**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 |
| 05-Jun-15 | |

**Mục lục**

[1 Giới thiệu vấn đề 5](#_Toc421196590)

[2 Các công trình liên quan 6](#_Toc421196591)

[2.1 Bệnh án điện tử 6](#_Toc421196592)

[2.2 Các hướng nghiên cứu trong bệnh án điện tử 6](#_Toc421196593)

[2.3 Thách thức i2b2 năm 2010 và 2011 8](#_Toc421196594)

[2.4 Nhận dạng thực thể có tên 9](#_Toc421196595)

[2.5 Phân giải đồng tham chiếu 10](#_Toc421196596)

[3 Kiến thức và công nghệ 13](#_Toc421196597)

[3.1 Các định nghĩa và thuật ngữ 13](#_Toc421196598)

[3.2 Support Vector Machine 13](#_Toc421196599)

[4 Bài toán đề xuất 13](#_Toc421196600)

[4.1 Phạm vi đề tài 13](#_Toc421196601)

[4.2 Thiết kế hệ thống 15](#_Toc421196602)

[5 Tập dữ liệu và phương pháp đánh giá 20](#_Toc421196603)

[5.1 Tập dữ liệu 20](#_Toc421196604)

[5.2 Phương pháp đánh giá 20](#_Toc421196605)

[6 Kết luận 21](#_Toc421196606)

[Tài liệu tham khảo 22](#_Toc421196607)

# Giới thiệu vấn đề

Trong hơn chục năm trở lại đây, với sự bùng nổ của kỉ nguyên công nghệ thông tin, việc số hóa dữ liệu trở nên phổ biến hơn bao giờ hết, và *bệnh án* cũng không phải là ngoại lệ. Bệnh án điện tử (Electronic Medical Record) đã và đang dần thay thế cho phương pháp ghi chép và lưu trữ truyền thống thông tin của bệnh nhân trong quá trình khám và chữa bệnh. Hầu hết bệnh viện ở những nước phát triển đã triển khai các *hệ thông tin bệnh viện* (HTTBV) để phục vụ cho việc số hóa loại tài liệu này.

Bên cạnh việc xây dựng bệnh án điện tử (BAĐT) thì việc khai thác nguồn dữ liệu lớn này cũng là một lĩnh vực đang rất được quan tâm trong những năm gần đây. Năm 2004, Viện y tế Quốc gia Hoa Kì (NIH: National Institute of Health) đã kêu gọi thành lập mạng lưới nghiên cứu cấp quốc gia về y sinh. Để đáp lại lời kêu gọi đó, bảy Trung tâm nghiên cứu công nghệ tính toán y sinh (NBCB: National Center for Biomedical Computing) đã được thành lập dưới sự tài trợ của NIH với nhiệm vụ xây dựng cơ sở hạ tầng phục vụ cho việc áp dụng khoa học máy tính vào lĩnh vực y sinh, hỗ trợ cho công việc nghiên cứu. Trong đó, i2b2 (Informatics for Integrating Biology and the Bedside), một NBCB được thành lập bởi sự hợp tác giữa hai trường đại học nổi tiếng Havard và MIT, bắt đầu từ năm 2006 đã tổ chức các cuộc thi hàng năm nhằm tìm kiếm các phương pháp phân tích và rút trích kiến thức trên dữ liệu BAĐT, gọi là các Thách thức (Challenges). Mỗi Thách thức đưa ra một vấn đề phân tích và một tập dữ liệu BAĐT được cung cấp bởi các bệnh viện trong và ngoài nước Mỹ. Hàng năm có trên dưới 100 nhóm nghiên cứu tham gia đề xuất giải pháp và gửi kết quả phân tích, và những giải pháp tốt được chọn lọc để công bố ở một hội thảo quốc tế và được áp dụng rộng rãi vào các dịch vụ chăm sóc sức khỏe.

Tại Việt Nam, các HTTBV cũng đang dần được triển khai, tiêu biểu là Bệnh viện đa khoa Vân Đồn tỉnh Quảng Ninh – cơ quan y tế đầu tiên có trang bị hệ thống bệnh án điện tử hiện đại và hoàn chỉnh với giải pháp MEDI SOLUTIONS của công ty phần mềm Hoa Sen. Cùng với việc xây dựng, 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 các phương pháp và phần mềm phục vụ cho khai thác bệnh án điện tử tiếng Việt.

Một trong những vấn đề của việc khai thác dữ liệu BAĐT đó là phân giải đồng tham chiếu. Thách thức lần thứ 5 (năm 2011) của i2b2 đã đưa ra một cái nhìn có hệ thống về vấn đề này. Một cách tổng quát, việc phân giải đồng tham chiếu các khái niệm trong văn bản là xác định liệu hai sự đề cập 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. Khi mà đa phần các văn bản được viết tay bằng ngôn ngữ tự nhiên, chứa đựng rất nhiều các khái niệm phụ thuộc vào ngữ cảnh thì việc phân giải đồng tham chiếu này giúp cho máy tính có một cái nhìn mang tính cấu trúc hơn về văn bản, từ đó làm nền tảng cho việc rút trích các kiến thức sâu từ những hiểu biết này.

Tuy vấn đề về phân giải đồng tham chiếu trong những năm gần đây đã được quan tâm nghiên cứu rất nhiều cho các loại văn bản khác (ví dụ các bài báo) thì ở phạm vi BAĐT vấn đề này vẫn còn ít được sự quan tâm. 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 BAĐT tiếng Anh và tiếng Việt.

# Các công trình liên quan

## Bệnh án điện tử

Bệnh án là văn bản ghi chép các thông tin sức khỏe của một cá nhân trong quá trình khám và chữa bệnh. Bệnh án điện tử chính là bệnh án được số hóa bằng HTTBV. BAĐT thông thường chứa những dữ liệu cơ bản cho quản lý, các dữ liệu cận lâm sàng và lâm sàng của người bệnh trong một lần nằm viện (Bảo, 2015). Dữ liệu lâm sàng là những *văn bản lâm sàng* (clinical text) do bác sĩ và y tá ghi chép hàng ngày về thông tin khám và chữa bệnh của người bệnh. Chính các văn bản lâm sàng này là nguồn dữ liệu quý giá cho việc khai thác và rút trích kiến thức, phục vụ cho việc chăm sóc sức khỏe và nghiên cứu y học.

Các văn bản lâm sàng trong bệnh án điện tử chủ yếu gồm ba loại (Bảo, 2015):

1. *Phiếu điều trị* (doctor daily notes): ghi chép các chuẩn đoán, nhận định và y lệnh hàng ngày của bác sĩ về bệnh nhân.
2. *Phiếu chăm sóc* (nurse narratives): là những ghi chép trong ngày của y tá trong quá trình chăm sóc và thực hiện y lệnh của bác sĩ.
3. *Hồ sơ xuất viện* (discharge summary): toàn bộ dữ liệu và thông tin cơ bản của bệnh nhân trong một lần điều trị.

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.

Ngoài các văn bản lâm sàng được lưu trữ dưới dạng phi cấu trúc, một số tiêu chuẩn được đưa ra để lưu trữ một cách có cấu trúc các BAĐT:

* IDC (International Classification of Diseases): bao gồm các loại mã cũng như thông tin về bệnh như tên bệnh, mô tả, triệu chứng, dấu hiệu, mức độ, …
* CPT (Current Procedural Terminology): bao gồm các mã mang tính thủ tục trong bệnh viện như mã xét nghiệm, gây tê, phẫu thuật, X quang, thuốc, cấp cứu, …

## Các hướng nghiên cứu 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.

Hình .. Các bài toán trong EMR

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.

Hình .. Các bài toán dự đoán

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.

Hình .. Các bài toán phát hiện

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.

Hình .. Các bài toán tìm hiểu sự kết hợp

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.

Hình .. Các bài toán rút trích thông tin

## Thách thức i2b2 năm 2010 và 2011

Năm 2010, i2b2 đưa ra Thách thức về vấn đề xử lý ngôn ngữ tự nhiên cho các văn bản lâm sàng, nó bao gồm ba tác vụ (Uzuner, South, Shen, & DuVall, 2011):

1. Trích xuất và nhận dạng các thực thể y học.
2. Phân loại bệnh vào một trong các dạng: đang xảy ra ở hiện tại, không xảy ra ở hiện tại, có thể xảy ra trong tương lai, …
3. Rút trích các quan hệ giữa các bệnh, phương pháp điều trị và thủ tục y tế.

Đối với Thách thức này, i2b2 đã định nghĩa các loại thực thể quan trọng trong một văn bản lâm sàng bao gồm vấn đề (Problem), thủ tục y tế (Test) và phương pháp điều trị (Treatment). Sự thành công của Thách thức năm 2010 đã làm nền tảng cho Thách thức 2011 về phân giải đồng tham chiếu cho BAĐT.

Các kết quả đạt được của năm 2010 được mở rộng ra cho phù hợp với việc phân giải đồng tham chiếu, lớp con người (Person) và đại từ (Pronoun) được thêm vào. Ở Thách thức 2011, có ba tác vụ được đưa ra (Uzuner, et al., 2012):

1. Tác vụ 1A: tập trung vào vấn đề trích xuất các khái niệm và phân giải đồng tham chiếu chúng cho tập dữ liệu ODIE (bao gồm các văn bản lâm sàng được viết tay bởi các bác sĩ, y tá và các loại văn bản khác như các báo cáo xuất ra ở các máy chụp CT, chụp X Quang, v.v…).
2. Tác vụ 1B: tập trung vào vấn đề phân giải đồng tham chiếu cho tập dữ liệu ODIE đã được nhận dạng và gán nhãn thực thể.
3. Tác vụ 1C: tập trung vào vấn đề phân giải đồng tham chiếu cho tập dữ liệu i2b2/VA đã được nhận dạng và gán nhãn thực thể. Tập dữ liệu này chỉ bao gồm các văn bản lâm sàng, cụ thể là các hồ sơ xuất viện (discharge summary).

Nhóm sẽ xây dựng một hệ thống xử lý tác vụ 1C.

## Nhận dạng thực thể có tên

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

Bài toán nhận dạng thực thể có tên thường bao gồm 2 bước: xác định thực thể và phân loại thực thể vào các nhóm mà thực thể đó thuộc về (như con người, nơi chốn, tổ chức, …). Trong đó, bước đầu tiên của bài toán thường được xem đơn giản như là một bài toán phân mảnh các từ trong câu thành các “tên”. Trong đó “tên” là một chuỗi các từ liên tục có ý nghĩa vả chỉ tới một thực thể có thật, không lồng ghép trong “tên” khác. Vì vậy từ “Ngân hàng Việt Nam” chỉ được xem như là một tên duy nhất, mặc dù từ “ngân hàng” và “Việt Nam” bản thân cũng mang ý nghĩa.

Một số khái niệm về thời gian và con số (như tỉ lệ %, lượng tiền, độ dài, …) cũng có thể được xem là một thực thể có tên tùy theo ngữ cảnh của bài toán nhận dạng. Tuy nhiên không phải, con số hay khái niệm thời gian nào cũng được xem là một thực thể có tên. Ví dụ như “năm 2015” được xem là một thực thể do chỉ tới năm thứ 2015 sau công nguyên, là một năm xác định. Trong khi đó “tháng 6” không được xem là một thực thể do ta không thể xác định được tháng 6 này là của năm nào. Từ ví dụ trên, ta có thể thấy khái niệm thực thể có tên không thể xác định chặt chẽ mà cần được giải thích theo ngữ cảnh của bài toán.

Có nhiều hướng tiếp cận trong việc giải quyết bài toán nhận dạng thực thể có tên, trong số đó kĩ thuật phân tích ngữ pháp ngôn ngữ và các mô hình thống kê sử dụng học máy là các hướng nổi bật. Kĩ thuật phân tích ngữ pháp ngôn ngữ thường cho độ chính xác cao hơn, nhưng thường tỉ lệ thực thể nhận diện được trên tổng số thực thể thấp. Đồng thời kĩ thuật phân tích ngữ pháp ngôn ngữ đòi hỏi một lượng lớn thời gian làm việc của các nhà ngôn ngữ học có kinh nghiệm. Trong khi đó mô hình thống kê sử dụng học máy thường yêu cầu một lượng lớn dữ liệu mẫu đã được gán nhãn để cung cấp cho quá trình học. Gần đây, mô hình học bán giám sát đã được để xuất nhằm giúp giảm bớt lượng dữ liệu mẫu cần phải có trong mô hình thống kê.

Các nghiên cứu đã xác định rằng, các hệ thống nhận diện thực thể có tên nếu hoạt động tốt trong một lĩnh vực cụ thể (như y tế, địa chất, ký sự, …) thì thường cho kết quả không tốt nếu đem ứng dụng ở lĩnh vực khác ngoài lĩnh vực hệ thống đó được thiết kế. Để có thể chỉnh sửa cho một hệ thống có sẵn hoạt động tốt trong một lĩnh vực mới thường tiêu tốn rất nhiều công sức.

Mặc dù các hệ thống được phát triển gần đây cho kết quả khá tốt nhưng bài toán nhận diện thực thể có tên vẫn mở ra nhiều hướng nghiên cứu mới. Các hướng nghiên cứu hiện nay trong bài toán nhận diện thực thể có tên bao gồm: sử dụng hệ thống học bán giám sát nhằm giúp giảm lượng dữ liệu mẫu cần có, cải thiện hiệu năng hệ thống trong nhiều lĩnh vực khác nhau, và tăng khả năng nhận diện khi có nhiều lớp thực thể.

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 |

## Phân giải đồng tham chiếu

Phân giải đồng tham chiếu là công việc xác định xem hai khái niệm trong một văn bản cùng ám chỉ, tham chiếu tới một thực thể trong thế giới thật hay không. Trong hầu hết các trường hợp thì những khái niệm này là danh từ, tên riêng (tên người, tên nơi chốn, …) hay đại từ (tôi, anh ấy, I, he, …).

**John** drove to *Judy*’s house. **He** made *her* dinner.

Trong câu trên, từ “John” và “he” cùng ám chỉ tới một thực thể là con người trong thế giới thật. Từ “Judy”, “her” ám chỉ tới một thực thể người khác. Từ đó ta có 2 chuỗi đồng tham chiếu là (John, he) và (Judy, her). Phân giải đồng tham chiếu là công việc tìm ra các chuỗi khái niệm đó.

Mặc dù việc phân giải đồng tham chiếu đã được nghiên cứu từ những năm 60 của thế kỉ trước, các thành tựu trong lĩnh vực này mới đạt được trong 16 năm trở lại đây. Sau khi các mô hình xác suất được áp dụng vào xử lý ngôn ngữ tự nhiên, các phương pháp học máy ra đời đã dần thay thế cho các phương pháp tìm kiếm heuristic. Bên cạnh đó, việc sử dụng các dữ liệu đã được gán nhãn dần trở nên phổ biến hơn vì tính hiệu quả của chúng, cho phép các nhà nghiên cứu dễ dàng xây dựng các hệ thống phân giải mới và so sánh với các hệ thống cũ sử dụng chung tập dữ liệu.

Có hai hướng tiếp cận cho bài toán phân giải đồng tham chiếu (Ng., Supervised noun phrase coreference research: the first fifteen years, 2010):

**Hướng tiếp cận về ngôn ngữ học**

Ở hướng tiếp cận này, các hệ thống phân giải lấy việc phân tích ngôn ngữ tự nhiên, cùng với các kiến thức chuyên biệt (domain knowledge), làm nền tảng. Một số giải thuật dựa trên ngôn ngữ học được sử dụng phổ biến như giải thuật Hobb (Hobbs, 1986), các nguyên lý về lý thuyết trung tâm (Centering Theory principles) (Brennan, Friedman, & Pollard, 1987) và các tham chiếu bắc cầu (bridging references) (Poesio, Mehta, Maroudas, & Hitzeman, 2004).

**Hướng tiếp cận sử dụng học máy**

Hướng tiếp cận này sử dụng các giải thuật học máy và các dữ liệu huấn luyện. Một số giải thuật như Naïve Bayes (Ge, Hale, & Charniak, 1998), Decision Tree (Soon, Ng, & Lim, 2001; Ng. & Cardie, Improving machine learning approaches to coreference resolution, 2002; Yang, Su, & Tan, 2008), Conditional Random Fields (Mccallum & Wellner, 2004) hoặc một số giải thuật gom cụm (Cardie & Wagstaf, 1999).

Hầu hết các giải thuật học máy được sử dụng là có giám sát, tuy nhiên giải thuật học máy có giám sát cần được huấn luyện trên một tập dữ liệu đã được gán nhãn, và các tập dữ liệu này thường là không có hoặc có rất ít (nhất là đối với các ngôn ngữ khác tiếng Anh). Bên cạnh đó, việc gán nhãn một tập dữ liệu cũng rất tốn kém. Do vậy, không dễ gì để có thể áp dụng các giải thuật này vào những tập dữ liệu của ngôn ngữ khác. Để giải quyết vấn đề trên, một số giải thuật học máy không giám sát đã được xây dựng, những giải thuật này có đặc điểm là không cần dữ liệu đã được gán nhãn hoặc chỉ cần được gán nhãn một phần (Cardie & Wagstaf, 1999).

Có 3 mô hình được đưa ra để giải quyết bài toán phân giải đồng tham chiếu (Ng., Supervised noun phrase coreference research: the first fifteen years, 2010) theo hướng học máy: mô hình cặp khái niệm (mention-pair), mô hình đề cập thực thể (entity-mention) và mô hình xếp hạng.

### Mô hình cặp khái niệm

Đây là một trong các mô hình đầu tiên được đề xuất cho các hệ thống phân giải đồng tham chiếu. Mô hình này phân loại đánh giá hai khái niệm được đưa ra là có đồng tham chiếu hay không. Vì đây là một hệ thống phân loại hai lớp cho các cặp khái niệm, nó không xây dựng được chuỗi đồng tham chiếu nên cần phải có thêm một giải thuật gom cụm để phân các cặp khái niệm vào các cụm, để cuối cùng xây dựng chuỗi đồng tham chiếu từ các cụm này.

Việc xây dựng các mẫu từ các cặp khái niệm từ tập dữ liệu thường là không thực tế. Một khái niệm thông thường chỉ đồng tham chiếu tới một số lượng nhỏ các khái niệm khác, thậm chí nhiều khái niệm chỉ đứng một mình mà không đồng tham chiếu với bất kì khái niệm nào (singleton). Vì vậy, số mẫu âm (tức hai khái niệm của mẫu không đồng tham chiếu với nhau) thường rất lớn so với số mẫu dương gây ảnh hưởng xấu đến kết quả huấn luyện.

Để giảm ảnh hưởng gây ra do có quá nhiều mẫu âm, một số công trình được đề xuất để lọc bỏ bớt các cặp khái niệm khó có khả năng là đồng tham chiếu để làm giảm số lượng các mẫu âm được sinh ra. Một trong những phương pháp thường được sử dụng nhất là theo hướng tiếp cận heuristic (Soon, Ng, & Lim, 2001). Hình 2.6 mô tả một đoạn văn bản và chuỗi đồng tham chiếu của nó. Có hai chuỗi đồng tham chiếu là (M1 – M2 – M3) và (D1 – D2). Hai khái niệm a và b là duy nhất, tức chúng không thuộc bất kì chuỗi đồng tham chiếu nào. Các mẫu dương được sinh là các cặp khái niệm kề nhau trực tiếp (M1 – M2), (M2 – M3) và (D1 – D2), ở đây cặp (M1 – M3) tuy thuộc chuỗi đồng tham chiếu thứ nhất nhưng không được xem là mẫu dương vì nó không kề nhau với hai cặp còn lại.

M1

M2

a

b

D1

D2

M3

Hình .. Ví dụ cho việc sinh các mẫu huấn luyện

Các mẫu âm được sinh ra từ các khái niệm nằm giữa các cặp anaphora và antecedent của các mẫu dương. Như ví dụ ở Hình 2.6, giữa D1 và D2 không có khái niệm nào, tuy nhiên giữa M2 và M3 có 4 khái niệm, các cặp (a – M3), (b – M3), (D1 – M3) và (D2 – M3) là các mẫu âm và ta chỉ cần lấy một bên là đủ (ở đây ta không xét đến các cặp bắt đầu từ M2). Như vậy có thể thấy số mẫu âm đã giảm đi đáng kể.

Ngoài ra, một số công trình khác được đề xuất để cải tiến phương pháp trên, như tối ưu việc sinh mẫu dương giữa các đại từ và phi đại từ (Müller, 2002) hoặc xem xét đến nhãn của các khái niệm. Ví dụ một khái niệm chỉ về người thì không đồng tham chiếu với một khái niệm chỉ nơi chốn, như vậy mẫu được sinh ra từ cặp hai khái niệm này là mẫu âm.

Các giải thuật học máy

Để huấn luyện mô hình này, một số giải thuật học máy đã được sử dụng. Một trong các giải thuật lâu đời nhất đó là các giải thuật sử dụng Cây quyết định, ví dụ C4.5 (Quinlan, 1993). Các giải thuật khác bao gồm học dựa trên luật (ví dụ RIPPER (Cohen, 1995)), học dựa trên trí nhớ (ví dụ TiMBL (Daelemans & Bosch, 2005)) được sử dụng trong thời kì đầu của việc áp dụng học máy vào phân giải đồng tham chiếu. Khi các giải thuật học máy dựa trên mô hình xác suất trở nên phổ biến, một số giải thuật đã được sử dụng vào lĩnh vực này gồm có mô hình entropy cực đại (Berger, Pietra, & Pietra, 1996), mạng neuron bầu cử (Freund & Schapire, 1999) và support vector machines (SVM) (Joachims, 1999). Một đặc điểm lợi thế của các giải thuật học máy dựa trên mô hình xác suất đó là chúng có thể tính được độ tin cậy đồng tham chiếu của các cặp khái niệm.

Các phương pháp gom cụm

Như đã nói ở trên, mô hình cặp khái niệm chỉ đánh giá tính đồng tham chiếu của một cặp khái niệm mà không có khả năng xây dựng chuỗi đồng tham chiếu. Do đó, một số giải thuật gom cụm được sử dụng để làm việc này.

**Gom cụm gần nhất trước (Closest-first Clustering)** (Soon, Ng, & Lim, 2001)

Giải thuật gom cụm gần nhất trước xem rằng mọi khái niệm đều có thể là một hồi chỉ (anaphora) và mọi khái niệm đứng trước hồi chỉ đó đều có thể là tiền đề (antecedent) của nó. Giải thuật này làm việc như sau: xét tất cả các khái niệm từ về , với mỗi khái niệm xét tất cả các khái niệm đứng trước nó từ về , sử dụng một giải thuật phân loại xác định tính đồng tham chiếu của cặp khái niệm , nếu gặp khái niệm đầu tiên mà giải thuật phân loại xác định cặp là đồng tham chiếu thì lấy làm tiền đề của và không xét tới những khái niệm khác nữa. Nếu đã duyệt về khái niệm mà giải thuật vẫn xác định cặp là không đồng tham chiếu thì không có tiền đề.

**Gọm cụm tốt nhất trước (Best-first Clustering)**

Một công trình nghiên cứu khác (Ng. & Cardie, Improving machine learning approaches to coreference resolution, 2002) cho rằng giải thuật gom cụm gần nhất trước không phải lúc nào cũng phân các cặp khái niệm vào những cụm tốt nhất, vì giải thuật này chỉ xét đến duy nhất một tiền đề đồng tham chiếu gần nhất mà không quan tâm đến xác suất đồng tham chiếu của nó, những tiền đề khác rất dễ bị bỏ qua ngay cả khi chúng có xác suất đồng tham chiếu cao hơn. Vì thế giải thuật gọm cụm tốt nhất trước xem xét tất cả các tiền đề khả thi, sử dụng một giải thuật phân loại dựa trên mô hình xác suất để tính độ tin cậy đồng tham chiếu của chúng và chọn ra khái niệm có độ tin cậy cao nhất làm tiền đề.

### Mô hình đề cập thực thể

Mô hình đề cập thực thể gom cụm các khái niệm bằng cách tính toán khoảng cách dựa trên sự tương thích. Khi khoảng cách giữa hai cụm không vượt quá một ngưỡng cho phép, hai cụm này sẽ được trộn vào nhau. Khoảng cách giữa hai khái niệm được định nghĩa như sau:

trong đó, là khái niệm, là tập các thuộc tính và là một thuộc tính trong tập . Hàm xác định sự không tương thích của thuộc tính giữa hai khái niệm với giá trị là 0 hoặc 1. Trọng số xác định mức độ quan trọng của thuộc tính đối với tính tương thích giữa hai khái niệm.

Ở bước khởi tạo, giải thuật xác định một bán kính cho phép của một cụm và ban đầu mỗi khái niệm là một cụm của chính nó. Trong quá trình duyệt lần lượt từ cuối lên, các cụm được trộn với nhau nếu khoảng cách của chúng không vượt quá . Nếu có ít nhất một cặp khái niệm không tương thích ở một hoặc cả hai cụm thì hai cụm này sẽ không được trộn.

### Mô hình xếp hạng

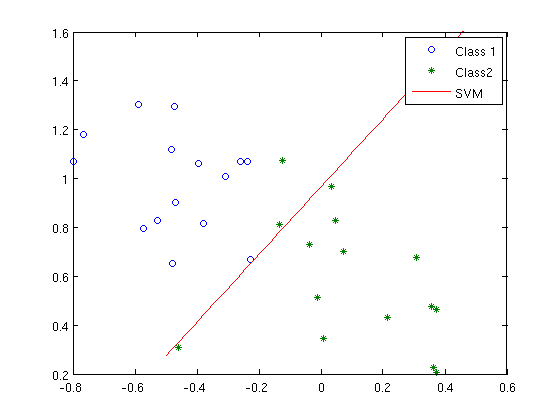
Mô hình xếp hạng xét tất cả các tiền đề và cố gắng chọn ra tiền đề tốt nhất bằng cách so sánh các tiền đề với nhau, hay nói một cách khác là xếp hạng chúng. Một ví dụ của mô hình xếp hạng là mô hình cặp ứng cử viên <ref>. Khác với mô hình một ứng cử viên, mô hình cặp ứng cử viên xét từng bộ ba gồm hồi chỉ, hai tiền đề khả thi và sử dụng một hệ phân loại để xác định xem tiền đề nào là tốt nhất cho hồi chỉ này. Kết quả cuối cùng là xác suất mà một tiền đề được xem là tốt nhất so với các tiền đề còn lại

# Kiến thức nền tảng

## Các định nghĩa và thuật ngữ

## Support Vector Machine

Trong lĩnh vực học máy, Support Vector Machine (SVM) là 1 mô hình học có giám sát dựa trên nền tảng là giải thuật phân tích dữ liệu và nhận diễn mẫu. Mô hình SVM thường được sử dụng cho các bài toán phân loại và phân tích hồi quy. SVM nhận vào một tập dữ liệu với mỗi điểm dữ liệu đã được đánh dấu thuộc 1 trong 2 lớp có sẵn và cố gắng xây dựng mô hình để phân định lớp khi xuất hiện điểm dữ liệu mới chưa biết. Mô hình SVM thường được minh họa bằng các điểm trong không gian, trong đó các điểm dữ liệu được phân chia bằng một đường thẳng tuyến tính sao cho khoảng cách giữa 2 cụm dữ liệu là lớn nhất có thể. Khi xuất hiện điểm dữ liệu mới chưa biết, điểm dữ liệu đó sẽ được ánh xạ vào không gian tương ứng và từ đó có thể giúp dự đoán được điểm dữ liệu mới thuộc vào lớp nào. Ngoài khả năng phân loại tuyến tính, SVM cũng cho kết quả tốt đối với phân loại phi tuyến tính nếu áp dụng kĩ thuật *kernel,* trong đó ngầm ánh xạ các điểm dữ liệu vào không gian cấp cao.



Hình .. Minh họa mô hình SVM

**Tối ưu hóa khoảng cách**

Mô hình SVM là mô hình phân loại tuyến tính các điểm dữ liệu trong không gian. Tuy nhiên, ta có thể tìm được nhiều hơn 1 đường thẳng có khả năng giúp phân biệt các điểm dữ liệu thuộc 2 lớp khác nhau (Hình 3.2). Nếu chúng ta chọn đường thẳng phân loại có khoảng cách tới điểm dữ liệu thuộc 2 lớp nhỏ (Hình 3.2b), trong thực tế, sai số của dữ liệu có thể khiến cho việc phân loại sai đối với những điểm dữ liệu thuộc biên của 2 lớp. Vì vậy chúng ta thường chọn đường phân loại có thể tối đa hóa khoảng cách từ đường phân loại đến điểm dữ liệu thuộc 2 lớp.



Hình .. Tối ưu hóa khoảng cách

Để dễ dàng cho việc định lượng, ta thường xem xét khoảng cách lớn nhất từ đường thẳng phân loại đến điểm dữ liệu gần nhất của 2 bên. Khoảng cách này là đối xứng qua 2 bên của đường thẳng phân loại và được gọi là M. Các điểm dữ liệu thuộc 2 lớp và có vị trí gần nhất với đường thẳng phân loại được gọi là *support vector.*

Từ đó công việc của mô hình SVM là tìm kiếm đường phân loại sao cho ta có thể tối đa hóa khoảng cách M.

**Kĩ thuật kernel**

Trong nhiều trường hợp, chúng ta không thể nào tìm được 1 đường thẳng có thể giúp phân loại các điểm dữ liệu thành 2 lớp. Vì vậy để giải quyết vấn đề đó, chúng ta cần suy nghĩ cách để thay đổi việc ánh xạ các điểm dữ liệu vào không gian. Ý tưởng cơ bản của việc này là chỉnh sửa các đặc trưng của dữ liệu sao cho các điểm dữ liệu xuất hiện trong không gian có thể được phân loại bằng 1 đường thẳng tuyến tính, hoặc ánh xạ các điểm dữ liệu vào không gian có số chiều cao hơn mà trong không gian đó, ta có thể tìm được một đường thẳng hoặc mặt phẳng có thể giúp phân loại các điểm dữ liệu. Tuy nhiên, ta không thể tự tạo ra dữ liệu mới, vì vậy ta chỉ có thể suy diễn đặc trưng mới từ các đặc trưng có sẵn.

Trong Hình 3.3, dữ liệu trong không gian 2 chiều không thể được phân loại nếu sử dụng đường thẳng, vì vậy ta cần biến đổi dữ liệu trong không gian 2 chiều lên không gian có chiều cao hơn để nhờ đó, ta có thể tìm được một mặt phẳng giúp phân loại chúng. Trong ví dụ minh họa, chiều thứ 3 được suy diễn từ 2 chiều có sẵn theo cách sau



Hình .. Kĩ thuật kernel giúp biến đổi không gian dữ liệu

Khi sử dụng kĩ thuật kernel, hàm ánh xạ không gian thường không cố định và được lựa chọn tùy theo tính chất của dữ liệu. Trong việc lựa chọn hàm ánh xạ, kiến thức nền tảng của lĩnh vực bài toán đóng vai trò hỗ trợ.

**Mô hình SVM mở rộng**

Mặc dù mô hình SVM chỉ sử dụng trong việc phân loại dữ liệu vào 2 lớp, chúng ta thường gặp phải bài toán trong đó số lượng lớp của dữ liệu là lớn hơn 2. Để giải quyết những bài toán có số lớp lớn hơn 2, mô hình SVM mở rộng được đưa ra trong đó sử dụng cùng lúc nhiều SVM để phân loại.

Trong mô hình SVM mở rộng, mỗi lớp dữ liệu sẽ có 1 SVM giúp phân loại lớp đó với các lớp còn lại. Như vậy, nếu bài toán bao gồm n lớp dữ liệu, ta cần n SVM trong đó SVM 1 giúp phân loại lớp x1 với (n-1) lớp còn lại, SVM 2 giúp phân loại lớp x2 với (n-1) lớp còn lại và tiếp tục như vậy với các SVM còn lại. Khi xuất hiện 1 điểm dữ liệu mới, điểm dữ liệu đó sẽ được phân loại qua tất cả n SVM đã được xây dựng. Lớp của điểm dữ liệu mới sẽ tùy thuộc vào SVM nào có độ tin cậy cao nhất.

# Phương pháp đề xuất

**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:

* 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:

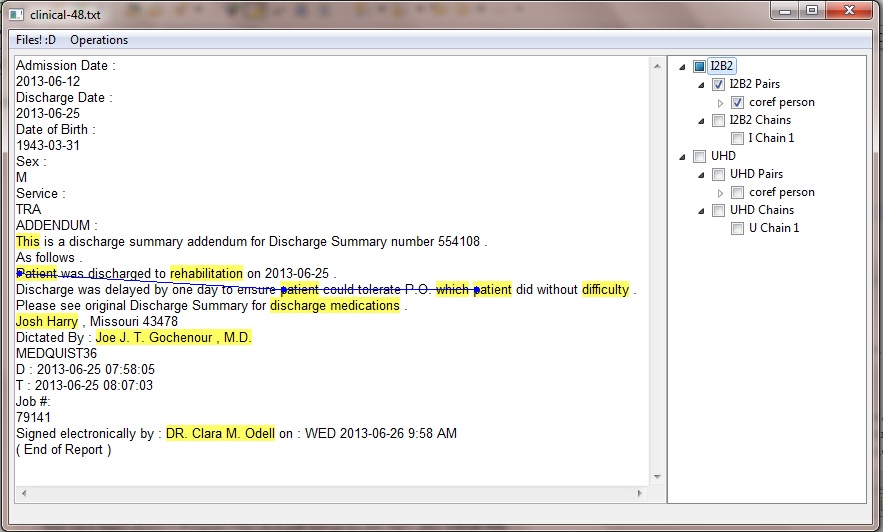
*c=”<mention>” <begin> <end> || c=….. || t=”<class>”*

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



**Chi tiết phương pháp**

Một cách tổng quan, hệ thống sẽ gồm các bước sau (Hình 4.1): 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**:

**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.

| Thuộc tính | Giá trị | Giải thích |
| --- | --- | --- |
| First previous mention type | 0, 1, 2, 4 | Các lớp Person, Problem, Treatment, Test, Null lần lượt tương ứng với các giá trị 0, 1, 2, 3, 4 |
| Second previous mention type | 0, 1, 2, 3, 4 | Các lớp Person, Problem, Treatment, Test lần lượt tương ứng với các giá trị 0, 1, 2, 3, 4 |
| First next mention type | 0, 1, 2, 3, 4 | Các lớp Person, Problem, Treatment, Test lần lượt tương ứng với các giá trị 0, 1, 2, 3, 4 |
| Sentence distance | 0, 1, 2,… | Khoảng cách giữa 2 câu chứ 2 khái niệm |
| Pronoun | 0,1,2,…14 | Chỉ số của đại từ trong bảng tra 15 đại từ |
| Part of speech | 0, 1, 2 | DT, WDT, PRP |
| First next verb after mention |  | Động từ đầu tiên liền sau khái niệm được xét |
| First word before mention is preposition | 0, 1 | Là đại từ chỉ nơi chốn (1), ngược lại (0) |
| First one/two/three words before mention | 0, 1 | 3 từ liền trước của khái niệm được xét |
| First one/two/three words after mention | 0, 1 | 3 từ liền sau của khái niệm được xét |

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

## 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

Trong giai đoạn thực tập tốt nghiệp vừa qua, nhóm đã thực hiện được:

* Tìm hiểu khái quát về bệnh án điện tử
* Tìm hiểu về các hướng nghiên cứu và bài toán hiện có trong bệnh án điện tử
* Tìm hiểu về bài toán đồng tham chiếu nói chung và các bài toán liên quan
* Thiết kế sơ bộ hệ thống
* Chuẩn bị tập dữ liệu tiếng Anh và tiếng Việt

Sau quá trình tìm hiểu, nhóm quyết định đề xuất đề tài “**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**”. Dữ liệu đầu vào của bài toán gồm hồ sơ xuất viện và các khái niệm được đề cập trong bài toán theo một định dạng có sẵn. Dữ liệu đầu ra của bài toán gồm các cặp khái niệm và chuỗi khái niệm đồng tham chiếu.

Giải pháp đề xuất cho bài toán gồm 4 bước: tiền xử lý dữ liệu, áp dụng hệ thống học có giám sát SVM, áp dụng hệ thống gom cụm theo chiến thuật best-first và xây dựng các chuỗi khái niệm đồng tham chiếu.

Ngoài ra, nếu thời gian cho phép, nhóm đề xuất khả năng mở rộng đề tài bao gồm việc xây dựng hệ thống tự động nhận diện các khái niệm được đề cập trong bệnh án điện tử. Từ đó nhóm có thể tích hợp để đưa ra 1 hệ thống duy nhất cả khả năng nhận đầu vào là bệnh án điện tử và đưa ra kết quả là chuỗi khá niệm đồng tham chiếu trong bệnh án đó.

Kế hoạch trong giai đoạn luận văn tốt nghiệm của nhóm là:

* Tiền xử lý dữ liệu để phù hợp với đầu vào của bài toán
* Xây dựng hệ thống học có giám sát SVM gồm 3 module person, non-person và pronoun
* Xây dựng hệ thống gom cụm theo chiến thuật best-first
* Hoàn thiện hệ thống
* Đánh giá hệ thống và so sánh với kết quả

Xây dựng và phát triển hướng mở rộng

# Tài liệu tham khảo

Bảo, H. T. (2015). Xây dụng và khai thác BAĐT: con đường mới trong khám chữa bệnh và nghiên cứu y học. *Khoa học & Công nghệ Việt Nam, 56*(3), 16-20.

Berger, A. L., Pietra, V. J., & Pietra, S. A. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics, 22*(1), 39-71.

Brennan, S. E., Friedman, M. W., & Pollard, C. J. (1987). A Centering Approach to Pronouns. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics* (pp. 155-162). Stroudsburg, PA, USA: Association for Computational Linguistics.

Cardie, C., & Wagstaf, K. (1999). Noun Phrase Coreference as Clustering. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 82-89).

Cohen, W. W. (1995). Fast Effective Rule Induction. In *In Proceedings of the Twelfth International Conference on Machine Learning* (pp. 115-123). Morgan Kaufmann.

Daelemans, W., & Bosch, A. v. (2005). *Memory-Based Language Processing.* New York, NY, USA: Cambridge University Press.

Freund, Y., & Schapire, R. E. (1999). Large Margin Classification Using the Perceptron Algorithm. *Machine Learning, 37*(3), 277-296.

Ge, N., Hale, J., & Charniak, E. (1998). A Statistical Approach to Anaphora Resolution. In *Sixth Workshop on Very Large Corpora* (pp. 161-170). Orlando, Florida.

Hobbs, J. (1986). Resolving Pronoun References. In *Readings in Natural Language Processing* (pp. 339-352). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Joachims, T. (1999). Making Large-scale Support Vector Machine Learning Practical. In *Advances in Kernel Methods* (pp. 169-184). Cambridge, MA, USA: MIT Press.

Marsland, S. (n.d.). *Machine Learning: An Algorithmic Perspective.*

Mccallum, A., & Wellner, B. (2004). Conditional Models of Identity Uncertainty with Application to Noun Coreference. In *Advances in Neural Information Processing Systems 17* (pp. 905-912). Cambridge, MA, USA: MIT Press.

Müller, M. S. (2002). The Influence of Minimum Edit Distance on Reference Resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (pp. 312-319). Association for Computational Linguistics.

Ng., V. (2010). Supervised noun phrase coreference research: the first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1396-1441). Uppsala, Sweden: Association for Computational Linguistics.

Ng., V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of 40th Annual Meeting on Association for Computational Linguistics* (pp. 104-111). Philadelphia, Pennsylvania: Association for Computational Linguistics.

Poesio, M., Mehta, R., Maroudas, A., & Hitzeman, J. (2004). Learning to Resolve Bridging References. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics* (p. 143). Stroudsburg, PA, USA: Association for Computational Linguistics.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Soon, W. M., Ng, H. T., & Lim, D. C. (2001, December). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics, 27*(4), 521-544.

Uzuner, Ö., Bodnari, A., Shen, S., Forbush, T., Pestian, J., & South, B. R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association : JAMIA, 19*(5), 786-791.

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA, 18*(5), 552-556.

Wikipedia. (n.d.). *Coreference.*

Wikipedia. (n.d.). *Electronic Health Record.*

Wikipedia. (n.d.). Named Entity Regconition.

Wikipedia. (n.d.). Support Vector Machine.

Xu, Y., Liu, J., Wu, J., Wang, Y., Tu, Z., Sun, J.-T., . . . I-Chao, E. (2012). A classification approach to coreference in discharge summaries: 2011 i2b2 challenge. *Journal of the American Medical Informatics Association : JAMIA, 19*(5), 897-905. doi:10.1136/amiajnl-2011-000734

Yang, X., Su, J., & Tan, C. L. (2008). A Twin-Candidate Model for Learning-Based Anaphora Resolution. *Computational Linguistics, 34*(3), 327-356.