên tổ chức (**ORGANIZATION - ORG**), tên địa điểm (**LOCATION - LOC**), tên các loại khác (**MISCELLANEOUS - MICS).** Tuy nhiên, số lượng thực thể này sẽ không cố định mà có thể được mở rộng ra bao gồm cả: ngày tháng, nghề nghiệp, tuổi, phương tiện, … Dưới đây là một câu ví dụ đã được gán nhãn thực thể định danh:

> Như đã nói, em [PER Trương Văn Khải] là sinh viên năm ba của một [ORG trường đại học Công Nghệ Thông Tin] ở [LOC Việt Nam].

Trong ví dụ, có 3 thực thể đã được gán nhãn trong đó có 1 thực thể tên người, 1 thực thể tên tổ chức, 1 thực thể tên địa điểm.

Ngoài ra, một phương pháp tiêu chuẩn đã đề xuất và được áp dụng trong khâu gán nhãn chuỗi cho bài toán Nhận dạng thực thể định danh đó là phương pháp gán nhãn **BIO** (**B**egin - **I**nside - **O**utside). Bằng việc sử dụng phương pháp này, có thể giải quyết bài toán Nhận dạng thực thể định danh như một bài toán gán nhãn chuỗi cho từng từ bởi hai giá trị: Ranh giới từ (Là vị trí bắt đầu **B** hay nằm trong **I**) và loại thực thể định danh tương ứng. Ví dụ cho bài toán Nhận dạng thực thể định danh với phương pháp gán nhãn chuỗi:

> Bệnh nhân [B-PER TVK] là phi công hãng [B-ORG Vietnam Airlines], được xác định là dương tính vào ngày [B-DATE 19/7], thường trú tại [B-LOC quận] [I-LOC hai].

`

Ví dụ trên đã trình bày cho cách gán nhãn sử dụng phương pháp gán nhãn **BIO**, ta gán nhãn tất cả những từ bắt đầu cụm thực thể bằng B, các từ xuất hiện trong cụm đó bằng nhãn I. Tóm tắt lại quá trình gán nhãn trên, ta có thể phát biểu bài toán tổng quát như sau:

- **Input**: $X = \{x_1, x_2, \ldots, x_n\}$ là chuỗi các từ tiếng việt x_i.
- **Output**: $Y = \{y_1, y_2, \ldots, y_n\}$ là chuỗi các nhãn y_i được gán nhãn tương ứng với x_i.

Thông thường, bài toán Nhận dạng thực thể định danh được chia thành 2 quy trình liên tiếp: 1 - Nhận dạng thực thể; 2 - Phân loại thực thể. Nhận dạng thực thể là quy trình tìm kiếm các thể thực thể đã được định danh có ở trong câu. Phân loại thực thể là quy trình phân loại các thực thể về các loại như: tên người, tên địa điểm, …

`

CHƯƠNG 2: NGỮ LIỆU

2.1 Giới thiệu bộ dữ liệu

Vào thời điểm năm 2020, tổng số ca nhiễm COVID 19 trên toàn cầu đã tăng chóng mặt và đạt một con số cực kỳ khủng khiếp. Số lượng ca nhiễm mới luôn được báo cáo cập nhập. Ở Việt Nam, các báo cáo văn bản chứa thông tin chính thức từ chính phủ về các ca bệnh Covid 19 luôn được cập nhập, chi tiết bao gồm về: thông tin cá nhân giấu tên, lịch trình đi lại, thông tin về những người tiếp xúc với ca bệnh. Do đó, việc xây dựng hệ thống để truy xuất và tóm tắt thông tin từ những nguồn chính thức này là rất quan trọng, giúp những người và tổ chức liên quan có thể nhanh chóng nắm bắt thông tin chính cho các nhiệm vụ phòng dịch, và hệ thống cũng phải có khả năng thích ứng và đồng bộ nhanh chóng với các đợt dịch sắp diễn ra trong tương lai.

Đó cũng là lí do ra đời của bộ dữ liệu PhoNER Covid-19 [1], một bộ dữ liệu có chứa thông tin liên quan đến Covid-19 được chú thích với các nhãn của thực thể được định nghĩa trước và có thể được áp dụng trong các đợt dịch bệnh trong tương lai.

Đây là bộ dữ liệu được phát hành với mục đích nghiên cứu hoặc giáo dục, cũng là bộ dữ liệu tiếng Việt đầu tiên được chú thích thủ công trong lĩnh vực COVID-19. Bộ dữ liệu của PhoNER Covid-19 được chú thích với 10 loại thực thể khác nhau liên quan đến bệnh nhân COVID-19 tại Việt Nam. Bộ dữ liệu bao gồm 35,000 thực thể trên 10,000 câu.

2.2 Cấu trúc bộ dữ liệu

2.2.1 Các loại thực thể

Bộ dữ liệu được xây dựng với 10 thực thể xác định để trích xuất thông tin có liên quan đến bệnh nhận Covid-19. Nhìn chung thì bộ dữ liệu này không chỉ giới hạn ở khuôn khổ đại dịch Covid-19 mà còn có thể được sử dụng trong các dịch bệnh khác trong tương lai. Mô tả ngắn gọn từng loại thực thể như sau:

`

Nhãn	Định nghĩa
PATIENT_ID	Mã định danh duy nhất của một bệnh nhân mắc Covid-19 tại Việt Nam. Chú thích PATIENT_ID phía trên "X" đề cập đến bệnh nhân thứ X mắc COVID-19 tại Việt Nam.
PERSON_NAME	Tên bệnh nhân hoặc người tiếp xúc với bệnh nhân.
AGE	Tuổi của bệnh nhân hoặc người tiếp xúc với bệnh nhân.
GENDER	Giới tính của bệnh nhân hoặc người tiếp xúc với bệnh nhân.
JOB	Công việc của bệnh nhân hoặc người tiếp xúc với bệnh nhân.
LOCATION	Địa điểm/nơi ở mà bệnh nhân đã đến.
ORGANIZATION	Các tổ chức liên quan đến bệnh nhân, ví dụ: công ty, tổ chức chính phủ, v.v., với cơ cấu và chức năng riêng của chúng.
SYMPTOM_DISEASE	Các triệu chứng mà bệnh nhân gặp phải và các bệnh mà bệnh nhân mắc phải trước khi mắc bệnh COVID-19 hoặc các biến chứng thường xuất hiện trong báo cáo tử vong.
TRANSPORTATION	Phương tiện vận chuyển mà bệnh nhân sử dụng. Ở đây, chúng tôi chỉ gắn thẻ số nhận dạng cụ thể của phương tiện, ví dụ: số chuyến bay và biển số xe buýt/ô tô.
DATE	Bất kỳ ngày nào xuất hiện trong câu.

Table 2.1: Bảng mô tả từng loại thực thể và định nghĩa

2.2.2 Quá trình gán nhãn dữ liệu

Toàn bộ quy trình bộ dữ liệu trên sẽ được tạo ra như sau:

- Thu thập dữ liệu liên quan đến COVID-19: Thu thập dữ liệu các bài viết được gắn thẻ với từ khóa "COVID-19" hoặc "COVID" từ các trang tin tức trực tuyến có uy tín của Việt Nam và phân đoạn nội dung văn bản chính của các bài báo được thu thập thông tin thành các câu bằng RDRSegmenter [2] từ VnCoreNLP. Các câu liên quan đến bệnh nhân COVID-19 được chọn bằng BM25Plus. Sau đó, lọc thủ công những thông tin không chứa thông tin liên quan đến bệnh nhân ở Việt Nam, kết quả là 10027 câu thô.
- Quy trình đánh nhãn dữ liệu:

`

- Trước tiên, phát triển một hướng dẫn chú thích ban đầu và lấy mẫu ngẫu nhiên một bộ thí điểm gồm 1000 câu để chú thích thủ công để sử dụng để kiểm soát chất lượng.
- Sau đó, chia toàn bộ bộ dữ liệu gồm 10027 câu thành 10 tập hợp con không chồng chéo và bằng nhau, mỗi tập hợp chứa 100 câu từ bộ thí điểm và sử dụng 10 chú thích. Chất lượng chú thích được đo bằng F1 được tính trên 100 câu đã có chú thích vàng từ bộ thí điểm. Tất cả các chú thích được yêu cầu sửa đổi chú thích của họ cho đến khi họ đạt được F1 ít nhất là 0,92, sau đó chúng tôi xem xét lại từng câu và sửa thêm nếu cần.
- Bộ dữ liệu kết quả bao gồm 35K thực thể trên 10027 câu.

2.2.3 Phân chia dữ liệu

Nhóm tác giả chia ngẫu nhiên bộ dữ liệu từ 10,027 câu thành các tập train/val/test với tỷ lệ là 5/2/3, đồng thời đảm bảo tỷ lệ phân phối tương đồng của các thực thể trên cả ba tập này. Thống kê của bộ dữ liệu như sau:

Entity Type	Train	Valid.	Test	All
PATIENT_ID	3240	1276	2005	6521
PERSON_NAME	349	188	318	855
AGE	682	361	582	1625
GENDER	542	277	462	1281
OCCUPATION	205	132	173	510
LOCATION	5398	2737	4441	12576
ORGANIZATION	1137	551	771	2459
SYMPTOM&DISEASE	1439	766	1136	3341
TRANSPORTATION	226	87	193	506
DATE	2549	1103	1654	5306
# Entities in total	15767	7478	11735	34984
# Sentences in total	5027	2000	3000	10027

Table 2.2: Bảng thống kê các loại thực thể trong bộ dữ liệu PhoNER

`

2.3 Phân tích việc gán nhãn dữ liệu

Chúng tôi sẽ tiến hành phân tích việc đánh nhãn dữ liệu của nhóm tác giả trên 60 câu dữ liệu được lấy từ cả 3 bộ train/val/test. Sau đó quá trình phân tích được thể hiện ở bảng bên dưới:

STT	Nguồn	Input	Nhãn	Lý Do
1	Train set - Dòng 1	"Đồng", "thời", ",", "bệnh", "viện", "tiếp", "tục", "thực", "hiện", "các", "biện", "pháp", "phòng", "chống", "dịch", "bệnh", "COVID", "-", "19", "theo", "hướng", "dẫn", "của", "Bộ", "Y", "tế", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O"	1. Các từ trước từ các từ "Bộ", "Y", "tế" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Cụm các từ "Bộ", "Y", "tế" là cụm từ có nghĩa là **Tên cơ quan đến việc xử lý dịch tể đồng thời cũng là Tên viết gọn của cơ quan ở cấp độ Quốc Gia: "Bộ Y tế" viết tắt cho "Bộ Y tế Việt Nam"**. Vì vậy, các từ "Bộ", "Y", "tế" lần lượt sẽ được đánh nhãn là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION". 3. Dấu "." không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O".
2	Train set - Dòng 2	"\"", "Số", "bệnh_viện", "có_thể", "tiếp_nhận", "bệnh_nhân", "bị", "sốt", "cao", "và", "khó", "thở", "đang", "giảm", "dần", "\"", ",", "thông_cáo", "có", "đoạn", ",", "cảnh_báo", "những", "bệnh_nhân", "này", "thay", "vào", "đó",	"O", "O", "O", "O", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE",	1. Các từ trước các từ "sốt", "cao" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Cụm các từ "sốt", "cao" là cụm từ có nghĩa là **Triệu chứng liên quan tới bệnh nhân COVID-19** nên sẽ lần lượt được đánh nhãn là "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 3. Từ "và" không liên quan đến các thực thể

`

		"được", "chuyển", "tới", "các", "phòng_khám", "khẩn_cấp", ",", "khiến", "những", "bệnh_nhân", "mắc", "bệnh", "hiểm_nghèo", "khác", "không", "có", "cơ_hội", "được", "điều_trị", "."	SE", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"]	đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 4. Cụm các từ "khó", "thở" là cụm từ có nghĩa là **Triệu chứng liên quan tới bệnh nhân COVID-19** nên sẽ lần lượt được đánh nhãn là "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 5. Các từ còn lại không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O".
3	Train set - Dòng 3	"Ngoài_ra", ",", "những", "người", "tiếp_xúc", "gián_tiếp", "(", "đã", "gặp", "những", "người", "tiếp_xúc", "gần", "với", "bệnh_nhân", ")", "được", "lập", "danh_sách", "và", "yêu_cầu", "cách_ly", "y_tế", "tại", "nơi", "ở", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O",	1. Tất cả các từ trong câu đều không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O" hết.

`

			"O", "O", "O", "O", "O", "O", "O", "O", "O"	
4	Train set - Dòng 4	["Bà", "này", "khi", "trở", "về", "quá_cảnh", "Doha", "(", "Qatar", ")", ",", "đáp", "xuống", "Tân_Sơn_Nhất", "sáng", "2/3", "cùng", "75", "hành_khách", ",", "trong", "đó", "có", "55", "người", "nước_ngoài", "."	"O", "O", "O", "O", "O", "O", "B-LOCATION", "O", "B-LOCATION", "O", "O", "O", "O", "B-LOCATION", "O", "B-DATE", "O", "O", "O", "O", "O", "O", "O", "O", "O"	1. Các từ trước từ "DoHa" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "DoHa" chỉ **tên một Quốc Gia** nên được đánh nhãn là B-Location 3. Từ "Qatar" tương tự trường hợp 2. 4. Từ "Tân Sơn Nhất" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: sân bay** nên được đánh nhãn B-Location. 5. Từ "2/3" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date 6. Các từ còn lại không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn “O”.
5	Train set - Dòng 5	"\"", "Bệnh_nhân", "523", "\"", "và", "chồng", "là", "\"", "bệnh_nhân", "522", "\"", ",", "67", "tuổi", ",", "được", "Bộ", "Y_tế", "ghi_nhận", "nhiễm", "nCoV", "hôm", "31/7", "."	"O", "O", "B-PATIENT_ID", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "O", "B-AGE", "O", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "O",	1. Các từ trước từ "523" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "523" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID** nên X ở đây là "523". 3. Các từ trước từ "522" đều không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O" hết.

`

			"O", "O", "O", "B-DATE", "O"	4. Từ "522" tương tự trường hợp 2 nên được đánh nhãn là B-PATIENT_ID. 5. Tiếp theo là cụm "67", "tuổi". Ta **chỉ đánh nhãn Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi"). Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân)**. Vì vậy, ta cần đánh nhãn từ "67" là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 6. Cụm các từ "Bộ", "Y_tế" là cụm từ có nghĩa là **Tên cơ quan đến việc xử lý dịch tể đồng thời cũng là Tên viết gọn của cơ quan ở cấp độ Quốc Gia: "Bộ Y tế" viết tắt cho "Bộ Y tế Việt Nam"**. Vì vậy, các từ "Bộ", "Y_tế" lần lượt sẽ được đánh nhãn là: "B-ORGANIZATION", "I-ORGANIZATION". 7. Từ "31/7" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date

`

6	Train set - Dòng 6	"Trường_hợp", "bệnh_nhân", "188", "L.T.H.", ",", "theo", "thông_tin", "từ", "cơ_quan", "y_tế", "địa_phương", ",", "bệnh_nhân", "về", "nhà", "ngày", "14", "-", "4", "và", "từ", "đó", "chỉ", "tiếp_xúc", "với", "chồng", "và", "con", ",", "đây", "không", "phải", "là", "tái_nhiễm", "mà", "do", "có_thể", "virus", "yếu", "ở", "thời_điểm", "lấy", "mẫu", "lần", "trước", ",", "hoặc", "vị_trí", "lấy", "mẫu", ",", "thời_điểm", "lấy", "mẫu", "dẫn", "đến", "âm_tính", "giả", "."	"O", "O", "B-PATIENT_ID", "B-NAME", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"	1. Các từ trước từ "188" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "188" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "188". 3. Từ "L.T.H": **Tên bệnh nhân (để bảo vệ quyền riêng tư, tên bệnh nhân COVID-19 thường được viết tắt)**, trong câu là "bà", "L.T.H" nên cụm "L.T.H" được đánh nhãn là "B-NAME". Từ "bà" không được đánh nhãn bởi vì: **Các danh xưng "ông", "bà", "anh", "chị", "giám đốc", "chủ tịch", … KHÔNG nằm trong tên riêng.** 4. Các từ trước cụm các từ "14", "-", "4" đều không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O" hết. 5. Cụm từ "14", "-", "4" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 6. Các từ còn lại không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O".

`

7	Train set - Dòng 7	"Riêng", "bệnh_nhân", "91", "là", "phi_công", "người", "Anh", "ngụ", "ở", "quận", "2", ",", "TP.", "HCM", "và", "có", "liên_quan", "ổ", "dịch", "quán", "bar", "Buddha", ",", "thông_tin", "cập_nhật", "ngày", "10", "-", "4", "cho", "biết", "diễn_biến", "bệnh", "của", "bệnh_nhân", "không", "xấu", "hơn", "nhưng", "cũng", "chưa", "có", "dấu_hiệu", "hồi_phục", "."	"O", "O", "B-PATIENT_ID", "O", "B-JOB", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "O", "O", "O", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"	1. Các từ trước từ "91" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "91" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "91". 3. Từ "phi_công" được gán nhãn là B-JOB là vì: **Chỉ gán nhãn nghề nghiệp của bệnh nhân và các cá nhân có liên quan trực tiếp (tiếp xúc, gặp mặt, ở gần). Ngoài ra, những từ chỉ nghề nghiệp cần phải được gắn với 1 cá nhân nhất định trong câu (có tên, có mã bệnh nhân).** 4. Các từ trước cụm các từ "quận", "2" đều không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O" hết. 5. Cụm các từ "quận", "2" là **Địa chỉ, đặc biệt hơn nó chỉ cấp bậc đơn vị hành chính nên được xem như một thực thể riêng biệt**. Nên cụm "quận", "2" lần lượt được gán nhãn là "B-LOCATION", "I-LOCATION". 6. Cụm các từ "TP.", "HCM" được gán nhãn là "B-LOCATION", "I-LOCATION" là vì: **Tên đơn vị hành chính của quốc gia.** 7. Cụm các từ "quán", "bar", "Buddha" chỉ **Tên các địa điểm mang tính thương mại:**

`

				nhà hàng, quán ăn, quán nước nên lần được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 8. Cụm từ "10", "-", "4" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 9. Các từ còn lại không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O".
8	Train set - Dòng 8	"Bệnh_nhân", "đã", "được", "xét_nghiệm", "có", "3", "kết_quả", "âm_tính", "vào", "các", "ngày", "19", ",", "21", "và", "23", "-", "8", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "B-DATE", "O", "B-DATE", "O", "B-DATE", "I-DATE", "I-DATE", "O"	1. Các từ trước từ "19" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "19" chỉ **Ngày** nên được đánh nhãn lần lượt là: "B-DATE". 3. Từ "21" chỉ **Ngày** nên được đánh nhãn lần lượt là: "B-DATE". 4. Cụm từ "23", "-", "8" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE".
9	Train set - Dòng 9	"Bà", "đã", "tiếp_xúc", "với", "người_thân", "xác_định", "mắc", "Covid", "-", "19", "trước", "khi", "về", "Việt_Nam", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-	1. Các từ trước từ "Việt_Nam" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "Việt_Nam" chỉ **Tên quốc gia: Việt Nam** nên được đánh nhãn là B-LOCATION.

`

			LOCATION", "O"	
10	Train set - Dòng 10	"Chiều", "22", "-", "4", ",", "bệnh_nhân", "được", "cho", "về", "theo_dõi", "cách_ly", "tại", "nhà", "."	"O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "O", "O", "O", "O", "O", "O", "O"	1. Cụm từ "22", "-", "4" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 2. Các từ còn lại không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O".
11	Train set - Dòng 11	"Hôm_qua", ",", "hai", "bệnh_nhân", "Covid", "-", "19", "cũng", "tử_vong", ",", "có", "bệnh", "nền", "suy", "thận", "mạn", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O"	1. Các từ trước từ "nền" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Cụm các từ "suy", "thận", "mạn" chỉ **Các loại bệnh khác mà bệnh nhân COVID-19 mắc phải** nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE".

`

12	Train set - Dòng 12	"8h", "ngày", "1", "-", "8", ",", "bệnh_nhân", "861", "chở", "con", "gái", "đến", "khám", "tại", "phòng_khám", "đa_khoa", "của", "bác_sĩ", "Hoàng_Đức_Dũng", "(", "số", "16", "-", "18", "B", "-", "22", "đường", "Lê_Duẩn", ",", "TP", "Đông_Hà", ")", "."	"O", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "B-PATIENT_ID", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-	1. Cụm từ "1", "-", "8" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 2. Từ "861" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "91". 3. Cụm các từ "phòng_khám", "đa_khoa", "của", "bác_sĩ", "Hoàng_Đức_Dũng" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng** và là **địa danh liên quan đến lịch trình di chuyển của bệnh nhân** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION". 4. Cụm các từ "16", "-", "18", "B", "-", "22", "đường", "Lê_Duẩn" chỉ **Địa chỉ: Số nhà phải bao gồm cả tên đường** nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION". 5. Cụm các từ "TP", "Đông_Hà" được gán nhãn là "B-LOCATION", "I-LOCATION" là vì: **Tên đơn vị hành chính của quốc gia.**

`

			LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "O"	
13	Train set - Dòng 13	["Cả", "hai", "đều", "thuộc", "diện", "xét_nghiệm", "sàng_lọc", ",", "lấy", "mẫu", "bệnh_phẩm", "ngày", "11/4", ",", "kết_quả", "dương_tính", "ngày", "13/4", ",", "điều_trị", "tại", "Bệnh_viện", "Bệnh", "Nhiệt_đới", "Trung_ương", "cơ_sở", "2", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-DATE", "O", "O", "O", "O", "B-DATE", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O"	1. Các từ trước từ "Việt_Nam" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "11/4" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date 3. Từ "13/4" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date. 4. Cụm các từ "Bệnh_viện", "Bệnh", "Nhiệt_đới", "Trung_ương", "cơ_sở", "2" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: bệnh viện, trạm y tế** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION".

`

14	Train set - Dòng 14	"Tỉnh", "đến", "ngày", "11", "-", "3", ",", "trên", "địa_bàn", "tỉnh", "Quảng_Ninh", "chưa", "phát_hiện", "thêm", "ca", "bệnh", "COVID", "-", "19", "ngoài", "4", "ca", "trước", "đó", "(", "đều", "là", "du_khách", "nước_ngoài", "đi", "cùng", "chuyến", "bay", "VN0054", "với", "bệnh_nhân", "thứ", "17", ")", "."	"O", "O", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-TRANSPORTATION", "O", "O", "O", "B-PATIENT_ID", "O", "O"	1. Cụm từ "11", "-", "3" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 2. Cụm các từ "Tỉnh", "Quảng_Ninh" được gán nhãn là "B-LOCATION", "I-LOCATION" là vì: **Tên đơn vị hành chính của quốc gia.** 3. Các từ trước cụm các từ "quận", "2" đều không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O" hết. 4. Từ "VN0054" chỉ **nhãn biển số, số hiệu của loại phương tiện di chuyển, không gán nhãn loại phương tiện di chuyển** nên được gán nhãn là B-TRANSPORTATION. 5. Từ "17" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "523".
15	Train set - Dòng 15	"Sở", "GD", "-", "ĐT", "đã", "quán_triệt", "tất_cả", "thí_sinh", "và", "những", "người", "làm", "công_tác", "thi", "sau", "khi", "xét_nghiệm", "thực_hiện", "nghiêm", "giãn", "cách", "xã_hội", "theo", "Chỉ_thị",	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O",	1. Tất cả các từ trong câu đều không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O" hết.

`

		"16", ",", "chỉ", "di_chuyển", "khi", "theo", "lịch", "thi", "và", "lịch", "làm", "nhiệm_vụ", "của", "kỳ", "thi", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"	
16	Train set - Dòng 16	"Những", "người", "vào", "trung_tâm", "cách_ly", "được", "xếp", "ở", "chung", "phòng", "một_cách", "ngẫu_nhiên", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"	1. Tất cả các từ trong câu đều không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O" hết.
17	Train set - Dòng 17	"Theo", "đó", ",", "bệnh_nhân", "thứ", "17", "có", "2", "lần", "xét_nghiệm", "cho", "kết_quả", "âm_tính", "(", "cùng", "bệnh_nhân", "24", "và", "27", ")", ",", "đủ", "tiêu_chuẩn", "xác_định", "khỏi", "bệnh", "."	"O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-PATIENT_ID", "O", "O", "O", "O", "O", "O", "O", "O"	1. Các từ trước từ "17" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "17" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "523". 3. Các từ trước từ "24" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 4. Từ "22" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số**

`

				thứ tự. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "523". 5. Từ "27" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "523". 6. Các từ còn lại không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O".
18	Train set - Dòng 18	"Công_tác", "khám", "bệnh", ",", "chẩn_đoán", ",", "điều_trị", ",", "thực_hiện", "thủ_thuật", ",", "phẫu_thuật", "...", "đối_với", "bệnh_nhân", "như", "với", "người", "nghi", "Covid", "-", "19", "đến", "khi", "có", "kết_quả", "xét_nghiệm", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"	1. Tất cả các từ trong câu đều không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O" hết.
19	Train set - Dòng 19	"Trung_tâm", "Kiểm_soát", "bệnh_tật", "tỉnh", "lấy", "mẫu", "gửi", "Viện", "Vệ_sinh", "dịch_tễ", "trung_ương", "xét_nghiệm", "."	"O", "O", "O", "O", "O", "O", "O", "O", "B-ORGANIZATION", "I-	1. Các từ trước từ "Viện" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Cụm các từ "Viện", "Vệ_sinh", "dịch_tễ", "trung_ương" chỉ **Tên các cơ quan liên quan tới việc xử lý dịch tễ** nên

`

			ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O"	được gán nhãn lần lượt là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION".
20	Train set - Dòng 20	"Bệnh_nhân", "là", "phi_công", "hãng", "Vietnam_Airlines", ",", "xác_định", "dương_tính", "ngày", "18/3", "."	"O", "O", "O", "O", "B-ORGANIZATION", "O", "O", "O", "O", "B-DATE", "O"	1. Các từ trước từ "Vietnam_Airlines" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "Vietnam_Airlines" chỉ **Tên các công ty, tổ chức nơi bệnh nhân làm việc** nên được đánh nhãn là "B-ORGANIZATION".
21	Val Set - Dòng 1	"Bác_sĩ", "Nguyễn_Trung_Nguyên", ",", "Giám_đốc", "Trung_tâm", "Chống", "độc", ",", "Bệnh_viện", "Bạch_Mai", ",", "cho", "biết", "bệnh_nhân", "được", "chuyển", "đến", "bệnh_viện", "ngày", "7/3", ",", "chẩn_đoán", "ngộ_độc", "thuốc", "điều_trị", "sốt_rét", "chloroquine", "."	"O", "O", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "O", "O", "O", "B-DATE", "O", "O", "B-	1. Không gán nhãn những người không liên quan trực tiếp đến lịch trình di chuyển hay không có liên hệ, không tiếp xúc với bệnh nhân nên cụm từ "Bác_sĩ", "Nguyễn_Trung_Nguyên" không được đánh bất kỳ nhãn nào. 2. "Trung_tâm", "Chống", "độc", ",", "Bệnh_viện", "Bạch_Mai" chỉ **Tên các cơ quan liên quan tới việc xử lý dịch tễ** nên lần lượt được gán nhãn là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION". 3. Các từ trước từ "7/3" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 4. Từ "7/3" chỉ **Ngày trong tiếng Việt**

`

			SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O", "O", "O", "O"	**thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date. 5. Cụm các từ "ngộ_độc", "thuốc" chỉ **Các loại bệnh khác mà bệnh nhân COVID-19** mắc phải nên được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE".
22	Val Set - Dòng 2	"\"", "Bệnh_nhân", "812", "\"", ",", "nam", ",", "62", "tuổi", ",", "là", "nhân_viên", "giao", "bánh", "tiệm", "pizza", "phố", "Trần_Thái_Tông", ",", "Hà_Nội", ",", "trú", "tại", "quận", "Bắc_Từ_Liêm", ",", "lây", "từ", "\"", "bệnh_nhân", "447", "\"", "(", "cũng", "là", "nhân_viên", "tiệm", "bánh", ",", "đi", "du_lịch", "Đà_Nẵng", ")", "."	"O", "O", "B-PATIENT_ID", "O", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "B-JOB", "I-JOB", "I-JOB", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O", "B-LOCATION", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "O",	1. Từ "812" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "812". 2. Từ "nam" chỉ **Chỉ gán nhãn giới tính của bệnh nhân và những người liên quan trực tiếp (tiếp xúc) với bệnh nhân**. Trong một câu, giới tính cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân) nên được gán nhãn là "B-GENDER". 3. Cụm "62", "tuổi". **Ta chỉ đánh nhãn Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi")**. Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ “62” là "B-AGE" vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 4. Cụm các từ "nhân_viên", "giao", "bánh" chỉ **Chỉ gán nhãn nghề nghiệp của bệnh**

`

			"O", "O", "O", "B-PATIENT_ID", "O", "O", "O", "O", "B-JOB", "I-JOB", "I-JOB", "O", "O", "O", "B-LOCATION", "O", "O"	**nhân và các cá nhân có liên quan trực tiếp (tiếp xúc, gặp mặt, ở gần)**. Ngoài ra, những từ chỉ nghề nghiệp cần phải được gắn với 1 cá nhân nhất định trong câu (có tên, có mã bệnh nhân) nên lần lượt được gán nhãn là: "B-JOB", "I-JOB", "I-JOB". 5. Cụm các từ "tiệm", "pizza", "phố", "Trần_Thái_Tông" chỉ **Địa điểm mang tính thương mại: quán ăn** nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION". 6. Từ "Hà_Nội" là **Tên đơn vị hành chính của quốc gia** nên được gán nhãn là "B-LOCATION". 7. Cụm các từ "quận", "Bắc_Từ_Liêm" chỉ **Tên đơn vị hành chính của quốc gia (gán nhãn cả các từ chỉ đơn vị hành chính: tỉnh, thành phố, quận, huyện, đường)** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION". 8. Từ "447" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "447". 9. Cụm các từ "nhân_viên", "tiệm", "bánh" tương tự trường hợp 4 nên lần lượt được gán nhãn là: "B-JOB", "I-JOB", "I-JOB". 10. Từ "Đà_Nẵng" là **Tên đơn vị hành chính của quốc gia** nên được gán nhãn là

`

				"B-LOCATION".
23	Val Set - Dòng 3	"Trong", "số", "những", "người", "mà", "cô", "ấy", "đã", "tiếp_xúc", "với", "có", "nhân_viên", "MGM", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-ORGANIZATION", "O"	1. Các từ trước từ "MGM" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "MGM" chỉ **Tên các công ty, tổ chức nơi bệnh nhân làm việc** nên được đánh nhãn là "B-ORGANIZATION".

`

24	Val Set - Dòng 4	"Trong", "số", "hành_khách", "nhiễm", "có", "3", "người", "Việt", "là", "\"", "bệnh_nhân", "17", "\"", "Nguyễn_Hồng_Nhung", ",", "\"", "bệnh_nhân", "21", "\"", "Nguyễn_Quang_Thuấn", "và", "một", "nữ", "tiếp_viên", "hàng_không", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-NAME", "O", "O", "O", "B-PATIENT_ID", "O", "B-NAME", "O", "O", "O", "O", "O"	1. Các từ trước từ "17" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Từ "17" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "17". 3. Từ "Nguyễn_Hồng_Nhung" chỉ **Tên bệnh nhân** nên được gán nhãn là "B-NAME". 4. Từ "21" tương tự trường hợp 2 nên được gán nhãn là "B-PATIENT_ID". 5. Từ "Nguyễn_Quang_Tuấn" chỉ **Tên bệnh nhân** nên được gán nhãn là "B-NAME". 6. Cụm các từ "nữ", "tiếp_viên", "hàng_không" **KHÔNG** được đánh nhãn là "B-JOB" vì nó không được gắn với 1 cá nhân nhất định trong câu (có tên, có mã bệnh nhân).

`

25	Val Set - Dòng 5	"Bệnh_viện", "đa_khoa", "Trung_ương", "Quảng_Nam", "công_bố", "khỏi", "bệnh", "và", "cho", "xuất_viện", "9", "bệnh_nhân", ",", "gồm", "bệnh_nhân", "598", "(", "8", "tuổi", ")", ",", "bệnh_nhân", "774", "(", "63", "tuổi", ")", ",", "bệnh_nhân", "911", "(", "79", "tuổi", ")", ",", "bệnh_nhân", "432", "(", "63", "tuổi", ")", ",", "bệnh_nhân", "835", "(", "26", "tuổi", ")", ",", "bệnh_nhân", "792", "(", "25", "tuổi", ")", ",", "bệnh_nhân", "463", "(", "42", "tuổi", ")", ",", "bệnh_nhân", "720", "(", "30", "tuổi", ")", "và", "bệnh_nhân", "736", "(", "39", "tuổi", ")", "."	"B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-AGE", "O", "O", "O", "B-PATIENT_ID", "O", "B-AGE", "O", "O", "O", "B-PATIENT_ID", "O", "B-AGE", "O", "O", "O", "B-PATIENT_ID", "O", "B-AGE", "O", "O", "O", "B-	1. Cụm các từ "Bệnh_viện", "đa_khoa", "Trung_ương", "Quảng_Nam" chỉ **Tên các cơ quan liên quan tới việc xử lý dịch tễ** nên được gán nhãn lần lượt là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION". Không gán nhãn "LOCATION" cho cụm các từ trên vì: Thực thể kiểu ORGANIZATION phải là tổ chức bao gồm một hay nhiều cá nhân và phải có chức năng, công việc nhất định. Thực thể kiểu ORGANIZATION thường đóng vai trò là chủ ngữ, thực hiện một hành động nào đó trong câu. 2. Từ "598" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "598". 3. Tiếp theo là cụm "8", "tuổi". Ta chỉ đánh nhãn **Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi")**. Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ “8” là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 4. Từ "911" tương tự trường hợp 2 nên được gán nhãn là "B-PATIENT_ID".

`

			PATIENT_ID", "O", "B-AGE", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-AGE", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-AGE", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-AGE", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-AGE", "O", "O", "O"	5. Từ "79" tương tự trường hợp 3 nên được gán nhãn là "B-AGE". 6. Từ "432" tương tự trường hợp 2, 4 nên được gán nhãn là "B-PATIENT_ID". 7. Từ "63" tương tự trường hợp 3, 5 nên được gán nhãn là "B-AGE". 8. Từ "835" tương tự trường hợp 2, 4, 6 nên được gán nhãn là "B-PATIENT_ID". 9. Từ "26" tương tự trường hợp 3, 5, 7 nên được gán nhãn là "B-AGE". 10. Từ "792" tương tự trường hợp 2, 4, 6, 8 nên được gán nhãn là "B-PATIENT_ID". 11. Từ "25" tương tự trường hợp 3, 5, 7, 9 nên được gán nhãn là "B-AGE". 12. Từ "463" tương tự trường hợp 2, 4, 6, 8, 10 nên được gán nhãn là "B-PATIENT_ID". 13. Từ "42" tương tự trường hợp 3, 5, 7, 9, 11 nên được gán nhãn là "B-AGE". 14. Từ "720" tương tự trường hợp 2, 4, 6, 8, 10, 12 nên được gán nhãn là "B-PATIENT_ID". 15. Từ "30" tương tự trường hợp 3, 5, 7, 9, 11, 13 nên được gán nhãn là "B-AGE". 16. Từ "736" tương tự trường hợp 2, 4, 6, 8, 10, 12, 14 nên được gán nhãn là "B-PATIENT_ID". 17. Từ "39" tương tự trường hợp 3, 5, 7, 9, 11,13, 15 nên được gán nhãn là "B-AGE".
26	Val Set - Dòng 6	"Kết_quả", "xét_nghiệm", "lần", "1", "vào", "ngày", "15", "-", "9", "là", "âm_tính", "với", "SARS", "-", "CoV", "-", "2", "."	"O", "O", "O", "O", "O", "O", "B-DATE", "I-DATE", "I-	1. Các từ trước từ "15" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Cụm từ "15", "-", "9" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X**

`

			DATE", "O", "O", "O", "O", "O", "O", "O", "O"	**là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 3. Các từ còn lạin tương tự trường hợp 1.
27	Val Set - Dòng 7	"Ngoài_ra", ",", "cô", "tới", "một_số", "nơi", "gồm", "quán", "ăn_ở", "TP", "Biên_Hoà", ",", "Đồng_Nai", "ngày", "13/3", ",", "siêu_thị", "An_Phú", "ngày", "16/3", ",", "nhà_máy", "Huệ_Phong", "(", "quận", "Gò_Vấp", ")", "ngày", "19/3", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "O", "B-DATE", "O", "B-LOCATION", "I-LOCATION", "O", "B-DATE", "O", "B-ORGANIZATION", "I-ORGANIZATION", "O", "B-LOCATION", "I-LOCATION", "O", "O", "B-DATE",	1. Các từ trước từ "TP" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Cụm các từ "TP", "Biên_Hòa" được gán nhãn là "B-LOCATION", "I-LOCATION" là vì: **Tên đơn vị hành chính của quốc gia.** 3. Từ "Đồng_Nai" chỉ **Tên đơn vị hành chính của quốc gia** nên được gán nhãn là "B-LOCATION". 4. Từ "13/3" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date. 5. Cụm các từ "siêu_thị", "An_Phú" chỉ **Tên các địa điểm mang tính thương mại: siêu thị** nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION". 6. Từ "16/3" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date. 7. Cụm các từ "nhà_máy", "Huệ_Phong" chỉ **Tên các công ty, tổ chức nơi bệnh nhân làm việc** nên được gán nhãn lần lượt là "B-ORGANIZATION", "I-ORGANIZATION". 8. Cụm các từ "quận", "Gò_Vấp" là **Địa chỉ, đặc biệt hơn nó chỉ cấp bậc đơn vị hành chính**. Nên cụm "quận", "2" lần lượt được

`

			"O"	gán nhãn là "B-LOCATION", "I-LOCATION". 9. Từ "19/3" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date.
28	Val Set - Dòng 8	"Cô", "vào", "Khoa_Nội", "tổng_hợp", "(", "Bệnh_viện", "Đà_Nẵng", ")", "chăm_sóc", "bố", "chồng", "và", "tiếp_xúc", "với", "chị", "của", "chồng", "là", "nữ", "\"", "bệnh_nhân", "510", "\"", "(", "61", "tuổi", ",", "ở", "phường", "Phú_Thọ", "Hoà", ",", "quận", "Tân_Phú", ",", "TP.", "HCM", ")", "được", "Bộ", "Y_tế", "công_bố", "ngày", "31/7", "."	"O", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "O", "B-AGE", "O", "O", "O", "B-LOCATION", "I-	1. Cụm các từ "Khoa_Nội", "tổng_hợp" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: khoa của bệnh viện nên** được đánh nhãn lần lượt là "B-LOCATION", "I-LOCATION". 2. Cụm các từ "Bệnh_viện", "Đà_Nẵng" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: bệnh viện** nên được đánh nhãn lần lượt là "B-LOCATION", "I-LOCATION". 3. Từ "510" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "510". 4. Tiếp theo là cụm "61", "tuổi". Ta chỉ **đánh nhãn Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi").** Trong

`

			LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "O", "O", "B-DATE", "O"	một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ "61" là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 5. Cụm các từ "phường", "Phú_Thọ", "Hoà" chỉ **Địa chỉ** nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 6. Cụm các từ "quận", "Tân_Phú" là **Địa chỉ, đặc biệt hơn nó chỉ cấp bậc đơn vị hành chính** nên được xem như một thực thể riêng biệt. Nên cụm "quận", "Tân_Phú" lần lượt được gán nhãn là "B-LOCATION", "I-LOCATION". 7. Cụm các từ "TP.", "HCM" được gán nhãn là "B-LOCATION", "I-LOCATION" là vì: **Tên đơn vị hành chính của quốc gia.** 8. Cụm các từ "Bộ", "Y_tế" là cụm từ có nghĩa là **Tên cơ quan đến việc xử lý dịch tể đồng thời cũng là Tên viết gọn của cơ quan ở cấp độ Quốc Gia:** "Bộ Y tế" viết tắt cho "Bộ Y tế Việt Nam". Vì vậy, các từ "Bộ", "Y_tế" lần lượt sẽ được đánh nhãn là: "B-ORGANIZATION", "I-ORGANIZATION". 9. Từ "31/7" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date.

`

29	Val Set - Dòng 9	"Công_ty", "PouYuen", "trở_thành", "nỗi", "lo", "bùng_phát", "\"", "ổ", "dịch", "\"", "của", "TP.HCM."]	"tags": ["B-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "O", "O", "O", "B-LOCATION"]	1. từ "Công_ty" và "Pouyen" là 1 cụm từ "Công ty Pouyen". Từ khóa "B-" thể hiện sự bắt đầu của cụm từ chỉ thực thể và từ khóa "I-" là thể hiện sự tiếp nối hay có liên quan bên trong của cụm từ sau "B-". 2. Từ "Công_ty" được đánh nhãn là "ORGANIZATION" bởi vì một từ được đánh nhãn thực thể là "ORGANIZATION" nếu nó thuộc nhiều trường hợp, trong đó có trường hợp như sau: **Là tên công ty hay tổ chức nơi bệnh nhận làm việc.** 3. Từ "PouYuen" là tên riêng của công ty, nhưng nó vẫn trực thuộc **Là tên công ty hay tổ chức nơi bệnh nhận làm việc** và nó là từ bắt đầu sau cụm từ "Công_ty" nên được gán nhãn "I-ORGANIZATION" 4. Các từ sau đó (từ từ thứ 3 đến 11) không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 5. Từ "TP.HCM" được đánh nhãn "B-LOCATION" là vì: **Đây là tên một đơn vị hành chính Quốc Gia (thành phố Hà Nội, quận 12, ….).**
30	Val Set - Dòng 10	"Hệ_miễn_dịch", "suy_yếu", "sẽ", "làm", "cho", "bệnh_nhân", "có", "nguy_cơ", "mắc", "nCoV", "cao", "hơn", ",", "có_thể", "dẫn", "tới", "các", "biến_chứng", "nghiêm_trọng", "."	"B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O", "O", "O", "O", "O", "O", "O", "O",	1. Cụm các từ "Hệ_miễn_dịch", "suy_yếu" chỉ **Các loại bệnh khác mà bệnh nhân COVID-19 mắc phải** nên được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 2. Các từ còn lại không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O".

`

			"O", "O", "O", "O", "O", "O", "O", "O", "O"	
31	Val Set - Dòng 11	"Ông", "bị", "suy", "thận", "mạn", "giai_đoạn", "cuối", ",", "từng", "ngừng", "tim", "nhiều", "lần", "tại", "Bệnh_viện", "Đà_Nẵng", "."	"O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O", "B-LOCATION"	1. Cụm các từ "suy", "thận", "mạn", "giai_đoạn", "cuối" chỉ **Các loại bệnh khác mà bệnh nhân COVID-19 mắc phải** nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 2. Cụm các từ "ngừng", "tim", "nhiều", "lần" chỉ **Triệu chứng liên quan tới bệnh nhân COVID-19** nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 3. Cụm các từ "Bệnh_viện", "Đà_Nẵng" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: bệnh viện** nên được đánh nhãn lần lượt là "B-LOCATION", "I-LOCATION".

`

			, "I-LOCATION", "O"	
32	Val Set - Dòng 12	"Bệnh_nhân", "từng", "mua", "thịt", "và", "cá", "tại", "chợ", "đầu_mối", "Tân_Phát_Địa", "8", "ngày", "trước", "khi", "có", "triệu_chứng", "."	"O", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "O", "O", "O", "O", "O", "O"	1. Cụm các từ "chợ", "đầu_mối", "Tân_Phát_Địa" chỉ **Tên các địa điểm mang tính thương mại: nhà hàng, quán ăn, khách sạn, chợ, siêu thị** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION".
33	Val Set - Dòng 13	"Từ", "hôm_nay", ",", "Bệnh_viện", "Đa_khoa", "huyện", "Đồng_Văn", "-", "nơi", "\"", "bệnh_nhân", "268", "\"", "điều_trị", ",", "tạm_thời", "dừng", "tiếp_nhận", "người_bệnh", "đến", "khám", "nội_trú", ",", "ngoại_trú", ",", "chỉ", "nhận", "ca", "cấp_cứu", "."	"O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O", "O", "O", "O", "B-PATIENT_ID", "O", "O",	1. Cụm các từ "Bệnh_viện", "Đa_khoa", "huyện", "Đồng_Văn" chỉ chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: bệnh viện** nên được đánh nhãn lần lượt là "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION". 2. Từ "268" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn

`

			"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"	PATIENT_ID nên X ở đây là "268".
34	Val Set - Dòng 14	"Ngày", "29", "-", "7", ",", "anh", "L.", "đi", "thăm", "chị_gái", "bị", "bệnh", "tại", "toà", "nhà", "G", "Bệnh_viện", "Đa_khoa", "tỉnh", "Quảng_Trị", "."	"O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "B-NAME", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O"	1. Cụm từ "29", "-", "7" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 2. Từ "L.": **Tên người có liên quan trực tiếp đến bệnh nhân (để bảo vệ quyền riêng tư, tên thường được viết tắt)**, trong câu là "L." nên cụm “L.” được đánh nhãn là “B-NAME”. Từ “bà” không được đánh nhãn bởi vì: Các danh xưng "ông", "bà", "anh", "chị", "giám đốc", "chủ tịch", … KHÔNG nằm trong tên riêng. 3. Cụm các từ "toà", "nhà", "G", "Bệnh_viện", "Đa_khoa", "tỉnh", "Quảng_Trị" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: bệnh viện** và **Trường hợp khoa, phòng, ban, hội... thuộc một tổ chức, khu vực thì chỉ gán nhãn ORG khi có đầy đủ cả tên của tổ chức, khu vực** nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION".

`

35	Val Set - Dòng 15	"Các", "trường_hợp", "tử_vong", "đều", "có", "bệnh_lý", "nền", "nặng", "với", "82,4%", "có", "nhiều", "hơn", "1", "bệnh_lý", "kèm", "theo", ",", "phổ_biến", "nhất", "là", "suy", "thận", "mạn", "(", "12", ")", ",", "tăng", "huyết_áp", "(", "8", ")", ",", "đái_tháo_đường", "(", "8", ")", ",", "tim_mạch", "(", "7", ")", "và", "ung_thư", "(", "3", ")", ",", "nên", "nguy_cơ", "tử_vong", "rất", "cao", "và", "có_thể", "tiếp_tục", "ghi_nhận", "thêm", "các", "trường_hợp", "tử_vong", "trong", "nhóm", "các", "bệnh_nhân", "này", "trong", "thời_gian", "tới", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "O", "O", "O",	1. Các từ trước từ "TP" không liên quan đến các thực thể đã xác định trong bộ dữ liệu nên được đánh nhãn "O". 2. Cụm các từ "suy", "thận", "mạn" chỉ **Các loại bệnh khác mà bệnh nhân COVID-19** mắc phải nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 3. Cụm các từ "tăng", "huyết_áp" tương tự trương hợp 2 nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 4. Từ "đái_tháo_đường" tương tự trương hợp 2, 3 nên được gán nhãn là: "B SYMPTOM_AND_DISEASE". 5. Từ "ung_thư" tương tự trương hợp 2, 3, 4 nên được gán nhãn là: "B SYMPTOM_AND_DISEASE". 6. Cụm các "nguy_cơ", "tử_vong" tương tự trương hợp 2, 3, 4, 5 nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". (Tuy nhiên, trong bộ dữ liệu PhoNer nãy đã gán nhãn dữ liệu sai vì phải mất 5 token "O" mới đến được cụm "nguy_cơ", "tử_vong", tuy nhiên trong label của bộ dữ liệu họ chỉ đánh 3 token "O" nên cụm "nguy_cơ", "tử_vong" đã bị đánh nhầm nhãn thành "O", "O").

`

			"O", "B-SYMPTOM_AND_DISEASE", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"	
36	Val Set - Dòng 16	"Gia_đình", "tổ_chức", "đám_tang", "và", "có", "tiếp_xúc", "với", "chị", "là", "\"", "bệnh_nhân", "681", "\"", "(", "72", "tuổi", ",", "trú", "phường", "Nại_Hiên_Đông", ")", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "O", "B-AGE", "O", "O", "O", "B-	1. Từ "681" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "681". 2. Cụm "72", "tuổi". Ta chỉ đánh nhãn **Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi")**. Trong một câu, tuổi cần

`

			LOCATION", "I-LOCATION", "O", "O"	được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ "72" là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 3. Cụm các từ "phường", "Nại_Hiên_Đông" chỉ chỉ **Địa chỉ** nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION".
37	Val Set - Dòng 17	"Bệnh_nhân", "quốc_tịch", "Anh", ",", "là", "chuyên_gia", "của", "Tập_đoàn", "Dầu_khí", "VN", "được", "nhập_cảnh", "để", "thực_hiện", "dự_án", "kinh_tế", "."	"O", "O", "O", "O", "O", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "O"	1. Cụm các từ "Tập_đoàn", "Dầu_khí", "VN" chỉ **Tên các công ty, tổ chức nơi bệnh nhân làm việc** nên được gán nhãn lần lượt là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION".
38	Val Set - Dòng 18	"Khoảng", "21h", "đêm", "26", "-", "7", ",", "bệnh_nhân", "sốt", ",", "tức", "ngực", "nên", "đến", "khám", "tại", "phòng", "cấp_cứu", "-", "Trung_tâm", "Y_tế", "Hoà_Vang", "."	"O", "O", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "B-SYMPTOM_AND_DISEASE", "O", "B-SYMPTOM_AND_DISEA	1. Cụm từ "26", "-", "7" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 2. Từ "sốt" chỉ **Các loại bệnh khác mà bệnh nhân COVID-19 mắc phải** nên được gán nhãn là: "B-SYMPTOM_AND_DISEASE". 3. Cụm các từ "tức", "ngực" chỉ **Các loại bệnh khác mà bệnh nhân COVID-19**

`

			SE", "I-SYMPTOM_AND_DISEASE", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O"	**mắc phải** nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 4. Cụm các từ "phòng", "cấp_cứu", "-", "Trung_tâm", "Y_tế", "Hoà_Vang" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng và là địa danh liên quan đến lịch trình di chuyển của bệnh nhân** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION".
39	Val Set - Dòng 19	"Ca", "bệnh", "517", "(", "bệnh_nhân", "517", ")", ":", "nữ", ",", "55", "tuổi", ",", "ở", "phường", "Lê_Hồng_Phong", ",", "TP.", "Quảng_Ngãi", "."	"O", "O", "B-PATIENT_ID", "O", "O", "B-PATIENT_ID", "O", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION"	1. Từ "517" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "517". 2. Từ "517" tiếp theo tương tự trường hợp 1 nên được gán nhãn là "B-PATIENT_ID". 3. Từ "nữ" chỉ **Chỉ gán nhãn giới tính của bệnh nhân và những người liên quan trực tiếp (tiếp xúc) với bệnh nhân**. Trong một câu, giới tính cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân) nên được gán nhãn là "B-GENDER". 4. Cụm "55", "tuổi". Ta **chỉ đánh nhãn Giá**

`

			, "I-LOCATION", "O"	**trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi**"). Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ "55" là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 5. Cụm các từ "phường", "Lê_Hồng_Phonng" chỉ chỉ **Địa chỉ** nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 6. Cụm các từ "TP.", "Quảng_Ngãi" được gán nhãn là "B-LOCATION", "I-LOCATION" là vì: **Tên đơn vị hành chính của quốc gia.**
40	Val Set - Dòng 20	"Trước", "đó", "tối", "7", "-", "7", ",", "tại", "Bệnh_viện", "Đa_khoa", "Bà_Rịa", "-", "Vũng_Tàu", ",", "ba", "bệnh_nhân", "số", "340", ",", "341", "và", "350", "đã", "được", "công_bố", "khỏi", "bệnh", "."	"O", "O", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O", "O", "O", "O", "B-	1. Cụm từ "7", "-", "7" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 2. Cụm các từ "Bệnh_viện", "Đa_khoa", "Bà_Rịa", "-", "Vũng_Tàu" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: bệnh viện, trạm y tế nên** được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION". 3 Từ "340" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số**

`

			PATIENT_ID", "O", "B-PATIENT_ID", "O", "B-PATIENT_ID", "O", "O", "O", "O", "O"	**thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "340". 4. Từ "341" tương tự trường hợp 3 nên được đánh nhãn là "B-PATIENT_ID".
41	Test Set - Dòng 1	"Từ", "24", "-", "7", "đến", "31", "-", "7", ",", "bệnh_nhân", "được", "mẹ", "là", "bà", "H.T.P", "(", "47", "tuổi", ")", "đón", "về", "nhà", "ở", "phường", "Phước_Hoà", "(", "bằng", "xe_máy", ")", ",", "không", "đi", "đâu", "chỉ", "ra", "Tạp_hoá", "Phượng", ",", "chợ", "Vườn_Lài", ",", "phường", "An_Sơn", "cùng", "mẹ", "bán", "tạp_hoá", "ở", "đây", "."	"tags": ["O", "B-DATE", "I-DATE", "I-DATE", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "O", "O", "O", "B-NAME", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-LOCATION", "I-	1. Từ khóa “B-“ thể hiện sự bắt đầu của cụm từ chỉ thực thể và từ khóa “I-“ là thể hiện sự tiếp nối hay có liên quan bên trong của cụm từ sau “B-“. 2. Cụm từ "24", "-", "7" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)**. nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 3. Từ “đến” không liên quan đến các khía cạnh cụ thể nên được đánh nhãn “O”. 4. Cụm từ "31", "-", "7" tương tự 2. 5. **Tên người có liên quan trực tiếp đến bệnh nhân (để bảo vệ quyền riêng tư, tên người thường được viết tắt)** trong câu là "bà", "H.T.P" nên cụm “H.T.P” được đánh nhãn là “B-NAME”. Từ “bà” không được đánh nhãn bởi vì: Các danh xưng "ông", "bà", "anh", "chị", "giám đốc", "chủ tịch", … KHÔNG nằm trong tên riêng. 6. Tiếp theo là cụm "47", "tuổi". Ta chỉ đánh nhãn **Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi")**. Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh

`

			LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "O", "B-JOB", "I-JOB", "O", "O", "O"	nhân). Vì vậy, ta cần đánh nhãn từ "47" vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 7. "phường", "Phước_Hoà" lần lượt được đánh nhãn là "B-LOCATION", "I-LOCATION" vì các thực thể này: **Các thực thể này chỉ địa chỉ**. 8. "Tạp_hoá", "Phượng" lần lượt được đánh nhãn là "B-LOCATION", "I-LOCATION" vì các thực thể này: **Mang tính thương mại: nhà hàng, quán ăn, khách sạn, chợ, siêu thị.** 9. "chợ", "Vườn_Lài" tương tự 8. 10. "phường", "An_Sơn" tương tự 7. 11. Các từ "bán", "tạp_hoá" lần lượt được đánh nhãn là "B-JOB", "I-JOB" vì: **Chỉ gán nhãn nghề nghiệp của bệnh nhân và các cá nhân có liên quan trực tiếp (tiếp xúc, gặp mặt, ở gần).** Ngoài ra, những từ chỉ nghề nghiệp cần phải được gắn với 1 cá nhân nhất định trong câu (có tên, có mã bệnh nhân).
42	Test Set - Dòng 2	"Bác_sĩ", "Trần_Thanh_Linh", ",", "từ", "Bệnh_viện", "Chợ_Rẫy", "chi_viện", "phụ_trách", "đơn_nguyên", "hồi_sức", "tích_cực", ",", "cho", "biết", "\"", "bệnh_nhân", "416", "\"", "vẫn", "đang", "duy_trì", "ECMO", ",", "thở", "máy", ",", "hiện", "xơ", "phổi",	"O", "O", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-	1. "Bệnh_viện", "Chợ_Rẫy": **Ở trường hợp này Cần chú ý tới ngữ cảnh để xác định một thực thể có phải là LOCATION hay không (tránh nhập nhằng với ORGANIZATION)**. Vì "Bác_sĩ", "Trần_Thanh_Linh", "từ" là **chỉ nơi công tác cảu bác sĩ nên "Bệnh_viện", "Chợ_Rẫy" trong ngữ cảnh này được xem như một tổ chức** => gán nhãn "B-ORGANIZATION", "I-

`

		"rất", "nhiều", "."	PATIENT_ID", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O"	ORGANIZATION". 3. Từ "416" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "416". 4. Cụm các từ "xơ", "phổi", "rất", "nhiều" chỉ **Triệu chứng liên quan tới bệnh nhân COVID-19** nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE".
43	Test Set - Dòng 3	"Theo", "đó", ",", "Sở", "Y_tế", "Bình_Thuận", "cho", "biết", "sau", "khi", "xác_định", "bệnh_nhân", "số", "34", "(", "nữ_giới", "51", "tuổi", ",", "từ", "Mỹ", "về", "Việt_Nam", "ngày", "29", "-", "2", "có", "quá_cảnh", "Qatar", ")", ",", "Trung_tâm", "Kiểm_soát", "bệnh_tật", "Bình_Thuận", "đã", "điều_tra", "dịch_tễ", ",", "khoanh", "vùng", ",", "khử", "khuẩn", ",", "tiến_hành", "cách_ly",	"O", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "B-AGE", "O", "O",	1. Cụm các từ "Sở", "Y_tế", "Bình_Thuận" chỉ **Tên các cơ quan liên quan tới việc xử lý dịch tễ** nên lần lượt được gán nhãn là "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION". 2. Từ "34" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "34". 3. Từ "nữ_giới" chỉ **Chỉ gán nhãn giới tính của bệnh nhân và những người liên quan trực tiếp (tiếp xúc) với bệnh nhân.** Trong

`

		"người", "liên_quan", "đến", "ca", "bệnh", "số", "34", "."	"O", "B-LOCATION", "O", "B-LOCATION", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "B-LOCATION", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O"	một câu, giới tính cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân) nên được gán nhãn là "B-GENDER". 4. Cụm "51", "tuổi". Ta **chỉ đánh nhãn Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi").** Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ "51" là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 5. Từ "Mỹ" chỉ **Tên quốc gia: "Mỹ"** nên được đánh nhãn là "B-LOCATION". 6. Từ "Việt_Nam" tương tự trương hợp 5 nên được đánh nhãn là "B-LOCATION". 7. Cụm từ "29", "-", "2" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)**. nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 8. Từ "Qatar" tương tự trương hợp 5, 6 nên được đánh nhãn là "B-LOCATION". 9. Cụm các từ Trung_tâm", "Kiểm_soát", "bệnh_tật", "Bình_Thuận" chỉ **Tên các cơ quan liên quan tới việc xử lý dịch tễ** nên lần lượt được gán nhãn là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION". 10. Từ "34" tiếp theo tương tự trường hợp 2 nên được gán nhãn là "B-PATIENT_ID".

`

44	Test Set - Dòng 4	"Bệnh_nhân", "218", ":", "nữ", ",", "43", "tuổi", ",", "quốc_tịch", "Việt_Nam", ",", "địa_chỉ", "tại", "Phú_Xá", ",", "Thái_Nguyên", ",", "về", "nước", "trên", "chuyến", "bay", "SU290", "(", "số", "ghế", "46", "G", ")", "ngày", "25", "-", "3", ",", "sau", "nhập_cảnh", "được", "cách_ly", "tập_trung", "tại", "Đại_học", "FPT", "ở", "Láng", "-", "Hoà_Lạc", "(", "Hà_Nội", ")", ".", "Từ", "31", "-", "3", "bệnh_nhân", "được", "cách_ly", ",", "điều_trị", "tại", "Bệnh_viện", "Bệnh", "nhiệt_đới", "trung_ương", "cơ_sở", "2", "."	"O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-LOCATION", "O", "B-LOCATION", "O", "O", "O", "O", "O", "O", "B-TRANSPORTATION", "O", "O", "O", "O", "O", "O", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "O", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "O", "B-LOCATION"	1. Từ "218" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "218". 2. Từ "nữ" chỉ **Chỉ gán nhãn giới tính của bệnh nhân và những người liên quan trực tiếp (tiếp xúc) với bệnh nhân.** Trong một câu, giới tính cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân) nên được gán nhãn là "B-GENDER". 3. Cụm "43", "tuổi". Ta **chỉ đánh nhãn Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi").** Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ “43” là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 4. Từ "Việt_Nam" **KHÔNG** được gán nhãn "B-LOCATION" vì Không gán nhãn quốc tịch. 5. Từ "Phú_Xá" chỉ **Địa chỉ: cấp bậc đơn vị hành chính** nên được gán nhãn là "B-LOCATION". 6. Từ "Thái_Nguyên" chỉ **Tên đơn vị hành chính của quốc gia** nên được gán nhãn là "B-LOCATION". 7. Từ "SU290" chỉ **nhãn biển số, số hiệu**

`

			, "O", "B-LOCATION", "O", "O", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O"	**của loại phương tiện di chuyển, không gán nhãn loại phương tiện di chuyển** nên được gán nhãn là B-TRANSPORTATION. 8. Cụm từ "25", "-", "3" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)**. nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 9. Cụm các từ "Đại_học", "FPT" chỉ **Địa chỉ** nên được gán nhãn là: "B-LOCATION", "I-LOCATION". 10. Từ "Láng" tương tự trường hợp 5 nên được gán nhãn là "B-LOCATION". 11. Từ "Hoà_Lạc" tương tự trường hợp 5, 10 nên được gán nhãn là "B-LOCATION". 12. Từ "Hà_Nội" chỉ Tên đơn vị hành chính của quốc gia nên được gán nhãn là "B-LOCATION". 13. Cụm từ "31", "-", "3" tương tự trường hợp 8 nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 14. Cụm các từ "Bệnh_viện", "Bệnh", "Nhiệt_đới", "Trung_ương", "cơ_sở", "2" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: bệnh viện, trạm y tế** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION".

`

45	Test Set - Dòng 5	"Ông", "cùng", "4", "người", "khác", "hôm", "4/3", "từ", "Malaysia", "về", "sân_bay", "Tân_Sơn_Nhất", "trên", "chuyến", "bay", "VJ", "826", "."	"O", "O", "O", "O", "O", "O", "B-DATE", "O", "B-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "O", "O", "B-TRANSPORTATION", "I-TRANSPORTATION", "O"	1. Từ "4/3" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date. 2.Từ "Malaysia" chỉ **Tên quốc gia** nên được đánh nhãn là B-LOCATION. 3. Cụm các từ "sân_bay", "Tân_Sơn_Nhất" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: sân bay** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION". 4. Cụm các từ "VJ", "826" chỉ **nhãn biển số, số hiệu của loại phương tiện di chuyển, không gán nhãn loại phương tiện di chuyển** nên lần lượt được gán nhãn là "B-TRANSPORTATION", "I-TRANSPORTATION".

`

46	Test Set - Dòng 6	"Ca", "bệnh", "1.035", ":", "nữ", "34", "tuổi", ",", "ở", "Nam_Sách", ",", "Hải_Dương", ",", "từ", "Đài_Loan", "nhập_cảnh", "sân_bay", "Cam_Ranh", "ngày", "7", "-", "8", "trên", "chuyến", "bay", "VJ2849", ",", "được", "cách_ly", "tập_trung", "tại", "Trung_tâm", "Giáo_dục", "quốc_phòng", "an_ninh", ",", "ĐH", "Nha_Trang", ",", "Khánh_Hoà", "."	"O", "O", "B-PATIENT_ID", "O", "B-GENDER", "B-AGE", "O", "O", "O", "B-LOCATION", "O", "B-LOCATION", "O", "O", "B-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "O", "B-TRANSPORTATION", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION"	1. Từ "1.035" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự**. Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "1.035". 2. Từ "nữ" chỉ **Chỉ gán nhãn giới tính của bệnh nhân và những người liên quan trực tiếp (tiếp xúc) với bệnh nhân**. Trong một câu, giới tính cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân) nên được gán nhãn là "B-GENDER". 3. Cụm "34", "tuổi". Ta **chỉ đánh nhãn Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi").** Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ “34” là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 4. Từ "Nam_Sách" chỉ **Địa chỉ: cấp bậc đơn vị hành chính** là một thực thể riêng biệt nên được gán nhãn là "B-LOCATION". 5. Từ "Hải_Dương" chỉ **Tên đơn vị hành chính của quốc gia** nên được gán nhãn là "B-LOCATION". 6. Từ "Đài_Loan" chỉ **Tên quốc gia** nên được đánh nhãn là "B-LOCATION". 7. Cụm các từ "sân_bay", "Cam_Ranh" chỉ

`

			, "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "O"	**Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: sân bay** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION". 8. Cụm từ "7", "-", "8" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng).** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 9. Từ "VJ2849" chỉ **nhãn biển số, số hiệu của loại phương tiện di chuyển, không gán nhãn loại phương tiện di chuyển** nên được gán nhãn là B-TRANSPORTATION. 10. Cụm các từ "Trung_tâm", "Giáo_dục", "quốc_phòng", "an_ninh" chỉ **Địa chỉ** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION". 11. Cụm các từ "ĐH", "Nha_Trang" tương tự trường hợp 10 nên được lần lượt gán nhãn là: "B-LOCATION", "I-LOCATION". 12. Từ "Khánh_Hòa" chỉ **Tên đơn vị hành chính của quốc gia** nên được gán nhãn là "B-LOCATION".
47	Test Set - Dòng 7	["Khi", "vào", "khoa", ",", "các", "bác_sĩ", "nhận_định", "tình_trạng", "viêm", "phổi", "trên", "bệnh_nhân", "64", "tuổi", "tuổi", ",", "sức_khoẻ", "suy_kiệt", "."	"O", "O", "O", "O", "O", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O",	1. Cụm các từ "viêm", "phổi" chỉ **Các loại bệnh khác mà bệnh nhân COVID-19 mắc phải** nên được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 2. Cụm các từ "64", "tuổi" **KHÔNG** được gán nhãn vì không có bệnh nhân xác định đi kèm (Không có tên, mã bệnh nhân). 3. Cụm các từ "sức_khoẻ", "suy_kiệt"

`

			"O", "O", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O"	tương tự trường hợp 1 nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE".

`

48	Test Set - Dòng 8	"Các", "bệnh_nhân", "được", "công_bố", "khỏi", "bệnh", "bao_gồm", ":", "bệnh_nhân", "21", "(", "nam", ",", "61", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", ";", "bệnh_nhân", "72", "(", "nữ", ",", "25", "tuổi", ",", "quốc_tịch", "Pháp", ")", ";", "bệnh_nhân", "84", "(", "nam", ",", "21", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", ";", "bệnh_nhân", "111", "(", "nữ", "25", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", ";", "bệnh_nhân", "116", "(", "nam", ",", "29", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", ";", "bệnh_nhân", "136", "(", "nữ", ",", "23", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", ";", "bệnh_nhân", "137", "(", "nam", ",", "36", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", ";", "bệnh_nhân", "192", "(", "nữ", ",", "23", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", ";", "bệnh_nhân", "197", "(", "nam", ",", "41", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", ";",	"O", "O", "O", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "B-AGE", "O", "O",	1. Từ "21" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "21". 2. Từ "nam" chỉ **Chỉ gán nhãn giới tính của bệnh nhân và những người liên quan trực tiếp (tiếp xúc) với bệnh nhân.** Trong một câu, giới tính cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân) nên được gán nhãn là "B-GENDER". 3. Cụm "61", "tuổi". Ta **chỉ đánh nhãn Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi")**. Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ "61" là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 4. Từ "Việt_Nam" **KHÔNG** được gán nhãn là "B-LOCATION" vì KHÔNG được gán quốc tịch. 5. Từ "72" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 6. Từ "nữ" tương tự trường 2 nên được gán nhãn là "B-GENDER". 7. Từ "25" tương tự trường 3 nên được gán nhãn là "B-AGE". 8. Từ "84" tương tự trường 1 nên được gán

`

		"bệnh_nhân", "200", "(", "nữ", ",", "61", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", ";", "bệnh_nhân", "222", "(", "nữ", ",", "28", "tuổi", ",", "quốc_tịch", "Việt_Nam", ")", "."	"O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O",	nhãn là "B-PATIENT_ID". 9. Từ "nam" tương tự trường 2 nên được gán nhãn là "B-GENDER". 10. Từ "21" tương tự trường 3 nên được gán nhãn là "B-AGE". 11.Từ "111" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 12. Từ "nữ" tương tự trường 2 nên được gán nhãn là "B-GENDER". 13. Từ "25" tương tự trường 3 nên được gán nhãn là "B-AGE". 14. Từ "116" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 15. Từ "nam" tương tự trường 2 nên được gán nhãn là "B-GENDER". 16. Từ "29" tương tự trường 3 nên được gán nhãn là "B-AGE". 17. Từ "136" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 18. Từ "nữ" tương tự trường 2 nên được gán nhãn là "B-GENDER". 19. Từ "23" tương tự trường 3 nên được gán nhãn là "B-AGE". 17. Từ "136" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 18. Từ "nữ" tương tự trường 2 nên được gán nhãn là "B-GENDER". 19. Từ "23" tương tự trường 3 nên được gán nhãn là "B-AGE". 20. Từ "137" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 21. Từ "nam" tương tự trường 2 nên được gán nhãn là "B-GENDER". 22. Từ "36" tương tự trường 3 nên được gán

`

			"O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "O"	nhãn là "B-AGE". 23. Từ "192" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 24. Từ "nữ" tương tự trường 2 nên được gán nhãn là "B-GENDER". 25. Từ "23" tương tự trường 3 nên được gán nhãn là "B-AGE". 26. Từ "197" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 27. Từ "nam" tương tự trường 2 nên được gán nhãn là "B-GENDER". 28. Từ "41" tương tự trường 3 nên được gán nhãn là "B-AGE". 29. Từ "200" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 30. Từ "nữ" tương tự trường 2 nên được gán nhãn là "B-GENDER". 31. Từ "61" tương tự trường 3 nên được gán nhãn là "B-AGE". 32. Từ "222" tương tự trường 1 nên được gán nhãn là "B-PATIENT_ID". 33. Từ "nữ" tương tự trường 2 nên được gán nhãn là "B-GENDER". 34. Từ "28" tương tự trường 3 nên được gán nhãn là "B-AGE".

`

49	Test Set - Dòng 9	"Liên_quan", "các", "trường_hợp", "tiếp_xúc", "người", "nhiễm", "COVI", "-", "19", ",", "sáng", "6", "-", "8", ",", "ông", "Nguyễn_Văn_Định", "-", "giám_đốc", "Trung_tâm", "Kiểm_soát", "bệnh_tật", "(", "CDC", ")", "Nghệ_An", "-", "cho", "biết", "kết_quả", "xét_nghiệm", "với", "ông", "T.V.D.", "(", "ngụ", "xã", "Viên", "Thành", ",", "huyện", "Yên_Thành", ")", "và", "3", "người", "tiếp_xúc", "với", "ông", "D.", "đều", "cho", "kết_quả", "âm_tính", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-DATE", "I-DATE", "I-DATE", "O", "O", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "O", "B-NAME", "O", "O", "B-LOCATION", "I-LOCATION", "I-	1. Cụm từ "6", "-", "8" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng).** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 2. Cụm các từ "Trung_tâm", "Kiểm_soát", "bệnh_tật", "(", "CDC", ")", "Nghệ_An" chỉ **Tên các cơ quan liên quan tới việc xử lý dịch tễ** nên lần lượt được gán nhãn là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION". 3. Từ "T.V.D": **Tên người có liên quan trực tiếp đến bệnh nhân (để bảo vệ quyền riêng tư, tên người thường được viết tắt)** nên được đánh nhãn là "B-NAME". 4. Cụm các từ "xã", "Viên", "Thành" chỉ **Địa chỉ** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 5. Cụm các từ "huyện", "Yên_Thành" tương tự trường hợp 4 nên được gán nhãn lần lượt là "B-LOCATION", "I-LOCATION". 6. Từ "D." tư tượng trường hợp 3 nên được gán nhãn là "B-NAME".

`

			LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "O", "O", "O", "O", "O", "O", "B-NAME", "O", "O", "O", "O", "O"	
50	Test Set - Dòng 10	"Theo", "đó", "bệnh_nhân", "tên", "N.M.C.", ",", "là", "nhân_viên", "ngân_hàng", "tại", "38", "Hàng", "Da", ",", "phường", "Hàng", "Bông", ",", "quận", "Hoàn_Kiếm", "."	"O", "O", "O", "O", "B-NAME", "O", "O", "B-JOB", "I-JOB", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION"	1. Từ "N.M.C": **Tên bệnh nhân (để bảo vệ quyền riêng tư, tên bệnh nhân COVID-19 thường được viết tắt)** nên được đánh nhãn là "B-NAME". 2. Các từ "nhân_viên", "ngân_hàng" lần lượt được đánh nhãn là "B-JOB", "I-JOB" vì: C**hỉ gán nhãn nghề nghiệp của bệnh nhân và các cá nhân có liên quan trực tiếp (tiếp xúc, gặp mặt, ở gần).** Ngoài ra, những từ chỉ nghề nghiệp cần phải được gắn với 1 cá nhân nhất định trong câu (có tên, có mã bệnh nhân). 3. Cụm các từ "38", "Hàng", "Da" chỉ **Địa chỉ: Số nhà phải bao gồm cả tên đường để tránh bị nhập nhằng** nên được gán nhãn lần lượt là "B-LOCATION", "I-LOCATION", "I-LOCATION". 4. Cụm các từ "phường", "Hàng", "Bông" chỉ **Địa chỉ: cấp bậc đơn vị hành chính** nên được gán nhãn lần lượt là "B-LOCATION", "I-LOCATION", "I-

`

			, "O"	LOCATION". 5. Cụm các từ "quận", "Hoàn_Kiếm" tư tượng trường hợp 4 nên được gán nhãn lần lượt là "B-LOCATION", "I-LOCATION".
51	Test Set - Dòng 11	"Theo", "đó", ",", "ca", "bệnh", "785", "(", "bệnh_nhân", "785", ")", "là", "nam", ",", "42", "tuổi", ",", "có", "địa", "chỉ", "tại", "Đức", "Thượng", ",", "Hoài", "Đức", ",", "Hà", "Nội", "."	"O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "O", "B-PATIENT_ID", "O", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION"	1. Từ "785" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "785". 2. Từ "785" tiếp theo tương tự trường hợp 1 nên được gán nhãn là "B-PATIENT_ID". 3. Từ "nam" chỉ **Chỉ gán nhãn giới tính của bệnh nhân và những người liên quan trực tiếp (tiếp xúc) với bệnh nhân.** Trong một câu, giới tính cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân) nên được gán nhãn là "B-GENDER". 4. Cụm "42", "tuổi". Ta **chỉ đánh nhãn Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi").** Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ "42" là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh.

`

			, "O"	5. Cụm các từ "Đức", "Thượng" chỉ **Địa chỉ: cấp bậc đơn vị hành chính** nên được gán nhãn lần lượt là "B-LOCATION", "I-LOCATION". 6. Cụm các từ "Hoài", "Đức" chỉ **Địa chỉ: cấp bậc đơn vị hành chính** nên được gán nhãn lần lượt là "B-LOCATION", "I-LOCATION". 7. Từ "Hà", "Nội" là **Tên đơn vị hành chính của quốc gia** nên được gán nhãn lần lượt là "B-LOCATION", "I-LOCATION".
52	Test Set - Dòng 12	"Phát_biểu", "tại", "cuộc", "họp", ",", "Chủ_tịch", "UBND", "tỉnh", "Thanh_Hoá", "Nguyễn_Đình_Xứng", "khẳng_định", "việc", "xuất_hiện", "trường_hợp", "bà", "Đ.T.H.", "tại", "Sầm_Sơn", "đã", "cảnh_báo", "lỗ_hổng", "trong", "công_tác", "giám_sát", ",", "cách_ly", "các", "ca", "bệnh", "ở", "Thanh_Hoá", "."	"O", "O", "O", "O", "O", "O", "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "B-NAME", "O", "B-LOCATION", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-	1. Cụm các từ "UBND", "tỉnh", "Thanh_Hoá" chỉ **Tên các cơ quan chính phủ: bộ ngành, uỷ ban nhân dân** nên được gán nhãn lần lượt là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION". 2. Từ "Đ.T.H": **Tên bệnh nhân (để bảo vệ quyền riêng tư, tên bệnh nhân COVID-19 thường được viết tắt)** nên được đánh nhãn là "B-NAME". 3. Từ "Sầm_Sơn" là **Địa chỉ: cấp bậc đơn vị hành chính** nên được gán nhãn lần lượt là "B-LOCATION". 4. Từ "Thanh_Hóa" **Chỉ đơn vị hành chính cấp Quốc Gia** nên được gán nhãn lần lượt là "B-LOCATION".

`

			LOCATION", "O"	
53	Test Set - Dòng 13	"Hiện", "hai", "bệnh_nhân", "điều_trị", "tại", "Bệnh_viện", "Lao", "và", "Bệnh", "phổi", "Thành_phố", "Cần_Thơ", "."	"O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O"	1. Cụm các từ "Bệnh_viện", "Lao", "và", "Bệnh", "phổi", "Thành_phố", "Cần_Thơ" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: bệnh viện, trạm y tế** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION".
54	Test Set - Dòng 14	"Trong", "thời_gian", "ở", "đây", ",", "em", "đi", "chơi", "công_viên", "SunWorld", "Đà_Nẵng", ",", "siêu_thị", "Lotte_Mart", "và", "ăn_ở", "một_số", "quán", "."	"O", "O", "O", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "O", "B-LOCATION"	1. Cụm các từ "công_viên", "SunWorld", "Đà_Nẵng" **chỉ Tên các địa điểm mang tính thương mại** nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 2. Cụm các từ "siêu_thị", "Lotte_Mart" tương tự trường hợp 1 nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION".

`

			, "I-LOCATION", "O", "O", "O", "O", "O"	
55	Test Set - Dòng 15	"Tối", "13/8", ",", "Ban", "chỉ_đạo", "phòng_chống", "dịch_bệnh", "Covid", "-", "19", "Quảng_Trị", "công_bố", "thông_tin", "trên", "cùng", "hành_trình", "của", "\"", "bệnh_nhân", "904", "\"", "."	"O", "B-DATE", "O", "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O", "O", "O", "O", "O", "B-PATIENT_ID", "O", "O"	1. Từ "13/8" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng)** nên được đánh nhãn là B-Date. 2. Cụm các từ "Ban", "chỉ_đạo", "phòng_chống", "dịch_bệnh", "Covid", "-", "19", "Quảng_Trị" chỉ **Tên các cơ quan liên quan tới việc xử lý dịch tễ** nên lần lượt được gán nhãn là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION". 3. Từ "904" được gán nhãn B-PATIENT_ID là vì T**ại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "904".

`

56	Test Set - Dòng 16	"Bệnh_nhân", "435", ",", "nữ", ",", "29", "tuổi", ",", "đang", "tạm_trú", "phường", "An_Hải_Đông", ",", "quận", "Sơn_Trà", ",", "TP", "Đà_Nẵng", "có", "quê", "ở", "xã", "An_Hoà", ",", "huyện", "Quỳnh_Lưu", "."	"O", "B-PATIENT_ID", "O", "B-GENDER", "O", "B-AGE", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O"	1. Từ "435" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "435". 2. Từ "nữ" chỉ **Chỉ gán nhãn giới tính của bệnh nhân và những người liên quan trực tiếp (tiếp xúc) với bệnh nhân.** Trong một câu, giới tính cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân) nên được gán nhãn là "B-GENDER". 3. Cụm "29", "tuổi". Ta **chỉ đánh nhãn Giá trị tuổi của bệnh nhân và những người có liên quan (tiếp xúc). (KHÔNG gán nhãn từ "tuổi").** Trong một câu, tuổi cần được gán với một đối tượng được định danh (có tên, có mã bệnh nhân). Vì vậy, ta cần đánh nhãn từ “29” là B-AGE vì từ này vừa thể hiện giá trị, vừa thỏa mãn điều kiện đã được gán với một đối tượng đã có định danh. 4. Cụm các từ "phường", "An_Hải_Đông" chỉ **Địa chỉ: cấp bậc đơn vị hành chính** nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 5. Cụm các từ "quận", "Sơn_Trà" tương tự trường hợp 4 nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 6. Cụm các từ "TP", "Đà_Nẵng" được gán

`

				nhãn là "B-LOCATION", "I-LOCATION" là vì: **Tên đơn vị hành chính của quốc gia.** 7. Cụm các từ "xã", "An_Hòa" tương tự trường hợp 4 nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 8. Cụm các từ "huyện", "Quỳnh_Lưu" tương tự trường hợp 4 nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION".
57	Test Set - Dòng 17	"Trước", "đó", ",", "khi", "điều_trị", "tại", "hai", "bệnh_viện", "ở", "Đà_Nẵng", ",", "bệnh_nhân", "viêm", "phổi", "nặng", "trên", "10", "năm", ",", "tràn", "khí", "màng", "phổi", "đã", "dẫn_lưu", "."	"O", "O", "O", "O", "O", "O", "O", "O", "O", "B-LOCATION", "O", "O", "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O", "O", "O", "O", "B-	1. Từ "Đà_Nẵng" được gán nhãn là "B-LOCATION" là vì: **Tên đơn vị hành chính của quốc gia**. 2. Cụm các từ "viêm", "phổi", "nặng" chỉ **Các triệu chứng của COVID-19 mắc phải** nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE". 3. Cụm các từ "tràn", "khí", "màng", "phổi", "đã", "dẫn_lưu" tương tự trường hợp 2 nên nên lần lượt được gán nhãn là: "B-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE".

`

			SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "I-SYMPTOM_AND_DISEASE", "O"	
58	Test Set - Dòng 18	"Sáng", "24", "-", "2", ",", "Trung_tâm", "kiểm_soát", "bệnh_tật", "tỉnh", "Thừa_Thiên_Huế", "đã", "tổ_chức", "họp_báo", "để", "công_bố", "thông_tin", "về", "nguyên_nhân", "tử_vong", "của", "nữ_sinh", "lớp", "12", "ở", "xã", "Vinh_Hiền", ",", "huyện", "Phú_Lộc", "sau", "khi", "bệnh_nhân", "này", "có", "triệu_chứng", "ho", ",", "sốt", ",", "ói", "."	"O", "B-DATE", "I-DATE", "I-DATE", "O", "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "O", "O", "O",	1. Cụm từ "24", "-", "2" chỉ **Ngày trong tiếng Việt thường có dạng X/Y, X-Y (X là ngày, Y là tháng).** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 2. Cụm các từ "Trung_tâm", "kiểm_soát", "bệnh_tật", "tỉnh", "Thừa_Thiên_Huế" chỉ **Tên các cơ quan liên quan tới việc xử lý dịch tễ** nên lần lượt được gán nhãn là: "B-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION", "I-ORGANIZATION". 3. Cụm các từ "xã", "Vinh_Hiền" chỉ **Địa chỉ:cấp bậc đơn vị hành chính** nên lần

`

			"O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-LOCATION", "O", "O", "O", "O", "O", "O", "B-SYMPTOM_AND_DISEASE", "O", "B-SYMPTOM_AND_DISEASE", "O", "B-SYMPTOM_AND_DISEASE", "O"	lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 4. Cụm các từ "huyện", "Phú_Lộc" tương tự trường hợp 4 nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 5. Từ "ho" chỉ **Các triệu chứng của COVID-19 mắc phải** nên được gán nhãn là "B-SYMPTOM_AND_DISEASE". 6. Từ "sốt" chỉ **Các triệu chứng của COVID-19 mắc phải** nên được gán nhãn là "B-SYMPTOM_AND_DISEASE". 7. Từ "ói" chỉ **Các triệu chứng của COVID-19 mắc phải** nên được gán nhãn là "B-SYMPTOM_AND_DISEASE".
59	Test Set - Dòng 19	"Thiếu_nữ", "trú", "phường", "Nghĩa", "Trung", ",", "TP", "Gia_Nghĩa", ",", "có", "yếu_tố", "dịch_tễ", "là", "tiếp_xúc", "với", "người", "từ", "vùng", "dịch", "trở", "về", "."	"O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "O", "B-LOCATION", "I-	1. Cụm các từ "phường", "Nghĩa", "Trung" chỉ **Địa chỉ: cấp bậc đơn vị hành chính** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION". 2. Cụm các từ "TP", "Gia_Nghĩa" được gán nhãn là "B-LOCATION", "I-LOCATION" là vì: **Tên đơn vị hành chính của quốc gia.**

`

			LOCATION", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"	
60	Test Set - Dòng 20	"Ca", "số", "20", "và", "161", "đang", "được", "điều_trị", "tại", "Bệnh", "nhiệt_đới", "Trung_ương", "cơ_sở", "2", "."	"O", "O", "B-PATIENT_ID", "O", "B-PATIENT_ID", "O", "O", "O", "O", "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "O"	1. Từ "20" được gán nhãn B-PATIENT_ID là vì **Tại Việt Nam, bệnh nhân COVID-19 được định danh bằng số thứ tự. X: số thứ tự.** Và Bệnh nhân X, Bệnh nhân thứ X, bệnh nhân số X, BN X, => chỉ gán "X" với nhãn PATIENT_ID nên X ở đây là "20". 2. Từ "161" tương tự trường hợp 1 nên được gán nhãn là "B-PATIENT_ID". 3. Cụm các từ "Bệnh", "Nhiệt_đới", "Trung_ương", "cơ_sở", "2" chỉ **Tên các công trình xây dựng, công trình kiến trúc mang tính công cộng: bệnh viện, trạm y tế** nên được gán nhãn lần lượt là: "B-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION", "I-LOCATION".

Table 2.3: Bảng phân tích 60 mẫu dữ liệu trong bộ dữ liệu. 20 mẫu trong tập Train, 20 mẫu trong tập Val và 20 mẫu trong tập Test.

`

CHƯƠNG 3: PHƯƠNG PHÁP TIẾP CẬN

3.1 Giới thiệu PhoBERT

PhoBERT [3] là mô hình ngôn ngữ tiền huấn luyện đầu tiên cho tiếng Việt, được phát triển bởi VinAI Reasearch. PhoBERT gồm hai phiên bản là base và large. Trong đồ án môn học này, chúng tôi sẽ sử dụng phiên bản PhoBERT-base để trích xuất đặc trưng cho mô hình. PhoBERT-base được tiền huấn luyện trên 20GB dữ liệu mức độ từ, bao gồm các văn bản trên Wikipedia và tin tức. Kiến trúc của PhoBERT-base dựa trên kiến trúc BERT-base [4] và phương pháp tiền huấn luyện dựa trên RoBERTa [5], giúp tối ưu quá trình huấn luyện BERT với hiệu quả mạnh mẽ hơn.

3.2 Kiến trúc mô hình

Mô hình sử dụng gồm 12 khối transformer encoder và một lớp point-wise linear layer ở cuối để dự đoán xác suất các lớp của các token đầu vào. Kiến trúc cụ thể như hình sau:

`

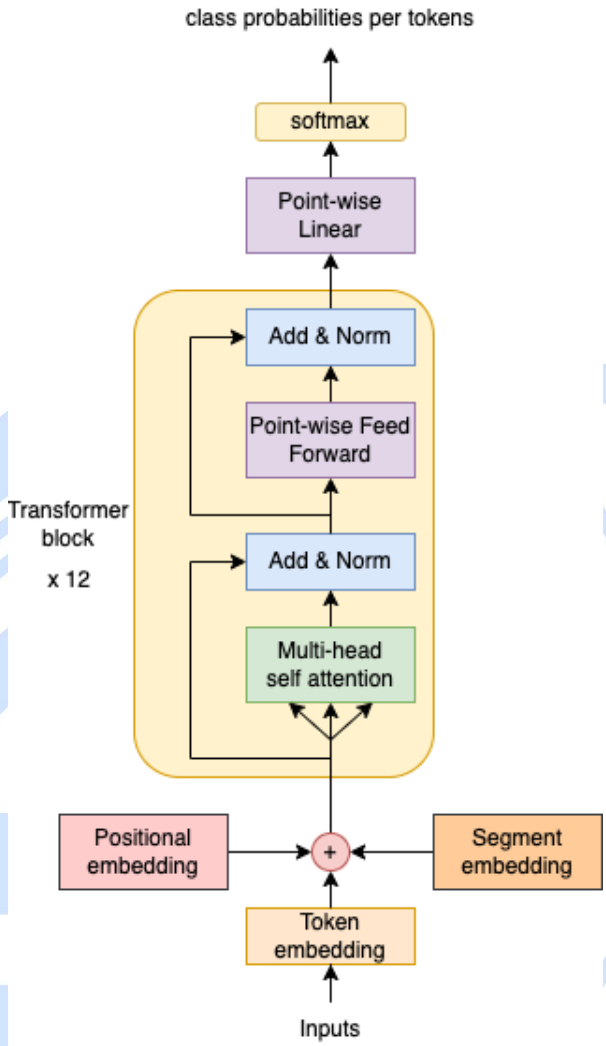

Figure 3.1: Kiến trúc mô hình PhoBERT cho cho bài toán NER

3.2.1 Embeddings

Để tạo được embeddings cho tokens của chuỗi đầu vào, mô hình cần 3 embedding layers sau:

- **Token embedding layer:** Gồm 64001 (kích thước vocab) embedding vector, mỗi vector $\in \mathbb{R}^{768}$. Layer này giúp chuyển đổi các token id (một số nguyên) của chuỗi đầu vào thành embedding vector.

`

- **Positional embedding layer:** Gồm 258 (max_position_embeddings) embedding vectors, mỗi vector $\in \mathbb{R}^{768}$. Layer này giúp biểu diễn thông tin vị trí của token trong chuỗi.
- **Token type embedding layer**: Gồm 1 (Kích thước token type vocab) embedding vector $\in \mathbb{R}^{768}$. Trước đây, mô hình BERT sẽ có kích thước token type vocab là 2 để phục vụ cho tác vụ next sentence prediction. Sau này, RoBERTa đã cải thiện BERT và loại bỏ tác vụ này lúc pretrained, nên việc biểu diễn token này thuộc câu nào là không cần thiết, nên có lẽ ta nên bỏ luôn cả layer này nếu không cần thiết trong quá trình finetune. Tuy nhiên, RoBERTa vẫn giữ lại layer này nên ta vẫn sẽ đề cập.

Với mỗi token, ta sẽ tính:

$$embedding = token\ embedding + positional\ embedding + token\ type\ embedding$$

Như vậy với chuỗi đầu vào có max_seq_len tokens, ta sẽ có một ma trận $E \in \mathbb{R}^{max\_seq\_len \times 768}$.

3.2.2 MHSA

MHSA [6] giúp mô hình có thể học được nhiều kiểu quan hệ giữa các từ với nhau, tức là nó sẽ xem xét câu đầu vào ở nhiều góc độ hơn. MHSA gồm các layer sau:

- **Query linear layer:** Là 1 linear layer có số node là 768, đầu vào là 768. Layer này biến đổi mỗi embeding vector thành một query vector hay một cách tổng quát hơn là biến ma trận embedding thành ma trận query Q .d
- **Key linear layer**: Số lượng node, số lượng đầu vào tương tự query linear layer. Layer này biến đổi ma trận embedding thành ma trận key K .
- **Value linear layer**: Số lượng node, số lượng đầu vào tương tự query linear layer. Layer này biến đổi ma trận embedding thành ma trận value V .

`

- **Dropout layer 1**: dropout rate 0.1, layer này bỏ đi 10% số lượng node sau layer masked softmax giúp tránh hiện tượng overfitting, mô hình generalize hơn.
- **Output linear layer**: Số lượng node, số lượng đầu vào tương tự query linear layer. Layer này giúp tổng hợp hợp thông tin từ nhiều head.
- **Layer normalization**: chuẩn hóa các phân phối của các lớp trung gian, cho phép gradients mượt hơn, huấn luyện nhanh hơn và tổng quát hoá tốt hơn.
- **Droupout layer 2**: dropout rate 0.1

Cách hoạt động như Figure 3.3: Embedding matrix sau khi qua query, key, value, linear layer cho ra 3 ma trận lần lượt là $Q, K, V \in \mathbb{R}^{max\_seq\_len \times 768}$. Với mỗi ma trận, ta chia ra thành $n\_head = 12$ ma trận con $\in \mathbb{R}^{max\_seq\_len \times 64}$ theo chiều embed size ($Q_i, K_i, V_i \in \mathbb{R}^{max\_seq\_len \times 64}$). Ở mỗi head, ta tính Attention score $A_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{head\_size}} + mask\right) \in \mathbb{R}^{max\_seq\_len \times max\_seq\_len}$. Sau đó, A_i đi qua dropout layer 1 và nhân với V_i tạo thành ma trận $\in \mathbb{R}^{max\_seq\_len \times 64}$. Ma trận đầu ra A_i của mỗi head được concat lại tạo thành ma trận $\in \mathbb{R}^{max\_seq\_len \times 768}$ rồi qua Output linear layer. Đầu ra tiếp tục cộng với đầu vào ban đầu của MHSA (residual connection), đi qua layer normalization và dropout layer 2.

`

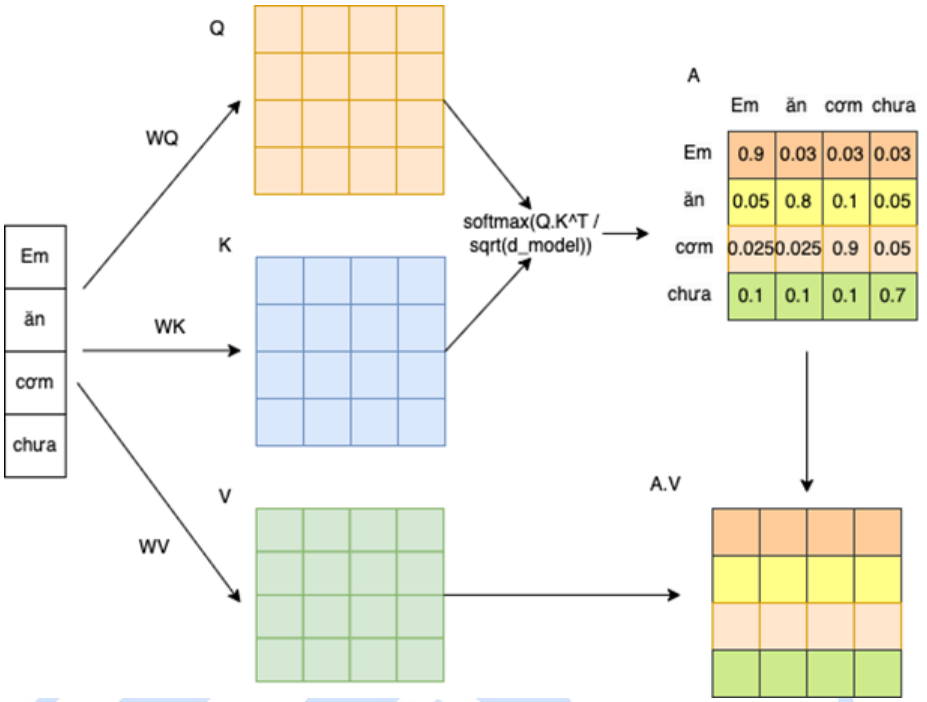

Figure 3.2: Cách hoạt động của self-attention. Mỗi từ sẽ tập trung vào những từ khác xung quanh nó (kể cả chính nó) với một mức độ nào đó.

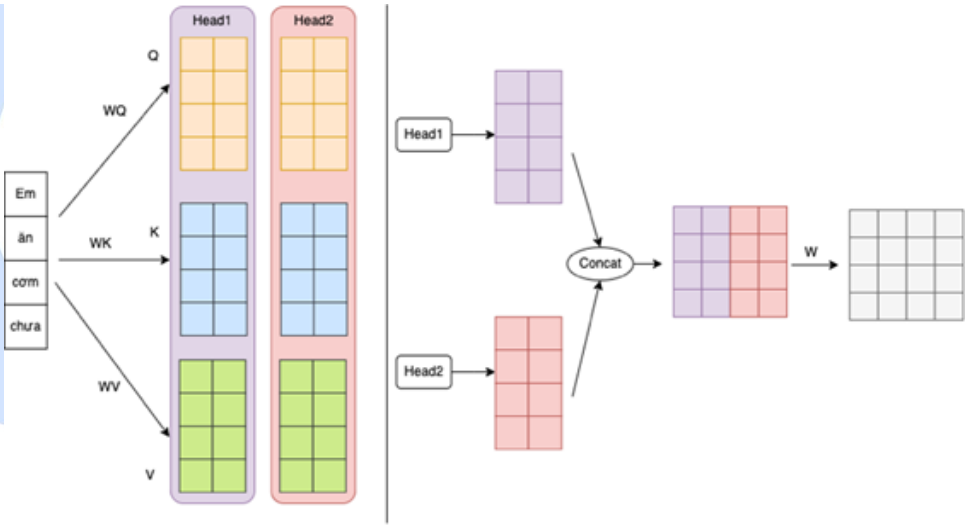

Figure 3.3: Cách hoạt động của Multi-head self attention. Tương tự như self-attention nhưng ma trận Q (K hoặc V) sẽ được tách ra theo chiều model dimension và đưa vào các head khác nhau.

`

3.2.3 Point-wise FFN

Point-wise FFN gồm các layer sau:

- **Up linear layer:** một linear layer có số node bằng 3072, số đầu vào là 768. Activation funcion GELU
- **Down linear layer:** một linear layer có số node là 3072, số đầu vào bằng 768
- **Dropout layer:** dropout rate 0.1
- **Layer normalization**

Đầu ra của MHSA đi qua up linear layer làm tăng số chiều của vector, sau đó qua down linear layer để giảm lại thành số chiều ban đầu. Tiếp đến qua dropout layer, cộng với đầu vào ban đầu của module (residual connection) rồi qua layer normalization.

3.2.4 Prediction layer:

Là một point-wise linear layer có số node = 21, số đầu vào = 768, activation function softmax. Layer này dự đoán xác suất thuộc 21 classses của token là bao nhiêu. Nếu đầu vào của mô hình là chuỗi có max$_seq_len$ tokens thì đầu ra sẽ là một ma trận $\in R^{\text{max\_seq\_len x 21}}$.

3.3 Tiền xử lý

Xét quá trình training, một mẫu dữ liệu đầu vào sẽ bao gồm văn bản đầu vào và chuỗi các slot labels. Kết thúc quá trình tiền xử lý, ta sẽ nhận được 4 trường thông tin sau:

- input_ids: danh sách các token ids của văn bản đầu vào
- attetion_mask:
- slot_label_ids: biểu diễn số slot labels
- val_pos_list: danh sách vị trí các subwords đầu tiên của mỗi từ.

`

3.3.1 Các bước thực hiện

B1: Khởi tạo tokens = []; slot_label_ids = []; val_pos_list = []

B2: Dùng rdrsegmenter để segment văn bản:

words = rdrsegmenter(words)

Do bộ dữ liệu đã được segment sẵn nên ta sẽ không cần dùng. Chỉ dùng cho inference.

B3: Chuyển slot_label trong slot_labels thành slot_label_id, tạo thành raw_slot_label_ids

B4: Lặp qua các cặp (word, label_slot_id) trong words. Với mỗi cặp ta thực hiện:

PhoBERT tokenizer tách word thành list các subwords

subwords = tokenizer.tokenizer(word)

tokens = tokens + subwords

Lưu lại vị trí của subword đầu tiên

val_pos_list = val_pos_list + [True] + [False] * (len(subwords) – 1))

Subword đầu tiên sẽ có nhãn là label_slot, các subwords còn lại sẽ có nhãn là pad_label_id (bị hàm loss bỏ qua)

slot_label_ids = slot_label_ids + [slot_label_id] + [pad_label_id] * (len(subwords) – 1)

B5: Thêm “[CLS]” và “[SEP]” tokens

Nếu len(tokens) > max_seq_len – 2 thì ta sẽ bỏ đi 2 phần tử cuối trong các list tokens, slot_label_ids, val_pos_list. Ngược lại thì đi tiếp.

Thêm “[CLS]” token

`

tokens = [“[CLS]”] + tokens

slot_label_ids = [pad_label_id] + slot_label_ids

val_pos_list = [False] + val_pos_list

Thêm “[SEP]” token

tokens = tokens + [“[SEP]”]

slot_label_ids = slot_label_ids + [pad_label_id]

val_pos_list = val_pos_list + [False]

B6: Tạo attention mask:

mask = [1] * len(tokens)

B7: Padding đến max_seq_len

pad_len = max_seq_len – len(tokens)

tokens = tokens + ([“[PAD]”] * pad_len)

slot_label_ids = slot_label_ids + (pad_label_id * pad_len)

val_pos_list = val_pos_list + (False * pad_len)

mask = mask + ([0] * pad_len)

B8: Chuyển tokens thành token ids với PhoBERT tokenizer

input_ids = tokenizer.convert_tokens_to_ids(tokens)

B9: Return (input_ids, slot_label_ids, val_pos_list_mask)

3.3.2 Ví dụ minh hoạ

Giả sử ta có mẫu dữ liệu sau: input – words: “Bệnh nhân N.V.A bị viêm khớp”, output - slot_labels: [‘O’, ‘B-NAME’, ‘O’, ‘B-SYMPTOM_AND_DISEASE’, ‘I-SYMPTOM_AND_DISEASE]. Max_seq_len = 10, pad_label_id = -100

`

B1: tokens = []; slot_label_ids = []; val_pos_list = []

B2: words = [‘Bệnh_nhân’, ‘N.V.A’, ‘bị’, ‘viêm’, ‘khớp’]

B3: Giả sử label id của O, B-NAME, B-SYMPTOM_AND_DISEASE, I_SYMPTOM_AND DISEASE lần lượt là 0, 4, 13, 14, ta có raw_label_slot_ids = [0, 4, 0, 13, 14]

B4:

Lần lặp 1: word = “Bệnh_nhân”, slot_label_id = 0

subwords = [‘Bệnh_nhân”]

➔ tokens = [‘Bệnh_nhân’]; slot_label_ids = [0]; val_pos_list = [True]

Lần lặp 2: word = “N.V.A”, slot_label_id = 4:

subwords = [‘N.V.@@’, ‘A’]

➔ tokens = [‘Bệnh_nhân’, ‘N.V.@@’, ‘A’]

slot_label_ids = [0, 4, -100]

val_pos_list = [True, True, False]

Lần lặp 3: word = ‘bị’; slot_label_id = 0:

subwords = [‘bị’]

➔ tokens = [‘Bệnh_nhân’, ‘N.V.@@’, ‘A’, ‘bị’]

slot_label_ids = [0, 4, -100, 0]

val_pos_list = [True, True, False, True]

Lần lặp 4: word = ‘viêm’; slot_label_id = 13:

subwords = [‘viêm’]

➔ tokens = [‘Bệnh_nhân’, ‘N.V.@@’, ‘A’, ‘bị’, ‘viêm’]

`

slot_label_ids = [0, 4, -100, 0, 13]

val_pos_list = [True, True, False, True, True]

Lần lặp 5: word = ‘khớp’; slot_label_id = 14:

subwords = [‘khớp’]

➔ tokens = [‘Bệnh_nhân’, ‘N.V.@@’, ‘A’, ‘bị’, ‘viêm’, ‘khớp’]

slot_label_ids = [0, 4, -100, 0, 13, 14]

val_pos_list = [True, True, False, True, True, True]

B5: Do len(tokens) = 6 <= 10 – 2 = 8 nên ta sẽ không cắt bỏ. Sau khi thêm 2 token đặc biệt vào ta có:

tokens = [‘[CLS]’, ‘Bệnh_nhân’, ‘N.V.@@’, ‘A’, ‘bị’, ‘viêm’, ‘khớp’, ‘[SEP]’]

slot_label_ids = [-100, 0, 4, -100, 0, 13, 14, -100]

val_pos_list = [False, True, True, False, True, True, True, False]

B6: mask = [1, 1, 1, 1, 1, 1, 1, 1]

B7: Padding đến độ dài 10

tokens = [‘[CLS]’, ‘Bệnh_nhân’, ‘N.V.@@’, ‘A’, ‘bị’, ‘viêm’, ‘khớp’, ‘[SEP]’, ‘[PAD]’, ‘[PAD]’]

slot_label_ids = [-100, 0, 4, -100, 0, 13, 14, -100, -100, -100]

val_pos_list = [False, True, True, False, True, True, True, False, False, False]

mask = [1, 1, 1, 1, 1, 1, 1, 1, 0, 0]

B8: input_ids = [0, 6207, 22290, 768, 45, 1743, 2819, 2, 1, 1]

B9: Vậy:

`

input_ids = [0, 6207, 22290, 768, 45, 1743, 2819, 2, 1, 1]

slot_label_ids = [-100, 0, 4, -100, 0, 13, 14, -100, -100, -100]

val_pos_list = [False, True, True, False, True, True, True, False, False, False]

mask = [1, 1, 1, 1, 1, 1, 1, 1, 0, 0]

Xét inference, các thao tác tương tự như trên nhưng không cần quan tâm đến slot_label_ids.

3.4 Hậu xử lý

Kết quả trả về của mô hình là 1 tensor all_val_pos_list (batch_size; max_seq_len; n_slots) và 1 ma trận all_preds (batch_size; max_seq_len). Xét riêng một mẫu dữ liệu, ta sẽ có val_pos_list (max_seq_len; n_slots) - đây là thứ mà chúng ta tiền xử lý, preds (max_seq_len).

3.4.1 Các bước thực hiện

B1: Chọn slot label id có xác suất lớn nhất

slot_preds = argmax(preds, axis = -1)

B2: Lọc bỏ slot_preds của các token đặc biệt và subwods tầm thường (không là subword đầu tiên của một từ):

res = []

For i = 0 → len(val_pos_list) – 1:

Nếu val_post_list[i] == True:

res = res + convert_id2label(slot_preds[i])

B3: Return res

`

3.4.2 Ví dụ minh hoạ

Giả sử ta đưa văn bản đầu vào là “Bệnh_nhân N.V.A”, mô hình trả về preds và val_pos_list = [False, True, True, False, False, False, False, False, False, False]. Sau khi tính argmax, ta có slot_preds = [int, 0, 4, int, int, int, int, int, int, int]. Nếu để ý ví dụ trước đó, thì “Bệnh nhân N.V.A” sẽ có tokens = [“[CLS]”, “Bệnh_nhân”, “N.V.@@”, “A”, “[SEP]”, “[PAD]”, “[PAD]”, “[PAD]”, “[PAD]”], “[PAD]”]]. Token “A” sẽ có slot_pred[3] là int nhưng giá val_pos_list[3] là False nên ta sẽ không quan tâm đến giá trị này. Như vậy, đầu ra sẽ là [0, 4] tương đương với [“O”, “B-NAME”].

`

CHƯƠNG 4: TRÌNH BÀY VÀ CÀI ĐẶT KIỂM THỬ

4.1 Môi trường

Colab thường, GPU Tesla T4, 16GB RAM

4.2 Hyperparameters

Độ dài chuỗi đầu vào tối đa (max_seq_len) là 256 (Do trong bộ dữ liệu độ dài chuỗi tối đa sau khi tách thành subwords là 182)

Epoch: 30

Early stopping với patience = 5

Train batch size: 32

Eval batch size: 128

Optimizer: AdamW, learning rate khởi tạo = 5e-5, epsilon: 1e-8,

Learning rate scheduler: Linear Scheduler - một scheduler đơn giản nhưng có thể giúp ta tăng hiệu quả của việc huấn luyện. Ban đầu, learning rate sẽ có giá trị lớn, đi tới điểm optimal nhanh hơn. Sau đó, leanring rate nhỏ dần giúp mô hình đạt tới.

Gradient clipping với max gradient norm = 1, giúp tránh bùng nổ gradient.

Dropout rate: 0.1

4.3 Sourcecode

Link: https://colab.research.google.com/drive/1XOMULfNn5eZOOiVa6ufPJKMLrI-blphV?authuser=2#scrollTo=IaVGjUvhxqvB

`

CHƯƠNG 5: KẾT QUẢ SƠ BỘ

5.1 Kết quả mô hình PhoBERT

Kết quả của mô hình PhoBERT khi được đánh giá trên bộ Test của tập dữ liệu PhoNER Covid-19 đưa ra với các chỉ số đánh giá là Precession, Recall và F1-score trọng số cho từng miền văn bản. Mô hình chúng tôi cài đặt đạt được macro-avg F1-score và weighted-avg F1 score lần lượt là **95.10%, 93.75%.** Dưới đây là bảng mô tả chi tiết kết quả đạt được của nhóm:

	precision	recall	f1-score	support
AGE	0.9911	0.9605	0.9756	582
DATE	0.9826	0.9909	0.9868	1654
GENDER	0.9846	0.9697	0.9771	462
JOB	0.8221	0.7746	0.7976	173
LOCATION	0.9451	0.9491	0.9471	4441
NAME	0.9401	0.9371	0.9386	318
ORGANIZATION	0.8896	0.9092	0.8993	771
PATIENT_ID	0.9826	0.9855	0.9841	2005
SYMPTOM_AND_DISEASE	0.8920	0.8873	0.8897	1136
TRANSPORTATION	0.9744	0.9845	0.9794	193
micro avg	0.9504	0.9517	0.9510	11735
macro avg	0.9404	0.9348	0.9375	11735
weighted avg	0.9504	0.9517	0.9510	11735

Figure 5.1: Kết quả đánh giá trên tập test

Theo đó, kết quả đạt được trên miền thực thể DATE đạt được kết quả cao nhất với **98.68%**, còn miền thực thể JOB đạt kết quả thấp nhất với **79.71%.** Các miền còn lại cũng đạt được kết quả rất tốt, đều trên **88%**.

`

5.2 Phân tích kết quả

Thường thì, một token (tức là một thực thể định danh có thể chứa nhiều hơn một từ) sẽ được trích xuất như là một thực thể đúng nếu xảy ra hai điều kiện đúng và đồng thời sau đây:

- Độ dài của từ (range) là đúng: Từ bắt đầu (B) và từ kết thúc (I) giống như True Label.
- Nhãn (tag) đúng: Nhãn giống như True Label.

Nếu như không thể đúng một trong hai trường hợp, nó sẽ là một thực thể sai. Thường thì một mô hình dự đoán ra các thực thể sẽ gặp 5 lỗi sai như sau:

- **No extraction**: Lỗi trong đó mô hình không trích xuất mã thông báo dưới dạng thực thể tên (Name Entity) (NE) mặc dù mã thông báo được chú thích là NE.

  Ví dụ: **Pred label**: Việt_Nam

  True Label: LOC Việt_Nam LOC

- **No annotation**: Lỗi mô hình trích xuất mã thông báo dưới dạng NE mặc dù các mã thông báo không được chú thích là NE.

  Ví dụ: **True label**: LOC Việt_Nam LOC

  Pred Label: Việt_Nam

- **Wrong range**: Lỗi trong đó mô hình trích xuất mã thông báo dưới dạng NE và chỉ sai phạm vi.

  Ví dụ: **Pred label**: JOB Bác_sĩ Trương_Văn_Khải JOB

  True label: Bác_sĩ JOB Trương_Văn_Khải JOB

- **Wrong tag**: Lỗi mô hình trích xuất token là NE và chỉ sai loại thẻ.

  Ví dụ: **Pred label**: LOC Bệnh_viện Quận_hai LOC

`

True label: ORG Bệnh_viện Quận_hai ORG

- **Wrong range and tag**: Lỗi trong đó mô hình trích xuất mã thông báo dưới dạng NE nhưng cả phạm vi và loại thẻ đều sai.

  Ví dụ: **Pred label**: LOC Cửa_hàng KFC LOC

  True label: Cửa_hàng ORG KFC ORG

5.2.1 Nhận xét một số TH đúng

Trong tổng số 11,888 thực thể trong tập Test (11,735 thực thể có định danh, 153 thực thể "O") thì mô hình của chúng tôi dự đoán dự đoán đúng 11,206 thực thể (chiếm tỷ lệ 94,26%). Bảng thống kê tỷ lệ dự đoán của chúng tôi như sau:

	Tag	Total	Errors	No Extraction	No Annotation	Wrong Range	Wrong Tag	Wrong Range and tag
0	PATIENT_ID	2005	29	5	0	19	5	0
1	NAME	318	20	17	0	2	1	0
2	AGE	582	22	13	0	1	8	0
3	GENDER	462	14	13	0	0	1	0
4	JOB	173	38	29	0	7	2	0
5	LOCATION	4441	205	41	0	112	44	8
6	ORGANIZATION	771	66	12	0	12	37	5
7	SYMPTOM_AND_DISEASE	1136	118	56	0	61	1	0
8	TRANSPORTATION	193	3	0	0	2	1	0
9	DATE	1654	14	2	0	8	4	0
10	O	153	153	0	153	0	0	0
11	Total	11888	682	188	153	224	104	13

Figure 5.2: Bảng thống kê những trường hợp sai

Mô hình của chúng tôi dự đoán đúng với tỷ lệ khá cao, khoảng 94%.

Phân tích một số trường hợp dự đoán đúng như sau:

STT	Sentence	True Label - Dự đoán	Lý do
1	['Theo', 'đó', ',', 'Sở', 'Y_tế', 'Bình_Thuận', 'cho', 'biết', 'sau', 'khi', 'xác_định',	[(['B-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION'], ['B-	Vì các thực thể xuất hiện trong câu trên đều là những thực thể mà

`

	'bệnh_nhân', 'số', '34', '(', 'nữ_giới', '51', 'tuổi', ',', 'từ', 'Mỹ', 'về', 'Việt_Nam', 'ngày', '29', '-', '2', 'có', 'quá_cảnh', 'Qatar', ')', ',', 'Trung_tâm', 'Kiểm_soát', 'bệnh_tật', 'Bình_Thuận', 'đã', 'điều_tra', 'dịch_tễ', ',', 'khoanh', 'vùng', ',', 'khử', 'khuẩn', ',', 'tiến_hành', 'cách_ly', 'người', 'liên_quan', 'đến', 'ca', 'bệnh', 'số', '34', '.']	ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION'], ['Sở', 'Y_tế', 'Bình_Thuận']), (['B-PATIENT_ID'], ['B-PATIENT_ID'], ['34']), (['B-GENDER'], ['B-GENDER'], ['nữ_giới']), (['B-AGE'], ['B-AGE'], ['51']), (['B-LOCATION'], ['B-LOCATION'], ['Mỹ']), (['B-LOCATION'], ['B-LOCATION'], ['Việt_Nam']), (['B-DATE', 'I-DATE', 'I-DATE'], ['B-DATE', 'I-DATE', 'I-DATE'], ['29', '-', '2']), (['B-LOCATION'], ['B-LOCATION'], ['Qatar']), (['B-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION'], ['B-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION'], ['Trung_tâm', 'Kiểm_soát', 'bệnh_tật', 'Bình_Thuận']), (['B-PATIENT_ID'], ['B-PATIENT_ID'], ['34'])]	model được học đi học lại rất nhiều lần trong tập train nên model đã dự đoán đúng hoàn toàn.
2	['Trong', 'thời_gian', 'ở', 'đây', ',', 'em', 'đi', 'chơi', 'công_viên', 'SunWorld', 'Đà_Nẵng', ',', 'siêu_thị', 'Lotte_Mart', 'và', 'ăn_ở', 'một_số', 'quán', '.']	tức là train trên số lượng dữ liệu lớn nhất bộ dữ liệu ()	Vì trong câu trên là một câu ngắn, đồng thời chỉ xuất hiện các thực thể LOCATION (mà thực thể LOCATION được model làm rất tốt (tổng 4441 thực thể) nhưng lại đạt được 94.71% điểm F1-score).
3	["", 'Bệnh_nhân', 'này', 'là', 'công_nhân', 'làm_việc', 'ở', 'Công_ty', 'Samsung', 'tại', 'huyện', 'Yên_Phong', ',', 'tỉnh', 'Bắc_Ninh', '.']	[(['B-ORGANIZATION', 'I-ORGANIZATION'], ['B-ORGANIZATION', 'I-ORGANIZATION'], ['Công_ty', 'Samsung']), (['B-LOCATION', 'I-LOCATION'], ['B-LOCATION',	Tương tự như trường hợp hai, trì trong câu chỉ xuất hiện thực thể LOCATION

`

		'I-LOCATION'], ['huyện', 'Yên_Phong']), (['B-LOCATION', 'I-LOCATION'], ['B-LOCATION', 'I-LOCATION'], ['tỉnh', 'Bắc_Ninh'])]	
4	["", 'Bệnh_nhân', '1013', "", ',', '51', 'tuổi', ',', 'tiểu_thương', 'bán', 'hải_sản', 'ở', 'chợ', 'Siêu_Thị', ',', 'từng', 'được', 'lấy', 'mẫu', 'xét_nghiệm', 'nCoV', 'lần', 'một', 'ngày', '16/8', 'và', 'có', 'kết_quả', 'âm_tính', '.']	[(['B-PATIENT_ID'], ['B-PATIENT_ID'], ['1013']), (['B-AGE'], ['B-AGE'], ['51']), (['B-JOB', 'I-JOB', 'I-JOB'], ['B-JOB', 'I-JOB', 'I-JOB'], ['tiểu_thương', 'bán', 'hải_sản']), (['B-LOCATION', 'I-LOCATION'], ['B-LOCATION', 'I-LOCATION'], ['chợ', 'Siêu_Thị']), (['B-DATE'], ['B-DATE'], ['16/8'])]	Theo kết quả dựa trên tập test, model luôn đạt kết quả rất cao ở 3 loại thực thể LOCATION; AGE; DATE (đều trên 97%); PATIENT_ID là 94.71%. Thực thể JOB thì không đạt kết quả cao như các thực thể còn lại nhưng vẫn dữ đoán đúng trong truòng hợp này.
5	['Bệnh_nhân', 'tử_vong', 'bên', 'đường', 'Đa_Phú', ',', '2', 'bệnh_nhân', 'còn', 'lại', 'ở', 'chợ', 'Đà_Lạt', '.']	[(['B-LOCATION', 'I-LOCATION'], ['B-LOCATION', 'I-LOCATION'], ['đường', 'Đa_Phú']), (['B-LOCATION', 'I-LOCATION'], ['B-LOCATION', 'I-LOCATION'], ['chợ', 'Đà_Lạt'])]	Có thể thấy, model luôn dự đoán đúng hoàn toàn ở những câu ngắn và các câu có xuất hiện nhiều thực thể LOCATION
6	['Kết_quả', 'xét_nghiệm', 'ngày', '17', '-', '9', 'cả', '2', 'dương_tính', 'với', 'virus', 'SARS', '-', 'CoV', '-', '2', '.']	[(['B-DATE', 'I-DATE', 'I-DATE'], ['B-DATE', 'I-DATE', 'I-DATE'], ['17', '-', '9'])]	Ở câu này, model dự đoán hoàn toàn vì model dự đoán rất tốt trên thực thể DATE (khoảng 98.68%).

Table 5.1: Bảng nhận xét một số trường hợp mô hình dự đoán đúng.

5.2.2 Nhận xét một số TH sai

Từ bảng tổng kết kết quả dự đoán của mô hình trên, chúng tôi quyết định vẽ biểu đồ tròn để kiểm chứng tỷ lệ lỗi của từng trường hợp sai trên tổng số lỗi sai, kết quả thu được như sau:

`

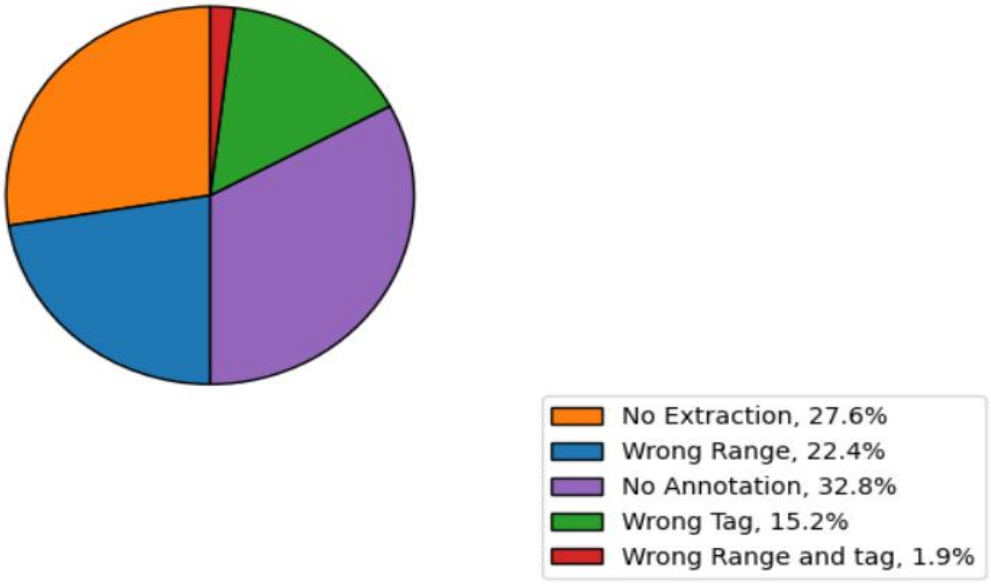

Figure 5.3: Biểu đồ cơ cấu tỷ lệ những trường hợp sai

Tỷ lệ các loại lỗi như No Extraction 27.6%, Wrong Range 22.4%, No Annotation 32.8%, Wrong Tag 15.2%, Wrong Range and tag 1.9%. Trong đó No Annotation là loại lỗi thường gặp phải nhất, chiếm 32.8%. Một phần là do bộ dữ liệu PhoNER Covid-19 có tới 10 loại thực thể, lại còn là bộ dữ liệu có số thực thể lớn nhất từ trước đến giờ nên đây cũng chính là thách thức ban đầu đến từ bộ dữ liệu. Ngoài ra, số lượng thực thể không cân bằng cũng là một phần nguyên nhân dẫn đến cho mô hình của chúng tôi không thể dự đoán chính xác hoàn toàn được.

Sau đây là một phần liệt kê các trường hợp sai của chúng tôi:

STT	Sentence	Dự đoán sai	True Label - Dự đoán đúng	Nhận xét sai
1	['Hai', 'người', 'có', 'tiếp_xúc', 'gần', 'với', 'nữ',	No Extracton: [(['B-GENDER'], ['O'], ['nữ'])]	[(['B-PATIENT_ID'], ['B-PATIENT_ID'], ['17']), (['B-LOCATION'], ['B-LOCATION'], ['Việt_Nam']), (['B-LOCATION', 'I-LOCATION', 'I-LOCATION'],	Trong trường hợp này, model dự đoán từ "nữ" là nhãn "O", trong

`

	'bệnh_nhân', 'nhiễm', 'COVID', 'thứ', '17', 'tại', 'Việt_Nam', 'đang', 'theo_dõi', 'tại', 'Bệnh_viện', 'Hữu_nghị', 'Việt_Tiệp', ',', 'bước_đầu', 'có', 'kết_quả', 'âm_tính', 'với', 'virus', 'corona', '.']		['B-LOCATION', 'I-LOCATION', 'I-LOCATION'], ['Bệnh_viện', 'Hữu_nghị', 'Việt_Tiệp'])]	khi True Label là "B-GENDER".
2	['Đặc_biệt', 'chống', 'chỉ_định', 'với', 'người', 'có', 'bệnh_lý', 'tim_mạch', ',', 'trào', 'ngược', 'dạ_dày', '-', 'tá_tràng', ',', 'nhiễm_khuẩn', '...']	No annotation: [(['O'], ['B-SYMPTOM_AND_DISEASE'], ['tim_mạch'])]	[(['B-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE'], ['B-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE'], ['trào', 'ngược', 'dạ_dày', '-', 'tá_tràng']), (['B-SYMPTOM_AND_DISEASE'], ['B-SYMPTOM_AND_DISEASE'], ['nhiễm_khuẩn'])]	Trong trường hợp này, model dự đoán từ "tim mạch" là 'B-SYMPTOM_AND_DISEASE' trong khi nhãn thực tế là "O. Ở trường hợp này, cá nhân tụi em nhận xét là TH này nhóm tác giả đã đánh nhãn sai, vì trong Annotation Guideline, 1 thực thể được đánh nhãn là SYMPTOM_AND_DISEASE khi nó có liên quan đến những bệnh lý mà bệnh nhân Covid-19 gặp phải.

`

3	['Ngày', '24/7', ',', 'bệnh_nhân', 'chăm_sóc', 'bố', 'là', "", 'bệnh_nhân', '428', "", 'tại', 'khoa', 'Nội', '-', 'Tiết_niệu', ',', 'Bệnh_viện', 'Đà_Nẵng', '.']	Wrong range: [(['B-LOCATION', 'I-LOCATION'], ['I-LOCATION', 'I-LOCATION'], ['Bệnh_viện', 'Đà_Nẵng'])]	[(['B-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE'], ['B-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE'], ['viêm', 'phổi'])]	Trong trường hợp này, model đã đánh nhãn sai vị trí bắt đầu của từ "Bệnh_viện" là "B-" thay vì "I-".
4	['Bác_sĩ', 'Trần_Thanh_Linh', ',', 'từ', 'Bệnh_viện', 'Chợ_Rẫy', 'chi_viện', 'phụ_trách', 'đơn_nguyên', 'hồi_sức', 'tích_cực', ',', 'cho', 'biết', "", 'bệnh_nhân', '416', "", 'vẫn', 'đang', 'duy_trì', 'ECMO', ',', 'thở', 'máy', ',', 'hiện', 'xơ', 'phổi', 'rất', 'nhiều', '.']	Wrong tag: [(['B-ORGANIZATION', 'I-ORGANIZATION'], ['B-LOCATION', 'I-LOCATION'], ['Bệnh_viện', 'Chợ_Rẫy'])]	[(['B-PATIENT_ID'], ['B-PATIENT_ID'], ['416']), (['B-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE'], ['B-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE', 'I-SYMPTOM_AND_DISEASE'], ['xơ', 'phổi', 'rất', 'nhiều'])]	Trong trường hợp này, model đã dự đoán sai nhãn, true label là ORGANIZATION trong khi đó nhãn mà model dự đoán là "LOCATION". Vì hay thực thể này gần giống nhau và chỉ có sự khác nhau về ngữ nghĩa nên rất nhọc nhằn cho model trong khâu dự đoán nhãn.
5	['Ca', 'bệnh', '157', '(', 'bệnh_nhân', '157', ')', ':', 'Bệnh_nhân', 'nữ', ',', 'quốc_tịch', 'Anh', ',', '31', 'tuổi', ',', 'giáo_viên',	Wrong Range and Tag: [(['B-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION'], ['I-JOB', 'O', 'O'], ['Eschool', '-', 'Eclass'])]	[(['B-PATIENT_ID'], ['B-PATIENT_ID'], ['157']), (['B-PATIENT_ID'], ['B-PATIENT_ID'], ['157']), (['B-GENDER'], ['B-GENDER'], ['nữ']), (['B-AGE'], ['B-AGE'], ['31']), (['B-JOB'], ['B-JOB'], ['giáo_viên']), (['B-LOCATION', 'I-LOCATION'], ['B-LOCATION', 'I-LOCATION'],	Trong trường hợp này, model đã dự đoán sai cả nhãn và vị trí bắt đầu. Ở cụm từ 'Eschool', '-', 'Eclass' có nhãn thực là 'B-ORGANIZATION', 'I-

`

	'Eschool', '-', 'Eclass', ',', 'hiện', 'ngụ', 'tại', 'đường', 'Tôn_Đản', ',', 'phường', '13', ',', 'quận', '4', ',', 'TP.HCM.']		['đường', 'Tôn_Đản']), (['B-LOCATION', 'I-LOCATION'], ['B-LOCATION', 'I-LOCATION'], ['phường', '13']), (['B-LOCATION', 'I-LOCATION'], ['B-LOCATION', 'I-LOCATION'], ['quận', '4']), (['B-LOCATION'], ['B-LOCATION'], ['TP.HCM.'])]	ORGANIZATION', 'I-ORGANIZATION' trong khi model dự đoán là 'I-JOB', 'O', 'O'.
6	['Ngày', '28', '-', '8', ',', 'Trung_tâm', 'CDC', 'xét_nghiệm', '(', 'lần', '3', ')', 'dương_tính', 'với', 'SARS', '-', 'CoV', '-', '2', ',', 'bệnh_nhân', 'được', 'chuyển', 'tới', 'Bệnh_viện', 'Nhiệt_đới', 'trung_ương', '2', '.']	No Annotation: [(['O', 'O'], ['B-ORGANIZATION', 'I-ORGANIZATION'], ['Trung_tâm', 'CDC'])] Wrong tag: [(['B-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION'], ['B-LOCATION', 'I-LOCATION', 'I-LOCATION', 'I-LOCATION'], ['Bệnh_viện', 'Nhiệt_đới', 'trung_ương', '2'])]	[(['B-DATE', 'I-DATE', 'I-DATE'], ['B-DATE', 'I-DATE', 'I-DATE'], ['28', '-', '8'])]	Trong trường hợp này, model đã dự đoán sai cả trường hợp dự đoán nhãn không được đánh nhãn và dự đoán sai nhãn.

Table 5.2: Bảng nhận xét một số trường hợp mô hình dự đoán sai.

Ngoài ra, chúng tôi có liệt kê đầy đủ các Trường Hợp dự đoán đúng, dự đoán sai của model trên toàn bộ 3000 câu trong tập Test.

`

KẾT LUẬN VÀ HƯỚNG CẢI TIẾN

Trong phần này, chúng tôi xin tổng kết lại những điều mà chúng tôi đã đạt được trong phạm vi đồ án cuối kỳ này. Ngoài ra, chúng tôi cũng sẽ trình bày những hạn chế và đưa ra hướng phát triển tiếp theo của chúng tôi trong tương lai.

Kết Luận

Chúng tôi đã tiến hành tìm hiểu và nghiên cứu về một quy trình cho bài toán "Nhận diện thực thể Covid-19 cho Tiếng Việt", cụ thể là chúng tôi sẽ giải quyết vấn đề tìm ra các thực thể đã được định danh ở trong câu. Ngoài ra, chúng tôi xây dựng đã xây dựng và tinh chỉnh mô hình PhoBERT để đánh giá xem liệu PhoBERT có thực hiện tác vụ nhận diện thực thể Tiếng Việt tốt hay không.

Mô hình PhoBERT của chúng tôi xây dựng đạt được kết quả macro-avg F1 score, weighted-avg F1 score lần lượt là 95.10%, 93.75% cho bộ dataset với 10 loại thực thể khác nhau.

Ngoài ra, chúng tôi cũng đồng thời thực hiện xây dựng một Web App Demo cho mô hình PhoBERT và các mô hình khác trên bài toán của chúng tôi.

`

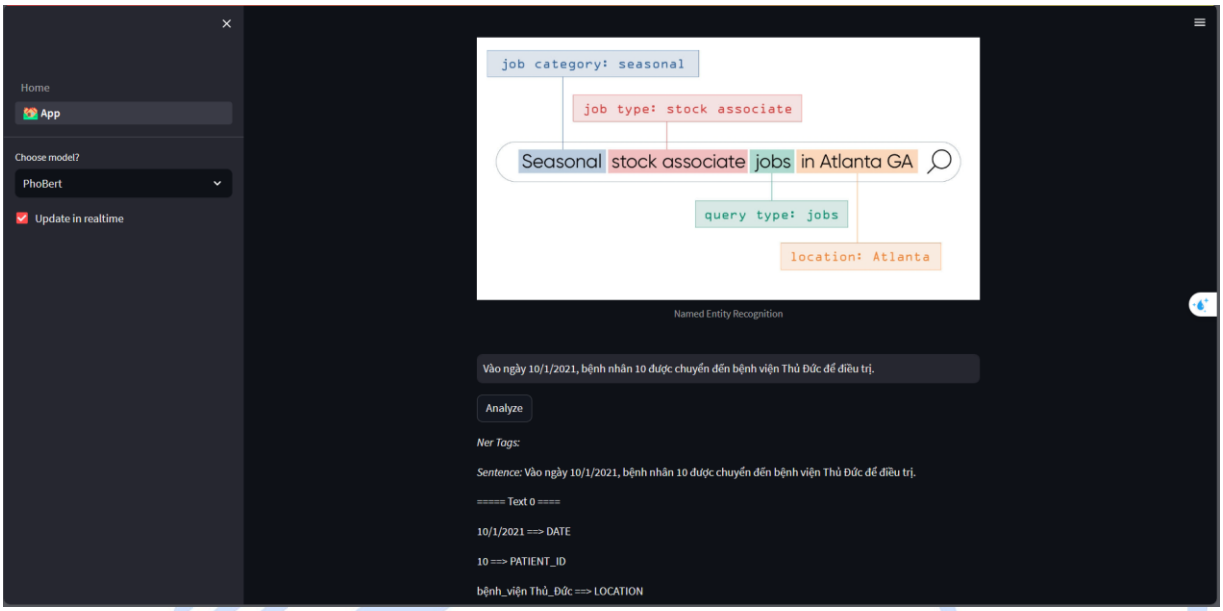

Figure 5.4:Hình ảnh mô tả Web App cho mô hình PhoBERT để dự đoán.

Hạn chế

Những thực thể có số lượng thực thể rất ít như "JOB", "NAME" vẫn chưa thể đạt kết quả cao như các thể thực thể chiếm số lượng rất nhiều "LOCATION", "PATIENT_ID".

Cuối cùng, việc xây dựng các bước tiền xử lý và hậu xử lý vẫn chưa tối ưu cho dữ liệu đầu vào nên dẫn đến hiệu suất của mô hình và kết quả dự đoán vẫn chưa là tốt nhất.

Hướng phát triển trong tương lai

Sử dụng các mô hình có hiệu quả tốt hơn và đồng thời tối ưu hóa giai đoạn tiền xử lý dữ liệu đầu vào và hậu xử lý.

Thực nghiệm mô hình trên các bộ dữ liệu về Nhận Dạng Thực Thể Tiếng Việt khác.

Mở rộng bộ dữ liệu với nhiều thực thể hơn.

`

Mở rộng sang ứng dụng mobile tiện lợi với người dùng.

`

TÀI LIỆU THAM KHẢO

[1] Thinh Hung Truong, Mai Hoang Dao, Dat Quoc Nguyen, "COVID-19 Named Entity Recognition for Vietnamese," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico, 2021.

[2] Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, Mark Johnson, "A Fast and Accurate Vietnamese Word Segmenter," in *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018.

[3] Dat Quoc Nguyen, Anh Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics.*, Minneapolis Institute of Art, 2019.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man- dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," in *arXiv preprint, arXiv:1907.11692*, 2019.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is All you Need," in *In*

`

Advances in Neural Information Pro- cessing Systems 30, Long Beach, CA, USA, 2017.