

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - CƠ - TIN HỌC



Nguyễn Chí Dũng

XÂY DỰNG DỮ LIỆU
VÀ TRÍCH XUẤT QUAN HỆ TRONG
VĂN BẢN Y TẾ TIẾNG VIỆT

Khóa luận tốt nghiệp đại học hệ chính quy
Ngành Khoa học dữ liệu
(Chương trình đào tạo chuẩn)

Hà Nội - 2024

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - CƠ - TIN HỌC

Nguyễn Chí Dũng

XÂY DỰNG DỮ LIỆU
VÀ TRÍCH XUẤT QUAN HỆ TRONG
VĂN BẢN Y TẾ TIẾNG VIỆT

Khóa luận tốt nghiệp đại học hệ chính quy
Ngành Khoa học dữ liệu
(Chương trình đào tạo chuẩn)

Cán bộ hướng dẫn: TS. Nguyễn Thị Minh Huyền
TS. Nguyễn Hải Vinh

18

Hà Nội - 2024

Lời cảm ơn

Để thực hiện và hoàn thành được khóa luận tốt nghiệp này, trước tiên em xin gửi lời cảm ơn sâu sắc đến TS. Nguyễn Thị Minh Huyền và TS. Nguyễn Hải Vinh là những người đã trực tiếp hướng dẫn truyền đạt cho em những kinh nghiệm quá báu trong suốt quá trình nghiên cứu.

Đồng thời em xin gửi lời cảm ơn chân thành tới tập thể các thầy cô, anh chị, các em sinh viên khóa dưới thuộc khoa Toán-Cơ-Tin học đã đồng hành giúp đỡ và tạo mọi điều kiện cho em trong suốt quá trình thực hiện đề tài.

Em cũng xin bày tỏ lòng biết ơn chân thành đến tập thể các thầy cô giáo Khoa Toán-Cơ-Tin học, Trường Đại học Khoa học Tự nhiên - Đại học Quốc gia Hà Nội truyền thụ những kiến thức quý báu cho em trong suốt 4 năm học qua.

Em cũng xin gửi lời cảm ơn chân thành đến Tổ chức VLSP vì đã cấp phép sử dụng dữ liệu tư nhân trong quá trình nghiên cứu. Sự hỗ trợ quý báu này là vô cùng quan trọng, giúp em có thể tiến hành nghiên cứu một cách thuận lợi.

Cuối cùng, em xin cảm ơn gia đình, bạn bè, những người luôn quan tâm giúp đỡ và động viên, khuyến khích em trong suốt thời gian qua để em hoàn thành khóa luận được tốt hơn.

Hà Nội, 20-05-2024

Sinh viên

Nguyễn Chí Dũng

Lời cam đoan

Em xin cam đoan bài báo cáo này là công trình học tập và nghiên cứu thật sự nghiêm túc của bản thân dưới sự hướng dẫn khoa học của TS. Nguyễn Thị Minh Huyền và TS. Nguyễn Hải Vinh. Kết quả nêu ra trong nghiên cứu này là trung thực và chưa chưa từng được sử dụng để hoàn thành bất kỳ khóa luận tốt nghiệp nào khác. Các số liệu trong bài nghiên cứu có nguồn gốc rõ ràng, được tổng hợp từ những nguồn thông tin đáng tin cậy. Tất cả các tài liệu tham khảo cần thiết và các nghiên cứu liên quan được trích dẫn chính xác, và đảm bảo rằng không có đạo văn.

Hà Nội, 20-05-2024

Sinh viên

Nguyễn Chí Dũng

Mục lục

Danh mục thuật ngữ	6
Giới thiệu	11
1 Giới thiệu các khái niệm cơ bản liên quan tới vấn đề trích rút thông tin trong văn bản và trích rút quan hệ trong văn bản y tế	13
1.1 Hệ thống Ngôn ngữ Y tế hợp nhất - Unified Medical Language System	13
1.1.1 UMLS Metathesaurus	13
1.1.2 UMLS Semantic Network	14
1.2 Trích rút thông tin trong văn bản y tế	14
1.2.1 Nhận dạng thực thể có tên	14
1.2.2 Trích rút quan hệ thực thể	16
1.2.3 Trích rút quan hệ thực thể trong văn bản y tế	17
2 Ứng dụng mô hình PhoBERT cho bài toán trích rút quan hệ	19
2.1 Mô hình Transformer	19
2.1.1 Tổng quan về mô hình Transformer	19
2.1.2 Kiến trúc tổng quát của mô hình Transformer	23
2.2 Mô hình PhoBERT	28
2.2.1 Mô hình BERT	28

2.2.2	Mô hình RoBERTa	29
2.2.3	Mô hình PhoBERT	30
2.3	Thử nghiệm	32
2.3.1	Bộ dữ liệu VLSP 2020 Relation Extraction	32
2.3.2	Bộ công cụ xử lý tiếng Việt Underthesea	34
2.3.3	Tiền xử lý dữ liệu	35
2.3.4	Tinh chỉnh mô hình PhoBERT	36
2.3.5	Phương pháp đánh giá	37
2.3.6	Kết quả thực nghiệm	38
3	Xây dựng dữ liệu trích rút quan hệ thực thể y tế Tiếng Việt	40
3.1	Xây dựng dữ liệu	40
3.1.1	Hệ thống phân loại bệnh quốc tế ICD-9 và ICD-10	40
3.1.2	Cơ sở dữ liệu bách khoa toàn thư mở trực tuyến - Wikipedia	41
3.1.3	Công cụ gán nhãn INCEption	42
3.1.4	Thu thập và tiền xử lý dữ liệu	42
3.1.5	Định nghĩa nhãn các thực thể ánh xạ đến UMLS	43
3.1.6	Định nghĩa nhãn quan hệ giữa các thực thể	52
3.1.7	Gán nhãn thực thể và quan hệ giữa các thực thể	54
3.2	Thực nghiệm và kết quả	56
3.2.1	Bộ dữ liệu trích xuất quan hệ cho văn bản y tế tiếng Việt	56
3.2.2	Tiền xử lý dữ liệu, tinh chỉnh mô hình PhoBERT và phương pháp đánh giá	56
3.2.3	Kết quả thực nghiệm	56
Kết luận		58
Tài liệu tham khảo		60

Danh mục thuật ngữ

UMLS	Unified Medical Language System - Hệ thống Ngôn ngữ Y tế hợp nhất
ICD-9	International Classification of Diseases, Ninth Revision - Hệ thống phân loại bệnh quốc tế, bản sửa đổi lần thứ 9
ICD-10	International Classification of Diseases, Tenth Revision - Hệ thống phân loại bệnh quốc tế, bản sửa đổi lần thứ 10
WHO	World Health Organization - Tổ chức Y tế Thế giới
BERT	Bidirectional Encoder Representations from Transformers - Mô hình học sẵn biểu diễn Thẻ hiện Mã hóa Hai chiều từ Transformer
NLP	Natural Language Processing - Xử lý ngôn ngữ tự nhiên
RoBERTa	Robustly Optimized BERT Approach - Phương pháp tiếp cận BERT được tối ưu hóa mạnh mẽ
TSV	Tab-Separated Values - Các giá trị được phân cách bằng tab
POS	Part-of-Speech - Từ loại
NER	Named entity recognition - Nhận diện thực thể có tên
RE	Relation Extraction - Trích xuất quan hệ
NLI	Natural language inference - Suy luận ngôn ngữ tự nhiên
DP	Dependency Parsing - Phân tích phụ thuộc
XLM-	Cross-Lingual Model - Robustly Optimized BERT
RoBERTa	Approach - Mô hình đa ngôn ngữ - Phương pháp tiếp cận BERT được tối ưu hóa mạnh mẽ

Danh sách bảng

2.1 Các quan hệ trong bộ dữ liệu VLSP 2020 Relation Extraction	33
2.2 Hiệu suất của mô hình trên các tập dữ liệu huấn luyện và kiểm thử của bộ dữ liệu VLSP 2020 Relation Extraction	38
2.3 Hiệu suất của mô hình trên các nhãn trong tập dữ liệu kiểm thử của bộ dữ liệu VLSP 2020 Relation Extraction	39
3.1 Các ví dụ về SYMPTOM AND DISEASE cần được gán nhãn	44
3.2 Các ví dụ về SYMPTOM AND DISEASE không được gán nhãn	45
3.3 Các ví dụ về THERAPEUTIC OR PREVENTIVE PROCEDURE cần được gán nhãn	46
3.4 Các ví dụ về THERAPEUTIC OR PREVENTIVE PROCEDURE không được gán nhãn	47
3.5 Các ví dụ về TESTS cần được gán nhãn	48
3.6 Các ví dụ về TESTS không được gán nhãn	49
3.7 Các ví dụ về MEDICINE cần được gán nhãn	50
3.8 Các ví dụ về MEDICINE không được gán nhãn	51
3.9 Các ví dụ về BODY LOCATION OR REGION cần được gán nhãn	52
3.10 Mối quan hệ ngữ nghĩa tổng quát giữa các thực thể	53
3.11 Hiệu suất của mô hình trên các tập dữ liệu huấn luyện và kiểm thử cho tập dữ liệu mẫu trích xuất quan hệ cho văn bản y tế tiếng Việt	57
12 Các ví dụ về SYMPTOM AND DISEASE cần được gán nhãn	64

13	Các ví dụ về SYMPTOM AND DISEASE không được gán nhãn	66
14	Các ví dụ về THERAPEUTIC OR PREVENTIVE PROCEDURE cần được gán nhãn	67
15	Các ví dụ về THERAPEUTIC OR PREVENTIVE PROCEDURE không được gán nhãn	68
16	Các ví dụ về TESTS cần được gán nhãn	69
17	Các ví dụ về TESTS không được gán nhãn	70
18	Các ví dụ về MEDICINE cần được gán nhãn	71
19	Các ví dụ về MEDICINE không được gán nhãn	72
20	Các ví dụ về BODY LOCATION OR REGION cần được gán nhãn	73
21	Mối quan hệ ngữ nghĩa giữa các thực thể	74

Danh sách hình vẽ

2.1	Mô hình ngôn ngữ nhân quả [1]	20
2.2	Mô hình ngôn ngữ có mặt nạ [1]	21
2.3	Kích thước của một số mô hình Transformer theo thống kê từ Đại học Carnegie Mellon	21
2.4	Lượng khí thải CO ₂ cho các hoạt động, [1]	22
2.5	Học chuyển giao [2]	22
2.6	Kiến trúc tổng quát của mô hình Transformer [2]	23
2.7	Quá trình sinh Q, K, V	25
2.8	Minh họa quá trình tính điểm chú ý cho tất cả các từ trong câu	26
2.9	Cấu trúc chú ý đa đầu	26
2.10	Quy trình của mô hình Transformer [2]	27
2.11	Mô hình BERT [3]	29
2.12	Sự khác biệt giữa các token âm tiết và từ của tiếng Việt	30
2.13	Kết quả cho POS tagging và NER [4]	31
2.14	Kết quả cho NLI và DP [4]	32
2.15	Tệp dưới định dạng WebAnno TSV phiên bản 3.2	34
3.1	Giao diện công cụ gán nhãn INCEpTION [5])	42
3.2	Danh sách các văn bản được gán nhãn trên INCEpTION)	43
3.3	Gán nhãn thực thể và quan hệ trong văn bản y tế tiếng Việt trên hệ thống INCEpTION	54

3.4 Kiểm định chéo các phiên bản trên hệ thống INCEPTION	55
3.5 Tệp dưới định dạng TSV phiên bản 3.3	55

Giới thiệu

Ngôn ngữ y tế là một lĩnh vực chuyên ngành phức tạp, việc xử lý và phân tích thông tin y tế tự động sẽ mang lại nhiều lợi ích cho việc chăm sóc sức khỏe và nghiên cứu y học. Xây dựng dữ liệu trích rút quan hệ thực thể y tế là một bước tiến quan trọng trong việc thúc đẩy ứng dụng công nghệ thông tin vào lĩnh vực này. Tuy nhiên, trong khi ngôn ngữ tiếng Anh sở hữu nguồn tài nguyên phong phú về dữ liệu y tế, các ngôn ngữ ít được quan tâm hơn như tiếng Việt lại gần như không có sẵn bộ dữ liệu trích rút quan hệ thực thể y tế nào.

Sự chênh lệch này xuất phát từ nhiều nguyên nhân, bao gồm:

- Sự phổ biến của tiếng Anh: tiếng Anh là ngôn ngữ chính trong nghiên cứu và xuất bản y học quốc tế, dẫn đến sự tập trung nguồn lực vào việc xây dựng dữ liệu cho ngôn ngữ này.
- Khả năng tiếp cận công nghệ: Các quốc gia sử dụng tiếng Anh thường có điều kiện tiếp cận công nghệ tiên tiến và nguồn lực tài chính dồi dào hơn, tạo điều kiện thuận lợi cho việc phát triển dữ liệu y tế.
- Sự khác biệt về cấu trúc ngôn ngữ: tiếng Việt với cấu trúc ngữ pháp và hệ thống từ vựng riêng biệt, đòi hỏi những phương pháp xử lý ngôn ngữ tự nhiên đặc thù, gây khó khăn cho việc áp dụng trực tiếp các mô hình được phát triển cho tiếng Anh.

Khóa luận này tập trung vào việc xây dựng dữ liệu trích rút quan hệ thực thể y tế tiếng Việt, với mục tiêu tạo ra một bộ dữ liệu đáng tin cậy và hiệu quả, góp phần thu hẹp khoảng cách về nguồn lực dữ liệu y tế giữa tiếng Việt và các ngôn ngữ phổ biến khác.

Các bước chính trong quá trình thực hiện khóa luận bao gồm:

- Xây dựng bộ nhãn thực thể và quan hệ y tế theo quy chuẩn quốc tế được định nghĩa trong hệ thống ngôn ngữ y tế thống nhất (*Unified Medical Lan-*

guage System, UMLS). Việc này đảm bảo tính thống nhất và khả năng tương tác với các hệ thống y tế quốc tế.

- Xây dựng dữ liệu dựa trên văn bản y tế đáng tin cậy được đánh dấu theo hệ thống phân loại bệnh tật quốc tế (*International Classification of Diseases, ICD*). Nguồn dữ liệu này đảm bảo tính chính xác và độ tin cậy của thông tin y tế.
- Áp dụng mô hình PhoBERT, một mô hình xử lý ngôn ngữ tự nhiên tiên tiến cho tiếng Việt, để trích xuất quan hệ giữa các thực thể y tế đã được xây dựng.

Khóa luận sẽ trình bày chi tiết về quy trình xây dựng dữ liệu, phương pháp trích rút và kết quả đạt được.

Nội dung chính của báo cáo bao gồm ba chương như sau:

- Chương 1: Giới thiệu các khái niệm cơ bản liên quan tới vấn đề trích rút thông tin trong văn bản và trích rút quan hệ trong văn bản y tế.
- Chương 2: Ứng dụng mô hình PhoBERT cho bài toán trích rút quan hệ.
- Chương 3: Xây dựng dữ liệu trích rút quan hệ thực thể y tế tiếng Việt.

Chương 1

Giới thiệu các khái niệm cơ bản liên quan tới vấn đề trích rút thông tin trong văn bản và trích rút quan hệ trong văn bản y tế

1.1 Hệ thống Ngôn ngữ Y tế hợp nhất - Unified Medical Language System

1.1.1 UMLS Metathesaurus

UMLS Metathesaurus¹ là một nguồn tài nguyên từ vựng y tế đa ngôn ngữ tích hợp được phát triển bởi Thư viện Y khoa Quốc gia Hoa Kỳ (*National Library of Medicine*, NLM). Siêu từ điển (*Metathesaurus*) này cung cấp một cấu trúc hợp nhất để đại diện các khái niệm về y tế từ hơn 200 nguồn từ vựng khác nhau, bao gồm từ điển, phân loại bệnh tật, nội dung y văn và nhiều nguồn khác. Các khái niệm trong siêu từ điển được gán một định danh duy nhất gọi là *Concept Unique Identifier* (CUI) và được liên kết với các thuật ngữ tương đương từ các nguồn khác nhau.

Mục đích chính của UMLS Metathesaurus là tăng cường khả năng trao đổi

¹<https://uts.nlm.nih.gov/uts/umls/home>

và chia sẻ dữ liệu giữa các hệ thống thông tin y tế khác nhau bằng cách cung cấp một ngôn ngữ chung để đại diện cho các khái niệm y khoa. Nó cho phép ánh xạ thuật ngữ từ nhiều nguồn khác nhau vào một tập hợp khái niệm hợp nhất, hỗ trợ việc truy xuất thông tin, xử lý ngôn ngữ tự nhiên và các ứng dụng y tế khác. UMLS Metathesaurus là một công cụ quan trọng trong lĩnh vực tin học y sinh, hỗ trợ việc nghiên cứu, phát triển và tích hợp hệ thống thông tin y tế.

1.1.2 UMLS Semantic Network

UMLS Semantic Network² là một phần quan trọng trong Hệ thống Ngôn ngữ Y tế hợp nhất (UMLS) do Thư viện Y khoa Quốc gia Hoa Kỳ phát triển. Mạng ngữ nghĩa - Semantic Network cung cấp một bộ khung ngữ nghĩa bao gồm các loại ngữ nghĩa (*semantic types*) được sử dụng để phân loại và mô tả các khái niệm y tế có trong UMLS Metathesaurus. Nó cũng xác định các mối quan hệ ngữ nghĩa (*semantic relations*) giữa các loại ngữ nghĩa này.

Mạng ngữ nghĩa (*Semantic Network*) hiện có hơn 130 loại ngữ nghĩa như "Bệnh hay Hội chứng (*Disease or Syndrome*)", "Phương pháp Điều trị (*Therapeutic or Preventive Procedure*)", "Thuốc lâm sàng (*Clinical Drug*)" và hơn 50 loại mối quan hệ như "gây ra (*cause*)", "phòng ngừa (*prevent*)", "điều trị (*treat*)". Những phân loại và mối liên kết này cho phép tổ chức, phân nhóm và kết nối các khái niệm y tế một cách có hệ thống theo ngữ nghĩa, giúp việc phân tích và truy xuất thông tin trở nên hiệu quả hơn.

1.2 Trích rút thông tin trong văn bản y tế

1.2.1 Nhận dạng thực thể có tên

Nhận dạng thực thể có tên (*Named entity recognition*, NER) đề cập đến nhiệm vụ xác định tiếp đến là phân loại các thực thể được gán nhãn khác nhau từ văn bản. Nhận dạng thực thể có tên có nhiều ứng dụng như hệ thống trả lời câu hỏi, hệ thống trích xuất thông tin. Việc trích xuất các thực thể cũng cho phép thực hiện các nhiệm vụ nghiên cứu khác nhau như thực hiện phân tích ngữ nghĩa cũng như sắc thái giữa các thực thể khác nhau hoặc biên soạn tất cả các tài liệu

²<https://uts.nlm.nih.gov/uts/umls/semantic-network/T200>

tham khảo liên quan cho một thực thể duy nhất. Nhận dạng thực thể có tên có thể cải thiện công cụ tìm kiếm bằng cách giúp công cụ tìm kiếm hiểu rõ ý nghĩa của thông tin cần được truy vấn và cung cấp thông tin liên quan dựa trên ngữ cảnh, từ đó mang đến trải nghiệm tìm kiếm chính xác và cá nhân hóa hơn cho người dùng.

Một số bộ dữ liệu nổi bật có thể được kể đến như:

- CoNLL-2003 [6] - bộ dữ liệu này đã được giới thiệu trong nhiệm vụ chia sẻ CoNLL-2003 (*CoNLL-2003 shared task*) và đã được sử dụng rộng rãi để đánh giá Nhận dạng thực thể có tên. Nó bao gồm các bài báo từ Reuters Corpus, được chú thích với bốn loại thực thể: Người, Tổ chức, Vị trí và Khác.
- WNUT 2017 [7] - bộ dữ liệu đã được giới thiệu cho nhiệm vụ chia sẻ nhận dạng thực thể mới nổi và mới lạ (*Emerging and Novel Entity Recognition shared task*). Nó bao gồm dữ liệu được chú thích từ các miền khác nhau, bao gồm phương tiện truyền thông xã hội, diễn đàn trực tuyến và đánh giá của người tiêu dùng. Nó bao gồm sáu loại thực thể: Người, Vị trí, Công ty, Sản phẩm, Công việc sáng tạo và Nhóm.
- GENIA corpus [8] - bộ dữ liệu phổ biến cho Nhận dạng thực thể có tên y sinh. Nó bao gồm các bản tóm tắt từ các tài liệu nghiên cứu y sinh, được chú thích với nhiều loại thực thể khác nhau, chẳng hạn như Protein, DNA, Dòng tế bào và các loại khác.

Ngoài ra một số bộ dữ liệu tiếng Việt trong văn bản bao gồm văn bản y tế có thể kể đến như:

- Bộ dữ liệu VLSP 2021 Named Entity Recognition for Vietnamese³ - bộ dữ liệu bao gồm các văn bản báo chí đã được gán nhãn thực thể đặc biệt bao gồm tên người, tổ chức, địa điểm, sự kiện, thời gian và số liệu quan trọng. Bộ dữ liệu cung cấp một nền tảng đáng tin cậy để nghiên cứu và phát triển các kỹ thuật NER cho tiếng Việt trong lĩnh vực tin tức báo chí.

³<https://vlsp.org.vn/vlsp2021/eval/ner>

- Bộ dữ liệu ViMQ [9] - bộ dữ liệu gồm một tập hợp câu hỏi y tế bằng tiếng Việt từ người bệnh, được gán nhãn ở cấp câu và cấp thực thể. Nó phục vụ cho các nhiệm vụ phân loại Ý định (*Intent Classification*) và Nhận dạng thực thể có tên (*Named Entity Recognition*) trong lĩnh vực y tế, nhằm hỗ trợ phát triển chatbot y tế với khả năng hiểu tốt hơn câu hỏi của bệnh nhân.

1.2.2 Trích rút quan hệ thực thể

Trích xuất quan hệ (*Relation Extraction*, RE) là một trong những nhiệm vụ chính liên quan đến trích xuất thông tin. Nó đề cập đến phân loại mỗi quan hệ ngữ nghĩa có thể tồn tại giữa các thực thể. Loại thông tin này rất cần thiết để xây dựng cơ sở kiến thức ngữ nghĩa (*knowledge bases* - KBs), thứ có thể được sử dụng để suy luận các mối quan hệ tồn tại giữa các thực thể khác nhau. Trích xuất quan hệ rất hữu ích cho việc phát triển hệ thống trả lời câu hỏi, thực hiện tóm tắt văn bản và xây dựng phân loại khái niệm.

Một số bộ dữ liệu nổi bật có thể được kể đến như:

- SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals [10] - bộ dữ liệu được sử dụng trong cuộc thi SemEval năm 2010, tập trung vào việc phân loại các quan hệ ngữ nghĩa giữa các cặp danh từ, góp phần vào việc hiểu và xử lý ngôn ngữ tự nhiên hiệu quả hơn. Điều này có ý nghĩa quan trọng trong việc cải thiện giao tiếp giữa con người và máy tính, làm tăng trải nghiệm người dùng và mang lại nhiều tiện ích trong cuộc sống hàng ngày.
- TACRED (TAC Relation Extraction Dataset) [11] - bộ dữ liệu lớn được xây dựng bởi các nhà nghiên cứu tại Đại học Stanford, tập trung vào lĩnh vực tin tức và văn bản web, giúp cải thiện khả năng trích rút thông tin từ các nguồn dữ liệu phức tạp và đa dạng. Điều này có tác động lớn trong việc tóm tắt và tổng hợp thông tin, hỗ trợ ra quyết định dựa trên dữ liệu, cũng như nâng cao hiệu quả trong các lĩnh vực như truyền thông, báo chí, và quản lý thông tin.
- SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers [12] - bộ dữ liệu này được sử dụng trong cuộc thi SemEval

năm 2018, tập trung vào việc trích rút và phân loại các quan hệ ngữ nghĩa trong các tài liệu khoa học. Điều này có tác động sâu rộng trong việc tận dụng và khai thác hiệu quả các nguồn tri thức khoa học, thúc đẩy sự phát triển của nghiên cứu khoa học và công nghệ, từ đó đóng góp vào sự tiến bộ của xã hội

1.2.3 Trích rút quan hệ thực thể trong văn bản y tế

Trong lĩnh vực y tế, việc xây dựng các bộ dữ liệu về trích rút quan hệ thực thể trong văn bản là vô cùng cần thiết và quan trọng. Điều này đóng vai trò then chốt trong việc nâng cao hiệu quả khai thác thông tin từ các tài liệu y khoa, hỗ trợ đắc lực cho các nghiên cứu và ứng dụng trong lĩnh vực này.

Thứ nhất, khả năng trích rút các quan hệ giữa các thực thể y tế như tên bệnh, triệu chứng, thuốc điều trị từ các văn bản như hồ sơ bệnh án, báo cáo y tế, hoặc tài liệu nghiên cứu khoa học có ý nghĩa quan trọng trong việc hỗ trợ ra quyết định lâm sàng, theo dõi và quản lý bệnh tật hiệu quả hơn.

Thứ hai, các bộ dữ liệu về trích rút quan hệ thực thể trong văn bản y tế là nền tảng thiết yếu cho việc nghiên cứu và phát triển các hệ thống trí tuệ nhân tạo (Artificial Intelligence, AI) trong lĩnh vực y tế. Chúng cho phép huấn luyện và đánh giá các mô hình học máy, đặc biệt là các mô hình học sâu, nhằm cải thiện khả năng hiểu và xử lý ngôn ngữ tự nhiên trong lĩnh vực chuyên ngành này.

Trên thế giới, một số bộ dữ liệu nổi bật về trích rút quan hệ thực thể trong văn bản y tế bao gồm:

- Bộ dữ liệu i2b2 2010 [13]: Được giới thiệu trong cuộc thi i2b2 2010, bộ dữ liệu này chứa các văn bản y tế được ghi chú về các quan hệ giữa các thực thể như bệnh lý, kiểm tra, điều trị,...
- Bộ dữ liệu BioNLP 2011 [14]: Được sử dụng trong cuộc thi BioNLP 2011, bộ dữ liệu này tập trung vào việc trích rút các quan hệ trong các văn bản liên quan đến lĩnh vực sinh học phân tử.

Các bộ dữ liệu này đã thúc đẩy quá trình nghiên cứu và phát triển hệ thống

trích rút quan hệ trong y tế, góp phần nâng cao chất lượng chăm sóc sức khỏe và thúc đẩy sự tiến bộ của ngành y học.

Tại thời điểm hiện tại, việc xây dựng các bộ dữ liệu về trích rút quan hệ trong văn bản y tế tiếng Việt vẫn còn hạn chế.

Bộ dữ liệu đáng chú ý nhất là bộ dữ liệu Trích rút Quan hệ từ VLSP 2020⁴ được xây dựng và công bố tại hội thảo về xử lý ngôn ngữ và tiếng nói tiếng Việt năm 2020. Ngoài ra, còn có một số bộ dữ liệu nhỏ hơn được xây dựng bởi các nhóm nghiên cứu tại các trường đại học nhưng vẫn đang trong giai đoạn nghiên cứu sơ khai và chưa được công bố rộng rãi.

⁴<https://vlsp.org.vn/vlsp2020/eval/re>

Chương 2

Ứng dụng mô hình PhoBERT cho bài toán trích rút quan hệ

2.1 Mô hình Transformer

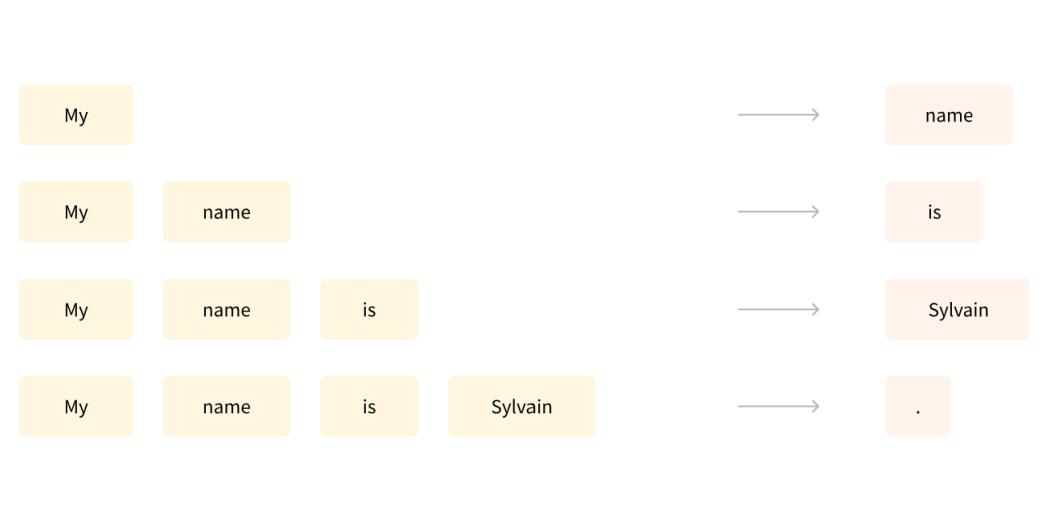
2.1.1 Tổng quan về mô hình Transformer

Transformer [2] là một loại kiến trúc mạng nơ ron (*Neural Networks*) đạt được hiệu suất mạnh mẽ trên nhiều tác vụ xử lý ngôn ngữ tự nhiên và cả trên một số tác vụ thị giác máy tính. Transformer thường được sử dụng làm cơ sở để giải quyết các vấn đề xử lý ngôn ngữ tự nhiên hiện nay.

Transformer là mô hình ngôn ngữ thường được xây dựng dựa trên hai phương pháp:

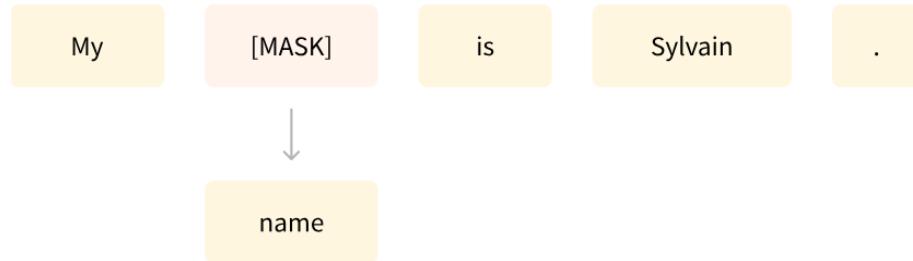
- Mô hình ngôn ngữ nhân quả (*Causal Language Modeling*) là một phương pháp tự hồi quy trong đó mô hình được huấn luyện để dự đoán token tiếp theo trong một chuỗi dựa vào các token trước đó. Mô hình ngôn ngữ nhân quả được sử dụng trong các mô hình GPT (*Generative Pre-trained Transformer*) và rất phù hợp cho các tác vụ như tạo và tóm tắt văn bản. Tuy nhiên, các mô hình ngôn ngữ nhân quả có bối cảnh đơn hướng, có nghĩa là chúng chỉ xem xét bối cảnh quá khứ chứ không phải bối cảnh tương lai khi đưa ra dự đoán. Ví dụ về một tác vụ là dự đoán từ tiếp theo trong một

câu sau khi đã đọc n từ trước đó. Đây được gọi là mô hình ngôn ngữ nhân quả bởi vì đầu ra phụ thuộc vào các đầu vào quá khứ và hiện tại, nhưng không phụ thuộc vào các đầu vào tương lai.



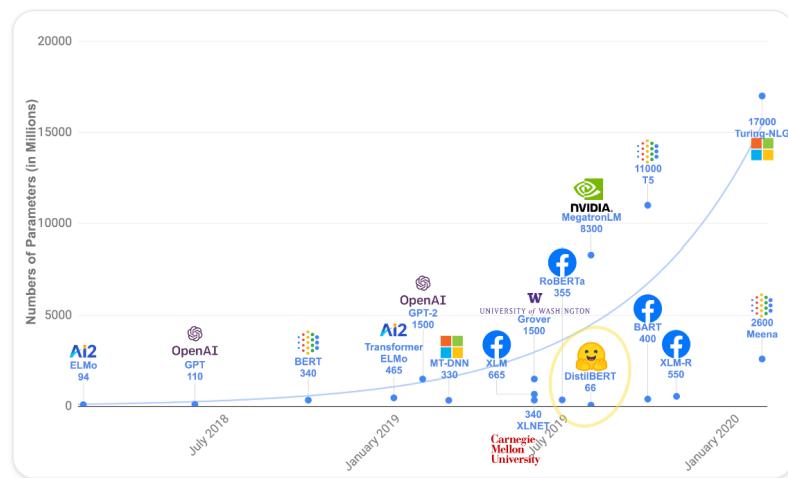
Hình 2.1: Mô hình ngôn ngữ nhân quả [1]

- Mô hình ngôn ngữ có mặt nạ (*Masked Language Modeling*) là một phương pháp huấn luyện được sử dụng trong các mô hình như BERT, trong đó một số token trong chuỗi đầu vào được che giấu (*masked*) và mô hình học cách dự đoán các *masked token* dựa trên bối cảnh xung quanh. Mô hình ngôn ngữ có mặt nạ có lợi thế về bối cảnh hai chiều, cho phép mô hình xem xét cả token trong quá khứ và tương lai khi đưa ra dự đoán. Cách tiếp cận này đặc biệt hữu ích cho các tác vụ như phân loại văn bản, phân tích cảm xúc và nhận dạng thực thể được gán nhãn.



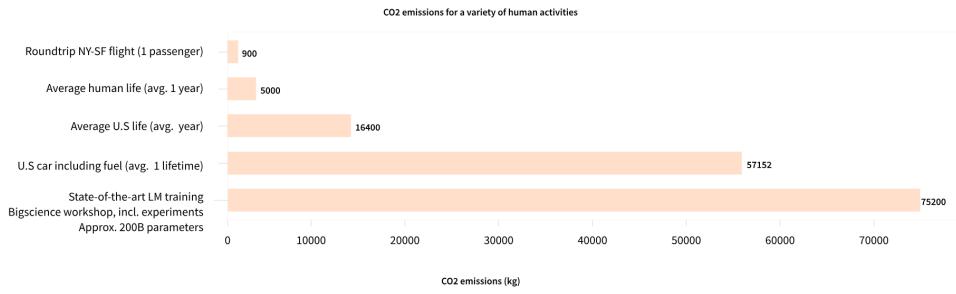
Hình 2.2: Mô hình ngôn ngữ có mặt nạ [1]

Transformer được huấn luyện trên lượng dữ liệu khổng lồ.



Hình 2.3: Kích thước của một số mô hình Transformer theo thống kê từ Đại học Carnegie Mellon

Huấn luyện các mô hình Transformer rất tốn kém.



Hình 2.4: Lượng khí thải CO₂ cho các hoạt động, [1]

Do vậy, chúng ta tận dụng các mô hình được huấn luyện sẵn (*pretrained models*) và học chuyển giao (*transfer learning*)

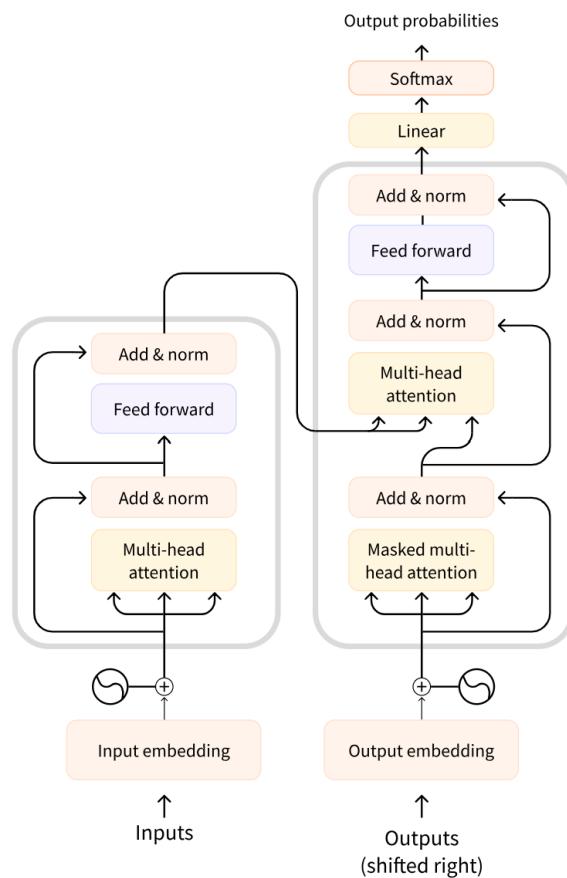
Quá trình tinh chỉnh (*fine-tuning process*) do đó có thể lấy lợi thế về kiến thức có được bởi mô hình ban đầu trong quá trình tiền huấn luyện. Do mô hình đã được tiền huấn luyện (*pretrained model*) đã được huấn luyện trên rất nhiều dữ liệu, quá trình tinh chỉnh (*fine-tuning*) yêu cầu dữ liệu ít hơn rất nhiều để có kết quả tốt và cũng dẫn đến lượng thời gian và tài nguyên cần thiết là ít hơn rất nhiều



Hình 2.5: Học chuyển giao [2]

2.1.2 Kiến trúc tổng quát của mô hình Transformer

- Bộ mã hóa - *Encoder* (khối bên trái) nhận đầu vào và xây dựng một biểu diễn của nó (các đặc trưng).
- Bộ giải mã - *Decoder* (khối bên phải) sử dụng biểu diễn của bộ mã hóa (các đặc trưng) cùng với các đầu vào khác để tạo ra chuỗi mục tiêu.



Hình 2.6: Kiến trúc tổng quát của mô hình Transformer [2]

Mô hình chỉ có bộ mã hóa (*Encoder-only*) thường được áp dụng trong các tác vụ phân loại câu, nhận dạng thực thể được gán nhãn.

Mô hình chỉ có bộ giải mã (*Decoder-only*) thường được áp dụng trong các tác vụ tạo văn bản.

Mô hình gồm cả bộ mã hóa - bộ giải mã (*Encoder-Decoder*) thường được áp dụng trong các tác vụ xử lý ngôn ngữ tự nhiên đòi hỏi dữ liệu đầu vào dưới dạng văn bản, trong đó có thể kể đến các ứng dụng như dịch máy hoặc tổng hợp tóm tắt.

Điều đặc biệt khiến các mô hình Transformer tốt đó là nhờ cơ chế chú ý - Attention

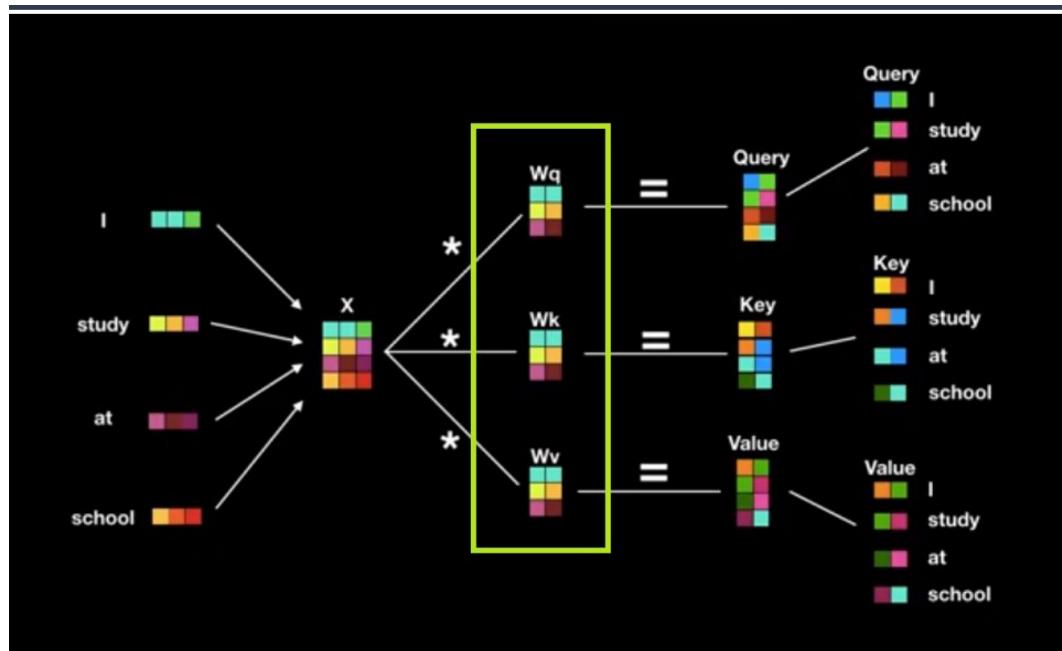
Cơ chế chú ý - Attention là một cơ chế quan trọng trong các mô hình Transformer, cho phép chúng tập trung vào các phần cụ thể của chuỗi đầu vào khi đưa ra dự đoán. Đó là điều khiến Transformer khác biệt với các mạng nơ ron hồi quy truyền thống (*Recurrent Neural Networks*) và cho phép chúng đạt được hiệu suất cao trong các nhiệm vụ khác nhau.

Cơ chế chú ý cho phép mô hình xem xét toàn bộ chuỗi đầu vào đồng thời thay vì xử lý tuần tự và đánh giá mối liên hệ giữa các từ trong một câu. Điều này giúp mô hình nắm bắt được sự phụ thuộc và mối quan hệ tầm xa giữa các từ trong một câu, bất kể khoảng cách của chúng với nhau.

Thành phần chính:

- Truy vấn (Q - Queries) biểu thị những gì mô hình đang tìm kiếm ở từng vị trí trong chuỗi.
- Khóa (K - Keys) biểu thị thông tin từ từng phần tử trong chuỗi.
- Giá trị (V - Values) biểu thị nội dung thực tế của từng phần tử sẽ được sử dụng để tạo đầu ra.

W_q, W_k, W_v là các tham số học



Hình 2.7: Quá trình sinh Q, K, V

Cơ chế:

- Tính điểm chú ý - Attention Score: Mô hình tính toán độ tương tự giữa mỗi truy vấn và tất cả các khóa, thường sử dụng tích vô hướng. Từ đó cho ra điểm attention score trong từng khóa, cho biết mức độ liên quan của khóa đó với truy vấn hiện tại.
- Softmax và Trọng số: Điểm chú ý được chuẩn hóa theo hàm softmax, đảm bảo chúng có tổng bằng 1 và thể hiện phân bố xác suất. Các trọng số này xác định tầm quan trọng của từng giá trị đối với truy vấn cụ thể.

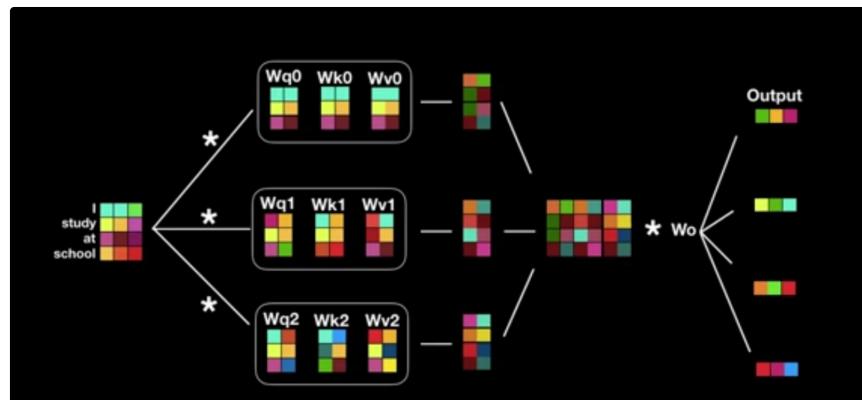
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (2.1)$$

- Tổng trọng số - Weighted Sum: Mô hình tính tổng trọng số của các giá trị dựa trên trọng số tương ứng của chúng, tạo đầu ra chú ý được sử dụng trong các lớp tiếp theo của mô hình.

	$\text{Query} * \text{Key}^T$	Score	Softmax	Value	Softmax * Value	$\sum \text{Softmax} * \text{Value}$ (Attention layer output)
I	I * I * = 130	0.92				
	I * study * = 50	0.05				
	I * at * = 20	0.02				
	I * school * = 10	0.01				
study	study * I * = 30	0.02				
	study * study * = 110	0.70				
	study * at * = 20	0.03				
	study * school * = 70	0.25				
at	at * I * = 30	0.03				
	at * study * = 50	0.10				
	at * at * = 90	0.80				
	at * school * = 40	0.07				
school	school * I * = 30	0.01				
	school * study * = 80	0.27				
	school * at * = 23	0.02				
	school * school * = 160	0.70				

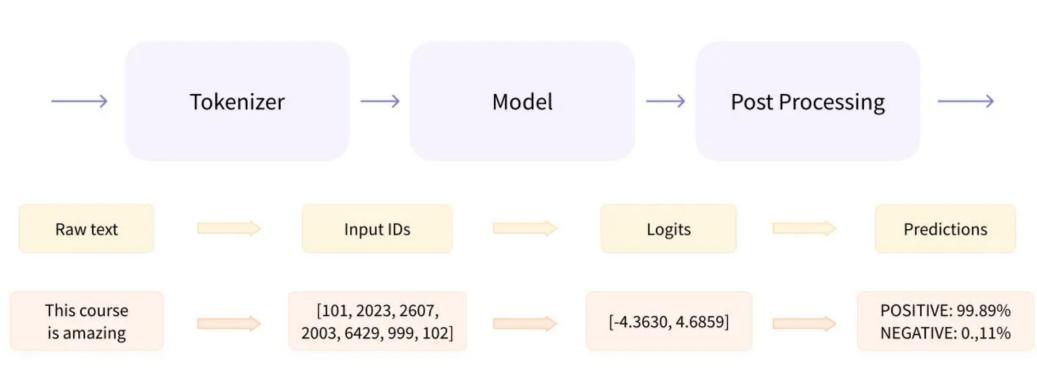
Hình 2.8: Minh họa quá trình tính điểm chú ý cho tất cả các từ trong câu

Lặp lại quá trình trên nhiều lần, với các bộ (W_q, W_k, W_v) khác nhau, ta có cơ chế chú ý đa đầu (*Multi-head attention*).



Hình 2.9: Cấu trúc chú ý đa đầu

Quy trình (Pipeline) của các mô hình Transformer



Hình 2.10: Quy trình của mô hình Transformer [2]

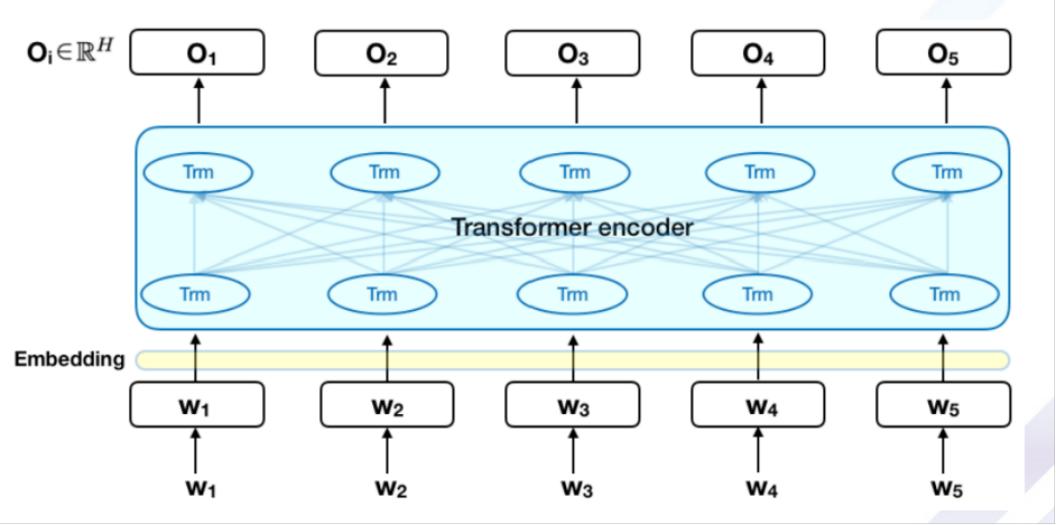
2.2 Mô hình PhoBERT

2.2.1 Mô hình BERT

BERT - Bidirectional Encoder Representations from Transformers [3] là một mô hình mang tính cách mạng trong lĩnh vực xử lý ngôn ngữ tự nhiên do Google AI phát triển vào năm 2018. Nó đã đạt được những kết quả đáng chú ý trong nhiều tác vụ xử lý ngôn ngữ tự nhiên, góp phần vào những tiến bộ vượt bậc và đặt nền móng cho sự phát triển của các mô hình ngôn ngữ lớn.

Đặc điểm chính:

- Huấn luyện hai chiều (*Bidirectional Training*): Không giống như các mô hình trước đây xử lý văn bản tuần tự (từ trái sang phải hoặc từ phải sang trái), BERT được đào tạo hai chiều, có nghĩa là nó xem xét cả ngữ cảnh trước và sau của mỗi từ. Điều này cho phép hiểu sâu hơn về các mối quan hệ từ và cấu trúc câu.
- Kiến trúc dựa trên Transformer: BERT sử dụng kiến trúc Transformer với cơ chế chú ý để đánh giá mối liên hệ giữa các từ trong một câu. Điều này cho phép mô hình nắm bắt các phụ thuộc tầm xa và thông tin theo ngữ cảnh một cách hiệu quả.
- Tiền huấn luyện (*Pre-training*) và tinh chỉnh (*Fine-tuning*): BERT được tiền huấn luyện trên một bộ dữ liệu khổng lồ để nắm bắt về ngữ cảnh ngôn ngữ chung. Mô hình được tiền huấn luyện này sau đó có thể được tinh chỉnh cho các tác vụ cụ thể với các bộ dữ liệu nhiệm vụ cụ thể có kích thước bé hơn.
- Mô hình ngôn ngữ có mặt nạ (*Masked Language Modeling*): Trong quá trình tiền huấn luyện, BERT sử dụng mô hình ngôn ngữ có mặt nạ, trong đó một số từ trong đầu vào được che giấu và mô hình học cách dự đoán các từ được che giấu dựa trên ngữ cảnh của chúng. Điều này giúp mô hình học được các biểu diễn phong phú của các từ và mối quan hệ của chúng.



Hình 2.11: Mô hình BERT [3]

2.2.2 Mô hình RoBERTa

RoBERTa [15], viết tắt của Robustly Optimized BERT Pretraining Approach, là một mô hình học máy dựa trên transformers được phát triển bởi Facebook AI vào năm 2019. Nó được xây dựng dựa trên kiến trúc BERT với một số sửa đổi chính để nâng cao hiệu suất và hiệu quả của nó.

Đặc điểm khác biệt chính so với BERT:

- **Mặt nạ động (Dynamic Masking):** Không giống như mặt nạ tĩnh (Static Masking) của BERT, RoBERTa sử dụng Dynamic Masking nơi các token bị che giấu (masked token) thay đổi trong mỗi epoch huấn luyện. Điều này giúp mô hình làm quen với nhiều ngữ cảnh hơn và cải thiện khả năng khái quát hóa của nó.
- **Kích thước bộ (Batch Size) lớn hơn:** RoBERTa sử dụng batch size lớn hơn đáng kể trong quá trình huấn luyện. Điều này cho phép sử dụng hiệu quả các tài nguyên tính toán hơn và có thể dẫn đến sự hội tụ mô hình tốt hơn.
- RoBERTa loại bỏ Dự đoán câu tiếp theo (Next Sentence Prediction) được sử dụng trong BERT. Phương pháp này nhằm mục đích dự đoán liệu hai câu có liên tiếp hay không, nhưng đã có những nghiên cứu chỉ ra rằng nó không quan trọng đối với hiệu suất và thậm chí có thể gây bất lợi.
- RoBERTa được huấn luyện trong thời gian dài hơn với nhiều dữ liệu hơn

so với BERT. Điều này cho phép mô hình tìm hiểu các mối quan hệ phức tạp hơn trong dữ liệu và cải thiện hơn nữa hiệu suất của nó.

- Text Encoding: RoBERTa sử dụng mã hóa cấp byte (*byte-level encoding*), hiệu quả hơn trong việc xử lý các bộ từ vựng lớn và các từ ít gặp so với mã hóa WordPiece (*WordPiece encoding*) được sử dụng trong BERT.

2.2.3 Mô hình PhoBERT

Sự thành công của BERT và các biến thể của nó phần lớn chỉ giới hạn ở ngôn ngữ tiếng Anh. Hầu hết các mô hình dựa trên BERT được đào tạo trước chỉ được học bằng cách sử dụng ngữ liệu tiếng Anh hoặc dữ liệu kết hợp từ các ngôn ngữ khác nhau (tức là các mô hình đa ngôn ngữ được tiền huấn luyện).

Các mô hình dựa trên BERT đa ngôn ngữ không nhận thức được sự khác biệt giữa các token âm tiết và từ của tiếng Việt, do đó sử dụng văn bản tiếng Việt tiền huấn luyện cấp âm tiết (*syllable-level pre-training Vietnamese texts*). Mặc dù 85% các dạng từ tiếng Việt gồm ít nhất 2 âm tiết (âm/tiếng).

Syllable-level	VinAI công bố các kết quả nghiên cứu khoa học tại hội nghị hàng đầu thế giới về trí tuệ nhân tạo
Word-level	VinAI công_bố các_kết_quả_nghiên_cứu_khoa_học_tại_hội_nghị_hàng_đầu_thế_giới_về_trí_tuệ_nhân_tạo (VinAI publishes research outputs at world-leading conferences in Artificial Intelligence)

Hình 2.12: Sự khác biệt giữa các token âm tiết và từ của tiếng Việt

PhoBERT [4] được huấn luyện trên bộ dữ liệu gồm:

- 20GB văn bản tiếng Việt.
- Thực hiện phân đoạn từ (*word segmentation*) tiếng Việt trước khi tiền huấn luyện. Bộ dữ liệu tiền huấn luyện gồm 145 triệu câu được phân đoạn từ (3 tỷ tokens từ).

Quy trình tiền huấn luyện PhoBERT dựa trên RoBERTa được tối ưu từ BERT cho hiệu suất cao hơn.

Hai phiên bản: PhoBERT-base (150 triệu tham số) và PhoBERT-large (350 triệu tham số).

PhoBERT có thể được sử dụng với các thư viện mã nguồn mở phổ biến: transformers và fairseq.

PhoBERT được đánh giá so với mô hình tiêu chuẩn XLM-R (mô hình pre-trained đa ngôn ngữ tốt nhất gần đây sử dụng 2,5 TB dữ liệu tiền huấn luyện bao gồm 137GB dữ liệu văn bản tiếng Việt cấp âm tiết) dựa trên 4 bài toán trong xử lý ngôn ngữ tự nhiên tiếng Việt gồm:

- Gán một thẻ danh mục từ vựng cho mỗi từ trong văn bản (*Part-of-Speech tagging*, POS).
- Nhận diện thực thể được gán nhãn (*Named entity recognition*, NER).
- Xác định một "giả thuyết" là đúng (entailment - đòi hỏi), sai (contradiction - đối lập), hoặc không xác định (neutral - trung lập) với một "premise - tiền đề" cho trước, là một nhiệm vụ phân loại cặp câu (*Natural language inference*, NLI).
- Kiểm tra sự phụ thuộc giữa các cụm từ trong câu để xác định cấu trúc ngữ pháp của nó (*Dependency Parsing*, DP).

Kết quả:

POS tagging (word-level)		NER (word-level)	
Model	Acc.	Model	F ₁
RDRPOSTagger (Nguyen et al., 2014a) [♣]	95.1	BiLSTM-CNN-CRF [♦]	88.3
BiLSTM-CNN-CRF (Ma and Hovy, 2016) [♣]	95.4	VnCoreNLP-NER (Vu et al., 2018) [♦]	88.6
VnCoreNLP-POS (Nguyen et al., 2017) [♣]	95.9	VNER (Nguyen et al., 2019b)	89.6
jPTDP-v2 (Nguyen and Verspoor, 2018) [★]	95.7	BiLSTM-CNN-CRF + ETNLP [♠]	91.1
jointWPD (Nguyen, 2019) [★]	96.0	VnCoreNLP-NER + ETNLP [♠]	91.3
XLM-R _{base} (our result)	96.2	XLM-R _{base} (our result)	92.0
XLM-R _{large} (our result)	96.3	XLM-R _{large} (our result)	92.8
PhoBERT _{base}	96.7	PhoBERT _{base}	93.6
PhoBERT _{large}	96.8	PhoBERT _{large}	94.7

Hình 2.13: Kết quả cho POS tagging và NER [4]

PhoBERT cho ra kết quả tốt hơn XLM-R trên tất cả 4 bài toán

- PhoBERT sử dụng ít tham số hơn nhiều so với XLM-R: 135 triệu (PhoBERT-base) so với 250 triệu (XLM-R-base); 370 triệu (PhoBERT-large) so với 560 triệu (XLM-R-large).

NLI (syllable- or word-level)		Dependency parsing (word-level)	
Model	Acc.	Model	LAS / UAS
BiLSTM-max (Conneau et al., 2018)	66.4	VnCoreNLP-DEP (Vu et al., 2018) [★]	71.38 / 77.35
mBiLSTM (Artetxe and Schwenk, 2019)	72.0	jPTDP-v2 [★]	73.12 / 79.63
multilingual BERT (Devlin et al., 2019) [■]	69.5	jointWPD [★]	73.90 / 80.12
XLM _{MLM+TLM} (Conneau and Lample, 2019)	76.6	Biaffine (Dozat and Manning, 2017) [★]	74.99 / 81.19
XLM-R _{base} (Conneau et al., 2020)	75.4	Biaffine w/ XLM-R _{base} (our result)	76.46 / 83.10
XLM-R _{large} (Conneau et al., 2020)	79.7	Biaffine w/ XLM-R _{large} (our result)	75.87 / 82.70
PhoBERT _{base}	78.5	Biaffine w/ PhoBERT _{base}	78.77 / 85.22
PhoBERT _{large}	80.0	Biaffine w/ PhoBERT _{large}	<u>77.85 / 84.32</u>

Hình 2.14: Kết quả cho NLI và DP [4]

- XLM-R sử dụng kho dữ liệu tiền huấn luyện đa ngôn ngữ 2,5TB chứa 137GB tiếng Việt văn bản, tức là lớn hơn xấp xỉ 7 lần so với kho dữ liệu tiền huấn luyện đơn ngữ của PhoBERT.

PhoBERT có thể phục vụ như một nền tảng vững chắc cho các nghiên cứu và ứng dụng xử lý ngôn ngữ của Việt Nam.

2.3 Thủ nghiệm

2.3.1 Bộ dữ liệu VLSP 2020 Relation Extraction

Lý do lựa chọn thử nghiệm trên bộ dữ liệu VLSP 2020 Relation Extraction là vì quá trình xây dựng và tiền xử lý dữ liệu y tế đòi hỏi rất nhiều thời gian. Việc thu thập và gán nhãn dữ liệu văn bản y tế cần sự tham gia và đối chiếu chéo một cách kỹ lưỡng. Hơn nữa, dữ liệu y tế thường chứa nhiều thuật ngữ chuyên ngành, viết tắt phức tạp, yêu cầu quá trình tiền xử lý kỹ lưỡng để đảm bảo chất lượng dữ liệu đầu vào cho các mô hình trích rút thông tin và quan hệ. Bộ dữ liệu VLSP 2020 Relation Extraction đã được xử lý sẵn, đảm bảo chất lượng và đáp ứng yêu cầu của nhiệm vụ trích rút quan hệ, giúp chúng tôi có thể tập trung vào phát triển và đánh giá các mô hình nhanh chóng và hiệu quả hơn.

Bộ dữ liệu VLSP 2020 Relation Extraction¹ bao gồm ba tập dữ liệu (huấn luyện chứa 507 tập, phát triển chứa 200 tập và kiểm tra chứa 210 tập) trong định dạng WebAnno TSV phiên bản 3.2. Mỗi tập chỉ chứa một văn bản thô

¹<https://vlsp.org.vn/vlsp2020/eval/re>

(báo điện tử) chưa được tách thành câu, có ba loại thực thể: Địa điểm (LOC), Tổ chức (ORG), Người (PER) và bốn loại quan hệ giữa các thực thể được mô tả trong Bảng 2.1.

Bảng 2.1: Các quan hệ trong bộ dữ liệu VLSP 2020 Relation Extraction

STT.	Quan hệ	Arguments	Directionality
1	LOCATED	PER-LOC, ORG-LOC	Directed
2	PART-WHOLE	LOC-LOC, ORG-ORG, ORG- LOC	Directed
3	PERSONAL-SOCIAL	PER-PER	Undirected
4	ORGANIZATION-AFFILIATION	PER-ORG, PER- LOC, ORG-ORG, LOC-ORG	Directed

- Quan hệ LOCATED bao hàm vị trí vật lý của một người, quan hệ giữa một tổ chức và địa điểm nơi nó đặt trụ sở hoặc kinh doanh.
- Quan hệ PART-WHOLE bao hàm quan hệ địa lý của vị trí của một địa điểm hoặc tổ chức trong hoặc tại hoặc là một phần của địa điểm hoặc tổ chức khác.
- Quan hệ PERSONAL-SOCIAL mô tả mối quan hệ giữa con người với nhau. Cả hai đối số phải là thực thể người. Kiểu quan hệ này là đối xứng. Ví dụ về loại quan hệ này bao gồm: mối quan hệ giữa hai thực thể trong bất kỳ mối quan hệ nghề nghiệp/chính trị/kinh doanh nào, mối quan hệ gia đình/họ hàng, mối quan hệ với cá nhân khác.
- Quan hệ ORGANIZATION-AFFILIATION bao hàm các quan hệ: quan hệ giữa một người và nhân viên của họ (tổ chức), quyền sở hữu giữa một người và một tổ chức thuộc sở hữu của người đó, quan hệ giữa người sáng lập/nhà đầu tư (cá nhân hoặc tổ chức) và một tổ chức, quan hệ giữa một người và cơ sở giáo dục mà người này theo học, quan hệ giữa một người và một tổ chức mà người đó là thành viên, quan hệ giữa một vị trí địa chính trị và một tổ chức mà nó là thành viên, Một người là công dân, cư dân,... của một địa điểm.

#FORMAT=WebAnno TSV 3.2
#T_SP=de.tudarmstadt.ukp.dkpro.core.api.ner.type.NamedEntity identifier value
#T_RL=webanno.custom.VLSPRelationextraction relation BT_de.tudarmstadt.ukp.dkpro.core.api.ner.type.NamedEntity
#Text=Mỹ kêu gọi Nga , Trung dừng phô biến vũ khí hạt nhân Ngoài trưởng Mỹ Rex Tillerson kêu gọi cộng đồng quốc tế ngăn chặn các nước sở hữu vũ khí hạt nhân. Trong lời kêu gọi dừng các hoạt động tạo điều kiện để vũ khí hạt nhân phổ biến, Ngoại trưởng Mỹ Rex Tillerson nêu đích danh Nga , Trung Quốc - ám chỉ các nước này nên già tăng áp lực nhằm buộc Triều Tiên phải từ bỏ tham vọng vũ khí hạt nhân. Ông Tillerson lấy Triều Tiên làm bài học điển hình cho việc thắt bại trong hoạt động ngăn chặn các nước sở hữu vũ khí tiêu diệt hàng loạt. Ngoại trưởng Mỹ đưa ra phát ngôn trên trong cuộc họp cấp Bộ trưởng tại Hội đồng Bảo an Liên Hợp Quốc do Mỹ kêu gọi, tổ chức nhằm giải quyết mâu thuẫn do các hoạt động phát triển các vũ khí sinh học, hóa học, hạt nhân. Ông Tillerson nhắc lại, Washington và Moscow từng hợp tác rất tốt trong thời kỳ Xô-viết về các biện pháp giải trừ vũ khí, dù là đối thủ Chiến tranh Lạnh với nhau. Vậy nên, bây giờ "hai nước nên hợp tác như vậy một lần nữa". Hiện nay, thế giới đang nóng ván đề vũ khí hạt nhân tên lửa khi Triều Tiên không ngừng thực hiện các vụ thử tên lửa, hạt nhân với sức mạnh không ngừng phát triển bất chấp các lệnh trừng phạt từ cộng đồng quốc tế
1-1 0-2 Mỹ * LOCATION - -
1-2 3-6 kêu - - - -
1-3 7-10 gọi - - - -
1-4 11-14 Nga * LOCATION - -
1-5 15-16 , - - - -
1-6 17-22 Trung * LOCATION - - - -
1-7 23-27 dừng - - - -
1-8 28-31 phô - - - -
1-9 32-36 biến - - - -
1-10 37-39 vũ - - - -
1-11 40-43 khí - - - -
1-12 44-47 hạt - - - -
1-13 48-52 nhân - - - -
1-14 53-58 Ngoại - - - -
1-15 59-65 trưởng - - - -
1-16 66-68 Mỹ * LOCATION - AFFILIATION 1-17[1_8]
1-17 69-72 Rex *[1] PERSON[1] - -
1-18 73-82 Tillerson *[1] PERSON[1] - -
1-19 83-86 kêu - - - -
1-20 87-99 gọi - - - -
1-21 91-95 công - - - -
1-22 96-100 đồng - - - -
1-23 101-105 quốc - - - -

Hình 2.15: Tệp dưới định dạng WebAnno TSV phiên bản 3.2

Định dạng tệp tin TSV 3.2 có cấu trúc sau:

Cột 1: Vị trí bắt đầu của token trong văn bản.

Cột 2: Vị trí kết thúc của token trong văn bản.

Cột 3: Nội dung văn bản.

Cột 4-6: Nhãn cho các thực thể.

Cột 6-9 : Quan hệ giữa các thực thể được xác định.

2.3.2 Bộ công cụ xử lý tiếng Việt Underthesea

Trong bối cảnh ngày càng nhiều ứng dụng xử lý ngôn ngữ tự nhiên (*Natural Language Processing*, NLP) cho tiếng Việt ra đời, "Underthesea"² đã trở thành một công cụ không thể thiếu. Đây là một bộ thư viện và công cụ NLP mã nguồn mở đầu tiên dành riêng cho tiếng Việt, được phát triển bởi Trung tâm Công nghệ Phần mềm (ITS) tại Đại học Công nghệ Quốc gia Hà Nội.

Underthesea tích hợp nhiều thư viện và mô hình máy học cho các tác vụ NLP quan trọng như:

- Gán một thẻ danh mục từ vựng cho mỗi từ trong văn bản (*Part-of-Speech Tagging*)

²<https://undertheseanlp.com>

- Tách từ (*Word Segmentation*)
- Nhận dạng thực thể có tên và trích rút quan hệ (*Named Entity Recognition and Relation Extraction*)
- Phân loại văn bản (*Text Classification*)
- Tóm tắt tự động (*Automatic Summarization*)

Một trong những ưu điểm lớn của Underthesea là mã nguồn mở, cho phép các nhà nghiên cứu và nhà phát triển:

- Mở rộng và cải tiến các mô hình hiện có.
- Huấn luyện mô hình mới trên tập dữ liệu của riêng họ.
- Tùy chỉnh và tích hợp vào các ứng dụng xử lý ngôn ngữ tự nhiên cho tiếng Việt.

Bộ công cụ Underthesea đóng vai trò quan trọng trong việc thúc đẩy các nghiên cứu và ứng dụng xử lý ngôn ngữ tự nhiên cho tiếng Việt. Nó cung cấp một nền tảng vững chắc để phát triển các giải pháp xử lý ngôn ngữ tự nhiên tiếng Việt chất lượng cao, đáp ứng nhu cầu ngày càng tăng của người dùng trong lĩnh vực này.

2.3.3 Tiền xử lý dữ liệu

Trước tiên, ta chia văn bản thô thành các câu sử dụng thư viện Underthesea vì tập dữ liệu chỉ chứa các mối quan hệ giữa các thực thể được gắn nhãn thuộc cùng một câu.

Giả sử rằng có tổng n thực thể trong một câu, ta tạo $\frac{n(n-1)}{2}$ câu tương ứng với $\frac{n(n-1)}{2}$ cặp thực thể. Từng câu là một điểm dữ liệu được truyền sang mô hình PhoBERT. Nhãn cho mỗi điểm dữ liệu là nhãn quan hệ giữa cặp thực thể trong câu.

Bộ dữ liệu gồm bốn loại quan hệ. Ba trong số đó là directed, vì vậy ta tạo hai quan hệ undirected cho từng quan hệ directed, tùy thuộc vào việc nhãn quan hệ directed là thực thể trước hoặc sau trong câu. Hai ví dụ dưới đây sẽ minh họa rõ ràng hơn cho vấn đề này.

Ví dụ 1: Trong câu: "Hà Nội là thủ đô của Việt Nam", mối quan hệ giữa hai thực thể ("Hà Nội" và "Việt Nam") là PART-WHOLE. Nhãn quan hệ này nằm ở thực thể "Việt Nam", là thực thể sau trong câu. Ta gán nhãn cho điểm dữ liệu này thành PART-WHOLE.

Ví dụ 2: Trong câu: "Việt Nam có thủ đô là Hà Nội", mối quan hệ giữa hai thực thể ("Hà Nội" và "Việt Nam") là PART-WHOLE. Nhãn quan hệ này nằm ở thực thể "Việt Nam", là thực thể xuất hiện đầu tiên trong câu. Ta gán nhãn cho điểm dữ liệu này thành WHOLE-PART.

Có nhiều thực thể trong cùng một câu nhưng không có quan hệ giữa chúng, vì vậy ta tạo ra một loại quan hệ mới gọi là "OTHERS" cho chúng.

Cuối cùng, ta chuyển các điểm dữ liệu này vào mô hình PhoBERT.

2.3.4 Tinh chỉnh mô hình PhoBERT

Ta tiến hành tinh chỉnh phiên bản phobert-base của mô hình PhoBERT [16] [17] với hàm tổn thất cross entropy và áp dụng kỹ thuật trung bình pooling (average pooling) để xử lý các đặc trưng đầu ra của mô hình. Các lớp embedding từ 10 đến 13 được trích xuất và kết hợp bằng phương pháp sum để tạo ra đặc trưng cuối cho từng từ. Đối với các cặp thực thể (ent1, ent2) trong cấu trúc dữ liệu, ta sử dụng các phép toán như mul (nhân), add (cộng) và abs sub (trị tuyệt đối của hiệu) để kết hợp các đặc trưng của chúng.

Trong quá trình tinh chỉnh, mô hình phobert-base được huấn luyện trong 3 epoch với tốc độ học (phobert lr) là 0.001 và không áp dụng phương pháp phân rã trọng số (phobert weight decay).

Mô hình cuối cùng bao gồm một lớp dropout với tỷ lệ 0.6 (dropout1 rate), một lớp tuyến tính với 1024 đơn vị (out linear1), và một lớp dropout khác với tỷ lệ 0.2 (dropout2 rate). Mô hình được huấn luyện với tổng số 100 epoch (total epochs), kích thước batch là 32 (batch size), và sử dụng bộ tối ưu hóa AdamW với các thông số như tốc độ học (linear lr) là 0.00001, hệ số phân rã trọng lượng (linear weight decay) là 0.15 và giá trị chuẩn hóa gradient (clip grad norm rate) là 5.0.

Quá trình huấn luyện được thực hiện trên GPU Tesla V4 trên Google Colab. Các kỹ thuật khác như khởi tạo hạt giống (seed) và ghi lại kết quả sau mỗi 30

batch (log batch) cũng được áp dụng để đảm bảo tính ổn định và khả năng theo dõi quá trình huấn luyện.

Nhìn chung, việc tinh chỉnh mô hình PhoBERT với các thông số và kỹ thuật được mô tả trên nhằm mục đích tối ưu hóa hiệu suất của mô hình cho tác vụ trích xuất quan hệ. Các kết quả thu được sẽ được đánh giá và phân tích để đưa ra những kết luận và hướng nghiên cứu tiếp theo.

2.3.5 Phương pháp đánh giá

Đối với bài toán RE, các định nghĩa và công thức của các độ đo như sau

- **Độ chính xác (Precision)**

- **Định nghĩa:** Độ chính xác là tỷ lệ giữa số lượng dự đoán đúng cho một lớp cụ thể so với tổng số dự đoán mà mô hình đã đưa ra cho lớp đó.
- **Công thức:**

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Ý nghĩa:** Precision đo lường mức độ chính xác của các dự đoán dương tính thực sự của mô hình. Độ chính xác cao nghĩa là ít có các mẫu dương tính giả (False Positives).

- **Độ phủ (Recall)**

- **Định nghĩa:** Độ phủ là tỷ lệ giữa số lượng dự đoán đúng cho một lớp cụ thể so với tổng số mẫu thực sự thuộc về lớp đó.
- **Công thức:**

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

- **Ý nghĩa:** Recall đo lường khả năng của mô hình trong việc tìm ra tất cả các mẫu dương tính thực sự. Độ phủ cao nghĩa là ít có các mẫu âm tính giả (False Negatives).

- **Độ đo F1**

- **Định nghĩa:** điểm F1 (*F1-Score*) là trung bình của độ chính xác và độ phủ. Nó cung cấp một thước đo cân bằng giữa hai yếu tố này.

- **Công thức:**

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

- **Ý nghĩa:** F1 hữu ích trong các tình huống mà cần cân bằng giữa precision và recall, và đặc biệt hữu ích khi có sự mất cân bằng giữa các lớp trong tập dữ liệu.

Trong báo cáo này, hiệu quả của mô hình được đánh giá thông qua hai độ đo là điểm số F1 trung bình vĩ mô (*macro-averaged F-score*) và điểm số F1 trung bình vi mô (*micro-averaged F-score*). Việc sử dụng cả hai biến thể F1 vi mô và F1 vĩ mô cho phép ta phân tích hiệu suất của mô hình từ các góc nhìn khác nhau, đảm bảo rằng đánh giá được thực hiện một cách toàn diện.

Điểm số F1 trung bình vi mô tính toán dựa trên tổng số dự đoán đúng và tổng số dự đoán trên toàn bộ tập dữ liệu, không phân biệt nhãn.

Mặt khác, điểm số F1 trung bình vĩ mô tính toán giá trị điểm số F1 cho từng nhãn quan hệ, sau đó lấy trung bình cộng của các giá trị đó. Phương pháp này đảm bảo rằng mỗi nhãn đóng góp đồng đều vào kết quả cuối cùng, bất kể số lượng mẫu của chúng trong tập dữ liệu. Điều này đặc biệt quan trọng trong trường hợp tập dữ liệu có sự mất cân bằng về phân bố giữa các nhãn, giúp tránh việc các nhãn ít phổ biến bị ảnh hưởng bởi các nhãn chiếm đa số.

2.3.6 Kết quả thực nghiệm

Bảng 2.2: Hiệu suất của mô hình trên các tập dữ liệu huấn luyện và kiểm thử của bộ dữ liệu VLSP 2020 Relation Extraction

Tập dữ liệu	Điểm số F1 trung bình vĩ mô	Điểm số F1 trung bình vi mô
Huấn luyện	0.99	0.99
Kiểm thử	0.59	0.93

Kết quả đánh giá hiệu suất của mô hình trích xuất quan hệ trên bộ dữ liệu VLSP 2020 Relation Extraction cho thấy sự tồn tại của hiện tượng quá khớp

(*overfitting*). Mặc dù đạt được điểm số F1 trung bình vĩ mô và vi mô gần như hoàn hảo trên tập huấn luyện (lần lượt là 0.99 và 0.99), mô hình chỉ đạt được điểm số 0.59 và 0.93 trên tập kiểm thử tương ứng.

Sự sụt giảm đáng kể về điểm số F1 trung bình vĩ mô, một chỉ số nhạy cảm với hiệu suất trên các lớp thiểu số, cho thấy mô hình gặp khó khăn trong việc khai quát hóa các đặc trưng học được từ tập huấn luyện sang dữ liệu chưa nhìn thấy trước đó. Ngược lại, điểm số F1 trung bình vi mô cao trên tập kiểm thử cho thấy mô hình vẫn đạt được độ chính xác tổng thể cao, nhưng hiệu suất có thể không đồng đều giữa các lớp.

Bảng 2.3: Hiệu suất của mô hình trên các nhãn trong tập dữ liệu kiểm thử của bộ dữ liệu VLSP 2020 Relation Extraction

Nhãn quan hệ	Độ chính xác	Độ phủ	Độ đo F1	Số lượng nhãn phân loại
LOCATED	0.77	0.69	0.73	320
PART WHOLE	0.87	0.89	0.87	468
PERSONAL SOCIAL	0.89	0.26	0.40	98
AFFILIATION	0.83	0.66	0.73	328
IS LOCATED	0.20	0.04	0.07	23
WHOLE PART	0.50	0.07	0.12	44
AFFILIATION TO	0.87	0.58	0.69	189
OTHERS	0.94	0.98	0.96	7765

Hiện tượng này có thể liên quan đến sự phân bố không đồng đều giữa số lượng các nhãn trong tập dữ liệu kiểm thử. Mô hình thể hiện hiệu quả tốt trong việc phân loại các nhãn chiếm đa số, tuy nhiên, các chỉ số lại thấp hơn đáng kể đối với các nhãn thiểu số như IS-LOCATED và WHOLE-PART.

Chương 3

Xây dựng dữ liệu trích rút quan hệ thực thể y tế Tiếng Việt

3.1 Xây dựng dữ liệu

3.1.1 Hệ thống phân loại bệnh quốc tế ICD-9 và ICD-10

Hệ thống phân loại bệnh quốc tế, phiên bản 9 (International Classification of Diseases, Ninth Revision, ICD-9) là hệ thống phân loại bệnh tật quốc tế phiên bản thứ 9 được Tổ chức Y tế Thế giới (WHO) phát triển và chính thức công bố vào năm 1977. Trong suốt nhiều thập kỷ, ICD-9¹ đã đóng vai trò quan trọng trong việc chuẩn hóa quá trình ghi chép và báo cáo dữ liệu y tế trên toàn cầu.

Cấu trúc của ICD-9 bao gồm 17 chương chính, tập trung vào các nhóm bệnh lý hoặc hệ cơ quan cụ thể trong cơ thể con người. Mỗi bệnh hoặc tình trạng sức khỏe sẽ được gán một mã số duy nhất, giúp quá trình phân loại và theo dõi trở nên đơn giản và nhất quán hơn.

Mặc dù đã mang lại nhiều lợi ích trong việc chuẩn hóa dữ liệu y tế, ICD-9 cũng gặp phải một số hạn chế nhất định. Số lượng mã bệnh trong ICD-9 là có giới hạn, đồng thời mô tả cho một số lĩnh vực còn thiếu chi tiết và cập nhật.

Nhằm khắc phục những hạn chế của ICD-9, WHO đã phát triển và chính thức công bố hệ thống phân loại bệnh quốc tế, phiên bản 10 (International

¹https://en.wikipedia.org/wiki/List_of_ICD-9_codes

Classification of Diseases, Tenth Revision, ICD-10) vào năm 1992. ICD-10² bắt đầu được áp dụng rộng rãi trong các hệ thống chăm sóc sức khỏe trên toàn cầu từ đầu những năm 2000.

Cấu trúc của ICD-10 gồm 22 chương chính, bao gồm hơn 14.000 mã bệnh khác nhau. So với phiên bản trước, ICD-10 cung cấp mô tả chi tiết và cụ thể hơn cho các bệnh lý, giúp quá trình chẩn đoán và ghi chép dữ liệu trở nên chính xác hơn.

Một đổi mới quan trọng trong ICD-10 là hệ thống sử dụng mã alphanumeric (kết hợp chữ cái và số) dài hơn. Điều này cho phép thêm nhiều mã mới khi cần thiết, đáp ứng nhu cầu phân loại các bệnh lý mới xuất hiện hoặc các trường hợp phức tạp hơn.

Với khả năng cung cấp thông tin chi tiết và có thể mở rộng, ICD-10 đã trở thành phiên bản tiêu chuẩn được áp dụng rộng rãi trong các hệ thống chăm sóc sức khỏe trên toàn thế giới. ICD-10 giúp ghi chép, theo dõi và báo cáo các bệnh tật cũng như nguyên nhân tử vong một cách chính xác và cập nhật hơn so với các phiên bản trước đó.

3.1.2 Cơ sở dữ liệu bách khoa toàn thư mở trực tuyến - Wikipedia

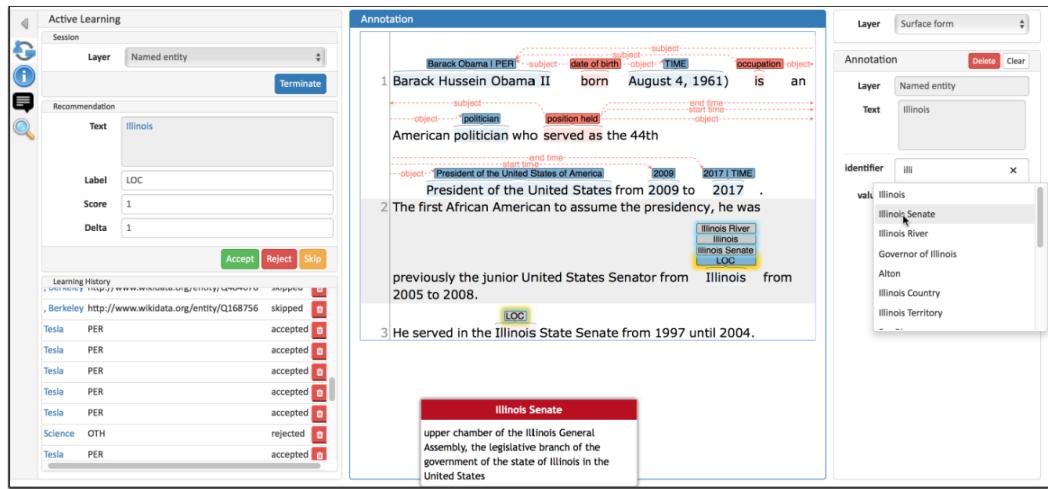
Wikipedia, nguồn tài nguyên tri thức mở không lồ, cung cấp cho người dùng khả năng tải dữ liệu xuống cơ sở dữ liệu của mình cho các mục đích khác nhau. Trang "Wikipedia: Database download"³ là nơi cung cấp thông tin và hướng dẫn để tải dữ liệu từ cơ sở dữ liệu của Wikipedia. Tại đây, người dùng có thể tìm hiểu về các bản tải xuống khác nhau, bao gồm bản đầy đủ chứa toàn bộ nội dung của Wikipedia hoặc các bản thu gọn chỉ chứa một phần nội dung. Các bản tải xuống này được cung cấp dưới dạng tệp nén hoặc bản ghi SQL có thể nhập vào hệ quản trị cơ sở dữ liệu. Ngoài ra, trang web cũng cung cấp hướng dẫn chi tiết về cách tải xuống và sử dụng dữ liệu Wikipedia cho các mục đích như nghiên cứu, phân tích hoặc tạo ứng dụng mới.

²<https://icd.kcb.vn/#/icd-10/icd10>

³https://en.wikipedia.org/wiki/Wikipedia:Database_download

3.1.3 Công cụ gán nhãn INCEpTION

INCEpTION [5] là một nền tảng phần mềm mã nguồn mở chuyên dụng cho việc gán nhãn và chú thích dữ liệu ngôn ngữ tự nhiên. Nó được phát triển bởi nhóm nghiên cứu Ubiquitous Knowledge Processing Lab tại Đại học Kỹ thuật Darmstadt, Đức. INCEpTION cung cấp một môi trường trực quan và linh hoạt để các nhà nghiên cứu, nhà phát triển có thể dễ dàng gán nhãn và chú giải dữ liệu văn bản, âm thanh hoặc hình ảnh cho các tác vụ như gán nhãn thực thể, phân loại quan hệ, phân đoạn văn bản và nhiều tác vụ thú vị khác.



Hình 3.1: Giao diện công cụ gán nhãn INCEpTION [5])

3.1.4 Thu thập và tiền xử lý dữ liệu

Bước đầu tiên là tải về bản dump ngôn ngữ tiếng Việt từ Wikipedia, chứa toàn bộ nội dung của Wikipedia dưới dạng tệp nén. Tiếp theo, ta sử dụng thư viện JWPL⁴ (Java Wikipedia Library) để trích xuất và xây dựng cơ sở dữ liệu từ bản dump này.

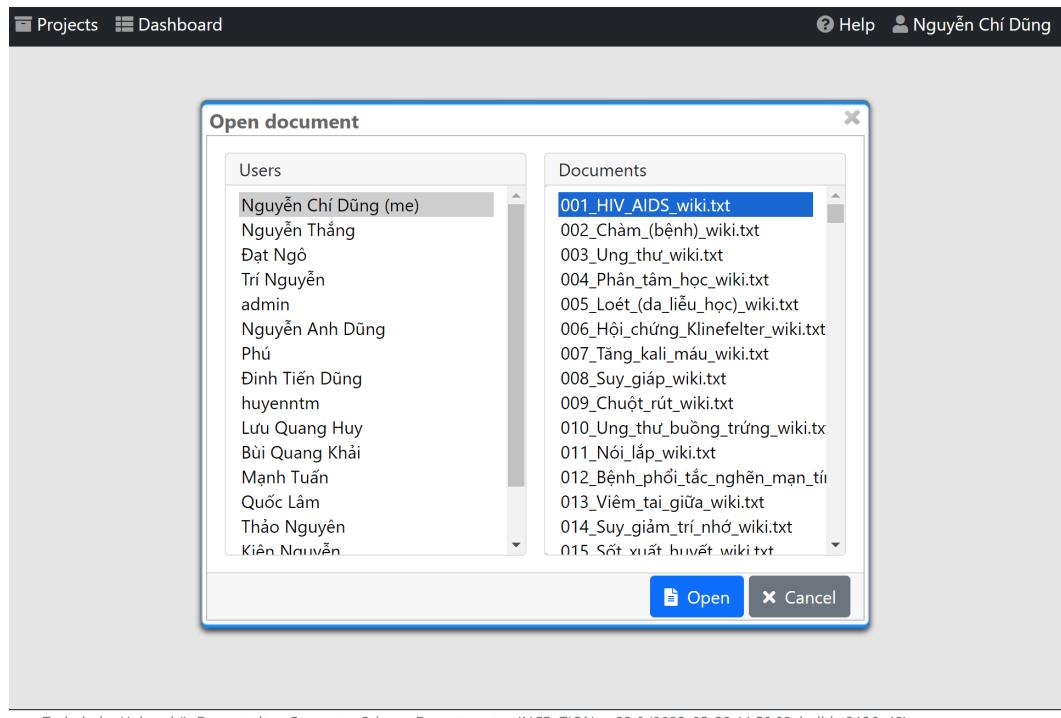
Nhằm thu hẹp phạm vi nghiên cứu vào lĩnh vực y tế, ta thực hiện truy vấn trên cơ sở dữ liệu để trích xuất các văn bản có chứa các mã bệnh thuộc hệ thống phân loại bệnh quốc tế ICD-9 hoặc ICD-10. Các văn bản này được xem là có liên quan đến chủ đề bệnh tật và y khoa.

Tiếp theo, ta tiến hành tiền xử lý dữ liệu văn bản thu được. Đầu tiên, các thẻ

⁴<https://github.com/dkpro/dkpro-jwpl>

HTML và định dạng được loại bỏ để chỉ giữ lại nội dung thuần văn bản. Sau đó, ta sử dụng thư viện Underthesea để thực hiện các tác vụ xử lý ngôn ngữ tự nhiên như tách từ cho tiếng Việt. Cuối cùng, các văn bản được lưu thành các tệp tin riêng biệt, với tên file là số thứ tự tăng dần và tiêu đề là tên bệnh được trích xuất từ văn bản.

Tập dữ liệu cuối cùng thu được gồm có 418 văn bản liên quan đến lĩnh vực y tế, được tiền xử lý và đặt tên một cách có hệ thống. Tập dữ liệu này sẽ được sử dụng làm đầu vào cho các công đoạn tiếp theo trong nghiên cứu.



Technische Universität Darmstadt -- Computer Science Department -- INCEpTION -- 23.6 (2022-05-29 11:50:02, build e3136a42)

Hình 3.2: Danh sách các văn bản được gán nhãn trên INCEpTION)

3.1.5 Định nghĩa nhãn các thực thể ánh xạ đến UMLS

Bộ nhãn thực thể bao gồm 5 nhãn

- SYMPTOM AND DISEASE
- THERAPEUTIC OR PREVENTIVE PROCEDURE
- TESTS
- MEDICINE

- BODY LOCATION OR REGION

SYMPTOM AND DISEASE chứa các quan sát của bệnh nhân hoặc bác sĩ lâm sàng về cơ thể hoặc tâm trí của bệnh nhân được cho là bất thường hoặc do bệnh gây ra. Chúng dựa trên các loại ngữ nghĩa UMLS của chức năng bệnh lý, bệnh tật hoặc hội chứng, rối loạn chức năng tâm thần hoặc hành vi, rối loạn chức năng tế bào hoặc phân tử, bất thường bẩm sinh, bất thường mắc phải, chấn thương hoặc ngộ độc, bất thường giải phẫu, quá trình ung thư, vi rút, vi khuẩn, dấu hiệu hoặc triệu chứng, nhưng không bị giới hạn bởi phạm vi phủ sóng của UMLS.

SYMPTOM AND DISEASE chứa sáu thực thể được định nghĩa bởi UMLS gồm Disease or Syndrome(Bệnh hay Hội chứng), Injury or Poisoning (Chấn thương hoặc Ngộ độc), Sign or Symptom (Dấu hiệu hoặc Triệu chứng), Virus (Vi rút), Fungus (Nấm), Bacterium (Vi khuẩn).

Dưới đây là một số ví dụ về các trường hợp cần được gán nhãn và trường hợp không cần gán nhãn, chi tiết xem tại phần phụ lục.

Bảng 3.1: Các ví dụ về SYMPTOM AND DISEASE cần được gán nhãn

Trường hợp	Ví dụ
Các cụm danh từ, tính từ mô tả bệnh, hội chứng, dấu hiệu, triệu chứng	<ul style="list-style-type: none"> Bệnh nhân bị chẩn đoán mắc hội chứng vàng da do suy gan.
Những nhận định về tình trạng tinh thần, hành vi	<ul style="list-style-type: none"> Bệnh nhân trầm cảm nặng có biểu hiện các triệu chứng mất hứng thú với mọi thứ, cảm giác tuyệt vọng và vô dụng.
Các loại vi rút, vi khuẩn gây bệnh	<ul style="list-style-type: none"> Bệnh nhân đang trong giai đoạn cấp của bệnh viêm gan do vi rút viêm gan B.

Chấn thương, tổn thương	<ul style="list-style-type: none"> Bệnh nhân bị đa chấn thương sau tai nạn giao thông, đặc biệt là chấn thương sọ não và gãy xương đùi phải.
Các bất thường về cấu trúc, chức năng cơ thể	<ul style="list-style-type: none"> Bệnh nhân bị teo cơ nặng ở chi dưới, dẫn đến liệt hoàn toàn 2 chi, mất khả năng đi lại và sinh hoạt.
Kết quả xét nghiệm được mô tả rõ ràng là bất thường	<ul style="list-style-type: none"> Xét nghiệm máu cho thấy bệnh nhân bị thiếu máu nặng với hồng cầu: 2.5 triệu/mm3, hemoglobin: 8 g/dL.

Bảng 3.2: Các ví dụ về SYMPTOM AND DISEASE không được gán nhãn

Trường hợp	Ví dụ
Các mô tả về trạng thái sức khỏe bình thường	<ul style="list-style-type: none"> Khám sức khỏe tổng quát cho kết quả bình thường, nhịp tim bình thường.
Các giá trị đo, kết quả xét nghiệm nếu không mô tả rõ bất thường	<ul style="list-style-type: none"> Đường huyết lúc đói của bệnh nhân là 5.5 mmol/L.
Các động từ mô tả kết quả/tiến triển bệnh	<ul style="list-style-type: none"> Cơn đau của bệnh nhân có giảm dần sau khi dùng thuốc.
Sử dụng rượu bia, thuốc lá	<ul style="list-style-type: none"> Bệnh nhân có tiền sử nghiện rượu nặng.

THERAPEUTIC OR PREVENTIVE PROCEDURE là thuật ngữ chỉ về các phương thức, biện pháp và thủ tục được thiết kế để điều trị một rối loạn sức khỏe hoặc ngăn ngừa các rối loạn sức khỏe tiềm ẩn. Những biện pháp này bao gồm các liệu pháp y tế và nha khoa, phẫu thuật, cũng như các phương pháp y tế, vật lý trị liệu và can thiệp tâm lý xã hội.

Bảng 3.3: Các ví dụ về THERAPEUTIC OR PREVENTIVE PROCEDURE cần được gán nhãn

Trường hợp	Ví dụ
Các thủ thuật, phẫu thuật, kỹ thuật y khoa được áp dụng cho bệnh nhân	<ul style="list-style-type: none"> Bệnh nhân được phẫu thuật cắt bỏ khối u ác tính ở đại tràng để ngăn chặn sự phát triển của các khối u di căn.
Các biện pháp can thiệp trị liệu	<ul style="list-style-type: none"> Bệnh nhân ung thư vú được chỉ định trị liệu bằng phương pháp hóa trị sử dụng thuốc 5-Fluorouracil nhằm tiêu diệt tế bào ung thư.
Các thủ tục, kỹ thuật hỗ trợ điều trị	<ul style="list-style-type: none"> Bệnh nhân bị sốc nhiễm khuẩn được chỉ định truyền dịch để bù lượng máu lưu thông và cân bằng điện giải trong cơ thể.

Bảng 3.4: Các ví dụ về THERAPEUTIC OR PREVENTIVE PROCEDURE
không được gán nhãn

Trường hợp	Ví dụ
Các dụng cụ, thiết bị y tế	<ul style="list-style-type: none"> Trước khi bắt đầu ca phẫu thuật, y tá sử dụng máy đo huyết áp để đo áp lực máu của bệnh nhân và đảm bảo rằng nó ổn định.
Các bộ phận, khoa, phòng y tế	<ul style="list-style-type: none"> Bệnh nhân đái tháo đường được tư vấn và theo dõi điều trị tại khoa Nội Tiết của bệnh viện.
Địa điểm điều trị	<ul style="list-style-type: none"> Người bệnh được nhân viên y tế tại trạm xá thôn bản khám, kê đơn và hướng dẫn sử dụng thuốc.

TESTS chứa các xét nghiệm, thủ thuật thăm dò được thực hiện nhằm mục đích chẩn đoán, đánh giá một tình trạng bệnh lý, xác định nguyên nhân gây bệnh để có biện pháp điều trị hợp lý. TESTS được thực hiện bằng việc lấy và phân tích mẫu sinh học như máu, nước tiểu, xét nghiệm trên mô bệnh phẩm, chụp cắt lớp, chụp cộng hưởng từ, siêu âm, nội soi....

TESTS chứa hai thực thể được định nghĩa bởi UMLS gồm Laboratory Procedure (Thủ thuật cận lâm sàng), Diagnostic Procedure (Thủ thuật chẩn đoán)

Bảng 3.5: Các ví dụ về TESTS cần được gán nhãn

Trường hợp	Ví dụ
Các xét nghiệm, thủ thuật y học nhằm mục đích chẩn đoán, theo dõi, đánh giá tình trạng bệnh	<ul style="list-style-type: none"> Bệnh nhân cao huyết áp được lấy máu xét nghiệm định lượng lipid máu nhằm đánh giá nguy cơ mắc bệnh tim mạch.
Các xét nghiệm cận lâm sàng trên mẫu máu, dịch cơ thể, mô	<ul style="list-style-type: none"> Người bệnh có biểu hiện sốt, nhức đầu, nôn ói, nghi ngờ viêm màng não do vi khuẩn. Bác sĩ lấy mẫu dịch não tủy để thực hiện xét nghiệm nuôi cấy phân lập vi khuẩn gây bệnh và kháng sinh đồ nhằm xác định tác nhân và phác đồ kháng sinh phù hợp.
Các kỹ thuật hình ảnh như chụp X-quang, cộng hưởng từ, siêu âm, nội soi	<ul style="list-style-type: none"> Bệnh nhân bị đau bụng được chỉ định chụp cắt lớp để phát hiện các khối u, tổn thương ở các cơ quan nội tạng trong ổ bụng.
Các thăm dò chức năng như điện tim, đo thính lực, thị lực	<ul style="list-style-type: none"> Bệnh nhân có các triệu chứng đau ngực, khó thở, nghi ngờ bị rối loạn nhịp tim. Bác sĩ cho bệnh nhân đo điện tâm đồ, ghi nhận đường điện tim trong 24 giờ để phát hiện các dấu hiệu loạn nhịp thất và đánh giá mức độ nguy hiểm.

Bảng 3.6: Các ví dụ về TESTS không được gán nhãn

Trường hợp	Ví dụ
Các thiết bị, dụng cụ y tế	<ul style="list-style-type: none"> Bác sĩ khoa chẩn đoán hình ảnh quan sát dưới kính hiển vi quang học các mẫu bệnh phẩm để phát hiện bất thường.
Kết quả đo lường	<ul style="list-style-type: none"> Điều dưỡng dùng máy đo huyết áp và nghe nhịp tim để theo dõi các chỉ số sinh tồn của bệnh nhân sau phẫu thuật tim là: nhip tim 92 lần/phút, huyết áp 140/90 mmHg.
Các khoa/phòng xét nghiệm và hình ảnh	<ul style="list-style-type: none"> Các kỹ thuật viên xét nghiệm tại phòng xét nghiệm thực hiện xét nghiệm sinh hóa, huyết học, miễn dịch trên các mẫu bệnh phẩm để phục vụ công tác chẩn đoán.

MEDICINE chứa cụm từ mô tả các hợp chất/chất hóa học, có nguồn gốc tự nhiên hoặc nhân tạo, có khả năng tương tác với sinh lý bình thường của tế bào hoặc cơ thể để tạo ra các phản ứng sinh học, được lực học mong muốn hoặc không mong muốn. Bao gồm thuốc/dược phẩm, chất độc, thực phẩm chức năng, vắc-xin, hormon, chất ức chế, kháng sinh...

MEDICINE chứa ba thực thể được định nghĩa bởi UMLS gồm Antibiotic (Kháng sinh), Clinical Drug (Thuốc lâm sàng), Pharmacologic Substance (Dược chất), Vitamin.

Bảng 3.7: Các ví dụ về MEDICINE cần được gán nhãn

Trường hợp	Ví dụ
Các loại thuốc được kê đơn cho bệnh nhân: tên thương hiệu, tên chung, tên nhóm thuốc	<ul style="list-style-type: none"> Bệnh nhân được kê đơn thuốc giảm đau paracetamol 500mg với liều dùng 1 viên khi cần, không quá 4 viên/ngày.
Các chế phẩm sinh học, huyết thanh, kháng thể được dùng với mục đích điều trị	<ul style="list-style-type: none"> Liệu pháp sinh học với các kháng thể đơn dòng được sử dụng để điều trị bệnh viêm khớp dạng thấp.
Thuật ngữ chung chỉ đơn thuốc/liệu trình điều trị của bệnh nhân	<ul style="list-style-type: none"> Bệnh nhân được kê đơn Ibuprofen (Thuốc chống viêm) để giảm viêm nhiễm và giảm đau.

Bảng 3.8: Các ví dụ về MEDICINE không được gán nhãn

Trường hợp	Ví dụ
Tên công ty sản xuất thuốc	<ul style="list-style-type: none"> "Paracetamol" được sản xuất bởi công ty XYZ.
Các dạng bào chế thuốc: viên nén, siro, thuốc tiêm	<ul style="list-style-type: none"> "Aspirin" có dạng viên nén hoặc dạng nước.
Quy trình, cách thức dùng thuốc	<ul style="list-style-type: none"> Uống sau khi ăn để giảm nguy cơ kích thích dạ dày.
Các thiết bị, dụng cụ y tế liên quan thuốc	<ul style="list-style-type: none"> Bình xịt phun thuốc để đưa thuốc vào họng.

BODY LOCATION OR REGION là các danh mục giải phẫu mô tả về cơ quan, vị trí cụ thể hoặc tập hợp (vùng/hệ) các mô trong cơ thể người. Các vị trí hay vùng cơ thể này có thể được xác định giải phẫu hoặc ý nghĩa chức năng trong cơ thể như:

- Các cơ quan rời rạc: não, phổi, gan, thận...
- Các vùng/khoang cơ thể: đầu, ngực, ổ bụng, tứ chi...
- Các mô mềm: cơ, xương, dây chằng...
- Các hệ cơ quan chức năng: hệ tuần hoàn, hệ bạch huyết, hệ miễn dịch...

Bảng 3.9: Các ví dụ về BODY LOCATION OR REGION cần được gán nhãn

Trường hợp	Ví dụ
Mô tả Vị Trí Cụ Thể trên Cơ Thể	<ul style="list-style-type: none"> Vết thương nằm ở phần trên của cánh tay trái đang gây đau đớn cho bệnh nhân.
Thực Thể Y Tế Liên Quan Đến Một Khu Vực Cụ Thể	<ul style="list-style-type: none"> Cơ sở y tế này chuyên về điều trị các vấn đề về hệ tiêu hóa, bao gồm bệnh dạ dày và ruột kết.

3.1.6 Định nghĩa nhãn quan hệ giữa các thực thể

Bộ nhãn quan hệ giữa các thực thể bao gồm 6 nhãn quan hệ được định nghĩa trong mạng ngữ nghĩa của UMLS.

- TREATS - Áp dụng một biện pháp khắc phục với mục đích chữa bệnh hoặc kiểm soát một tình trạng.
- PREVENTS - Ngăn ngừa, cản trở hoặc loại bỏ một tác nhân hoặc tình trạng.
- CAUSES - Mang lại một tình trạng hoặc một hiệu ứng. Ở đây ngụ ý rằng một tác nhân, chẳng hạn như một dược chất hoặc một sinh vật đã gây ra tác dụng đó. Điều này bao gồm phát sinh, tác động, gây ra và nguồn gốc.
- DIAGNOSES - Phân biệt hoặc xác định bản chất hoặc đặc điểm của đối tượng
- USES - Sử dụng trong việc thực hiện một số hoạt động. Điều này bao gồm áp dụng, sử dụng và tận dụng.
- LOCATION OF - Vị trí, địa điểm hoặc khu vực của một thực thể hoặc nơi diễn ra của một quy trình.

Dưới đây là mối quan hệ ngữ nghĩa tổng quát giữa các thực thể, chi tiết xem tại phụ lục.

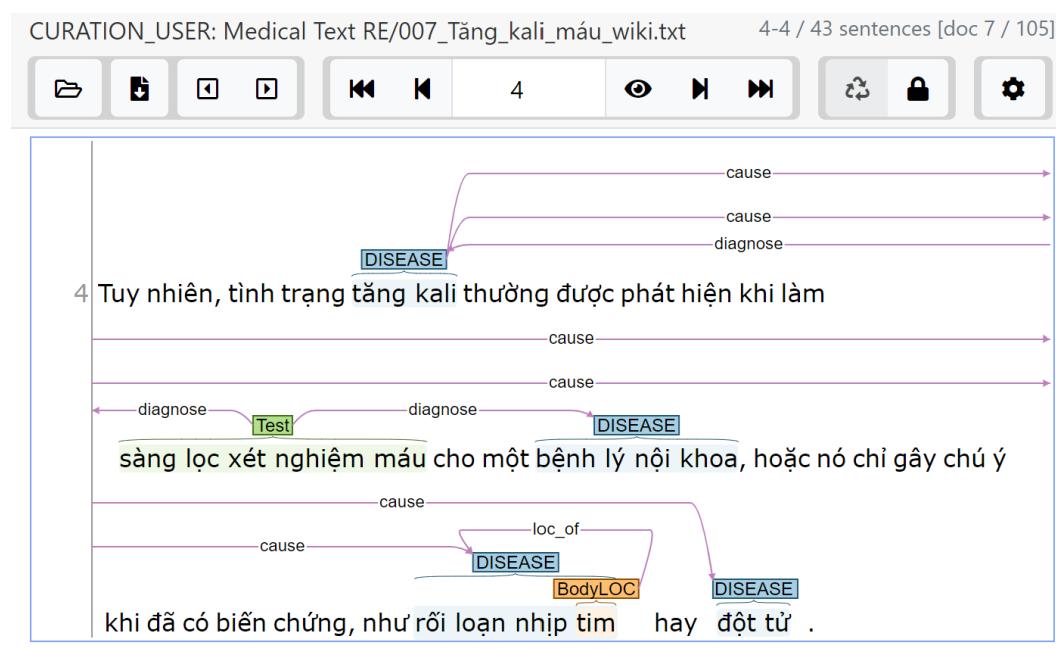
Bảng 3.10: Mối quan hệ ngữ nghĩa tổng quát giữa các thực thể

Thực thể	Mối quan hệ tới các thực thể
SYMPTOM AND DISEASE	<ul style="list-style-type: none"> • SYMPTOM AND DISEASE causes SYMPTOM AND DISEASE.
THERAPEUTIC OR PREVENTIVE PROCEDURE	<ul style="list-style-type: none"> • THERAPEUTIC OR PREVENTIVE PROCEDURE treats SYMPTOM AND DISEASE. • THERAPEUTIC OR PREVENTIVE PROCEDURE prevents SYMPTOM AND DISEASE. • THERAPEUTIC OR PREVENTIVE PROCEDURE uses MEDICINE.
MEDICINE	<ul style="list-style-type: none"> • MEDICINE treats SYMPTOM AND DISEASE. • MEDICINE prevents SYMPTOM AND DISEASE. • MEDICINE causes SYMPTOM AND DISEASE. • MEDICINE causes SYMPTOM AND DISEASE.
TESTS	<ul style="list-style-type: none"> • TESTS diagnoses SYMPTOM AND DISEASE. • TESTS uses MEDICINE.
BODY LOCATION OR REGION	<ul style="list-style-type: none"> • BODY LOCATION OR REGION location of SYMPTOM AND DISEASE. • BODY LOCATION OR REGION location of THERAPEUTIC OR PREVENTIVE PROCEDURE. • BODY LOCATION OR REGION location of TESTS.

3.1.7 Gán nhãn thực thể và quan hệ giữa các thực thể

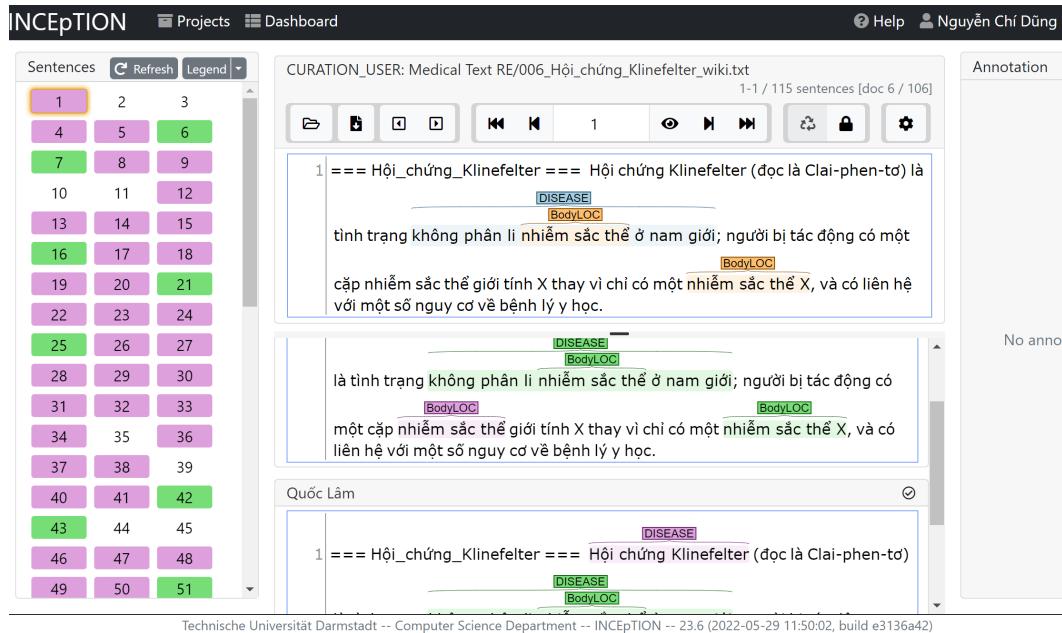
418 văn bản liên quan đến lĩnh vực y tế đã trải qua quá trình tiền xử lý, bao gồm việc chuẩn hóa, loại bỏ dữ liệu nhiễu và đặt tên một cách có hệ thống được đẩy lên hệ thống INCEpTION để phục vụ cho các bước xử lý tiếp theo.

Bước tiếp theo là gán nhãn các thực thể đã được định nghĩa trong dữ liệu và kiểm tra để đảm bảo tính chính xác và nhất quán. Sau khi gán nhãn thực thể, ta tiến hành gán nhãn cho các quan hệ giữa các thực thể đã được xác định. Quá trình này cũng được kiểm tra kỹ lưỡng để đảm bảo tính chính xác và nhất quán của dữ liệu.



Hình 3.3: Gán nhãn thực thể và quan hệ trong văn bản y tế tiếng Việt trên hệ thống INCEpTION

Quá trình kiểm tra bao gồm đánh giá mức độ thống nhất giữa các phiên bản được gán nhãn khác nhau bằng cách kiểm tra thủ công trên từng câu được gán nhãn, sau đó thống nhất trên một phiên bản cuối cùng.



Hình 3.4: Kiểm định chéo các phiên bản trên hệ thống INCEpTION

Cuối cùng, 138 văn bản tương ứng với 4578 câu đã được xử lý và gán nhãn được xuất ra dưới định dạng TSV phiên bản 3.3 , đây là một định dạng phổ biến để lưu trữ và trao đổi dữ liệu giữa các hệ thống và ứng dụng khác nhau.

#FORMAT=WebAnno TSV 3.3
#T_SP=webanno.custom.MedNameEntity MedNameEntity
#T_RL=webanno.custom.MedRelation MedDependency MedRelation BT_webanno.custom.MedNameEntity
#Text=Hội chứng Klinefelter (đọc là Clai-phen-to) là tình trạng không phân li nhiễm sắc thể ở nam giới; người bị tác động có một cặp nhiễm sắc thể giới tính X thay vì chỉ có một nhiễm sắc thể X, và có liên hệ với một số nguy cơ về bệnh lý y học.
1-1 0-1 = - - - - -
1-2 1-2 = - - - - -
1-3 2-3 = - - - - -
1-4 4-25 Hội_chứng_Klinefelter DISEASE - - - - -
1-5 26-27 = - - - - -
1-6 27-28 = - - - - -
1-7 28-29 = - - - - -
1-8 31-34 Hội - - - - -
1-9 35-40 chứng - - - - -
1-10 41-52 Klinefelter DISEASE - - - - -
1-11 53-54 (- - - - -
1-12 54-57 đọc - - - - -
1-13 58-60 là - - - - -
1-14 61-73 Clai-phen-to DISEASE - - - - -
1-15 73-74) - - - - -
1-16 75-77 là - - - - -
1-17 78-82 tình - - - - -
1-18 83-88 trạng - - - - -
1-19 89-94 không DISEASE[1] * loc_of 1-22[2_1]
1-20 95-99 phân DISEASE[1] - - - - -
1-21 100-102 lí DISEASE[1] - - - - -
1-22 103-108 nhiễm DISEASE[1] BodyLOC[2] - - - - -
1-23 109-112 sắc DISEASE[1] BodyLOC[2] - - - - -
1-24 113-116 thể DISEASE[1] BodyLOC[2] - - - - -
1-25 117-118 ó DISEASE[1] - - - - -
1-26 119-122 nam DISEASE[1] - - - - -
1-27 123-127 giới DISEASE[1] - - - - -
1-28 127-128 ; - - - - -
1-29 129-134 người - - - - -
1-30 135-137 bị - - - - -

Hình 3.5: Tệp dưới định dạng TSV phiên bản 3.3

Quá trình thu thập, xử lý và chuẩn bị dữ liệu này đóng vai trò quan trọng trong việc đảm bảo chất lượng và tính nhất quán của dữ liệu đầu vào, từ đó góp phần nâng cao hiệu suất và độ chính xác của mô hình học máy được đào tạo trên tập dữ liệu này.

3.2 Thực nghiệm và kết quả

3.2.1 Bộ dữ liệu trích xuất quan hệ cho văn bản y tế tiếng Việt

Bộ dữ liệu bao gồm hai tập dữ liệu (huấn luyện và kiểm thử) chứa các tệp trong định dạng WebAnno TSV phiên bản 3.3. Mỗi tệp chỉ chứa một văn bản thô đã được tách thành câu với các loại thực thể và quan hệ đã được định nghĩa ở phần xây dựng dữ liệu. Bộ dữ liệu chỉ chứa các mối quan hệ giữa các thực thể được gắn nhãn thuộc cùng một câu.

Trong quá trình thực hiện nghiên cứu, chúng tôi đã gặp phải một số hạn chế nhất định về thời gian và nguồn lực. Do vậy, trong phạm vi nghiên cứu này, chúng tôi chỉ có thể xử lý một tập dữ liệu mẫu bao gồm 200 câu tương ứng với 400 thực thể và 200 quan hệ từ dữ liệu đầu vào. Tập dữ liệu mẫu này được chia thành hai phần khác nhau gồm 150 mẫu cho huấn luyện và 50 mẫu cho kiểm thử để phục vụ cho quá trình huấn luyện và đánh giá mô hình cũng như bộ dữ liệu.

3.2.2 Tiền xử lý dữ liệu, tinh chỉnh mô hình PhoBERT và phương pháp đánh giá

Quá trình tiền xử lý dữ liệu, tinh chỉnh mô hình PhoBERT và phương pháp đánh giá được thực hiện tương tự như ở phần thử nghiệm trong chương 2: Ứng dụng mô hình PhoBERT trong bài toán trích rút quan hệ.

3.2.3 Kết quả thực nghiệm

Bảng 3.11: Hiệu suất của mô hình trên các tập dữ liệu huấn luyện và kiểm thử cho tập dữ liệu mẫu trích xuất quan hệ cho văn bản y tế tiếng Việt

Tập dữ liệu	Điểm số F1 trung bình vĩ mô	Điểm số F1 trung bình vi mô
Huấn luyện	0.67	0.85
Kiểm thử	0.65	0.78

Kết quả đánh giá trên mô hình trích xuất quan hệ trên văn bản y tế tiếng Việt cho thấy hiệu suất ban đầu khả quan nhưng vẫn còn tồn tại một số hạn chế.

Điểm số F1 trung bình vĩ mô trên tập huấn luyện (0.67) và tập kiểm thử (0.65) tương đối gần nhau, cho thấy khả năng khai quát hóa tương đối ổn định của mô hình trên các lớp quan hệ. Tuy nhiên, điểm F1 trung bình vi mô lại cho thấy sự chênh lệch đáng kể giữa hai tập dữ liệu (0.85 so với 0.78), cho thấy dấu hiệu của hiện tượng quá khớp với dữ liệu huấn luyện.

Để nâng cao hiệu suất của mô hình, đặc biệt là khả năng khai quát hóa trên dữ liệu mới, cần tập trung vào các giải pháp như tăng cường dữ liệu, tinh chỉnh siêu tham số, áp dụng các kỹ thuật điều chỉnh, và phân tích lỗi chi tiết. Việc giải quyết các hạn chế này sẽ góp phần xây dựng một mô hình trích xuất quan hệ hiệu quả và đáng tin cậy cho lĩnh vực y tế tiếng Việt.

Kết luận

Nghiên cứu hiện tại đã thành công trong việc xây dựng một bộ dữ liệu bao gồm 138 văn bản y tế với định dạng TSV phiên bản 3.3 được gán nhãn thực thể và quan hệ, đồng thời minh chứng khả năng ứng dụng mô hình PhoBERT kết hợp với các kỹ thuật xử lý ngôn ngữ tự nhiên tiếng Việt thực hiện trích xuất quan hệ trong bộ dữ liệu mẫu gồm 200 câu. Bộ dữ liệu sẽ góp phần phát triển các mô hình xử lý ngôn ngữ tự nhiên chuyên biệt cho lĩnh vực y tế, đặc biệt là cho tiếng Việt. Tuy nhiên, vẫn còn không ít tiềm năng để hoàn thiện bộ dữ liệu, đồng thời cải thiện hiệu suất và mở rộng khả năng ứng dụng của mô hình trong thực tế.

Về phương hướng phát triển trong tương lai, việc khai thác sức mạnh tính toán của phần cứng GPU hiệu năng cao có thể được xem xét. Điều này sẽ tạo điều kiện thuận lợi để huấn luyện các mô hình lớn hơn như phiên bản PhoBERT large được tinh chỉnh song song với các mô hình tiềm năng khác như XLMR.

Ngoài ra, việc hoàn thiện quy trình chuẩn bị và xử lý dữ liệu cũng đóng vai trò quan trọng. Điều này bao gồm tối ưu quá trình gán nhãn bằng cách kết hợp với các framework như RASA hay các mô hình ngôn ngữ lớn như ChatGPT, xử lý các ngoại lệ và trường hợp đặc biệt trong dữ liệu văn bản y tế tiếng Việt, cũng như cải tiến công cụ chuyển đổi dữ liệu từ định dạng TSV 3.3 sang định dạng tương thích với các mô hình Transformer. Các bước này sẽ góp phần nâng cao chất lượng dữ liệu đầu vào và đảm bảo tính ổn định của mô hình.

Bên cạnh đó, việc kiểm tra và điều chỉnh bộ dữ liệu cũng đóng vai trò không kém phần quan trọng. Cần đánh giá phân phối của các nhãn và làm rõ các trường hợp mơ hồ nhằm tăng độ tin cậy của bộ dữ liệu, đảm bảo mô hình có thể học tập một cách hiệu quả.

Bằng cách tập trung vào các hướng phát triển trên, hiệu suất và khả năng ứng dụng của mô hình trong lĩnh vực y tế có thể được nâng cao đáng kể. Kết quả nghiên cứu này sẽ đóng góp vào sự phát triển của các hệ thống xử lý ngôn

ngữ tự nhiên tiếng Việt chuyên sâu, hỗ trợ các ứng dụng y tế và mang lại lợi ích thiết thực cho cộng đồng.

Tài liệu tham khảo

- [1] H. F. Course, “Causal language modeling.” <https://huggingface.co/learn/nlp-course/chapter1/4>, 2023. Accessed: 2024-05-28.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.
- [4] D. Q. Nguyen and A. T. Nguyen, “PhoBERT: Pre-trained language models for Vietnamese,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1037–1042, 2020.
- [5] J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, and I. Gurevych, “The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation,” in *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, (Santa Fe, New Mexico), pp. 5–9, 2018.
- [6] E. F. Tjong Kim Sang and F. De Meulder, “Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition,” in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003.

- [7] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham, “Results of the WNUT2017 shared task on novel and emerging entity recognition,” in *Proceedings of the 3rd Workshop on Noisy User-generated Text* (L. Derczynski, W. Xu, A. Ritter, and T. Baldwin, eds.), (Copenhagen, Denmark), pp. 140–147, Association for Computational Linguistics, Sept. 2017.
- [8] E. A. Mendonca, J. Haas, L. Shagina, E. Larson, and C. Friedman, “Extracting information on pneumonia in infants using natural language processing of radiology reports,” in *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, (Sapporo, Japan), pp. 81–88, Association for Computational Linguistics, July 2003.
- [9] T. D. Huy, N. A. Tu, T. H. Vu, N. P. Minh, N. Phan, T. H. Bui, and S. Q. H. Truong, “Vimq: A vietnamese medical question dataset for healthcare dialogue system development,” in *Neural Information Processing* (T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, eds.), (Cham), pp. 657–664, Springer International Publishing, 2021.
- [10] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, “SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals,” in *Proceedings of the 5th International Workshop on Semantic Evaluation* (K. Erk and C. Strapparava, eds.), (Uppsala, Sweden), pp. 33–38, Association for Computational Linguistics, July 2010.
- [11] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning, “Position-aware attention and supervised data improve slot filling,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (M. Palmer, R. Hwa, and S. Riedel, eds.), (Copenhagen, Denmark), pp. 35–45, Association for Computational Linguistics, Sept. 2017.
- [12] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna, and T. Charnois, “SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers,” in *Proceedings of the 12th International Workshop on Semantic Evaluation* (M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, and M. Carpuat, eds.), (New Orleans,

- Louisiana), pp. 679–688, Association for Computational Linguistics, June 2018.
- [13] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, “2010 i2b2/va challenge on concepts, assertions, and relations in clinical text,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.
 - [14] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii, “Overview of BioNLP shared task 2011,” in *Proceedings of BioNLP Shared Task 2011 Workshop* (J. Tsujii, J.-D. Kim, and S. Pyysalo, eds.), (Portland, Oregon, USA), pp. 1–6, Association for Computational Linguistics, June 2011.
 - [15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pre-training approach,” *arXiv preprint arXiv:1907.11692*, 2019.
 - [16] T. Nguyễn and H. Mᾶn, “Vietnamese relation extraction with BERT-based models at VLSP 2020,” in *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, (Hanoi, Vietnam), pp. 30–34, Association for Computational Linguistics, Dec. 2020.
 - [17] V. Godbole, G. E. Dahl, J. Gilmer, C. J. Shallue, and Z. Nado, “Deep learning tuning playbook,” 2023. Version 1.0.

Phụ lục

Bảng 12: Các ví dụ về SYMPTOM AND DISEASE cần được gán nhãn

Trường hợp	Ví dụ
Các cụm danh từ, tính từ mô tả bệnh, hội chứng, dấu hiệu, triệu chứng	<ul style="list-style-type: none"> Bệnh nhân bị chẩn đoán mắc hội chứng vàng da do suy gan. Sau khi khám, bác sĩ phát hiện ra dấu hiệu sốt cao, ho và khó thở ở bệnh nhân. Bệnh Alzheimer là một hội chứng thoái hóa thần kinh tiến triển.
Những nhận định về tình trạng tinh thần, hành vi	<ul style="list-style-type: none"> Bệnh nhân trầm cảm nặng có biểu hiện các triệu chứng mất hứng thú với mọi thứ, cảm giác tuyệt vọng và vô dụng. Người bị PTSD có những hành vi hoảng loạn, sợ hãi khi đối mặt với các kích thích làm hồi tưởng lại sự kiện đau thương. Hội chứng rối loạn ám ảnh cưỡng chế khiến bệnh nhân có những suy nghĩ ám ảnh và các hành động lặp đi lặp lại.
Các loại vi rút, vi khuẩn gây bệnh	<ul style="list-style-type: none"> Bệnh nhân đang trong giai đoạn cấp của bệnh viêm gan do vi rút viêm gan B. Bệnh bạch hầu do vi khuẩn bạch hầu gây ra, đặc trưng bởi các triệu chứng viêm họng, sốt cao, khó thở. Uốn ván do độc tố của vi khuẩn Clostridium tetani xâm nhập vào cơ thể qua vết thương hở.

Chấn thương, tổn thương	<ul style="list-style-type: none"> Bệnh nhân bị đa chấn thương sau tai nạn giao thông, đặc biệt là chấn thương sọ não và gãy xương đùi phải. Trẻ nhỏ bị ngộ độc thức ăn với biểu hiện buồn nôn, nôn và tiêu chảy gây mất nước và rối loạn điện giải trong cơ thể. Tắc ruột do di chuyển tự phát của một số cấu trúc trong ổ bụng gây ra chứng tắc ruột cấp tính có nguy cơ hoại tử.
Các bất thường về cấu trúc, chức năng cơ thể	<ul style="list-style-type: none"> Bệnh nhân bị teo cơ nặng ở chi dưới, dẫn đến liệt hoàn toàn 2 chi, mất khả năng đi lại và sinh hoạt. Người bệnh bị suy giảm chức năng gan nặng do xơ gan khiến cơ thể dễ bị vàng da, chán ăn, mệt mỏi. Bé trai được chẩn đoán mắc hội chứng Down với các đặc điểm bất thường về ngoại hình và chậm phát triển trí tuệ.
Kết quả xét nghiệm được mô tả rõ ràng là bất thường	<ul style="list-style-type: none"> Xét nghiệm máu cho thấy bệnh nhân bị thiểu máu nặng với hồng cầu: 2.5 triệu/mm3, hemoglobin: 8 g/dL. Kết quả chụp CT cho thấy khối u kích thước 8 x 5 cm ở thùy phải gan, có đặc điểm ác tính. Xét nghiệm gen phát hiện đột biến gen BRCA1, là nguyên nhân gây ra bệnh ung thư vú di truyền.

Bảng 13: Các ví dụ về SYMPTOM AND DISEASE không được gán nhãn

Trường hợp	Ví dụ
Các mô tả về trạng thái sức khỏe bình thường	<ul style="list-style-type: none"> Khám sức khỏe tổng quát cho kết quả bình thường, nhịp tim bình thường.
Các giá trị đo, kết quả xét nghiệm nếu không mô tả rõ bất thường	<ul style="list-style-type: none"> Đường huyết lúc đói của bệnh nhân là 5.5 mmol/L. Kết quả xét nghiệm creatinine và ure trong giới hạn bình thường.
Các động từ mô tả kết quả/tiến triển bệnh	<ul style="list-style-type: none"> Cơn đau của bệnh nhân có giảm dần sau khi dùng thuốc. Bướu cổ của bệnh nhân có xu hướng phát triển nhanh.
Sử dụng rượu bia, thuốc lá	<ul style="list-style-type: none"> Bệnh nhân có tiền sử nghiện rượu nặng. Người bệnh đang hút khoảng 1 gói thuốc lá mỗi ngày.

Bảng 14: Các ví dụ về THERAPEUTIC OR PREVENTIVE PROCEDURE cần được gán nhãn

Trường hợp	Ví dụ
Các thủ thuật, phẫu thuật, kỹ thuật y khoa được áp dụng cho bệnh nhân	<ul style="list-style-type: none"> Bệnh nhân được phẫu thuật cắt bỏ khối u ác tính ở đại tràng để ngăn chặn sự phát triển của các khối u di căn. Trẻ bị sẹo co kéo ở tay được kỹ thuật phẫu thuật kéo dài gân để cải thiện khả năng cử động của tay bị hạn chế.
Các biện pháp can thiệp trị liệu	<ul style="list-style-type: none"> Bệnh nhân ung thư vú được chỉ định trị liệu bằng phương pháp hóa trị sử dụng thuốc 5-Fluorouracil nhằm tiêu diệt tế bào ung thư. Bệnh nhân bị thoát vị đĩa đệm được châm cứu kết hợp với xoa bóp bấm huyệt để giảm đau và cải thiện cử động. Bệnh nhân trầm cảm được điều trị bằng liệu pháp tâm lý kết hợp với dùng thuốc chống trầm cảm nhằm ổn định tâm trạng.
Các thủ tục, kỹ thuật hỗ trợ điều trị	<ul style="list-style-type: none"> Bệnh nhân bị sốc nhiễm khuẩn được chỉ định truyền dịch để bù lượng máu lưu thông và cân bằng điện giải trong cơ thể. Bệnh nhân thiếu máu nặng sau tai nạn giao thông cần được thực hiện truyền máu khẩn cấp để bù lượng hồng cầu và hemoglobin. Bệnh nhân hôn mê sâu sau chấn thương sọ não được đặt nội khí quản, thở máy để duy trì chức năng sống.

Bảng 15: Các ví dụ về THERAPEUTIC OR PREVENTIVE PROCEDURE
không được gán nhãn

Trường hợp	Ví dụ
Các dụng cụ, thiết bị y tế	<ul style="list-style-type: none"> Trước khi bắt đầu ca phẫu thuật, y tá sử dụng máy đo huyết áp để đo áp lực máu của bệnh nhân và đảm bảo rằng nó ổn định. Bệnh nhân sau phẫu thuật tim được theo dõi nhịp tim liên tục bằng máy đo điện tâm đồ nhằm phát hiện sớm các rối loạn nhịp thất. Trong trường hợp bệnh nhân bị sốt, y tá sử dụng nhiệt kế để đo nhiệt độ và quyết định liệu cần gấp bác sĩ hay không.
Các bộ phận, khoa, phòng y tế	<ul style="list-style-type: none"> Bệnh nhân đái tháo đường được tư vấn và theo dõi điều trị tại khoa Nội Tiết của bệnh viện. Bệnh nhi bị viêm phổi nặng được chuyển đến khoa Hồi sức tích cực - Chống độc để theo dõi sát và can thiệp hồi sức kịp thời. Bệnh nhân sau mổ được chuyển về phòng hậu phẫu để tiếp tục được theo dõi chăm sóc, xử trí kịp thời biến chứng.
Địa điểm điều trị	<ul style="list-style-type: none"> Người bệnh được nhân viên y tế tại trạm xá thôn bản khám, kê đơn và hướng dẫn sử dụng thuốc. Bệnh nhân cao huyết áp được tư vấn và kê đơn thuốc điều trị tại phòng khám đa khoa gần nhà.

Bảng 16: Các ví dụ về TESTS cần được gán nhãn

Trường hợp	Ví dụ
Các xét nghiệm, thủ thuật y học nhằm mục đích chẩn đoán, theo dõi, đánh giá tình trạng bệnh	<ul style="list-style-type: none"> Bệnh nhân cao huyết áp được lấy máu xét nghiệm định lượng lipid máu nhằm đánh giá nguy cơ mắc bệnh tim mạch. Bệnh nhân nghi ngờ mắc bệnh bạch cầu cấp được chỉ định chọc dò tủy sống để làm xét nghiệm tế bào học, sinh hóa tủy nhằm xác định chính xác chẩn đoán và phân loại mô bệnh học bệnh bạch cầu.
Các xét nghiệm cận lâm sàng trên mẫu máu, dịch cơ thể, mô	<ul style="list-style-type: none"> Người bệnh có biểu hiện sốt, nhức đầu, nôn ói, nghi ngờ viêm màng não do vi khuẩn. Bác sĩ lấy mẫu dịch não tủy để thực hiện xét nghiệm nuôi cấy phân lập vi khuẩn gây bệnh và kháng sinh đồ nhằm xác định tác nhân và phác đồ kháng sinh phù hợp.
Các kỹ thuật hình ảnh như chụp X-quang, cộng hưởng từ, siêu âm, nội soi	<ul style="list-style-type: none"> Bệnh nhân bị đau bụng được chỉ định chụp cắt lớp để phát hiện các khối u, tổn thương ở các cơ quan nội tạng trong ổ bụng. Thai phụ mang thai 28 tuần được siêu âm 4D để theo dõi sự phát triển của thai nhi, kiểm tra các bất thường bên trong tử cung.
Các thăm dò chức năng như điện tim, đo thính lực, thị lực	<ul style="list-style-type: none"> Bệnh nhân có các triệu chứng đau ngực, khó thở, nghi ngờ bị rối loạn nhịp tim. Bác sĩ cho bệnh nhân đo điện tâm đồ, ghi nhận đường điện tim trong 24 giờ để phát hiện các dấu hiệu loạn nhịp thất và đánh giá mức độ nguy hiểm.

Bảng 17: Các ví dụ về TESTS không được gán nhãn

Trường hợp	Ví dụ
Các thiết bị, dụng cụ y tế	<ul style="list-style-type: none"> Bác sĩ khoa chẩn đoán hình ảnh quan sát dưới kính hiển vi quang học các mẫu bệnh phẩm để phát hiện bất thường.
Kết quả đo lường	<ul style="list-style-type: none"> Điều dưỡng dùng máy đo huyết áp và nghe nhịp tim để theo dõi các chỉ số sinh tồn của bệnh nhân sau phẫu thuật tim là: nhip tim 92 lần/phút, huyết áp 140/90 mmHg.
Các khoa/phòng xét nghiệm và hình ảnh	<ul style="list-style-type: none"> Các kỹ thuật viên xét nghiệm tại phòng xét nghiệm thực hiện xét nghiệm sinh hóa, huyết học, miễn dịch trên các mẫu bệnh phẩm để phục vụ công tác chẩn đoán.

Bảng 18: Các ví dụ về MEDICINE cần được gán nhãn

Trường hợp	Ví dụ
Các loại thuốc được kê đơn cho bệnh nhân: tên thương hiệu, tên chung, tên nhóm thuốc	<ul style="list-style-type: none"> Bệnh nhân được kê đơn thuốc giảm đau paracetamol 500mg với liều dùng 1 viên khi cần, không quá 4 viên/ngày. Bác sĩ kê đơn cho bệnh nhân kháng sinh cephalosporin thế hệ 3 là Ceftriaxone 1g tiêm truyền tĩnh mạch mỗi ngày. Bệnh nhân cao huyết áp được kê đơn thuốc đổi kháng thụ thể angiotensin (losartan) và thuốc lợi tiểu quai (furosemide).
Các chế phẩm sinh học, huyết thanh, kháng thể được dùng với mục đích điều trị	<ul style="list-style-type: none"> Liệu pháp sinh học với các kháng thể đơn dòng được sử dụng để điều trị bệnh viêm khớp dạng thấp. Huyết thanh kháng độc tố botulinum được dùng để điều trị ngộ độc thực phẩm do vi khuẩn <i>Clostridium botulinum</i> gây ra. Các yếu tố kích thích tạo máu như erythropoietin và filgrastim là những chế phẩm sinh học quan trọng trong điều trị suy tủy xương.
Thuật ngữ chung chỉ đơn thuốc/liệu trình điều trị của bệnh nhân	<ul style="list-style-type: none"> Bệnh nhân được kê đơn Ibuprofen (Thuốc chống viêm) để giảm viêm nhiễm và giảm đau.

Bảng 19: Các ví dụ về MEDICINE không được gán nhãn

Trường hợp	Ví dụ
Tên công ty sản xuất thuốc	<ul style="list-style-type: none"> "Paracetamol" được sản xuất bởi công ty XYZ.
Các dạng bào chế thuốc: viên nén, siro, thuốc tiêm	<ul style="list-style-type: none"> "Aspirin" có dạng viên nén hoặc dạng nước.
Quy trình, cách thức dùng thuốc	<ul style="list-style-type: none"> Uống sau khi ăn để giảm nguy cơ kích thích dạ dày.
Các thiết bị, dụng cụ y tế liên quan thuốc	<ul style="list-style-type: none"> Bình xịt phun thuốc để đưa thuốc vào họng.

Bảng 20: Các ví dụ về BODY LOCATION OR REGION cần được gán nhãn

Trường hợp	Ví dụ
Mô tả Vị Trí Cụ Thể trên Cơ Thể	<ul style="list-style-type: none"> Vết thương nằm ở phần trên của cánh tay trái đang gây đau đớn cho bệnh nhân. Kiểm tra áp lực máu ở vùng cổ của bệnh nhân để đánh giá tình trạng cường độ máu của não. Bệnh nhân có đau ở vùng bụng dưới, phía bên phải, có thể là dấu hiệu của vấn đề về gan hoặc tụy.
Thực Thể Y Tế Liên Quan Đến Một Khu Vực Cụ Thể	<ul style="list-style-type: none"> Cơ sở y tế này chuyên về điều trị các vấn đề về hệ tiêu hóa, bao gồm bệnh dạ dày và ruột kết. Thực thể y tế này tập trung vào nghiên cứu và điều trị các bệnh lý của hệ thống hô hấp, bao gồm cả hen suyễn và viêm phế quản. Trung tâm y tế này chuyên về các phương pháp phòng và điều trị các bệnh về hệ thống thần kinh trung ương, bao gồm cả đột quỵ và bệnh Parkinson.

Bảng 21: Mối quan hệ ngữ nghĩa giữa các thực thể

Thực thể	Mối quan hệ tới các thực thể
SYMPTOM AND DISEASE	<p>SYMPTOM AND DISEASE causes SYMPTOM AND DISEASE</p> <ul style="list-style-type: none"> • Virus causes Disease or Syndrome. • Fungus causes Disease or Syndrome. • Bacterium causes Disease or Syndrome.
THERAPEUTIC OR PREVENTIVE PROCEDURE	<p>THERAPEUTIC OR PREVENTIVE PROCEDURE treats SYMPTOM AND DISEASE.</p> <ul style="list-style-type: none"> • Therapeutic or Preventive Procedure treats Injury or Poisoning. • Therapeutic or Preventive Procedure treats Sign or Symptom. • Therapeutic or Preventive Procedure treats Sign or Symptom. <p>THERAPEUTIC OR PREVENTIVE PROCEDURE prevents SYMPTOM AND DISEASE.</p> <ul style="list-style-type: none"> • Therapeutic or Preventive Procedure prevents Disease or Syndrome. <p>THERAPEUTIC OR PREVENTIVE PROCEDURE uses MEDICINE</p> <ul style="list-style-type: none"> • Therapeutic or Preventive Procedure uses Pharmacologic Substance. • Therapeutic or Preventive Procedure uses Antibiotic. • Therapeutic or Preventive Procedure uses Clinical Drug.

	<p>MEDICINE treats SYMPTOM AND DISEASE</p> <ul style="list-style-type: none"> • Antibiotic treats Injury or Poisoning. • Antibiotic treats Disease or Syndrome. • Antibiotic treats Sign or Symptom. • Pharmacologic Substance treats Injury or Poisoning • Pharmacologic Substance treats Sign or Symptom • Pharmacologic Substance treats Disease or Syndrome <p>MEDICINE prevents SYMPTOM AND DISEASE</p> <ul style="list-style-type: none"> • Antibiotic prevents Disease or Syndrome. • Pharmacologic Substance prevents Disease or Syndrome. <p>MEDICINE causes SYMPTOM AND DISEASE</p> <p>MEDICINE</p> <ul style="list-style-type: none"> • Antibiotic causes Injury or Poisoning. • Antibiotic causes Disease or Syndrome. • Pharmacologic Substance causes Injury or Poisoning. • Pharmacologic Substance causes Disease or Syndrome. • Clinical Drug causes Injury or Poisoning. • Clinical Drug causes Disease or Syndrome. • Vitamin causes Injury or Poisoning. • Vitamin causes Disease or Syndrome <p>MEDICINE diagnoses SYMPTOM AND DISEASE</p> <ul style="list-style-type: none"> • Antibiotic diagnoses Disease or Syndrome. • Pharmacologic Substance diagnoses Disease or Syndrome.
--	---

TESTS	<p>TESTS diagnoses SYMPTOM AND DISEASE</p> <ul style="list-style-type: none"> ● Laboratory Procedure diagnoses Injury or Poisoning. ● Laboratory Procedure diagnoses Disease or Syndrome. ● Diagnostic Procedure diagnoses Injury or Poisoning. ● Diagnostic Procedure diagnoses Disease or Syndrome <p>TESTS uses MEDICINE</p> <ul style="list-style-type: none"> ● Diagnostic Procedure uses Pharmacologic Substance. ● Diagnostic Procedure uses Antibiotic. ● Diagnostic Procedure uses Clinical Drug.
BODY LOCATION OR REGION	<p>BODY LOCATION OR REGION location of SYMPTOM AND DISEASE</p> <ul style="list-style-type: none"> ● Body Location or Region location of Injury or Poisoning. ● Body Location or Region location of Disease or Syndrome <p>BODY LOCATION OR REGION location of THERAPEUTIC OR PREVENTIVE PROCEDURE</p> <ul style="list-style-type: none"> ● Body Location or Region location of Therapeutic or Preventive Procedure. <p>BODY LOCATION OR REGION location of TESTS</p> <ul style="list-style-type: none"> ● Body Location or Region location of Diagnostic Procedure.