Tạp chí Tin học Y sinh 42 (2009) 923–936



Danh sách nội dung có sẵn tại [ScienceDirect](http://www.sciencedirect.com/science/journal/15320464)

Tạp chí Tin học Y sinh

Trang chủ tạp chí: [www.elsevier.com/locate/yjbin](http://www.elsevier.com/locate/yjbin)

Trích xuất thông tin dựa trên quy tắc từ dữ liệu lâm sàng của bệnh nhân

Agnieszka Mykowiecka \*, Małgorzata Marciniak, Anna Kups ́c ́ ́

*Viện Khoa học Máy tính PAS, ul. J.K. Ordona 21, 01-237 Warszawa, Ba Lan*

a r t i c l e

i n f o

a b s t r a c t

*Lịch sử bài viết:*

Nhận 27 Tháng Năm 2008

Có sẵn trực tuyến 29 Tháng Bảy 2009

*Từ khoá:*

Trích xuất thông tin dựa trên quy tắc Dữ liệu lâm sàng của Ba Lan

Phân tích ngôn ngữ

Bài báo mô tả một hệ thống trích xuất thông tin dựa trên quy tắc (IE) được phát triển cho các văn bản y khoa Ba Lan. Chúng tôi giới thiệu hai ứng dụng được thiết kế để chọn dữ liệu từ tài liệu y tế bằng tiếng Ba Lan: báo cáo vú và hồ sơ bệnh viện của bệnh nhân tiểu đường. Đầu tiên, chúng tôi đã thiết kế một bản thể học đặc biệt mà sau đó các khái niệm của nó được dịch thành hai mô hình riêng biệt, được biểu diễn dưới dạng hệ thống phân cấp cấu trúc tính năng kiểu (TFS), tuân thủ định dạng theo yêu cầu của nền tảng IE mà chúng tôi đã áp dụng. Sau đó, chúng tôi sử dụng ngữ pháp IE để xử lý tài liệu và điền vào các mẫu do các mô hình cung cấp. Đặc biệt, trong ngữ pháp, chúng tôi đề cập đến các vấn đề ngôn ngữ như: từ khóa mơ hồ, phủ định, phối hợp hoặc biểu thức đồng dụ. Việc giải quyết một số vấn đề này đã được trì hoãn sang giai đoạn xử lý hậu kỳ, nơi thông tin được trích xuất được nhóm thêm và cấu trúc thành các mẫu phức tạp hơn. Để đạt được mục tiêu này, chúng tôi đã xác định các thuật toán heuristic đặc biệt trên cơ sở dữ liệu mẫu. Việc đánh giá các quy trình được thực hiện cho thấy khả năng sử dụng của chúng đối với các nhiệm vụ trích xuất dữ liệu lâm sàng. Đối với hầu hết các tấm tem được đánh giá, độ chính xác và khả năng thu hồi trên 80% đã thu được.

© 2009 Công ty Elsevier Đã đăng ký Bản quyền.

1. Giới thiệu

Dữ liệu lâm sàng tạo thành một nguồn thông tin phong phú về các bệnh tật và kết quả điều trị. Một số tác giả đã chỉ ra rằng việc tiếp cận thông tin này sẽ có giá trị lớn cho nghiên cứu lâm sàng và cho các nghiên cứu theo chiều dọc và dịch tễ học. Tuy nhiên, việc tự động xử lý dữ liệu lâm sàng không phải là điều dễ dàng.

Tại các cơ sở chăm sóc sức khỏe Ba Lan, mô tả khám nghiệm và hồ sơ xuất viện của bệnh nhân được lưu trữ dưới dạng tệp văn bản. Mặc dù chúng có cấu trúc chung chung, nhưng có nhiều sự khác biệt trong cách chúng trình bày dữ liệu bệnh nhân. Nội dung của các tài liệu này phụ thuộc nhiều vào phong cách viết của tác giả và các tiêu chuẩn khác nhau. Do đó, các tệp rất khó phân tích cả thủ công (rất khó để tìm thấy các sự kiện thú vị trong nhiều văn bản dài) và tự động (rất khó để chỉ định các quy tắc để xen kẽ dữ liệu). Do đó, đối với các mục đích hành chính hoặc dịch tễ học, các văn bản phải được chuyển đổi thành các hình thức tiêu chuẩn hơn, tức là được mã hóa và chèn vào các bảng hoặc cơ sở dữ liệu. Tự động hóa quy trình này sẽ rất được hoan nghênh vì nó sẽ tăng cường cả khả năng tiếp cận và độ tin cậy của dữ liệu (mã hóa thủ công là một quy trình dễ xảy ra lỗi).

Tự động xử lý dữ liệu tường thuật y tế đã là một chủ đề tìm kiếm lại trong nhiều thập kỷ, xem [[36,37].](#_bookmark30) Tóm tắt các kết quả đạt được vào năm 1995 được đưa ra trong [[41],](#_bookmark42) trong khi các hệ thống đã

\* Tác giả tương ứng.

*Địa chỉ email:* [agn@ipipan.waw.pl(](mailto:agn@ipipan.waw.pl) A. Mykowiecka), [mm@ipipan.waw.pl(](mailto:mm@ipipan.waw.pl) M. Marciniak), [aniak@ipipan.waw.pl(](mailto:aniak@ipipan.waw.pl) A. Kups'c').

được báo cáo cho đến năm 1999 được mô tả trong [[19]](#_bookmark51). Các hệ thống nổi bật nhất đã được xây dựng trong thời kỳ đó là LSP (Linguistic String Pro- ject, [[36]](#_bookmark30)) được áp dụng cho nhiều lĩnh vực lâm sàng và MedLEE [[18]](#_bookmark49). Hệ thống thứ hai được sử dụng hàng ngày tại Bệnh viện Trưởng lão New York và được áp dụng để xử lý nhiều loại dữ liệu y tế, trong số những loại dữ liệu khác: báo cáo X quang, tóm tắt xuất viện và báo cáo chụp nhũ ảnh [[27].](#_bookmark21) Đây chắc chắn là một ví dụ rất tốt về tính hữu ích của MLP (Xử lý ngôn ngữ tự nhiên y tế). Ngoài một vài ứng dụng thành công, xử lý tự động dữ liệu lâm sàng vẫn là một nhiệm vụ rất khó khăn. Trong mười năm qua, nó đã được đề cập trong nhiều dự án được mô tả trong các bài báo và sách khác nhau, ví dụ, [[26,13],](#_bookmark21) xem thư mục trong [[12](#_bookmark31)]. Các dự án nổi tiếng nhất liên quan đến việc xử lý tài liệu tiếng Anh, nhưng cũng có công việc xử lý tài liệu y tế bằng tiếng Đức [[21],](#_bookmark21) tiếng Hà Lan [[42]](#_bookmark43) và tiếng Bulgaria [[3]](#_bookmark22). Chúng tôi không biết về bất kỳ kinh nghiệm tương tự nào về dữ liệu lâm sàng của Ba Lan.

Các kỹ thuật xử lý ngôn ngữ tự nhiên có thể được chia thành hai luồng chính: phương pháp tiếp cận dựa trên quy tắc theo định hướng ngôn ngữ hơn và xử lý thống kê. Ngày nay, phương pháp thứ hai phổ biến hơn và được quan tâm nhiều hơn, vì kết quả có thể đạt được nhanh hơn. Ví dụ, trong lĩnh vực y sinh có một số lượng lớn các dự án về chiết xuất thuật ngữ y sinh học bằng phương pháp học máy (ML) (ví dụ: [[4,47,44,5,6]](#_bookmark23)). Xử lý ngôn ngữ tự nhiên dựa trên quy tắc trở nên ít phổ biến hơn vì nó đòi hỏi nhiều công việc thủ công và không dễ sử dụng lại. Không bao giờ, đặc biệt là để trích xuất các mẫu phức tạp, có cấu trúc, các phương pháp tiếp cận dựa trên quy tắc chính thức cũng được sử dụng và cho kết quả đáng tin cậy. MedLEE [[13]](#_bookmark41)  là một hệ thống dựa trên quy tắc hoạt động dựa trên kết quả của phân tích cú pháp nông và đề cập đến ngữ nghĩa

1532-0464 / $ - xem vấn đề © trước 2009 Elsevier Inc. Đã đăng ký Bản quyền. doi:[10.1016/j.jbi.2009.07.007](http://dx.doi.org/10.1016/j.jbi.2009.07.007)

từ vựng. Việc tự động học các cấu trúc phức tạp rất khó khăn, đặc biệt nếu không có nhiều dữ liệu lâm sàng để đào tạo và thử nghiệm, và không có kho dữ liệu lâm sàng được chú thích ngữ nghĩa. Tuy nhiên, một số nỗ lực để tìm hiểu các mẫu phức tạp để xử lý dữ liệu lâm sàng cũng đang được tiến hành. Một hệ thống được xây dựng bởi Taira và Soderland [[43,40]](#_bookmark45)  sử dụng một bộ phân loại entropy tối đa để xác định ranh giới câu; một máy phân tích từ vựng dựa trên các từ vựng được tạo thủ công với các đặc điểm cú pháp và ngữ nghĩa; một trình phân tích cú pháp thống kê xây dựng các cấu trúc phụ thuộc; và hai loại trình thông dịch ngữ nghĩa (quy trình học quy tắc và bộ phân loại entropy tối đa); và cuối cùng là một bộ lấp đầy khung dựa trên quy tắc. Kết hợp cả phương pháp tượng trưng và máy học có lẽ là giải pháp hứa hẹn nhất. Một trong những hệ thống như vậy là MPLUS [[14]](#_bookmark44)  cung cấp phân tích cú pháp dựa trên ngữ pháp không có ngữ cảnh và phân tích ngữ nghĩa bằng cách sử dụng mạng Bayes. Một trong những ứng dụng của hệ thống là trích xuất thông tin y tế từ các báo cáo CT đầu.

Dữ liệu bệnh nhân có thể chứa thông tin về rất nhiều chủ đề nhưng việc trích xuất tất cả chúng sẽ đòi hỏi các mô hình miền rất phong phú. Đó là lý do tại sao trong các bài báo liên quan đến xử lý và cấu trúc thông tin lâm sàng thường chỉ có một số đặc điểm được chọn được đề cập, ví dụ: [[17,25,21].](#_bookmark48) Trong ứng dụng của chúng tôi, chúng tôi trích xuất một số lượng tương đối lớn các khái niệm - trên 60. Một ví dụ về công việc tương tự như của chúng tôi là [[11],](#_bookmark32) trình bày một quy trình gán giá trị cho khoảng 70 biến (tính năng) bằng cách áp dụng hệ thống MetaMap gán các khái niệm từ UMLS (Hệ thống ngôn ngữ y tế thống nhất) cho các phân đoạn của văn bản tường thuật. Cách tiếp cận đó sau đó đã được làm phong phú bởi một chương trình độc lập có thể xác định những thay đổi ngữ cảnh và phạm vi phủ định [[10,12]](#_bookmark33). Trong ngữ pháp trích xuất của chúng tôi, toàn bộ các cụm từ thể hiện dữ liệu thích hợp được xác định cùng một lúc.

Tương tự như những người khác xử lý dữ liệu lâm sàng (ví dụ: [[24]](#_bookmark21)), chúng tôi đã quyết định xây dựng một hệ thống dựa trên quy tắc cổ điển. Quyết định này được thúc đẩy bởi những lý do được đề cập ở trên: sự phức tạp của các mẫu khó điền bằng cách sử dụng các phương pháp học máy và thực tế là không có kho dữ liệu y tế có chú thích đầy đủ cho tiếng Ba Lan. Tuy nhiên, chúng tôi có kế hoạch sử dụng hệ thống IE của mình để tăng tốc độ xây dựng kho dữ liệu y tế có chú thích có thể được sử dụng để đào tạo và thử nghiệm các phương pháp ML khác nhau trong các thí nghiệm tiếp theo.

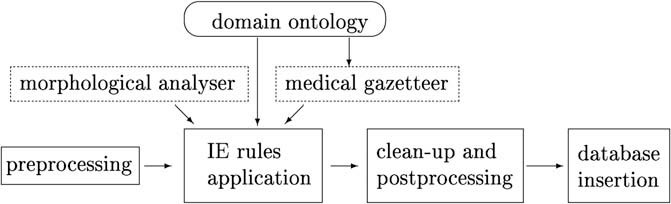
Các ứng dụng IE cho các miền tương tự được tạo cho văn bản bằng các ngôn ngữ khác không thể được sử dụng lại cho dữ liệu tiếng Ba Lan. Mặc dù nghiên cứu về các ứng dụng độc lập ngôn ngữ đã được thực hiện trong một thời gian, nhưng không có kỹ thuật nào cho phép chuyển giao một hệ thống (đặc biệt là hệ thống dựa trên quy tắc) được viết cho các văn bản bằng ngôn ngữ này sang ngôn ngữ khác. Ngoài ra, trong các ứng dụng lâm sàng, chúng ta cũng phải đối phó với các tiêu chuẩn mô tả khác nhau và đặc tính uốn của ngôn ngữ Ba Lan khiến nhiệm vụ trở nên phức tạp hơn so với tiếng Anh. Do đó, đối với các văn bản được viết bằng tiếng Ba Lan, chúng tôi phải phát triển một hệ thống IE chuyên dụng.

Trong bài báo, chúng tôi trình bày hai thí nghiệm IE khác nhau.[1](#_bookmark1) Mục tiêu của ứng dụng đầu tiên của chúng tôi là trích xuất thông tin chi tiết về mô vú và các phát hiện bệnh lý được mô tả trong các báo cáo chụp nhũ ảnh. Nhiệm vụ thứ hai liên quan đến việc lựa chọn thông tin y tế quan trọng về bệnh nhân tiểu đường từ các tài liệu xuất viện của họ bao gồm: mô tả tình trạng sức khỏe khi bắt đầu và kết thúc thời gian nhập viện; mô tả bệnh tật của bệnh nhân và cách điều trị; và kết quả của tất cả các kỳ thi.

Chiến lược xử lý được áp dụng dựa trên kiến thức miền và phân tích ngôn ngữ nông cạn. Luồng thông tin chung và các giai đoạn xử lý được thể hiện trong [Hình 1](#_bookmark0).

Tổ chức phần còn lại của bài báo như sau. Phần [2](#_bookmark2) bao gồm mô tả chung về phương pháp của chúng tôi. Phần [3–5](#_bookmark3) trình bày các khía cạnh thú vị nhất của tất cả các giai đoạn, hệ thống xử lý

1 Công việc này được tài trợ một phần bởi dự án quốc gia Ba Lan 3 T11C 007 27.



Hình 1. Các giai đoạn xử lý dữ liệu.

đánh giá được cung cấp trong Phần [6](#_bookmark14) trong khi Phần [7](#_bookmark19) bao gồm một phân tích về các kết quả. Phần [8](#_bookmark20) kết luận bài báo.

1. Phương pháp
   1. *Tạo hệ thống*

Mô hình tên miền và các ứng dụng IE được xây dựng trên cơ sở phân tích dữ liệu thực tế từ các tổ chức chăm sóc sức khỏe Warsaw. Dữ liệu chụp nhũ ảnh bao gồm 2439 báo cáo được viết bởi các bác sĩ chuyên về giải thích chụp X-quang tuyến vú. Từ bộ này, chúng tôi đã chọn một bộ thử nghiệm gồm 867 tài liệu, không được kiểm tra trong quá trình tạo ra ứng dụng. Dữ liệu bệnh tiểu đường bao gồm 606 tài liệu xuất viện từ những năm: 2001–2006. Tất cả các tài liệu đều từ một khu vực chuyên điều trị bệnh tiểu đường và được viết bởi các bác sĩ chuyên khoa. Tất cả 169 tài liệu từ năm 2006 không được kiểm tra cho đến giai đoạn đánh giá.

Để xác định thông tin cần được trích xuất, chúng tôi đã định nghĩa một bản thể y học. Mô hình miền cung cấp cơ sở để cấu trúc thông tin văn bản thành các mẫu được sử dụng trong IE. Một số khái niệm bản thể được đưa vào từ vựng miền (công báo) cùng với nền tảng xử lý. Các từ vựng chứa, trong số các mục khác, các thuật ngữ chuyên biệt đóng vai trò là từ khóa cho IE. Ontology được xây dựng dựa trên dữ liệu và kiến thức chuyên môn.

Xử lý văn bản đã được chia thành nhiều bước. Đầu tiên, các tài liệu chúng tôi có được, yêu cầu xử lý ban đầu. Điều này chủ yếu liên quan đến việc kiểm tra chính tả và chuyển đổi định dạng. Cốt lõi của phân tích văn bản của chúng tôi được thực hiện bằng cách xử lý nông bằng cách sử dụng nền tảng IE có mục đích chung. Để xử lý tiếng Ba Lan, nền tảng được tích hợp với một máy phân tích hình thái. Điều này cho phép chúng ta đối phó với biến đổi phong phú của Ba Lan (7 trường hợp, 2 số, 3 giới tính) và sử dụng các dạng cơ bản của các từ trong các quy tắc. Các dạng biến đổi của các thuật ngữ chuyên ngành, nếu chúng không được công cụ phân tích hình thái nhận ra, được liệt kê chính xác trong các từ vựng miền được đề cập ở trên. Trong IE, các quy tắc ngữ pháp nhằm mục đích lưu trữ các phần tử được nhận dạng trực tiếp trong các tấm nhưng điều này không phải lúc nào cũng có thể thực hiện được. Các quy tắc chỉ tính đến bối cảnh địa phương, trong khi thông tin cần thiết thường nằm rải rác trong văn bản. Chúng tôi trì hoãn việc thu thập tất cả các mẩu thông tin được công nhận tại địa phương sang giai đoạn tiếp theo, tức là xử lý hậu kỳ. Ở giai đoạn này, thông tin được hoàn thành và chỉnh sửa để điền vào các mẫu phức tạp hơn. Đây hóa ra là một nhiệm vụ khá khó khăn, đặc biệt là đối với các báo cáo chụp nhũ ảnh, do sự phức tạp của các mẫu và dữ liệu. Cuối cùng, các mẫu được chèn vào cơ sở dữ liệu quan hệ để có thể dễ dàng truy cập và tìm kiếm.

* 1. *Quy trình đánh giá*

Cả hai ứng dụng của chúng tôi đều được đánh giá trên dữ liệu thực. Các tài liệu được sử dụng để đánh giá được chọn ngẫu nhiên từ các bộ dữ liệu thử nghiệm và bao gồm 705 báo cáo chụp nhũ ảnh và

100 tài liệu xuất viện bệnh tiểu đường của các bệnh nhân khác nhau. Vì chú thích thủ công rất tốn kém (về thời gian và công việc của con người), chúng tôi quyết định áp dụng quy trình chú thích có sự hỗ trợ của máy tính cho

chuẩn bị bộ tham chiếu. Trong bước đầu tiên, dữ liệu xét nghiệm được chú thích bằng các hệ thống IE tương ứng, sau đó các chuyên gia chỉnh sửa chú thích thủ công (có một chuyên gia cho dữ liệu vú và một chuyên gia cho các tài liệu bệnh tiểu đường). Trong khi chỉnh sửa, chuyên gia đã xem xét tài liệu gốc và kiểm tra xem tất cả thông tin có được trích xuất chính xác hay không. Nếu có điều gì đó bị thiếu trong kết quả, chuyên gia sẽ thêm thông tin thích hợp vào tệp và đánh dấu nó bằng '+n'. Nếu một cái gì đó được nhận dạng không chính xác, nó được đánh dấu bằng '—n'. Nếu một thuộc tính được nhận dạng đúng cách nhưng nó được gán sai giá trị hoặc đường viền khối được đặt không chính xác, nó cũng được đánh dấu. Cuối cùng, chúng tôi đếm tất cả các chỉnh sửa được thực hiện bởi các chuyên gia để đánh giá chất lượng của kết quả chiết xuất tự động. Chúng tôi đã đo lường mức độ nhận dạng của các hệ thống các cụm từ và mẫu ký hiệu. Trong lĩnh vực chụp nhũ ảnh, chúng tôi cũng kiểm tra mức độ chèn các điểm đánh dấu khối (không chỉ số lượng mà còn vị trí của chúng).

Quy trình đánh giá có thể thiên về hệ thống

bởi vì người chú thích con người có thể bị ảnh hưởng bởi chú thích tự động. Mặc dù vậy, phương pháp xác minh này dường như là tốt nhất khi rất nhiều thuộc tính phải được nhận dạng - việc nhận dạng thủ công của chúng sẽ mất rất nhiều thời gian và yêu cầu xác minh bổ sung. Sử dụng quy trình này, chúng tôi đánh giá mức độ hệ thống của chúng tôi nhận ra các phần tử của mô hình miền của chúng tôi. Trong trường hợp báo cáo chụp nhũ ảnh, việc đánh giá hệ thống cũng là đánh giá một phần và gián tiếp của chính mô hình. Đôi khi các nhận thức không được công nhận vì chúng tôi không đưa chúng vào mô hình miền của mình. Một vài trường hợp như vậy đã xảy ra và tất cả đều thuộc về "miền border", ví dụ, bóng của máy tạo nhịp tim. Trong bệnh tiểu đường, nơi nhiệm vụ của chúng tôi là nhận ra các khái niệm được chỉ ra bởi những người sử dụng tiềm năng của hệ thống, chúng tôi chỉ quan tâm đến các thuộc tính được xác định trước và giá trị của chúng.

Để đánh giá các phương pháp được trình bày, chúng tôi đã sử dụng độ chính xác, khả năng nhớ lại và kết hợp điểm F của chúng. Đối với mỗi tính năng được thử nghiệm, các biện pháp được xác định như sau:

độ chính xác = Không. của các cụm từ được công nhận đúng là đại diện cho tính năng (TP) Không. của tất cả các cụm từ được công nhận là đại diện cho tính năng (TP + FP)

recall = Không. của các cụm từ được công nhận đúng là đại diện cho tính năng (TP) Không. các cụm từ đại diện cho tính năng (TP + FN)

*Điểm F* = 2 \* (độ chính xác \* nhớ lại)

Độ chính xác + Nhớ lại

Đánh giá của chúng tôi liên quan đến các giá trị của các thuộc tính và dấu ấn được chèn vào để nhóm thông tin. Trong trường hợp thuộc tính, chúng tôi chỉ tính TP những thuộc tính được nhận dạng đúng và được gán các giá trị thích hợp. Những giá trị được gán sai giá trị được tính cùng với các thuộc tính được nhận dạng sai. Để tính toán việc thu hồi, chúng tôi đã tính đến tất cả các phân tích không được chương trình của chúng tôi nhận dạng (được điền thủ công bởi trình chú thích kiểm tra đầu ra) và các thuộc tính được ký kết sai giá trị. Trong trường hợp đánh dấu, TP đếm tất cả các điểm đánh dấu được chèn vào đúng vị trí. FP đếm cả hai điểm đánh dấu được đưa sai vị trí và các điểm đánh dấu hoàn toàn không cần thiết. FN đếm tất cả các điểm đánh dấu hoàn toàn không được nhận ra cũng như những điểm được chèn sai vị trí (các điểm đánh dấu được đặt sai vị trí được tính là cả FP và FN).

1. Mô hình miền

Đối với bất kỳ ứng dụng trích xuất thông tin nào, một định nghĩa thích hợp về mô hình miền là rất quan trọng. Trong trường hợp dữ liệu y tế phức tạp, nhiệm vụ này khá khó khăn và người ta phải quyết định xem có nên sử dụng (điều chỉnh) bất kỳ mô hình nào hiện có hay xây dựng

cái mới. Ngày nay, tồn tại rất nhiều từ vựng có kiểm soát, ví dụ: SNOMED (<http://www.ihtsdo.org/snomed-ct/>) hoặc UMLS (<http://www.nlm.nih.gov/research/umls/>) và bản thể học, ví dụ: BI-RADS ([http://www.birads.at](http://www.birads.at/), [[28]](#_bookmark21)) bao gồm các lĩnh vực khác nhau của kiến thức y sinh. Thật không may, không có nguồn lực nào hiện có (và có sẵn) dường như đủ cho nhiệm vụ của chúng tôi. Trước hết, chúng rất phức tạp (chứa một số lượng lớn các khái niệm) và không được các bác sĩ Ba Lan biết đến. Học chúng sẽ đòi hỏi rất nhiều thời gian trong khi các khái niệm được sử dụng trong thực tế bởi một ứng dụng IE sẽ tương ứng với một tập hợp con rất nhỏ trong số này. Hơn nữa, không có tài nguyên nào có sẵn chứa tất cả các khái niệm cần thiết cho nhiệm vụ hiện tại (để thảo luận ngắn về bản thể vú học, xem [[33]](#_bookmark24)), vì vậy chúng sẽ phải được sửa đổi, điều này không dễ dàng đối với các bản thể y học lớn và đòi hỏi kiến thức chi tiết về cấu trúc của chúng. Một vấn đề khác là các tài nguyên có sẵn được xây dựng dựa trên thuật ngữ khác với tiếng Ba Lan. Các biến thể thuật ngữ rõ ràng hơn nhiều trong dữ liệu lâm sàng so với các bài báo khoa học sử dụng ngôn ngữ chuẩn mực hơn nhiều. Mặt khác, việc xác định một mô hình miền mới cho phép kiểm soát hoàn toàn phạm vi, độ chi tiết và cấu trúc của nó. Do đó, chúng tôi quyết định xác định bản thể OWL nhỏ của riêng mình phản ánh quan điểm của chúng tôi về tên miền phụ đã chọn.

Bản thể được tạo ra [[32]](#_bookmark25)  chủ yếu dựa trên dữ liệu mẫu và kiến thức chuyên môn. Một đoạn trích của các cấp cao nhất của hệ thống phân cấp lớp được đưa ra trong [Hình 2](#_bookmark6). Hình ảnh không hiển thị toàn bộ cấu trúc liên kết - các lớp có các lớp con không được hiển thị được đánh dấu bằng một hình tam giác màu đen. Các lớp được xác định chính là *HumanAnatomy*, *Medicine* và *PhysicalFeature*. Bản thể chỉ bao gồm các tập hợp con của các tên miền phụ y tế có liên quan đến các ứng dụng của chúng tôi. Chúng tôi cũng định nghĩa các mảnh của các bản thể tổng quát hơn, tức là: *PhysicalFeature*, *PhysicalFeatureComparison* và Person. *Lớp Vật lý* bao gồm, trong số những lớp khác, kích thước, đường viền, tổng hợp, mật độ, hình chiếu, tính đều đặn và so sánh số lượng, mức độ, độ bão hòa và thời gian. Bản  *thể thời gian* khá đơn giản và chỉ liên quan đến các biểu thức thời gian xuất hiện trong các tài liệu được phân tích. Nó bao gồm các khoảng thời gian tính bằng năm, tháng và tuần; ngày tháng chính xác và không chính xác; và cả những biểu hiện lặp đi lặp lại như *mọi năm*.

Dưới đây, trong (1) chúng tôi trình bày một danh sách các thuộc tính được xác định cho  *lớp Anat-Finding*.

(1) *Tìm kiếm*:

hasAcompFinding (nhiều AnatFinding) có Appendices (boolean đơn) có AppendicesofShape (một AppendicesShape) hasContour (đường viền đơn)

hasInterpretation (nhiều diễn giải) có Multiplicity (số lượng đơn) hasPalpability (sờ nắn đơn) hasSaturation (độ bão hòa đơn)

hasShape (hình dạng đơn)

Nền tảng IE được áp dụng, SProUT [[15],](#_bookmark46) không thể đọc trực tiếp các bản thể OWL, vì vậy đối với cả hai thí nghiệm, chúng tôi đã dịch thủ công các phần liên quan của bản thể thành hệ thống phân cấp TFS (cấu trúc tính năng được nhập, [[8]](#_bookmark34)) theo định dạng theo yêu cầu của nền tảng. Có hai loại TFS: đơn giản (nguyên tử) và phức tạp. TFS phức tạp là một cấu trúc tính năng được gõ (ma trận thuộc tính-giá trị) chứa các thuộc tính và các giá trị (loại) của chúng, xem ví dụ trong (2). Mỗi cấu trúc được gán một kiểu, tương ứng gần với một lớp từ bản thể, trong khi các thuộc tính tương ứng với các thuộc tính của lớp. Ngoài ra, các cấu trúc TFS phức tạp xác định các mẫu để trích xuất. Một TFS đơn giản chỉ bao gồm một loại (không có cấu trúc). Ví dụ, TFS như vậy được sử dụng để đại diện cho các bộ phận cơ thể hoặc tên thuốc. Một TFS phức tạp đại diện cho một mẫu cho một phát hiện được minh họa trong (2): cấu trúc được cung cấp một kiểu (*finding*\_*str*) và các thuộc tính thích hợp (tính năng). Chẳng hạn

thuộc tính ANAT\_CHANGE có giá trị của  *loại finding*\_*t* có các kiểu con biểu thị các phát hiện khác nhau được chỉ định trong bản thể, ví dụ: tu- mor, mật độ hoặc bóng tối.[2](#_bookmark4)

(2)

2

3

6

6

*Tìm str*

ANAT CHANGE *tìm t* CONTOUR *contour t* MULT *mult t* PALPABILITY *bool t*

BÃO HÒA *SATUR T*

4

7

HÌNH DẠNG *hình dạng t*

77

5

Hệ thống phân cấp TFS chứa các tương đương của tất cả các lớp cấp thấp (khái niệm) từ bản thể và các thuộc tính của chúng. Ví dụ: lớp Contour, một lớp con của *GeneralPhysicalFeature*, đã được dịch thành kiểu *contour*\_*t* và các lớp con của nó được chuyển đổi thành các kiểu con tương ứng (TFS cụ thể hơn). Thuộc tính *hasContour* (được định nghĩa cho  *lớp AnatFinding*) đã được dịch thành TFS dưới dạng thuộc tính CONTOUR nhận giá trị contour\_*t*. Mô hình được tạo cho ứng dụng IE dựa trên dữ liệu: nó chỉ bao gồm thông tin liên quan đến các ứng dụng của chúng tôi. Sau khi chuyển đổi, 176 loại và 66 thuộc tính cho miền chụp nhũ ảnh, và 139 loại (bao gồm 75 tên thuốc) và 65 thuộc tính cho miền tiểu đường, đã được giữ lại trong hệ thống phân cấp TFS. Trong SProUT, để liên kết trực tiếp các chuỗi văn bản và các khái niệm mô hình miền, người ta có thể xác định các từ vựng chuyên biệt (công báo). Các từ vựng này chứa các thuật ngữ cụ thể (ví dụ: tên thuốc), chữ viết tắt dị nguyên và các trường hợp của các khái niệm chính, tức là các từ được sử dụng làm yếu tố kích hoạt cho các quy tắc IE. Đối với cả hai lĩnh vực, chúng tôi đã định nghĩa các từ vựng như vậy trên cơ sở phân tích dữ liệu. Một phần nhỏ của công báo được xác định cho lĩnh vực bệnh tiểu đường được đưa ra trong (3). Mỗi dòng tương ứng với một mục từ vựng riêng biệt và chứa dạng bề mặt (một chuỗi) và hai thuộc tính với các giá trị từ hệ thống phân cấp kiểu. Cấu  trúc CONCEPT đại diện cho kiểu chính xác của chuỗi, trong khi TYPE đại diện cho một kiểu tổng quát hơn, thường là siêu kiểu trực tiếp trong hệ thống phân cấp.

(4) 0: nr\_ksiegi :>

;; Tên quy tắc

1:

2:

3:

4:

5:

6:

7:

(token & [SURFACE ''nr"] | token & [SURFACE ''Nr"] |

morph & [STEM ''numer"]) mã thông báo ?

morph & [STEM ''book ̨ ga"] morph & [STEM ''main"] @seek(liczba\_nat) & [COUNT #nr] ((token & [TYPE slash] |

token & [TYPE back\_slash]) @seek (liczba\_nat) & [LICZ #nr1])?

;; 'số'

;; Mã thông báo tùy chọn

;; 'sách'

;; 'chính'

;; số

8:

9: ->id\_str & [ID #nr, ID\_YEAR #nr1].

;; rạch

;; số

;; đầu ra quy tắc

* mã hóa văn bản (đại khái, tách dấu câu và tách văn bản thành các mã thông báo, được biểu thị dưới dạng cấu trúc TFS của loại *mã thông báo*),
* Phân tích hình thái (các mã thông báo riêng lẻ được giải thích bằng cách sử dụng

một máy phân tích hình thái tổng quát và sau đó được chuyển đổi thành các cấu trúc kiểu biến *hình*),

* tra cứu từ vựng miền (mỗi mã thông báo được kiểm tra dựa trên

từ vựng miền và, nếu tìm thấy một mục tương ứng, nó được liên kết với cấu trúc TFS của kiểu công *báo*).

Tiếp theo, các quy tắc ngữ pháp hoạt động trên các thực thể này và kết hợp chúng thành các cấu trúc phức tạp hơn (TFS) tương ứng với tem- plates. Về mặt kỹ thuật, các quy tắc ngữ pháp là biểu thức chính quy trên TFS, được mở rộng với sự thống nhất. Chúng được mã hóa bằng cách sử dụng XTDL for- malism, vốn có của nền tảng. Mỗi quy tắc XTDL bắt đầu bằng một quy tắc

Tên. Biểu thức chính quy (nội dung quy tắc), được đặt sau ':>'

ký hiệu, mô tả các chuỗi đầu vào, tức là các phần tử phải được xác định trong văn bản, trong khi ký hiệu '->' cho biết cấu trúc đầu ra kết quả. Biểu thức chính quy có thể đề cập đến ba loại cấu trúc đầu vào được đề cập ở trên: *token*, *morph* hoặc *gazet-*

*Teer*. Quy tắc có thể chứa một phần tử thay thế ('j'), tùy chọn ('?') hoặc được sửa lại ('\*'). Các quy tắc ngữ pháp đã xác định trước đó có thể được tham chiếu thông qua toán tử @seek.

Dưới đây, chúng tôi trình bày một ví dụ về quy tắc ngữ pháp *nr*\_*ksiegi* nhận dạng số nhận dạng của bệnh nhân trong bệnh viện được đưa ra trong (4). Quy tắc này nắm bắt, trong số những cụm từ khác, các cụm từ sau:

* *Sổ cái tổng hợp số 11125* 'Số tài liệu chính 11125'
* *Sổ cái tổng hợp số 12354*/*2006* 'Số của tài liệu chính

12354/2006'

* *Số tài liệu chính 13578 'Số tài liệu chính 13578'*

(3) Thần kinh | LOẠI: complication\_t | KHÁI TƯỞNG: neuropathy\_t bệnh thần kinh | LOẠI: complication\_t | KHÁI NIỆM : neuropathy\_t obwodowa | KIỂU: neuropathy\_t | KHÁI NIỆM : peripheral\_polyneuro autonomiczna | KIỂU: neuropathy\_t | KHÁI NIỆM : autonomic\_neuro Gluformin | LOẠI: oral\_t | KHÁI TƯỞNG: gluformin\_t

Glurenorm | LOẠI: oral\_t | KHÁI TƯỞNG: glurenorm\_t

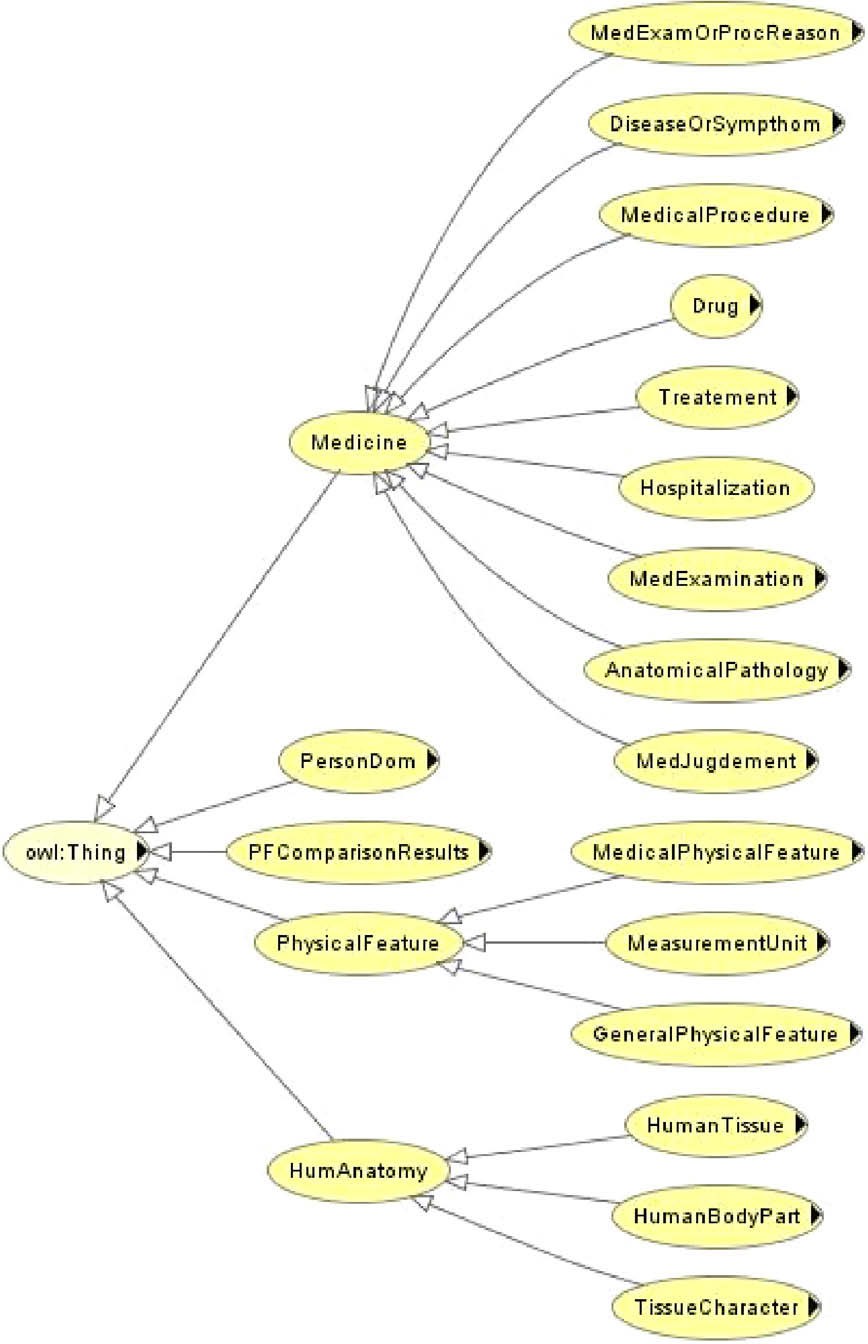
Vì từ vựng không được kết hợp với một máy phân tích hình thái, nên các hình thức được thay đổi và các cách viết khác nhau khi chúng xuất hiện trong văn bản, phải được liệt kê trực tiếp. Để tìm kiếm một loại thông tin chung (ví dụ: thuốc uống), có thể được khởi tạo bởi các khái niệm khác nhau trong bản thể, các quy tắc IE đề cập đến các giá trị TYPE (thay vì CONCEPT). Chi tiết về cách thức hoạt động của các quy tắc IE được trình bày trong Phần [4](#_bookmark5).

1. Ngữ pháp trích xuất

Trong các thí nghiệm của mình, chúng tôi đã trích xuất thông tin cụ thể bằng cách sử dụng nền tảng IE có mục đích chung, SProUT, đã được điều chỉnh để xử lý tiếng Ba Lan, [[35]](#_bookmark35). Trong SProUT, dữ liệu ở tất cả các cấp độ của quá trình ngôn ngữ được thể hiện dưới dạng cấu trúc đặc điểm được đánh máy. Ở mức thấp nhất, điều này liên quan đến ba thành phần chính:

2 Trong tất cả các số liệu tiếp theo có trong giấy, tên của các loại được bỏ qua.

Dòng 1–2 chỉ định ba phần tử thay thế: một từ, được xác định bởi trình phân tích hình thái (*morph*) thông qua dạng cơ sở (STEM) số 'số', hoặc hai chữ viết tắt của từ này (được xác định là *mã thông báo). Dòng 3 chỉ ra rằng mã thông báo sau đây là tùy chọn: điều này cho phép chúng ta bỏ qua dấu chấm (một mã thông báo riêng biệt) trong các chữ viết tắt mà* nó không có (xem các ví dụ ở trên). Dòng 4–5 nhận ra hai từ thông qua các dạng cơ sở của chúng: *ksie ̨ga* 'book', 'document' và *główny* 'main'. Lưu ý rằng ký hiệu thay thế ('j') không được chèn vào giữa hai phần tử và cả hai đều phải được xác định trong văn bản. Sau đó, (dòng 6) số nhận dạng của tài liệu được nhận dạng bằng cách gọi đến một quy tắc khác, liczba\_nat (một quy tắc nhận dạng số tự nhiên). Số được nhận dạng được coi là một biến ('#nr') và thống nhất với (gán cho) giá trị của thuộc tính ID trong cấu trúc đầu ra. Hai dòng tiếp theo (7–8) nhận ra một năm ('#nr1') sau dấu gạch chéo hoặc dấu gạch chéo ngược, nếu có thông tin này. Cuối cùng, cấu trúc đầu ra được cung cấp (dòng 9). Nó chứa ID và ngày (năm) của bệnh nhân đến bệnh viện.



Hình 2. Cấp cao nhất của hệ thống phân cấp lớp.

Đối với mỗi mẫu được chỉ định bởi mô hình miền của chúng tôi, chúng tôi xác định một quy tắc thích hợp (hoặc một tập hợp các quy tắc) cho phép chúng tôi nhận ra các giá trị của nó. Ngữ pháp cho các báo cáo chụp nhũ ảnh chứa 190 quy tắc, trong khi ngữ pháp cho lĩnh vực bệnh tiểu đường chứa 150 quy tắc. Đối với lĩnh vực chụp nhũ ảnh, gần như toàn bộ báo cáo được bao phủ bởi các quy tắc, trong khi trong lĩnh vực bệnh tiểu đường, chúng tôi chỉ nhắm mục tiêu các tập hợp con của thông tin. Việc lựa chọn của họ được thực hiện bởi bác sĩ tham gia vào việc phát triển ứng dụng.

Các quy tắc trích xuất hoạt động trực tiếp trên văn bản, nhưng trong cả hai kinh nghiệm, dữ liệu gốc của chúng tôi yêu cầu xử lý trước. Đối với các báo cáo của mammogra-phy, hóa ra các văn bản được viết bởi các bác sĩ chứa rất nhiều lỗi chính tả làm giảm chất lượng của dữ liệu được trích xuất: một từ sai chính tả có thể hiểu được đối với con người nhưng không phải đối với bộ xử lý tự động. Để giải quyết vấn đề này, chúng tôi đã triển khai một chương trình sửa lỗi chính tả tự động sử dụng từ vựng dành riêng cho tên miền [[31]](#_bookmark26). Hồ sơ bệnh nhân tiểu đường nhiều hơn

được gõ cẩn thận, vì vậy trong trường hợp này, trình kiểm tra chính tả là không cần thiết và tiền xử lý chủ yếu liên quan đến chuyển đổi định dạng (từ MS Word sang tệp văn bản). Trước khi sử dụng chúng trong các thí nghiệm, tất cả các hồ sơ của người được ẩn danh: thông tin về tên và địa chỉ được loại bỏ và một mã nhận dạng biểu tượng được gán cho mọi bệnh nhân. Quá trình này dựa trên cấu trúc tương đối nghiêm ngặt của các tệp (thông tin cá nhân được cung cấp trong các dòng đầu tiên của tài liệu), vì vậy không có sự nhận dạng thực thể được đặt tên phức tạp nào được áp dụng (trái ngược với [[45]](#_bookmark50)).

Dưới đây, chúng tôi mô tả các vấn đề ngôn ngữ điển hình phải được giải quyết trong khi xác định các quy tắc trích xuất và chúng tôi trình bày các giải pháp mà chúng tôi đề xuất cho các ứng dụng của mình. Vì trích xuất dựa trên quy tắc phụ thuộc mạnh vào miền, các giải pháp của chúng tôi không chung chung. Tuy nhiên, nhiều vấn đề phổ biến đối với loại tiến trình văn bản này và các cách tiếp cận tương tự có thể được áp dụng trong nhiều bối cảnh khác.

* 1. *Từ khóa mơ hồ*

Một trong những vấn đề chính khi giải thích các văn bản ngôn ngữ tự nhiên là đa nghĩa: các từ thường mơ hồ liên quan đến cách giải thích ngữ nghĩa (và hình thái cú pháp). Trong ngôn ngữ học tính toán, quá trình xác định cảm giác được gọi là Word Sense Disambiguation (WSD). Một cái nhìn tổng quan tuyệt vời về các thuật toán và ứng dụng WSD được đưa ra trong [[1]](#_bookmark27). Một nghiên cứu về các vấn đề WSD trong tài liệu y sinh và ghi chú lâm sàng, sử dụng phương pháp học máy có giám sát, được trình bày trong [[39].](#_bookmark36)

Trong các ứng dụng của chúng tôi, những khó khăn đáng kể duy nhất là với việc giải thích các từ khóa, tức là các từ quan trọng đối với việc làm chính được xác định bởi các quy tắc ngữ pháp. Một số trong số chúng, ví dụ, thuốc insulin, được chỉ định trong từ vựng miền, trong khi những người khác, như *cukrzyca* 'tiểu đường', được đề cập trong các quy tắc thông qua dạng cơ bản của chúng. Trong khi xây dựng mô hình miền và viết các quy tắc ngữ pháp IE, chúng tôi chỉ xác định được một vài sự mơ hồ về ý nghĩa như vậy. Vì không có công cụ thích hợp để định hướng ý nghĩa cho tiếng Ba Lan, nó phải được xử lý bởi các ứng dụng của chúng tôi. Trên cơ sở phân tích dữ liệu, chúng tôi đã áp dụng hai cách để giải quyết sự mơ hồ: trực tiếp trong các quy tắc ngữ pháp hoặc trong giai đoạn xử lý hậu kỳ.

Một ví dụ về sự mơ hồ có thể được giải quyết bằng quy tắc ngữ pháp là việc sử dụng từ khóa *nieregularny* 'không đều': trong lĩnh vực chụp nhũ ảnh, nó có thể mô tả hình dạng của một phát hiện, hoặc được hiểu là một thuộc tính của mô. Trong cụm từ *nieregularne zage ̨szczenie* 'mật độ không đều', từ mơ hồ được tìm thấy bên cạnh một từ khóa khác *zage ̨szczenie* 'mật độ' biểu thị một phát hiện, vì vậy quy tắc ngữ pháp tương ứng sẽ giải thích *nieregularne* là đề cập đến hình dạng. Một ví dụ về cùng một loại mơ hồ trong lĩnh vực bệnh tiểu đường là *mikroalbuminuria*, đề cập đến một biến chứng trong cụm từ *wystpiła mikroalbuminuria* 'microal- buminuria đã xảy ra', trong khi nó biểu thị một xét nghiệm trong cụm từ *Mikroalbuminuria: 25 mg / dobce ̨* 'Microalbuminuria: 25 mg / ngày'. Do đó, khi từ *mikroalbuminuria* được theo sau bởi một đại lượng cho biết kết quả xét nghiệm, từ khóa được coi là một xét nghiệm; nếu không, nó là một phức tạp.

Các trường hợp không thể giải quyết bằng các quy tắc ngữ pháp IE được giải quyết trong giai đoạn xử lý hậu kỳ. Trong (5), các cụm từ mô tả sự định vị và mật độ mô được chèn vào giữa từ khóa 'không đều' và ngữ cảnh định hướng của nó - 'mô tuyến'. Trong trường hợp này, ngữ cảnh không được tính đến trực tiếp, và từ khóa 'không quy định' được coi là một siêu kiểu của hai cách giải thích có thể xảy ra. Trong giai đoạn xử lý hậu kỳ, nếu từ khóa thuộc về một khối tis-sue, nó được hiểu là một thuộc tính của phân loại mô, trong khi nếu nó thuộc về một khối tìm kiếm, nó có nghĩa là một hình dạng, xem Phần [5](#_bookmark9) để biết chi tiết.

|  |  |  |  |
| --- | --- | --- | --- |
| (5) [Mô tuyến] | [nhú] | [ge ̨ sta], | [bất thường] |
| [Mô tuyến] | [dưới areolar] | [dày đặc], | [bất thường] |

* 1. *Phủ định*

Nhận biết chính xác sự phủ định trong các báo cáo lâm sàng là rất quan trọng đối với sự hiểu biết của họ. Các cách tiếp cận khác nhau cho vấn đề đã được đề xuất. Ví dụ, [[9]](#_bookmark37)  mô tả thuật toán NegEx xác định phủ định đối với các phát hiện hoặc bệnh. NegEx sử dụng một tập hợp các biểu thức chính quy và một danh sách các thuật ngữ phủ định để phân biệt một phủ định thực sự (ví dụ: *vắng mặt; từ chối*) từ các biểu thức liên quan đến một thuật ngữ phủ định không phủ nhận phát hiện (ví dụ: *không chỉ*). Xử lý phủ định trong tiếng Bungari, được thảo luận trong [[3],](#_bookmark22) dựa trên sự kết hợp của các phương pháp nông (để xác định phủ định cục bộ) với phân tích ngữ nghĩa sâu (để nhận biết phủ định từ xa). [[16]](#_bookmark47)  mô tả việc xử lý phủ định trong một hệ thống sử dụng thuật ngữ snomed-CT để lập chỉ mục các tài liệu lâm sàng. Các quy tắc được xác định cho hệ thống dựa trên việc nhận dạng các từ ngụ ý phủ nhận (ví dụ: *không, từ chối, loại trừ*) và các thuật ngữ ngăn chặn

lan truyền gán phủ định (ví dụ: *khác với*). Trong hệ thống của chúng ta, việc nhận ra cả phủ định cục bộ và xa xôi đã được đưa vào các quy tắc ngữ pháp, đôi khi làm tăng độ phức tạp của chúng.

Cả hai bộ tài liệu của chúng tôi đều thể hiện sự phủ định. Nếu phủ định xuất hiện ngay trước từ khóa, nó có thể dễ dàng được nắm bắt bởi một quy tắc nhận ra một biểu thức phủ định và một từ khóa thích hợp. Thật không may, phủ định không phải lúc nào cũng trực tiếp đứng trước từ khóa.

Hãy xem xét câu sau: *Nie stwierdzono póz'nych powikłan' cukrzycy o typie mikroangiopatii.* 'Không có biến chứng bệnh tiểu đường kéo dài của loại microagiopathy'. Biểu thức phủ định  *nie stwierdzono* 'there were no' nằm ở đầu câu, trong khi từ khóa *mikroangiopatii* 'microangiopa- thy' là từ cuối cùng trong câu. Cụm từ này có thể được nhận ra bằng quy tắc được đưa ra trong (6). Quy tắc đề cập đến các dạng cơ sở của các từ gợi ý và sự phức tạp (được biểu thị bằng biến #t) được tra cứu trong từ vựng miền. Kết quả là, biến chứng được xác định chính xác là không có mặt.

(6)

no\_complication :> morph & [STEM ''nie"]

(morph & [STEM ''state ́"] j

morph & [STEM ''xảy ra ́"] j

morph & [STEM ''detect ́"]) (morph & [STEM ''presence ́c ́"])? (morph & [STEM ''muộn'])? (morph & [STEM ''phức tạp''] j

[STEM ''thay đổi'])

(morph & [STEM ''bệnh tiểu đường'] j

morph & [STEM ''cukrzyca"]) (morph & [STEM ''w"] j

morph & [STEM ''pod"] j

morph & [STEM ''o"]) (morph & [STEM ''postac'"] j

morph & [STEM ''typ"] j

morph & [STEM ''charakter"]) gazetteer & [TYPE complication\_t,

KHÁI NIỆM #t]

-> no\_comp\_str & [N\_COMP #t].

;; 'không'

;; 'công nhận'

;; 'hiện tại'

;; 'lâu dài'

;; 'biến chứng'

;; 'bệnh tiểu đường'

;; giới từ

;; 'loại'

;; loại

;; biến chứng

Quy tắc trên công nhận, trong số những cụm từ khác, các cụm từ sau (có cùng ý nghĩa như trước đây):

* + - *không có biến chứng tiểu đường muộn*  *của dạng vi mạch*,
    - *sự hiện diện của các biến chứng tiểu đường muộn dưới*

*vi mạch,*

* + - *Không có thay đổi do bệnh tiểu đường muộn ở dạng vi mạch được tìm thấy*.

Trong ví dụ rất giống với ví dụ trước: *Nie stwierdzono póz'nych powikłañ cukrzycy z wyja ̨ tkiem mikroangiopatii* 'không có biến chứng bệnh tiểu đường kéo dài, ngoại trừ bệnh vi mạch', ý nghĩa khác với các ví dụ trước: 'vi mạch máu' là một biến chứng thực sự và cần được công nhận. Cụm từ này bị bỏ qua chính xác bởi quy tắc (6) vì *z wyja ̨ tkiem* 'except for' không phải là một biểu thức phủ định. Tuy nhiên, trong trường hợp này và các trường hợp tương tự, chúng ta phải phân tích toàn bộ câu để xác định chính xác liệu có hay không có biến chứng, hoặc nếu một thuộc tính nhất định xuất hiện.

* 1. *Phối hợp*

Trong cả hai bộ tài liệu, có các ví dụ về thông tin được chỉ định bởi các cụm từ phối hợp. Chúng tôi xác định ba loại phối hợp:

* + - phối hợp hình elip,
    - phối hợp các công cụ sửa đổi liên quan đến một từ khóa,
    - phối hợp các từ khóa.

Một ví dụ về sự phối hợp hình elip là cụm từ z *neuropatia ̨ autonomiczna ̨ i obwodow*a ̨ 'with autonomic and peripheral neurop- athy', mô tả hai biến chứng. Cụm từ này bao gồm từ khóa *neuropatia ̨ 'bệnh thần kinh' và hai tính từ phối hợp chỉ ra* hai loại bệnh thần kinh. Phối hợp hình elip đã được giải quyết ví dụ trong [[7]](#_bookmark38)  bằng các kỹ thuật học máy. Vì chúng tôi không có đủ dữ liệu để đào tạo các thuật toán như vậy, chúng tôi nhận ra các cấu trúc như vậy trong các quy tắc ngữ pháp IE của chúng tôi. Trong (7) chúng tôi chỉ ra quy tắc để nhận biết sự phối hợp của các bệnh thần kinh. Quy tắc này gần như chỉ đề cập đến các mục từ từ vựng miền.

(7) neuropat\_coord:>

gazetteer & [TYPE complication\_t, CONCEPT neuropathy\_t] (công báo & [TYPE neuropathy\_t, CONCEPT #r1])

(token & [SURFACE ''i"]| token & [SURFACE ''oraz"]) (công báo & [LOẠI neuropathy\_t, CONCEPT #r2])

->feature\_l\_str & [TÍNH NĂNG feature\_list &

[FIRST comp\_str & [COMP #r1],

REST feature\_list & [ FIRST comp\_str & [COMP#r2], REST \*null\* ]]].

Loại phối hợp tiếp theo liên quan đến các trường hợp khi một số thuộc tính được liên kết với một khái niệm, ví dụ: nhiều thuộc tính của một trường hợp bệnh tiểu đường cụ thể hoặc của một phát hiện. Trong (8) chúng tôi chỉ ra một ví dụ từ lĩnh vực bệnh tiểu đường.

(8) *Biến chứng lâu dài* do bệnh võng mạc, bệnh tiểu đường loại 2 không kiểm soát được

Mỗi từ trong (8) mang thông tin quan trọng: *wieloletnia* 'lâu dài', *powikłana retinopatia ̨ 'phức tạp* bởi bệnh võng mạc', *niekontrolowany* 'không kiểm soát', loại *2* 'loại 2'. Hầu hết chúng chỉ liên quan đến ứng dụng của chúng tôi trong ngữ cảnh của từ khóa *cukrzyca* 'tiểu đường' và nên bỏ qua nếu không. Ví dụ, *wieloletni* trong *wieloletni pacjent szpitala* 'một bệnh nhân lâu dài của bệnh viện' không chỉ ra bệnh tiểu đường kéo dài. Mặt khác, thông tin về các biến chứng bệnh võng mạc (cũng như các biến chứng khác hoặc thông tin về cân nặng của bệnh nhân) không yêu cầu bất kỳ từ khóa ngữ cảnh nào và có thể được giải thích chính xác. Để truyền đạt sự khác biệt này, chúng tôi phân biệt hai nhóm thuộc tính tùy thuộc vào sự hiện diện của từ 'tiểu đường' trong ngữ cảnh. Các thuộc tính không yêu cầu từ ngữ cảnh này có thể được nhận dạng riêng biệt. Trong ví dụ này, chúng ta nhận ra toàn bộ cụm từ bằng một quy tắc duy nhất, cho phép chúng ta điền vào mẫu cục bộ với tối đa năm thuộc tính, sử dụng một từ khóa 'tiểu đường'.

Loại cụm từ phối hợp thứ ba liên quan đến sự phối hợp của các cụm từ chính mà cùng một thông tin phải được đính kèm. Ví dụ, trong một số báo cáo nhất định, một phần thông tin (ví dụ: một phát hiện hoặc chẩn đoán) có thể được liên kết với một số khu vực hóa, xem (9) và (10). Kết quả phân tích của (9) được thể hiện trong [Hình 3](#_bookmark7).[3](#_bookmark8) Mỗi cấu trúc là kết quả của một quy tắc khai thác.

(9) *Ở góc phần tư trên-ngoài của cả hai núm vú và trong vùng nhú của núm vú phải, một đường viền dày lên 5 mm. Tổn thương lành tính. (chúng tôi ̨ bạch huyết xấu? nang?)* Ở góc phần tư trên-ngoài và ở dưới quầng vú bên phải, có mật độ giới hạn duy nhất có đường kính 5 mm. Những thay đổi lành tính (hạch bạch huyết? u nang?)

3 Trong (9), [Hình 3](#_bookmark7), và một số hình sau đây, chúng tôi sử dụng màu sắc để chỉ ra đoạn văn bản mà mỗi cấu trúc được gán cho phần văn bản.

(10) *Rỗ nách, da và núm vú không có.*

Nách và da và núm vú không ác tính.

Để đối phó với nhiều biểu thức bản địa hóa như trong (9), chúng tôi giới thiệu một quy tắc ngữ pháp, có thể xử lý một số bản địa hóa. Quy tắc gán các giá trị cho thuộc tính LOC cũng như cho thuộc tính LOC2 (và nếu cần LOC3) được thêm vào.

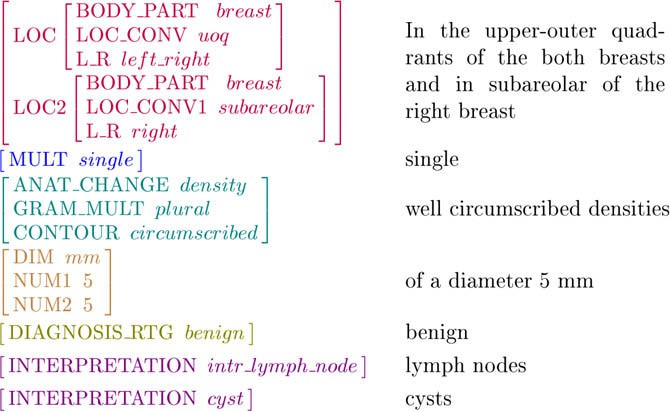
* 1. *Biểu thức Anaphoric*

Các cách diễn đạt ẩn dụ, đề cập đến một đối tượng được đề cập trước đó trong văn bản, có tầm quan trọng khác nhau trong các văn bản khác nhau. Ví dụ, [[38]](#_bookmark39)  gợi ý rằng trong các bài báo sinh học khoảng 5% mô tả về tương tác protein-protein chứa các biểu hiện tương tác. Có rất nhiều tài liệu về anaphora và giải pháp đồng tham chiếu trong các văn bản và diễn ngôn, Hội nghị Discourse Anaphora và Anaphora Resolution (DAARC) dành cho những vấn đề này. Trong lĩnh vực y sinh, một thuật toán chú thích các mối quan hệ phân loại trong các bản tóm tắt bài báo được trình bày trong [[34].](#_bookmark40) Phương pháp này dựa trên các đặc điểm cú pháp và thông tin ngữ nghĩa từ UMLS. Một cách tiếp cận khác, dựa trên học có giám sát, được trình bày trong [[46],](#_bookmark52) trong khi một phương pháp xác suất để giải quyết anaphora trong các bài báo toàn văn được mô tả trong [[20].](#_bookmark53) Rất ít đã được công bố về độ phân giải anaphora trong các báo cáo lâm sàng. Một ví dụ về hệ thống thực hiện phân giải đồng tham chiếu là MedSyndikate [[22].](#_bookmark21)

Trong các tài liệu của chúng tôi, các biểu hiện tương thích thú vị chỉ xuất hiện trong các báo cáo chụp nhũ ảnh và chủ yếu liên quan đến việc xác định các bản địa hóa và mô tả tương đối về những thay đổi giải phẫu. Mặc dù chúng không thường xuyên lắm, nhưng sự công nhận của chúng rất quan trọng đối với kết quả cuối cùng. Cách tiếp cận của chúng tôi tương tự như [[2],](#_bookmark28) vì chúng tôi kết hợp nhận dạng anaphora vào các quy tắc ngữ pháp. Chúng tôi giới thiệu các thuộc tính đặc biệt đại diện cho các cụm từ chỉ ra các mối quan hệ đồng cảm. Giải quyết Anaphora được thực hiện ở giai đoạn sau quá trình. Trong dữ liệu của chúng tôi, các biểu thức anaphoric liên quan đến việc bản địa hóa các thay đổi hoặc chính những thay đổi, và có thể được chia thành các lớp sau:

* + - Anaphora của bản địa hóa
      * Anaphora của một bản địa hóa chung,
      * Anaphora của một bản địa hóa chi tiết,
    - Anaphora của một phát hiện và mô tả của nó
      * Anaphora của một phát hiện,
      * Anaphora về kích thước của phát hiện,
      * Anaphora về các tính năng của finding,
      * Tham chiếu anaphoric đến một thành viên của một tập hợp.

Anaphora của bản địa hóa là một cụm từ trong đó bản địa hóa của, ví dụ, một phát hiện được chỉ định bằng cách tham chiếu đến một đề cập trước đó



Hình 3. Giải thích báo cáo được trích dẫn trong (9).

Localization. Anaphora bên chung chỉ xảy ra khi thông tin bên hóa bị bỏ qua. Điều này xảy ra nếu có hai phát hiện được xác định trong vú, ví dụ: w *kwadrancie górno-zewne ̨trznym tego sutka* 'ở góc phần tư trên-ngoài (uoq) của vú này', *w tym sutku w KGZ* 'trong vú này trong uoq'. Trong những

cụm từ, đại từ biểu thị *tym, tego* 'thisinst/*gen*' nên

được hiểu là 'trong cùng một bộ ngực đã được mô tả trước đó'. Biểu thức bản địa hóa anaphoric cũng có thể đại diện cho thông tin chi tiết hơn. Các cụm từ *sa ̨ siaduja ̨cy* 'láng giềng', obok 'bên cạnh' đề cập đến một bản địa hóa trước đó. Để thể hiện anaphora bản địa hóa, trong trường hợp bỏ qua sự phân bổ bên (trường hợp đầu tiên được đề cập ở trên), đường dẫn LOCjL\_R được gán kiểu *loc*\_*l\_r\_last*. Để tham khảo bản địa hóa đã đề cập trước đó (trường hợp thứ hai), thuộc tính LOC\_REF đặc biệt được giới thiệu. Trong cả hai trường hợp, chúng tôi tính toán bản địa hóa thích hợp trong giai đoạn xử lý hậu kỳ.

Loại anaphora thứ hai liên quan đến các phát hiện và các đặc điểm của chúng. Trong các cụm từ sau: *podobna zmiana* 'một phát hiện tương tự', druga 'thứ hai (một)', hoặc *zmiana o tej samej wielko'sci i charakterze* 'một phát hiện có cùng kích thước và loại (như phát hiện trước)'. Trong những trường hợp như vậy, hệ thống nên tham khảo phát hiện trước đó để có được thông tin còn thiếu.

Trong một số trường hợp, chỉ một phần của mô tả phát hiện được giải quyết bằng một cách diễn đạt tương phận. Trong (11) phát hiện thứ hai được đề cập bằng cụm từ elip drugie 'thứ hai (một)'. Bản địa hóa và kích thước của phát hiện được đưa ra một cách rõ ràng, nhưng các thông tin khác phải được suy ra từ mô tả của phát hiện trước đó.

(11) *W sutku prawym w KGW 5 mm dobrze ograniczone zage ̨ szczenie (łagodne), drugie w KGZ o 'sr. 10 mm równiez*\_ *o podobnym charakterze.* Ở ngực phải, ở góc phần tư bên trong trên 5 mm có mật độ giới hạn tốt (lành tính), mật độ thứ hai ở góc phần tư trên-ngoài, có kích thước đường kính 10 mm và một loại tương tự.

Các mẩu thông tin cô lập thu được cần được tiến hành thêm để điền vào các mẫu phức tạp hơn được xác định trên cơ sở mô hình miền (bản thể học). Đối với các báo cáo chụp nhũ ảnh, nhiệm vụ khó khăn nhất là tách biệt đúng các mô tả về các loại mô khác nhau và những thay đổi giải phẫu khác nhau, nếu một số phát hiện được phát hiện. Ví dụ, trong (12), cấu trúc kích thước, phân tích và diễn giải có thể được gắn vào cả hai mô tả tìm kiếm. Vì vậy, nhóm dữ liệu thích hợp là điều cần thiết để hiểu văn bản chính xác.

Thông tin liên quan đến một phát hiện luôn liền kề, do đó nhóm có thể đạt được bằng cách giới thiệu các điểm đánh dấu bắt đầu và kết thúc. Chúng tôi đã sử dụng các điểm đánh dấu để chỉ ra một khuôn mẫu phức tạp (một khối) tương ứng với một phát hiện duy nhất (fb / fe) và cấu trúc mô (tbb / tbe).

Các khối cụ thể hơn trong mô tả mô được gắn thẻ TB / te. Các điểm đánh dấu này được chèn vào trong giai đoạn xử lý hậu kỳ.

Giai đoạn xử lý hậu kỳ bao gồm các bước sau:

* xác định các thuộc tính chính được sử dụng để nhóm thông tin thêm và các điểm đánh dấu chỉ định đường viền của ba khối thông tin chính,
* phân đoạn khối tìm kiếm,
* phân chia mô tả về thành phần vú,
* thêm thông tin tóm tắt.

Các thuật toán mà chúng tôi đề xuất phụ thuộc rất mạnh vào dữ liệu và được phát triển sau khi phân tích nhiều báo cáo. Giả định chính là mô tả về một phát hiện hoặc mô vú có thể được phân biệt, nếu chúng ta nhận ra các từ khóa chỉ ra một le-

Để giải thích các biểu thức anaphoric đề cập đến những thay đổi, chúng tôi đã giới thiệu ba thuộc tính, cho biết loại thông tin được tham chiếu. Việc nhận dạng các cụm từ sau dẫn đến việc cung cấp các cấu trúc tương ứng:

* *druga zmiana* 'phát hiện thứ hai' - [CHANGE\_REF *ref*\_*cmp*],
* *zmiana o podobnej wielko'sci* 'tìm thấy kích thước tương tự' – [SIZE\_REF *có*]
* *zmiana o podobnym charakterze* 'tìm kiếm một loại tương tự' - [TYPE\_REF *có*]
* *najwie ̨ksza z nich* 'người lớn nhất' – [CHANGE\_REF *ref*\_*max*].

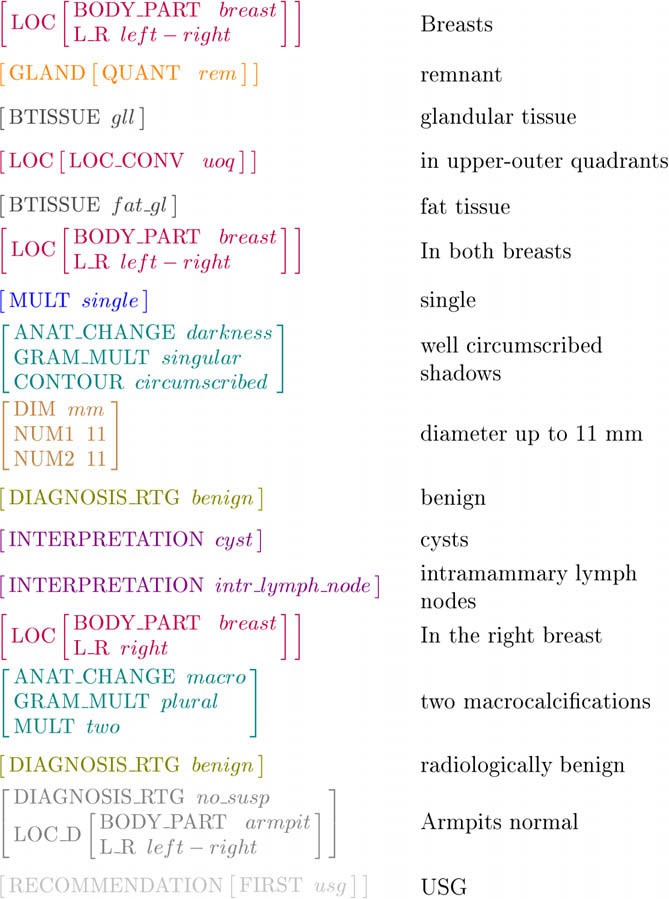
Vì vậy, phân tích anaphora từ Ví dụ (11) cho kết quả là cấu trúc sau: [CHANGE\_REF *ref*\_*cmp*, TYPE\_REF *yes*].

1. Xử lý hậu kỳ – nhóm và lựa chọn thông tin

Kết quả của IE là một tệp XML chứa các giá trị của tất cả các thuộc tính được nhận dạng trong toàn bộ tài liệu. Ví dụ (12) chứa một báo cáo chụp nhũ ảnh. Kết quả của quá trình xử lý của nó bởi gram-mars IE được đưa ra trong [Hình 4](#_bookmark10).

(12) *Núm vú có mô tuyến còn sót lại ở góc phần tư trên-ngoài. Ưu thế của mỡ trong cơ thể. Ở cả hai núm vú đều có bóng râm duy nhất có đường viền tốt với độ dài lên đến 11 mm. tổn thương lành tính về mặt X quang (u nang vú trong, u nang bạch huyết?). Ở núm vú bên phải, hai vôi hóa lớn có tính chất lành tính. Nách bình thường. Kiểm soát siêu âm được khuyến khích.*

Vú có mô tuyến còn sót lại ở góc phần tư trên-ngoài. Mô mỡ trội. Ở cả hai ngực, trải ra các bóng được giới hạn tốt với đường kính lên đến 11 mm. Thay đổi lành tính X quang (u nang? hạch bạch huyết trong vú?). Ở vú phải có hai đại vôi hóa lành tính. Nách bình thường. Khuyến nghị kiểm tra USG.



Hình 4. Giải thích báo cáo chụp nhũ ảnh được trích dẫn trong (12).

sion hoặc một loại mô, tương ứng. Quan sát quan trọng tiếp theo là một số đặc điểm chỉ có thể xuất hiện trong một khối một lần, trong khi các đặc điểm khác có thể được lặp lại. Chúng tôi cũng nhận thấy rằng chúng ta nên bắt đầu phân biệt việc tìm các khối từ đầu báo cáo, trong khi đối với các khối mô, một chiến lược tốt hơn là xử lý dữ liệu từ cuối. Thông tin đầu tiên về tổn thương là loại (hoặc khu trú) của nó và sau đó một số chi tiết được đưa ra. Trong trường hợp mô tả, mô trội thường được mô tả là mô cuối cùng và mô tả chi tiết về các loại mô khác được đưa ra ở đầu báo cáo.

Trong trường hợp nhóm thông tin, các thuộc tính chính để tìm các khối là ANAT\_CHANGE và INTERPRETATION, trong khi BTISSUE là thuộc tính chính cho thành phần ngực. Trong giai đoạn sơ bộ, các cấu trúc chứa các thuộc tính này được gắn thẻ bổ sung bằng nhãn a\_ch, i\_ch và tis tương ứng. Các cấu trúc có thuộc tính không thể thuộc bất kỳ khối tìm kiếm nào được đánh dấu là dloc. Các cấu trúc với thông tin được trích xuất từ phần cuối cùng của báo cáo, chứa các khuyến nghị chung, được đánh dấu bằng thẻ rb. Quá trình xác định các khối tìm kiếm bắt đầu từ cấu trúc đầu tiên được đánh dấu bằng thẻ a\_ch, i\_ch hoặc tis. Từ cấu trúc đó, chúng ta quay trở lại ranh giới khối gần nhất hoặc đến đầu báo cáo, nếu có thể thêm cấu trúc vào khối mới. Một số thuộc tính, ví dụ: bản địa hóa hoặc kích thước, chỉ có thể xuất hiện một lần trong khối. Do đó, nếu một khối đã chứa một cấu trúc với một thuộc tính như vậy, nó không thể được kết hợp vào khối. Khi một cấu trúc-

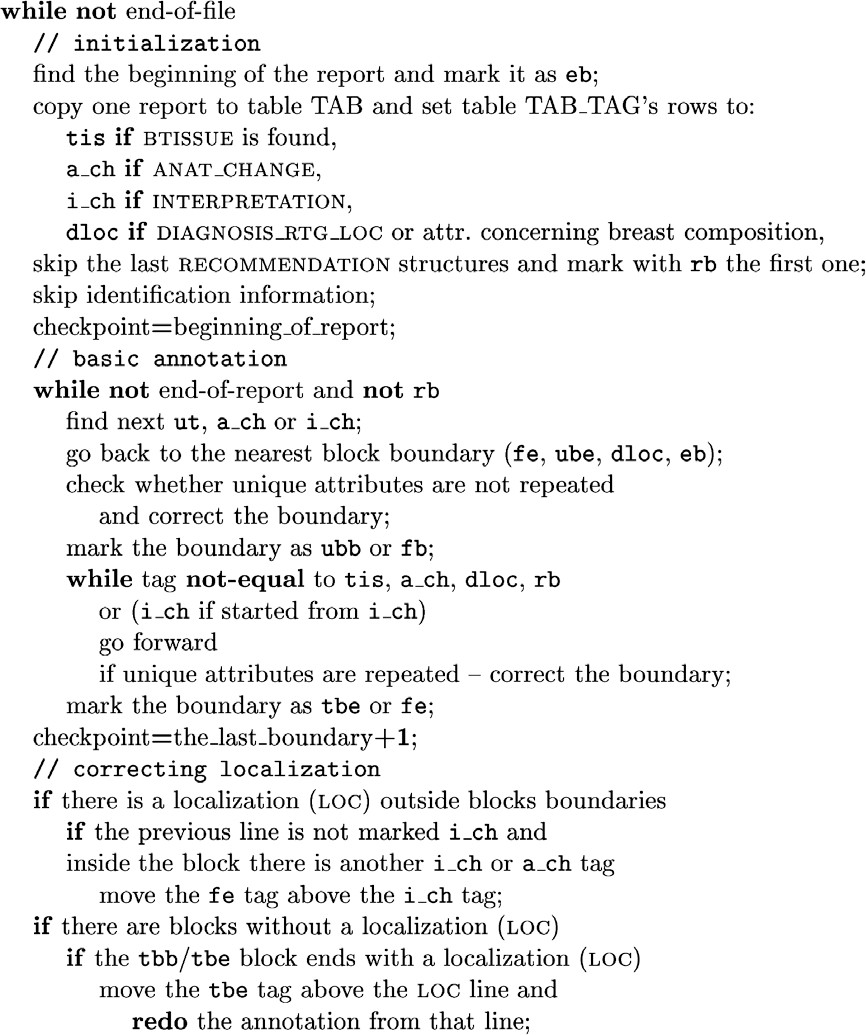
ture được đánh dấu bằng thẻ DLOC hoặc đã chứa một thuộc tính không thể lặp lại, thẻ bắt đầu khối fb được chèn. Quy trình tương tự được lặp lại từ cấu trúc chứa thuộc tính tìm kiếm chính ở cuối báo cáo và thẻ kết thúc fe được chèn vào. Cuối cùng, chương trình xác minh kết quả và sửa phân đoạn nếu cần. Một mã giả của thuật toán (được trình bày trong [[29]](#_bookmark21)) được đưa ra trong [Hình 5](#_bookmark11).

Kết quả chèn ranh giới tìm kiếm cho dữ liệu trong [Hình 4](#_bookmark10) được trình bày trong [Hình 6](#_bookmark12).

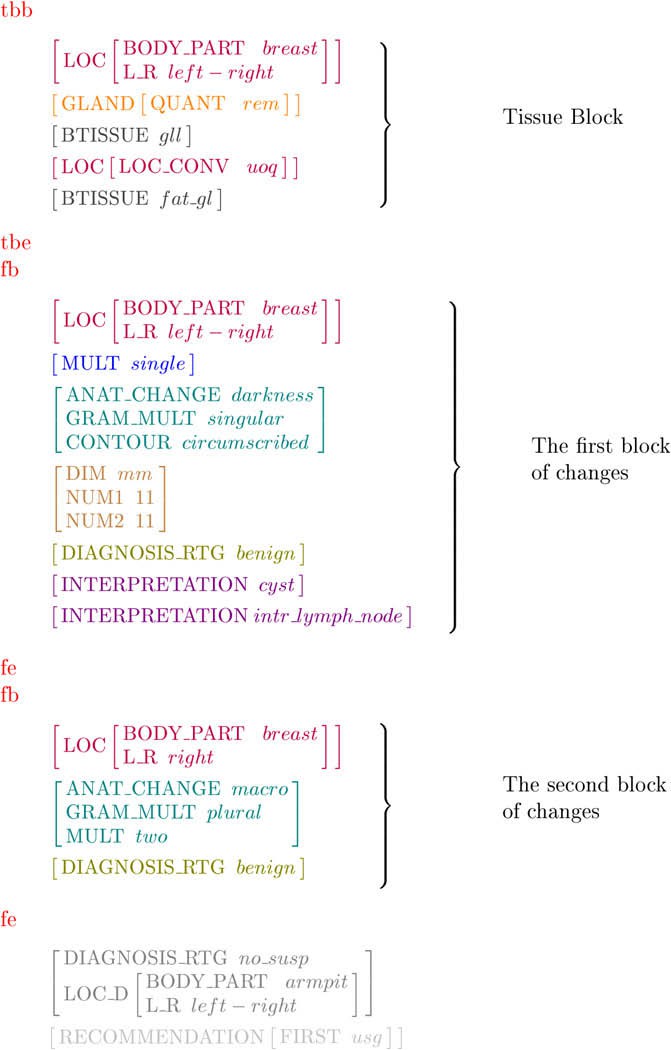
Bước xử lý hậu kỳ tiếp theo là chia khối cấu trúc của vú, được phân tách bằng thẻ tbb/tbe như trong [Hình 6](#_bookmark12), thành các khối phụ logic. Thuật toán phức tạp hơn so với phân đoạn các phát hiện vì thông tin không liên tục.

Thuật toán xác định các khối phụ mô chèn các thẻ tb / te phân định thông tin mô cụ thể hơn. Nó cũng đã được chia thành hai bước. Đầu tiên, các cấu trúc của khối cấu trúc chung được chú thích bằng các thẻ cho biết loại mô, tính chất mô hoặc loại khu trú. Chúng tôi phân biệt hai loại tiễn: đặc hiệu (xơ hóa tuyến và tuyến) và gen (các loại mô khác). Ngoài ra, chúng tôi sử dụng ba loại bản địa hóa:

* bản địa hóa chung: nếu cấu trúc LOC chỉ chứa thuộc tính BODY\_PART có giá trị *vú* và nếu thông tin bên (L\_R) được xác định,



Hình 5. Thuật toán xác định ranh giới phát hiện.



Hình 6. Giải thích báo cáo với các khối.

* bản địa hóa cụ thể: nếu chỉ có các thuộc tính chỉ ra một bộ phận giải phẫu hoặc bản địa hóa thông thường trong cấu trúc, tức là LOC\_CONV, LOC\_CONV1 hoặc LOC\_A,
* Bản địa hóa hoàn toàn: nếu cả thuộc tính chung và cụ thể đều

có trong cấu trúc, tức là cấu trúc LOC đã xác định các thuộc tính BODY\_PART và L\_R cũng như một trong các thuộc tính LOC\_CONV, LOC\_ CONV1 hoặc LOC\_A.

Tiếp theo, khối thành phần mô chung được xử lý từ cuối. Các cấu trúc được nhóm thành các khối logic. Nếu cần thiết, một bản địa hóa thích hợp được sao chép cho khối mô tả tis- sue: nếu khối chỉ chứa bản địa hóa cụ thể, chúng tôi nhìn lên trên để tìm bản địa hóa chung gần nhất; Tương tự, nếu có một cấu trúc mô tổng quát mà không có bất kỳ sự khu trú nào, chúng ta nhìn lên trên để tìm sự khu trú chung gần nhất.

Thuật toán phân biệt các khối mô cụ thể, cùng với các trường hợp có vấn đề của nó, đã được mô tả trong [ [30].](#_bookmark29) Kết quả thu được cho dữ liệu trong [Hình 6](#_bookmark12) được đưa ra trong [Hình 7](#_bookmark13). Cấu trúc đầu tiên, mô tả bản địa hóa chung, nằm ngoài bất kỳ khối nào và không được tính đến trong quá trình xử lý dữ liệu tiếp theo.

Ngoài các thuộc tính được gán giá trị trực tiếp trong quá trình xử lý văn bản, ba thuộc tính đã được xác định để cung cấp một bản tóm tắt ngắn gọn của báo cáo: REPORT\_CLASS – để chẩn đoán tổng thể; MMG\_REL – cho độ tin cậy của hình ảnh; và REPORT\_WITH\_FIND(ings) – cho một thông số kỹ thuật nhị phân nếu có bất kỳ phát hiện nào đã được phát hiện. Giá trị

của REPORT\_CLASS được suy ra từ các xét nghiệm khuyến cáo và chẩn đoán từng phần (nếu có) như sau:

* *diag*\_*no*\_*susp*\_*changes* - không có thay đổi hoặc chẩn đoán đáng ngờ khác,
* *diag*\_*benign* – các phát hiện trong báo cáo được phân loại là lành tính,
* *diag*\_*susp* – sinh thiết được khuyến nghị hoặc các phát hiện được phân loại là đáng ngờ,
* *diag*\_*mal* – phát hiện có thể là ác tính được phát hiện hoặc

cần tư vấn ung thư.

Độ tin cậy của hình ảnh (MMG\_REL) được xác định theo loại thành phần vú. Nếu mô vú rất dày đặc hoặc loạn sản, hoặc nếu được nêu rõ rằng hình ảnh khó đọc, MMG\_REL không *đáng tin cậy*. Nếu mô mỡ chiếm ưu thế, báo cáo là *đáng tin cậy*; trong tất cả các trường hợp khác, MMG\_REL lấy *giá trị* avg*\_*reliable.

Giai đoạn cuối cùng của quá trình xử lý hậu kỳ cho các báo cáo chụp nhũ ảnh liên quan đến việc chuyển đổi (sao chép) dữ liệu cho các biểu hiện phối hợp và phân tích. Ví dụ, trong các cụm từ sau: *podobna zmiana* 'một phát hiện tương tự', druga 'thứ hai (một)', hoặc *zmiana o tej samej wielko'sci i charakterze* 'một phát hiện có cùng kích thước và loại (như phát hiện trước)', các phát hiện và tính chất của chúng được mô tả bằng các biểu thức tham chiếu. Trong những trường hợp như vậy, hệ thống IE nhận ra rằng phát hiện xảy ra, nhưng không thể đưa ra các đặc điểm của nó. Để tính đến các biểu thức tham chiếu, chúng tôi đã giới thiệu ba thuộc tính: (CHANGE\_REF, TYPE\_REF, SIZE\_REF) cho biết các giá trị tham chiếu của các thuộc tính đã chọn. Ví dụ: TYPE\_REF *yes* chỉ ra rằng mô tả của phát hiện mới giống với mô tả của phát hiện trước đó. Tương tự, SIZE\_REF chỉ định rằng kích thước của phát hiện đồng nhất với giá trị của SIZE hoặc SIZE\_TEXT (tùy theo giá trị nào có) trong phát hiện trước đó. Trong giai đoạn xử lý hậu kỳ, các khe mẫu thích hợp của phát hiện anaphoric được lấp đầy với các giá trị được chỉ định cho phát hiện trước đó.

Các tài liệu xuất viện của bệnh nhân tiểu đường cũng trải qua quá trình sau khi xét xử, nhưng nó đơn giản hơn đáng kể so với các báo cáo chụp nhũ ảnh và chủ yếu là thông tin dư thừa. Dưới đây chúng tôi mô tả chiến lược để chọn các phần thông tin quan trọng về cùng một chủ đề có trong kết quả:

* chỉ có một bản sao của thông tin được lưu giữ (ví dụ: thông tin về bệnh tiểu đường không kiểm soát được hoặc một biến chứng được lặp lại nhiều lần trong tài liệu xuất viện),
* nếu thông tin không giống nhau về cùng một chủ đề được công nhận,

chúng tôi:













Hình 7. Xác định các khối thành phần của vú.

* + Lưu giữ thông tin cụ thể hơn nếu mối quan hệ này được giữ giữa các kết quả, ví dụ: từ hai biến chứng được phát hiện: *retinopathy*\_*t* và *proliferative*\_retinochúng tôi chọn biến chứng thứ hai vì loại sau là một loại phụ của (cụ thể hơn) biến chứng trước,
  + Chọn kết quả cao nhất, như đối với một số bài kiểm tra có số

giá trị (ví dụ: đối với mức cholesterol LDL),

* Để biết thông tin về thuốc và chế độ ăn uống được khuyến nghị, chúng tôi chọn thông tin có trong phần tóm tắt.

Sau khi xử lý hậu kỳ, các mẫu đã điền được chèn vào cơ sở dữ liệu quan hệ tiêu chuẩn. Quá trình này khá đơn giản

Bảng 1

Kết quả đánh giá cho các thuộc tính được tìm thấy trong 705 báo cáo chụp nhũ ảnh thử nghiệm, phần 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Thuộc tính | | Trường hợp | Trước. | Nhớ | F |
| TẤT CẢ CÁC THUỘC TÍNH |  | 22742 | 99.64 | 99.53 | 99.58 |
| Loại tìm kiếm | ANAT\_THAY ĐỔI | 277 | 93.81 | 98.56 | 96.13 |
| Tìm kiếm các tính năng |  |  |  |  |  |
| Đường nét | ĐƯỜNG NÉT | 51 | 100 | 100 | 100 |
| Bão hòa | BÃO HÒA | 8 | 100 | 100 | 100 |
| Hình dạng | HÌNH DẠNG | 39 | 100 | 100 | 100 |
| Hình dạng phụ lục | SH\_APP | 4 | 100 | 100 | 100 |
| Sự gợi ý | SỰ GỢI Ý | 8 | 100 | 100 | 100 |
| Đồng hành. Vôi hóa | VỚI\_CALC | 27 | 100 | 100 | 100 |
| Khả năng sờ thấy | KHẢ NĂNG BẢO VỆ | 2 | 100 | 100 | 100 |
| Vô cùng |  |  |  |  |  |
| Cho bằng chữ số | MULT | 114 | 100 | 99.12 | 99.56 |
| Ngữ pháp | GRAM\_MULT | 287 | 94.10 | 100 | 96.96 |
| Kích thước |  |  |  |  |  |
| Đơn vị | MỜ | 103 | 98.10 | 100 | 99.04 |
| Kích thước trong 1 chiều | Số 1 | 59 | 92.98 | 89.83 | 91.38 |
| Kích thước trong 2 chiều | Số 2 | 67 | 100 | 100 | 100 |
| Kích thước trong văn bản | KÍCH THƯỚC\_VĂN BẢN | 19 | 100 | 100 | 100 |
| Visib. trên ảnh MMG khác. | DỰ KIẾN | 5 | 100 | 80.00 | 88.89 |
| Giải thích | GIẢI THÍCH | 173 | 98.80 | 95.38 | 97.06 |
| Tìm anaphora |  |  |  |  |  |
| Tổng quát | THAY ĐỔI\_REF | 13 | 100 | 100 | 100 |
| Kiểu | TYPE\_REF | 6 | 100 | 66.67 | 80.00 |
| Kích thước | KÍCH THƯỚC\_REF | 3 | 100 | 100 | 100 |
| Loại mô vú | BTISSUE | 818 | 99.39 | 100 | 99.70 |
| Xơ hóa Stroma | SỢI | 17 | 100 | 100 | 100 |
| Tính năng mô |  |  |  |  |  |
| Mật độ | MẬT ĐỘ GLANDjMẬT ĐỘ | 84 | 98.82 | 100 | 99.41 |
| Sắc sụt | GLANDjMACUL | 71 | 100 | 98.59 | 99.29 |
| Lượng | VÒNG QUAY | 208 | 99.52 | 100 | 99.76 |
| Đều đặn | GLANDjTHƯỜNG XUYÊN | 80 | 100 | 100 | 100 |
| Nhân vật | CHAR | 93 | 98.92 | 98.92 | 98.92 |
| Các khối phụ rel. |  |  |  |  |  |
| sự so sánh |  |  |  |  |  |
| Mật độ | CMP\_MẬT ĐỘ | 17 | 100 | 100 | 100 |
| Số lượng | CMP\_QUANT | 7 | 100 | 85.71 | 92.31 |
| Localization |  |  |  |  |  |
| Anaphoric | LOC\_REF | 9 | 100 | 44.44 | 61.54 |
| Bộ phận cơ thể | LOCjBODY\_PART | 2005 | 99.85 | 99.60 | 99.73 |
| Phần giải phẫu | LOCjLOC\_A | 7 | 83.33 | 71.43 | 76.92 |
| Phần thông thường | LOCjLOC\_CONV | 273 | 100 | 98.90 | 99.45 |
| 2 chiều conv. | LOCjLOC\_CONV1 | 99 | 100 | 97.98 | 98.98 |
| phần |  |  |  |  |  |
| Bên | LOCjL\_R | 2009 | 99.85 | 99.55 | 99.70 |
| Hình ảnh mmg | CHIẾU | 14 | 100 | 100 | 100 |
| phối cảnh |  |  |  |  |  |
| Chẩn đoán | CHẨN ĐOÁN\_RTG | 1517 | 100 | 99.80 | 99.90 |

Bảng 2

Kết quả đánh giá cho các thuộc tính được tìm thấy trong 705 báo cáo chụp nhũ ảnh thử nghiệm, phần 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Thuộc tính | | Trường hợp | Trước. | Nhớ | F |
| Bản địa hóa với |  |  |  |  |  |
| Chẩn đoán |  |  |  |  |  |
| Bộ phận cơ thể | LOC\_DjBODY\_PHẦN | 988 | 100 | 99.90 | 99.95 |
| Phần giải phẫu | LOC\_DjLOC\_A | 61 | 100 | 100 | 100 |
| Bên | LOC\_DjL\_R | 989 | 100 | 100 | 100 |
| Phẫu thuật vú |  |  |  |  |  |
| Kiểu | PHẪU THUẬT | 18 | 100 | 77.78 | 87.50 |
| Lý do | LÝ DO | 5 | 100 | 100 | 100 |
| Ngày thi trước | TRƯỚC\_THI | 309 | 100 | 99.03 | 99.51 |
| Thay đổi so với trước. |  |  |  |  |  |
| Thi |  |  |  |  |  |
| Về kích thước | KÍCH THƯỚC\_THAY ĐỔI | 321 | 100 | 100 | 100 |
| Thay đổi số lượng | QUANT\_THAY ĐỔI | 258 | 100 | 100 | 100 |
| Trong độ bão hòa | SATURAT\_THAY ĐỔI | 321 | 100 | 100 | 100 |
| Khuyến nghị |  |  |  |  |  |
| Loại kiểm tra | KHUYÊN | 769 | 99.35 | 99.87 | 99.61 |
| Lý do | KHUYẾN NGHỊ\_LÝ DO | 64 | 100 | 100 | 100 |
| Tính năng tóm tắt |  |  |  |  |  |
| Lớp báo cáo | REPORT\_CLASS | 705 | 99.72 | 99.72 | 99.72 |
| Nơi nào có | REPORT\_WITH\_FIND | 705 | 99.57 | 99.57 | 99.57 |
| Kết quả? |  |  |  |  |  |
| Độ tin cậy của MMG | MMG\_REL | 705 | 100 | 100 | 100 |

nhưng một số xử lý dữ liệu bổ sung đã được thực hiện. Ví dụ: ngày tuyệt đối cho các khoảng thời gian tương đối được tính.

1. Kết quả

Trong [Bảng 1 và 2,](#_bookmark15)  chúng tôi bao gồm đánh giá cho tất cả các thuộc tính xảy ra ít nhất một lần trong 705 báo cáo chụp nhũ ảnh được lấy từ bộ thử nghiệm.

Bộ thử nghiệm không lớn lắm nên nhiều thuộc tính có tần số bằng hoặc gần bằng 0 (từ 66 thuộc tính 28 xảy ra ít hơn 10 lần). Nhưng đồng thời một số khá nhiều - 21 vụ nổ xảy ra hơn 100 lần.

Kết quả của nhiệm vụ trích xuất giá trị thuộc tính rất tốt. Hai mươi tám thuộc tính được công nhận chính xác 100%, 45 thuộc tính có điểm *F* trên 99%. Đây cũng là kết quả cho tất cả các phân tích được tính cùng nhau. Chỉ có năm thuộc tính được công nhận dưới 80%. Kết quả tồi tệ nhất (61,5% điểm *F* cho chín trường hợp) đạt được để nhận biết vị trí tương đối. Ví dụ, các quy tắc ngữ pháp của chúng ta mô tả *w jego okolicy* 'trong môi trường xung quanh' nhưng chúng bỏ lỡ các cách diễn đạt như *w jego obre ̨bie* 'trong giới hạn của nó' hoặc *wokół niego* 'xung quanh nó'.

Nhiệm vụ hợp nhất thông tin khó khăn hơn. Các khối phụ mô và khối bắt đầu được công nhận với điểm F trên 99%. Kết quả tồi tệ nhất thu được khi tách các mô tả phát hiện (84,48% điểm F để tìm kết thúc khối) và điều này ảnh hưởng đến kết quả cho các đầu khối mô (94,6%).

Bảng 3

Kết quả đánh giá nhóm dữ liệu cho 705 báo cáo chụp nhũ ảnh.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Marker | | Trường hợp | Trước. | Nhớ | F |
| Bắt đầu khối mô | muỗng canh | 706 | 99.86 | 99.58 | 99.72 |
| Kết thúc khối mô | Tbe | 706 | 94.74 | 94.48 | 94.61 |
| Bắt đầu khối phụ mô | Tb | 966 | 99.69 | 99.90 | 99.79 |
| Kết thúc khối phụ mô | Te | 966 | 97.22 | 97.62 | 97.42 |
| Bắt đầu tìm chặn | Fb | 347 | 82.07 | 87.03 | 84.48 |
| Tìm kết thúc khối | Fe | 347 | 86.38 | 91.35 | 88.80 |
| Phần đề xuất bắt đầu | Rb | 610 | 99.84 | 99.84 | 99.84 |

Khuyến nghị khối bắt đầu rất dễ nhận biết, do đó kết quả là 99,8% F-score, xem [Bảng 3](#_bookmark16).

Chúng tôi đã sử dụng phiên bản hiệu chỉnh thủ công của đầu ra hệ thống làm tiêu chuẩn tham chiếu. Các chú thích cần thiết cho nhiệm vụ của chúng tôi chi tiết và phức tạp hơn nhiều so với chú thích cho các nhiệm vụ phân loại, và sửa đầu ra hệ thống là một cách để đảm bảo tính nhất quán của người đánh giá. Một nhược điểm của giải pháp này là dữ liệu như vậy có khả năng thiên vị về hệ thống. Để kiểm tra ngắn gọn tiêu chuẩn tham chiếu của chúng tôi, chúng tôi đã thực hiện một thí nghiệm trong đó 20 báo cáo chụp nhũ ảnh (được chọn ngẫu nhiên nhưng ưu tiên các ghi chú mô tả các phát hiện) được phân tích thủ công bởi hai người chú thích. Đối với 20 báo cáo này, tiêu chuẩn tham chiếu của chúng tôi bao gồm 550 giá trị thuộc tính. Một người chú thích nhận ra 11 giá trị của các thuộc tính bổ sung và bỏ qua năm, người chú thích thứ hai nhận ra thiếu 11 giá trị giống nhau và bỏ qua 16. Chúng tôi đã tính toán phần trăm thỏa thuận và Kappa giữa hai người chú thích và giữa mỗi người chú thích và tiêu chuẩn tham chiếu được đặt trên 36 thuộc tính xảy ra ít nhất một lần trong tập hợp 20 báo cáo, như

Bảng 4

Nghiên cứu thỏa thuận so sánh các chú thích từ hai người chú thích độc lập với nhau và so với các chú thích tiêu chuẩn tham chiếu. A1 = Chú thích 1 chú thích; A2 = chú thích 2; ref = chú thích tiêu chuẩn tham chiếu. PA = phần trăm độ chính xác; K = Hệ số Kappa.

thể hiện trong [Bảng 4](#_bookmark17). Đối với 20 thuộc tính, cả ba so sánh đều cho thấy sự đồng ý giống hệt nhau. Hệ số Kappa cho bất kỳ cặp nào được so sánh đều thấp hơn 0,7 đối với bốn thuộc tính. Đối với ba người trong số họ, sự bất đồng là do một thiếu sót có hệ thống của một người chú thích. Đối với thuộc tính thứ tư (WITH\_CALC), một trình chú thích tốt hơn cả trình chú thích thứ hai và tiêu chuẩn tham chiếu. Chỉ đối với mô tả bản địa hóa rất chi tiết, sự đồng thuận giữa hai người chú thích mới vượt quá sự thỏa thuận giữa con người và chú thích tham chiếu. Hệ số Kappa giữa mỗi người chú thích và tập tham chiếu cho thuộc tính này gần 0,8, trong khi sự đồng ý của con người bằng 1. Mặc dù bộ tiêu chuẩn tham chiếu của chúng tôi được tạo ra thông qua một đầu ra hệ thống hiệu chỉnh duy nhất của con người, nhưng các chú thích độc lập chú thích mà không có bất kỳ phản hồi hệ thống nào cũng đồng ý với bộ tiêu chuẩn tham chiếu

Bảng 5

Kết quả đánh giá cho các thuộc tính được tìm thấy trong 100 tài liệu xả thử nghiệm, phần 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Thuộc tính | | Trường hợp | Trước. | Nhớ | F |
| TẤT CẢ CÁC THUỘC TÍNH |  | 4021 | 99.26 | 96.49 | 97.86 |
| Dữ liệu tài liệu |  |  |  |  |  |
| Bắt đầu | DOC\_BEG | 100 | 100 | 100 | 100 |
| Ngày | DOC\_DAT | 100 | 100 | 100 | 100 |
| Xác định | ID | 100 | 100 | 98 | 98.99 |
| Tiếp tục | CONT | 98 | 100 | 97.96 | 98.97 |
| Nhập viện | H\_TỪ | 99 | 100 | 98.99 | 99.49 |
| từ |  |  |  |  |  |
| Nhập viện đến | H\_ĐẾN | 99 | 100 | 98.99 | 99.49 |
| Bắt đầu | KHỦNG HOẢNG EPICRISIS | 100 | 100 | 100 | 100 |
| tóm tắt |  |  |  |  |  |
| Dữ liệu bệnh nhân |  |  |  |  |  |
| Xác định | Mã số\_P | 100 | 99.01 | 100 | 99.50 |
| Tình dục | ID\_P\_SEX | 101 | 100 | 100 | 100 |
| Tuổi | ID\_TUỔI | 192 | 100 | 98.44 | 99.21 |
| Chiều cao | CHIỀU CAO | 87 | 100 | 98.85 | 99.42 |
| Trọng lượng | TRỌNG LƯỢNG | 88 | 100 | 100 | 100 |
| BMI | BMI | 88 | 98.82 | 95.45 | 97.11 |
| Trọng lượng bằng lời nói | W\_TRONG\_TỪ | 62 | 98.41 | 100 | 99.20 |
| Đặc điểm của bệnh tiểu đường |  |  |  |  |  |
| Kiểu | D\_LOẠI | 173 | 98.84 | 98.27 | 98.55 |
| Cân | D\_ĐIỀU KHIỂN | 212 | 98.54 | 95.75 | 97.13 |
| Acetonuria | ACET | 11 | 91.67 | 100 | 95.65 |
| Phương pháp điều trị | D\_ĐIỀU TRỊ | 50 | 100 | 92 | 95.83 |
| Khi được chẩn đoán |  |  |  |  |  |
| Năm | Y\_DAT | 6 | 100 | 100 | 100 |
| Dữ liệu tương đối | D\_NUM | 41 | 100 | 97.56 | 98.77 |
|  | D\_ĐƠN VỊ | 41 | 100 | 97.56 | 98.77 |
| Năm của cuộc đời | NĂM\_CỦA CUỘC SỐNG | 3 | 100 | 100 | 100 |
| Bằng lời nói | TỪ\_TRONG\_W | 32 | 93.94 | 96.88 | 95.38 |
| Bệnh |  |  |  |  |  |
| Tất cả bệnh nhân tiểu đường | COMP | 369 | 97.35 | 99.46 | 98.39 |
| đồng lõa. |  |  |  |  |  |
| Loại bệnh võng mạc | BỆNH VÕNG MẠC\_T | 120 | 98.36 | 100 | 99.17 |
| Với bệnh điểm vàng | VỚI\_bệnh điểm vàng | 32 | 100 | 90.63 | 95.08 |
| Không có compl. của loại | N\_COMP | 34 | 100 | 100 | 100 |
| Đi kèm | ACC\_BỆNH | 141 | 100 | 100 | 100 |
| Tự miễn dịch | BỆNH TỰ ĐỘNG | 4 | 100 | 100 | 100 |
| Kết quả kiểm tra |  |  |  |  |  |
| Creatinin lev. 1 | CREATIN1 | 96 | 100 | 100 | 100 |
| Creatinin lev. 2 | CREATIN2 | 3 | 60 | 100 | 75 |
| HbA1C | HBA1C | 146 | 100 | 93.15 | 96.45 |
| Cholesterol LDL | LDL | 81 | 100 | 100 | 100 |
| Vi albuminury | LEV1 |  | 100 | 92.59 | 96.15 |
| Chế độ ăn uống được khuyến nghị |  |  |  |  |  |
| Lượng calo tối thiểu | CAL\_MIN | 102 | 100 | 94.12 | 96.97 |
| Tối đa lượng calo | CAL\_TỐI ĐA | 4 | 100 | 50 | 66.67 |
| Bữa ăn tối thiểu | BỮA ĂN\_PHÚT | 95 | 100 | 87.37 | 93.26 |
| Tối đa bữa ăn | BỮA ĂN\_TỐI ĐA | 20 | 100 | 80 | 88.89 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Thuộc tính | #ref | A1+tham khảo  PA | K | A2+tham khảo  PA | K | A1+A2  PA | K |
| Tìm kiếm các tính năng đầu tiên |  |  |  |  |  |  |  |
| ANAT\_CHANGE (loại) | 9 | 0.95 | 0.93 | 0.95 | 0.93 | 1 | 1 |
| ĐƯỜNG NÉT | 4 | 0.95 | 0.83 | 1 | 1 | 1 | 1 |
| HÌNH DẠNG | 0 | 0.95 | 0 | 0.95 | 0 | 1 | 1 |
| VỚI\_CALC | 2 | 0.90 | 0.46 | 0.95 | 0.78 | 0.95 | 0.78 |
| Vô cùng | 15 | 0.95 | 0.93 | 0.4 | 0 | 0.40 | 0 |
| Kích thước | 7 | 1 | 1 | 1 | 1 | 1 | 1 |
| GIẢI THÍCH | 13 | 1 | 1 | 1 | 1 | 1 | 1 |
| Localization |  |  |  |  |  |  |  |
| LOCjBODY\_PART, LOCjL\_R | 16 | 1 | 1 | 0.95 | 0.94 | 0.95 | 0.94 |
| LOCjLOC\_CONV | 9 | 0.85 | 0.78 | 0.85 | 0.78 | 1 | 1 |
| LOCjLOC\_CONV1 | 2 | 0.95 | 0.78 | 0.95 | 0.78 | 1 | 1 |
| CHẨN ĐOÁN\_RTG | 13 | 1 | 1 | 1 | 1 | 1 | 1 |
| Các tính năng tìm kiếm thứ hai |  |  |  |  |  |  |  |
| GIẢI THÍCH | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| LOCjLOC\_CONV | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Mô tả chung về vú |  |  |  |  |  |  |  |
| BTISSUE (loại mô) | 19 | 1 | 1 | 0.95 | 0.94 | 0.95 | 0.94 |
| LOCjBODY\_PART, LOCjL\_R | 19 | 1 | 1 | 1 | 1 | 1 | 1 |
| Xơ hóa chất nền (FIBER) | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| Đặc điểm mô tuyến |  |  |  |  |  |  |  |
| GLANDjMACUL | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| VÒNG QUAY | 7 | 1 | 1 | 1 | 1 | 1 | 1 |
| GLANDjTHƯỜNG XUYÊN | 4 | 1 | 1 | 1 | 1 | 1 | 1 |
| CHAR | 1 | 1 | 1 | 0.95 | 0 | 0.95 | 0 |
| Đặc điểm của mảnh mô |  |  |  |  |  |  |  |
| BTISSUE (loại mô) | 9 | 1 | 1 | 1 | 1 | 1 | 1 |
| LOCjBODY\_PART,LOCjL\_R | 16 | 1 | 1 | 1 | 1 | 1 | 1 |
| LOCjLOC\_CONV | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| LOCjLOC\_CONV1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 |
| CHIẾU | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| CHẨN ĐOÁN\_RTG | 18 | 0.95 | 0.91 | 0.90 | 0.83 | 0.85 | 0.75 |
| LOC\_D | 13 | 1 | 1 | 0.80 | 0.58 | 0.80 | 0.58 |
| PHẪU THUẬT (loại phẫu thuật) | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| REASON (lý do phẫu thuật) | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| PREV\_EXAM(ngày) | 11 | 0.95 | 0.94 | 0.80 | 0.74 | 0.85 | 0.81 |
| Thay đổi so với kỳ thi trước | 10 | 0.95 | 0.92 | 0.95 | 0.92 | 1 | 1 |
| Khuyến nghị |  |  |  |  |  |  |  |
| Kiểm tra đầu tiên | 13 | 1 | 1 | 1 | 1 | 1 | 1 |
| Thời gian đầu tiên | 15 | 1 | 1 | 0.95 | 0.91 | 1 | 1 |
| Kiểm tra lần thứ hai | 11 | 1 | 1 | 1 | 1 | 1 | 1 |
| Thời gian thứ hai | 3 | 1 | 1 | 1 | 1 | 1 | 1 |
| Lý do | 1 | 1 | 1 | 0.90 | 0.47 | 0.90 | 0.47 |
| TẤT CẢ CÁC THUỘC TÍNH | 270 | 0.98 | 0.97 | 0.95 | 0.91 | 0.96 | 0.93 |

Bảng 6

Kết quả đánh giá cho các thuộc tính được tìm thấy trong 100 tài liệu kiểm tra bệnh tiểu đường, phần 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Thuộc tính | | Trường hợp | Trước. | Nhớ | F |
| Liệu pháp insulin  *Mô tả liều lượng*  Thuốc insulin Min. liều lượng  Tối đa liều lượng  *Truyền liên tục*  Thuốc insulin Tối thiểu cơ bản mỗi ngày Tối đa cơ bản mỗi ngày Tối thiểu bolus mỗi bữa tối đa bolus mỗi bữa  Thuốc uống  Khác nhau  Lý do của bệnh viện. Đào tạo bệnh nhân Bắt đầu điều trị insulin  Sửa đổi liệu pháp Liều insulin modif. Tự giám sát  Điều chỉnh chế độ ăn uống Quan sát chế độ ăn uống | I\_LOẠI | 298 | 100 | 92.62 | 96.17 |
| DOSE\_MIN | 298 | 100 | 90.60 | 95.07 |
| DOSE\_MAX | 35 | 100 | 97.14 | 98.55 |
| INS\_TYPE | 2 | 100 | 100 | 100 |
| TOT\_MIN\_BASE | 1 | 100 | 100 | 100 |
| TOT\_MAX\_CƠ SỞ | 2 | 100 | 100 | 100 |
| B\_MIN | 2 | 100 | 100 | 100 |
| B\_TỐI ĐA | 2 | 100 | 100 | 100 |
| ORAL\_ĐIỀU TRỊ | 76 | 96.15 | 98.68 | 97.40 |
| LÝ DO | 83 | 98.73 | 93.98 | 96.30 |
| GIÁO DỤC | 46 | 100 | 91.30 | 95.45 |
| TRỊ LIỆU\_XIN LỖI | 3 | 66.67 | 66.67 | 66.67 |
| TRỊ LIỆU\_ MODIF | 23 | 100 | 91.30 | 95.45 |
| DOSE\_MODIF | 10 | 90 | 90 | 90 |
| TỰ GIÁM SÁT | 1 | 0 | 0 | 0 |
| CHẾ ĐỘ ĂN KIÊNG\_ĐIỀU CHỈNH | 1 | 100 | 100 | 100 |
| CHẾ ĐỘ ĂN UỐNG\_QUAN SÁT | 1 | 0 | 0 | 0 |

như họ đã làm với nhau, cho thấy chất lượng của bộ tiêu chuẩn tham chiếu của chúng tôi.

Đối với thí nghiệm bệnh tiểu đường, đánh giá 55 thuộc tính (tổng cộng 68), xảy ra trong bộ thử nghiệm gồm 100 tài liệu được trình bày trong [Bảng 5 và 6](#_bookmark18). Một số thuộc tính hiếm trong bộ thử nghiệm, vì vậy việc xen kẽ kết quả nhận dạng của chúng là không đáng tin cậy. Mười bốn dị ứng xảy ra ít hơn 10 lần: chín trong số đó được nhận biết chính xác 100%; nhưng sự xuất hiện đơn lẻ của hai thuộc tính không được công nhận, cho điểm *F* 0%. Đối với 41 thuộc tính còn lại: 16 thuộc tính có điểm *F* trên 99% và ba thuộc tính có điểm *F* dưới 95%. Thuộc tính COMPlication thường xuyên nhất xảy ra 369 lần và có 98,39%  *điểm F*.

1. Sự thảo luận

Kết quả của các hệ thống trích xuất dựa trên quy tắc phụ thuộc nhiều vào chất lượng của văn bản. Lỗi đánh máy, có thể hiểu được đối với một người, ví dụ: 'I0 mm' thay vì '10 mm', hầu như không thể sửa chữa bằng một chương trình máy tính. Do đó, các văn bản sẽ trải qua quá trình xử lý tự động nên được viết rất cẩn thận. Trong các quy tắc ngữ pháp thủ công, một số lỗi chính tả rất điển hình, ví dụ, thiếu khoảng trắng sau một chữ viết tắt, có thể được tính toán, nhưng không thể lường trước được tất cả các loại lỗi. Ví dụ, thông tin về chế độ ăn uống được khuyến nghị bao gồm hai phần thông tin: bao nhiêu calo và bao nhiêu bữa ăn được khuyến nghị. Cả hai giá trị đều có thể được biểu thị dưới dạng phạm vi. Vì vậy, cuối cùng thông tin được thể hiện bằng bốn thuộc tính. Cả bốn thuộc tính đều có độ chính xác 100% nhưng khả năng nhớ lại thấp chủ yếu là do các vấn đề về đánh máy (chữ viết tắt không chuẩn, dấu câu lạ: sáu trong số mười hai trường hợp cụm từ không được nhận dạng là do thiếu khoảng cách giữa số và từ *posiłek* 'meal'), vì vậy khả năng nhớ thuộc tính MEALS\_MIN là 87,37%.

Một hạn chế đã biết của các hệ thống dựa trên quy tắc là sự cần thiết phải nhìn thấy trước tất cả các cách có thể diễn đạt thông tin được rút ra. Nếu ngữ pháp không bao gồm tất cả các khả năng, khả năng nhớ lại hệ thống sẽ giảm, ví dụ: khả năng nhớ lại thấp hơn đối với thuộc tính INTERPRETATION (94,25%) phản ánh thực tế là một số cụm từ được các bác sĩ sử dụng không được dự đoán bởi ngữ pháp của chúng ta (ví dụ: viết tắt *fa* cho *u tuyến xơ).* Thực tế này cũng có thể là lý do cho độ chính xác thấp hơn, nếu một số thay đổi ngữ cảnh trong ý nghĩa không được công nhận. Tồi tệ nhất

Kết quả thu được cho việc nhận biết các thay đổi giải phẫu (độ chính xác 93,8%) chủ yếu là do phân loại không chính xác các phát hiện trước đó là vẫn còn tồn tại (phủ định không cục bộ không được công nhận hoặc bối cảnh kiểm tra trước đó). Hai trong số cả ba trường hợp mắc các loại bệnh tiểu đường không được công nhận là do các biểu hiện không lường trước được gây ra. Cụm từ: *cukrzyca spowodowana leczeniem sterydami* 'bệnh tiểu đường do điều trị ste-roid ' không được hệ thống của chúng tôi công nhận. Ví dụ thứ hai chưa được công nhận là *cukrzyca typu 3* 'bệnh tiểu đường loại 3'.

Độ chính xác thấp hơn thu được cho nhiệm vụ xác định khối tìm kiếm bắt đầu một phần là do lỗi trong việc nhận dạng các tính năng ANAT\_CHANGE và INTERPRETATION và một số lỗi nhận dạng bản địa hóa. Lý do quan trọng khác của việc chèn đường viền mô tả tìm thấy không chính xác là tách mô tả mô không chính xác khỏi phân đoạn mô tả phát hiện. Lỗi này tương đối dễ mắc phải, vì rất thường xuyên các mô thay đổi thành một mô tả của một phát hiện rất trơn tru mà không lặp lại bản địa hóa vượt qua biên giới.

Một hệ thống dựa trên quy tắc không thể nhận ra một tên mới cho đến khi nó được xác định rõ ràng. Ví dụ, 3 tên mới của thuốc insulin dẫn đến 9 trường hợp thuộc tính I\_TYPE không được công nhận. Trong trường hợp các hệ thống IE dựa trên các phương pháp thống kê, có thể nhận ra các loại insulin chưa được biết đến, đơn giản vì hệ thống có thể xác định chúng trên cơ sở các ví dụ có sẵn thông qua đào tạo. Các phương pháp thống kê rất tốt trong việc nhận biết các thông tin đơn giản như bệnh xuất hiện, nhưng không đơn giản để chỉ ra những biến chứng nào trong số đó là biến chứng của bệnh tiểu đường. Các phương pháp dựa trên quy tắc tốt hơn trong việc nhận biết thông tin hiếm và phức tạp, ví dụ như mô tả liệu pháp truyền insulin liên tục (14 ví dụ trong dữ liệu đào tạo), bao gồm: loại insulin; mô tả liều cơ bản được mô tả là liều mỗi giờ hoặc tổng liều hàng ngày; và mô tả liều bolus mỗi bữa ăn hoặc liều cho các bữa ăn cụ thể. Cả hai cách tiếp cận dường như bổ sung cho một định nghĩa nhất định, vì vậy, có lẽ giải pháp tốt nhất là kết hợp cách tiếp cận dựa trên quy tắc với một số phương pháp thống kê.

Lưu ý rằng quy trình phát triển tập tham chiếu có thể ảnh hưởng đến kết quả. Như chúng tôi đã đề cập, một tập hợp dựa trên dữ liệu được chú thích tự động thiên về hệ thống - một công cụ sửa chữa có xu hướng bảo toàn các quyết định của hệ thống và có thể không có sự kiện nào đã bị hệ thống bỏ qua. Mặt khác, chú thích thủ công có thể không nhất quán trong cách chú thích cùng một thông tin và trong các quyết định về những gì đáng và những gì không đáng để chú thích. Các hướng dẫn chú thích càng chi tiết, chú thích thủ công càng gần với chú thích tự động. Để đánh giá tiêu chuẩn tham chiếu của chúng tôi, chúng tôi đã thực hiện một nghiên cứu chú thích có chọn lọc, cho thấy rằng trong trường hợp cụ thể này, chú thích có sự hỗ trợ của hệ thống rất gần với mức đạt được bằng tay.

1. Kết luận

Mục tiêu của nghiên cứu được trình bày là xác định xem liệu các kỹ thuật IE dựa trên kiến thức về một ngôn ngữ tự nhiên cụ thể và miền ứng dụng kết hợp với các quy trình xử lý sau lĩnh vực cụ thể, có thể cho kết quả khả quan trong việc trích xuất dữ liệu từ các tài liệu lâm sàng văn bản tự do hay không. Trong bài báo, chúng tôi đã mô tả một hệ thống IE mã hóa thông tin chi tiết từ các văn bản y học được viết bằng tiếng Ba Lan. Ngay cả đối với các văn bản tiếng Anh, chỉ có một số hệ thống MLP cung cấp số lượng chi tiết này.

Do sự phức tạp của các mẫu và thiếu dữ liệu đào tạo, chúng tôi đã chọn phương pháp IE dựa trên quy tắc. Phương pháp này, mặc dù đã được sử dụng trong nhiều ứng dụng, nhưng vẫn chưa được áp dụng cho dữ liệu Ba Lan. Thật thú vị khi thấy sự đa dạng của các dạng từ và trật tự từ tự do có thể ảnh hưởng đến khả năng sử dụng của các kỹ thuật IE như thế nào. Trong hệ thống của chúng tôi, sự thay đổi của các dạng biến đổi phải được giải quyết thông qua việc sử dụng cả hai: từ vựng chung và từ vựng miền bao gồm các dạng từ hình thái khác nhau (thậm chí cả tên thuốc

có thể được biến tấu). Trật tự từ tự do phải được tính đến trong các quy tắc ngữ pháp (một số hoán vị là cần thiết) hoặc trong giai đoạn sau quá trình (khi các nhận thức bề mặt khác nhau không cho phép thông tin hợp nhất trực tiếp trong ngữ pháp).

Việc tạo ra một hệ thống dựa trên quy tắc đòi hỏi rất nhiều kiến thức về lĩnh vực và tốn thời gian, nhưng chúng tôi tin rằng các kết quả được hiển thị trong bài báo hỗ trợ luận điểm rằng các hệ thống như vậy có thể đáng tin cậy và hữu ích cho việc xử lý dữ liệu lâm sàng tự động. Sau khi kiểm tra kỹ lưỡng và thích ứng với các nguồn dữ liệu cụ thể, chúng có thể được áp dụng cho các tài liệu lịch sử để trích xuất dữ liệu cho mục đích thống kê. Đặc biệt, chúng cho phép trích xuất thông tin có trong các tệp này, nhưng trước đây được coi là ít quan trọng hơn và không được đưa vào cơ sở dữ liệu có cấu trúc. Phương pháp được gửi trước cũng có thể được sử dụng trên dữ liệu của bệnh nhân hiện đang được thu thập để trích xuất các sự kiện quan trọng nhất và để chọn các nhóm bệnh cần được chú ý đặc biệt. Ngoài ra, chất lượng tài liệu cao hơn có thể đạt được bằng cách sử dụng các hệ thống như vậy để kiểm tra tính nhất quán của dữ liệu dạng bảng và văn bản, và để tìm kiếm các tài liệu trong đó thiếu một loại thông tin cụ thể.

Tham khảo

1. Agirre E, Edmonds P, biên tập viên. Định nghĩa từ. Thuật toán và ứng dụng. Công nghệ văn bản, giọng nói và ngôn ngữ. Springer-Verlag; 2007.
2. Bontcheva K, Dimitrov M, Maynard D, Tablan V, Cunningham H. Phương pháp nông cho độ phân giải đồng tham chiếu thực thể được đặt tên. Trong: Proceedings of traitement automatique des langues naturelles. TALN Nancy; 2002.
3. Boytcheva S, Strupchanska A, Paskaleva E, Tcharaktchiev D. Một số khía cạnh của xử lý phủ định trong hồ sơ sức khỏe điện tử. Trong: Kỷ yếu cơ sở hạ tầng ngôn ngữ và lời nói của hội thảo quốc tế để tiếp cận thông tin ở các nước Balkan; 2005. trang 1–8.
4. Bunescu R, Ge R, Kate RJ, Mooney RJ, Wong YW. Học cách chiết xuất protein và tương tác của chúng từ các bản tóm tắt medline. Trong: Kỷ yếu của hội thảo ICML-2003 về học máy trong tin sinh học; 2003. trang 46–53.
5. Burnside E, Rubin D, Shachter R. Một mạng Bayes cho chụp nhũ ảnh. Trong: Kỷ yếu của hội nghị chuyên đề về hiệp hội tin học y tế Hoa Kỳ; 2000. tr. 16–110.
6. Burnside E, Strasberg H, Rabin D. Lập chỉ mục tự động chụp nhũ ảnh bằng cách sử dụng phù hợp bình phương nhỏ nhất tuyến tính. Trong: Hội nghị quốc tế CARS 2000 về X quang và phẫu thuật có sự hỗ trợ của máy tính, San Francisco, CA; 2000.
7. Buyko E, Tomanek K, Hahn U. Độ phân giải của các dấu chấm lửng phối hợp trong các thực thể được đặt tên sinh học bằng cách sử dụng trường ngẫu nhiên có điều kiện. Trong: Kỷ yếu của hội nghị lần thứ 10 của hiệp hội ngôn ngữ học tính toán Thái Bình Dương; 2007.
8. Thợ mộc B. Logic của cấu trúc tính năng được nhập. Cambridge tract về khoa học máy tính lý thuyết. Nhà xuất bản Đại học Cambridge; 1992.
9. Chapman WW, Hanbury B, Cooper P, Buchanan G. Một thuật toán đơn giản để xác định các phát hiện và bệnh bị phủ nhận trong tóm tắt xuất viện. J Biomed Thông báo 2001; 34:204–30.
10. Chapman WW, Dowling JN, Chu D. Bối cảnh: một thuật toán để xác định các đặc điểm ngữ cảnh từ văn bản lâm sàng. Trong: ACL 2007. Kỷ yếu hội thảo về BioNLP 2007; 2007.
11. Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Xác định các phát hiện hô hấp trong báo cáo khoa cấp cứu để giám sát sinh học bằng metamap. Trong: Medinfo; 2004. trang 487–91.
12. Chapman WW, Dowling JN, Ivanov O, Gesteland P, Olszewski R, Espino J, và cộng sự. Đánh giá các ứng dụng xử lý ngôn ngữ tự nhiên được áp dụng cho giám sát dịch bệnh và bùng phát. Trong: Kỷ yếu của hội nghị chuyên đề lần thứ 36 về giao diện: khoa học máy tính và thống kê, Baltimore, MD; Năm 2004.
13. Chen H, Fuller S, Friedman C, Hersh W, biên tập viên. Tin học y tế: quản lý tri thức và khai thác dữ liệu trong y sinh. Springer; 2005.
14. Christensen LM, Haug PJ, Fiszman M. MPLUS: hiểu ngôn ngữ y tế xác suất. Trong: Kỷ yếu hội thảo ACL về xử lý ngôn ngữ tự nhiên trong lĩnh vực y sinh; 2002.
15. Droz\_ dz\_ yñski W, Krieger HU, Piskorski J, Schäfer U, Xu F. Xử lý nông

với cấu trúc tính năng thống nhất và đánh máy – nền tảng và ứng dụng. AI J KI-Zeitschrift của Đức 2004;01/04.

1. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, và cộng sự. Một thử nghiệm có kiểm soát về phân loại phủ định tự động từ ghi chú lâm sàng. BMC Med Thông báo Ra quyết định 2005;5.
2. Fiszman M, Chapman W, Aronsky D, Evans S, Haug P. Tự động phát hiện viêm phổi cấp tính do vi khuẩn từ các báo cáo chụp X-quang ngực. J Am Med Thông báo Assoc 2000;7.
3. Friedman C, Alderson P, Austin J, Cimino J, Johnson S. Một bộ xử lý văn bản ngôn ngữ tự nhiên chung cho X quang lâm sàng. J Am Med Inform Assoc 1994;1.
4. Friedman C, Hripcsak G. Xử lý ngôn ngữ tự nhiên và tương lai của nó trong y học. Acad Med 1999; 74(8):890–5.
5. Gasperin C, Briscoe T. Độ phân giải anaphora thống kê trong các văn bản y sinh. Trong: Kỷ yếu của hội nghị quốc tế lần thứ 22 về ngôn ngữ học tính toán (Coling 2008), Manchester, Vương quốc Anh, tháng 8 năm 2008. Tổ chức Coling 2008

Uỷ ban; 2008. trang 257–64.

1. Hahn U, Romacker M, Schultz S. MEDSYNDIKATE – một hệ thống ngôn ngữ tự nhiên để trích xuất thông tin y tế từ các báo cáo phát hiện. Int J Med Inform 2002; 67:63–74.
2. Hahn U, Romacker M, Schulz S. Tạo kho kiến thức từ các báo cáo y sinh: hệ thống khai thác văn bản MEDSYNDIKATE. Pac Symp Biocomput 2002; 338–49.
3. Harkema H, Gaizauskas R, Hepple M, Roberts A, Roberts I, Davis N, và cộng sự. Một tài nguyên thuật ngữ quy mô lớn để xử lý văn bản y sinh. Trong: Hội thảo HLT-NAACL 2004: BioLINK 2004 liên kết tài liệu sinh học. Bản thể và cơ sở dữ liệu; 2004.
4. Harkema H, Roberts I, Gaizauskas R, Hepple M. Trích xuất thông tin từ hồ sơ lâm sàng. Trong: Kỷ yếu cuộc họp tất cả các tay khoa học điện tử lần thứ 4 của Vương quốc Anh, Nottingham, Vương quốc Anh; 2005.
5. Hripcsak G, Austin J, Anderson P, Friedman C. Sử dụng xử lý ngôn ngữ tự nhiên để dịch thông tin lâm sàng từ cơ sở dữ liệu gồm 889.921 báo cáo chụp X quang ngực. X quang 2002; 224:157–63.
6. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Mở khóa dữ liệu lâm sàng từ các báo cáo tường thuật: một nghiên cứu về xử lý ngôn ngữ tự nhiên. Ann Thực tập sinh Med 1995; 122 (9): 681–8.
7. Jain NL, Friedman C. Xác định các phát hiện nghi ngờ ung thư vú dựa trên xử lý ngôn ngữ tự nhiên của các báo cáo chụp X-quang tuyến vú. Trong: Kỷ yếu của hội nghị chuyên đề mùa thu hàng năm của hiệp hội tin học y tế Hoa Kỳ; 1997.

trang 829–33.

1. Kopans DB, D'Orsi C, Adler D, Bassett LW, Brenner RJ, Dodd GD, và cộng sự. Hệ thống dữ liệu và báo cáo hình ảnh vú (BI-RADS). Trong: Trường Cao đẳng X quang Hoa Kỳ, Reston, Va; 1993.
2. Marciniak M, Mykowiecka A, Kups'c' A, Piskorski J. Trích xuất nội dung thông minh từ các báo cáo y tế của Ba Lan. Trong: Hội thảo quốc tế về công nghệ truyền thông thông minh cho trí tuệ giao tiếp. LNCS 3490. Springer-Verlag; Năm 2005.
3. Mykowiecka A, Kups'c' A, Marciniak M. Trích xuất và phân loại nội dung y tế dựa trên quy tắc. Trong: Kỷ yếu của ISMIS 2005, Gdan' sk. Springer-Verlag; 2005.
4. Mykowiecka A, Marciniak M. Sửa chính tả tự động theo tên miền cho các báo cáo chụp nhũ ảnh. Trong: Kỷ yếu của hệ thống thông tin thông minh 2006. p. Springer.
5. Mykowiecka A, Marciniak M. Mô hình miền để trích xuất thông tin y tế - Ontology LightMedOnt. Trong: Các khía cạnh của xử lý ngôn ngữ tự nhiên, Festschrift để vinh danh L. Bolc, LNAI 5070. Springer-Verlag.

trang 341–67.

1. Mykowiecka A, Marciniak M, Podsiadły-Marczynkowska T. Một bản thể dựa trên dữ liệu cho một hệ thống trích xuất thông tin từ các báo cáo chụp nhũ ảnh. Trong: Kỷ yếu của hội nghị Protégé Quốc tế lần thứ 10, Budapest; 2007.
2. Không JC, Zhang J, Pustejovsky J. Anaphora resolution trong tài liệu y sinh. Trong: Kỷ yếu của hội nghị chuyên đề quốc tế về giải pháp tham chiếu cho NLP, Alicante, Tây Ban Nha; 2002.
3. Piskorski J, Homola P, Marciniak M, Mykowiecka A, Przepiórkowski A, Wolin' ski M. Trích xuất thông tin cho tiếng Ba Lan bằng nền tảng SProUT. Trong: Xử lý thông tin thông minh và khai thác web. Kỷ yếu của IIS: IIPWM'04. Springer-Verlag; Năm 2004.
4. Sager N, Friedman C, Lyman M. Xử lý ngôn ngữ y tế: quản lý máy tính dữ liệu tường thuật. Công ty xuất bản Addison-Wesley; 1987.
5. Sager N, Lyman M, Bucknall C, Nhân N, Tick LJ. Xử lý ngôn ngữ tự nhiên và biểu diễn dữ liệu lâm sàng. J Am Med Inform Assoc 1994; 1 (2): 142–60.
6. Sanchez-Graillet O, Poesio M, Kabadjov M, Tesar R. Tương tác protein nêu ra loại vấn đề nào để giải quyết anaphora? – một phân tích sơ bộ. Trong: Kỷ yếu của hội nghị chuyên đề quốc tế lần thứ hai về khai thác ngữ nghĩa trong y sinh, Jena, ngày 9–12 tháng 4 năm 2006.
7. Savova G, Coden A, Sominsky I, Johnson R, Ogren P, Groen P, và cộng sự. Phân định nghĩa từ trên hai lĩnh vực: tài liệu y sinh và ghi chú lâm sàng. J Biomed Inform 2008; 41 (6): 1088–100.
8. Soderland SG. Xây dựng hệ thống hiểu văn bản dựa trên máy học. Trong: Kỷ yếu của hội thảo IJCAI-2001 về trích xuất và khai thác văn bản thích ứng; 2001.
9. Spyns P. Xử lý ngôn ngữ tự nhiên trong y học: tổng quan. Phương pháp thông báo Med 1996; 35:285–301.
10. Spyns P, Nhân TN, Baert E, Sager N, Moor GD. Xử lý ngôn ngữ y tế được áp dụng để trích xuất thông tin lâm sàng từ các tài liệu y tế của Hà Lan. Trong: Kỷ yếu của đại hội thế giới lần thứ 9 về tin học y tế (MEDINFO98); 1998. trang 685–9.
11. Taira RK, Soderland SG, Jakobovits RM. Tự động cấu trúc các báo cáo văn bản tự do X quang. Chụp X quang 2001; 21:237–45.
12. Tveit A, Saetre R. Protchew: tự động chiết xuất tên protein từ tài liệu y sinh. Trong: Kỷ yếu của hội nghị quốc tế lần thứ 21 về hội thảo kỹ thuật dữ liệu; 2005.
13. Uzuner O, Sibanda TC, Luo Y, Szolovits P. Một danh hiệu hủy định danh cho tóm tắt xuất viện y tế. Artif Intell Med 2007; 42:13–35.
14. Yang X, Su J, Zhou G, Tan CL. Một cách tiếp cận dựa trên cụm np để phân giải đồng tham chiếu. Trong: Kỷ yếu Coling 2004, Geneva, Thụy Sĩ, 23–27 tháng 8 năm 2004; COLING. trang 226–32.
15. Yangarber R. Tiếp thu kiến thức miền. Trong: Pazienza T. (biên tập). Trích xuất thông tin: trong Kỷ nguyên Web. Giao tiếp ngôn ngữ tự nhiên để tiếp thu kiến thức và các tác nhân thông tin thông minh, LNAI 2700. Springer-Verlag; 2003. tr. 1–28.