Machine Translated by Google

Tạp chí Tin học Y sinh 55 (2015) 73–81

Danh sách nội dung có sẵn tại ScienceDirect

Tạp chí Tin học Y sinh

trang chủ tạp chí: www.elsevier.com/locate/yjbin

CADEC: Tập hợp các chú thích về biến cố bất lợi của thuốc

Sarvnaz Karimi , Alejandro Metke-Jimenez, Madonna Kemp, Chen Wang

CSIRO, Úc

thông tin bài viết

Lịch sử bài viết:
Nhận ngày 24 tháng 10 năm 2014
Sửa đổi ngày 9 tháng 2 năm 2015
Chấp nhận ngày 20 tháng 3 năm 2015
Có sẵn trực tuyến ngày 27 tháng 3 năm 2015

từ khóa:
Phản ứng có hại của thuốc
diễn đàn y tế
SNOMED CT
MedDRA
văn bản chú thích
an toàn thuốc
Truyền thông xã hội
Khai thác thông tin
đánh giá của người tiêu dùng

trừu tượng

Tập dữ liệu về Tác dụng phụ của Thuốc CSIRO (CADEC) là một kho ngữ liệu mới có chú thích phong phú gồm các bài đăng trên diễn đàn y tế về các Biến cố Có hại của Thuốc do bệnh nhân báo cáo (ADE). Kho ngữ liệu được lấy từ các bài đăng trên mạng xã hội và chứa văn bản phần lớn được viết bằng ngôn ngữ thông tục và thường khác với các quy tắc chấm câu và ngữ pháp tiếng Anh chính thức. Các chú thích chứa đề cập đến các khái niệm như thuốc, tác dụng phụ, triệu chứng và bệnh liên quan đến các khái niệm tương ứng của chúng trong các từ vựng được kiểm soát, ví dụ: Thuật ngữ lâm sàng SNOMED và MedDRA. Chất lượng của các chú thích được đảm bảo bởi các hướng dẫn chú thích, chú thích nhiều giai đoạn, đo lường sự đồng thuận giữa các chú thích và đánh giá cuối cùng về các chú thích bởi một nhà thuật ngữ lâm sàng. Kho dữ liệu này rất hữu ích cho các nghiên cứu trong lĩnh vực khai thác thông tin, hay nói chung là khai thác văn bản, từ phương tiện truyền thông xã hội để phát hiện các phản ứng có hại của thuốc có thể xảy ra từ các báo cáo trực tiếp của bệnh nhân. Kho dữ liệu được cung cấp công khai tại https://data.csiro.au.[1]

2015 Elsevier Inc. Bảo lưu mọi quyền.

1. Giới thiệu

Do những hạn chế trong các thử nghiệm lâm sàng, không phải tất cả các tác dụng phụ tiềm ẩn của thuốc đều được phát hiện trước khi thuốc được tung ra thị trường [3]. Các phản ứng có hại của thuốc vẫn chưa được biết đến tạo ra mối quan tâm lớn đối với sức khỏe cộng đồng [7]. Họ phải chịu trách nhiệm cho hàng ngàn trường hợp tử vong hoặc thương tích nghiêm trọng, cũng như hàng triệu ca nhập viện. Do đó, chúng tiêu tốn hàng tỷ đô la cho các hệ thống chăm sóc sức khỏe trên toàn thế giới [1,14,27].

Giám sát sau khi đưa ra thị trường, còn được gọi là cảnh giác dược, đóng một vai trò quan trọng trong việc xác định những tác dụng phụ bất lợi của thuốc không được phát hiện trong khi thuốc có mặt trên thị trường [3,6]. Thực hành giám sát truyền thống, được thực thi bởi các cơ quan quản lý như Cục Quản lý Thực phẩm và Dược phẩm (FDA) ở Hoa Kỳ và Cơ quan Quản lý Sản phẩm Trị liệu (TGA) ở Úc, là thu thập các báo cáo tự nguyện về tác dụng phụ bất lợi của thuốc, điều tra các báo cáo và đưa ra các tín hiệu an toàn nếu một loại thuốc bị nghi ngờ gây ra tác dụng phụ.

Gần đây hơn, giám sát tích cực đã được nghiên cứu trong đó các nguồn dữ liệu khác nhau được tự động giám sát để báo cáo về các phản ứng bất lợi có thể xảy ra; Sáng kiến Sentinel của FDA là một ví dụ về giám sát tích cực. Một trong những nguồn thông tin có thể được theo dõi tích cực là các diễn đàn y tế nơi người tiêu dùng thảo luận về trải nghiệm trực tiếp của họ với thuốc.

Một kho văn bản được chú thích bởi con người là một nguồn tài nguyên quý giá trong việc phát triển và đánh giá các phương pháp khai thác văn bản dựa trên học máy. Những chú thích như vậy cần tính đến thuốc và tác dụng phụ, cũng như tình trạng bệnh nhân và dữ liệu nhân khẩu học.

Một kho văn bản như vậy có thể tốn kém để xây dựng nhưng một khi được tạo ra sẽ phục vụ nhiều nghiên cứu nhằm phát triển các thuật toán khai thác văn bản của họ. Do đó, chúng tôi đã phát triển một tập hợp các bài đăng trên diễn đàn y tế lấy từ AskaPatient[2] để thu thập xếp hạng và đánh giá về thuốc từ người tiêu dùng của họ. Những bài đăng này được chú thích cho các thực thể như tên của các loại thuốc được tiêu thụ và tác dụng phụ của chúng.

Ngoài ra, các thực thể được chú thích này được liên kết với các từ vựng được kiểm soát: Thuật ngữ lâm sàng về danh pháp y học được hệ thống hóa (SNOMED CT), AMT (Thuật ngữ về thuốc của Úc) và MedDRA (Từ điển y tế cho các hoạt động quản lý).

Theo hiểu biết của chúng tôi, kho văn bản của chúng tôi là kho văn bản đầu tiên được chú thích phong phú và có sẵn công khai về các bài đăng trên diễn đàn y tế có thể được áp dụng cho các nhiệm vụ khai thác văn bản liên quan đến cảnh giác dược.

2. Công việc liên quan

Chúng tôi xem xét các kho dữ liệu có liên quan hiện có và thông số kỹ thuật của chúng, cũng như tổng quan ngắn gọn về các phương pháp được đề xuất trong tài liệu để trích xuất thông tin về tác dụng phụ của thuốc từ phương tiện truyền thông xã hội.

Tác giả tương ứng.
Địa chỉ email: sarvnaz.karimi@csiro.au (S. Karimi).

[1] Dữ liệu chỉ có thể được sử dụng cho mục đích nghiên cứu, theo giấy phép dữ liệu CSIRO.

[2] http://www.askaopathy.com/.

http://dx.doi.org/10.1016/j.jbi.2015.03.010
1532-0464/ 2015 Elsevier Inc. Bảo lưu mọi quyền.

Machine Translated by Google

2.1. tập đoàn liên quan

Các tài liệu báo cáo về khối văn bản được chú thích để tạo điều kiện khai thác thông tin từ tài liệu y sinh, hồ sơ sức khỏe điện tử và dữ liệu văn bản khác. Dưới đây, chúng tôi giới thiệu một số kho ngữ liệu này cùng với thông số kỹ thuật của chúng, chẳng hạn như các thực thể được chú thích và nguồn dữ liệu của chúng, cũng như nêu rõ những khác biệt đối với kho ngữ liệu của chúng tôi. Chúng tôi chỉ xem xét những bộ dữ liệu chia sẻ một thực thể quan tâm với kho dữ liệu của chúng tôi hoặc có liên quan chặt chẽ đến việc khai thác văn bản trong lĩnh vực cảnh giác dược.

Leaman et al. [17] đã phát triển Tập dữ liệu bệnh Arizona (AZDC) trong đó tài liệu y sinh (bản tóm tắt MEDLINE) được chú thích để đề cập đến các bệnh. Các đề cập về bệnh sau đó được chuẩn hóa thành các khái niệm UMLS tương ứng của chúng. Họ cũng chứng minh rằng các đề cập đến bệnh tật có những đặc điểm tương tự như các thực thể như gen và protein. Đó là, chúng có các biến thể tên, một tên có thể đề cập đến các thực thể thuộc các loại ngữ nghĩa khác nhau và có xu hướng cấu trúc cú pháp phức tạp. Những đặc điểm này có thể dẫn đến sự mơ hồ trong việc tự động nhận ra các đề cập đến bệnh tật trong văn bản.

Vào năm 2014, một phần mở rộng của AZDC được gọi là kho ngữ liệu bệnh NCBI [9] đã được giới thiệu để chú thích các đề cập đến bệnh trong phần tóm tắt PubMed bằng cách sử dụng một công cụ dựa trên web có tên là PubTator [31]. Những đề cập này sau đó được liên kết với Tiêu đề chủ đề y tế (MeSH) hoặc thuật ngữ tiêu chuẩn Mendelian Inheritance in Man (OMIM) trực tuyến. Các chú thích được thực hiện bởi 14 người chú thích, mỗi người chú thích bốn đến năm lô gồm 30 bản tóm tắt với mỗi lô nhận được ít nhất hai chú thích. Các nhà chú thích đã được hướng dẫn chi tiết.

Roberts và cộng sự. [26] hồ sơ bệnh nhân ung thư đã qua đời được chú thích cho một số lượng lớn các thực thể như can thiệp, thuốc hoặc thiết bị và tình trạng (danh sách đầy đủ trong Bảng 1, hàng thứ hai). Dữ liệu của họ, được sử dụng trong một nhiệm vụ được chia sẻ, được tạo dựa trên các hướng dẫn được thiết kế cẩn thận và được hoàn thiện sau một tập hợp chú thích ban đầu. Họ đã thuê 25 người chú thích và yêu cầu mỗi tài liệu được chú thích hai lần để tính toán sự đồng thuận giữa các người chú thích. Định nghĩa của họ về các thực thể nói chung bao hàm và lấy cảm hứng từ các khái niệm UMLS.
Ví dụ, họ chú thích các loại thuốc và thiết bị y tế trong cùng một thực thể. Loại thực thể tình trạng của họ cũng bao gồm một loạt các khái niệm như triệu chứng, biến chứng và chấn thương.

Gurulingappa và cộng sự. [10] đã mô tả một tập hợp các bản tóm tắt MEDLINE có chú thích cho các bệnh ở người và các tác dụng phụ. Các phần tóm tắt được tìm thấy bằng cách truy vấn PubMed về ''bệnh HOẶC tác dụng phụ'' và sau đó một tập con gồm 400 phần tóm tắt được chọn ngẫu nhiên để chú thích. Hai người chú thích đã được hướng dẫn để xác định các đề cập đến bệnh tật và tác dụng phụ dựa trên bối cảnh của họ. Máu mủ cuối cùng chứa 813 đề cập đến tác dụng phụ và 1428 đề cập đến bệnh tật. Kho ngữ liệu được cung cấp công khai với các chú thích ở định dạng IOB (Bên trong, Bên ngoài, Bắt đầu).

Gurulingappa và cộng sự. [11] đã tạo một kho ngữ liệu để trích xuất thông tin liên quan đến an toàn thuốc từ các báo cáo ca bệnh trên MEDLINE. Tóm tắt MEDLINE được chọn ngẫu nhiên từ nhóm tóm tắt được trả về bằng cách truy vấn PubMed với các thuật ngữ MeSH (Tiêu đề chủ đề y tế) ''điều trị bằng thuốc'' và ''tác dụng phụ'' . Tập văn bản được chú thích bởi ba người chú thích cho các đề cập đến thuốc (tên thương hiệu, tên tầm thường, chữ viết tắt), tác dụng phụ, liều lượng (số lượng và tần suất), cũng như mối quan hệ giữa các thực thể này. Tác dụng phụ của một loại thuốc nhất định bao gồm một loạt các dấu hiệu, triệu chứng, bệnh tật, rối loạn, bất thường, tổn thương cơ quan và thậm chí tử vong do loại thuốc đó gây ra.

Trong các chú thích của họ, Gurulingappa et al. [11] loại trừ tên của các thiết bị y tế hoặc hóa chất bệnh viện. Ngoài ra, họ chỉ chú thích đề cập đến tên thuốc liên quan đến một sự kiện bất lợi. Phần cuối cùng của corpus bao gồm những câu từ phần tóm tắt có ít nhất một lần đề cập đến tác dụng phụ.

Van Mulligen và cộng sự. [30] báo cáo một kho tài liệu y sinh được chú thích có sẵn công khai trong đó các trường hợp về thuốc, rối loạn, gen và mối quan hệ giữa các thực thể được xác định là chú thích. Dữ liệu cho kho văn bản được thu thập bằng cách truy vấn PubMed. Các chiến lược tìm kiếm được liệt kê trong bài viết của họ. Ban đầu, họ chú thích kho văn bản bằng cách sử dụng Bộ nhận dạng thực thể được đặt tên (NER) và sau đó yêu cầu người chú thích của họ sửa các chú thích tự động.

Deleger và cộng sự. [8] đã tạo ra một tập hợp các hồ sơ lâm sàng có chú thích. Kho dữ liệu được tạo ra với hai mục đích khác nhau: (1) nhiệm vụ khử nhận dạng mà dữ liệu được chú thích cho thông tin sức khỏe cá nhân như tuổi của bệnh nhân hoặc địa chỉ email; và (2) chú thích về các thực thể liên quan đến thuốc, chẳng hạn như tên và loại thuốc, cũng như bệnh và triệu chứng. Các chú thích trong bước này dựa trên các khái niệm SNOMED CT và UMLS. Người chú thích được hướng dẫn để chú thích các thực thể nếu tồn tại một khái niệm SNOMED CT hoặc UMLS tương ứng. Deleger và cộng sự. [8] đã chú thích một tập hợp các ghi chú lâm sàng và nhãn thuốc của FDA bằng cách sử dụng hai công cụ chú thích. Một hướng dẫn đã được phát triển và cập nhật sau bước chú thích thử nghiệm. Họ đã tính toán sự thống nhất giữa các người chú thích để đảm bảo chỉ những tài liệu có sự nhất trí đầy đủ mới được đưa vào kho văn bản cuối cùng.

Một hướng công việc trong lĩnh vực cảnh giác dược là nghiên cứu tương tác thuốc-thuốc (DDIs) dẫn đến phản ứng có hại của thuốc. Herrero-Zazo et al. [13] đã tạo một kho văn bản hỗ trợ khai thác văn bản trong lĩnh vực này. Dữ liệu cho kho văn bản được lấy từ DrugBank [32] và phần tóm tắt của MEDLINE . Các chú thích tuân theo các hướng dẫn được viết đầy đủ cẩn thận bởi hai dược sĩ. Dữ liệu ban đầu được MetaMap chú thích để xác định các đề cập đến các thực thể y sinh và sau đó được chuyển cho người chú thích để tiếp tục quản lý và chỉnh sửa. Thỏa thuận giữa các chú thích được báo cáo riêng cho các thực thể và mối quan hệ. Kho văn bản này đã được sử dụng trong nhiệm vụ chia sẻ SemEval3 2013. Kho dữ liệu của chúng tôi khác vì nó không xem xét các tương tác giữa thuốc và thuốc mà thay vào đó tập trung vào các tác dụng phụ được báo cáo của một loại thuốc.

Chúng tôi tóm tắt kho ngữ liệu hiện có trong Bảng 1. Cột thứ hai xác định nguồn gốc của tập dữ liệu. Loại dữ liệu và kích thước kho văn bản được liệt kê trong cột thứ ba và cột thứ tư. Các thực thể được chú thích và mối quan hệ giữa các thực thể này được liệt kê trong cột cuối cùng của bảng. Lưu ý rằng thường thì cùng một loại thực thể, ví dụ như thuốc, khác nhau về định nghĩa giữa các tập hợp khác nhau.

Tất cả các kho ngữ liệu hiện có được tạo ra trên cơ sở tài liệu y sinh, ví dụ, từ tóm tắt MEDLINE . Kho ngữ liệu CADEC , mà chúng tôi giới thiệu thêm trong các phần sau, là kho dữ liệu duy nhất được lấy từ các báo cáo của người tiêu dùng trên mạng xã hội, giới thiệu các đặc điểm ngôn ngữ độc đáo và các thách thức xử lý.

2.2. Khai thác tác dụng phụ của thuốc từ phương tiện truyền thông xã hội

Khai thác tín hiệu phản ứng có hại của thuốc từ mạng xã hội đã được nghiên cứu từ năm 2010. Leaman et al. [18] khai thác nhận xét của bệnh nhân trên một diễn đàn y tế có tên là DailyStrength4 để tìm các đề cập đến tác dụng phụ của thuốc. Dữ liệu của họ được chú thích về tác dụng phụ, tác dụng có lợi, chỉ định và những thứ khác. Họ đã sử dụng một từ vựng kết hợp COSTART5 và một số tài nguyên khác để trích xuất thông tin về tác dụng phụ từ nhận xét của bệnh nhân bằng cách sử dụng phương pháp cửa sổ trượt.

Che và cộng sự. [4] áp dụng phân loại để xác định các loại thuốc có khả năng trở thành một phần của danh sách theo dõi của FDA. Họ đã sử dụng các bài đăng của bệnh nhân trên Yahoo! Các nhóm.

Benton et al. [2] đã trích xuất các tác dụng phụ tiềm ẩn từ một số diễn đàn ung thư vú khác nhau, chẳng hạn như breastcancer.org, sử dụng tần suất đếm các thuật ngữ trong từ vựng được kiểm soát. Họ

[3] SemEval (Đánh giá ngữ nghĩa) là một nhiệm vụ được chia sẻ để đánh giá ngữ nghĩa các hệ thống phân tích trên một tập dữ liệu được chia sẻ và các chỉ số đánh giá đã được thống nhất. www.dailysức mạnh.org.

[5] Các ký hiệu mã hóa cho từ điển đồng nghĩa về phản ứng có hại (COSTART) được FDA phát triển để mã hóa phản ứng có hại của thuốc trong các báo cáo sau khi đưa ra thị trường. Nó hiện được thay thế bằng MedDRA.

Machine Translated by Google

Bảng 1
Thông số kỹ thuật của kho dữ liệu liên quan hiện có và CADEC.

kho văn bản	Nguồn gốc	Kiểu	Kích cỡ	Thực thể/mối quan hệ
Leaman et al. [17]	MEDLINE	Văn học	749 tóm tắt (2784 câu) Bệnh tật	
Roberts và cộng sự. [26]	Hoàng gia Marsden Bệnh viện	Hồ sơ bệnh nhân ung thư 150 tài liệu (50 báo cáo lâm sàng, 50 báo cáo mô bệnh học, 50 báo cáo hình ảnh) 400 tóm tắt		Điều kiện, can thiệp, điều tra, kết quả, thuốc hoặc thiết bị, quỹ tích, tín hiệu phủ định, tín hiệu bên, tín hiệu vị trí phụ và các mối quan hệ
Gurulingappa và cộng sự. [10]	MEDLINE	Văn học		Dịch bệnh, ảnh hưởng xấu
Gurulingappa và cộng sự. [11]	MEDLINE	Báo cáo trường hợp y tế	2972 tóm tắt (4272 câu) Thuốc, tác dụng phụ, liều lượng, mối quan hệ giữa các thực thể này	
Deleger và cộng sự. [số 8]	Bệnh viện nhi Cincinnati, ClinicalTrials.gov, DailyMed	Nhãn thuốc của FDA, thông báo thử nghiệm lâm sàng, ghi chú lâm sàng	3503 ghi chú lâm sàng cho thông tin sức khỏe cá nhân, 1655 cho bệnh tật và rối loạn	Tên thuốc, loại thuốc, ngày, liều lượng, thời gian, dạng, tần suất, lộ trình, thay đổi trạng thái, độ mạnh, công cụ điều chỉnh, bệnh/rối loạn, dấu hiệu/triệu chứng và thông tin sức khỏe cá nhân (tuổi, ngày, email, v.v.)
Van Mulligen và cộng sự. [30]	MEDLINE	Văn học	300 tóm tắt	Thuốc, rối loạn, gen, mối quan hệ giữa các thực thể này
Herrero-Zazo et al. [13]	Cơ sở dữ liệu MEDLINE và DrugBank	Tài liệu và thông tin thuốc	233 tóm tắt MEDLINE và 792 văn bản từ DrugBank	Dược chất (tên chung của thuốc, nhãn hiệu, nhóm thuốc và hoạt chất không được phép sử dụng cho người), bốn loại mối quan hệ DDI
Dogan et al. [9]	tóm tắt PubMed	Văn học	793 tóm tắt (6881 câu) Đề cập đến bệnh tật	
Tập đoàn của chúng tôi (CADEC)	HỏiBệnh nhân	diễn đàn y tế	1253 bài viết (7398 câu)	Thuốc, tác dụng phụ, bệnh, triệu chứng, phát hiện

sau đó sử dụng khai thác quy tắc kết hợp để thiết lập mối quan hệ giữa các điều khoản phù hợp. Khai thác quy tắc kết hợp là một phương pháp khai thác dữ liệu phổ biến để khai thác các tác động bất lợi từ cơ sở dữ liệu quản lý và quy định. Dương và cộng sự. [33] đã nghiên cứu phát hiện tín hiệu từ một diễn đàn y tế có tên là MedHelp6 bằng cách mở rộng các thuật toán khai thác quy tắc kết hợp hiện có bằng cách thêm các số liệu về mức độ thú vị và ấn tượng. Để tìm các đề cập đến tác dụng phụ của thuốc trong văn bản, họ đã sử dụng cửa sổ trượt và từ vựng do người tiêu dùng kiểm soát để khớp với các thuật ngữ.

Liu và Chen [21] đã triển khai một hệ thống có tên là AZDrugMiner. Dữ liệu được thu thập bằng trình thu thập thông tin từ Cộng đồng trực tuyến về bệnh tiểu đường.7 Để tìm các đề cập đến tác dụng phụ và thông tin liên quan hoặc các mối quan hệ như tác dụng phụ của thuốc, trước tiên họ sử dụng MetaMap để ánh xạ văn bản tới các khái niệm UMLS, sau đó trích xuất các mối quan hệ bằng cách sử dụng đồng thời Phân tích.

Xem xét đầy đủ các kỹ thuật này có thể được tìm thấy trong [16]. Tất cả các nghiên cứu này được đánh giá trên các tập đoàn tư nhân khác nhau, điều này khiến việc so sánh hiệu quả của chúng trở nên khó khăn. Việc cung cấp một kho dữ liệu công khai như CADEC sẽ tạo điều kiện thuận lợi cho việc so sánh và đánh giá các kỹ thuật khai thác này.

3. Vật liệu Corpus

Dữ liệu cho kho văn bản CADEC được lấy từ một diễn đàn y tế có tên là AskaPatient, diễn đàn dành riêng cho các đánh giá của người tiêu dùng về thuốc. Bệnh nhân có thể đánh giá thuốc bằng cách điền vào biểu mẫu chi tiết về một loại thuốc cụ thể dựa trên tên thương hiệu của chúng, ví dụ Tamiflu. Biểu mẫu này yêu cầu tỷ lệ hài lòng, lý do dùng thuốc, liều lượng và tần suất, thời gian dùng thuốc, tác dụng phụ gặp phải ở dạng văn bản tự do, nhận xét ở dạng văn bản tự do, cũng như nhân khẩu học của bệnh nhân bao gồm tuổi và giới tính. Không phải tất cả các thông tin trong biểu mẫu đánh giá là bắt buộc. Ngoài ra, có một chính sách đánh giá cho mỗi loại thuốc trong trang web nơi người tiêu dùng được yêu cầu không nhập nhiều đánh giá trừ khi họ muốn cập nhật bài đăng trước đó của mình.

AskaPatient đã cung cấp cho chúng tôi các bài đăng của người tiêu dùng về 12 loại thuốc sau: Voltaren (Diclofenac Sodium), Cataflam (Diclofenac Kali), Voltaren-XR (Diclofenac Natri), Arthrotec (Diclofenac Natri; Misoprostol), Pennsaid (Diclofenac Natri), Solaraze (Diclofenac Natri), Flector (Diclofenac Epolamine), Cambia (Diclofenac Kali), Zipsor (Diclofenac Kali), Diclofenac Natri, Diclofenac Kali và Lipitor (Atorvastatin Canxi). Chúng tôi chia các loại thuốc này thành hai loại: Diclofenac, bao gồm những loại thuốc có hoạt chất Diclofenac, và Lipitor. Theo biểu mẫu xếp hạng được giải thích ở trên, các bài đăng này chứa thông tin nhân khẩu học của bệnh nhân, đánh giá mức độ hài lòng về thuốc từ 1 (thấp) đến 5 (cao), lý do dùng thuốc, cách dùng thuốc, nhận xét của bệnh nhân về hiệu quả của thuốc. thuốc và nếu có bất kỳ tác dụng phụ nào xảy ra.

Một bài đăng mẫu cho Voltaren được hiển thị trong Bảng 2. Trong CADEC, chúng tôi chỉ chú thích và cung cấp các phần văn bản miễn phí của mỗi bài đăng. Ngôn ngữ của tất cả các bài đăng này là tiếng Anh. Hầu hết các bài đăng được viết bằng ngôn ngữ thông tục và không tuân theo các quy tắc chấm câu và ngữ pháp tiếng Anh chính thức. Họ chủ yếu báo cáo kinh nghiệm cá nhân của bệnh nhân, tuy nhiên, đôi khi điều kiện của một thành viên gia đình đã được báo cáo.

Thống kê về các bài đăng được sử dụng để tạo kho dữ liệu được trình bày trong Bảng 3. Số liệu thống kê được liệt kê cho toàn bộ kho dữ liệu, cũng như từng danh mục thuốc (Diclofenac và Lipitor) riêng biệt. Số lượng bài viết cho dữ liệu ban đầu là 1321, tuy nhiên, 71 bài viết không chứa bất kỳ văn bản nào. Những bài đăng này đã bị loại khỏi kho văn bản cuối cùng.
Độ dài của bài viết và kích thước trung bình của chúng trong câu và từ cũng được báo cáo trong Bảng 3, hàng từ bốn đến bảy. Lipitor có số lượng bài đăng cao hơn đáng kể trong một khoảng thời gian tương tự so với Diclofenac với mỗi bài đăng trung bình dài hơn. Giới tính của người tiêu dùng báo cáo gần như được chia đều giữa nam và nữ, với 42 bài đăng thiếu thông tin giới tính. Độ tuổi của bệnh nhân là từ 17 đến 84 tuổi, với độ tuổi trung bình là 52.

4. Chú thích

Chúng tôi đã chú thích kho văn bản theo hai giai đoạn chính: (1) nhận dạng thực thể và (2) liên kết thuật ngữ, còn được gọi là thiết lập thông thường, để liên kết các thực thể đã xác định với các từ vựng được kiểm soát.

Dưới đây chúng tôi giải thích các nguyên tắc chú thích, các công cụ được sử dụng trong quá trình chú thích và quy trình chú thích.

4.1. hướng dẫn

Nguyên tắc chú thích là một phương pháp kiểm soát chất lượng của việc xây dựng văn bản, với một số gợi ý được cung cấp trong

6 www.medhelp.org/.
7 http://community.diabetes.org.

Machine Translated by Google

Bảng 2
Một bài viết mẫu về Voltaren trên AskaPatient.com.

Lý do xếp hạng		Phản ứng phụ	Bình luận	Giới tính Tuổi Thời lượng/liều lượng	Ngày thêm vào
4	Viêm xương khớp hông	Nó giúp giảm đau mãn tính nhưng theo thời gian, gây đau và chảy máu đường ruột. Tôi có các triệu chứng tương tự như viêm túi thừa: có máu trong phân, đau	Tôi sẽ thận trọng khi chú ý đến chuột rút, đau ruột/dạ dày có thể dẫn đến các tình trạng rất nghiêm trọng	M 63 1,5 năm 100MG ER 1X D	4/12/ 2013

bàn số 3
Thống kê về dữ liệu được sử dụng trong CADEC.

	kho văn bản	diclofenac	lipit
Số bài viết	1321	264	1057
Số bài viết có văn bản	1250	250	1000
Số câu 7632		1263	6369
Trung bình độ dài bài đăng (câu)	6 56		
Số từ Trung bình. độ dài bài đăng (từ)	101,486 81	16,778 67	84,708 85
khoảng thời gian	Tháng 1 năm 2001– Tháng 9 năm	Tháng 2 năm 2002–tháng 8 năm	Tháng 1 năm 2001– Tháng 9 năm
Giới tính	2013 F 662 (50,1%) M 617 (49,9%)	2013 F 181 (68,6%) M 76 (28,8%)	2013 F 481 (45,6%) M 541 (51,2%)
Độ tuổi	17–84	17–78	19–84
Trung bình tuổi	52	47	54

tài liệu về cách đưa ra hướng dẫn cho người chú thích [5,12,19].
Chúng tôi đã điều chỉnh một số gợi ý của Zazo et al. [12] trong việc tạo ra các hướng dẫn. Chú thích đã được thực hiện ở cấp độ câu. Không có thực thể nào kéo dài qua các câu được chú thích. Chú thích có thể không liên tục trong cùng một câu. Các thực thể trùng lặp trong một câu được chú thích độc lập, nghĩa là tất cả các lần xuất hiện của cùng một thực thể đều được chú thích. Các đề cập chung về một thực thể không được chú thích, chẳng hạn như thuật ngữ tác dụng phụ. Các thực thể nhúng không được chú thích riêng.

Nghĩa là, nếu một phần của thực thể chứa một thực thể khác, thì chỉ có thực thể chính được chú thích. Ví dụ: nếu bài đăng đề cập đến đau cơ như một tác dụng phụ bất lợi, thì cơn đau không được chú thích riêng dưới dạng một thực thể khác. Tham chiếu đồng tham chiếu/tương tự không được chú thích. Giới từ hàng đầu, từ hạn định hoặc tính từ sở hữu đã bị loại trừ để thúc đẩy các nhịp nhất quán hơn. Ví dụ, viêm khớp đã được chú thích thay vì viêm khớp của tôi.

Cụ thể, đối với giai đoạn đầu tiên của chú thích, chúng tôi đã xác định các thực thể quan tâm như bên dưới. Các thực thể này và định nghĩa của chúng là các sửa đổi của các thực thể được đề xuất bởi Karimi et al. [15].

Thuốc Đề cập đến tên của một loại thuốc hoặc thuốc được chú thích với thuốc. Các nhóm thuốc, chẳng hạn như thuốc chống viêm không steroid (NSAIDS), đã bị loại trừ.
Các thiết bị y tế cũng bị loại trừ. Ví dụ, trong câu ''Tôi phải nghiện Diclofenac'', ''Diclofenac'' được chú thích là Thuốc.

ADR Các đề cập đến các phản ứng có hại của thuốc mà theo văn bản, rõ ràng có liên quan đến một loại thuốc được chú thích bằng nhãn ADR. Ví dụ, trong câu ''Đôi khi gây buồn ngủ.'', ''buồn ngủ'' là tác dụng phụ. Tất cả ngữ cảnh cần thiết cho khái niệm ADR đều được chú thích. Ví dụ: nếu câu nói ''Tôi bị đau dạ dày cấp tính...'', thì ''đau dạ dày cấp tính'' được chú thích không chỉ là ''đau dạ dày'', trong khi trong một câu như ''Tôi cảm thấy trống rỗng như một tờ giấy trắng'', chỉ chú thích ''cảm thấy trống rỗng''.

Bệnh Thực thể này chỉ định lý do dùng thuốc.
Bệnh nhân có thể đề cập đến tên của một bệnh mà họ dùng thuốc. Nếu đó là một tên bệnh cụ thể, nó được gắn thẻ là Bệnh. Ví dụ, trong câu ''...sau 3 năm sử dụng Ativan để kiểm soát sự lo lắng & hung hăng...'', cả ''lo lắng'' và ''hung hăng'' đều được chú thích (riêng) là Bệnh .

Triệu chứng Thực thể này chỉ định lý do dùng thuốc.
Bệnh nhân có thể đề cập đến các triệu chứng của một căn bệnh khiến họ phải dùng thuốc. Ví dụ, trong câu ''Tim tôi đập nhanh và...'', triệu chứng ''tim đập nhanh'' được nhấn mạnh.

Phát hiện Phát hiện lâm sàng là bất kỳ tác dụng phụ, bệnh hoặc triệu chứng bất lợi nào mà bệnh nhân báo cáo không trực tiếp gặp phải hoặc bất kỳ khái niệm lâm sàng nào khác có thể thuộc bất kỳ loại nào trong số này nhưng người chú thích không rõ nó thuộc loại nào.

Các định nghĩa này đã được hoàn thiện sau khi nhận được phản hồi từ các nhà chú thích và chuyên gia trong lĩnh vực này, bao gồm cả dược sĩ. Một tác vụ chú thích thử nghiệm đã được thiết lập để chú thích một số lượng nhỏ bài đăng và cài đặt chú thích đã được sửa đổi dựa trên phản hồi.

Đối với giai đoạn thứ hai của chú thích, liên kết thuật ngữ, các hướng dẫn sau đây đã được sử dụng. Chi tiết được giải thích trong Phần 4.3.

SNOMED CT (SCT) Bất kỳ khoảng văn bản nào được chú thích bằng bất kỳ thẻ nào không phải là Thuốc phải được ánh xạ tới khái niệm SNOMED CT tương ứng từ hệ thống phân cấp Tìm kiếm lâm sàng. Nếu không có khái niệm nào tồn tại thì hãy gán thẻ concept_less.

AMT Bất kỳ khoảng văn bản nào được chú thích bằng Thuốc phải được ánh xạ tới khái niệm AMT tương ứng.
Nếu không tồn tại khái niệm phù hợp thì hãy gán thẻ concept_less.

MedDRA Bất kỳ khoảng văn bản nào được chú thích bằng ADR phải được ánh xạ tới thuật ngữ MedDRA tương ứng.

4.2. nhận dạng thực thể

Giai đoạn đầu tiên của quy trình chú thích là xác định các đề cập đến các thực thể được quan tâm, ví dụ: các phản ứng bất lợi, trong các bài đăng trên diễn đàn. Chúng tôi đã sử dụng Brat—một công cụ chú thích văn bản dựa trên web được phát triển tại Đại học Tokyo [28]—để thiết lập các chú thích cho giai đoạn này.

Các bài đăng trên diễn đàn về các loại thuốc trong danh mục Diclofenac đã được bốn sinh viên y khoa chú thích. 22% bài đăng của Lipitor được chú thích bởi các sinh viên y khoa và phần còn lại bởi hai nhà khoa học máy tính. Ba trong số các tác giả đã sàng lọc các chú thích và sửa những lỗi rõ ràng. Ví dụ: nếu người chú thích đã bỏ lỡ các phần của từ (pai thay vì đau) trong quá trình chú thích, chúng tôi đã sửa khoảng cách. Tất cả các chú thích này đã được xem xét thêm bởi một nhà thuật ngữ lâm sàng trong giai đoạn chuẩn hóa.

Hình 1 cho thấy các ví dụ về chú thích sử dụng Brat. Những ví dụ này cho thấy sự đa dạng về ngôn ngữ trong cách các bệnh nhân khác nhau thể hiện tình trạng của họ. Trong ví dụ (a), Trazodone bị viết sai thành

Machine Translated by Google

Trazadone của bệnh nhân. Ngoài ra còn có hai nhãn thuốc tuyến giáp và testostrone (viết sai thành testonrone) không phải là tên thuốc nên không được chú thích. Những trường hợp như vậy đã được khắc phục trong các đánh giá của các thẻ. Ví dụ (b) chứa một thuật ngữ thông tục charley horse để chỉ những cơn co thắt hoặc chuột rút gây đau đớn. Đây cũng là một trường hợp chú thích phức tạp khi có nhiều thẻ không liên tục có chung một thuật ngữ. Ví dụ (c) cho thấy một bài đăng được viết bằng chữ in hoa, trong đó danh sách các sự kiện bất lợi được đưa ra mà không có bất kỳ dấu chấm câu nào ở giữa. Những ví dụ này nhấn mạnh sự cần thiết của việc phát triển các kỹ thuật xử lý ngôn ngữ vừa có khả năng xử lý văn bản thông tục và bất thường, vừa xử lý các khái niệm y tế trong dữ liệu đó.

Các bài đăng trên diễn đàn được chia đều giữa các chú thích, ngoại trừ 55 tài liệu được trao cho tất cả các chú thích nhằm mục đích tính toán thỏa thuận giữa các chú thích. Hai số liệu được sử dụng cho tính toán này: thỏa thuận nghiêm ngặt và thỏa thuận thoải mái, như được mô tả bởi Metke-Jimenez et al. [22]. Cả hai chỉ số này đều dựa trên mức trung bình của thỏa thuận theo cặp giữa các người chú thích như

$$bÞ \text{ thỏa thuậnði; } jÞ \frac{\text{phù hợpðAi; Aj; Một;}}{¼\text{maxðnAi ; nAj Þ}};$$

trong đó Ai đại diện cho tập hợp các chú thích của chú thích i; Aj repre gửi tập hợp các chú thích bởi chú thích j; nAi là kích thước của tập Ai và nAj là kích thước của tập Aj. phù hợpðAi; Aj; Một; bÞ là hàm đếm số thẻ trùng khớp. Hàm khớp có hai tham số nhị phân: độ nghiêm ngặt của nhịp a và độ nghiêm ngặt của thẻ b. Cả hai tham số này có thể nghiêm ngặt hoặc thoải mái. Nếu đối sánh nhịp là nghiêm ngặt, thì các chú thích được so sánh phải khớp chính xác. Hãy xem xét câu ''Tôi bị căng cơ tăng lên''. Nếu một người chú thích đoạn văn bản ''độ căng cơ'' và một người chú thích khác đoạn văn bản ''tăng độ căng cơ'' thì chức năng khớp với khớp nhịp nghiêm ngặt sẽ không trả về kết quả khớp. Nếu đối sánh nhịp được định cấu hình để nới lỏng, thì các chú thích trùng lặp sẽ được tính là đối sánh, với hạn chế là mỗi chú thích chỉ có thể được đối sánh với một chú thích khác. Đối với tính nghiêm ngặt của thẻ, nếu cả hai người chú thích chú thích cùng một đoạn văn bản, ví dụ: ''cơ căng'', nhưng một trong số họ sử dụng thẻ ADR và người kia sử dụng thẻ Triệu chứng, thì hàm sẽ chỉ trả về kết quả khớp hợp lệ nếu độ nghiêm ngặt của thẻ được nới lỏng.

Bảng 4 cho thấy thỏa thuận giữa các chú thích sử dụng các cấu hình khác nhau của chỉ số thỏa thuận. Khi cả cài đặt nhịp và chú thích đều được nới lỏng, mức độ đồng ý trung bình đối với Diclofenac là khoảng 78% và đối với Lipitor là 95%. Lưu ý rằng các thỏa thuận dành cho bốn người chú thích cho Diclofenac và hai người cho Lipitor.
Do đó, chúng tôi không thể so sánh trực tiếp các thỏa thuận này giữa hai loại thuốc.

4.3. Hiệp hội thuật ngữ

Mặc dù bản thân việc chú thích các thực thể trong kho văn bản đã có giá trị, nhưng việc liên kết các thực thể này với các thuật ngữ tiêu chuẩn cung cấp một mức độ thông tin khác trong kho văn bản. Quá trình như vậy được các nhà nghiên cứu khác gọi là chuẩn hóa, ví dụ, Pradhan et al. [25]. Một ví dụ về ánh xạ như vậy trong kho ngữ liệu được xem xét trong Phần 2 là Arizona Disease Corpus [17] trong đó tên bệnh được ánh xạ tới các khái niệm UMLS tương ứng của chúng.

Để chuẩn hóa các thực thể CADEC , một nhà thuật ngữ lâm sàng đã xem xét các thực thể được xác định trong giai đoạn trước để ánh xạ chúng tới khái niệm đại diện của chúng trong SNOMED CT, AMT và MedDRA.
Trong quá trình này, các thực thể có thể đã được chú thích sai, đã được sửa chữa.

4.3.1. Liên kết với SNOMED CT

SNOMED CT (Systematized Nomenclature of Medicine–Clinical Terms) là một thuật ngữ lâm sàng cung cấp các mã, từ đồng nghĩa và định nghĩa của các thuật ngữ lâm sàng và có thể được truy cập thông qua UMLS Metathesaurus.

Lợi ích chính của việc sử dụng từ vựng chuẩn để bình thường hóa thuật ngữ được sử dụng trong các diễn đàn là thu hẹp khoảng cách giữa ngôn ngữ của người bình thường và chuyên gia y tế. Người ta đã lập luận rằng việc mã hóa các tài liệu lâm sàng, chẳng hạn như hồ sơ lâm sàng, với SNOMED mang lại lợi ích cho việc thu thập dữ liệu thống kê bằng cách cung cấp các thuật ngữ tiêu chuẩn, chính thức, rõ ràng mô tả thông tin quan trọng về mặt lâm sàng [29] . Do đó, chúng tôi đã chọn SNOMED CT làm từ vựng được kiểm soát mục tiêu để ánh xạ các thực thể.

Các thực thể được phân loại thành các loại sau: ADR, Thuốc, Bệnh, Phát hiện và Triệu chứng. Tất cả các thực thể từ mỗi danh mục ngoại trừ Thuốc được ánh xạ tới SNOMED CT-AU (SCT-AU) v20140531 bởi một nhà thuật ngữ lâm sàng sử dụng công cụ CSIRO Snapper [23] . Hầu hết các thực thể được ánh xạ theo cách một đối một; tuy nhiên, cách tiếp cận ánh xạ một đến nhiều được thực hiện trong các tình huống mà thực thể có thể được mô tả tốt nhất bằng cách sử dụng nhiều hơn một khái niệm SCT. Ví dụ: thực thể ''đau lưng vài lần'' được ánh xạ tới các khái niệm SCT-AU 76948002 (Đau dữ dội) và 161891005 (Đau lưng). Bảng 5 liệt kê một số ví dụ từ các thực thể CADEC (cột cuối cùng, Thực thể gốc) được ánh xạ tới một hoặc nhiều khái niệm trong SNOMED CT.

4.3.2. Liên kết với AMT

Thuật ngữ Thuốc Úc (AMT), được phát triển và duy trì bởi NeHTA Australia, là một thuật ngữ được thiết kế để mô tả và xác định rõ ràng các loại thuốc có sẵn trong hệ thống chăm sóc sức khỏe của Úc. Mục đích sử dụng của nó cụ thể là trong các ứng dụng phần mềm được sử dụng trong môi trường chăm sóc sức khỏe của Úc.
Chỉ những sản phẩm đã đăng ký với TGA cho mục đích điều trị bệnh nhân mới được xác định trong AMT [24].

Các thực thể từ danh mục Thuốc được ánh xạ tới AMT V2.56 theo cách một đối một. Phần lớn, các loại thuốc được ánh xạ tới sản phẩm thương mại vì hầu hết chúng được mô tả trong các mục nhập văn bản theo tên thương mại của chúng. Tuy nhiên, do tính chất quốc tế của bộ sưu tập, một số loại thuốc có sẵn ở các quốc gia khác không có sẵn ở Úc. Ví dụ: Advicor là sự kết hợp của Niacin và Lovastatin không có trong AMT ở phiên bản thương mại hoặc sản phẩm chung, do đó được gán giá trị concept_less. Nếu không thể tìm thấy thuốc trong AMT sử dụng tên thương mại, thì dạng chung của thuốc sẽ được tìm kiếm và nếu có mặt được sử dụng làm bản đồ tương đương. Ví dụ: thuốc Aciphex được ánh xạ tới dạng chung 21296011000036107 (Rabeprazole) hoặc thuốc Cataflam được ánh xạ tới 21288011000036105 (Diclofenac). AMT không được cấu trúc để bao gồm các khái niệm cho các nhóm thuốc nên các thực thể được mô tả là thuốc kháng sinh hoặc statin, chẳng hạn, được gán giá trị concept_less. Một số loại thuốc quá mơ hồ để gán khái niệm từ AMT, chẳng hạn như benadryl; AMT chứa nhiều phiên bản benadryl nên không thể chỉ định ánh xạ một cách dứt khoát.

Bảng 5 cho thấy các ánh xạ được gán cho Cataflam và Demerol.

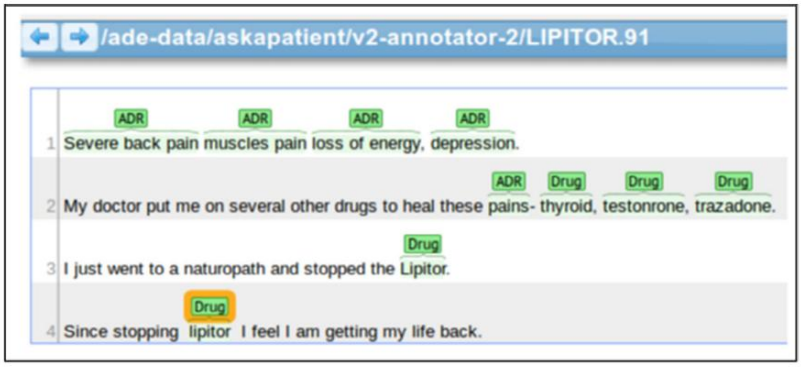
4.3.3. Liên kết với MedDRA

MedDRA8 (Từ điển Y tế về Hoạt động Điều tiết) là từ điển đồng nghĩa tiêu chuẩn được ngành dược phẩm và các cơ quan quản lý như FDA sử dụng. Nó chứa từ vựng được sử dụng cho các phản ứng có hại của thuốc được cấu trúc theo thứ bậc. Thuật ngữ Cấp độ Thấp nhất (LLT) là thuật ngữ cụ thể nhất thể hiện tình trạng của một cá nhân, chẳng hạn như ''cảm thấy buồn nôn''. Tăng một cấp là Điều khoản ưu tiên (PT), cùng với LLT thường được sử dụng trong

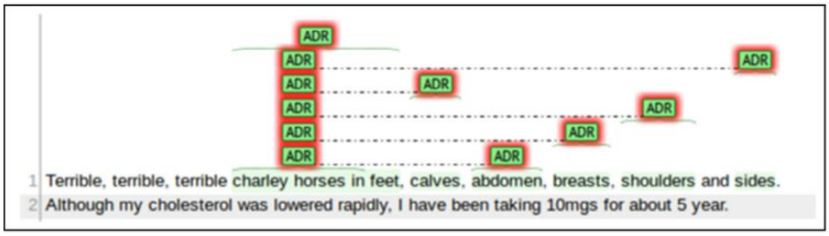
[8] http://www.meddra.org/.

Machine Translated by Google

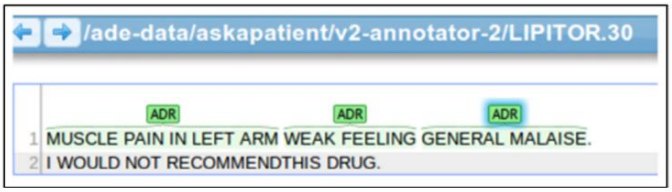
(a) Thuốc Trazodone sai chính tả



(b) Nhiều thẻ không liên tục chồng lên nhau



(c) Tất cả các chữ in hoa và không sử dụng dấu chấm câu



Hình 1. Ví dụ về chú thích thực thể. (a) Một ví dụ về lỗi chính tả thuốc; (b) Một ví dụ về các chú thích phức tạp; và (c) Một ví dụ về văn bản bất quy tắc.

Bảng
kích thước đầy đủ

kéo dài một	Thẻ b	Hiệp định	
		diclofenac	lipit
Nghiêm ngặt	Nghiêm ngặt	46,6	74.2
Nghiêm ngặt	thư giãn	49.1	81,6
thư giãn	Nghiêm ngặt	68,7	85.1
thư giãn	thư giãn	77,9	94,8

Bảng 5
Các ví dụ chuẩn hóa SNOMED CT và AMT.

mã số thuế TTDB	Ý tưởng	Thực thể ban đầu
76948002	đau dữ dội	Đau dữ dội ở cả hai vai
45326000	Đau vai	
284140004	Không thể di chuyển cánh tay	Không thể với tay lên đỉnh đầu
70733008	giới hạn của thể cử động khớp để nắm tay	Chuyển động bị hạn chế và không
76948002	đau dữ dội	đau đớn cùng cực
271599002	Cảm thấy hài lòng	
68962001	đau cơ	Đau cơ
mã AMT		
21288011000036105	Diclofenac	Cataflam
34839011000036106	Pethidine	Demerol
concept_less	-	Cảm thấy già hơn nhiều so với tôi

báo cáo về các sự kiện bất lợi của thuốc bởi cả các công ty dược phẩm và cơ quan quản lý. Các cấp cao hơn trong hệ thống phân cấp MedDRA chung chung hơn.

Chúng tôi đã sử dụng MedDRA V16.0 để chú thích các khái niệm được xác định trong giai đoạn ánh xạ SNOMED CT. Những khái niệm này được chú thích ở cấp độ LLT, để nắm bắt các thuật ngữ cụ thể thể hiện tình trạng của bệnh nhân. Bảng 6 cho thấy các ví dụ về các phiên bản bình thường của MedDRA từ kho văn bản CADEC . Cột cuối cùng, Thực thể gốc, là thực thể được xác định trong giai đoạn đầu tiên của chú thích và cột đầu tiên là định danh khái niệm MedDRA tương ứng. Chúng tôi liệt kê các thông tin khác có thể được truy xuất từ MedDRA được cung cấp ID MedDRA liên quan đến thực thể đó trong cột thứ hai và thứ ba.
Hàng thứ hai đến hàng cuối cùng là một ví dụ về một thực thể được ánh xạ tới hai ID MedDRA.

Nếu một thực thể được ánh xạ tới nhiều hơn một khái niệm SCT, thì tất cả các khái niệm SCT sau đó được ánh xạ tới MedDRA. Ví dụ: ''đau lưng nặng'' được ánh xạ tới hai khái niệm SCT Đau dữ dội và Đau lưng, những khái niệm này sau đó được ánh xạ tới các khái niệm MedDRA LLT 10003993 (Đau lưng) và 10033371 (Đau).
Tuy nhiên, cần lưu ý rằng trong trường hợp cụ thể này do tính chất phân loại của MedDRA, bản đồ thứ hai là không cần thiết vì nó không bổ sung thêm định nghĩa.

4.3.4. Thách thức bình thường hóa

Mặc dù việc ánh xạ các mục nhập văn bản miễn phí như những mục này chắc chắn mang lại một lớp bổ sung cho ngân hàng kiến thức, nhưng đôi khi nó có thể mang tính chủ quan. Nỗi đau là một ví dụ điển hình về tính chủ quan như vậy. Một thực thể chẳng hạn như ''đau đến mức tôi nghĩ mình sắp chết'', có thể được mã hóa thành cơn đau dữ dội hoặc cơn đau dữ dội. xét về

Machine Translated by Google

Bảng 6
Các ví dụ về chuẩn hóa MedDRA.

ID MedDRA ưu tiên	tên	Được phân loại là	Thực thể ban đầu
10040617	Đau vai	Đau cơ xương khớp	Đau dữ dội ở cả hai vai
10033407	đau đớn	Đau bụng trên	Đói bụng
10038742	chán bồn chồn	Hội chứng chân tay bồn chồn	Đôi chân bồn chồn
10043890	mệt mỏi	Mệt mỏi	Rất mệt mỏi
10069830	Không có khả năng ăn	mất ngôn ngữ	Không thể ăn hoặc uống
10069830	Không có khả năng ăn	mất ngôn ngữ	Không thể ăn uống bình thường
10013781	Khô miệng	9 hạng, ví dụ, Khô miệng 10003068	Khô miệng
Aptyalism	4 hạng, ví dụ, Asialia	concept_less –	Phổi cảm thấy nặng nề

gán một khái niệm MedDRA, ví dụ cụ thể này về tính chủ quan không phải là vấn đề vì MedDRA về bản chất là phân loại, chỉ quan tâm đến nỗi đau. Các ví dụ khác như ''Tôi không thể ra khỏi giường'' được gán giá trị không_khái niệm do không chắc chắn về điều mà người đó đang cố gắng truyền đạt. Không rõ ý của họ là ''Tôi mệt quá tôi không thể ra khỏi giường'' hay ''Tôi không muốn ra khỏi giường'' hay ''Tôi không thể ra khỏi giường được' do những khiếm khuyết về thể chất''. Khái niệm thuế TTĐB đề cập đến việc ''ra khỏi giường'' được sử dụng để mô tả khả năng thể chất của một người để ''ra/xuống giường'', tức là nếu những người này có thể tự ngồi dậy, đu đưa chân của họ sang một bên và có được một vị trí đứng. Tuy nhiên, MedDRA không có thuật ngữ ''không thể ra khỏi giường''.

Có một số thực thể đã được phân tích cú pháp khiến chúng bỏ sót ngữ cảnh tinh tế. Ví dụ: ''Mất trí nhớ/khả năng tập trung'' được phân tích thành ''mất trí nhớ'' và ''khả năng tập trung''. Tuy nhiên, nhiều khả năng mục này nhằm truyền đạt ''mất trí nhớ'' và ''mất khả năng tập trung'', do đó mục sau được ánh xạ thành một khái niệm tương đương với ''mất khả năng tập trung''.

5. Thống kê kho văn bản

Trong giai đoạn đầu tiên của chú thích, 64 bài đăng (5,1%) không nhận được bất kỳ chú thích thực thể nào và do đó không yêu cầu cũng như không làm sai. Sau giai đoạn chuẩn hóa, thêm 78 bài đăng (tổng cộng 142 bài đăng hoặc 11,4%) không nhận được chú thích MedDRA. Một ví dụ về bài đăng không nhận được bất kỳ chú thích thực thể nào là bài đăng trên Lipitor với nội dung: Tôi không gặp bất kỳ tác dụng phụ nào trong số vô số tác dụng phụ có thể xảy ra của thuốc.

Chúng tôi đã tạo một kho văn bản hợp nhất, trong đó đối với phần kho văn bản nhận được nhiều chú thích, chỉ một bộ dựa trên sự lựa chọn ngẫu nhiên của một người chú thích, được sử dụng để tính toán số liệu thống kê. Một giải pháp tốt hơn cho một kho văn bản được chú thích bởi nhiều người chú thích sẽ theo phương pháp centroid của Lewin et al. [20].

Bảng 7 liệt kê tần suất xuất hiện của các thực thể được chú thích trong toàn bộ kho dữ liệu cũng như trong từng loại thuốc (Lipitor và Diclofenac) riêng biệt. Tổng cộng có 9111 thực thể đã được xác định. ADR bao gồm 69,3% tổng số thực thể, tiếp theo là thuốc (1800 hoặc 19,8%). Từ tất cả 9111 thực thể được chú thích, chỉ có 39,4% (3591 thực thể) là duy nhất; mọi người thường báo cáo phản ứng tương tự.

Số lượng thực thể được chú thích là Triệu chứng đối với Diclofenac lớn hơn so với Lipitor (239 so với 38) mặc dù số lượng bài viết về Diclofenac nhỏ hơn nhiều. Lý do là bệnh nhân thường đề cập đến cơn đau như là triệu chứng của họ hoặc lý do dùng thuốc, trong khi đối với Lipitor/Atorvastatin thì

Bảng 7
Số lượng thực thể được chú thích trong toàn bộ kho văn bản và từng loại thuốc, cũng như số lượng thực thể duy nhất cho mỗi loại.

thực thể	kho văn bản		diclofenac		lipit	
	Tất cả	Độc nhất	Tất cả	Độc nhất	Tất cả	Độc nhất
ADR	6318	2713	888	508	5445	2316
Bệnh	283	162	59	42	226	129
Thuốc	1800	321	246	113	1542	222
triệu chứng	275	130	239	103	38	30
Phát hiện	435	265	41	34	394	238
Tất cả	9111	3591	1473	800	7645	2935

Bảng 8
Số lượng thẻ liên tục và không liên tục của mỗi loại.

	ADRs	Bệnh	Triệu chứng	Phát hiện	Thuốc	Tất cả
Tiếp diễn	5318	280	255	397	1797	8047
Không liên tục (chồng chéo)	918	2	13	34	2	969
Không liên tục (không chồng chéo)	82	1	7	4	1	95
Tổng cộng	6318	283	275	435	1800	9111

Bảng 9
Các giá trị thực thể thường xuyên nhất cho từng loại thuốc với tần suất của chúng trong ngoặc.

thực thể	Top 5 cho Diclofenac	Top 5 cho Lipitor
ADR	Tiêu chảy (28), buồn nôn (25), chảy máu âm đạo (17), chuột rút (16), chóng mặt (14)	Đau đớn (185), mệt mỏi (84), trầm cảm (83), đau cơ (82), giảm trí nhớ (62)
Bệnh	Viêm khớp (10), lạc nội mạc tử cung (3), đau nửa đầu (2), hai cực (2), viêm cân gan chân (2), Ibs (2), viêm xương khớp sau chấn thương (1), viêm khớp thắt lưng dưới (1)	Đau tim (14), viêm khớp (13), tiểu đường (8), đau cơ xơ hóa (8), MS (6)
Thuốc	Arthrotec (53), voltaren (25), celebrex (9), cataflam (6), advil (6), ibuprofen (6), aleve (5)	Lipitor (1034), zocor (44), pravachol (22), crestor (20), coq10 (19)
Phát hiện	Đột quỵ (2), mãn kinh (2), đau tim (2), bệnh dạ dày (1), viêm khớp (1)	Viêm khớp (17), cholesterol cao (15), đau tim (13), căng thẳng (8), đau (7)
Triệu chứng	Đau (89), đau đầu gối (8), đau lưng (6), viêm nhiễm (4), đau đớn (3), đau lưng dưới (3)	Đau (3), viêm (2), mãn kinh (2), huyết áp và nhịp tim thấp (1), lo lắng (1)

các triệu chứng liên quan đến Tăng cholesterol máu (cholesterol cao) không được đề cập thường xuyên trong các bài viết.

Bảng 8 báo cáo số lượng thẻ liên tục và không liên tục được chú thích trong bước chú thích thực thể.
Các thẻ liên tục đại diện cho một tập hợp các từ liên tục, trong đó các thẻ liên tục dis được chia thành nhiều nhịp trong một câu.
Ví dụ về các thẻ không liên tục như vậy được hiển thị trong Hình 1 (b).

Sau đó, chúng tôi xem xét các thực thể được chú thích bằng bất kỳ nhãn nào. Bảng 9 liệt kê các giá trị thực thể thường xuyên nhất cho từng loại trong số năm loại thực thể (Thuốc, ADR, Bệnh, Triệu chứng và Phát hiện) cho hai loại thuốc. Với hai loại thuốc trong kho văn bản điều trị các tình trạng rất khác nhau, giá trị thực thể của chúng đối với hầu hết các loại thực thể cũng khác nhau. Ví dụ, có sự phân biệt rõ ràng giữa các nhóm thuốc được đề cập cho Diclofenac và Lipitor. Tuy nhiên, các loại thuốc được đề cập trong bài đăng thường là thuốc thay thế hoặc bổ sung cho bệnh nhân và họ hiếm khi đề cập đến các loại thuốc khác mà họ đang dùng trong khi sử dụng loại thuốc là chủ đề của bài đăng.

Các chú thích từ giai đoạn chuẩn hóa cũng được phân tích với số liệu thống kê được hiển thị trong Bảng 10. Bộ cuối cùng chứa 1655 khái niệm từ AMT, trong đó chỉ có 123 (7,4%) là duy nhất. trong khác

Machine Translated by Google

Bảng 10
Thống kê về các thực thể được chuẩn hóa bằng AMT, SNOMED CT (SCT) và MedDRA.

	AMT	thuế TTBB	MedDRA
Số khái niệm được chú thích	1655	7259	6569
Số khái niệm duy nhất	123	924	686
Trung bình mỗi	1.2	5.4	5.2
bài đăng Tối đa không. khái niệm mỗi bài	17	42	34
1 tần số cao nhất. ý tưởng	Lipitor (1073)	Nỗi đau (371)	Nỗi đau (485)
tần suất nhiều thứ 2 ý tưởng	Thuốc giảm đau (62)	Đau cơ (288) Đau cơ (311)	
Tần suất cao thứ 3 ý tưởng	Zocor (48)	Đau dữ dội (196)	Đau khớp (311)

từ, cùng một tên thuốc (chủ yếu là Lipitor) đã được lặp lại trên tất cả các bài đăng. Trung bình, có 1,2 khái niệm tên thuốc trên mỗi bài đăng được liên kết với AMT. Trong một trường hợp cực đoan, chúng tôi đã có một bài đăng với 14 biệt dược có liên quan đến AMT. Số liệu thống kê tương tự được báo cáo cho SCT và MedDRA trong Bảng 10. Chúng tôi cũng báo cáo ba khái niệm được liên kết thường xuyên nhất cùng với tần suất xuất hiện của chúng trong kho văn bản cho từng thuật ngữ ở cuối Bảng 10.

6. Bài học kinh nghiệm

Trong quá trình tạo văn bản CADEC , một số quyết định đã được đưa ra liên quan đến những gì nên được chú thích, các quy tắc liên quan đến điều này và các công cụ phù hợp nhất để hoàn thành bài tập. Công cụ dựa trên web chính được sử dụng là Brat. Nó cung cấp một số lợi thế: tạo các chú thích dựa trên độ lệch ký tự trong các tệp văn bản thuần túy dễ xử lý, cũng như giao diện dễ cài đặt và chú thích. Tuy nhiên, một thiếu sót của công cụ này là tự động cung cấp thẻ mặc định trong trường hợp người chú thích tô sáng một phần văn bản nhưng quên chọn thẻ phù hợp. Trong trường hợp của chúng tôi, Brat đã chọn thẻ Thuốc cho những trường hợp như vậy. Do đó, chúng tôi đã giới thiệu một bước bổ sung để xem xét các chú thích và sửa đổi các thẻ không chính xác.

7. Hạn chế của CADEC

CADEC áp đặt các hạn chế sau đây có thể ảnh hưởng đến các nghiên cứu sử dụng bộ dữ liệu này:

Loại dữ liệu Văn bản phương tiện truyền thông xã hội thường ồn ào và chứa thông tin không chính xác, không đầy đủ hoặc thậm chí sai lệch.

Nguồn dữ liệu Có thể có những hạn chế kế thừa từ nguồn dữ liệu của chúng tôi: AskaPatient. Có thể có một nhóm người tiêu dùng cụ thể sử dụng AskaPatient để xem xét thuốc của họ. Cũng có thể có những hàm ý vốn có từ cấu trúc dữ liệu được sử dụng trong AskaPatient để thu thập các bài đánh giá.

Kích thước dữ liệu Kích thước của kho dữ liệu, 1253 bài đăng, bị giới hạn, không đại diện cho tất cả các loại thuốc và đánh giá của người tiêu dùng hiện có trên web.

Phạm vi giới hạn của các loại thuốc và tác dụng phụ CADEC chỉ bao gồm các loại thuốc có chứa Diclofenac và Atorvastatin trong thành phần hoạt tính của chúng. Do đó, phạm vi các tác dụng phụ trong ngữ liệu là của 12 loại thuốc có trong ngữ liệu và không có tác dụng phụ đặc hiệu đối với các loại thuốc khác. Ngay cả đối với các loại thuốc có trong bộ dữ liệu, phạm vi bao quát của các tác dụng phụ hiếm gặp được gọi là phản ứng thuốc đặc ứng9 vẫn còn hạn chế.

Thiếu tương tác thuốc-thuốc và chú thích quá liều thuốc Thông tin được cung cấp trong các báo cáo của người tiêu dùng thường tập trung vào một loại thuốc cụ thể. Do đó, thông tin về các loại thuốc khác do người tiêu dùng sử dụng thường bị thiếu. Điều này gây ra tình trạng thiếu thông tin về khả năng tương tác giữa thuốc và thuốc gây ra các phản ứng bất lợi được báo cáo. Điều tương tự cũng áp dụng cho liều lượng và tần suất sử dụng thuốc. Bệnh nhân có gặp tác dụng phụ do quá liều hay không vẫn chưa được biết do bản chất của các báo cáo này. Chúng tôi chưa chú thích kho dữ liệu về tương tác thuốc-thuốc hoặc nhiễm độc do dùng quá liều.

Lỗi của người chú thích Bất chấp những nỗ lực của chúng tôi đối với các chú thích chính xác, bộ dữ liệu có thể chứa lỗi của con người trong việc xác định các khái niệm hoặc liên kết chúng với các thuật ngữ y tế. Lỗi chú thích sẽ ảnh hưởng đến việc đánh giá bất kỳ hệ thống tự động nào sử dụng CADEC làm tiêu chuẩn vàng.

8. Kết luận và công việc trong tương lai

Chúng tôi đã tạo một kho dữ liệu truy cập mở có tên CADEC (CSIRO Adverse Drug Event Corpus) dành cho các nhà nghiên cứu khai thác văn bản để cảnh giác dược. Kho dữ liệu bao gồm các bài đăng của người tiêu dùng thuốc từ một diễn đàn y tế, AskaPatient, được chú thích với các khái niệm như tên thuốc, phản ứng bất lợi, bệnh tật và triệu chứng. Những khái niệm này được liên kết với các khái niệm tương ứng của chúng trong các từ vựng được kiểm soát. Kho dữ liệu CADEC tạo cơ hội cho các nhà nghiên cứu trong một số lĩnh vực (1) phát triển và đánh giá các hệ thống tự động trích xuất các tác dụng phụ của thuốc từ các báo cáo của người không chuyên; (2) phát triển các hệ thống trích xuất thuốc từ văn bản tự do; (3) phát triển các hệ thống tự động ánh xạ văn bản tự do tới SNOMED CT hoặc MedDRA, bởi vì nó chứa một ánh xạ phong phú từ thuật ngữ không chính thức sang chính thức, như được thể hiện trong SNOMED CT và MedDRA; và (4) sử dụng các biến thể để diễn đạt một tình trạng y tế của người dân, hoặc cách viết tên thuốc, được ghi lại trong CADEC, ngoài Từ vựng Sức khỏe Người tiêu dùng (CHV) trong thuật toán khai thác văn bản dựa trên từ điển.

Những thách thức của việc chú thích dữ liệu từ các diễn đàn y tế bao gồm văn bản bất quy tắc và ngôn ngữ thông tục cũng được thảo luận. Đáng chú ý nhất là trong những dữ liệu như vậy, các khoảng văn bản thường không rõ ràng, tạo ra những thách thức đối với các kỹ thuật xử lý văn bản tiêu chuẩn.

Chúng tôi đang trong quá trình chú thích kho dữ liệu với các mối quan hệ cũng như bao gồm các thực thể khác như liều lượng và tần suất dùng thuốc.

Sự nhìn nhận

AskaPatient vui lòng cung cấp dữ liệu được sử dụng trong nghiên cứu này chỉ cho mục đích nghiên cứu. Kho dữ liệu CADEC theo Giấy phép Dữ liệu CSIRO . Giấy phép này cho phép người dùng sử dụng dữ liệu cho các mục đích phi thương mại với sự ghi công phù hợp.

Phê duyệt đạo đức cho dự án này được lấy từ ủy ban đạo đức CSIRO đã phân loại công việc là rủi ro thấp (CSIRO Ecoscatics #07613).

Chúng tôi xin cảm ơn các sinh viên y khoa của Đại học Queensland, Timothy Sladden, Thomas Souchen, Warren Brown và Digvijay Khangarot, những người đã đóng góp vào việc phát triển các hướng dẫn và chú thích của kho văn bản CADEC .

Người giới thiệu

[1] ACSQHC, Tiêu chuẩn quốc gia về an toàn và chất lượng dịch vụ y tế, 2012.
[2] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. Leonard, J. Holmes, Xác định các tác động bất lợi tiềm ẩn khi sử dụng web: một cách tiếp cận mới để tạo giả thuyết y tế , J. Sinh học. Thông báo. 44 (6) (2011) 989–996.

9 Phản ứng thuốc đặc ứng còn được gọi là tác dụng phụ Loại B và rất hiếm khi xảy ra ở những người mà hệ thống miễn dịch của họ phản ứng với một số loại thuốc ngay cả với liều lượng nhỏ.

Machine Translated by Google

[3] J. Berlin, S. Glasser, S. Ellenberg, Phát hiện sự kiện bất lợi trong phát triển thuốc: khuyến nghị và nghĩa vụ sau giai đoạn 3, Am. J. Y tế Công cộng 98 (8) (2008) 1366–1371.

[4] B. Chee, R. Berlin, B. Schatz, Dự đoán các tác dụng phụ của thuốc từ thông điệp sức khỏe cá nhân, trong: Kỷ yếu chuyên đề hàng năm AMIA, Washington, DC, 2011, trang 217–226.

[5] K. Cohen, P. Ogren, L. Fox, L. Hunter, Thiết kế Corpus cho xử lý ngôn ngữ tự nhiên y sinh, trong: Hội thảo ACL-ISMB về Liên kết Văn bản Sinh học, Bản thể luận và Cơ sở dữ liệu: Khai thác Ngữ nghĩa Sinh học, Detroit, Michigan , 2005, trang 38–45.

[6] P. Coloma, G. Trifiró, V. Patadia, M. Sturkenboom, Giám sát an toàn sau khi đưa ra thị trường , An toàn thuốc 36 (3) (2013) 183–197.

[7] J. Couzin, An toàn thuốc: lỗ hổng trong mạng lưới an toàn, Science 307 (5707) (2005) 196–198.

[8] L. Deleger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo, I. Solti, Xây dựng kho ngữ liệu tiêu chuẩn vàng cho các nhiệm vụ xử lý ngôn ngữ tự nhiên y tế, trong: Hội nghị chuyên đề thường niên AMIA, Washington, DC, 2012, trang 144–153.

[9] R. Dogan, R. Leaman, Z. Lu, NCBI disease corpus: nguồn tài nguyên để nhận dạng tên bệnh và chuẩn hóa khái niệm, J. Biomed. Thông báo. 47 (2014) 1–10.

[10] H. Gurulingappa, R. Klinger, M. Hofmann-Apitius, J. Fluck, Đánh giá thực nghiệm các nguồn lực để xác định bệnh tật và tác dụng phụ trong tài liệu y sinh, trong: Hội thảo lần thứ 2 về Xây dựng và Đánh giá các nguồn lực cho Y sinh học Khai thác văn bản (ấn bản lần thứ 7 của Hội nghị Đánh giá và Tài nguyên Ngôn ngữ), 2010, trang 15–22.

[11] H. Gurulingappa, A. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, L. Toldo, Phát triển kho dữ liệu chuẩn để hỗ trợ trích xuất tự động các tác dụng phụ liên quan đến thuốc từ các báo cáo ca bệnh, sinh học. Thông báo. 45(5) (2012) 885–892.

[12] M. Herrero-Zazo, I. Segura-Bedmar, P. Martinez, Các vấn đề chú thích trong văn bản dược học, Proc. - Hành vi xã hội. Khoa học. 95 (2013) 211–219.

[13] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, Văn bản DDI: văn bản chú thích với dược chất và tương tác thuốc-thuốc , Biomed. Thông báo. 46(5) (2013) 914–920.

[14] B. Hug, C. Keohane, D. Seger, C. Yoon, D. Bates, Chi phí của các biến cố bất lợi do thuốc trong các bệnh viện cộng đồng, Ủy ban Hỗn hợp J. Chất lượng An toàn cho Bệnh nhân 38 (3) (2012) 120–126 .

[15] S. Karimi, S. Kim, L. Cavedon, Tác dụng phụ của thuốc: Diễn đàn bệnh nhân tiết lộ điều gì? trong: Hội thảo quốc tế lần thứ 2 về khoa học web và trao đổi thông tin trên web y tế, Glasgow, Vương quốc Anh, 2011, trang 14–15.

[16] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, C. Paris, Kỹ thuật khai thác dữ liệu và văn bản trong phát hiện phản ứng có hại của thuốc, ACM Comput. Khảo sát (2015) trong nhấn.

[17] R. Leaman, C. Miller, G. Gonzalez, Cho phép nhận biết các bệnh trong văn bản y sinh với học máy: ngữ liệu và điểm chuẩn, trong: Hội nghị chuyên đề quốc tế lần thứ 3 về ngôn ngữ trong sinh học và y học, đảo Jeju, Hàn Quốc, 2009 , trang 82–89.

[18] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, G. Gonzalez, Hướng tới cảnh giác dược thời đại Internet: Trích xuất các phản ứng có hại của thuốc từ các bài đăng của người dùng lên các mạng xã hội liên quan đến sức khỏe, trong: Hội thảo Về xử lý ngôn ngữ tự nhiên trong y sinh, Uppsala, Thụy Điển, 2010, trang 117–125.

[19] G. Leech, Corpus annotation scheme, Nhà ngôn ngữ học văn học. Điện toán. 8 (4) (1993) 275–281.

[20] I. Lewin, S. Kafkas, D. Rebholz-Schuhmann, Centroids: tiêu chuẩn vàng với các biến thể phân phối, trong: Hội nghị quốc tế lần thứ tám về tài nguyên và đánh giá ngôn ngữ, Istanbul, Thổ Nhĩ Kỳ, 2012, trang 3894–3900.

[21] X. Liu, H. Chen, AZDrugminer: một hệ thống trích xuất thông tin để khai thác các biến cố bất lợi do thuốc do bệnh nhân báo cáo trên các diễn đàn bệnh nhân trực tuyến, trong: Hội nghị Quốc tế về Sức khỏe Thông minh 2013, Bắc Kinh, Trung Quốc, 2013, trang 134– 150.

[22] A. Metke-Jimenez, S. Karimi, C. Paris, Đánh giá các thuật toán xử lý văn bản để trích xuất sự kiện bất lợi của thuốc từ phương tiện truyền thông xã hội, trong: Hội thảo quốc tế đầu tiên về phân tích và truy xuất phương tiện truyền thông xã hội, Gold Coast, Australia, 2014 , trang 15–20.

[23] J. Michel, M. Lawley, A. Chu, J. Barned, Lập bản đồ hồ sơ thuốc iPharmacy của cơ quan y tế Queensland với thuật ngữ thuốc của Úc sử dụng Snapper, Stud. Công nghệ Y tế Thông báo. 168 (2010) 104–116.

[24] NEHTA, Thuật ngữ Thuốc Úc v3 model-common v1.4, Tech. trả lời. EP-1825:2014, Cơ quan Chuyển đổi Y tế Điện tử Quốc gia, 2014.

[25] S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, G. Savova, Nhiệm vụ 7 của Semeval 2014: phân tích văn bản lâm sàng, trong: Hội thảo quốc tế lần thứ 8 về đánh giá ngữ nghĩa, Dublin, Ireland, 2014, trang .54–62.

[26] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, A. Setzer, Xây dựng kho ngữ liệu lâm sàng chú thích ngữ nghĩa, Biomed. Thông báo. 42 (5) (2009) 950–966.

[27] E. Roughhead, S. Semple, An toàn thuốc trong chăm sóc cấp tính ở Úc: chúng ta đang ở đâu? Phần 1: Đánh giá về mức độ và nguyên nhân của các vấn đề về thuốc 2002–2008, Australia New Zealand Health Policy 6 (1) (2009) 18.

[28] P. Stenetorp, S. Pyysalo, G. Topic´, T. Ohta, S. Ananiadou, J. Tsujii, Brat: một công cụ dựa trên web dành cho chú thích văn bản được NLP hỗ trợ, trong: The Demonstrations at the 13th Conference of Chương Châu Â u của Hiệp hội Ngôn ngữ học Máy tính, Avignon, Pháp, 2012, trang 102–107.

[29] D. Truran, P. Saad, M. Zhang, K. Innes, SNOMED CT và vị trí của nó trong thực hành quản lý thông tin y tế, Health Inform. Quản lý. 39 (2) (2010) 37– 39.

[30] E. Van Mulligen, A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. Kors, L. Furlong, Khối liệu EU-ADR: chú thích thuốc, bệnh, mục tiêu và mối quan hệ của chúng, Biomed. Thông báo. 45(5) (2012) 879–884.

[31] C.-H. Nguy, H.-Y. Kao, Z. Lu, PubTator: một công cụ khai thác văn bản dựa trên web để hỗ trợ xử lý sinh học, Nucl. Axit Res. 41 (W1) (2013) 518–522.

[32] D. Wishart, C. Knox, AC Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, Ngân hàng Dược phẩm: một nguồn tài nguyên toàn diện để phát hiện và thăm dò thuốc silico , Nucl. Axit Res. 34 (2006) 668–672.

[33] C. Yang, L. Jiang, H. Yang, X. Tang, Phát hiện các dấu hiệu phản ứng có hại của thuốc từ nội dung đóng góp của người tiêu dùng về sức khỏe trên mạng xã hội, trong: Hội thảo ACM SIGKDD về Tin học Y tế, Bắc Kinh, Trung Quốc, 2012.