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:
GS. TS. Cao Hoàng Trụ

Sinh viên thực hiện:
Nguyễn Duy Hưng – 51101475
Vương Anh Tuấn – 51104040

TP. Hồ Chí Minh, 06/2015

Mục lục

1 Giới thiệu vấn đề

Trong hơn mười 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 tắt 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, 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 sự vật hoặc hiện tượng 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 cho dữ liệu BAĐT.

2 Các công trình liên quan

2.1 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 [1]. 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 [1]:

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:

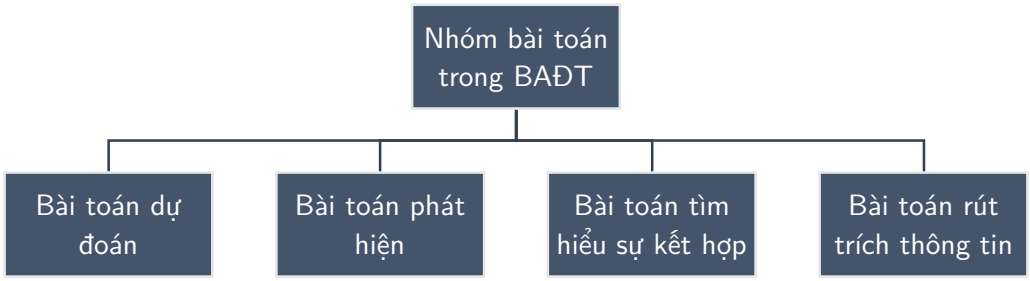
- Lưu trữ chính xác và đầy đủ thông tin bệnh nhân, tránh trùng lặp dữ liệu.
- Hỗ trợ quá trình tìm kiếm và truy xuất thông tin nhanh chóng.
- Dữ liệu có thể được chia sẻ hoặc tích hợp.

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, …

2.2 Các hướng khai thác dữ liệu bệnh án điện tử

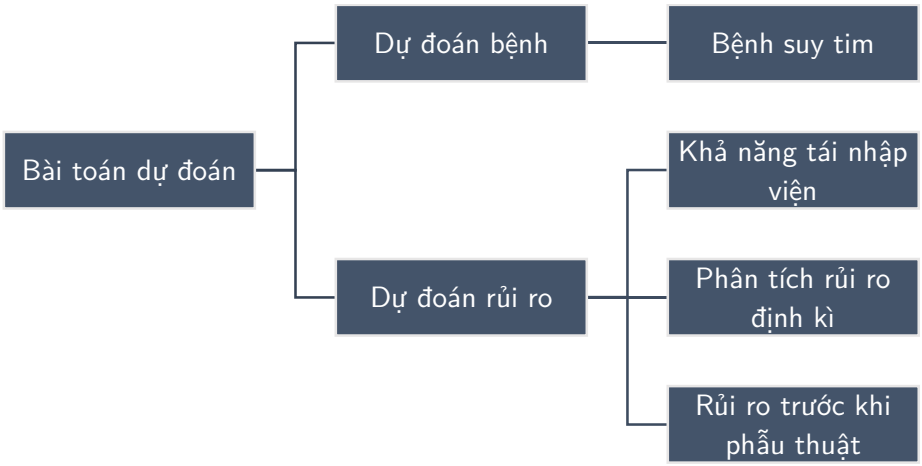
Với sự phát triển mạnh mẽ của bệnh án điện tử, nhiều hướng nghiên cứu khai thác bệnh án điện tử đã được đưa ra để sử dụng tối đa nguồn thông tin đã có. Tuy vậy, ta có thể phân chia thành 4 hướng nghiên cứu chính như sau: bài toán dự đoán bệnh hoặc các biến cố trong y học, bài toán phát hiện bệnh, bài 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 lâm sàng trong bệnh án điện tử như chẩn đoán của bác sĩ, ghi chú chăm sóc của y tá, báo cáo xuất viện,… để nhận diện các thực thể, tìm mối quan hệ và trích xuất các thông tin theo yêu cầu. Một trường hợp đặc biệt của bài toán xác định mối quan hệ là bài toán phân giải đồng tham chiếu.



Hình 2.1. Các nhóm bài toán khai thác BAĐT

Bài toán dự đoán

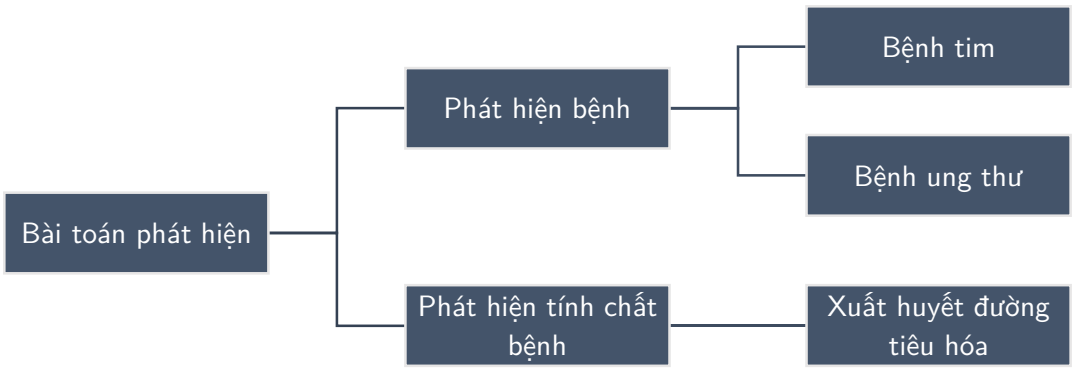
Bài toán dự đoán là những bài toán giúp đoán trước khả năng mắc bệnh hoặc những biến cố y tế có thể xảy ra với bệnh nhân trong tương lai. Việc dự đoán được dựa trên những thông tin có được hiện tại hoặc quá khứ của bệnh nhân được xét từ đó giúp phòng tránh hoặc có những biện pháp thích hợp. Bài toán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



Hình 2.2. Các bài toán dự đoán

Bài toán phát hiện

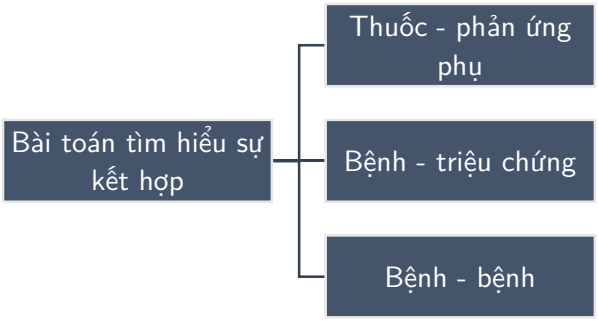
Bài toán phát hiện là những bài toán giúp tìm ra được những vấn đề tiềm ẩn trong cơ thể bệnh nhân mà chưa được biết đến. Bài toán phát hiện có hai nhánh chính là: phát hiện bệnh tiềm ẩn và phát hiện tính chất bệnh.



Hình 2.3. Các bài toán phát hiện

Bài toán tìm hiểu sự kết hợp

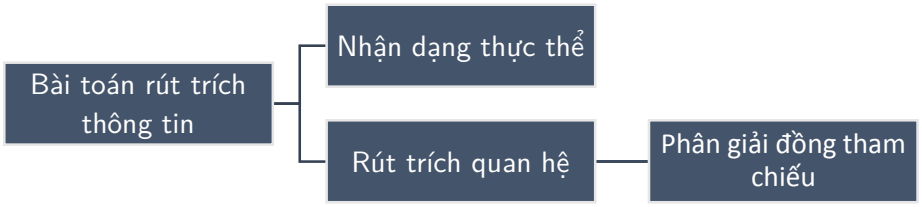
Bài toán tìm hiểu sự kết hợp là những bài toán cho biết mối quan hệ giữa các thực thể trong y học, thường được áp dụng cho việc ra quyết định hoặc phát hiện tác dụng phụ của thuốc. Nhóm bài toán gồm các nhánh chính là: sự kết hợp của thuốc – phản ứng phụ, bệnh – triệu chứng, bệnh – bệnh.



Hình 2.4. 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. Phân giải đồng tham chiếu là một trường hợp đặc biệt của bài toán rút trích quan vệ, vì đồng tham chiếu cũng được xem như là một quan hệ giữa hai thực thể.



Hình 2.5. Các bài toán rút trích thông tin

2.3 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 [2]:

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 tập trung vào giải quyết nhóm bài toán rút trích thông tin vì đây là nhóm bài toán nền tảng, tạo tiền đề để nghiên cứu cho các hướng đi khác. Các loại thực thể quan trọng trong một văn bản lâm sàng được i2b2 định nghĩa bao gồm vấn đề (Problem), thủ tục y tế (Test) và phương pháp điều trị (Treatment). Tuy Thách thức i2b2 năm 2010 có đề cập

Lớp	Ví dụ	Định nghĩa
Person	Dr.Lightman, the patient, cardiology, he, she, ...	Những chủ thể người hoặc một nhóm người được đề cập trong bệnh án và các đại từ nhân xưng
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	Which, it, that,...	Những đại từ có thể tham chiếu đến bất kì lớp nào trong bốn lớp kể trên nhưng không phải là đại từ nhân xưng

Bảng 2.1. Ý nghĩa các lớp thực thể được đề xuất bởi i2b2

đến việc rút trích các quan hệ giữa các thực thể trong bệnh án (tác vụ thứ 3), nhưng mối quan hệ đồng tham chiếu lại không được bao gồm trong số đó. Chính vì thế, năm 2011 i2b2 tổ chức Thách thức lần thứ 5 dành riêng cho việc giải quyết vấn đề phân giải đồng tham chiếu trên dữ liệu BAĐT.

Các kết quả đạt được của năm 2010 được mở rộng cho phù hợp với việc phân giải đồng tham chiếu, trong đó lớp con người (Person) và đại từ (Pronoun) được thêm vào (Bảng 2.1). Ở Thách thức 2011, có ba tác vụ được đưa ra [3]:

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, 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).

Thách thức i2b2 2011 đã thu hút được sự quan tâm của nhiều bệnh viện và có hơn 20 nhóm nghiên cứu trên toàn thế giới tham gia. Một cách tổng quát, các giải pháp phân giải đồng tham chiếu đề xuất bởi các nhóm dự thi được chia làm ba loại: *hệ thống dựa trên luật, hệ thống học máy có giám sát* và *hệ thống lai.*

Theo như quan sát [3], các hệ thống trên có một điểm chung là đều phát triển các module phân giải khác nhau cho ba lớp Person, Problem/Treatment/Test, Pronoun; phân loại các sự đề cập chỉ người trong BAĐT thành ba loại: *bệnh nhân, người thân của bệnh nhân* và *nhân sự của bệnh viện*; sử dụng các kiến thức nền như WordNet, Wikipedia hay UMLS để xác định các từ viết tắt, từ đồng nghĩa hay trích xuất các đặc trưng về ngữ cảnh cho các sự đề cập thuộc lớp Problem/Treatment/Test. Ở tác vụ 1C, hệ thống sử dụng giải pháp học máy có giám sát [4] cho

kết quả cao nhất với độ F bằng 0.915. Vì kết quả đó, ở luận văn này nhóm sẽ tập trung tìm hiểu các phương pháp học máy có giám sát cho phân giải đồng tham chiếu.

2.4 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 bài toán rút trích thông tin. Nhiệm vụ của nhận dạng thực thể là xác định những thực thể trong câu và gán nhãn cho nó. Ví dụ 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 và 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 vì nó 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 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 này 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 đề xuất nhằm giúp giảm bớt lượng dữ liệu mẫu cần phải có trong mô hình thống kê.

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 sẽ cho kết quả không tốt nếu đem ứng dụng vào lĩnh vực khác. 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 2010 đã định nghĩa 5 loại thực thể cần được quan tâm, đó là vấn đề (Problem), phương pháp điều trị (Treatment), các xét nghiệm (Test).

2.5 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, …). Ví dụ:

John drove to **Judy**'s house. **She** made *him* dinner.

Trong câu trên, từ "John" và "him" cùng ám chỉ tới một thực thể là con người trong thế giới thật. Từ "Judy", "she" ám chỉ tới một thực thể người khác. Từ đó ta có 2 chuỗi đồng tham chiếu là (John, him) và (Judy, she). 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 gần 20 năm trở lại đây. Sau khi các mô hình thống kê 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 [5]:

1. *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 Hobbs [6], các nguyên lý về lý thuyết trung tâm (Centering Theory principles) [7] và các tham chiếu bắc cầu (bridging references) [8].
2. *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 [9], Cây quyết định [10,11,12], Conditional Random Fields [13] hoặc một số giải thuật gom cụm [14].

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 [14].

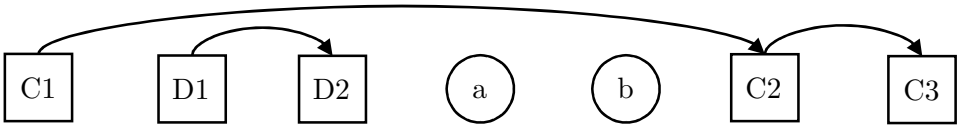
Có 3 mô hình được đưa ra để giải quyết bài toán phân giải đồng tham chiếu [5] theo hướng học máy: *mô hình cặp khái niệm* (mention-pair), *mô hình đề cập thực thể* (entity-mention) và *mô hình xếp hạng* (ranking).

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 của 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 [10]. Hình 2.6 mô tả một đoạn văn bản và chuỗi đồng tham chiếu của nó. Có hai chuỗi đồng tham chiếu là (C1 – C2 – C3) 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 ra là các cặp khái niệm kề nhau trực tiếp (C1 – C2), (C2 – C3) và (D1 – D2), ở đây cặp (C1 – C3) tuy thuộc chuỗi đồng tham chiếu thứ nhất nhưng không được xem là mẫu dương vì hai khái niệm của cặp không kề nhau trực tiếp.



Hình 2.6. Ví dụ cho việc sinh các mẫu huấn luyện

Các mẫu âm được sinh ra từ các khái niệm nằm giữa các cặp hồi chỉ (khái niệm đứng sau) và tiền đề (khái niệm đứng trước) của các mẫu dương. Như ví dụ ở Hình 2.6, giữa D1 và D2 không có khái niệm nào, tuy nhiên giữa C1 và C2 có 4 khái niệm, do đó các cặp (D1 – C2), (D2 – C2), (a – C2) và (b – C2) là các mẫu âm và ta chỉ cần lấy một bên là đủ (ở đây ta không xét đến các cặp bắt đầu từ C1). 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 [15] 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.

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 [16]. Các giải thuật khác bao gồm học dựa trên luật (ví dụ RIPPER [17]), học dựa trên trí nhớ (ví dụ TiMBL [18]) 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 thống kê trở nên phổ biến, một số giải thuật đã được

sử dụng vào lĩnh vực này gồm có mô hình entropy cực đại [19], mạng neuron bầu cử [20] và support vector machines (SVM) [21]. Một đặc điểm lợi thế của các giải thuật học máy dựa trên mô hình thống kê đó là chúng có thể tính được độ tin cậy đồng tham chiếu của các cặp khái niệm.

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)

Giải thuật gom cụm gần nhất trước [10] xem rằng mọi khái niệm đều có thể là một hồi chỉ và mọi khái niệm đứng trước hồi chỉ đó đều có thể là tiền đề của nó. Giải thuật này làm việc như sau: xét tất cả các khái niệm từ j_n về j_2, với mỗi khái niệm j xét tất cả các khái niệm i đứng trước nó từ i_j về i_1, sử dụng một giải thuật phân loại xác định tính đồng tham chiếu của cặp khái niệm (i, j), nếu gặp khái niệm i đầu tiên mà giải thuật phân loại xác định cặp (i, j) là đồng tham chiếu thì lấy i làm tiền đề của j và không xét tới những khái niệm khác nữa. Nếu đã duyệt về khái niệm i_1 mà giải thuật vẫn xác định cặp (i_1, j) là không đồng tham chiếu thì j không có tiền đề.

Gom cụm tốt nhất trước (Best-first Clustering)

Một công trình nghiên cứu khác [11] 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 được đề xuất, nó 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.

Mặc dù các giải pháp gom cụm được để xuất để cải tiến việc phân giải đồng tham chiếu thực sự cho kết quả tốt, chúng vẫn không giải quyết được hai nhược điểm lớn của mô hình cặp khái niệm:

1. Mô hình cặp khái niệm xem xét đến các cặp khái niệm một cách độc lập, tức nó chỉ xác định được tiền đề ứng viên nào là có khả năng đồng tham chiếu nhất với một hồi chỉ mà không xét được mức độ đồng tham chiếu với hồi chỉ giữa các tiền đề ứng viên khác nhau. Nói cách khác, mô hình cặp khái niệm không trả lời được cho câu hỏi: ứng viên nào là tốt nhất để lấy làm tiền đề cho một hồi chỉ.

2. Việc chỉ xem xét đến cặp hai khái niệm độc lập không cung cấp đủ thông tin để đảm bảo việc xác định đúng mối quan hệ đồng tham chiếu giữa chúng, nhất là với các khái niệm là đại từ hay các danh từ ko có bổ từ xác định (như số lượng hay giới tính).

Vì lý do đó, hai mô hình khác được đề xuất để giải quyết các khuyết điểm này.

Mô hình đề cập thực thể

Mô hình này giải quyết được khuyết điểm thứ hai của mô hình cặp khái niệm. Ta xem xét một ví dụ như sau: một văn bản có chứa ba khái niệm “Mr. Taylor”, “Taylor” và “she”. Hai khái niệm “Mr. Taylor” và “Taylor” được hệ thống phân loại xác định là đồng tham chiếu do có sự trùng lắp chuỗi “Taylor”, cặp “Taylor” và “she” cũng được xác định là đồng tham chiếu vì

chúng nằm gần nhau trong văn bản và thiếu đi một số thông tin xác định giới tính hay số lượng. Như vậy theo tính chất bắc cầu, hai khái niệm "Mr. Taylor" và "she" đồng tham chiếu với nhau, tuy nhiên có thể dễ dàng nhận thấy điều này là không đúng.

Vì lý do đó, mô hình đề cập thực thể cố gắng xem xét tính đồng tham chiếu của một khái niệm NP_k với một cụm khái niệm C_j trước đó, thay vì là các cặp hai khái niệm độc lập. Để làm được điều này, khoảng cách giữa hai khái niệm được tính toán dựa trên sự tương thích giữa chúng. Khi khoảng cách giữa NP_k và cụm C_j không vượt quá một ngưỡng cho phép, NP_k có thể được đưa vào cụm. Khoảng cách giữa hai khái niệm được định nghĩa như sau:

$$dist(NP_1, NP_2) = \sum_{f \in F} w_f \cdot incompatibility_f(NP_1, NP_2)$$

trong đó, NP là khái niệm, F là tập các thuộc tính và f là một thuộc tính trong tập F. Hàm $incompatibility_f$ xác định sự không tương thích của thuộc tính f giữa hai khái niệm, nó có thể trả về giá trị là 0 (không tương thích) hoặc 1 (tương thích). Trọng số w_f xác định mức độ quan trọng của thuộc tính f đối với tính tương thích giữa hai khái niệm.

Ở bước khởi tạo, giải thuật xác định một bán kính cho phép r của 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 r. 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.

Ngoài tính toán khoảng cách, mô hình này có thể được huấn luyện bằng cách sử dụng tập thuộc tính ở mức cụm (Cluster-level feature). Giả sử ta có khái niệm NP_k và cụm C_j được tạo thành một mẫu, thuộc tính của mẫu này có thể được xác định bằng cách áp dụng một mệnh đề lên các thuộc tính của cặp hai khái niệm. Ví dụ với thuộc tính đồng thuận *giới tính*, mệnh đề ALL có thể được sử dụng để xét xem tất cả các khái niệm trong cụm C_j có cùng giới tính với khái niệm NP_k hay không. Ngoài mệnh đề ALL, hai mệnh đề khác có thể được sử dụng như MOST (NP_k có cùng giới tính với hơn một nửa khái niệm trong C_j) hay ANY (NP_k có cùng giới tính với ít nhất một khái niệm trong C_j).

Mô hình xếp hạng

Mô hình xếp hạng giải quyết được nhược điểm thứ nhất, tức nó 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 [12]. Khác với mô hình một ứng cử viên, mô hình cặp ứng cử viên xét từng bộ ba gồm hồi chỉ, hai tiền đề ứng viên 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

$$\ln p(ante(C_k) \mid ana, C_1, C_2, \dots, C_n) = \textstyle\sum_{1<i<n, i \neq k} \ln p(C_k \succ C_i \mid ana, C_k, C_i)$$

3 Kiến thức nền tảng

3.1 Các định nghĩa và thuật ngữ

Trong các công trình nghiên cứu về phân giải đồng tham chiếu, các tác giả thường sử dụng từ *markable* để chỉ đến những từ/cụm từ mà có thể chỉ về một từ/cụm từ khác. Ở một số tài liệu khác lại sử dụng từ *mention* để chỉ đến các thực thể trong văn bản, nó mang ý nghĩa như một

sự đề cập. Tuy nhiên để tiện lợi trong việc diễn đạt bằng tiếng Việt, nhóm sử dụng từ **khái niệm** (concept) để chỉ đến những thực thể trong một văn bản mà ta cần phân giải đồng tham chiếu cho chúng. Một lý do khác mà nhóm sử dụng từ này bắt nguồn từ việc trung tâm i2b2 gọi các tập tin chứa những thực thể đã được gán nhãn là *concept files.*

Các khái niệm thông thường là danh từ hoặc cụm danh từ. Một khái niệm có thể được lồng ở trong một khái niệm khác, đa phần sự lồng nhau này xuất hiện ở các cụm danh từ mang ý nghĩa sở hữu, ví dụ như cụm *ngôi nhà của anh ta* chứa hai khái niệm *(ngôi nhà của anh ta)* và *(anh ta)*. Một số tài liệu định nghĩa các khái niệm lồng nhau là *khái niệm đầy đủ* và các hệ thống phân giải đồng tham chiếu ở những bài báo đó xem xét đến sự lồng nhau này. Tuy nhiên phương pháp mà nhóm đề xuất ở đây bỏ qua sự lồng nhau đó và xem cụm *ngôi nhà của anh ta* chỉ ám chỉ tới một thực thể duy nhất đó là ngôi nhà.

Hai khái niệm được xem là **đồng tham chiếu** nếu cả hai khái niệm đó cùng ám chỉ về một thực thể trong thế giới thực, ví dụ *thủ tướng Việt Nam* và *Nguyễn Tấn Dũng*. Một đặc điểm cần lưu ý của tính đồng tham chiếu đó là nó phụ thuộc vào ngữ cảnh và thời điểm mà hai khái niệm được đề cấp đến. Như ví dụ trên *thủ tướng Việt Nam* và *Nguyễn Tấn Dũng*, hai khái niệm này chỉ đồng tham chiếu khi Nguyễn Tấn Dũng đang là thủ tướng Việt Nam hiện tại. Mọi khái niệm đều có những thuộc tính về mặt ngữ pháp và ngữ nghĩa, như giới tính, số lượng hay lớp ngữ nghĩa (sematic class), v.v…

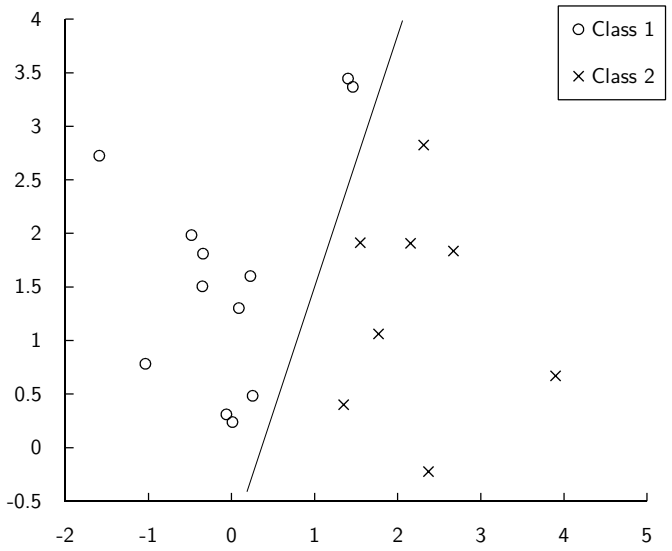
Một **cặp khái niệm** gồm hai khái niệm có thể có hoặc không đồng tham chiếu với nhau. Khái niệm đứng trước được gọi là *tiền đề* (antecedent), khái niệm đứng sau gọi là *hồi chỉ* (anaphora). Thông thường hồi chỉ phụ thuộc vào tiền đề về mặt ngữ pháp. Tuy nhiên ở phạm vi phân giải đồng tham chiếu, hồi chỉ không nhất thiết phải tuân theo các quy tắc ngữ pháp mà nó có thể là những khái niệm độc lập, như tên riêng hoặc danh từ không xác định. Những thuộc tính của một cặp khái niệm là những thuộc tính đồng thuận (agreement feature), nó áp dụng cho cả hai khái niệm của cặp, ví dụ như sự đồng thuận về số lượng hay giới tính. Trong một văn bản, nhiều khái niệm có thể cùng tham chiếu tới một thực thể, khi đó chúng tạo thành một chuỗi đồng tham chiếu.

Mặc dù nhiều chuỗi đồng tham chiếu có thể được phân giải từ một văn bản, có một số lượng lớn các khái niệm không đồng tham chiếu với bất kì khái niệm nào khác. Các khái niệm loại này được gọi là *khái niệm duy nhất* (singleton). Chúng có thể là những từ chỉ được đề cập một lần duy nhất hay là những khái niệm không diễn giải hay ám chỉ tới một thực thể nào trong thế giới thực. Số lượng các khái niệm duy nhất có thể rất nhiều tùy thuộc vào tập dữ liệu.

Về mối quan hệ đồng tham chiếu, đây là mối quan hệ tương đương và nó có những tính chất sau:

- *Tính phản xạ:* một khái niệm thì luôn đồng tham chiếu với chính nó.
- *Tính đối xứng:* nếu khái niệm C_1 đồng tham chiếu với C_2 thì C_2 cũng đồng tham chiếu với C_1.
- *Tính bắc cầu:* nếu khái niệm C_1 đồng tham chiếu với C_2, C_2 đồng tham chiếu với C_3 thì C_1 đồng tham chiếu với C_3.

Ta có thể xem việc phân giải các chuỗi đồng tham chiếu như việc gom các cặp khái niệm lại thành từng cụm, mỗi cụm ứng với một chuỗi đồng tham chiếu. Các chuỗi đồng tham chiếu được



Hình 3.1. Minh họa mô hình SVM

phân giải bằng con người từ các dữ liệu được gán nhãn gọi là các *chuỗi kết quả*. Các chuỗi đồng tham chiếu được xuất ra bởi các giải thuật gom cụm gọi là các *chuỗi hệ thống*.

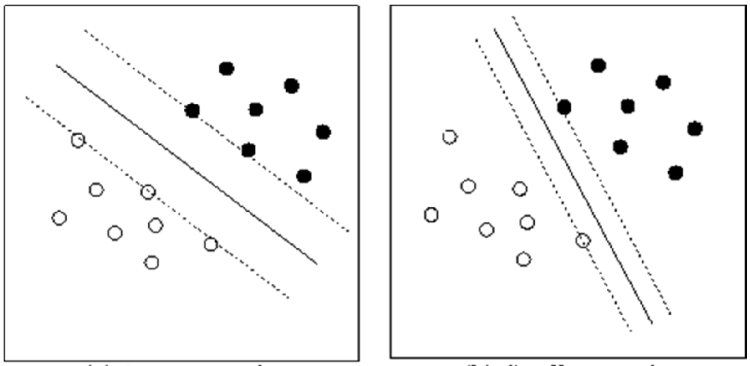
3.2 Support Vector Machine

Trong lĩnh vực học máy, Support Vector Machine (SVM) [22] là một mô hình học có giám sát dựa trên nền tảng là giải thuật phân tích dữ liệu và nhận diện 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 một trong hai 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 đường thẳng phân cách tới các điểm dữ liệu gần nhất ở 2 bên là lớn nhất (Hình 3.1). 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.

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 một đường thẳng có khả năng giúp phân biệt các điểm dữ liệu thuộc hai lớp khác nhau (Hình 3.2). Nếu chúng ta chọn đường thẳng phân loại có khoảng cách tới điểm dữ liệu thuộc hai 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 hai 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 hai lớp.

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 hai bên. Khoảng cách này là đối xứng qua hai bên của đường thẳng phân loại và được gọi là M. Các điểm dữ liệu thuộc hai lớp và có vị trí gần nhất với



Hình 3.2. Tối ưu hóa khoảng cách

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 hai lớp. Vì vậy để giải quyết vấn đề đó, chúng ta cần tìm 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 một 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.

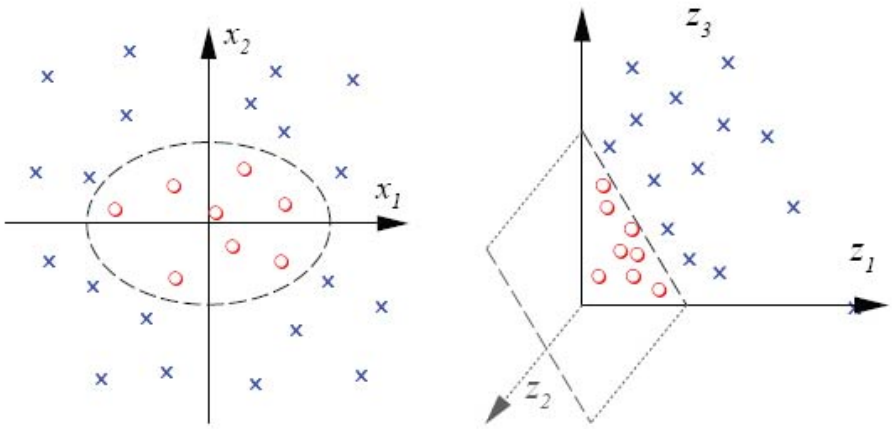
Trong Hình 3.3, dữ liệu trong không gian 2 chiều không thể được phân loại nếu sử dụng đường thẳng, vì vậy ta cần biến đổi dữ liệu trong không gian 2 chiều lên không gian có chiều cao hơn để nhờ đó, ta có thể tìm được một mặt phẳng giúp phân loại chúng. Trong ví dụ minh họa, chiều thứ 3 được suy diễn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

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 những 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 x_1 với $(n-1)$ lớp còn lại, SVM 2 giúp phân loại lớp x_2 với $(n-1)$ lớp còn lại và tiếp tục như vậy với các SVM còn lại. Khi xuất hiện một 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.



Hình 3.3. Kĩ thuật kernel giúp biến đổi không gian dữ liệu

3.3 Phân tích các thuộc tính đặc trưng cho phân giải đồng tham chiếu

Để sử dụng các mô hình phân loại hay gọm cụm, hệ thống phân giải đồng tham chiếu cần một tập các thuộc tính đặc trưng. Thuộc tính đặc trưng là các giá trị mô tả các đặc điểm đặc trưng của một khái niệm, một cặp khái niệm hay thậm chí là một cụm các khái niệm. Tập các thuộc tính được hệ thống trích xuất từ văn bản lấy làm đầu vào cho các mô hình học máy. Các thuộc tính đặc trưng được chia làm hai loại:

1. *Các đặc trưng nội tại:* là các thuộc tính đặc trưng được trích xuất từ chính bản thân các các khái niệm trong văn bản, ví dụ như các thuộc tính về so trùng chuỗi kí tự, các thuộc tính về mặt ngữ pháp, về cú pháp hay về ngữ nghĩa.
2. *Các đặc trưng bên ngoài:* là các thuộc tính đặc trưng mà cần phải được trích xuất từ một nguồn kiến thức bên ngoài, vì bản các thân khái niệm không cho ta biết được chúng. Một số các kiến thức nền có thể sử dụng như Wikipedia, WordNet, Freebase hay Yago.

Ở phần này nhóm sẽ trình bày một cách tổng quan một số thuộc tính đặc trưng tiêu chuẩn được sử dụng ở các hệ thống phân giải đồng tham chiếu cho các văn bản nói chung [5]. Các thuộc tính ở đây được mô tả một cách tương đối, chúng có thể được sử dụng theo các cách khác nhau ở các hệ thống khác nhau.

Đặc trưng về từ ngữ

Trùng chuỗi hoàn toàn

Chuỗi kí tự biểu diễn hai khái niệm là như nhau. Thuộc tính này có thể được định nghĩa cho toàn bộ khái niệm hoặc chỉ cho từ đầu của khái niệm.

Trùng chuỗi con

Một phần chuỗi kí tự của khái niệm này (có thể là phần đầu) trùng với toàn bộ chuỗi kí tự của khái niệm khác.

Trùng lặp từ ngữ

Hai khái niệm có sự trùng lặp về ít nhất một từ trong chuỗi kí tự biểu diễn của chúng

Tên giả
: Khái niệm này là tên giả của khái niệm kia, ví dụ *Việt Nam* và *VN*.

Khoảng cách Leveinstein
: Khoảng cách Leveinstein giữa hai khái niệm. Thuộc tính này được sử dụng cho các ngôn ngữ chứa nhiều hình thái của các từ như tiếng Anh.

Đặc trưng về ngữ pháp

Tính xác định
: Một hoặc cả hai khái niệm là danh từ xác định (ví dụ các từ có mạo từ *the* phía trước). Khi đó chúng có khả năng cao là đồng tham chiếu.

Tính không xác định
: Một hoặc cả hai khái niệm là danh từ không xác định (ví dụ các từ có mạo từ *a/an* phía trước). Khi đó chúng có khả năng cao là không đồng tham chiếu, vì các danh từ không xác định thường ám chỉ tới một thực thể khác hoặc một khái niệm chung.

Cả hai đều là tên riêng
: Cả hai khái niệm đều là tên riêng.

Cả hai đều là đại từ
: Cả hai khái niệm đều là đại từ.

Quan hệ về mặt ngữ pháp
: Cả hai khái niệm có chung một mối quan hệ về ngữ pháp, ví dụ hai khái niệm đều là *chủ ngữ*.

Sự đồng thuận
: Hai khái niệm đồng thuận về số lượng hay giới tính (thường xuất hiện trong tiếng Anh).

Đặc trưng về ngữ nghĩa

Lớp ngữ nghĩa
: Hai khái niệm thuộc cùng một lớp ngữ nghĩa. Lớp ngữ nghĩa này thường được xác định bởi bước nhận dạng thực thể có tên hoặc sử dụng một kiến thức nền như Wikipedia.

Quan hệ ngữ nghĩa
: Hai khái niệm có quan hệ về ngữ nghĩa, ví dụ khái niệm này đồng nghĩa với khác niệm khác mặc dù cách viết của hai khái niệm là khác nhau. Tính đồng nghĩa này có thể được xác định bằng cách sử dụng một kiến thức nền như Wikipedia.

Đặc trưng về vị trí

Khoảng cách câu
: Số câu xuất hiện giữa hai khái niệm. Đa phần các hồi chỉ thường rất gần với tiền đề của chúng. Một số hệ thống còn tính số khái niệm xuất hiện giữa một cặp khái niệm.

Đồng vị ngữ
: Khái niệm sau là đồng vị ngữ của khái niệm trước.

Đại từ quan hệ
: Khái niệm sau là đại từ quan hệ (ví dụ which, who, whom,…) của khái niệm trước.

Lớp	Ví dụ
Person	**The patient** was started on Pseudomonas for treatment of **her** right lower lobe pneumonia .
Problem	**Serratia urosepsis** treated with ceftizoxime .
	Azotemia presumed secondary to **sepsis** and dehydration, creatinine decreased to 2.1 with intravenous fluids and antibiotics .
Treatment	L2 to L5 laminectomy, **L2 through L5 posterior fusion** with segmental instrumentation .
	She was taken to the operating room for laminectomy with **posterior instrumented fusion at L3 to L5** .
Test	**An echocardiogram** was scheduled and Cardiology was consulted .
	Echocardiogram showed moderate anterior pericardial effusion of approximately 600 cc with diastolic indications of the right ventricle and low velocity paradox .
Pronoun	Her deep wound culture showed (**MRSA**) (**which**) is sensitive to Bactrim and rifampin .

Bảng 3.1. Một số ví dụ về đồng tham chiếu trong BAĐT

Một số công trình nghiên cứu [23] xem xét đến ảnh hưởng của các thuộc tính đặc trưng lên *tính đúng đắn* và *tính đầy đủ* của hệ thống. Tuy nhiên nhóm không đi quá chi tiết vào phần này vì đa phần các công trình đó giải quyết các vấn đề cho phạm vi văn bản nói chung còn BAĐT lại có những đặc trưng riêng của nó. Việc thiết kế và xây dựng tập thuộc tính cho các khái niệm/cặp khái niệm cho hệ thống phân giải đồng tham chiếu trên BAĐT của nhóm sẽ được trình bày chi tiết ở phần sau.

3.4 Các vấn đề về phân giải đồng tham chiếu cho bệnh án điện tử

Như được đề cập ở phần trước, có ba mô hình được đề xuất cho việc phân giải đồng tham chiếu. Mô hình cặp khái niệm là mô hình ra đời đầu tiên và đơn giản nhất, tuy nhiên cũng chính vì thế mà nó có một số nhược điểm khi áp dụng cho những dữ liệu phức tạp. Hai mô hình sau là mô hình đề cập thực thể và mô hình xếp hạng được đề xuất để giải quyết các điểm yếu đó. Các mô hình này được xây dựng cho các văn bản nói chung, khi mà ở đó có nhiều hơn một chuỗi đồng tham chiếu lớn thuộc về một lớp cùng tồn tại những lại không chứa các khái niệm mang nhiều ngữ nghĩa mà đa phần chỉ là các đại từ hay danh từ xác định. Tuy nhiên phân giải đồng tham chiếu trong phạm vi BAĐT lại có những tính chất khác biệt so với các văn bản chung, đòi hỏi chúng ta cần có một sự phân tích chuyên biệt cho dữ liệu BAĐT, từ đó có thể áp dụng phương pháp phân giải phù hợp nhất.

Trong quá trình tìm hiểu các phương pháp phân giải đồng tham chiếu cho dữ liệu BAĐT [4], nhóm nhận thấy rằng các khái niệm thuộc về các lớp khác nhau thì có các đặc tính khác nhau. Như ví dụ ở Bảng 3.1, nếu hai khái niệm thuộc lớp Problem/Treatment/Test có sự trùng lắp chuỗi thì rất có khả năng hai khái niệm này là đồng tham chiếu (*Serratia urosepsis* và *sepsis*). Mặt khác, đối với lớp Person khi mà đa phần cái khái niệm là đại từ thì sự trùng lắp chuỗi này không phải là một tác nhân quan trọng. Ngoài ra, lớp Problem/Treatment/Test có rất nhiều từ đồng nghĩa như các từ viết tắt, biệt ngữ hay tên giả. Chúng là tác nhân chính gây nên sự khó khăn trong việc phân giải đồng tham chiếu ở ba lớp này. Nhận thấy điều đó, nhóm cho rằng cần thiết phải sử dụng một kiến thức thực tế làm nền tảng (ví dụ Wikipedia) để hỗ trợ việc xác định các từ đồng nghĩa.

Một vấn đề quan trọng nữa ở lớp Problem/Treatment/Test là tính đồng tham chiếu của các khái niệm thuộc lớp này thường phụ thuộc vào ngữ cảnh, do vậy để phân giải đúng ta cần xem

xét đến ngữ cảnh mà các khái niệm được đề cập tới (như thời gian và không gian). Đối với BAĐT, tính phụ thuộc ngữ cảnh không tồn tại ở lớp Person.

Thông thường trong một BAĐT, tồn tại một chuỗi đồng tham chiếu lớn được tạo thành bởi các khái niệm thuộc lớp Person. Đó chính là chuỗi các khái niệm tham chiếu đến bệnh nhân của bệnh án. Trong khi đó, một chuỗi đồng tham chiếu lớn như vậy lại không tồn tại cho các khái niệm thuộc lớp Problem/Treatment/Test.

Từ những quan sát trên nhóm đưa ra hai đặc điểm lợi thế ở BAĐT mà các loại văn bản khác không có:

1. Chỉ có một chuỗi đồng tham chiếu lớn, và chuỗi này chứa các khái niệm cùng tham chiếu đến bệnh nhân của bệnh án.
2. Các lớp ngữ nghĩa trong BAĐT (các lớp Problem, Treatment, Test và Person) phân định rõ các khái niệm và chúng đóng một vai trò quan trọng trong phân giải đồng tham chiếu ở BAĐT.

Hai đặc điểm trên khiến cho một số vấn đề khó khăn gặp phải ở các văn bản nói chung không còn tồn tại, tuy nhiên có những vấn đề ở BAĐT mà ta cần phải xem xét tới như:

1. Cùng một khái niệm nhưng được đề cập ở các ngữ cảnh khác nhau thì thường không đồng tham chiếu đến cùng một thực thể, như vậy cần được phân biệt rõ.
2. Cần phải nhận định được một số lượng lớn các từ đồng nghĩa trong chuyên ngành y khoa và các danh từ thiếu mạo từ xác định.

Từ các nhận định trên, nhóm cho rằng chỉ cần sử dụng phương pháp phân giải đồng tham chiếu đơn giản nhất làm nền tảng và tập trung vào việc thiết kế tập thuộc tính đặc trưng cho các lớp khái niệm.

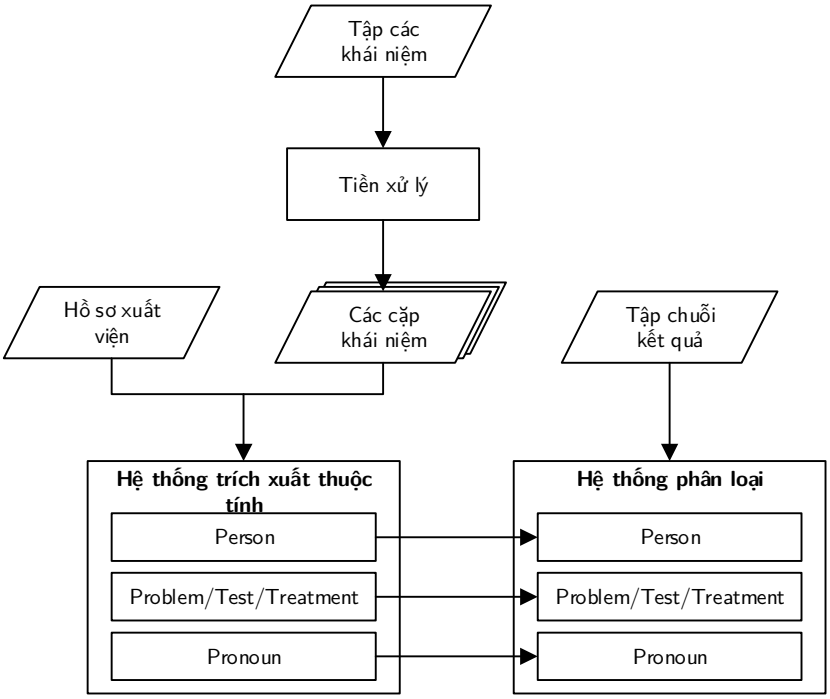
4 Phương pháp đề xuất

4.1 Nội dung bài toán

Dựa vào tác vụ 1C trong các Thử thách lần thứ 5 (2011) của Trung tâm nghiên cứu công nghệ tính toán y sinh i2b2, nhóm quyết định đề xuất bài toán "*Phân giải đồng tham chiếu trên bệnh án điện tử với các khái niệm đã được trích xuất và gán nhãn cho dữ liệu tiếng Anh*".

Dữ liệu đầu vào của nhóm gồm 2 phần:

1. *Tập các hồ sơ xuất viện*
   Đây là những văn bản lâm sàng được viết tay bằng ngôn ngữ tự nhiên bởi các bác sĩ, y tá. Chúng mô tả lại toàn bộ thông tin của bệnh nhân trong một lần điều trị, bao gồm các thông tin về tên bệnh mà bệnh nhân mắc phải, các thủ tục y tế được thực hiện và các phương pháp điều trị được áp dụng lên bệnh nhân.
2. *Tập các khái niệm đã được trích xuất và gán nhãn từ các hồ sơ xuất viện*
   Mỗi hồ sơ xuất viện đi kèm với một văn bản chứa toàn bộ các khái niệm được đề cập trong hồ sơ đó. Các khái niệm này đã được gán nhãn cho phù hợp với loại thực thể mà nó đề cập tới. Có tất cả năm nhãn là Problem, Treatment, Test, Person và Pronoun. Bảng 2.1 mô tả chi tiết ý nghĩa của năm nhãn này.



Hình 4.1. Quy trình huấn luyện

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

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

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

4.2 Tổng quan quy trình

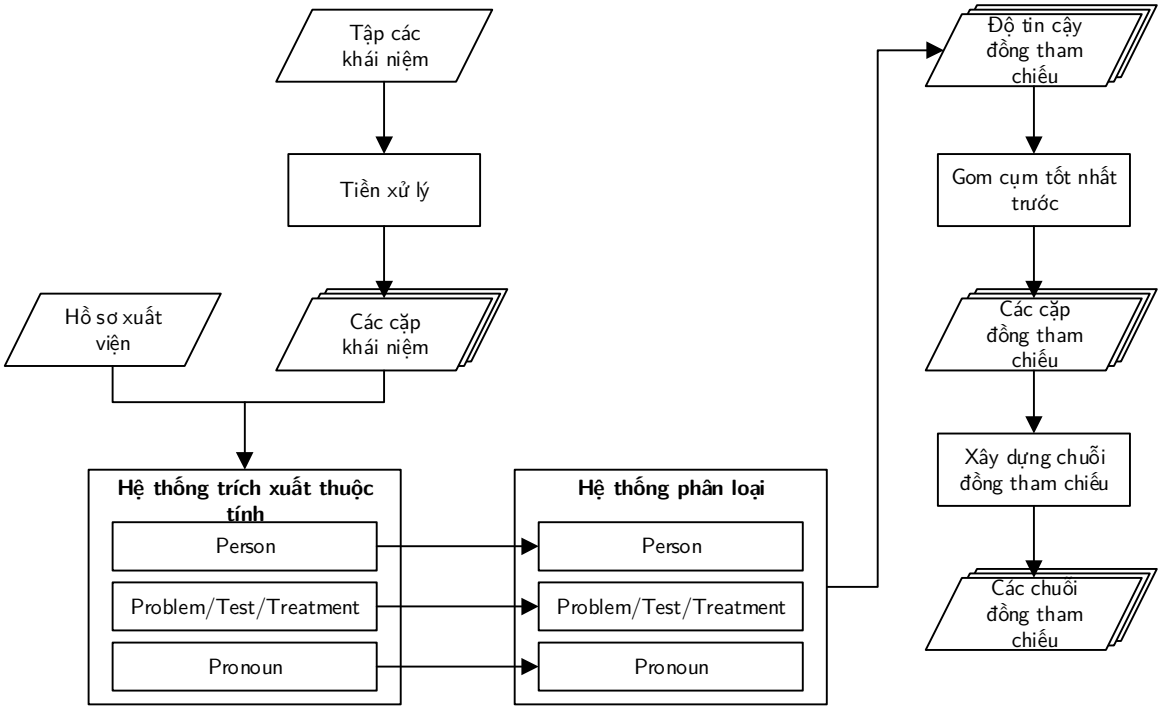
Từ các kiến thức thu được thông qua các công trình đã tìm hiểu, nhóm quyết định sử dụng mô hình cặp khái niệm để giải quyết mục tiêu của bài toán. Cụ thể phương pháp mà nhóm đề xuất bao gồm hai quy trình: *quy trình huấn luyện hệ thống phân loại* dựa trên dữ liệu huấn luyện và *quy trình phân giải đồng tham chiếu* cho dữ liệu kiểm thử hoặc dữ liệu mới.

Quy trình huấn luyện hệ thống phân loại

Quy trình huấn luyện (Hình 4.1) là quy trình giúp xây dựng hệ thống phân loại có khả năng đánh giá độ tin cậy đồng tham chiếu của một cặp hai khái niệm. Đây cũng chính là quá trình học của mô hình học có giám sát. Quy trình huấn luyện hệ thống phân loại nhận đầu vào là tập các khái niệm, hồ sơ xuất viện trong tập dữ liệu i2b2/VA và các chuỗi kết quả của việc phân giải đồng tham chiếu để sử dụng cho mô hình học máy.

Tập các khái niệm ban đầu sau quá trình tiền xử lý sẽ được ghép đôi thành các cặp hai khái niệm. Các khái niệm chỉ được ghép đôi nếu chúng cùng một lớp nhằm lọc bớt các mẫu âm. Việc chia cặp khái niệm này nhằm mục đích xây dựng các mẫu huấn luyện.

Các cặp khái niệm sau khi được ghép đôi sẽ được đưa vào hệ thống trích xuất thuộc tính cùng với văn bản gốc của hồ sơ xuất viện. Trong đó, văn bản gốc của hồ sơ xuất viện đóng vai trò



Hình 4.2. Quy trình phân giải đồng tham chiếu

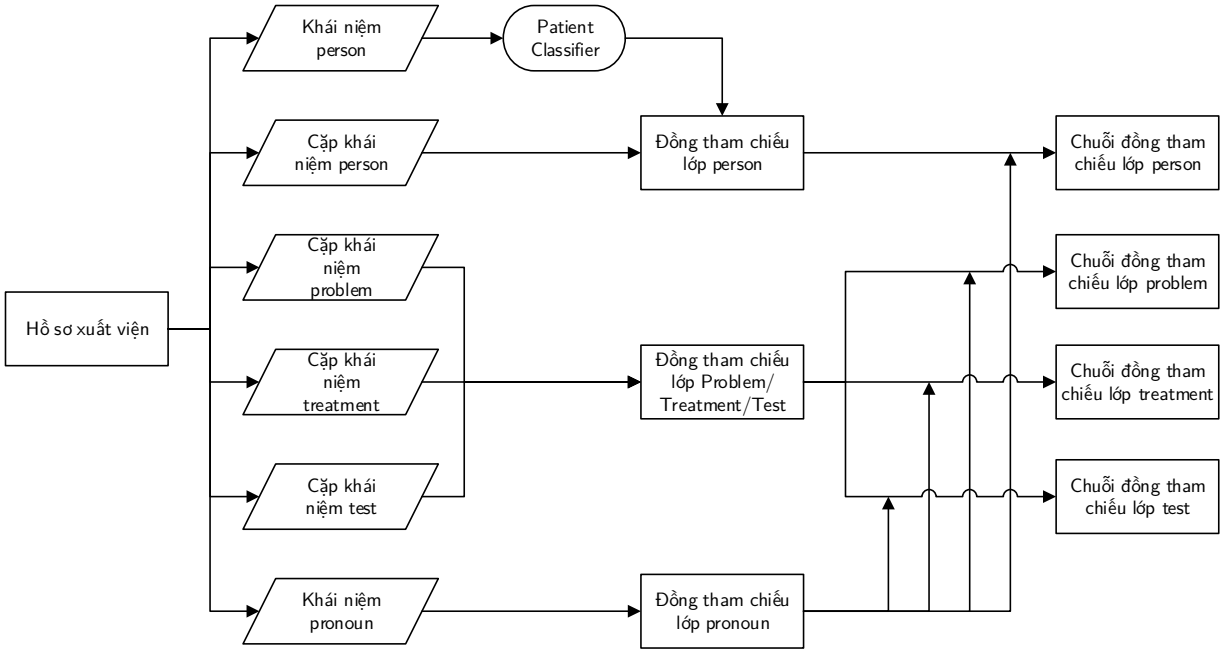
cung cấp thông tin ngữ pháp và ngữ cảnh cho cặp khái niệm. Thuộc tính là các giá trị độc lập giúp ta biết được các thông tin đặc trưng của đối tượng được quan sát khi đặt trong ngữ cảnh. Việc lựa chọn các thuộc tính có chứa nhiều thông tin và có khả năng phân loại dữ liệu cao là một bước quan trọng trong mô hình học máy. Vì mỗi lớp khái niệm có những đặc trưng riêng khác nhau, hệ thống trích xuất thuộc tính sẽ chia làm ba module riêng cho ba nhóm lớp: module Person, module Problem/Test/Treatment và module Pronoun. Hệ thống trích xuất thuộc tính sẽ cho kết quả là các vector chứa giá trị của những thuộc tính được quy định trước.

Hệ thống phân loại là trung tâm của quy trình huấn luyện và là hệ thống cần được xây dựng từ các kết quả biết trước. Ở đây, hệ thống phân loại sẽ được huấn luyện dựa trên các vector chứa giá trị các thuộc tính đặc trưng của cặp khái niệm và kết quả đúng sai về việc cặp khái niệm đó có đồng tham chiếu hay không. Tương tự như hệ thống trích xuất thuộc tính, hệ thống phân loại cũng bao gồm 3 module riêng cho 3 nhóm lớp: module Person, module Problem/Test/Treatment, module Pronoun. Mô hình phân loại mà nhóm lựa chọn để sử dụng trong các hệ thống này là SVM.

Quy trình phân giải đồng tham chiếu

Sau khi xây dựng được hệ thống phân loại, quy trình phân giải đồng tham chiếu (Hình 4.2) là quy trình tìm ra các chuỗi đồng tham chiếu trong một văn bản mới không có trong tập huấn luyện. Đây cũng chính là quy trình được sử dụng trong thực tế khi ta cần phân giải đồng tham chiếu trong văn bản bất kì.

Quy trình phân giải đồng tham chiếu có các bước đầu giống như quy trình phân loại bao gồm: tiền xử lý để xây dựng các cặp khái niệm, sử dụng hệ thống trích xuất thuộc tính để tìm ra các vector thuộc tính, sử dụng hệ thống phân loại để đánh giá khả năng đồng tham chiếu của cặp khái niệm. Tuy nhiên, điểm khác biệt với quy trình huấn luyện là hệ thống phân loại trong quy



Hình 4.3. Phân chia hệ thống phân giải đồng tham chiếu ra ba loại Person, Problem/Treatment/Test và Pronoun

trình phân giải đồng tham chiếu đã được huấn luyện và có khả năng đưa ra độ tin cây đồng tham chiếu của cặp khái niệm mới.

Sau khi có được độ tin cậy đồng tham chiếu, nhóm sử dụng giải thuật gom cụm theo chiến lược tốt nhất trước để gom các cặp khái niệm cùng trỏ tới một thực thể lại với nhau. Sau đó các cặp khái niệm có đồng tham chiếu sẽ được liên kết lại với nhau và xây dựng thành các chuỗi đồng tham chiếu. Đây cũng là bước cuối cùng của quy trình phân giải đồng tham chiếu và cho ta được kết quả là các chuỗi đồng tham chiếu của hồ sơ xuất viện bất kì.

4.3 Xây dựng các cặp khái niệm

Ở bước này, đầu tiên các khái niệm ở dữ liệu đầu vào sẽ được tiền xử lý 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.

Sau khi các khái niệm đã được tiền xử lý, các 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. Ở đây nhóm không sinh các cặp mà có một trong hai khái niệm thuộc lớp Pronoun, vì việc phân giải đồng tham chiếu của lớp này có một chút khác biệt mà nhóm sẽ nói rõ hơn ở phần sau.

4.4 Thiết kế tập thuộc tính đặc trưng

Từ các phân tích được đề cập ở phần 3, ngoài các thuộc tính chung về mặt ngôn ngữ (như ngữ pháp hay ngữ nghĩa), các lớp thực thể ở BAĐT còn mang những đặc tính khác nhau. Việc này

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), các trường hợp 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	Nếu khái niệm đứng trước là từ "who" (1), không phải (0)
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	Cả hai cùng chứa cùng một chức vụ bác sĩ (1), nếu không (0)
Doctor general match	0, 1	Cả hai cùng đề cập đến bác sĩ chung (1), không (0)
Twin/triplet	0, 1	Cả hai đều chỉ về cùng cặp sinh đôi/sinh ba (1), ngược lại (0)
We	0, 1	Cả hai đều chứa thông tin về "chúng tôi" (1), ngược lại (0)
You	0, 1	Cả hai đều chứa thông tin về "tôi" (1), ngược lại (0)
I	0, 1	Cả hai đều chứa thông tin về "bạn" (1), ngược lại (0)
Pronoun match	0, 1	Khái niệm đứng trước là một đại từ (1), ngược lại (0)

Bảng 4.1. Tập thuộc tính của cặp hai khái niệm lớp Person

đòi hỏi nhóm phải thiết kế ba hệ thống trích xuất đặc trưng và ba hệ thống phân loại tương ứng khác nhau cho ba lớp Person, các khái niệm thuộc lớp Problem/Treatment/Test và lớp Pronoun. Hình 4.3 mô tả tổng quan ba hệ thống này, trong đó các module "Đồng tham chiếu cho lớp X" mang ý nghĩa bao hàm cả Hệ thống rút trích đặc trưng và Hệ thống phân loại cho lớp tương ứng. Sau đây là chi tiết các phương pháp trích xuất đặc trưng của nhóm cho mỗi lớp thực thể trong hồ sơ xuất viện tiếng Anh.

Đặc trưng lớp Person

Một đặc tính nổi bật của mối quan hệ đồng 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 nói chung 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 BAĐT thì công việc này sẽ dễ hơn rất nhiều. Một BAĐT 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 lớn 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.

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

Bảng 4.2. Tập thuộc tính của một khái niệm lớp Person
dùng cho việc xác định có là bệnh nhân hay không

Các đặc trưng thường được sử dụng ở các hệ thống phân giải đồng tham chiếu cho văn bản nói chung thường là không đủ, vì đa phần chúng không xét đến các tính chất khác biệt của một phạm vi văn bản cụ thể. 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 chỉ thuộc một trong ba lớp: *bệnh nhân*, *người thân của bệnh nhân* hoặc *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) cho cặp hai khái niệm chỉ người. Bảng 4.1 trình bày tập các thuộc tính dùng cho lớp Person. Mô hình phân loại được sử dụng cho cặp các khái niệm Person là SVM.

Thuộc tính Patient-class

Để xác định một khái niệm có đề cập đến bệnh nhân hay không, nhóm huấn luyện một hệ thống SVM để phân loại nó. 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 chắc chắn sẽ được đưa vào chuỗi đồng tham chiếu duy nhất về bệnh nhân đó. Do vậy mục đích của thuộc tính này là xác định xem một khái niệm có ám chỉ đến bệnh nhân hay không.

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 Bảng 4.2. Kết quả của việc phân loại được lấy làm giá trị cho thuộc tính Patient-class ở Bảng 4.1.

Đặc trưng lớp Problem/Treatment/Test

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 vì có sự ảnh hưởng bởi ngữ cảnh mà chúng được đề cập đến. 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.

Nhóm lớp Problem/Treatment/Test là nhóm lớp 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. Việc xác định các từ đồng nghĩa có thể giúp giảm sai sót và tăng độ chính xác cho quá trình học máy. Để tìm được các từ đồng nghĩa không có trong tập huấn luyện, ta cần sử dụng nguồn thông tin có sẵn từ bên ngoài (kiến thức nền). 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 lý do đó, nhóm cho rằng cần xây dựng các bộ trích xuất ngữ cảnh cho lớp Problem/Treatment/Test, phục vụ việc phân giải đúng đồng tham chiếu cho các khái niệm thuộc lớp này. Một số nguồn kiến thức nền có thể được sử dụng như:

Wikipedia
: Đây là một bộ bách khoa toàn thử mở và miễn phí. Các thông tin trong này có thể được sử dụng để xác định tên giả (alias), các từ viết tắt hay từ đồng nghĩa.

WordNet
: Được sử dụng để tìm kiếm các từ đồng nghĩa xuất hiện trong các khái niệm.

UMLS
: Viết tắt của Unified Medical Language System, là một hệ thống ngôn ngữ y khoa đồng nhất. Nó được phát triển bởi Thư viện y tế quốc gia Hoa Kì nhằm mang đến một cơ sở dữ liệu chung về các thuật ngữ y sinh cũng như các mối quan hệ về ngữ nghĩa giữa chúng.

Các 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 dù có cùng cách viết nhưng lại không đồng tham chiếu với "Đau" ở chân, hai 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 và các thông tin liên quan để giúp phân biệt các khái niệm có ngữ nghĩa hoặc vị trí thời gian khác nhau.

Trích xuất các thông tin về cơ quan trên cơ thể (Anatomy)
: Hai khái niệm tuy có cũng ngữ nghĩa nhưng lại được đề cập ở hai vị trí khác nhau trên cơ thể thì không đồng tham chiếu. Ví dụ trong một hồ sơ xuất viện có chứa hai câu sau: "The patient continued to suffer from edema of the left upper extremity and a vascular radiology consult revealed a *thrombosis of the left subclavian vein* extending into the axillary vein." và " There was some *thrombosis of the left internal jugular vein* as well.".

Mặc dù chứa cùng một triệu chứng là *thrombosis* (chứng huyết khối) nhưng triệu chứng này lại xuất hiện ở hai nơi khác nhau là *subclavian vein* (tĩnh mạch dưới đòn) và *jugular vein* (tĩnh mạnh cổ) nên hai khái niệm này không đồng tham chiếu với nhau. Vì thế để phân biệt rõ ở trường hợp này, so trùng chuỗi là không đủ, ta cần thiết phải xác định được cơ quan trên cơ thể mà hai khái niệm được đề cập tới. Việc xác định cơ quan này có thể thực hiện bằng cách sử dụng nguồn kiến thức ngoài như UMLS.

Trích xuất thông tin về vị trí (Position)

Một số cơ quan khác trên cơ thể tuy khác nhau về mặt ngữ nghĩa nhưng lại chỉ phân biệt được bằng một số tính từ chỉ vị trí đi kèm theo nó, ví dụ "The patient has *burning sensation* in the upper left leg" và "The patient has *burning sensation* in the lower right leg". Mặc dù cùng một triệu chứng là *burning sensation* (cảm giác rát) nhưng nó lại xuất hiện ở hai cơ quan khác nhau là *upper left leg* (phía trên chân trái) và *lower right leg* (phía dưới chân phải), và hai cơ quan này chỉ có thể được phân biệt bởi các từ chỉ vị trí đi kèm theo nó (về mặt từ ngữ). Chính vì thế, nhóm sẽ xây dựng một từ điển tra cứu các từ chỉ vị trí dựa trên tập dữ liệu.

Trích xuất thông tin về thuốc (Medical Information)

Các thông tin về thuốc như tên thuốc, liều lượng, tần suất sử dụng, đường hấp thụ (ví dụ đường uống, tiêm hay bôi), thời gian sử dụng, lý do sử dụng, dạng "liệt kê" hay "tường thuật" cần phải được xác định khi phân giải đồng tham chiếu cho hai khái niệm thuộc lớp Treatment. Khi một trong các thông tin trên khác nhau thì hai khái niệm sẽ không đồng tham chiếu. Vì thế, một bộ từ điển tra cứu cần được xây dựng để phục vụ cho việc trích xuất các thông tin thuốc này.

Trích xuất các chỉ số của thủ tục y tế (Indicator)

Một thủ tục y tế (tức khái niệm thuộc lớp Test) thường đi kèm với một hoặc một vài chỉ số mô tả chi tiết thủ tục đó, ví dụ "wbc", "rbc", "hct" và "hgb". Khi hai khái niệm lớp Test chứa chỉ số có giá trị khác nhau, chúng sẽ không đồng tham chiếu với nhau. Như vậy cần thiết phải xác định rõ các chỉ số khi phân giải đồng tham chiếu cho các khái niệm thuộc lớp Test.

Trích xuất thông tin về thời gian (Temporal)

Thông tin thời gian là một thông tin rất có ích cho phân giải đồng tham chiếu ở ba lớp Problem/Treatment/Test. Một thủ tục y tế được tiến hành ở hai thời điểm khác nhau thì không đồng tham chiếu hay cùng một loại thuốc nhưng được kê khai ở các thời điểm khác nhau thì độc lập. Thông tin thời gian ở tác vụ này được chia làm hai loại: thời gian tường minh (ví dụ "03/06/2015") và thời gian suy diễn (là các mốc thời gian được trích xuất từ một số từ khóa đứng gần khái niệm đang xét, như *admission date* (ngày nhập viện) hay *transfer date* (ngày thuyên chuyển)).

Trích xuất thông tin về không gian (Spatial)

Cũng như thời gian, không gian là một ngữ cảnh quan trọng mà ta cần phải xét tới khi phân giải đồng tham chiếu các khái niệm thuộc lớp Problem/Treatment/Test. Ví dụ như cùng một loại thuốc nhưng một khái niệm xuất hiện ở phòng phẫu thuật và một khái niệm xuất hiện ở phòng hồi sức thì hai khái niệm này không đồng tham chiếu với nhau.

Thuộc tính	Giá trị	Giải thích
Kiến thức nền		
Wiki page match	0, 1	Hai khái niệm cùng chỉ đến một trang Wiki
Wiki anchor match	0, 1	Hai khái niệm cùng chỉ đến một liên kết trên trang Wiki
Wiki bold name match	0, 1	Hai khái niệm có cùng tên thực thể trên Wiki
WordNet match	0, 1	Hai khái niệm đồng nghĩa trên WordNet
Trích xuất thông tin ngữ cảnh		
Anatomy	0, 1, 2	Không cùng cơ quan (0), cùng cơ quan (1), không xác định (2)
Position	0, 1, 2	Không cùng vị trí (0), cùng vị trí (1), không xác định (2)
Indicator	0, 1, 2	Không cùng chỉ số (0), cùng chỉ số (1), không xác định (2)
Temporal	0, 1, 2	Không cùng thời gian (0), cùng thời gian (1), không xác định (2)
Spatial	0, 1, 2	Không cùng không gian (0), cùng không gian(0), không xác định (2)
Section	1, 2, ..., n^2	Hai khái niệm thuộc về mục i và j của văn bản
Equipment	0, 1, 2	Không cùng thiết bị (0), cùng thiết bị (1), không xác định (2)
Operation	0, 1, 2	Không cùng hành động (0), cùng hành động (1), không xác định (2)
Assertion	0, 1, 2, ..., 6^2	Trạng thái lớp khẳng định của hai khái niệm
Trích xuất thông tin thuốc		
Drug	0, 1	Cùng tên thuốc (1), ngược lại (0)
Mode	0, 1, 2, ..., 29	Chỉ mục của 29 đường hấp thụ, hoặc không xác định (29)
Dosage	0, 1	Cùng liều lượng (1), ngược lại (1)
Duration	0, 1	Cùng khoảng thời gian sử dụng (1), ngược lại (0)
Frequency	0, 1	Cùng tần suất sử dụng (1), ngược lại (0)
"List" or "Narrative"	0, 1	Cùng là dạng liệt kê (0) hay tường thuật (1)
Time of first mention	0, 1, 2, 3	Thời gian của lần đề cập đầu tiên: Quá khứ (0), hiện tại (1), tương lai (2), không xác định (3)
Time of second mention	0, 1, 2, 3	Thời gian của lần đề cập thứ 2: Quá khứ (0), hiện tại (1), tương lai (2), không xác định (3)
Episode of first mention	0, 1, 2, 3	Chương của lần đề cập đầu tiên: bắt đầu (0), tiếp diễn (1), tạm ngưng (2), không xác định (3)
Episode of second mention	0, 1, 2, 3	Chương của lần đề cập thứ 2: bắt đầu (0), tiếp diễn (1), tạm ngưng (2), không xác định (3)
Condition of first mention	0, 1, 2, 3	Tình trạng của lần đề cập đầu tiên: khẳng định (0), gợi ý (1), có điều kiện (2), không xác định (3)
Condition of second mention	0, 1, 2, 3	Tình trạng của lần đề cập thứ 2: khẳng định (0), gợi ý (1), có điều kiện (2), không xác định (3)
Khoảng cách		
Sentence distance	0, 1, 2...	Khoảng cách giữa 2 khái niệm
Ngữ pháp		
Article	1, 2, ..., 3^2	Trạng thái các từ hạn định (determiner) đứng trước hai khái niệm, bao gồm 3 loại: không xác định (a|an), xác định (the|his|her...) hoặc không có (NULL)
So trùng chuỗi		
Head noun match	0, 1	Cùng tiếp đầu ngữ (1), ngược lại (0)
Contains	0, 1	
Capital match	0, 1	Các kí tự đầu tiên trùng nhau (1), ngược lại (0)
Substring match	0, 1	Có cùng chuỗi con (1), ngược lại (0)
Cos distance	(0, 1)	Khoảng cách cos (góc) giữa hai khái niệm
Ngữ nghĩa		
Word match	$\mathbb{Z}$	Số cặp từ trùng nhau của hai khái niệm

Bảng 4.3. Tập thuộc tính của cặp khái niệm lớp Problem/Treatment/Test

	Problem	Treatment	Test
Anatomy	+		+
Position	+	+	+
Medication		+	
Indicator			+
Temporal		+	+
Spatial		+	
Section	+	+	+
Modifier			+
Equipment			+
Operation		+	
Assertion	+		

Bảng 4.4. Phân chia các đặc trưng được sử dụng
ở ba lớp Problem, Treatment và Test

Trích xuất thông tin về vị trí xuất hiện của khái niệm trong bệnh án (Section)

Một hồ sơ xuất viện thường được chia làm các phần (section) như: tiền sử bệnh, tiền sử dùng thuốc, tiền sử nhập viện,... Hai khái niệm xuất hiện ở hai phần khác nhau thì thường không đồng tham chiếu với nhau cho dù chúng có cùng cách viết. Ví dụ cụm "CT scan" xuất hiện ở phần tiền sử điều trị và phần xét nghiệm thể chất thì hai khái niệm "CT scan" này là độc lập với nhau.

Trích xuất bổ từ (Modifier)

Tính đồng tham chiếu của một số khái niệm thuộc lớp Test có thể được xác định bởi các từ bổ nghĩa đi kèm theo chúng. Ví dụ các từ "recent", "prior" hay "initial".

Trích xuất thông tin về thiết bị y tế (Equipment)

Một số thủ tục y tế thường được đặt tên theo các thiết bị có liên quan, các từ này thường có phần hậu tố "-gram", "-metry", "-scopy" hay "-graphy".

Trích xuất thông tin về hành động (Operation)

Các phương pháp điều trị cho bệnh nhân có thể là các hành động phẫu thuật. Các phương pháp này thường kết thúc bởi "-plasty" hay "-tomy".

Trích xuất thông tin về sự khẳng định (Assertion)

Dựa vào Thách thức i2b2 năm 2010 về việc phân loại các khái niệm thuộc lớp Problem thành các phân lớp khẳng định bao gồm *có thể* (possible), *vắng mặt* (absent), *hiện tại* (present), *có điều kiện* (conditional) và *giả thuyết* (hypothetical), nhóm nhận thấy thông tin về lớp khẳng định cũng đóng vai trò quan trọng trong việc phân giải đúng tính đồng tham chiếu giữa hai khái niệm lớp Problem. Vì vậy một trong những cách hiện thực bộ trích xuất này là sử dụng lại các giải pháp ở Thách thức i2b2 năm 2010.

Từ các thông tin về ngữ cảnh trên cùng với các thuộc tính đặc trưng về mặt ngữ pháp, nhóm tổng kết tập các thuộc tính dùng cho phân giải đồng tham chiếu ở lớp Problem/Treatment/Test trên Bảng 4.3. Vì các thuộc tính đặc trưng được trình bày ở đây không phải thuộc tính nào cũng dùng chung cho cả ba lớp Problem, Treatment và Test (ví dụ đặc trưng về thông tin thuốc chỉ có ở các khái niệm thuộc lớp Treatment) nên Bảng 4.4 mô tả một cách trực quan về vấn đề này.

Thuộc tính	Giá trị	Giải thích
First 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 hoặc không thuộc lớp nào (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 hoặc không thuộc lớp nào (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 hoặc không thuộc lớp nào (4)
Sentence distance	0, 1, 2,...	Khoảng cách giữa 2 câu chứa 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 4.5. Tập thuộc tính cho một khái niệm lớp Pronoun

Đặc trưng lớp Pronoun

Khác với các module đánh giá đồng tham chiếu của các lớp Person, Problem/Test/Treatment, module đánh giá đồng tham chiếu của lớp Pronoun nhận vào chỉ 1 khái niệm duy nhất thay vì một cặp 2 khái niệm. Mỗi đại từ được đưa vào module có thể đồng tham chiếu tới các khái niệm khác hoặc độc lập. Để giải quyết việc đồng tham chiếu cho lớp Pronoun, ta cần xác định đại từ được đưa vào có đồng tham chiếu với khái niệm khác hay không, nếu có thì đồng tham chiếu đến khái niệm thuộc lớp nào. Sau khi xác định được lớp mà đại từ được xét đồng tham chiếu tới, ta sẽ chọn khái niệm gần đại từ được xét nhất và cùng thuộc một lớp để kết luận là 2 khái niệm đồng tham chiếu. Ví dụ "Hepatitis C cirrhosis for **which** the patient was on the liver transplant list", ở đây ta cần xác định đại từ "which" thuộc lớp nào trong các lớp Person, Problem/Test/Treatment. Sau khi xác định được đại từ "which" thuộc về nhóm Problem, khái niệm "Hepatitis C cirrhosis" thuộc cùng lớp "Problem" và ở gần "which" nhất sẽ là khái niệm mà "which" đồng tham chiếu đến. Nhóm quyết định sử dụng mô hình SVM nhiều lớp để phân loại đại từ. Bảng 4.5 mô tả chi tiết tập thuộc đặc trưng của khái niệm lớp Pronoun.

4.5 Xây dựng các cụm khái niệm đồng tham chiếu

Như đã được tìm hiểu ở mục 2.5, một hệ thống phân loại sau khi đã được huấn luyện không có khả năng xây dựng chuỗi đồng tham chiếu mà nó chỉ có thể xác định được một cặp khái niệm là *có đồng tham chiếu* hay *không*. Như vậy cần thiết phải có một giải thuật có chức năng nhóm các khái niệm được xác định là đồng tham chiếu với nhau lại thành từng cụm, từ đó có thể xây dựng các chuỗi đồng tham chiếu từ các cụm này.

Có hai giải thuật gom cụm được đề xuất: *gom cụm gần nhất trước* và *gom cụm tốt nhất trước*. Nhóm quyết định sử dụng giải thuật gom cụm tốt nhất trước cho hệ thống của mình vì hai lý do:

1. Giải thuật gom cụm tốt nhất trước cho kết quả tốt hơn giải thuật gom cụm gần nhất trước [11].

2. Tận dụng một điểm mạnh của mô hình phân loại SVM là nó có khả năng tính toán mức độ tin cậy đồng tham chiếu của các cặp khái niệm.

Giải thuật gom cụm tốt nhất trước làm việc như sau:

1. Từ danh sách các khái niệm, duyệt từ cuối về đầu theo thứ tự xuất hiện, với mỗi khái niệm C_j, xét tất cả các cặp khái niệm (C_i, C_j) mà C_i đứng trước C_j.
2. Chọn ra cặp khái niệm (C_k, C_j) được hệ thống phân loại xác định là đồng tham chiếu và độ tin cậy đồng tham chiếu của cặp là lớn nhất, ta lấy C_k làm tiền đề cho C_j và tất cả các cặp (C_i, C_j) bị loại bỏ khỏi danh sách.
3. Trong trường hợp đã duyệt hết danh sách cặp khái niệm mà không có cặp nào được xác định là đồng tham chiếu, giải thuật cho rằng C_j là khái niệm duy nhất và tất cả các cặp (C_i, C_j) được loại bỏ.
4. Lặp lại bước 1 cho đến khi đã duyệt hết các khái niệm trong văn bản.

Có thể nhận thấy rằng mức độ chính xác của giải thuật gom cụm phụ thuộc rất nhiều vào độ chính xác của hệ thống phân loại. Mặt khác, một số nhược điểm của mô hình cặp khái niệm cũng ảnh hưởng đến kết quả phân giải đồng tham chiếu. Tuy nhiên với những đặc điểm khác biệt của BAĐT mà nhóm đã rút ra được trong phần 3, thay vì sử dụng một mô hình phân giải tốt hơn (như mô hình đề cập thực thể hay mô hình xếp hạng), nhóm quyết định tập trung vào tối ưu công đoạn thiết kế thuộc tính với mục tiêu là tăng hiệu xuất của hệ thống phân loại.

5 Thí nghiệm đánh giá

5.1 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 có thể lấy 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. Tập dữ liệu mà nhóm nhận được bao gồm: *251 hồ sơ xuất viện* cho tập huấn luyện và *175 hồ sơ* cho tập kiểm tra. Trong đó mỗi hồ sơ xuất viện đi kèm với một tập tin chứa danh sách các khái niệm được đề cập trong hồ sơ đó và được gán nhãn theo mẫu:

c="<concept>" <begin> <end> || t="<class>"

Ví dụ: *c="which" 20:5 20:5 || t="pronoun"* có ý nghĩa là khái niệm “which” xuất hiện ở dòng 20 kí tự thứ 5, kết thúc ở dòng 20 kí tự thứ 5 và thuộc lớp Pronoun.

Ngoài danh sách khái niệm, mỗi hồ sơ còn đi kèm với một tập tin chứa chuỗi đồng tham chiếu đã được phân giải (*ground truth*) nhằm huấn luyện các hệ thống phân loại có giám sát, các chuỗi kết quả này có định dạng:

c="<concept>" <begin> <end> || c="<concept>" <begin> <end> || … || t="<class>"

5.2 Phương pháp đánh giá

Hiệu năng của hệ thống được đánh giá qua ba độ đo: *độ đúng đắn* (precision), *độ đầy đủ* (recall) và *độ F* (F-measure). Các độ đo này được tính bằng ba hệ đo khác nhau [3]: MUC, B-CUBED và CEAF, mỗi hệ có ưu điểm và nhược điểm riêng. Trung bình không trọng số của mỗi độ đo được tính bởi ba hệ đo trên sẽ được lấy làm kết quả cuối cùng để đánh giá các chuỗi đồng tham chiếu xuất ra bởi 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 G là tập các chuỗi kết quả, S là tập các chuỗi được xuất ra bởi hệ thống, g và s là chuỗi đồng tham chiếu từ tập G và S tương ứng. Các độ đúng đắn (P) và độ đầy đủ (R) của hệ MUC được tính như sau:

$$P = \frac{\sum_s \big(|s| - m(s, G)\big)}{\sum_s (|s| - 1)}$$

$$R = \frac{\sum_g \big(|g| - m(g, S)\big)}{\sum_g (|g| - 1)}$$

trong đó $m(s, G)$ được định nghĩa là số chuỗi trong G có giao nhau với chuỗi s, $|g|$ số khái niệm tạo thành chuỗi g.

Độ F của hệ MUC là trung bình điều hòa của độ chính xác và độ đầy đủ:

$$F = \frac{2 \times R \times P}{R + P}$$

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 C là tập N tài liệu, d là một tài liệu trong C và m là một khái niệm trong d. Ta định nghĩa G_m là chuỗi kết quả có chứa m và S_m là chuỗi của hệ thống có chứa m. O_m là chuỗi giao nhau giữa G_m và S_m. Độ đúng đắn và đầy đủ của hệ B-CUBED được tính như sau:

$$P = \frac{1}{N} \sum_{d \in C} \sum_{m \in d} \frac{|O_m|}{|S_m|}$$

$$R = \frac{1}{N} \sum_{d \in C} \sum_{m \in d} \frac{|O_m|}{|G_m|}$$

Độ F 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 $\Phi(g^*)$ 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, φ_3 và φ_4; nhóm sử dụng φ_4.

Gọi tập các chuỗi kết quả của một tài liệu d là $G(d) = \{G_i : i = 1,2,\dots,|G(d)|\}$, và tập các chuỗi của hệ thống cho một tài liệu d là $S(d) = \{S_i : i = 1,2,\dots,|S(d)|\}$, G_i và S_i là một chuỗi trong $G(d)$ và $S(d)$ tương ứng. Độ tương tự dựa trên chuỗi được tính như sau:

$$\varphi_3(G_i, S_i) = |G_i \cap S_i|$$

$$\varphi_4(G_i, S_i) = \frac{2|G_i \cap S_i|}{|G_i| + |S_i|}$$

Độ đúng đắn và độ đầy đủ của hệ CEAF được tính như sau:

$$P = \frac{\Phi(g^*)}{\sum_i \varphi(S_i, S_i)}$$

$$R = \frac{\Phi(g^*)}{\sum_i \varphi(G_i, G_i)}$$

Độ F được tính tương tự như hệ MUC.

6 Tổng kết

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.

Sau quá trình tìm hiểu, nhóm quyết định đề xuất bài toán cụ thể "*Phân giải đồng tham chiếu trên bệnh án điện tử với các khái niệm đã được biết trước cho dữ liệu tiếng Anh*". 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 2 quy trình: quy trình huấn luyện và quy trình phân giải đồng tham chiếu. Quy trình huấn luyện có mục đích là xây dựng hệ thống phân loại dựa trên mô hình học máy có giám sát có khả năng đánh giá độ tin đồng tham chiếu của các cặp khái niệm dựa trên tập huấn luyện. Hệ thống phân loại sau khi được huấn luyện sẽ được sử dụng cùng với giải thuật gom cụm để xác định các chuỗi đồng tham chiếu trong bệnh án điện tử bất kì.

Ngoài ra, trong giai đoạn luận văn, nhóm sẽ xây dựng hệ thống hỗ trợ tiếng Việt cho thiết kế đề xuất. Trong quá trình tìm hiểu, nhóm nhận thấy rằng dữ liệu tiếng Việt và tiếng Anh chỉ khác nhau ở các đặc trưng ngữ cảnh cần được trích xuất. Vì vậy khi xây dựng hệ thống hỗ trợ tiếng Việt, nhóm chỉ cần thay thế module trích xuất đặc trưng của dữ liệu tiếng Anh bằng các đặc trưng phù hợp cho dữ liệu tiếng Việt. Khi đã trích xuất được các đặc trưng, nhóm có thể sử dụng mô hình đã được đề xuất với thiết kế không thay đổi. Từ đó, hệ thống có thể hỗ trợ bệnh án điện tử của cả 2 ngôn ngữ tiếng Việt và tiếng Anh.

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

- Tìm hiểu thiết kế tập thuộc tính đặc trưng cho dữ liệu tiếng Việt.
- Trích xuất các thực thể và gán nhãn cho dữ liệu tiếng Việt.
- Xây dựng hệ thống phân giải đồng tham chiếu.
- Hoàn thiện hệ thống.
- Đánh giá hệ thống và so sánh với một phương pháp nền [4]

Tài liệu tham khảo

[1] Hồ Tú Bảo, "Xây dựng và khai thác BAĐT: con đường mới trong khám chữa bệnh và nghiên cứu y học," *Khoa học & Công nghệ Việt Nam*, vol. 56, no. 3, pp. 16-20, 2015.

[2] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, no. 5, pp. 552-556, 2011.

[3] Özlem Uzuner et al., "Evaluating the state of the art in coreference resolution for electronic medical records," *Journal of the American Medical Informatics Association : JAMIA*, vol. 19, no. 5, pp. 786-791, 2012.

[4] Yan Xu et al., "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.

[5] Vincent Ng., "Supervised noun phrase coreference research: the first fifteen years," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 1396-1441.

[6] J Hobbs, "Resolving Pronoun References," in *Readings in Natural Language Processing*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1986, pp. 339-352.

[7] Susan E Brennan, Marilyn W. Friedman, and Carl J. Pollard, "A Centering Approach to Pronouns," in *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1987, pp. 155-162.

[8] Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman, "Learning to Resolve Bridging References," in *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004, p. 143.

[9] Niyu Ge, John Hale, and Eugene Charniak, "A Statistical Approach to Anaphora Resolution," in *Sixth Workshop on Very Large Corpora*. Orlando, Florida, 1998, pp. 161-170.

[10] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim, "A machine learning approach to coreference resolution of noun phrases," *Computational Linguistics*, vol. 27, no. 4, pp. 521-544, December 2001.

[11] Vincent Ng. and Clair Cardie, "Improving machine learning approaches to coreference resolution," in *Proceedings of 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 104-111.

[12] Xiaofeng Yang, Jian Su, and Chew Lim Tan, "A Twin-Candidate Model for Learning-Based Anaphora Resolution," *Computational Linguistics*, vol. 34, no. 3, pp. 327-356, 2008.

[13] Andrew Mccallum and Ben Wellner, "Conditional Models of Identity Uncertainty with Application to Noun Coreference," in *Advances in Neural Information Processing Systems 17*. Cambridge, MA, USA: MIT Press, 2004, pp. 905-912.

[14] Claire Cardie and Kiri Wagstaf, "Noun Phrase Coreference as Clustering," in *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*, 1999, pp. 82-89.

[15] Michael Strube and Stefan Rapp and Christoph Müller, "The Influence of Minimum Edit Distance on Reference Resolution," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing.*: Association for Computational Linguistics, 2002, pp. 312-319.

[16] J. Ross Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[17] William W. Cohen, "Fast Effective Rule Induction," in *Proceedings of the Twelfth International Conference on Machine Learning.*: Morgan Kaufmann, 1995, pp. 115-123.

[18] Walter Daelemans and Antal van den Bosch, *Memory-Based Language Processing*. New York, NY, USA: Cambridge University Press, 2005.

[19] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39-71, 1996.

[20] Yoav Freund and Robert E. Schapire, "Large Margin Classification Using the Perceptron Algorithm," *Machine Learning*, vol. 37, no. 3, pp. 277-296, 1999.

[21] Thorsten Joachims, "Making Large-scale Support Vector Machine Learning Practical," in *Advances in Kernel Methods*. Cambridge, MA, USA: MIT Press, 1999, pp. 169-184.

[22] Stephen Marsland, *Machine Learning: An Algorithmic Perspective.*: Chapman & Hall/CRC, 2009.

[23] Eric Bengtson and Dan Roth, "Understanding the Value of Features for Coreference Resolution," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii, USA: Association for Computational Linguistics, 2008, pp. 294-303.