Mô tả đồ án môn học Xử lý ngôn ngữ tự nhiên – CS221.O11

Thành viên nhóm:

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

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 thực hiện một mô hình NER trên bộ dữ liệu mới nhất của VINAI là PHONER COVID19 được công bố tại hội nghị NAACL 2021, bộ dữ liệu được coi là một trong những bộ dữ liệu NER có số lượng thực thể lớn.

MỤC LỤC

1 Ngữ liệu

1.1 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, 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.

1.2 Cấu trúc bộ dữ liệu

1.2.1 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. 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.
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_AND_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.

Bảng 1.1 Bảng định nghĩa các loại thực thể

1.2.2 Quá trình gán nhãn dữ liệu

Quy trình gán nhãn bộ dữ liệu trên 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 từ VnCoreNLP. Các câu liên quan đến bệnh nhân COVID-19 được chọn bằng BM25Plus. Sau đó, lọc thủ công những câu 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 và 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 người 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 người chú thích được yêu cầu sửa đổi chú thích của họ cho đến

khi đạt được F1 ít nhất là 0,92, sau đó nhóm tác giả xem xét lại từng câu và sửa thêm nếu cần.

- Bộ dữ liệu kết quả bao gồm 35K thực thể trên 10027 câu.

- Quy trình đánh nhãn dữ liệu được thực hiện trên văn bản ở mức độ tiếng. Để tạo phiên bản mức độ từ, nhóm tác giả sử dụng RDRSegmenter để phân đoạn từ tự động.

1.2.2. 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

Hình 1.1: Bảng thống kê số lượng thực thể

1.3. 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. 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 B-ORG I-ORG 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 liên 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-ORG", "I-ORG". 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 được chuyển tới các phòng_khám khẩn_cấp , khiến những bệnh_nhân mắc bệnh hiểm_nghèo khác không có cơ_hội được điều_trị .	O O O O O O O B-SYMP_DIS I-SYMP_DIS O B-SYMP_DIS I-SYMP_DIS 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 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 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	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".

		y_tế tại nơi ở .	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-LOC O B-LOC O O O O B-LOC O B-DATE O O 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 O B-PATIENT_ID O O B-AGE O O O B-ORG I-ORG O O O O B-DATE 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. 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-ORG", "I-ORG". 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 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	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)**, vì vậy "L.T.H" được đánh nhãn là "B-NAME". 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 -	O O B-PATIENT_ID O B-JOB O O O O B-LOC I-LOC O B-LOC I-LOC O O O O O B-LOC I-LOC I-LOC O O O O B-DATE I-DATE	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).**

		4 cho biết diễn_biến bệnh của bệnh_nhân không xấu hơn nhưng cũng chưa có dấu_hiệu hồi_phục .	I-DATE O O O O O O O O O O O O O O	**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 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 là: "B-DATE". 3. Từ "21" chỉ **Ngày** nên được đánh nhãn 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 B-LOC 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 "Việt_Nam" chỉ **Tên quốc gia: Việt Nam** nên được đánh nhãn là B-LOCATION.

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 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 B-SYMP_DIS I-SYMP_DIS I-SYMP_DIS O	1. Các từ trước từ "suy" 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 O B-LOC I-LOC I-LOC I-LOC I-LOC O O B-LOC I-LOC I-LOC I-LOC I-LOC I-LOC I-LOC I-LOC O B-LOC I-LOC O O	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** 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". 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.**

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 O O B-DATE O O O O B-DATE O O O B-LOC I-LOC I-LOC I-LOC I-LOC I-LOC O	1. 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 2. 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. 3. 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** 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".
14	Train set - Dòng 14	Tính 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-LOC I-LOC O O O O O O O O O O O O O O O O O O B-TRANS 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** và **liên quan đến lịch trình di chuyển của bệnh nhân.** 3. Các từ trước cụm các từ "VN0054" 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 "17".
15	Train set - Dòng 15	Sở GD - ĐT đã quán_triệt tất_cả thí_sinh và những người làm công_tác thi sau khi	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.

		xét_nghiệm thực_hiện nghiêm giãn cách xã_hội theo Chỉ_thị 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 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	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. 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 "24" 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 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 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 ,	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.

		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	
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 B-ORG I-ORG I-ORG I-ORG O O	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". Cụm từ "Trung_tâm Kiểm_soát bệnh_tật tỉnh" do không có địa chỉ cụ thể nên không được đánh nhãn xác định thực thể ORG 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 được gán nhãn lần lượt là: "B-ORG", "I-ORG", "I-ORG", "I-ORG".
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-ORG 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". Từ "phi_công" không được đánh nhãn là B-JOB do nó không gắn liền với bệnh nhân xác định. 2. Từ "Vietnam_Airlines" chỉ Tên các công ty, tổ chức nơi bệnh nhân làm việc nên được đánh nhãn là "B-ORG". 3. Từ "18/3" chỉ ngày trong tiếng Việt nên được đánh nhãn là B-DATE
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-ORG I-ORG I-ORG I-ORG I-ORG I-ORG O O O O O O O O O B-DATE O O B-SYMP_DIS I-SYMP_DIS O O O O	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-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG". 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 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-LOC I-LOC I-LOC I-LOC O B-LOC O O O B-LOC I-LOC O O O O O B-PATIENT_ID O O O O B-JOB I-JOB I-JOB O O O B-LOC 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 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 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 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-ORG 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-ORG".
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 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 . 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-ORG I-ORG I-ORG I-ORG O O O O O O O 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 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 O B-PATIENT_ID O B-AGE O O O	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-ORG", "I-ORG", "I-ORG", "I-ORG". Không gán nhãn "LOCATION" cho cụm các từ trên vì: Thực thể kiểu ORG phải là tổ chức bao gồm một hay nhiều cá nhân và phải có chức năng, công việc nhất định. Thực thể kiểu ORG 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 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". 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-DATE O O O O O O O O O	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 là ngày, Y là tháng)** nên được đánh nhãn lần lượt là: "B-DATE", "I-DATE", "I-DATE". 3. Các từ còn lại 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-LOC I-LOC O B-LOC O B-DATE O B-LOC I-LOC O B-DATE O B-ORG I-ORG O B-LOC I-LOC O O B-DATE 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 "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-ORG", "I-ORG". 8. Cụm các từ "quận", "Gò_Vấp" là Địa chỉ, đặc biệt hơn nó chỉ cấp bậc đơn vị hành chính. Nên cụm "quận", "Gò_Vấp" lần lượt được 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-LOC I-LOC O B-LOC I-LOC 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-LOC I-LOC I-LOC O B-LOC I-LOC O B-LOC I-LOC O O B-ORG I-ORG O O B-DATE O	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. 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 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-ORG", "I-ORG". 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.	B-ORG I-ORG O O O O O O O O O B-LOC	1. Cụm các từ "Công_ty", "Huệ_Phong" chỉ Tên các công ty, tổ chức nơi bệnh nhân làm việc nên được gán nhãn lần lượt là "B-ORG", "I-ORG". 2. 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, ….). 3. Các từ còn lại không liên quan đến thực thể định danh nên được đánh nhãn là O
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-SYMP_DIS I-SYMP_DIS O O O O O O O O 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".
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-SYMP_DIS I-SYMP_DIS I-SYMP_DIS I-SYMP_DIS I-SYMP_DIS O O B-SYMP_DIS I-SYMP_DIS I-SYMP_DIS I-SYMP_DIS O B-LOC I-LOC O	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".
32	Val Set - Dòng 12	Bệnh_nhân từng mua thịt và cá	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**

		tại chợ đầu_mối Tân_Phát_Địa 8 ngày trước khi có triệu_chứng .	O B-LOC I-LOC I-LOC O O O O O O O	**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-LOC I-LOC I-LOC I-LOC O O O O B-PATIENT_ID O O O O O O O O O O O O O O O O O O	1. Cụm các từ "Bệnh_viện", "Đa_khoa", "huyện", "Đồng_Văn" 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 PATIENT_ID
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 O B-LOC I-LOC I-LOC I-LOC I-LOC I-LOC I-LOC I-LOC 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)**, nên được đánh nhãn là “B-NAME”. 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". Tuy nhiên, trong câu này đã đánh nhãn sai từ tại thành B-LOCATION.
35	Val Set - Dòng 15	Các trường_hợp tử_vong đều có bệnh_lý nền nặng với 82,4%	O O O O O O O O O O	1. Các từ trước từ "suy" 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**

		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 B-SYMP_DIS I-SYMP_DIS I-SYMP_DIS O O O O B-SYMP_DIS I-SYMP_DIS O O O O B-SYMP_DIS O O O O B-SYMP_DIS O O O B-SYMP_DIS I-SYMP_DIS O O O O O O O O O O O O O O O O O O O O O O O O	**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 "tim_mạch" tương tự trường hợp 2 nên được đánh nhãn là "B SYMPTOM_AND_DISEASE". 6. Từ "ung_thư" tương tự trương hợp 2 nên được gán nhãn là: "B SYMPTOM_AND_DISEASE". Tuy nhiên, từ này đã được đánh nhãn sai thành "I-SYMPTOM_AND_DISEASE". Từ "và" đứng trước cũng bị đánh nhãn sai trong khi nó phải là nhãn 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-LOC I-LOC O O	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. 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 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-ORG I-ORG I-ORG O 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-ORG", "I-ORG", "I-ORG", "I-ORG".
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-SYMP_DIS O B-SYMP_DIS I-SYMP_DIS O O O O B-LOC I-LOC I-LOC I-LOC I-LOC I-LOC O	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 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".
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-LOC I-LOC O B-LOC I-LOC O	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. 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 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". 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-LOC I-LOC I-LOC I-LOC I-LOC O O O O B-PATIENT_ID O B-PATIENT_ID O B-PATIENT_ID O O O O O O	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", "341", "350" 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.
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 (	O B-DATE I-DATE I-DATE O B-DATE I-DATE I-DATE O O O O O O B-NAME O B-AGE O O O O O O B-LOC I-LOC O	1. 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". 2. Từ “đến” không liên quan đến các khía cạnh cụ thể nên được đánh nhãn “O”. 3. Cụm từ "31", "-", "7" tương tự 2. 4. **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

		bằng xe_máy) , không đi đâu chỉ ra Tạp_hoá Phượng , chợ Vườn_Lài , phường An_Sơn cùng mẹ bán tạp_hoá ở đây .	O O O O O O O O O B-LOC I-LOC O B-LOC I-LOC O B-LOC I-LOC O O B-JOB I-JOB O O O	được đánh nhãn bởi vì: Các danh xưng "ông", "bà", "anh", "chị", "giám đốc", "chủ tịch", … KHÔNG nằm trong tên riêng. 5. 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 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. 6. "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**. 7. "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.** 8. "chợ", "Vườn_Lài" tương tự 7. 9. "phường", "An_Sơn" tương tự 7. 10. 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 rất nhiều .	O O O O B-ORG I-ORG O O O O O O O O O O B-PATIENT_ID O O O O O O O O O O B-SYMP_DIS I-SYMP_DIS I-SYMP_DIS I-SYMP_DIS O	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 ORG). Vì "Bác_sĩ", "Trần_Thanh_Linh", "từ" là chỉ nơi công tác của bác sĩ nên "Bệnh_viện", "Chợ_Rẫy" trong ngữ cảnh này được xem như một tổ chức => gán nhãn "B-ORG", "I-ORG". 2. 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. 3. 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 người liên_quan đến ca bệnh số 34 .	O O O B-ORG I-ORG I-ORG O O O O O O O B-PATIENT_ID O B-GENDER B-AGE O O O B-LOC O B-LOC O B-DATE I-DATE I-DATE O O B-LOC O O B-ORG I-ORG I-ORG I-ORG O O O O O O O O O O O O O O O O O O B-PATIENT_ID 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-ORG", "I-ORG", "I-ORG". 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. 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 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-ORG", "I-

				ORG", "I-ORG", "I-ORG". 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 O B-LOC O B-LOC O O O O O O B-TRANS O O O O O O O B-DATE I-DATE I-DATE O O O O O O O B-LOC I-LOC O B-LOC O B-LOC O B-LOC O O O B-DATE I-DATE I-DATE O O O O O O B-LOC I-LOC I-LOC I-LOC I-LOC I-LOC O	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. 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 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-LOC O B-LOC I-LOC O O O B-TRANS I-TRANS 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	O O B-PATIENT_ID O B-GENDER B-AGE O O O B-LOC O B-LOC O O B-LOC O B-LOC I-LOC O B-DATE I-DATE I-DATE O O O B-TRANS	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. 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".

		, được cách_ly tập_trung tại Trung_tâm Giáo_dục quốc_phòng an_ninh , ĐH Nha_Trang , Khánh_Hoà .	O O O O O B-LOC I-LOC I-LOC I-LOC O B-LOC I-LOC O B-LOC O	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 **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	O O O O O O O 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-

		viêm phổi trên bệnh_nhân 64 tuổi tuổi , sức_khoẻ suy_kiệt .	B-SYMP_DIS I-SYMP_DIS O O O O O O B-SYMP_DIS I-SYMP_DIS O	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" 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)	O O O O O O O O O B-PATIENT_ID O B-GENDER O B-AGE O O O O O O O B-PATIENT_ID O B-GENDER O B-AGE O O O O O O O B-PATIENT_ID O B-GENDER O B-AGE O O O O O O O B-PATIENT_ID O B-GENDER B-AGE O O O O O O O B-PATIENT_ID O B-GENDER O B-AGE O O O O O O O B-PATIENT_ID O B-GENDER O B-AGE O O O 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 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".

		; 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) ; bệnh_nhân 200 (nữ , 61 tuổi , quốc_tịch Việt_Nam) ; bệnh_nhân 222 (nữ , 28 tuổi , quốc_tịch Việt_Nam) .	O O B-PATIENT_ID O B-GENDER O B-AGE O O O O O O O B-PATIENT_ID O B-GENDER O B-AGE O O O O O O O B-PATIENT_ID O B-GENDER O B-AGE O O O O O O O B-PATIENT_ID O B-GENDER O B-AGE O O O O O O O B-PATIENT_ID O B-GENDER O B-AGE O O O O O O	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 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 O B-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG O O O O O O O B-NAME O O B-LOC I-LOC I-LOC O B-LOC I-LOC O O O O O O O B-NAME O O O O O	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-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG". 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".
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-LOC I-LOC I-LOC O B-LOC I-LOC I-LOC O B-LOC I-LOC O	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-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 O B-LOC I-LOC O B-LOC I-LOC O B-LOC I-LOC O	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. 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. 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	O O O O O O	1. Cụm các từ "UBND", "tỉnh", "Thanh_Hoá" chỉ Tên các cơ quan chính phủ: bộ ngành, uỷ ban nhân dân nên

		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 .	B-ORG I-ORG I-ORG O O O O O O B-NAME O B-LOC O O O O O O O O O O O O B-LOC O	được gán nhãn lần lượt là: "B-ORG", "I-ORG", "I-ORG". 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".
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-LOC I-LOC I-LOC I-LOC I-LOC I-LOC I-LOC 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-LOC I-LOC I-LOC O B-LOC I-LOC O O O O O	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".
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-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG O 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-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG". 3. Từ "904" 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.
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-LOC I-LOC O B-LOC I-LOC O B-LOC I-LOC O O O B-LOC I-LOC O B-LOC I-LOC 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. 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". 5. Cụm các từ "quận", "Sơn_Trà" tương tự trường hợp 4 nên lần lượt được gán nhãn là: "B-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". 8. Cụm các từ "huyện", "Quỳnh_Lưu" tương tự trường hợp 4 nên lần lượt được gán nhãn là: "B-LOCATION", "I-LOCATION".
57	Test Set - Dòng 17	Trước đó , khi điều_trị tại hai bệnh_viện ở Đà_Nẵng , bệnh_nhân viêm phổi	O O O O O O O O O B-LOC O O B-SYMP_DIS I-SYMP_DIS	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-

		nặng trên 10 năm , tràn khí màng phổi đã dẫn_lưu .	I-SYMP_DIS O O O O B-SYMP_DIS I-SYMP_DIS I-SYMP_DIS I-SYMP_DIS I-SYMP_DIS I-SYMP_DIS O	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".
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-ORG I-ORG I-ORG I-ORG I-ORG O O O O O O O O O O O O O O B-LOC I-LOC O B-LOC I-LOC O O O O O O B-SYMP_DIS O B-SYMP_DIS O B-SYMP_DIS 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-ORG", "I-ORG", "I-ORG", "I-ORG", "I-ORG". 3. Cụm các từ "xã", "Vinh_Hiền" chỉ Địa chỉ:cấp bậc đơn vị hành chính nên lần lượt được gán nhãn là: "B-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". 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-LOC I-LOC I-LOC O B-LOC I-LOC O O O O O O O O O O O O O O	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.**

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-LOC I-LOC I-LOC I-LOC I-LOC 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".

1.4. Nhận xét về ngữ liệu

Nhìn chung, 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. Khi phân tích 60 mẫu dữ liệu, nhóm thấy chúng được được đánh nhãn đầy đủ và tuân theo guideline một cách nghiêm ngặt, chỉ có 2 mẫu bị đánh nhãn sai không đáng kể (Mẫu thứ 14, 15 trong tập val, ứng với dòng thứ 34, 35 trong bảng phân tích). Bên cạnh đó, nhãn LOCATION và ORGANIZATION thường bị nhập nhằng, dễ bị nhầm lẫn với nhau

2 Phương pháp

2.1 Đầu vào, đầu ra mong đợi

Đầu vào: là một văn bản Tiếng Việt

Đầu ra: danh sách các thực thể có trong văn bản, mỗi thực thể được xác định bằng vị trí bắt đầu, kết thúc và nhãn loại thực thể tương ứng. Các loại thực thể bao gồm: PATIENT_ID, PERSON_NAME, AGE, GENDER, JOB, LOCATION, ORGANIZATION, SYMPTOM_AND_DISEASE, TRANSPORTATION, DATE.

VD:

- **Đầu vào:** 'Bệnh_nhân N.V.A bị viêm khớp'
- **Đầu ra:** ('NAME', 1, 1), ('SYMPTOM_AND_DISEASE', 3, 4). Các thực thể được biểu diễn theo định dạng (Nhãn loại thực thể, vị trí bắt đầu, vị trí kết thúc)

2.2 Các bước thực hiện chính:

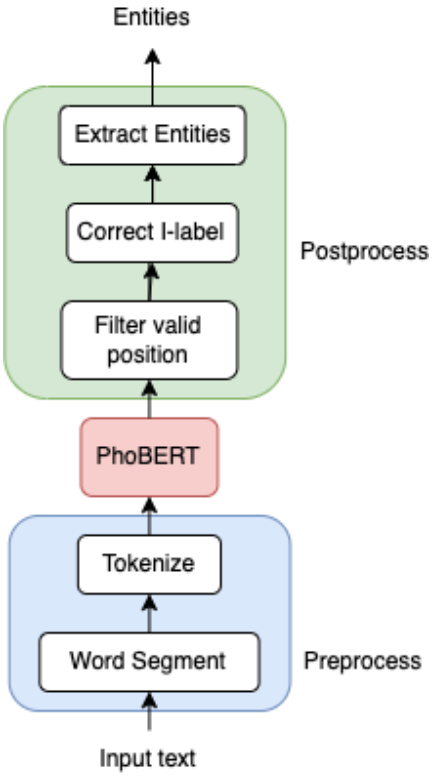


Figure 2.1: Sơ đồ các bước thực hiện phương pháp

B1: Sử dụng RDRSegmenter để segment văn bản đầu vào.

B2: Tokenize:

- Sử dụng PhoBERT Tokenizer để tách văn bản thành list các subwords
- Nếu độ dài list subwords > max_seq_len = 256 thì tiến hành bỏ 2 token cuối
- Thêm token [CLS] vào đầu list và token [SEP] vào cuối list. Nếu list subwords chưa đủ max_seq_len thì padding cho đủ độ dài. Cuối cùng chuyển các token thành token ids.
- Trong quá trình này, ta cần phải tạo list val_pos_list, slot_label_ids và mask. Trong đó:
    - val_pos_list: mảng một chiều có max_seq_len phần tử, mỗi phần tử val_pos_list[i] = True cho biết vị trí thứ i là vị trí của subword đầu tiên được tách ra của mỗi từ, ngược lại thì đó là vị trí của các token đặc biệt và subwords phụ.
    - slot_label_ids (Chỉ dùng trong training): mảng một chiều có max_seq_len phần tử, mỗi phần tử slot_label_ids[i] $\in [0; 20]$ cho biết nhãn BIO tương ứng với token thứ i trong chuỗi subwords, nếu slot_label_ids[i] = -100 thì nó sẽ không đóng góp vào hàm loss. Các token có nhãn -100 là các token đặc biệt, subwords phụ.
    - mask: mảng một chiều có max_seq_len phần tử, dùng để che các pad token trong chuỗi subwords
- Như vậy, đầu ra của bước này sẽ gồm 4 thông tin: input_ids (chứa các token ids), mask, val_pos_list, slot_label_ids (optional).

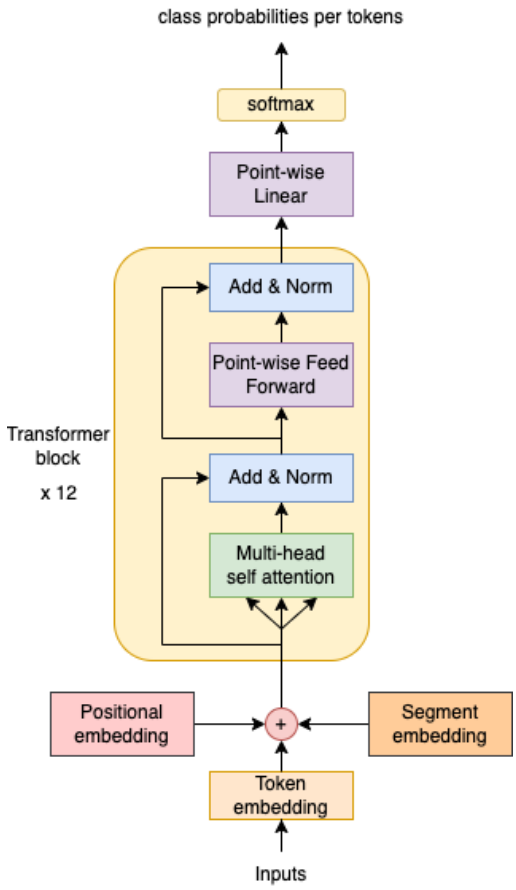


Figure 2.2: Kiến trúc PhoBERT cho bài toán NER

B3: Từ inputs_ids ở bước 2, ta chuyển thành các token embedding vectors. Sau đó bổ sung thêm các thông tin về vị trí của token trong chuỗi và token type để nhận được được các embedding vectors đại diện cho các token trong câu.

B4: Đưa các embedding vectors và mask vào các Transformer Block.

Mỗi embedding vector sẽ được biến đổi thành bộ 3 vector query, key, value $\in R^{768}$. Dựa trên một query vector, các key vectors và mask, ta sẽ tính được attention scores cho biết token ứng với query tập trung vào các tokens trong chuỗi như thế nào. Sau đó tính trung bình có trọng số (dựa trên attention scores) các value vectors để tính được vector đại diện cho token. (Mô tả trên đây dùng cho Multi head self attention có 1 head, có thể mở rộng ra cho nhiều head hơn).

Sau khi thu được các vector đại diện, ta cho chúng đi qua lớp Point-wise Feed Forward.

B5: Lấy đầu ra của khối transformer block cuối cho vào lớp point-wise linear với activation softmax để thu được xác suất 21 nhãn của từng token. Nếu trong quá trình training, ta sẽ dùng slot_label_ids để tính giá trị hàm mất mát rồi back propagate để cập nhật tham số mô hình.

B6: Với ma trận xác suất $\in R^{\max\_seq\_len\ x\ 21}$ ở bước 5, ta tính argmax để lấy ra được slot label ids. Sau đó dựa vào val_pos_list để lọc những slot label ids hợp lệ và chuyển thành slot labels (thay vì là id).

B7: Sửa nhãn I thành B nếu rơi vào 1 trong 3 trường hợp sau:

- Nhãn liền trước nhãn I là O. VD: [‘O’, ‘I-NAME’] → [‘O’, B-NAME’]
- Nhãn liền trước nhẫn I là nhãn chỉ loại thực thể khác. VD: [‘B-AGE’, ‘I-NAME’] → [‘B-AGE’, ‘B-NAME’]
- Nhãn I đứng đầu câu. VD: [‘I-NAME’, ‘O’] → [‘B-NAME’, ‘O’]

B8: Trích xuất các thực thể: Gôm các nhãn liên quan với nhau thành một thực thể (Nhãn loại thực thể, vị trí bắt đầu, vị trí kết thúc). VD: [‘B-NAME’, ‘I-NAME’, ‘O’, ‘B-JOB, ‘I-JOB, ‘O’] → (NAME, 0, 1); (JOB, 3, 4)

2.3 Phương pháp đánh giá

Với mỗi loại thực thể ta sẽ tính precision, recall và f1-score:

- Precision: cho biết tỷ lệ giữa số thực thể được dự đoán đúng so với số thực thể mà mô hình dự đoán nhãn là loại thực thể đó:

$$Pre\ = \frac{NE_{true}}{NE_{sys}}$$

- Recall: cho biết tỷ lệ số thực thể được dự đoán đúng so với số thực thể có nhãn thực sự là loại thực thể đó:

$$Rec = \frac{NE_{true}}{NE_{ref}}$$

- F1: là trung bình điều hoà của precision và recall

$$F1 = \frac{2 * Pre * Rec}{Pre + Rec}$$

Sau đó, ta tính macro f1 cho tất cả các loại thực thể bằng cách lấy trung bình cộng f1 của các loại thực thể.

3 Cài đặt

3.1 Môi trường cài đặt

- Google Colab bản thường
- GPU: Tesla T4
- RAM: 16GB

3.2 Thông số mô hình PhoBERT

3.2.1 Embedding

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 của 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 token type vocab size chỉ còn 1.

3.2.2 Multi-head self attention

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.

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, 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 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

3.2.3 Point-wise Feed Forward

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 768, số đầu vào bằng 3072

Dropout layer: dropout rate 0.1

Layer normalization

3.2.4 Classification layer

Là một point-wise linear layer có số node = 21, số đầu vào = 768, activation function softmax. Layer này dự đoán xác suất 21 nhãn của từng token. 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 Hàm mất mát

Mục đích của chúng ta là tối thiểu hàm mát mát sau:

$$L(\theta;\ X, y) = -\frac{1}{N}\sum_{k=0}^{N-1}\ \sum_{i=0}^{\text{max\_seq\_len}-1}\ \sum_{j=0}^{n\_slots-1} y_{k,i,j}\log\left(\hat{y}_{k,i,j}\right)$$

Trong đó:

- θ, (X, y) lần lượt là tham số mô hình, là dữ liệu trong một batch
- N, n_slots = 21 lần lượt là số mẫu dữ liệu trong một batch, số nhãn.

- $y_{k,i}, \hat{y}_{k,i}$ lần lượt là vector nhãn ground truth được biểu diễn dưới dạng one-hot coding và vector xác suất các nhãn mà mô hình dự đoán, tương ứng với token thứ i trong mẫu dữ liệu thứ k.

Lưu ý: các nhãn groundtruth không có nghĩa (pad_label_id = -100) sẽ không đóng góp vào hàm loss.

3.4 Các hyperparameters khác

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 → Sau 5 lần validate mà val loss không cải thiện thì sẽ dừng huấn luyện.

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 điểm optimal.

Dropout rate: 0.1

3.5 Source Code

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

4 Kết quả sơ bộ

4.1 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 cho macro f1 **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

Hình 4.1 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%**.

4.2 Phân tích kết quả đạt được

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**: $_{O}$Việt_Nam $_{O}$

  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: **Pred label**: $_{LOC}$ Việt_Nam $_{LOC}$

  True Label: $_{O}$Việt_Nam $_{O}$

- **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: $_{JOB}$ Bác_sĩ $_{JOB}$ Trương_Văn_Khải

- **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 $_{ORG}$

4.2.1 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

Hình 4.2 Bảng thống kê những TH 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 đúng	Lý do
1	['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', 'người',	[(['B-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION'], ['B-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']),	Vì các thực thể xuất hiện trong câu trên đều là những thực thể mà model được học đi học lại rất nhiều lần trong tập train nên model đã dự đoán đúng hoàn toàn.

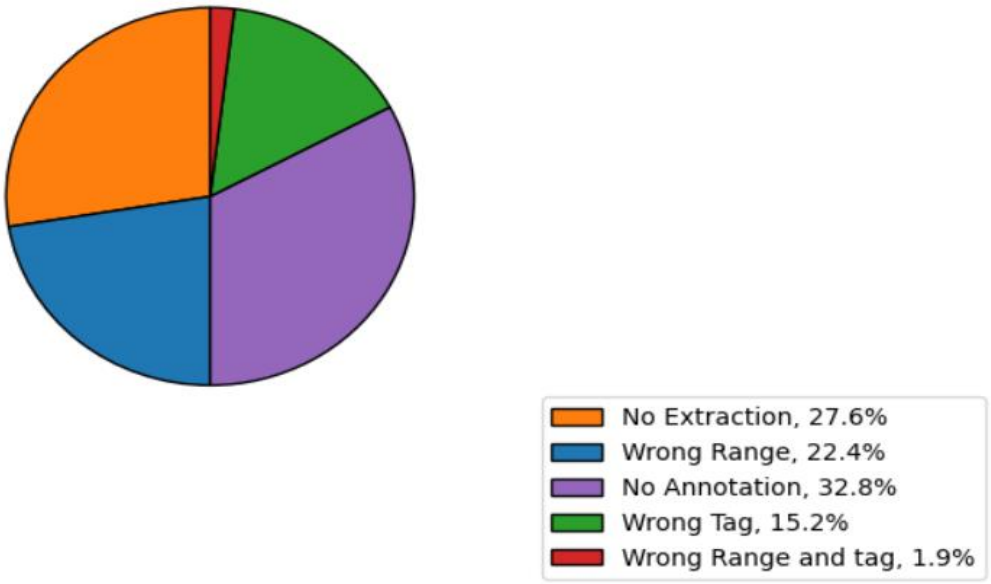
	'liên_quan', 'đến', 'ca', 'bệnh', 'số', '34', '.']	(['B-DATE', 'I-DATE', 'I-DATE'], ['B-DATE', 'I-DATE', 'I-DATE'], ['29', '-', '2']), (['B-LOCATION'], ['B-LOCATION'], ['Qatar']), (['B-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION'], ['B-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION'], ['Trung_tâm', 'Kiểm_soát', 'bệnh_tật', 'Bình_Thuận']), (['B-PATIENT_ID'], ['B-PATIENT_ID'], ['34'])]	
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', 'I-LOCATION'], ['huyện', 'Yên_Phong']), (['B-LOCATION', 'I-LOCATION'], ['B-LOCATION', 'I-LOCATION'], ['tỉnh', 'Bắc_Ninh'])]	Tương tự như trường hợp hai. Ngoài ra, trường hợp này mô hình đã dự đoán đúng nhãn ORGANIZATION so với các nhãn LOCATION, thông thường 2 nhãn này thường bị nhọc nhằn với nhau.
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',	[(['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-	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

	'có', 'kết_quả', 'âm_tính', '.']	LOCATION'], ['chợ', 'Siêu_Thị']), (['B-DATE'], ['B-DATE'], ['16/8'])]	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%).

Bảng 4.1 Bảng nhận xét một số trường hợp mô hình dự đoán đúng.

4.2.2 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:



Hình 4.3 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', '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', '.']	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'], ['B-LOCATION', 'I-LOCATION', 'I-LOCATION'], ['Bệnh_viện', 'Hữu_nghị', 'Việt_Tiệp'])]	Trong trường hợp này, model dự đoán từ "nữ" là nhãn "O", trong khi True Label là "B-GENDER".
2	['Đặc_biệt', 'chống', 'chỉ_định', 'với', 'người', 'có', 'bệnh_lý',	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-	Trong trường hợp này, model dự đoán từ "tim mạch" là 'B-SYMPTOM_AND_DISEASE'

	'tim_mạch', ',', 'trào', 'ngược', 'dạ_dày', '-', 'tá_tràng', ',', 'nhiễm_khuẩn', '...']		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 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',	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',	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

	'hồi_sức', 'tích_cực', ',', 'cho', 'biết', "", 'bệnh_nhân', '416', "", 'vẫn', 'đang', 'duy_trì', 'ECMO', ',', 'thở', 'máy', ',', 'hiện', 'xơ', 'phổi', 'rất', 'nhiều', '.']		'I-SYMPTOM_AND_DISEASE'], ['xơ', 'phổi', 'rất', 'nhiều'])]	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', 'Eschool', '-', 'Eclass', ',', 'hiện', 'ngụ', 'tại', 'đường', 'Tôn_Đản', ',', 'phường', '13', ',', 'quận', '4', ',', 'TP.HCM.']	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'], ['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.'])]	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-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	No Annotation: [(['O', 'O'], ['B-ORGANIZATION', 'I-ORGANIZATION'], ['Trung_tâm', 'CDC'])] Wrong tag: [(['B-ORGANIZATION', 'I-ORGANIZATION', 'I-ORGANIZATION', 'I-	[(['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.

	', 'với', 'SARS', '-', 'CoV', '-', '2', ',', 'bệnh_nhân', 'được', 'chuyển', 'tới', 'Bệnh_viện', 'Nhiệt_đới', 'trung_ương', '2', '.']	ORGANIZATION'], ['B-LOCATION', 'I-LOCATION', 'I-LOCATION', 'I-LOCATION'], ['Bệnh_viện', 'Nhiệt_đới', 'trung_ương', '2'])]		

Bảng 4.2 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. Có thể tham khảo ở file Excel này:

Link: https://1drv.ms/x/s!AvIJzwvRlszaqnlPJW0thtiJRevc?e=f4rgii