**ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC BÁCH KHOA**

KHOA KHOA HỌC VÀ KĨ THUẬT MÁY TÍNH





**BÁO CÁO THỰC TẬP TỐT NGHIỆP**

Phân giải đồng tham chiếu cho bệnh án điện tử

|  |  |
| --- | --- |
| **Giáo viên hướng dẫn:** | **Sinh viên thực hiện:** |
| Cao Hoàng Trụ | Nguyễn Duy Hưng – 51101475  Vương Anh Tuấn – 51104040 |
| 03-Jun-15 | |

**Mục lục**

[1 Giới thiệu vấn đề 5](#_Toc420781480)

[2 Các công trình liên quan 5](#_Toc420781481)

[2.1 Bệnh án điện tử 5](#_Toc420781482)

[2.2 Phân giải đồng tham chiếu 7](#_Toc420781483)

[2.3 Phân giải đồng tham chiếu cho bệnh án điện tử 8](#_Toc420781484)

[2.4 Named-Entity-Regconition 9](#_Toc420781485)

[2.4.1 Định nghĩa nhãn 9](#_Toc420781486)

[3 Kiến thức và công nghệ 10](#_Toc420781487)

[3.1 Những vấn đề trong phân giải đồng tham chiếu trong bệnh án điện tử 10](#_Toc420781488)

[4 Bài toán đề xuất 10](#_Toc420781489)

[4.1 Phạm vi đề tài 10](#_Toc420781490)

[4.1.1 Nội dung bài toán 10](#_Toc420781491)

[4.1.2 Dữ liệu đầu vào 10](#_Toc420781492)

[4.1.3 Kết quả đầu ra 11](#_Toc420781493)

[4.2 Thiết kế hệ thống 11](#_Toc420781494)

[4.2.1 Tiền xử lý 12](#_Toc420781495)

[4.2.2 Xây dựng các cặp khái niệm 12](#_Toc420781496)

[4.2.3 Các phương pháp phân loại 13](#_Toc420781497)

[4.2.4 Best-first clustering 16](#_Toc420781498)

[4.2.5 Xây dựng chuỗi đồng tham chiếu 16](#_Toc420781499)

[5 Tập dữ liệu và phương pháp đánh giá 16](#_Toc420781500)

[5.1 Tập dữ liệu 16](#_Toc420781501)

[5.2 Phương pháp đánh giá 16](#_Toc420781502)

[5.2.1 Hệ đo MUC 17](#_Toc420781503)

[5.2.2 Hệ đo B-CUBED 17](#_Toc420781504)

[5.2.3 Hệ đo CEAF 17](#_Toc420781505)

[6 Kết luận 18](#_Toc420781506)

[7 Tài liệu tham khảo 18](#_Toc420781507)

# Giới thiệu vấn đề

Bệnh án điện tử (Electronic Medical Record) là hệ thống lưu trữ hầu hết thông tin của bệnh nhân trong các bệnh viện. Trong hơn mười năm trở lại đây, bệnh án điện tữ đã thu thập được một lượng lớn dữ liệu liên quan đến các bệnh, triệu chứng, cách điều trị … Lượng dữ liệu to lớn và có giá trị này trở nên vô cùng hữu ích cho việc nghiên cứu về y tế cho cả ngành y tế lẫn ngành khoa học máy tính.

Đến nay, các nghiên cứu đã đạt được nhiều kết quả. Đặc biệt từ những nước nói tiếng Anh. Tổng thống mỹ Barack Obama trong kế hoạch hồi sinh nền kinh tế đã yêu cầu việc chăm sóc y tế dựa trên việc chuẩn hóa và số hóa mọi bệnh án trong bệnh viện, chính phủ Nhật Bản có chính sách ưu tiên phát triển bệnh án điện tử cho các bệnh viện vừa và lớn. Bệnh án điện tử được lưu trữ dưới hai dạng: có cấu trúc và không có cấu trúc. Đa phần những dữ liệu hữu ích đều được lưu trữ ở dạng không có cấu trúc như: hồ sơ bệnh án, phiếu điều trị, phiếu chăm sóc,…Vì thế, việc trích xuất thông tin từ những dữ liệu không có cấu trúc là bước cơ bản và quan trọng cho các nghiên cứu sau.

Đối với Việt Nam, bệnh án điện tử có thể chia làm 2 loại: loại cơ bản và loại hoàn chỉnh. Loại cơ bản chỉ bao gồm các thông tin quản lý cơ bản, trong kho loại hoàn chỉnh được bao gồm thêm các văn bản lâm sàng như phiếu điều trị, ghi chép các chẩn đoán, nhận định, phiếu chăm sóc, hồ sơ suất viện,.. và các thông tin chữa bệnh khác. Tuy nhiên, hầu hết các bệnh viện không được trang bị bệnh án điện tử hoặc chỉ trang bị bệnh án điện tử loại cơ bản. Lưu ý trong số đó, Bệnh viện đa khoa Vân Đồn tỉnh Quảng Ninh là cơ quan y tế đầu tiên có trang bị hệ thống bệnh án điện tử hiện đại và hoàn chỉnh với hệ thống MEDI SOLUTIONS.

Cùng với việc hệ thống bệnh án điện tử đang dần được xây dựng tại Việt Nam, tập thể nghiên cứu “Học máy và ứng dụng” của viện John von Neumann thuộc đại học Quốc Gia TP Hồ Chí Minh đã tiến hành phát triển phương pháp và phần mềm công cụ cho việc khai thác bệnh án điện tử tiếng Việt. Việc xây dựng kế hoạch nghiên cứu này không chỉ bắt nguồn từ nhận thức về vai trò thiết yếu của bệnh án điện tử mà còn từ việc đánh giá các điều kiện để thực hiện việc khai thác bệnh án điện tử tại nước ta.

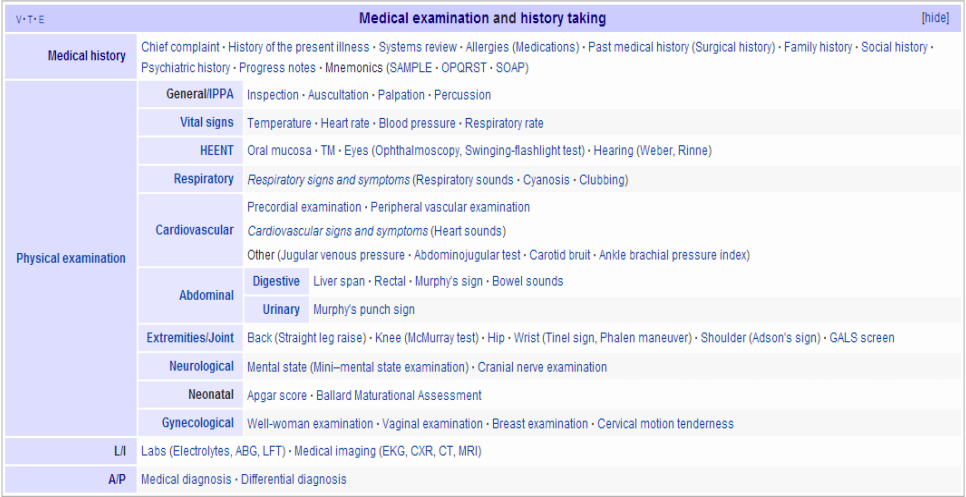
Bên cạnh đó, vấn đề đồng tham chiếu cũng là một bài toán nổi bật đang thu hút được sự quan tâm của cộng đồng khi khai thác bệnh án điện tử. Đồng tham chiếu là khi hai hay nhiều khái niệm trong văn bản cùng ám chỉ tới một người hoặc một vật. Việc phân giải đồng tham chiếu các khái niệm trong văn bản là xác định liệu hai khái niệm trong cùng văn bản có ám chỉ tới cùng một người hoặc một vật hay không, từ đó xây dựng các chuỗi đồng tham chiếu. Vấn đề phân giải đồng tham chiếu đã nhận được nhiều sự quan tâm từ cộng đồng, tuy nhiên việc áp dụng phân giải đồng tham chiếu để rút trích thông tin trong bệnh án điện tử chưa được đào sâu nghiên cứu.

Đứng trước nhu cầu đó, nhóm quyết định bắt tay vào phát triển một hệ thống phân giải đồng tham chiếu các khái niệm trong bệnh án tiếng Anh và tiếng Việt.

# Các công trình liên quan

## Bệnh án điện tử

Bệnh án điện tử là hệ thống lưu trữ các thông tin sức khỏe của một cá nhân hoặc một cộng đồng. Bệnh án đó được lưu trữ dưới dạng số và có thể được chia sẻ giữa các trung tâm y tế. Việc chia sẻ có thể được thực hiện thông qua mạng kết nối hoặc hệ thống thông tin của trung tâm. Bệnh án điện tử bao gồm lịch sử khám bệnh, dị ứng, thuốc áp dụng, tình trạng sức khỏe , cân nặng, chiều cao, lí do vào viện kết quả xét nghiệm và nhiều thông tin y tế khác.

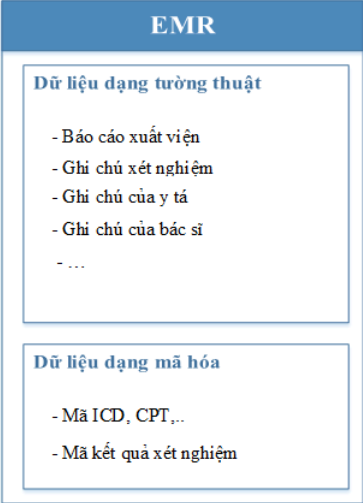


So với bệnh án được lưu trữ bằng giấy, bệnh án điện tử có nhiều ưu điểm như

* Lưu trữ chính xác và đầy đủ thông tin bệnh nhân, tránh trùng lặp dữ liệu
* Hỗ trợ quá trình tìm kiếm và truy xuất thông tin nhanh chóng
* Dữ liệu có thể được chia sẽ hoặc tích hợp

Dữ liệu trong bệnh án điện tử thường tồn tại dưới 2 dạng: có cấu trúc và không có cấu trúc. Dữ liệu không có cấu trúc thường tồn tại dưới dạng tường thuật và đây cũng là dạng có nhiều thông tin y tế nhất

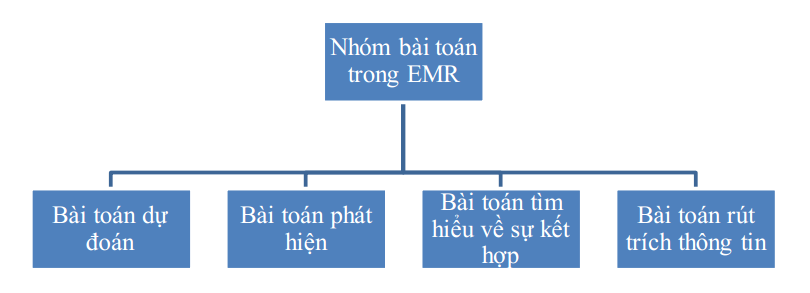
* Dạng không có cấu trúc: các ghi chú tường thuật của bác sĩ, y tá trong quá trình điều trị, các báo cáo xuất viện
  + Lý do nhập viện, tiền sử bệnh
  + Các ghi chú trong phiếu chăm sóc và phiếu điều trị được ghi chép hàng ngày
  + Các ghi chú kết quả xét nghiệm
  + Báo cáo xuất viện
  + ….
* Dạng có cấu trúc: các thông tin được mã hóa theo một tiêu chuẩn nhất định nào đó



## Các hướng nghiên cứu trong bệnh án điện tử

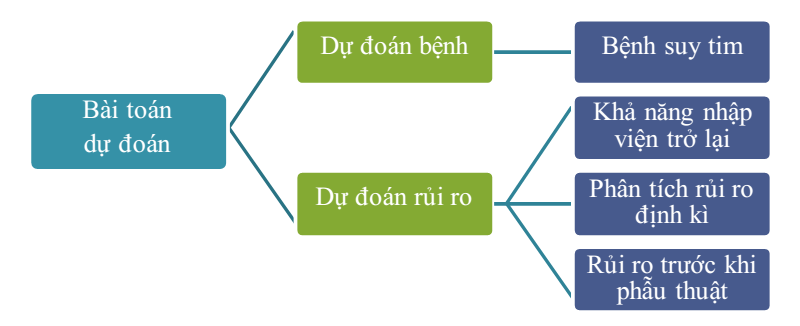
**Các nhóm bài toán trong bệnh án điện tử**

Với sự phát triển mạnh mẽ của bệnh án điện tử, nhiều hướng nghiên cứu khai thác bệnh án điện tử đã được đưa ra để sử dụng tối đa nguồn thông tin đã có. Tuy vậy, ta có thể phân chia thành 4 hướng nghiên cứu chính như sau: bài toán dự đoán bệnh hoặc các biến cố trong y học, bài toán phát hiện bệnh, bài toán tìm hiểu sự kết hợp của các thực thể trong y tế như mối quan hệ giữa bệnh-bệnh, giữa bệnh-triệu chứng, giữa bệnh-thuốc điều trị, …., và bài toán rút trích thông tin. Đặc biệt bài toán rút trích thông tin tập trung xử lý các dữ liệu không có cấu trúc trong bệnh án điện tử như chẩn đoán của bác sĩ, ghi chú chăm sóc của y tá, báo cáo xuất việnm,… để nhận diện các thực thể, tìm mối quan hệ và trích xuất các thông tin theo yêu cầu.



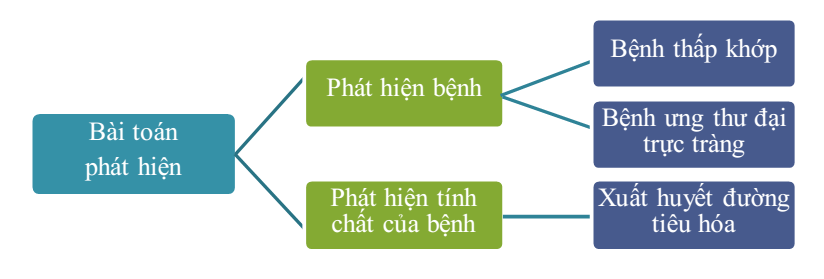
**Bài toán dự đoán**

Bài toán dự đoán là những bài toán giúp đoán trước khả năng mắc bệnh hoặc những biến cố y tế có thể xảy ra với bệnh nhân trong tương lai. Việc dự đoán được dựa trên những thông tin có được tại hiện hoặc quá khứ của bệnh nhận được xét từ đó giúp phòng tránh hoặc có những biện pháp thích hợp. Bài toán được chia làm 2 nhánh chính là: dự đoán bệnh và dự đoán những rủi ro có thể xảy ra.



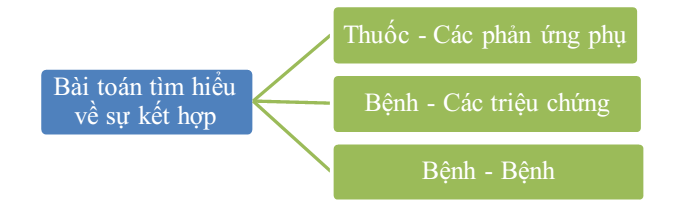
**Bài toán phát hiện**

Bài toán phát hiện là những bài toán giúp tìm ra được những vấn đề có thể tồn tại trên bệnh nhân mà chưa được nhận ra hoặc phát hiện. Bài toán phát hiện có 2 nhánh chính là: phát hiện bệnh tiềm ẩn và phát hiện tính chất của bệnh đang mắc phải



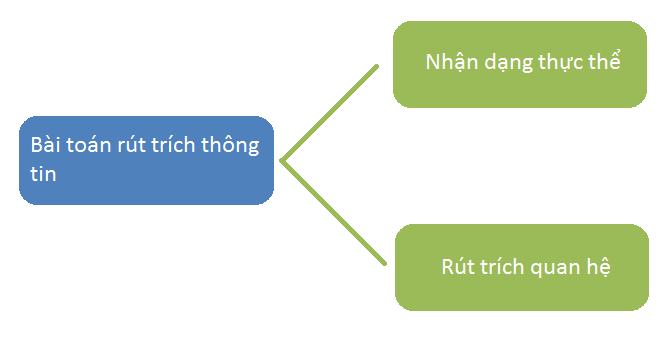
**Bài toán tìm hiểu sự kết hợp**

Bài toán tìm hiểu sự kết hợp là những bài toán cho biết mối liên hệ giữa các thực thể trong y học, thường được áp dụng cho việc ra quyết định hoặc phát hiện tác dụng phụ của thuốc. Nhóm bài toán gồm các nhánh chính là: sự kết hợp của thuốc – phản ứng phụ, bệnh – triệu chứng, bệnh – bệnh



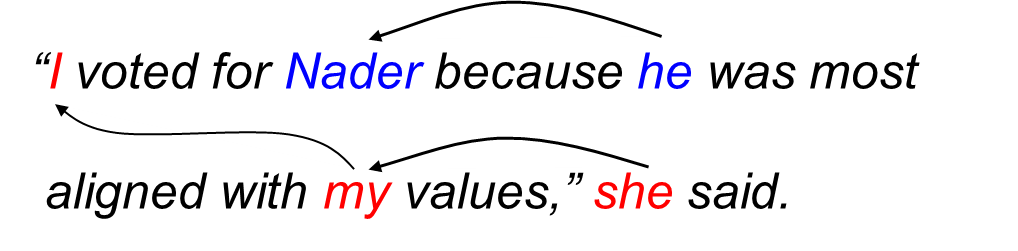
**Bài toán rút trích thông tin**

Bài toán rút trích thông tin là nhóm bài toán giúp trích xuất các thông tin đặc thù, thực thể, tham chiếu trong phần dữ liệu không có cấu trúc của bệnh án điện tử. Đầu vào của bài toán là các báo cáo y khoa, trong khi đó đầu ra của bài toán là các dữ liệu đã được định dạng theo một cấu trúc nhất định và có thể được sử dụng bởi máy tính để phục vụ cho các bài toán khác. Nhóm bài toán này có các nhánh chính là: nhận dạng thực thể và rút trích quan hệ



## Phân giải đồng tham chiếu

Phân giải đồng tham chiếu là công việc xác định các khái niệm trong ngôn ngữ cùng ám chỉ, tham chiếu tới một thực thể trong thế giới thật. Một ví dụ như sau



Trong ví dụ trên, từ I, my, she cùng ám chỉ tới một thực thể là con người trong thế giới thật. Từ Nader, he ám chỉ tới một thực thể người khác. Từ đó ta có 2 chuỗi đồng tham chiếu là (I, my, she) và (Nader, he). Phân giải đồng tham chiếu là công việc tìm ra các chuỗi khái niệm đó.

Bài toán phân giải đồng tham chiếu đã được cộng đồng quan tâm nghiên cứu trên 15 năm. Sau nhiều năm nghiên cứu, 3 mô hình quan trọng đã được đưa ra để giải quyết bài toán: mô hình mention-pair, mô hình entity-mention, mô hình ranking

**Mô hình mention-pair**

Mô hình mention-pair chú trong việc phân loại đánh giá 2 khái niệm được đưa ra có cùng tham chiếu tới một thực thể trong thế giới thật hay không. Nhiều chiến lược khác nhau đã được đề xuất trong việc xây dựng một tập dữ liệu trainning.

Phương pháp đơn giản nhất là kiến tạo toàn bộ cặp khái niệm từ tất cả các khái niệm có sẵn trong văn bản. Tuy nhiên cách làm này tạo ra quá nhiều cặp dữ liệu negative và có thể dẫn tới thành kiến cho bộ phân loại. Để giảm sự thành kiến bị gây ra do có quá nhiều dữ liệu negative, một số công trình chỉ sử dụng các cặp khái niệm negative chen giữa các cặp khái niệm positive trong văn bản. Một số công trình khác lại lọc bớt các cặp khái niệm để làm giảm số lượng các cặp khái niệm được sinh ra. Một ví dụ như khái niệm thuộc lớp Person sẽ không bao giờ đồng tham chiếu với khái niệm thuộc lớp Place, vì vậy ta có thể loại bỏ tất cả các cặp khái niệm không thuộc cùng một lớp.

Vấn đề lớn thứ 2 trong mô hình là làm thể nào để đánh giá 2 khái niệm có cùng tham chiếu hay không. Các biện pháp thông thường sử dụng bộ phân loại lớp (classification resolver), bộ phân cụm (clustering resolver) hoặc kết hợp cả 2 bộ phân loại lớp và cụm.

**Mô hình entity-mention**

Mô hình entity-mention lại xem xét bài toán phân giải đồng tham chiếu như một bài toán gom cụm, thay vì phân loại một cặp khái niệm có đồng tham chiếu hay không. Mô hình entity-mention xác định một khái niệm đang được xem xét có thuộc về các cụm khái niệm có trước hay không. Có 3 khả năng khi xem xét khái niệm thuộc về cụm khác niệm đã có trước: liên quan toàn bộ (all relevance), không liên quan (any relevance), liên quan một phần (most relevance)

**Mô hình ranking**

Mô hình ranking cố gắng tìm các khái niệm tốt nhất đã được đề cập ở phần trước của văn bản cho khái niệm đang được xem xét.

## Phân giải đồng tham chiếu cho bệnh án điện tử

Dữ liệu không có cấu trúc của bệnh án điện tử là một nguồn thông tin phong phú có thể bổ sung được cho phần dữ liệu có cấu trúc. Việc áp dụng các biện pháp xử lý ngôn ngữ tự nhiên không chỉ giúp ta lấy được thông tin từ dữ liệu không có cấu trúc, mà còn giúp cải thiện việc sử dụng phần dữ liệu có cấu trúc của bệnh án điện tử.

Bài toán phân giải đồng tham chiếu đã được phát triển lâu trong công đồng, tuy nhiên việc áp dụng cho lĩnh vực y tế nói chung và bệnh án điện tử nói riêng chưa được quan tâm đào sâu nghiên cứu. Phân giải đồng tham chiếu cho bệnh án điện tử đóng một vai trò then chốt trong việc kết nối các thông tin liên quan lại với nhau, giúp cho việc khai thác và sử dụng thông tin trong bệnh án điện tử tốt hơn.

Trong thử thách 2011 i2b2/VA, có 3 hướng tiếp cận chính để giải quyết bài toán phân giải đồng tham chiếu cho lĩnh vực bệnh án điện tử là: hệ thống rule-base, hệ thống học có giám sát, hệ thống lai

## Named-Entity-Regconition

Rút trích thông tin gồm 2 bước con là nhận dạng thực thể và rút trích quan hệ. Nhận dạng thực thể là bước đầu tiên của IE. Nhiệm vụ của nhận dạng thực thể là nhận dạng ra những thực thể trong câu và gắn nhãn cho nó. Đầu ra (output) của bước nhận dạng thực thể là:

* “Anh Tuấn” – Person
* “Duy Hưng” – Person
* “Bách Khoa” - Organization

Tùy theo mỗi lĩnh vực quan tâm cụ thể, các loại thực thể sẽ được định nghĩa khác nhau. Với những vấn đề không đặc thù, những nhóm thực thể thường được nhắc đến như: động vật, người, tổ chức, vật, … Khi nghiên cứu về nhận dạng thực thể trong bệnh án điện tử, i2b2 2011 đã định nghĩa 5 loại thực thể cần được quan tâm. Đó là vấn đề (problems), phương pháp điều trị (treatment), các xét nghiệm (test), con người (person) và đại từ (pronoun).

### Định nghĩa nhãn

Các lớp được quy định trong thử thách i2b2/VA 2011 như sau:

|  |  |  |
| --- | --- | --- |
| **Lớp** | **Ví dụ** | **Định nghĩa** |
| Person | Dr.Lightman, the patient, cardiology,.. | Những chủ thể người hoặc một nhóm người được để cập trong bệnh án. Có thể là bệnh nhân, người nhà hoặc nhân viên y tế |
| Problem | Heart attack, blood pressure, cancer, ... | 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 |
| Treatment | Surgery, ice pack, Tylenol,… | 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 |
| Test | CT scan, Temperature,…. | 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” |
| Pronoun | He, she, it, that,… | Những đại từ có thể tham chiếu đến bất kì lớp nào trong bốn lớp kể trên |

# Kiến thức và công nghệ

## Những vấn đề trong phân giải đồng tham chiếu trong bệnh án điện tử

* Đưa ra các key observation trong bài báo
* Đưa ra các feature design
* Nói rõ về coreference là gì và phân giải nó là như thế nào

# Bài toán đề xuất

## Phạm vi đề tài

### Nội dung bài toán

Dựa vào thử thách 1C trong các 3 thử thách cần giải quyết của website I2B2/VA 2011 (Informatics for Intergrating Biology and the Bedside), nhóm quyết định đề xuất bài toán “**Phân giải đồng tham chiếu trên bệnh án điện tử với các khái niệm đã được biết trước cho dữ liệu tiếng Anh và tiếng Việt**”.

Nội dung bài toán bao gồm việc đánh dấu các khái niệm đã được cho trước để biết được các khái niệm đó có đồng tham chiếu với nhau hay không. Từ đó xây dựng các chuỗi đồng tham chiếu của các khái niệm đó.

### Dữ liệu đầu vào

Dữ liệu đầu vào của nhóm gồm 2 phần:

1. Bệnh án điện tử dưới dạng text, có hoặc không có định dạng
2. Danh sách các khái niệm được đề cập trong bệnh án kèm theo lớp của khái niệm đó

**Bệnh án điện tử**

Tập hợp các thông tin về sức khỏe của một cá nhân đã được số hóa. Bệnh án điện tử có thể có định dạng tùy theo bệnh viên lưu trữ bệnh án đó.

Đầu vào của bài toán đề xuất là bệnh án điện tử được viết dưới ngôn ngữ là tiếng anh và được lưu trữ dưới dạng file txt

**Danh sách khái niệm**

Là file chứa danh sách toàn bộ các khái niệm được đề cập trong 1 một bệnh án cho trước. Mỗi file sẽ được đính kèm theo bệnh án.

Các khái niệm được đề cập sẽ kèm theo thông tin như vị trí bắt đầu, vị trí kết thúc và lớp phân loại của khái niệm đó

Được viết theo một định dạng quy định trước như sau:

*c=”<mention>” <begin> <end> || t=”<class>”*

Một số ví dụ như sau:

*c="which" 20:5 20:5||t="pronoun"*

Ví dụ trên được hiểu là khái niệm “which” xuất hiện tại dòng 20 từ thứ 5, kết thúc tại dòng 20 từ thứ 5 và được phân vào lớp Pronoun

### Kết quả đầu ra

Là chuỗi đồng tham chiếu của các khái niệm đã được đề cập ở trên

Được viết dưới một định dạng quy định trước như sau:

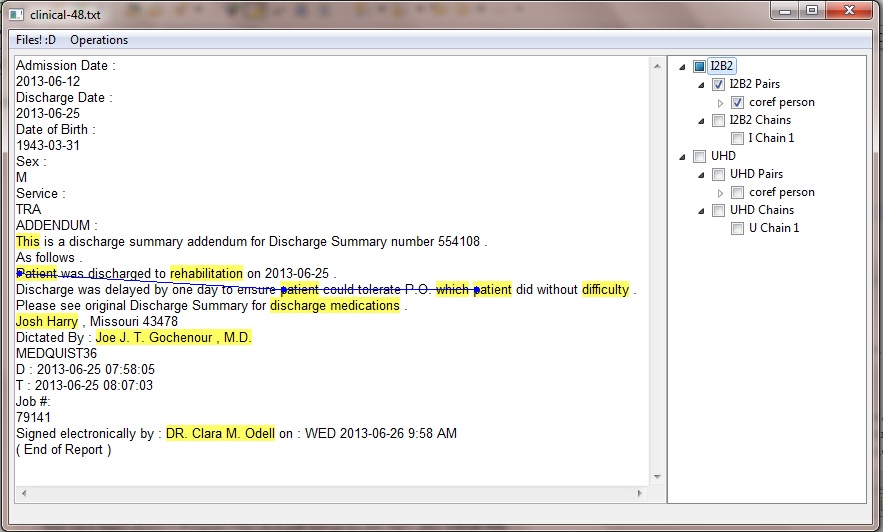
*c=”<mention>” <begin> <end> || c=….. || t=”<class>”*

Một số ví dụ như sau:

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

Ví dụ trên được hiểu là 2 khái niệm “the patient” và “he” đồng tham chiếu tới cùng 1 người

Dữ liệu đầu ra có thể được trực quan hóa để người dùng dễ nắm bắt



## Thiết kế hệ thống

Ở phần này, nhóm sẽ trình bày một cách chi tiết các bước hiện thực hệ thống phân giải đồng tham chiếu cho bệnh án điện tử. Một cách tổng quan, hệ thống sẽ gồm các bước sau (Hình 1): tiền xử lý, sinh các cặp khái niệm và lọc ra những cặp có khả năng là đồng tham chiếu, sử dụng SVM để học trên tập các cặp khái niệm để đánh giá mức độ tin cậy về việc đồng tham chiếu của chúng, áp dụng giải thuật gom cụm best-first để xác định các cặp đồng tham chiếu và cuối cùng là xây dựng chuỗi đồng tham chiếu từ các cặp đồng tham chiếu đó.



Hình . Sơ đồ khối

### Tiền xử lý

Ở bước tiền xử lý này, các khái niệm (concept) ở dữ liệu đầu vào sẽ được loại bỏ đi các bổ từ xung quanh nó (nếu có). Mục đích là để so trùng chuỗi kí tự giữa các cặp danh từ, giữa các danh từ và các kiến thức nền (Wikipedia), sử dụng làm thuộc tính cho hệ thống phân loại. Ví dụ cụm từ “her CT scan” và “a CT scan” sau khi qua bước tiền xử lý đều trở thành “CT scan”. Đối với các cụm từ có chứa giới từ, giới từ cùng với nội dung phía sau nó sẽ được loại bỏ.

### Xây dựng các cặp khái niệm

Từ danh sách khái niệm ở dữ liệu đầu vào, cặp khái niệm sẽ được xây dựng. Như đã được đề cập ở phần trước, loại bỏ đi các cặp ít có khả năng là đồng tham chiếu sẽ tránh đi ảnh hưởng tiêu cực của chúng lên hệ thống phân loại. Một số phương pháp được đề xuất để làm điều này, nhóm quyết định sẽ chọn phương pháp đơn giản nhất, đó là loại bỏ đi các cặp mà hai khái niệm thuộc về hai lớp khác nhau.

### Các phương pháp phân loại



Hình . Sơ đồ khối

Các khái niệm được chia làm năm lớp: Person, Problem, Treatment, Test và Pronoun. Mỗi mối quan hệ đồng tham chiếu được phân vào một trong các lớp trên, trừ Pronoun. Một đặc tính nổi bật của một mối quan hệ tham chiếu thuộc lớp Person là các khái niệm tham gia vào quan hệ đó có thể là một trong rất nhiều các đại từ nhân xưng (he, she, it, they, …), đại từ sở hữu (his, her, its, their, …) hoặc đại từ phản thân (himself, herself, itself, themselves, …). Việc phân giải đồng tham chiếu cho tên người và đại từ nói chung là một công việc khó, vì thông tin có được từ các đại từ là rất ít, chúng chỉ có thể cho ta biết về số lượng (số ít hay nhiều) hay ngôi thứ (ngôi thứ nhất, thứ hai), v.v… Mặt khác, các tài liệu thường chứa nhiều sự đề cập đến nhiều hơn một người khiến cho việc phát hiện đúng chuỗi đồng tham chiếu cho các đề cập này là một thách thức lớn. Tuy nhiên, nếu chúng ta chỉ giới hạn lại trong phạm vi bệnh án điện tử thì công việc này sẽ dễ hơn rất nhiều. Một bệnh án điện tử thông thường chỉ đề cập đến một bệnh nhân, và nếu một khái niệm được phát hiện là một sự đề cập đến bệnh nhân thì khái niệm đó gần như chắc chắn thuộc vào chuỗi đồng tham chiếu duy nhất đến bệnh nhân đó. Do vậy, việc xác định xem một khái niệm có phải là một sự đề cập đến bệnh nhân hay không là một công việc cực kì quan trọng trong phạm vi bệnh án điện tử.

Đối với lớp Problem/Test/Treatment, mặc dù cùng một sự kiện y khoa có thể xảy ra nhiều lần nhưng chúng không đồng tham chiếu mà mang nhiều ý nghĩa khác nhau. Việc xây dựng chính xác chuỗi đồng tham chiếu của nhóm lớp này cần nhiều gợi ý ngữ nghĩa từ ngữ cảnh trong văn bản.

Đối với lớp Pronoun, việc quan trọng nhất là xác định Pronoun được xét thuộc lớp ngữ nghĩa nào (Person, Problem, Treatment, Test). Khi đã xác định được lớp, ta sẽ chọn khái niệm gần nhất cùng lớp được để cập để xác định là 2 khái niệm đồng tham chiếu. Mặc dù cách làm rất đơn giản, nhưng phương pháp này lại cho hiệu quả khá tốt. Điều này cho thấy, việc phân giải đồng tham chiếu trong bệnh án điện tử rất khác với vấn để phân giải đồng tham chiếu nói chung. Đối với các lớp Person và Problem/Treatment/Test, 1 cặp 2 khái niệm được xem xét cùng 1 lúc, trong khi đó với lớp Pronoun, 1 khái niệm được xem xét cùng lúc.

Đồng tham chiếu lớp Person

Như đã đề cập ở trên, các phương pháp phân giải đồng tham chiếu cho các tài liệu nói chung không thể được áp dụng trực tiếp cho bệnh án điện tử. Các hệ thống phân giải cho tài liệu là các bài báo cho rằng có nhiều hơn một người hay nhóm người được đề cập đến và họ đều đóng vai trò quan trọng như nhau trong bài. Tuy nhiên, ở phạm vi bệnh án điện tử, những cá nhân được đề cập đến thường chỉ thuộc một trong ba lớp: bệnh nhân, người thân của bệnh nhân và nhân sự của bệnh viện. Việc xác định xem một sự đề cập đến người (bao gồm tên và đại từ) thuộc lớp nào trong ba lớp trên đóng một vai trò quan trọng trong việc phân giải đúng chuỗi đồng tham chiếu cho sự đề cập đó. Do vậy, nhóm quyết định giới thiệu thêm thuộc tính Patient-class (được giải thích rõ hơn bên dưới). <B> trình bày tập các thuộc tính dùng cho lớp Person. Phương pháp học máy được sử dụng ở phần này là SVM.

| Thuộc tính | Giá trị | Giải thích |
| --- | --- | --- |
| Patient-class | 0, 1, 2 | Không khái niệm nào là bệnh nhân (0), cả hai là đều là bệnh nhân (1), khác (2) |
| Distance between sentences | 0, 1, 2, 3, … | Số câu xuất hiện giữa hai khái niệm |
| Distance between mentions | 0, 1, 2, 3, … | Số khái niệm xuất hiện giữa hai khái niệm của cặp |
| String match | 0, 1 | Trùng hoàn toàn (1), ngược lại (0) |
| Levenshtein distance between two mentions | (0, 1) | 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 | ??? |
| Name match | 0, 1 | Loại bỏ các “stop word” (dr, dr., mr, ms, mrs, md, m.d., m.d, “,”, m, m., :), 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 (1), ngược lại (0) |
| Doctor title match | 0, 1 | ??? |
| Doctor general match | 0, 1 | ??? |
| 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 | ??? |

Bảng . Các thuộc tính dùng cho phân giải đồng tham chiếu lớp Person

Thuộc tính Patient-class

Để xác định một khái niệm có đề cập đến bệnh nhân hay không, nhóm sử dụng SVM để học và phân loại chúng. Trong một bệnh án điện tử, thường chỉ có một bệnh nhân đóng vai trò là chủ thể của bệnh án. Như vậy nếu như một khái niệm được xác định là một sự đề cập đến bệnh nhân, thì khái niệm đó sẽ được đưa vào chuỗi đồng tham chiếu duy nhất về bệnh nhân đó.

Bằng cách xem xét kĩ dữ liệu, nhóm nhận thấy việc xác định xem một khái niệm có đề cập đến bệnh nhân hay không tương đối dễ thông qua một số từ khóa. Để huấn luyện hệ thống phân loại ở phần này, tất cả những khái niệm thuộc vào chuỗi đồng tham chiếu về bệnh nhân được lấy làm mẫu dương, và những khái niệm không thuộc vào chuỗi này là mẫu âm. Tập các thuộc tính được mô tả ở <B>. Kết quả của việc phân loại sẽ được lấy làm giá trị cho thuộc tính là-bệnh-nhân ở <B>.

Đồng tham chiếu các lớp Problem/Treatment/Test

Nhóm nhãn Prblem/Treatment/Test là nhóm nhãn đặc biệt của lĩnh vực y khoa. Trong lĩnh vực này, rất nhiều cụm từ khác nhau có thể ám chỉ cùng một khái niệm. Xác định các từ đồng nghĩa này có thể giúp giảm những sai sót và tăng độ chính xác cho quá trình học máy. Để tìm được các từ đồng nghĩa không có trong tập trainning, ta cần sử dụng nguồn thông tin có sẵn từ Wikipedia. Tuy nhiên trong mặt khác, nhiều khái niệm lại không đồng tham chiếu mặc dù chúng được viết giống nhau vì có ngữ cảnh khác nhau. Phân biệt các khái niệm này cũng có thể giúp giảm sai sót và tăng độ chính xác. Vì các lí do trên, nhóm quyết định đề xuất các bộ trích xuất ngữ nghĩa:

Kiến thức thức nền tảng: sử dụng nguồn thông tin có sẵn từ Wikipedia hoặc WordNet

Bộ trích xuất ngữ nghĩa: các khái niệm của lớp Problem/Treatment/Test cần được phân biệt dựa trên ngữ cảnh của tài liệu. Ví dụ như “Đau” ở đầu mặc dùng có cùng cách viết nhưng lại không đồng tham chiếu với “Đau” ở chân, 2 bài kiểm tra y khoa có giá trị kết quả khác nhau thường không đồng tham chiếu. Vì vậy, nhóm đề xuất một tập các bộ trích xuất ngữ nghĩa để giúp phân biệt các khái niệm có ngữ nghĩa hoặc vị trí thời gian khác nhau:

1. **Anatomy Extractor**:
2. **Position Extractor**: cùng một khái niệm có thể xuất hiện ở nhiều vị trí khác nhau: “Tay **trái**” và “Tay **phải**”. Trong đó cùng khái niệm là tay nhưng lại có 2 vị trí khác nhau nên 2 khái niệm này không thể đồng tham chiếu.
3. **Medication information Extractor**: bộ trích xuất này làm việc với các khái niệm thuốc, liều lượng, tần suất, lí do, … Khi 2 khái niệm có thuốc, liều lượng, tần suất,… khác nhau, chúng thường không đồng tham chiếu với nhau.
4. **Indicator Extractor:** trong lớp Test, nhiều đơn vị khác nhau có thể được sử dụng như “litre”, “mililitre”, “wbc”, “hgb”,… Khi một khái niệm Test có đơn vị khác nhau thì chúng thường không đồng tham chiếu với nhau
5. **Temporal Extractor:** thông tin về thời gian là một thông tin quan trọng về ngữ cảnh. Đối với việc đồng tham chiếu lớp Treatment, các hành động và thuốc được sử dụng tại các thời điểm khác nhau được cho là khác nhau. Đối với đồng tham chiếu lớp Test, cùng một bài kiểm tra có tên khác nhau nhưng được tiến hành tại các thời điểm khác nhau là khác nhau. Khái niệm thời gian được diễn tả dưới 2 cách: cụ thể hoặc hiểu ngầm
6. **Spatial Extractor:** thông tin về không gian là thông tin quan trọng đối với lớp Treament. Ví dụ như 2 khái niệm có cùng cách viết, nhưng một khái niệm xuất hiện trong ngữ cảnh phòng phẫu thuật và một khái niệm xuất hiện trong ngữ cảnh phòng hồi sức thường không đồng tham chiếu.
7. **Section mapping Engine**: bệnh án điện tử thường bao gồm nhiều phần như: tiền sử bệnh, tiền sử dùng thuốc, tiền sử nhập viện,… Các khái niệm xuất hiện ở các phần khác nhau của bệnh án là độc lập dù chúng có cùng cách viết
8. **Modifier Extractor**: 2 khái niệm có đồng tham chiếu hay không đôi khi bị ảnh hưởng bởi từ bổ nghĩa. Một ví dụ cho các từ bổ nghĩa là “gần đây”, “bắt đầu”, “trước đó”
9. **Equipment Extractor**: thiết bị y tế đôi khi cũng là gợi ý cho ngữ cảnh vì các bài Test thường được đặt tên theo thiết bị.
10. **Operation Extractor**: nhiều khái niệm thuộc lớp Treatment là hành động phẫu thuật.
11. **Assertion Extractor**:
12. Đồng tham chiếu lớp Pronoun

Lớp Pronoun bao gồm 15 đại từ, trong đó “this”, “that”, “which”, “it” là các từ xuất hiện thường xuyên và chiếm phần lớn. Mỗi đại từ xuất hiện có thể là một khái niệm độc lập hoặc có tham chiếu tới khái niệm thuộc lớp khác. Để giải quyết vấn đề này, nhóm đề xuất sử dụng mô hình multi-class SVM để xác định đại từ được xem xét có tham chiếu tới các khái niệm thuộc 4 lớp khác hay không. Khi đã xác định được đại từ xem xét tham chiếu tới lớp nào, ta sẽ chọn khái niệm gần nhất của lớp đó để xác định 2 khái niệm là đồng tham chiếu.

### Best-first clustering

* Giải thích thuật toán best-first clustering

### Xây dựng chuỗi đồng tham chiếu

* Ghép các cặp khái niệm đồng tham chiếu để xây dựng chuỗi đồng tham chiếu

# Tập dữ liệu và phương pháp đánh giá

## Tập dữ liệu

Tập dữ liệu của nhóm được cung cấp kèm theo challenge i2b2/VA 2011 Coreference resolution, được cung cấp bởi Partners Healthcare, Beth Israel Deaconess Medical Center (MIMIC II Database), University of Pittsburgh, và Mayo Clinic. Tất cả dữ liệu được cung cấp đã được bỏ định danh và đánh dấu bằng tay bởi các chuyên gia y tế.

Để đạt được bộ dữ liệu, các nhóm hoặc tổ chức nghiên cứu cần đồng ý với cam kết về việc sử dụng dữ liệu (Data Use Agreement) và chỉ sử dụng cho mục đích nghiên cứu. Bản cảm kết cần được ký và gửi lại cho website i2b2 qua email hoặc fax.

Về tập dữ liệu nhận được bao gồm: **251 mẫu** cho trainning set và **175 mẫu** cho test set.

## Phương pháp đánh giá

Hiệu năng của hệ thống được đánh giá qua ba hệ đo: MUC, B-CUBED và CEAF. Mỗi hệ có ưu điểm và nhược điểm khác nhau. Trung bình không trọng số của ba hệ đo trên sẽ được lấy làm kết quả cuối cùng để đánh giá các chuỗi đồng tham chiếu của hệ thống so với các chuỗi ở tập kết quả.

### Hệ đo MUC

Hệ đo MUC đánh giá hệ thống dựa trên số lượng ít nhất các cặp khái niệm cần được thêm vào và loại bỏ để chuỗi đồng tham chiếu của hệ thống trùng với chuỗi ở tập kết quả. Các cặp được thêm vào là mẫu âm sai (false negative), các cặp được loại bỏ ra là mẫu dương sai (false positive). Gọi là tập các chuỗi kết quả, là tập các chuỗi được xuất ra bởi hệ thống, và là chuỗi đồng tham chiếu từ tập và tương ứng. Các độ đo recall và precision của hệ MUC được tính như sau:

được định nghĩa là số chuỗi trong có giao nhau với chuỗi .

Độ đo F-measure của hệ MUC được tính như sau:

### Hệ đo B-CUBED

Hệ đo B-CUBED đánh giá hệ thống đựa trên tính toán sự trùng lắp giữa chuỗi được xuất ra bởi hệ thống và chuỗi kết quả. Gọi là tập tài liệu, là một tài liệu trong và là một khái niệm trong . Ta định nghĩa là chuỗi kết quả có chứa và là chuỗi của hệ thống có chứa . là chuỗi giao nhau giữa và . Precision và recall của hệ B-CUBED được tính như sau:

F-measure của hệ B-CUBED được tính như hệ MUC.

### Hệ đo CEAF

Hệ đo CEAF đầu tiên sẽ tính toán một sự sắp xếp tối ưu giữa các chuỗi của hệ thống và chuỗi kết quả dựa trên mức độ tương tự (similarity score), mức độ này có thể tính dựa trên các khái niệm hoặc các chuỗi đồng tham chiếu. Độ tương tự dựa trên chuỗi đồng tham chiếu có hai phiên bản, và ; nhóm sử dụng .

Gọi tập các chuỗi kết quả của một tài liệu là , và tập các chuỗi của hệ thống cho một tài liệu là , và là một chuỗi trong và tương ứng. Độ tương tự dựa trên chuỗi được tính như sau:

Precision và recall của hệ CEAF được tính như sau:

F-measure được tính tương tự như hệ MUC.

# Kết luận

# Tài liệu tham khảo

|  |  |
| --- | --- |
| [1] | Y. Xu, J. Liu, J. Wu, Y. Wang, Z. Tu, J.-T. Sun, J. Tsujii and E. I-Chao, "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] | Wikipedia, Electronic Health Record. |
| [3] | A. B. S. S. T. F. J. P. B. R. S. Ozlem Uzuner, "Evaluating the state of the art in coreference resolution for electronic medical records," JAMIA Journal of the American Medical Informatics Association, 2012. |
| [4] | Wikipedia, "Coreference". |