**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 |
| 23/05/2015 | |

**Mục lục**

[1 Giới thiệu vấn đề 5](#_Toc420133873)

[2 Các công trình liên quan 5](#_Toc420133874)

[2.1 Bệnh án điện tử 5](#_Toc420133875)

[2.2 Phân giải đồng tham chiếu 5](#_Toc420133876)

[2.3 Phân giải đồng tham chiếu cho bệnh án điện tử 5](#_Toc420133877)

[3 Kiến thức và công nghệ 5](#_Toc420133878)

[3.1 Named-Entity-Regconition 5](#_Toc420133879)

[3.1.1 Định nghĩa nhãn 5](#_Toc420133880)

[3.2 Những vấn đề trong phân giải đồng tham chiếu trong bệnh án điện tử 6](#_Toc420133881)

[4 Bài toán đề xuất 6](#_Toc420133882)

[4.1 Phạm vi đề tài 6](#_Toc420133883)

[4.1.1 Nội dung bài toán 6](#_Toc420133884)

[4.1.2 Dữ liệu đầu vào 7](#_Toc420133885)

[4.1.3 Kết quả đầu ra 7](#_Toc420133886)

[4.2 Thiết kế hệ thống 8](#_Toc420133887)

[4.2.1 Tiền xử lý 9](#_Toc420133888)

[4.2.2 Xây dựng các cặp khái niệm 9](#_Toc420133889)

[4.2.3 Các phương pháp phân loại 9](#_Toc420133890)

[4.2.4 Best-first clustering 11](#_Toc420133891)

[4.2.5 Xây dựng chuỗi đồng tham chiếu 11](#_Toc420133892)

[5 Tập dữ liệu và phương pháp đánh giá 11](#_Toc420133893)

[5.1 Tập dữ liệu 11](#_Toc420133894)

[5.2 Phương pháp đánh giá 12](#_Toc420133895)

[5.2.1 Hệ đo MUC 12](#_Toc420133896)

[5.2.2 Hệ đo B-CUBED 12](#_Toc420133897)

[5.2.3 Hệ đo CEAF 13](#_Toc420133898)

[6 Kết luận 13](#_Toc420133899)

[7 Tài liệu tham khảo 13](#_Toc420133900)

# Giới thiệu vấn đề

* Giới thiệu về bệnh án điện tử và xu thế của nó trên thế giới
* Nêu lên vấn đề về trích xuất các kiến thức từ nguồn dữ liệu lớn như bệnh án điện tử
* Một trong những vấn đề liên quan là coref, tuy rất được quan tâm nghiên cứu cho các lĩnh vực khác nhưng cho lĩnh vực bệnh án điện tử thì vẫn chưa được xem xét tới
* Giới thiệu cụ thể bài toán: phân giải đồng tham chiếu trên bệnh án điện tử

# Các công trình liên quan

## Bệnh án điện tử

* Giới thiệu bệnh án điện tử là gì
* Đưa ví dụ

## Phân giải đồng tham chiếu

* Giới thiệu bài toán coreference là gì
* Coreference resolution nói chung có 3 kiểu hệ thống
  1. Mention-pair model
  2. Entity-mention model
  3. Ranking model

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

* Coreference cho văn bản y khoa, cụ thể là bệnh án điện tử
* Có 3 hướng tiếp cận:
  1. Rule-based learning system
  2. Supervised learning system
  3. Hybrid system

# Kiến thức và công nghệ

## Named-Entity-Regconition

* Giải thích, giới thiệu, đưa ví dụ về NER

### 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 |

## 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**”.

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:

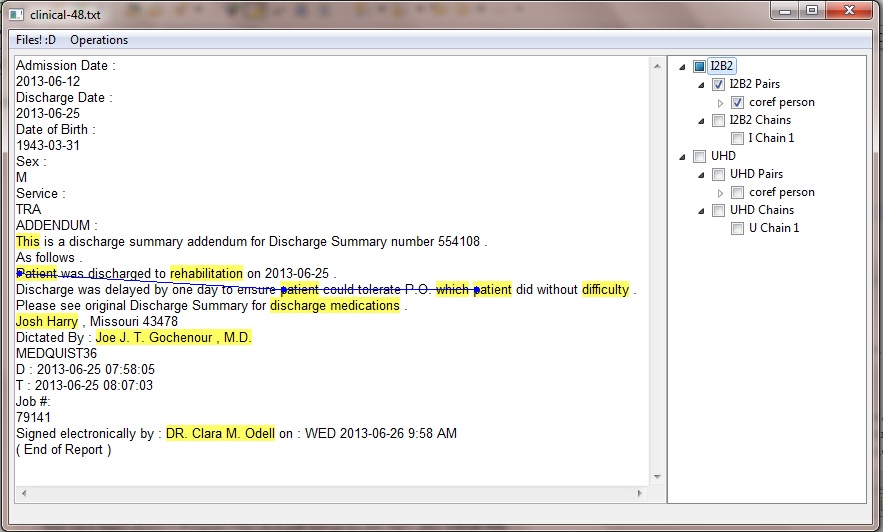
*c=”<mention>” <begin> <end> || c=….. || t=”<class>”*

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 1. 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 2. 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.

<Problem, Treatment, Test> here.

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 1. 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

…

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. |