Lời cam đoan

Lời cảm ơn

Mục lục

Danh sách hình vẽ

Danh sách bảng

Chương 1

Tổng quan

1.1 Giới thiệu đề tài

1.2 Mục tiêu và phạm vi đề tài

1.3 Cấu trúc luận văn

Chương 2

Các công trình liên quan

2.1 Bệnh án điện tử

2.2 Nhận dạng thực thể có tên

Chương 3

Kiến thức nền tảng

Chương 4

Hiện thực hệ thống

4.1 Nội dung bài toán

Bài toán mà chúng tôi giải quyết là bài toán: "Phân giải đồng tham chiếu cho các hồ sơ xuất viện tiếng Anh với các khái niệm đã được trích xuất và gán nhãn". Đầu vào của bài toán bao gồm hai phần:

1. *Tập các hồ sơ xuất viện:* Đây là những văn bản lâm sàng được viết tay bằng ngôn ngữ tự nhiên bởi các bác sĩ, y tá. Chúng mô tả lại toàn bộ thông tin của bệnh nhân trong một lần điều trị, bao gồm các thông tin về tên bệnh mà bệnh nhân mắc phải, các thủ tục y tế được thực hiện và các phương pháp điều trị được áp dụng lên bệnh nhân.

2. *Tập các khái niệm đã được trích xuất và gán nhãn:* Mỗi hồ sơ xuất viện đi kèm với một văn bản chứa toàn bộ các khái niệm được đề cập trong hồ sơ đó. Các khái niệm này đã được gán nhãn cho phù hợp với loại thực thể mà nó đề cập tới. Có tất cả năm nhãn là Problem, Treatment, Test, Person và Pronoun được i2b2 định nghĩa. Bảng 4.1 mô tả chi tiết ý nghĩa của năm nhãn này.

Mục tiêu của chúng tôi là phân giải đồng tham chiếu cho các khái niệm trong tập các khái niệm ứng với mỗi hồ sơ xuất viện. Cụ thể kết quả đầu ra là danh sách các chuỗi đồng tham chiếu của khái niệm đó, ví dụ

c="the patient" 13:0 13:1||c="he" 14:0 14:0||c="his" 14:7 14:7||t="coref person"

mô tả một chuỗi đồng tham chiếu bao gồm các khái niệm "the patient" (xuất hiện ở dòng thứ 13, từ vị trí 0 đến 1), "he" và "his". Các khái niệm này đồng tham chiếu tới cùng một người (*t="coref person"*).

4.2 Ý tưởng hiện thực

Dựa vào hệ thống có hiệu năng tốt nhất của thử thách i2b2 năm 2011 (hệ thống I), mô hình phân giải đồng tham chiếu mà chúng tôi sử dụng để hiện thực hệ thống là mô hình

Bảng 4.1: Ý nghĩa các lớp thực thể được đề xuất bởi i2b2

Lớp	Định nghĩa	Ví dụ
Person	Những chủ thể người hoặc một nhóm người được đề cập trong bệnh án và các đại từ nhân xưng	Dr.Lightman, the patient, cardiology, he, she, ...
Problem	Những bất thường về sức khỏe thân thể hoặc tinh thần của bệnh nhân, được mô tả bởi bệnh nhân hoặc quan sát của bác sĩ	Heart attack, blood pressure, cancer, ...
Test	Những thủ tục y tế như xét nghiệm, đo đạc, kiểm tra trên cơ thể bệnh nhân để cung cấp thêm thông tin cho "Problem"	CT scan, Temperature, ...
Treatment	Những thủ tục y tế hoặc quy trình áp dụng để chữa trị cho "Problem", bao gồm thuốc, phẫu thuật hoặc phương pháp điều trị	Surgery, ice pack, Tylenol, ...
Pronoun	Những đại từ có thể tham chiếu đến bất kì lớp nào trong bốn lớp kể trên nhưng không phải là đại từ nhân xưng	Which, it, that, ...

cặp thực thể. Tư tưởng cơ bản của mô hình này là xác định xem hai khái niệm bất kì có đồng tham chiếu với nhau hay không, sau đó gom nhóm các cặp đồng tham chiếu có một khái niệm chung lại để tạo thành các chuỗi đồng tham chiếu. Như vậy kiến trúc tổng quát của hệ thống chúng tôi hiện thực gồm 2 quy trình: *quy trình huấn luyện hệ thống phân loại* và *quy trình phân giải đồng tham chiếu*. Trong đó quy trình huấn luyện là bước huấn luyện các model phân loại dựa trên dữ liệu mẫu đã được phân giải đồng tham chiếu. Quy trình phân giải sử dụng các model phân loại đã được huấn luyện để xác định tính đồng tham chiếu của các cặp khái niệm, từ đó sử dụng một giải thuật gom nhóm các cặp đồng tham chiếu lại để tạo thành các chuỗi đồng tham chiếu.

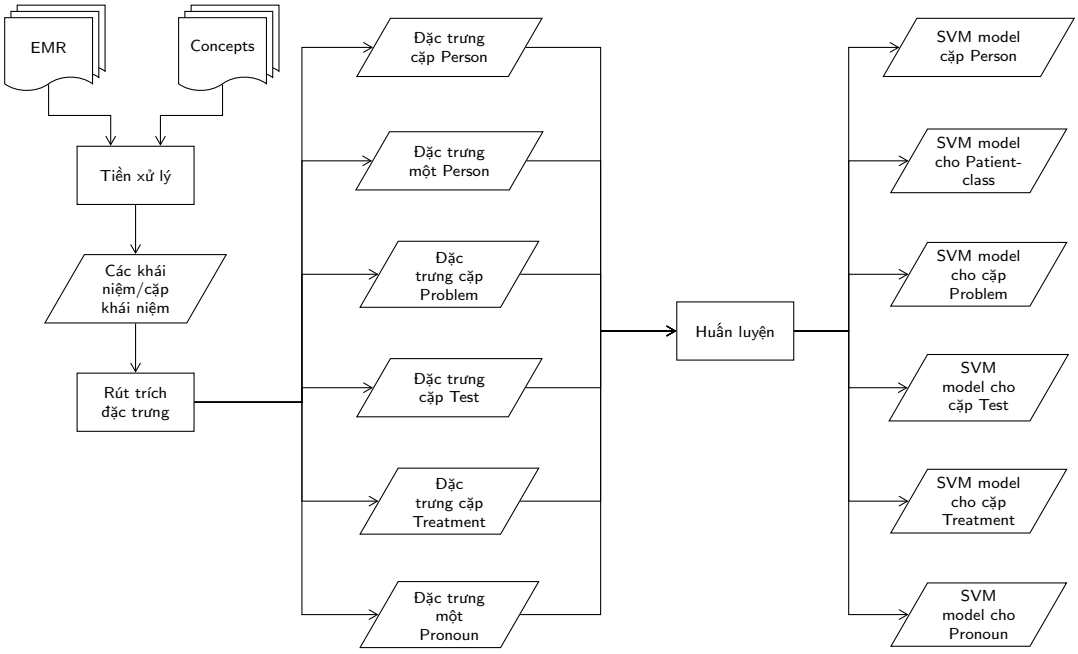
Quy trình huấn luyện

Để xác định tính đồng tham chiếu giữa hai khái niệm bất kì, ta cần huấn luyện một model phân loại dựa trên dữ liệu mẫu. Vì đầu vào của quy trình là các văn bản BAĐT và danh sách các khái niệm đã được gán nhãn, hệ thống cần trích xuất đặc trưng của các dữ liệu thô này rồi mới có thể đưa vào để huấn luyện. Bên cạnh đó, các khái niệm đã được phân loại vào 4 nhóm chính là Person, Problem, Test và Treatment, còn các đại từ được phân vào nhóm Pronoun nên để giảm bớt số cặp khái niệm được sinh ra, chúng tôi huấn luyện 4 model để xác định tính đồng tham chiếu của riêng các cặp Person-Person, Problem-Problem, Test-Test và Treatment-Treatment (vì hai khái niệm thuộc hai lớp khác nhau thì nghiễm nhiên không đồng tham chiếu với nhau). Đối với các đại từ thì thường chỉ tới một khái niệm ở trước đó, nên việc xác định xem một đại từ thực chất mang ý nghĩa của lớp nào trong 4 lớp chính Person, Problem, Test, Treatment là một việc quan trọng. Sau khi xác định được lớp chính của đại từ, chúng tôi chọn khái niệm thuộc lớp tương ứng ở gần nhất trước đó làm tiền đề cho nó. Các ý này đều là của các tác giả hệ thống I.

Ngoài ra cũng theo các tác giả này, thông tin một khái niệm lớp Person có chỉ về bệnh nhân hay không góp một phần quan trọng trong việc phân loại đúng tính đồng tham chiếu của cặp các khái niệm lớp này. Trong miền văn bản BAĐT, các khái niệm chỉ người thường chỉ đề cập đến một trong ba loại: bệnh nhân, người thân của bệnh nhân và

nhân sự của bệnh viện. Do một BADT, mà cụ thể là hồ sơ xuất viện, thông thường chỉ đề cập đến một bệnh nhân nên những khái niệm nào chỉ về bệnh nhân thì thường chắc chắn nằm trong cùng một chuỗi đồng tham chiếu lớn nhất và duy nhất chỉ về bệnh nhân đó. Từ nhận định này, nhóm tác giả của hệ thống I đã thêm vào đặc trưng lớp Patient (Patient-class) cho cặp hai khái niệm lớp Person, nó mang giá trị 1 khi hai khái niệm đều chỉ về bệnh nhân và 0 trong các trường hợp khác. Ở bước huấn luyện, thông tin "một khái niệm Person có chỉ về bệnh nhân hay không" được lấy từ tập chuỗi kết quả (ground truth), còn ở bước phân giải đồng tham chiếu thông tin này được xác định nhờ một model phân loại đã được huấn luyện.

Như vậy mục đích của quy trình huấn luyện là xây dựng tổng cộng 6 SVM model, trong đó 4 SVM model nhằm mục đích phân loại và đánh giá độ tin cậy đồng tham chiếu của các cặp khái niệm Person-Person, Problem-Problem, Test-Test và Treatment-Treatment; 1 SVM model để xác định các khái niệm Person có là bệnh nhân hay không (Patient-class) và 1 SVM model để phân loại các đại từ (các khái niệm lớp Pronoun) vào một trong bốn lớp Person, Problem, Test và Treatment. Đầu vào của quy trình này là toàn bộ các văn bản BADT với các khái niệm đã được trích xuất và gán nhãn. Sau khi tiền xử lý, hệ thống xây dựng các mẫu huấn luyện bao gồm: Person, Person-Person, Problem-Problem, Test-Test, Treatment-Treatment và Pronoun từ danh sách các khái niệm. Sáu tập mẫu này được trích xuất thuộc tính và đưa vào để huấn luyện 6 SVM model (Hình 4.1). Thư viện SVM được nhóm sử dụng là LibSVM.



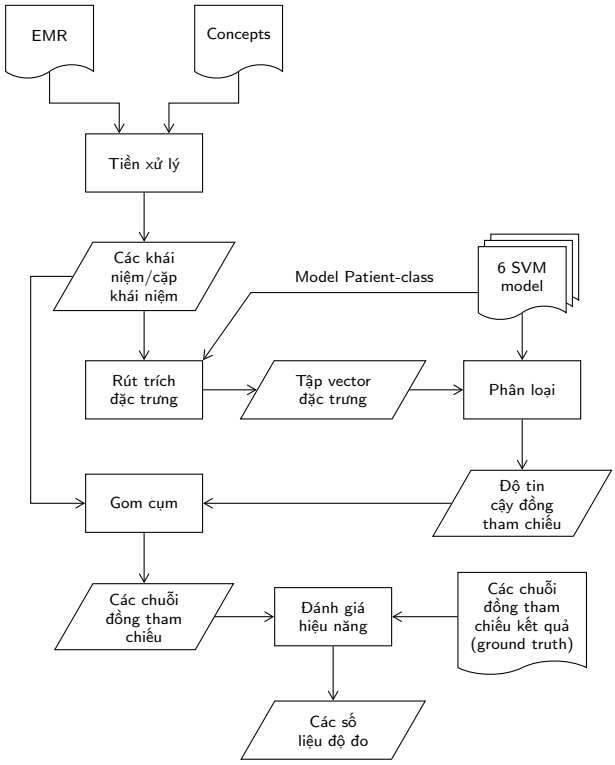
Hình 4.1: Sơ đồ huấn luyện

Quy trình phân giải

Quy trình phân giải đồng tham chiếu sử dụng 6 SVM model đã được huấn luyện ở trên, cùng với đó là một giải thuật gom nhóm các cặp khái niệm đã được phân loại là đồng tham chiếu với nhau lại để cuối cùng tạo thành các chuỗi đồng tham chiếu. Có thể xem

đây là quy trình mang đi ứng dụng thực tế để phân giải cho những văn bản BADT mới. Dựa vào hệ thống I, chúng tôi sử dụng giải thuật gom cụm tốt nhất trước để lựa chọn các cặp đồng tham chiếu có độ tin cậy cao nhất, sau đó xây dựng các chuỗi đồng tham chiếu bằng cách nối các cặp có một khái niệm chung. Đối với lớp Pronoun, sau khi đã xác định được lớp chính của một đại từ bất kì, chúng tôi tạo một cặp đồng tham chiếu giữa đại từ đó và khái niệm thuộc lớp chính tương ứng ở gần nhất trước đó trong văn bản. Theo nhận định của các tác giả hệ thống I, tuy cách làm này đơn giản nhưng lại tỏ ra rất hiệu quả.

Hình 4.2 mô tả trực quan quy trình phân giải đồng tham chiếu. Ở bước trích xuất thuộc tính của các cặp Person, chúng tôi sử dụng model phân loại bệnh nhân để xác định giá trị cho đặc trưng lớp Patient đã được đề cập ở trên. Theo kết quả đánh giá các hệ thống dự thi thử thách i2b2 2011, ba hệ đo được sử dụng là: MUC, B-CUBED và CEAF. Chúng tôi cũng hiện thực các hệ đo này để đánh giá hệ thống của mình bằng cách so sánh với kết quả của hệ thống I.



Hình 4.2: Sơ đồ phân giải đồng tham chiếu

4.3 Tiền xử lý

Trong quá trình rút trích đặc trưng, một số khái niệm được miêu tả cụ thể làm cho việc so trùng chuỗi hoặc tìm kiếm từ các nguồn tri thức nhân loại thiếu chính xác [1]. Ví dụ như khái niệm "her CT scan" và khái niệm "a CT scan". Mặc dù hai khái niệm này cùng chỉ một thủ tục y tế nhưng không trùng chuỗi. Ngoài ra các mạo từ "her", "a" làm việc tìm kiếm tri thức nhân loại từ các nguồn tri thức như Wikipedia, WordNet không được chính xác hoặc không thể tìm được kết quả. Vì vậy trước khi rút trích đặc trưng, các

khái niệm được tiền xử lý để loại bỏ mạo từ và các thông tin ngữ cảnh. Tuy nhiên, quá trình tiền xử lý chỉ được áp dụng cho các đặc trưng liên quan so trùng chuỗi và tìm kiếm tri thức nhân loại, các đặc trưng khác không cần qua quá trình tiền xử lý mà nhận vào nguyên gốc khái niệm được xác định.

Quá trình tiền xử lý gồm hai bước. Đầu tiên khái niệm sẽ được loại bỏ tất cả mạo từ. Sau đó, nếu khái niệm có bao gồm giới từ thì giới từ đó và toàn bộ nội dung theo sau sẽ được lược bỏ. Ví dụ như khái niệm "an MRI of the knee" sau quá trình tiền xử lý sẽ trở thành "MRI". Danh sách mạo từ được xây dựng từ tập dữ liệu và các mạo từ thông dụng của tiếng Anh.

Đặc biệt các khái niệm thuộc lớp Problem/Treatment/Test thường được kèm thêm thông tin về định lượng như 10mg, 5 lit và các thông tin về vị trí giải phẫu học như "upper", "left", "right". Để tăng khả năng tìm kiếm tri thức nhân loại, chúng tôi đề xuất loại bỏ các thông tin ngữ cảnh về số, định lượng và vị trí giải phẫu khỏi khái niệm. Các thông tin ngữ cảnh được loại bỏ bằng cách sử dụng biểu thức chính quy và các từ vựng được xây dựng từ tập dữ liệu. Các đặc trưng liên quan so trùng chuỗi không áp dụng bước tiền xử lý loại bỏ thông tin ngữ cảnh này.

4.4 Xây dựng các cặp khái niệm

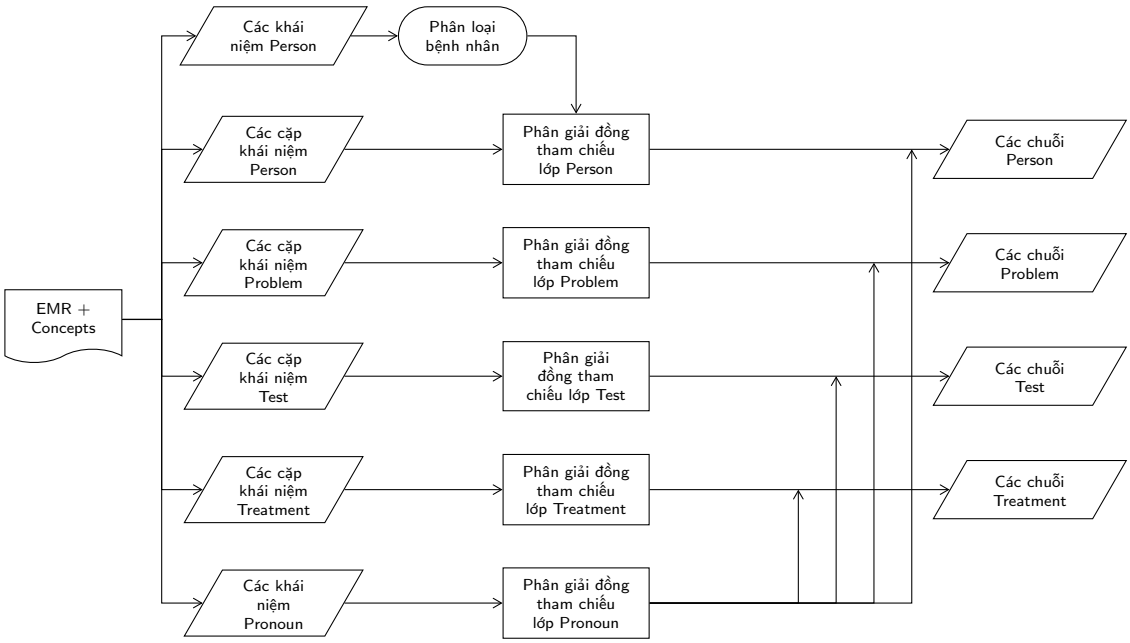
4.5 Rút trích đặc trưng

Từ các phân tích được đề cập ở Phần 3, ngoài các thuộc tính chung về mặt ngôn ngữ (như ngữ pháp hay từ vựng), từng lớp khái niệm ở BAĐT còn mang những đặc tính khác nhau. Việc này đòi hỏi chúng tôi phải thiết kế ba hệ thống rút trích đặc trưng và phân loại tương ứng khác nhau cho lớp Person, lớp Problem/Treatment/Test và lớp Pronoun. Hình 4.3 mô tả tổng quan ba hệ thống này, trong đó các khối "Đồng tham chiếu lớp X" bao hàm cả Hệ thống rút trích đặc trưng và Hệ thống phân loại cho lớp tương ứng.

Nhóm Person

Tổng quát, các khái niệm thuộc lớp Person có thể là các đại từ nhân xưng (he, she, it, they, ...), đại từ sở hữu (his, her, its, their, ...), đại từ phản thân (himself, herself, itself, themselves, ...) hoặc tên người (Stephanie I Sept, Mr. Anders, Heidi Laura Md, ...). Việc phân giải đồng tham chiếu cho tên người và đại từ là công việc khó, vì thông tin có được từ các đại từ và tên người là rất ít. Ngoài ra trong một văn bản thường đề cập đến nhiều hơn một người, khiến cho việc phát hiện chính xác chuỗi đồng tham chiếu cho các khái niệm này là một thách thức lớn.

Dựa vào hệ thống I, việc giới hạn vấn đề lại trong phạm vi BAĐT giúp công việc này trở nên đơn giản hơn. Trong BAĐT, các khái niệm thuộc lớp Person thường được chia vào ba nhóm chính: bệnh nhân, người thân của bệnh nhân hoặc nhân sự của bệnh viện. Trong đó bệnh nhân là nhóm có số lượng khái niệm được đề cập nhiều nhất và chiếm phần lớn tổng số khái niệm lớp Person. Do vậy việc xác định một khái niệm thuộc vào nhóm nào đóng vai trò quan trọng trong việc phân giải chính xác chuỗi đồng tham



Hình 4.3: Tổng quan hệ thống phân giải đồng tham chiếu

chiếu cho khái niệm đó [1]. Từ lí do trên, đặc trưng có phải là bệnh nhân hay không được thêm vào hệ thống. Đặc trưng lớp Patient được xác định bằng phương pháp phân loại nhị phân SVM. Hai nhóm người thân của bệnh nhân và nhân sự của bệnh viện được xác định bằng các đặc trưng từ vựng. Bảng 4.2 trình bày đầy đủ các đặc trưng dùng cho lớp Person.

Với các đặc trưng "Name match", "Relative match", "Department match", "Doctor title match", "Doctor general match", "Twin/Triplet", "We", "You", "I", "Pronoun match", chúng tôi hiện thực bằng cách xây dựng tập từ điển tương ứng với từng đặc trưng dựa trên việc khảo sát tập dữ liệu và sử dụng các biểu thức chính quy.

Đặc trưng về Giới tính được chúng tôi xác định dựa trên ba bước phân loại [2]. Bước thứ nhất: kiểm tra khái niệm có chứa các đại từ xác định giới tính như "Mr", "Ms", "she", "he", ... hay không. Nếu có, xác định giới tính dựa trên đại từ xuất hiện. Nếu không thực hiện bước thứ hai: kiểm tra khái niệm có xuất hiện nhiều hơn một lần hay không. Nếu xuất hiện nhiều hơn một lần thì các lần xuất hiện có chứa đại từ xác định giới tính hay không. Ví dụ khái niệm "Peter H. Diller" có thể xuất hiện nhiều lần, trong đó có xuất hiện dưới hình thức "Mr. Diller". Nếu không thể xác định giới tính qua hai bước kiểm tra, khái niệm sẽ được phân loại bằng cách sử dụng cơ sở dữ liệu về tên tiếng Anh của hệ thống Apache OpenNLP.

Nhóm Patient-class

Từ nhận định trong việc rút trích đặc trưng của lớp Person, chúng tôi xây dựng một hệ thống SVM nhị phân để phân loại khái niệm thuộc lớp Person có phải là bệnh nhân hay không. Trong BAĐT thường chỉ có một bệnh nhân đóng vai trò là chủ thể của bệnh án.Vì vậy, các khái niệm nếu được xác định là bệnh nhân, thì sẽ được đưa vào một chuỗi đồng tham chiếu duy nhất về bệnh nhân đó. Thông qua phân tích tập dữ liệu, chúng tôi nhận

Bảng 4.2: Tập đặc trưng cho lớp Person

Đặc Trưng	Giá trị	Giải thích
Patient-class	0, 1, 2	Không có khái niệm nào là bệnh nhân (0), cả hai khái niệm đều là bệnh nhân (1), trường hợp khác (2)
Distance between sentences	0, 1, 2, 3, ...	Số câu xuất hiện giữa hai khái niệm được xét
Distance between mentions	0, 1, 2, 3, ...	Số khái niệm xuất hiện giữa hai khái niệm được xét
String match	0, 1	Trùng chuỗi hoàn toàn (1), ngược lại (0)
Levenshtein distance between two mentions	0, 1, 2, 3, ...	Khoảng cách Levenshtein giữa hai khái niệm
Number	0, 1, 2	Cả hai đều là số ít hoặc nhiều (1), ngược lại (0), không xác định (2)
Gender	0, 1, 2	Cùng giới tính (1), khác giới tính (0), không xác định (2)
Apposition	0, 1	Là đồng vị ngữ (1), ngược lại (0)
Alias	0, 1	Là từ viết tắt hoặc cùng nghĩa (1), ngược lại (0)
Who	0, 1	Nếu hai khái niệm liền kề nhau và được phân cách bởi dấu ":"
Name match	0, 1	Loại bỏ các "stop word" (dr, dr., mr, ...), so trùng chuỗi con, trùng (1), không trùng (0)
Relative match	0, 1	Cả hai đều cùng chỉ đến một thân nhân (1), ngược lại (0)
Department match	0, 1	Cả hai cùng chỉ đến một lĩnh vực y học (1), ngược lại (0)
Doctor title match	0, 1	Cả hai có cùng một chức vụ bác sĩ (1), ngược lại (0)
Doctor general match	0, 1	Cả hai cùng đề cập đến bác sĩ nói chung (1), ngược lại (0)
Twin/triplet	0, 1	Cả hai đều chỉ về cùng cặp sinh đôi/sinh ba (1), ngược lại (0)
We	0, 1	Cả hai đều chứa thông tin về "chúng tôi" (1), ngược lại (0)
You	0, 1	Cả hai đều chứa thông tin về "tôi" (1), ngược lại (0)
I	0, 1	Cả hai đều chứa thông tin về "bạn" (1), ngược lại (0)
Pronoun match	0, 1	Cả hai đều là đại từ chỉ người (1), ngược lại (0)

thấy việc xác định một khái niệm thuộc lớp Person hay không có thể đạt được bằng cách xác định tập từ khóa chỉ về bệnh nhân như "patient", "pt", ... và tập từ khóa chỉ về nhóm người không phải bệnh nhân như "doctor", "dr", "wife", ...

Vì tập dữ liệu không có thông tin xác định một khái niệm thuộc lớp Person có phải là bệnh nhân hay không, dựa theo hệ thống I chúng tôi xác định bằng cách chọn chuỗi đồng tham chiếu có nhiều khái niệm nhất trong tập kết quả làm chuỗi đồng tham chiếu chỉ bệnh nhân. Các khái niệm thuộc chuỗi đồng tham chiếu này sẽ được xem là khái niệm chỉ bệnh nhân và được chọn làm mẫu dương trong quá trình huấn luyện. Các khái niệm thuộc lớp Person còn lại không thuộc vào chuỗi đồng tham chiếu này sẽ được chọn làm mẫu âm trong quá trình huấn luyện. Tuy nhiên, chúng tôi nhận thấy phương pháp xác định bệnh nhân này có một nhược điểm là các BADT nhỏ, có nội dung ngắn sẽ tồn tại nhiều chuỗi đồng tham chiếu lớp Person có kích thước tương tự nhau. Trong đó chuỗi đồng tham chiếu chỉ bệnh nhân không chắc chắn là chuỗi đồng tham chiếu có kích thước lớn nhất.

Bảng 4.3 trình bày đầy đủ các đặc trưng được sử dụng cho việc xác định khái niệm có phải là bệnh nhân hay không. Kết quả của việc phân loại này sẽ được sử dụng làm giá trị cho đặc trưng "Patient-class" khi rút trích đặc trưng cho lớp Person.

Bảng 4.3: Tập đặc trưng cho lớp Patient

Đặc Trưng	**Giá trị**	**Giải thích**
Keyword of patient	0, 1	Các từ khóa về bệnh nhân (như mr., mr, ms., ms, yo-, y.o., y/o, year-old, ...)
Keyword of doctor	0, 1	Các từ khóa về bác sĩ (dr, dr., md, m.d., m.d,. . .)
Key word of doctor title	0, 1	Các từ khóa về chức vụ của bác sĩ (dentist, orthodontist, . . .)
Key word of department	0, 1	Các từ khóa về chuyên ngành bác sĩ (electrophysiology, . . .)
Key word of general deparment	0, 1	Các từ khóa chung về phòng ban (team, service)
Key word of general doctor	0, 1	Các từ khóa chung về bác sĩ (doctor, dict, author, pcp, attend, provider)
Key word of relative	0, 1	Các từ khóa về người thân (wife, brother, sibling, nephew)
Name	0, 1	Có phải là tên riêng hay không
Last n line doctor	0, 1	Là tên bác sĩ ở n dòng cuối cùng
Twin or triplet information	0, 1	Thông tin về cặp sinh đôi, sinh ba (baby 1, 2, 3,. . .)
Preceded by non-patient	0, 1	Khái niệm đứng trước không phải là bệnh nhân.
Signed information	0, 1	Có liên quan đến việc kí/xác nhận bệnh án
Previous sentence		Câu hoàn chỉnh liền trước khái niệm
Next sentence		Câu hoàn chỉnh liền sau khái niệm
Pronouns we	0, 1	Là đại từ chỉ chúng tôi (we, us, our, ourselves)
Pronouns I	0, 1	Là đại từ chỉ tôi (I, my, me, myself)
Pronouns you	0, 1	Là đại từ chỉ bạn (you, your, yourself)
Pronouns they	0, 1	Là đại từ chỉ họ (they, them, their, themselves)
Pronouns he/she most	0, 1	Thuộc phần đa số của đại từ chỉ cô ấy/anh ấy (he, his, her)
Who	0, 1	Là đại từ "who" hoặc liền kề với khái niệm đứng trước
Appositive	0, 1	Là đồng vị ngữ

Các đặc trưng về từ khóa được chúng tôi hiện thực bằng cách khảo sát tập dữ liệu và xây dựng bộ từ điển thích hợp cho từng đặc trưng.

Các đặc trưng "Previous sentence" và "Next sentence" được hiện thực bằng cách khảo sát toàn bộ các khái niệm thuộc lớp Person, sau đó xây dựng bộ từ điển các câu có thể đứng trước hoặc đứng sau khái niệm đang xét. Giá trị của đặc trưng được lấy bằng chỉ mục của câu đứng trước (hoặc đứng sau) trong bộ từ điển các câu.

Đặc trưng "Pronouns he/she most" mang ý nghĩa giới tính chiếm đa số trong BAĐT được xét. Việc xác định giới tính chiếm đa số trong BAĐT được hiện thực bằng cách xác định giới tính cho từng khái niệm thuộc lớp Person, sau đó chọn giới tính có số lượng khái niệm lớn hơn. Phương pháp xác định giới tính được thực hiện theo miêu tả trong đặc trưng của nhóm Person. Nếu trong BAĐT có giới tính Nam chiếm đa số thì những khái niệm là đại từ chỉ về giới tính Nam như "he", "him", "himself", ... sẽ có đặc trưng "Pronouns he/she most" mang giá trị là 1. Tương tự cho BAĐT có giới tính Nữ chiếm đa số.

Nhóm Pronoun

Nhóm Problem/Test/Treatment

4.6 Gom cụm và xây dựng chuỗi đồng tham chiếu

Ở mô hình cặp thực thể, hệ thống phân loại không có khả năng xây dựng chuỗi đồng tham chiếu mà nó chỉ có thể xác định một cặp khái niệm là có đồng tham chiếu hay không. Mặt khác, đối với một văn bản HSXV, số cặp khái niệm được sinh ra rất nhiều và trong số đó có nhiều cặp có chung khái niệm đứng sau, ví dụ hai cặp "Dr. John"-"his" và "Mr. Brown"-"his" có chung khái niệm đứng sau là "his" mà hai cặp này đều được hệ thống phân loại xác định là đồng tham chiếu, tuy nhiên chỉ một trong hai khái niệm "Dr. John" và "Mr. Brown" được chọn làm tiền đề cho khái niệm "his" này. Như vậy cần thiết phải có một giải thuật lựa chọn các cặp đồng tham chiếu và xây dựng các chuỗi đồng tham chiếu từ chúng.

Như đã được đề cập ở mục, có hai giải thuật được đề xuất là: *gom cụm gần nhất trước* và *gom cụm tốt nhất trước*. Chúng tôi lựa chọn thực hiện giải thuật gom cụm tốt nhất trước cho hệ thống của mình vì hai lý do:

1. Theo [x], giải thuật gom cụm tốt nhất trước cho kết quả tốt hơn giải thuật gom cụm gần nhất trước.

2. Các tác giả hệ thống [I] cũng hiện thực giải thuật này cho hệ thống của họ.

Về cơ bản, giải thuật gom cụm tốt nhất trước lựa chọn các cặp khái niệm được xác định là đồng tham chiếu và có độ tin cậy cao nhất ứng với mỗi hồi chỉ; đối với đại từ, giải thuật sử dụng module phân loại xác định lớp chính của đại từ đó và tạo một cặp đồng tham chiếu giữa nó với khái niệm thuộc lớp tương ứng gần nhất trước đó (theo thứ tự xuất hiện) trong văn bản HSXV. Sau khi có được tập các cặp đồng tham chiếu, giải thuật nối các cặp có chung một khái niệm lại để tạo thành các chuỗi đồng tham chiếu. Đây chính là kết quả cuối cùng của hệ thống phân giải.

Chi tiết thuật toán gom cụm tốt nhất trước, BEST-CLUSTER, được trình bày ở Hình 4.4. BEST-CLUSTER nhận đầu vào là một văn bản HSXV, danh sách các khái niệm đã được gán nhãn của văn bản này và xuất ra danh sách các chuỗi đồng tham chiếu của nó.

Algorithm 1: BEST-CLUSTER

Đầu vào: văn bản HSXV E, danh sách khái niệm C
Đầu ra : danh sách chuỗi đồng tham chiếu của E

```
1  M: số khái niệm;
2  pairs ← ∅;
3  for i ← 1 to M do
4      ana ← C[i];
5      if TypeOf(ana) ≠ PRONOUN then
6          bestConf ← null;
7          bestAnte ← null;
8          for j ← i − 1 downto 0 do
9              ante ← C[j];
10             if TypeOf(ante) ≠ TypeOf(ana) then continue;
11             f ← ExtractFeature(E, ante, ana);
12             r ← Classify(f);
13             if IsCoref(r) then
14                 if bestConf = null or bestConf < Confidence(r) then
15                     confidence ← Confidence(r);
16                     bestAnte ← ante;
17                 end
18             end
19         end
20         if bestConf ≠ null and bestAnte ≠ null then
21             pairs ← pairs ∪ (bestAnte, ana);
22         end
23     else
24         f ← ExtractFeature(E, ana);
25         r ← Classify(f);
26         type ← TypeOf(r);
27         for j ← i − 1 downto 0 do
28             ante ← C[j];
29             if TypeOf(ante) = type then
30                 pairs ← pairs ∪ (ante, ana);
31                 break;
32             end
33         end
34     end
35 end
```

Hình 4.4: Giải thuật gom cụm tốt nhất trước

Chương 5

Thí nghiệm đánh giá

5.1 Tập dữ liệu

5.2 Các hệ đo

5.3 Kết quả

Chương 6

Tổng kết

Tài liệu tham khảo

[1] Y. Xu, "A classification approach to coreference in discharge summaries: 2011 i2b2 challenge," *Journal of the American Medical Informatics Association : JAMIA*, vol. 19, no. 5, pp. 897–905, 2012.

[2] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A machine learning approach to coreference resolution of noun phrases," *Computational Linguistics*, vol. 27, pp. 521–544, December 2001.