Một hướng tiếp cận rút trích mối quan hệ y tế

Huỳnh Hữu Nghĩa, Hồ Bảo Quốc, Nguyễn An Tế

***Tóm tắt*—Rút trích mối quan hệ giữa các khái niệm y tế có ý nghĩa rất quan trọng trong lĩnh vực y tế. Các mối liên hệ biểu thị các sự kiện, các quan hệ có thể có giữa các khái niệm. Thông tin về các mối quan hệ này giúp cho người dùng (bác sĩ, bệnh nhân, nhà nghiên cứu y tế, người chăm sóc bệnh nhân, …) có một cái nhìn đầy đủ về vấn đề y tế. Điều này hỗ trợ cho các bác sĩ và những người chăm sóc bệnh nhân đưa ra những quyết định hiệu quả và hạn chế những sai sót trong quá trình điều trị. Bài báo tổng hợp các phương pháp rút trích mối quan hệ trên văn bản y tế và trình bày một hướng tiếp cận được đề xuất để rút trích mối quan hệ trên một loại mối quan hệ (template filling) cụ thể. Hướng tiếp cận kết hợp các phương pháp gồm dựa trên tự điển, luật và máy học. Phương pháp dựa trên luật sử dụng mối quan hệ ngữ nghĩa phụ thuộc giữa các khái niệm để rút trích luật. Phương pháp máy học sử dụng thuật toán SVM (Support Vector Machine) và tập đặc trưng. Kết quả của hướng tiếp cận được đánh giá hiệu quả dựa trên độ đo tính đúng (accuracy) là 0.849.**

***Từ khóa*—Rút trích mối quan hệ, rút trích thông tin, khai thác thông tin lâm sàng, khai thác văn bản.**

1 GIỚI THIỆU

TIN học y tế (medical informatics) là lĩnh vực ứng dụng công nghệ thông tin vào y khoa và chăm sóc sức khỏe. Mục đích của tin học y tế là nghiên cứu tìm kiếm các phương pháp tối ưu hóa việc sử dụng thông tin nhằm cải thiện chất lượng chăm sóc y tế, giảm chi phí, cung cấp cho giáo dục và nghiên cứu y khoa hiệu quả. Thời gian qua, lĩnh vực tin học y tế có những tiến bộ và phát triển.

Những tiến bộ trong tin học y tế như hồ sơ bệnh án điện tử (EHR - Electronic Health Record), các hệ thống chăm sóc y tế và những ứng dụng trong y sinh học (biomedical) đã sinh ra khối lượng dữ liệu lớn được lưu trữ trong hàng trăm cơ sở dữ liệu. Ngoài ra, việc số hóa dữ liệu y tế quan trọng như các báo cáo phòng thí nghiệm, tài liệu nghiên cứu và hình ảnh giải phẫu cũng đã tạo ra dữ liệu chăm sóc bệnh nhân khổng lồ được lưu trữ trên máy tính. Sự phát triển của Internet cũng làm xuất hiện nhiều trang web tư vấn cách chăm sóc sức khỏe và đặc biệt là sự phát triển của bách khoa toàn thư mở Wikipedia chia sẻ thông tin và hình ảnh y khoa. Càng ngày càng có nhiều tạp chí y tế điện tử đăng tải những thành tựu khoa học kỹ thuật y khoa. Đây là nguồn dữ liệu lớn có thể cung cấp những thông tin bổ ích cho người dùng trong lĩnh vực y tế.

Nhu cầu thông tin đối với người dùng trong lĩnh vực y tế là rất đa dạng. Bác sĩ cần thông tin hỗ trợ trong quá trình chẩn đoán và điều trị. Sinh viên và nhà nghiên cứu cần tài liệu huấn luyện, những trường hợp điều trị cụ thể đã thực hiện, kết quả xét nghiệm và chẩn đoán, tạp chí, bài báo hoặc sách có liên quan hay những tóm tắt thông tin quan trọng. Bệnh nhân cần hiểu biết về nguyên nhân bệnh, điều kiện điều trị y khoa, hợp tác hỗ trợ việc điều trị, theo dõi quá trình điều trị. Một khả năng ứng dụng khác như công ty bảo hiểm cần giám sát việc sử dụng các điều kiện điều trị với chi phí thấp, kiểm soát rủi ro và hỗ trợ mức dịch vụ tốt, xác minh các thủ tục chẩn đoán và theo dõi kết quả điều trị.

Với lượng dữ liệu lớn và nhu cầu thông tin của người dùng mang đến cho lĩnh vực tin học y tế nhiều thách thức. Các nhà quản lý đang tìm kiếm giải pháp quản lý dữ liệu phù hợp và hiệu quả để phục vụ điều trị. Các tổ chức chăm sóc y tế gặp khó khăn khi đọc-hiểu đúng các thuật ngữ trong hồ sơ bệnh nhân liên quan đến những bệnh, các triệu chứng và nguyên nhân để điều trị hiệu quả. Dữ liệu y tế cũng có rất nhiều thách thức do hầu hết là dữ liệu văn bản không có cấu trúc. Các văn bản được định dạng khác nhau liên quan đến từng loại báo cáo, một số báo cáo chứa các bảng biểu với các hình thức khác nhau và sự xuất hiện của

Bài nhận ngày 04 tháng 04 năm 2017, hoàn chỉnh sửa chữa ngày 02 tháng 06 năm 2017.

Tác giả Huỳnh Hữu Nghĩa công tác tại Trường Đại học Lao động – Xã hội (CSII) (email: huynhnghiavn@gmail.com)

Tác giả Hồ Bảo Quốc công tác tại Trường Đại học Khoa học Tự nhiên, ĐHQGHCM (email: hbquoc@fit.hcmus.edu.vn)

Tác giả Nguyễn An Tế công tác tại Trường Đại học Kinh tế TP HCM (email: tena@ueh.edu.vn).

rất nhiều ký tự/chữ viết tắt. Các ký tự/chữ viết tắt là nguyên nhân rất lớn dẫn đến sự nhập nhằng và tính mơ hồ trong việc hiểu nội dung của văn bản. Để hiểu rõ nội dung tài liệu người dùng phải tìm đọc nhiều tài liệu khác có liên quan.

Hiện nay, người dùng tìm kiếm thông tin thông qua một số nguồn trực tuyến phổ biến như các công cụ tìm kiếm thông thường (Google, Bing và Yahoo!), các cơ sở dữ liệu nghiên cứu y tế (PubMed) và Wikipedia. Kết quả tìm kiếm là những tài liệu liên quan đến nội dung tìm kiếm, người dùng phải tự đọc tất cả tài liệu có để nắm bắt thông tin cần thiết phục vụ cho nhu cầu nên người dùng mất rất nhiều thời gian để đọc nghiên cứu tài liệu. Để nắm bắt tri thức mới trong lĩnh vực y tế đối với người dùng thật khó khăn trong điều kiện khối lượng lớn dữ liệu mới phát sinh hàng ngày.

Vấn đề được đặt ra là "Làm thế nào để đáp ứng nhu cầu thông tin y tế cho người dùng trong trường hợp bùng nổ dữ liệu?". Để giải quyết vấn đề này, một mô hình khai thác thông tin y tế ở mức khái niệm là rất cần thiết. Những yêu cầu đối với mô hình bao gồm: (1) Phân tích tự động nội dung tài liệu để nhận diện, gán nhãn và rút trích các thông tin quan trọng xuất hiện trong tài liệu sau đó chuẩn hóa các thông tin được rút trích đến các khái niệm đã định nghĩa trong các ontology lĩnh vực y tế; (2) Xác định hoặc rút trích mối quan hệ giữa các khái niệm xuất hiện trong tài liệu, nhằm tạo ra liên kết ngữ nghĩa giữa các khái niệm xuất hiện trên một hay nhiều tài liệu; (3) Tổ chức lưu trữ khái niệm và mối quan hệ thành kho tri thức phục vụ nhu cầu khai thác thông tin của người dùng; và (4) Hệ thống hỏi – đáp của người dùng. Kho trí thức này còn là nguồn cơ sở cung cấp tri thức để phát triển các hệ thống hỗ trợ ra quyết định trong lĩnh vực y tế. Một trường cụ thể về nhu cầu người dùng được minh họa ý nghĩa của mô hình như sau: Bệnh nhân hoặc người thân gặp khó khăn trong việc hiểu những thuật ngữ/khái niệm xuất hiện trong tóm tắt xuất viện. Ví dụ: một tài liệu xuất viện có nội dung "AP: 72 yo f w/ ESRD on HD, CAD, HTN, asthma p/w significant hyperkalemia & associated arrythmias." trong đó xuất hiện nhiều ký tự/chữ viết tắt và các khái niệm mà người dùng không hiểu được. Việc hiểu biết khái niệm sẽ giúp quá trình tự chăm sóc và điều trị được tốt hơn. Như vậy, hệ thống đầu tiên sẽ làm nổi bật lên những khái niệm trong tóm tắt xuất viện, liên kết đến các nguồn tri thức để giải thích ý của khái niệm mà người dùng quan tâm, có thể mở rộng giải thích mối quan hệ liên quan giữa các khái niệm từ các nguồn tri thức như: UMLS[1], Wikipedia, v.v… hoặc liên kết đến các trang web hay tài liệu liên quan.

Bài toán rút trích thông tin được xem là bài toán cơ bản đầu tiên trong mô hình khai thác thông tin y tế. Rút trích thông tin đề cập đến quá trình xử lý tự động trích xuất thông tin từ các văn bản phi cấu trúc hoặc bán cấu trúc để xây dựng các sự kiện có cấu trúc. Trong lĩnh vực tin học y tế, văn bản phi cấu trúc phổ biến gồm các bài báo khoa học, những tài liệu văn bản trong các hồ sơ bệnh án điện tử hoặc các hệ thống thông tin lâm sàng. Rút trích thông tin có 2 bài toán chính liên quan đến quá trình xử lý văn bản y tế. Thứ nhất, nhận diện khái niệm là bài toán xác định và phân lớp các khái niệm y tế vào các loại được định nghĩa trước chẳng hạn như: tên Protein, Genes, Bệnh, v.v… (Bài toán này được trình bày trong bài báo khác). Sau đó, các khái niệm được chuẩn hóa và biểu diễn rõ ràng thông qua các nguồn tài nguyên ontology và tiếp theo là phân lớp khái niệm vào các loại ngữ nghĩa. Bài toán thứ hai là rút trích mối quan hệ nhằm mục đích phát hiện mối quan hệ giữa các khái niệm. Ví dụ: các mối quan hệ giữa Gene-Bệnh, sự tương tác giữa Protein-Protein và các mối quan hệ giữa Điều trị - Vấn đề y tế.

Mục tiêu của bài báo là hệ thống các hướng tiếp cận cho bài toán rút trích mối quan hệ trên tài liệu y tế và trình bày một thực nghiệm xác định mối quan hệ cụ thể. Bố cục phần còn lại của bài báo gồm: mô tả toán rút trích mối quan hệ y tế, các phương pháp rút trích mối quan hệ đã được đề xuất, kết quả thực nghiệm và kết luận.

2 CÁC BÀI TOÁN

Bài toán rút trích mối quan hệ là xác định và rút ra các mối quan hệ ngữ nghĩa giữa những khái niệm được thể hiện trong văn bản. Các quan hệ có thể là mối quan hệ xã hội như quan hệ giữa người với người, giữa người với tổ chức, giữa các tổ chức, v.v… Trong lĩnh vực y tế, các mối quan hệ có thể là sự tương tác giữa protein-protein, mối quan hệ giữa vấn đề y tế và điều trị, …

Một số bài toán liên quan đến rút trích mối quan hệ bao gồm: xác định mối quan hệ giữa hai khái niệm (mối quan hệ nhị phân), sự kiện (mối quan hệ phức tạp), xác định giá trị cho các thuộc tính của khái niệm (điền mẫu), đồng tham chiếu, mối quan hệ thời gian, … Một vài trường hợp cụ

[1] *https://www.nlm.nih.gov/research/umls/*

thể trong lĩnh vực y tế được trình bày như sau:

Trong i2b2 năm 2010 đã định nghĩa các mối quan hệ nhị phân gồm mối quan hệ giữa vấn đề y tế - điều trị (ví dụ: điều trị làm cải thiện vấn đề y tế, điều trị làm xấu đi vấn đề y tế, điều trị giải quyết vấn đề y tế và điều trị không giải quyết vấn đề y tế), mối quan hệ giữa vấn đề y tế - xét nghiệm (ví dụ: xét nghiệm để phát hiện vấn đề y tế, xét nghiệm được thực hiện để điều tra vấn đề y tế) và mối quan hệ giữa vấn đề y tế - vấn đề y tế (ví dụ: vấn đề y tế chỉ ra vấn đề y tế).

Năm 2011, i2b2 đã xác định các mối quan hệ đồng tham chiếu giữa các khái niệm (treatment, problem, test, person và pronoun). Các đồng tham chiếu yêu cầu xác định gồm coref_person, coref_problem, coref_treatment và coreftest. Các cặp đồng tham chiếu được liên kết tạo thành một chuỗi khái niệm liên quan đến bệnh nhân, từ đó tạo ra cách nhìn đầy đủ về tình trạng lâm sàng.

Phần tiếp theo chúng tôi trình bày khái quát các phương pháp rút trích mối quan hệ.

3 CÁC ĐẶC ĐIỂM DỰ ĐOÁN MỐI QUAN HỆ

Việc rút trích mối quan hệ không đơn giản như rút trích trích khái niệm, để rút trích mối quan hệ giữa hai khái niệm trên cùng một câu yêu cầu sự kết hợp khéo léo từ cấu trúc cú pháp và ngữ nghĩa đa dạng trong câu. Một số đặc điểm có thể sử dụng để dự đoán mối quan hệ như sau:

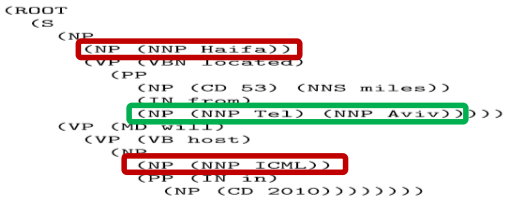
Mặt chữ (surface tokens): Các từ (token) xung quanh và bên trong giữa hai khái niệm là những đầu mối để xác định mối quan hệ. Ví dụ: Sự hiện diện của từ đơn *epidemic* giữa hai khái niệm *Disease* và *Location* thể hiện khả năng dự đoán mối quan hệ "*outbreak*" trong câu như sau:

> The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly <*Disease*>Ebola </*Company*> epidemic in <*Location*>Zaire </*Location*>.

Nhãn từ loại (part-of-speech tags): Nhãn từ loại đóng vai trò quan trọng trong rút trích mối quan hệ. Các động từ trong câu chính là những từ khóa để xác định mối quan hệ giữa các khái niệm. Ví dụ: Từ *hosts* xuất hiện giữa hai khái niệm *Conferences* và *Location* được gán nhãn là động từ (VBZ), từ đó có thể rút trích mối quan hệ "*held in*" trong câu sau đây:

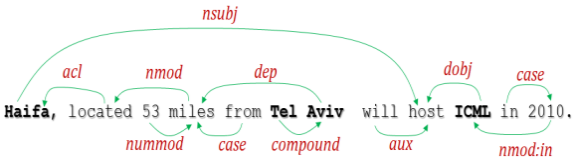
> <*Location*> The/DT University/NNP of/IN Helsinki/NNP </*Location*> hosts/VBZ <*Conference*> ICML/NNP </*Conference*> this/DT year/NN

Cấu trúc cây phân tích cú pháp (systactic parse tree structure): Cây phân tích cú pháp nhóm các từ trong câu thành những cụm từ như: Các cụm danh từ, cụm giới từ và cụm động từ. Nó có giá trị trong việc hiểu mối quan hệ giữa các khái niệm trong câu hơn là nhãn từ loại. Ví dụ: Xét câu "<*Location*> Haifa </*Location*> located 53 miles from <*Location*> Tel Aviv </*Location*> will host <*Conference*> ICML </*Conference*> in 2010". Dựa trên mối quan hệ gần thì cặp (Tel Aviv, ICML) thể hiện mối quan hệ "*held in*" phù hợp hơn cặp (Haifa, ICML). Nhưng xét trên cây phân tích cú pháp (hình 1) thì *ICML* gần *Haifa* hơn là *Tel Aviv* do *Haifa* đứng đầu cụm danh từ "*Haifa located 53 miles from Tel Aviv*" nó tạo thành chủ ngữ của cụm động từ "*will host ICML in 2010*".



Hình 1. Biểu diễn cây phân tích cú pháp cho câu "<*Location*> Haifa </*Location*> located 53 miles from <*Location*> Tel Aviv </*Location*> will host <*Conference*> ICML </*Conference*> in 2010"

Đồ thị phụ thuộc (dependency graph): Đồ thị phụ thuộc biểu diễn các mối liên kết mỗi từ đến các từ mà phụ thuộc vào nó. Ví dụ: Xem đồ thị phụ thuộc trong hình 2. Trên đồ thị rõ ràng động từ *host* được liên kết trực tiếp đến bởi cả hai khái niệm *Haifa* và *ICML*. Điều này đã tạo nên mối liên kết chặt chẽ giữa các khái niệm. Ngược lại đường dẫn giữa *ICML* và *Tel Aviv* phải đi qua *Haifa – located – miles*.



Hình 2. Biểu diễn đồ thị phụ thuộc cho câu "<*Location*> Haifa </*Location*> located 53 miles from <*Location*> Tel Aviv </*Location*> will host <*Conference*> ICML </*Conference*> in 2010"

4 CÁC PHƯƠNG PHÁP

Nhiều thập kỷ qua, có nhiều hướng tiếp cận đề xuất cho bài toán rút trích mối quan hệ trên tài liệu y khoa. Các hướng tiếp cận hiện nay gồm dựa trên luật, dựa trên máy học giám sát và bán giám sát. Các hướng tiếp cận lần lượt được trình bày chi tiết ở phần tiếp theo sau đây.

4.1 Hướng tiếp cận dựa trên luật

Các hướng tiếp cận dựa trên luật áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên và các mẫu được xây dựng bằng thủ công trên lĩnh vực cụ thể để nắm bắt các kiểu mối quan hệ khác nhau xuất hiện trong văn bản. Khi xây dựng tập luật tốn nhân công và chi phí cao cũng như yêu cầu phải có chuyên môn sâu. Ví dụ: chương trình rút trích mối quan hệ mã nguồn mở RelEx [6]. RelEx dựa trên trúc phụ thuộc để xây dựng tập luật và rút trích các mối quan hệ. Hệ thống RelEx sau đó được sử dụng rút trích các mối quan hệ tương tác giữa gene và protein trên tập dữ liệu hơn 1 triệu tóm tắt MedLine. Kết quả rút trích được trên 150 ngàn mối quan hệ với hiệu quả đánh giá là 80%. Một số nhóm nghiên cứu đề xuất các hướng tiếp cận khác nhau dựa trên luật nhằm xác định các mối quan hệ giữa các thực thể y sinh học như [3, 9]. Gần đây, công trình [2] đề xuất hệ thống dựa trên luật gọi là MeTAE (Medical Texts Annotation and Exploration) cho phép rút trích và gán nhãn thực thể và mối quan hệ trên văn bản y tế. Hướng tiếp cận của hệ thống dựa trên qui tắc ngôn ngữ để rút trích các mối quan hệ giữa triệu chứng và vấn đề y tế.

4.2 Hướng tiếp cận máy học

Hướng tiếp cận máy học dựa trên các thuật toán học có giám sát để huấn luyện và xác định những mối quan hệ trong văn bản. Tuy nhiên, hướng tiếp cận máy học yêu cầu dữ liệu huấn luyện được gán nhãn để xây dựng một bộ phân lớp tin cậy. Hướng tiếp cận máy học rút trích mối quan hệ có thể chia làm hai hướng tiếp cận là dựa trên đặc trưng và dựa trên Kernel. Các kỹ thuật dựa trên đặc trưng thì rút trích đặc trưng văn bản từ tài liệu đầu vào (ví dụ: những từ xuất hiện giữa các thực thể) và sử dụng những thuật toán học có giám sát để huấn luyện. Phương pháp dựa trên Kernel là mã hóa cấu trúc biểu diễn văn bản (ví dụ: chuỗi từ liên tục (word sequence) và hàm kernel) được thiết kế để nắm bắt và phân biệt giữa các cấu trúc có nghĩa.

Phân lớp dựa trên đặc trưng

Hướng tiếp cận rút trích mối quan hệ xem bài toán như vấn đề phân lớp. Cụ thể, bất kỳ một cặp khái niệm đồng xuất hiện trong cùng một câu thì được xem là một thể hiện mối quan hệ ứng viên. Mục tiêu là gán một nhãn phân lớp cho thể hiện trong đó nhãn phân lớp là một trong những kiểu quan hệ được định nghĩa trước hoặc nil (không) cho cặp khái niệm không liên quan. Quá trình xử lý có thể được thực hiện qua hai giai đoạn, ở giai đoạn đầu tiên là xác định hai khái niệm (cho dù có liên quan hay không) và giai đoạn thứ hai là xác định loại quan hệ cho từng cặp khái niệm liên quan.

Hướng tiếp cận phân lớp giả định rằng kho ngữ liệu huấn luyện có sẵn, trong đó tất cả những mối quan hệ cho từng kiểu quan hệ được định nghĩa trước đã được gán nhãn bằng thủ công. Những mối quan hệ được sử dụng như các mẫu huấn luyện đáng tin cậy. Từng sự thể hiện mối quan hệ ứng viên được biểu diễn bởi một tập đặc trưng được chọn lựa một cách cẩn thận. Các thuật toán học chuẩn như SVM và hồi qui lôgíc (logistic regression) có thể được sử dụng để huấn luyện các phân lớp mối quan hệ.

Xác định đặc trưng là một bước quan trọng cho hướng tiếp cập phân lớp. Những người nghiên cứu phải khảo sát hàng loạt các đặc trưng về từ vựng, cú pháp và ngữ nghĩa. Các đặc trưng được sử dụng phổ biến được giới thiệu như sau:

Đặc trưng khái niệm: Thường thì hai khái niệm có sự tương quan với các loại mối quan hệ nào đó gồm những từ bên trong khái niệm và các loại khái niệm. Ví dụ: trong các tập dữ liệu ACE, các khái niệm như: *father*, *mother*, *brother* và *sister* và loại khái niệm **person** là những chỉ định tốt cho loại quan hệ con **family**.

Đặc trưng ngữ cảnh từ vựng: Ngữ cảnh trực tiếp xung quanh hai khái niệm là quan trọng. Cách đơn giản nhất để kết hợp dấu hiệu (bằng chứng) từ những ngữ cảnh là sử dụng các đặc trưng từ vựng. Ví dụ: nếu từ *founded* xuất hiện giữa hai khái niệm, chúng có nhiều khả năng có mối quan hệ *FounderOf*.

Đặc trưng ngữ cảnh cú pháp: Các mối quan hệ cú pháp giữa hai khái niệm hoặc giữa một khái niệm và từ khác có thể có ít. Ví dụ: nếu thực thể đầu tiên là chủ ngữ của động từ *founded* và thực thể thứ hai là túc từ của động từ *founded* thì ngay lập tức có thể khẳng định rằng tồn tại mối quan hệ *FounderOf* giữa hai kh. Các đặc trưng cú pháp có được phải dựa trên cây phân tích cú pháp của câu chứa thể hiện mối quaái niệm hệ.

Tri thức cơ sở (Background knowledge): Công trình [20] đã nghiên cứu sử dụng tri thức cơ sở cho bài toán rút trích mối quan hệ.

Phương pháp Kernel

Một phương pháp quan trọng rút trích mối quan hệ là phân lớp dựa trên *kernel*. *Kernel* có thể được xem như độ đo sự tương đồng giữa các quan sát. Hiện nay có ba kiểu *kernel* phổ biến gồm các *kernel dựa trên chuỗi tuần tự*, các *kernel dựa trên cây* và các *kernel ghép*.

Kernel dựa trên chuỗi tuần tự. Tác giả công trình [16] định nghĩa một *kernel* đơn giản dựa trên

các hướng đi phụ thuộc ngắn nhất giữa hai khái niệm. Hai hướng đi phụ thuộc là tương đồng nếu chúng có cùng chiều dài và chia sẽ nhiều nút (node) chung. Ở đây, một nút có thể được biểu diễn bằng chính từ đó, nhãn từ loại và kiểu khái niệm của nó. Do đó hai hướng đi phụ thuộc "protestors → seized ← stations" và "troops → raided ← churches" có giá trị tương đồng khác 0 bởi vì cả hai có thể được biểu diễn như "Person → VBD ← Facility" mặc dùng chúng không chia sẽ bất kỳ từ chung nào. Một hạn chế của *kernel* này là bất kỳ hai hướng đi phụ thuộc với chiều dài khác nhau có độ tương tự là 0. Công trình [17] đã giới thiệu *kernel* chuỗi tuần tự con (subsequence) trong đó sự tương đồng giữa hai chuỗi tuần tự được định nghĩa trên chuỗi tuần tự con tương đồng của chúng. Tác giả đã thử nghiệm *kernel* chuỗi tuần tự con cho việc phát hiện sự tương tác giữa protein-protein.

Kernel dựa trên cây. Sử dụng cấu trúc con chung để đo độ tương đồng. Tác giả công trình [4] đã định nghĩa một *kernel* dựa trên các cây cú pháp thể hiện mối quan hệ. Ý tưởng chính là nếu hai cây phân tích cú pháp chia sẽ nhiều cấu trúc cây con chung thì hai thể hiện mối quan hệ tương đồng nhau. Sau đó, công trình [1] đã mở rộng ý tưởng trên cây phân tích cú pháp phụ thuộc. Bên cạnh đó, công trình [10] đã áp dụng *kernel* cây tích chập được đề xuất lần đầu bởi [11] nhằm rút trích mối quan hệ. Phương pháp dựa trên *kernel* cây tích chập sau đó được [8] cải tiến và đạt được hiệu quả mới nhất với độ đo F-1 gần 77% trên tập dữ liệu chuẩn của ACE 2004.

Kernel ghép. Là sự kết hợp nhiều *kernel* khác nhau hình thành nên một *kernel ghép*. Điều này được thực hiện khi mà không thể tìm ra tất cả những đặc trưng cần thiết để hình thành một kernel duy nhất. Công trình [18] đã định nghĩa một số kernel cú pháp như kernel tham số và kernel đường dẫn phụ thuộc sau đó kết hợp thành một kernel ghép. Các tác giả [12] kết hợp một kernel khái niệm với một *kernel* cây tích chập hình thành nên một kernel ghép.

4.3 Hướng tiếp cận học bán giám sát

Cả hai phương pháp phân lớp dựa trên đặc trưng và dựa trên kernel cho bài toán rút trích mối quan hệ dựa trên một số lượng lớn dữ liệu huấn luyện, tốn kém nhiều công sức và thời gian. Một giải pháp cho vấn đề này là phương pháp học bán giám sát làm việc với dữ liệu huấn luyện ít hơn nhiều. Phương pháp học bán giám sát đáng chú ý cho việc rút trích mối quan hệ là *hạt giống* (*bootstrapping*), nó bắt đầu từ một tập nhỏ các thể hiện mối quan hệ ban đầu gọi là hạt giống và lặp đi lặp lại để học nhiều thể hiện mối quan hệ và các mẫu rút trích. Nó đã được nghiên cứu mở rộng ở hai công trình [5, 19]. Sau đó, một mô hình khác được gọi là *giám sát từ xa (distant supervision)*, phương pháp được đề xuất để thực hiện sử dụng một số lượng lớn những thể hiện mối quan hệ đã biết trong các cơ sở tri thức lớn có sẵn để tạo ra dữ liệu huấn luyện [13]. Cả hai phương pháp hạt giống và giám sát từ xa có một khuyết điểm là tự động tạo ra dữ liệu huấn luyện nhiễu. Vì vậy, cần phải có giải pháp chọn đặc trưng và lọc mẫu.

Phần tiếp theo của bài báo sẽ trình bày một đề xuất hướng tiếp cận rút trích mối quan hệ cho bài toán cụ thể là xác định giá trị cho các thuộc tính liên quan đến khái niệm (hay gọi là bài toán điền mẫu).

5 HƯỚNG TIẾP CẬN RÚT TRÍCH MỐI QUAN HỆ Y TẾ

Bài toán xác định giá trị cho các thuộc tính của khái niệm y tế được đề xuất bởi ShARe/ CLEFe Health 2014[2]. Mỗi tài liệu y tế có một danh sách các khái niệm y tế gồm những bệnh/rối loạn xuất hiện trong tài liệu tương ứng. Mỗi bệnh/rối loạn được định nghĩa 10 thuộc tính. Ý nghĩa của từng thuộc tính và các giá trị chuẩn hóa cho thuộc được trình bày ở bảng 1 như sau:

BẢNG 1
Ý NGHĨA CỦA TỪNG THUỘC TÍNH VÀ GIÁ TRỊ CHUẨN HÓA.

Các thuộc tính **Bệnh/Rối loạn**	**Các định nghĩa từ ShARe guidelines**	**Giá trị chuẩn hóa**
Negation Indicator	Cho biết **bệnh/rối loạn** là âm tính	*no, yes
Subject Class	Cho biết người đã trải qua **bệnh/rối loạn**.	*patient, family_member, donor_family_member, donor_other, null, other
Uncertainty Indicator	Cho biết sự đánh giá nghi ngờ nhận xét về **bệnh/rối loạn**.	*false, true
Course Class	Cho biết diễn tiến của **bệnh/rối loạn**.	*unmarked, changed, increased, decreased, improved, worsened, resolved
Severity Class	Cho biết khả năng của **bệnh/rối loạn** như thế nào.	*unmarked, slight, moderate, severe
Conditional Class	Cho biết có tồn tại điều kiện của **bệnh/rối loạn** trong những hoàn cảnh nào đó.	*false, true
Generic Class	Cho biết có đề cập đến đặc điểm chung của **bệnh/rối loạn**.	*false, true
Body Location	Trình bày vị trí giải phẫu thuộc các kiểu ngữ nghĩa UMLS.	*NULL, CUI
DocTime Class	Cho biết mối quan hệ thời gian giữa **bệnh/rối loạn** và thời điểm viết tài liệu	*unknown, before, after, overlap, before-overlap
Temporal Expression	Biểu diễn bất kỳ biểu thức thời gian TIMEX (TimeML) liên quan đến **bệnh/rối loạn**: ngày bắt đầu, khoảng thời gian hoặc ngày kết thúc.	*none, date, time, duration, set

[2] http://clefehealth2014.dcu.ie/

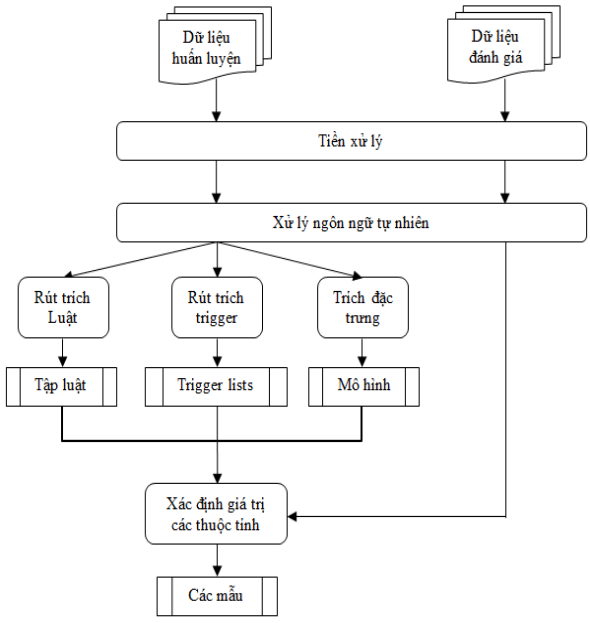
Mỗi thuộc tính được yêu cầu xác định 2 giá trị gồm giá trị chuẩn hóa cho thuộc tính và dấu hiệu nhận biết (trigger) giá trị chuẩn hóa thuộc tính, xem bảng 2.

BẢNG 2. CÁC CÂU VÍ DỤ VÀ GIÁ TRỊ CHUẨN HÓA VÀ VỊ TRÍ DẤU HIỆU NHẬN BIẾT GIÁ TRỊ THUỘC TÍNH.

Các thuộc tính của Bệnh/Rối loạn	Câu ví dụ	Giá trị chuẩn hóa	Vị trí của dấu hiệu
Negation Indicator	*Denies* numbness	yes	*0-5*
Subject Class	*Son* has schizophrenia.	family_member	*0-2*
Uncertainly Indicator	*Evaluation of* MI.	yes	*0-9*
Course Class	The cough *worsened* over the next two weeks.	worsened	*11-18*
Severity Class	He noted a *slight* bleeding.	slight	*12-17*
Conditional Class	Return *if* fever.	true	*8-9*
Generic Class	pain *while* standing	true	*6-10*
Body Location	Patient has *facial* rash.	*C0015450: Face*	*13-18*
DocTime Class	Patient had tumor removed.	before	--
Temporal Expression	The rash was present *for 3 days*.	duration	*22-31*

5.1 Kiến trúc hệ thống

Mỗi bệnh/rối loạn gồm 10 thuộc tính (được nêu trên), mỗi thuộc tính có yêu cầu khác nhau, cho nên hệ thống có sự kiết hợp các phương pháp bao gồm dựa trên từ điển, dựa trên luật và máy học để giải quyết bài toán. Cụ thể, các thuộc tính từ 1 đến 8 áp dụng luật và từ điển, thuộc tính 9 áp dụng máy học và luật, còn thuộc tính 10 áp dụng luật (biểu thức chính qui). Kiến trúc tổng quát của hệ thống được trình bày ở hình 3.



Hình 3. Kiến trúc của hệ thống.

Tiền xử lý

Trong quá trình tạo ra tài liệu lâm sàng, các bác sĩ và những người chăm sóc y tế thường hay sử dụng một số ký hiệu thể hiện ý nghĩa ngữ dụng như: dấu "–" hoặc "+" đứng trước một bệnh/rối loạn có ý nghĩa âm tính hoặc dương tính (ví dụ: –lymphadenopathy hoặc +thyromegaly), dấu "?" đứng bên trái hoặc phải của bệnh/rối loạn có ý nghĩa nghi ngờ (ví dụ: duplex?thrombus/clot). Xét về mặt ngữ dụng các dấu trên đều là những dấu hiệu để nhận biết giá trị cho thuộc tính. Trong thành phần xử lý ngôn ngữ tự nhiên, hệ thống có sử dụng một số thư viện xử lý ngôn ngữ tự nhiên có sẵn (xem thành phần xử lý ngôn ngữ tự nhiên). Đối với các thư viện này xem các dấu nêu trên là những dấu câu nên chúng bị bỏ qua khi phân tích cú pháp và phân tích phụ thuộc, điều này làm mất đi dấu hiệu để xác định giá trị cho thuộc tính. Cho nên, ở bước tiền xử lý hệ thống thay thế những dấu nêu trên thành những ký tự khác để chúng có thể xuất hiện trên cây cú pháp hoặc đồ thị phụ thuộc.

Xử lý ngôn ngữ tự nhiên

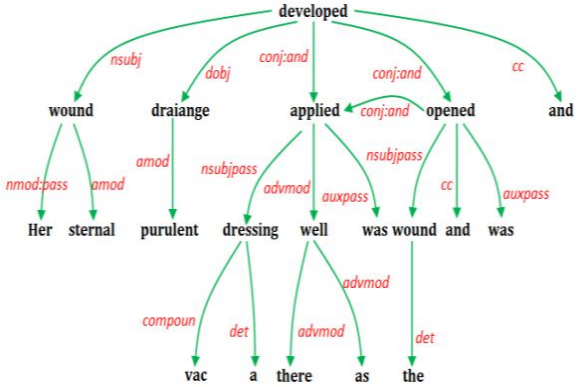
Xử lý ngôn ngữ tự nhiên được thực hiện nhằm mục đích là chuyển tài liệu dạng không có cấu trúc hoặc bán cấu trúc về các dạng có cấu trúc để có thể xử lý được trên máy tính. Hệ thống sử dụng thư viện xử lý ngôn ngữ tự nhiên có sẵn của Stanford NLP3 để thực hiện những việc như: tách câu từ tài liệu, mỗi câu tách thành từng token, xử lý cú pháp và đồ thị phụ thuộc trên câu văn bản. Stanford NLP đã thiết kế biểu diễn phụ thuộc (đồ thị phụ thuộc) để mô tả những mối quan hệ ngữ pháp giữa các từ trong câu nhằm giúp cho những người không có chuyên môn về ngôn ngữ học có thể dễ dàng hiểu và sử dụng để trích xuất các mối quan hệ trong văn bản. Hiện tại việc biểu diễn bao gồm xấp xỉ 50 mối quan hệ ngữ pháp. Những phụ thuộc đều là mối quan hệ nhị phân, như là một bộ ba gồm một mối quan hệ ngữ pháp chứa một từ chính (governor/head) và một từ phụ thuộc (dependent). Ví dụ: xét câu "Her sternal wound developed purulent draiange, and the wound was opened and a vac dressing was applied there as well.", kết quả biểu diễn phụ thuộc là:

nmod:poss(wound-3, Her-1)
amod(wound-3, sternal-2)
nsubj(developed-4, wound-3)
root(ROOT-0, developed-4)
amod(draiange-6, purulent-5)
dobj(developed-4, draiange-6)
cc(developed-4, and-8)
det(wound-10, the-9)

[3] *http://stanfordnlp.github.io/CoreNLP/*

nsubjpass(opened-12, wound-10)
auxpass(opened-12, was-11)
conj:and(developed-4, opened-12)
cc(opened-12, and-13)
det(dressing-16, a-14)
compound(dressing-16, vac-15)
nsubjpass(applied-18, dressing-16)
auxpass(applied-18, was-17)
conj:and(developed-4, applied-18)
conj:and(opened-12, applied-18)
advmod(well-21, there-19)
advmod(well-21, as-20)
advmod(applied-18, well-21)

Để dễ dàng hiểu các mối quan hệ ngữ pháp trong câu, những phụ thuộc được ánh xạ trên một đồ thị có hướng, trong đó các từ trong câu là các nút trên đồ thị và các mối quan hệ ngữ pháp là các nhãn cạnh, hình 4 biểu diễn đồ thị phụ thuộc cho câu ví dụ trên. Các mối quan hệ được định nghĩa trong [14], những định nghĩa sử dụng các nhãn từ loại (POS) và các nhãn cụm từ của Penn Treebank[4].



Hình 4. Biểu diễn đồ thị phụ thuộc

Rút trích trigger

Các thuộc tính liên quan đến bệnh/rối loạn thì mỗi thuộc tính có ý nghĩa khác nhau nên các *dấu hiệu nhận biết* (trigger) và các giá trị chuẩn hóa cho thuộc tính cũng khác nhau. Để thuận lợi cho quá trình xử lý mỗi thuộc tính được xây dựng một danh sách các dấu hiệu nhận biết và giá trị chuẩn hóa riêng biệt được trích xuất từ dữ liệu huấn luyện. Cấu trúc của mỗi danh sách gồm hai cột: cột đầu tiên chứa các dấu hiệu nhận biết và cột thứ hai chứa giá trị chuẩn hóa tương ứng. Ví dụ: bảng 3 minh họa cho danh sách dấu hiệu nhận biết và giá trị chuẩn hóa cho thuộc tính Course Class. Sau khi có được các danh sách dấu hiệu nhận biết của các thuộc tính, hệ thống tiếp tục làm giàu danh sách trigger từ các nguồn tài nguyên như: NegEx, WordNet, UMLS, … thông qua các nhóm từ đồng nghĩa tương ứng. Ví dụ: hệ thống sử dụng nguồn tài nguyên NegEx[5] để bổ sung những nhóm từ mang nghĩa phủ định cho thuộc tính Negation Indicator. Việc làm giàu danh sách trigger, với mong muốn sẽ xác định được nhiều trường hợp mà trong dữ luyện huấn luyện không có nhằm nâng cao hiệu quả cho hệ thống.

BẢNG 3
DANH SÁCH DẤU HIỆU NHẬN BIẾT GIÁ TRỊ THUỘC TÍNH THỨ 4 (COURSE CLASS)

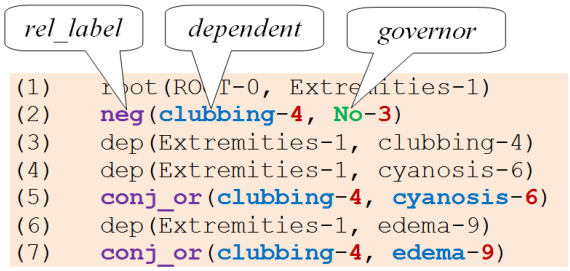
Dấu hiệu nhận biết	Giá trị chuẩn hóa
rapidly responded	improved
poorly controlled	worsened
getting better	improved
fluctuating up	increased
trending down	decreased
slow resolution	improved
…	…

Rút trích luật

Cơ sở để xây dựng tập luật là dựa trên mối quan hệ ngữ pháp giữa các từ trong câu. Bài toán xác định giá trị cho thuộc tính được chuyển về bài toán xác định mối quan hệ giữa *dấu hiệu* và *bệnh/rối loạn* sau khi có được danh sách *dấu hiệu* cho từng thuộc tính. Dựa trên những đặc trưng thể hiện của các nút trung gian giữa các từ thuộc *dấu hiệu* và *bệnh/rối loạn* trên đồ thị phụ thuộc để xây dựng luật thể hiện mối quan hệ, các đặc trưng được sử dụng bao gồm từ chính (*governor*), từ phụ thuộc (*dependent*) và mối quan hệ ngữ pháp (*nhãn cạnh – rel_label*) (xem hình 5). Các luật xây dựng dựa trên mối quan hệ ngữ pháp giữa các từ trong câu được chia là hai trường hợp: trường hợp 1 là các luật xác định mối quan hệ giữa *dấu hiệu* và *bệnh/rối loạn* và trường hợp 2 là các luật xác định mối quan hệ giữa *bệnh/rối loạn* và *bệnh/rối loạn*. Ví dụ: xét câu "*Extremities: No clubbing, cyanosis, or edema*.", trong câu này có 3 bệnh/rối loạn (*clubbing, cyanosis và edema*) có cùng *dấu hiệu* nhận biết *No* cho thuộc tính Negation Indicator. Dựa trên kết quả đầu ra của biểu diễn mối quan hệ ngữ pháp phụ thuộc trong hình 5 cho thấy bệnh/rối loạn *clubbing* có quan hệ trực tiếp với *dấu hiệu nhận biết No* ở dòng (2), còn hai bệnh/rối loạn *cyanosis và edema* có mối quan hệ với bệnh/rối loạn *clubbing* ở dòng (5) và (7).

[4] https://www.cis.upenn.edu/~treebank/

[5] https://healthinformatics.wikispaces.com/NegEx+Algorithm



Hình 5. Kết quả đầu ra phân tích phụ thuộc cho câu văn bản "*Extremities: No clubbing, cyanosis, or edema*."

Tập luật được hình thức hóa như sau:

Case 1: If *rel_label = "neg"* **and** dependent $\subseteq$ *{list of disorder}* **and** *governor* $\subseteq$ *{list of trigger}* **then 1 else 0**

Case 2: If *rel_label = "conj_or"* **and** dependent $\subseteq$ *{list of disorder}* **and** *governor* $\subseteq$ *{list of disorder}* **then 1 else 0**

Từ dữ liệu huấn luyện gồm 298 tài liệu (discharge summary, radiology report, ECHO report và ECG report) rút trích được 773 luật (xem bảng 4).

BẢNG 4
KẾT QUẢ TẬP LUẬT ĐƯỢC RÚT TRÍCH TỪ TẬP DỮ LIỆU HUẤN LUYỆN.

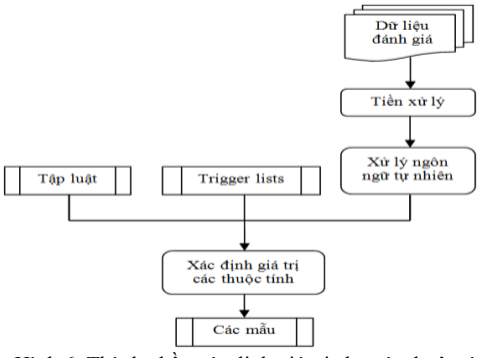
TT	Thuộc tính	Số luật/thuộc tính
1	Negation Indicator	131
2	Subject Class	16
3	Uncertainty Indicator	113
4	Course Class	120
5	Severity Class	84
6	Condition Class	108
7	Generic Class	21
8	Body Location	180
	Tổng số luật	**773**

Rút trích đặc trưng

Đối với thuộc tính DocTime Class, hệ thống sử dụng hướng tiếp cận lai ghép kết hợp giữa máy học và luật được chia làm hai giai đoạn. Giai đoạn 1: sử dụng phương pháp máy học để phân lớp giá trị cho thuộc tính. Giai đoạn 2: sử dụng phương pháp dựa trên luật để điều chỉnh lại kết quả của giai đoạn máy học. Đối với phương pháp máy học, vấn đề quan trọng là chọn ra tập đặc trưng để sử dụng huấn luyện mô hình phân lớp. Hệ thống sử dụng thuật toán SVM (Support vector machine) và dựa trên tập đặc trưng mà chúng tôi đã đề xuất trong [15] để tạo ra mô hình phân lớp.

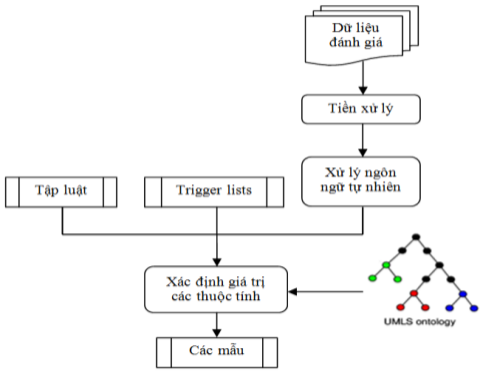
Xác định giá trị các thuộc tính

Đối với các thuộc tính Negation Indicator, Subject Class, Uncertainty Indication, Course Class, Condition Class, Generic Class và Temporal Expression có cùng cách thức xử lý và hệ thống tiến hành xử lý lần lượt trên từng thuộc tính. Đầu vào của thành phần này gồm Tập luật, các danh sách dấu hiệu nhận biết (Trigger lists) và tài liệu đánh giá đã qua các bước tiền xử lý và xử lý ngôn ngữ tự nhiên (xem hình 6).



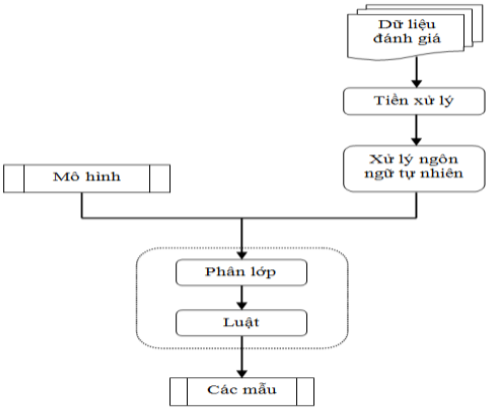
Hình 6. Thành phần xác định giá trị cho các thuộc tính.

Đối với thuộc tính Body Location. Ngoài phần xử lý giống như các thuộc tính trên, hệ thống có sử dụng thêm một nguồn tài nguyên UMLS để trong trường hợp Trigger list không có chứa trigger cần tìm thì hệ thống sẽ tìm trên UMLS (xem hình 7).



Hình 7. Thành phần xác định giá trị thuộc tính Body Location.

Đối với thuộc tính DocTime Class. Để xác định giá trị cho thuộc tính DocTime Class, hệ thống sử dụng hướng tiếp cận lai ghép giữa máy học và luật, xem hình 8. Trong thành phần Phân lớp, hệ thống sử dụng mô hình đã được huấn luyện để phân lớp giá trị cho thuộc tính trên tập tài liệu đánh giá. Sau đó, phân tích kết quả đầu ra của bước phân lớp máy học để tìm đặc trưng xây dựng luật nhằm cải hiện hiệu quả.



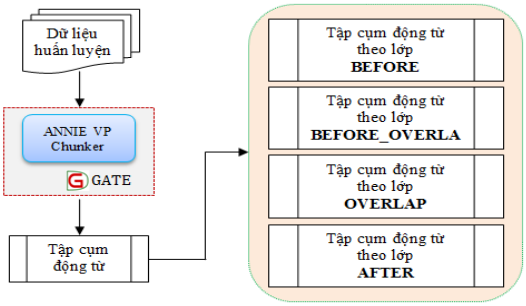
Hình 8. Thành phần xác định giá trị thuộc tính DocTime Class.

Ở bước xây dựng luật, tập luật được xây dựng dựa trên việc phân tích các đặc điểm của cụm động từ và đặc điểm của các từ trong câu có chứa bệnh/rối loạn.

Đặc điểm cụm động từ

Đặc điểm cụm động từ được xây dựng dựa trên một nhận định sau: "Mỗi một lớp sẽ có một tập các cụm động từ phổ biến thường xuyên đi kèm với nó." Ví dụ: lớp AFTER thường có các cụm động từ phổ biến đi kèm như: *be evaluated*, *please*, *recommended*, *to evaluate*, *to be removed*, *to follow*, *to arrange*, *to check*, *may want*, *to prevent*, *prescribed*, *should return*, v.v… Lớp BEFORE thường có các cụm động từ đi kèm như: *reported*, *was treated*, *had been removed*, *had reported*, …

Những cụm động từ phổ biến được xác định bằng cách tính trọng số (tf-idf) của mỗi cụm động từ. Các cụm động từ đại diện cho một lớp khi nó xuất hiện nhiều trong lớp này, nhưng xuất hiện ít ở lớp khác.



Hình 9. Xử lý cụm động từ trên tập dữ liệu huấn luyện.

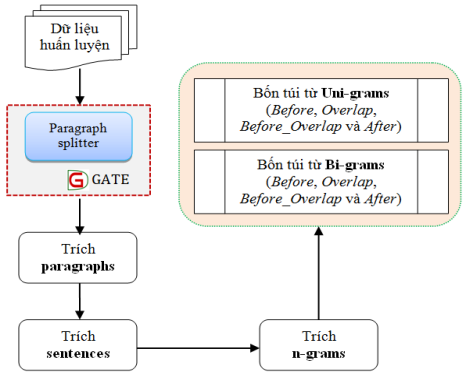
Quá trình thực hiện tính trọng số cho cụm động từ như sau: Từ dữ liệu huấn luyện, sử dụng Processing Resource ANNIE VP Chunker trong GATE để rút trích các cụm động từ trong các câu chứa bệnh/rối loạn. Sau đó, gom nhóm các cụm động từ theo các lớp (BEFORE, OVERLAP, BEFORE_OVERLAP và AFTER). Hệ thống tích hợp thư viện mở của Lucene6 để tính trọng số tf-idf, lập chỉ mục và quản lý tập cụm động từ phổ biến. Như vậy, sau giai đoạn này, sẽ có bốn tập cụm động từ đại diện cho bốn phân lớp nêu trên (xem hình 9).

Đối với dữ liệu đánh giá, hệ thống rút trích cụm động từ liên quan đến bệnh/rối loạn, cụm động từ này được so khớp với bốn tập phân lớp (BEFORE, OVERLAP, BEFORE_OVERLAP và AFTER) để tìm ra phân lớp mà có trọng số của cụm động từ cao nhất. Đầu ra của bước này là các cụm động từ với phân lớp tương ứng.

Đặc điểm n-grams

Đặc điểm n-grams được xây dựng dựa trên ý tưởng như sau: "Nếu một phân đoạn văn bản (paragraph) chứa phần lớn các bệnh/rối loạn thuộc một lớp, và chỉ một vài bệnh/rối loạn thuộc lớp khác, thì câu chứa một vài bệnh/rối loạn thuộc lớp khác có khả năng chứa các từ đặc biệt để phân lớp". Ví dụ: Phân đoạn văn bản với phần lớn các bệnh/rối loạn thuộc phân lớp BEFORE_OVERLAP, trong khi đó chỉ có một vài bệnh/rối loạn thuộc BEFORE, thì câu chứa bệnh/rối loạn này có thể chứa đựng các từ mang dấu hiệu đặc biệt cho lớp BEFORE.

Từ dữ liệu huấn luyện, mỗi tài liệu được tách thành các phân đoạn. Trong mỗi phân đoạn, thống kê số lượng bệnh/rối loạn theo các lớp và xác định được câu chứa bệnh/rối loạn thuộc lớp chiếm thiểu số trong phân đoạn văn đó. Trên các câu này rút trích các uni-gram và bi-gram, sau đó gom thành từng nhóm và tính tần số theo từng lớp tương ứng. Kết quả là các túi từ (bag-of-words) đại diện cho các lớp. Sau bước này, sẽ có bốn túi từ uni-gram cho bốn lớp và bốn túi từ bi-gram cho bốn lớp tương ứng (BEFORE, OVERLAP, BEFORE_OVERLAP và AFTER) (xem hình 10).



Hình 10. Xử lý n-gram trên dữ liệu huấn luyện.

[6] https://lucene.apache.org/

Đối với dữ liệu đánh giá, ứng với mỗi rối loạn, cũng trích ra uni-gram và bi-gram trên câu chứa rối loạn. Tiếp theo là tính toán xem các uni-gram và bi-gram này thuộc túi từ của lớp nào nhiều nhất. Sau quá trình này, tìm được lớp nào là khớp nhất theo uni-gram và bi-gram.

Rules

Luật được xây dựng dựa trên sự kết hợp giữa hai đặc điểm cụm động từ và n-gram. Gọi v là lớp có điểm số cao nhất theo đặc điểm cụm động từ; u và b là lớp có điểm số cao nhất theo đặc điểm uni-gram và bi-grams. Luật được phát biểu như sau:

Nếu v là BEFORE và (u hoặc b) là BEFORE thì kết luận là BEFORE.

Nếu v là BEFORE_OVERLAPS và (u hoặc b) là BEFORE_OVERLAPS thì kết luận là BEFORE_OVERLAPS.

Nếu v là OVERLAP và (u hoặc b) là OVERLAP thì kết luận là OVERLAP.

Nếu v là AFTER và (u hoặc b) là AFTER thì kết luận là AFTER

5.2 Phương pháp đánh giá kết quả

Hiệu quả của hướng tiếp cận được đánh giá dựa trên độ chính xác (accuracy) và độ hài hòa (F1-measure) như sau:

Dự đoán giá trị chuẩn hóa cho từng thuộc tính:

Phương pháp đánh giá: Accuracy (tổng thể và từng thuộc tính)

Accuracy (Acc) = Correct/Total.

Correct: Số lượng giá trị chuẩn hóa đúng.

Total: Tổng số giá trị.

Dự đoán dấu hiệu nhận diện giá trị thuộc tính:

Phương pháp đánh giá: F1-score (tổng thể và từng thuộc tính).

F1-score (F1) = (2 * R * P) / (R + P)

Recall (R) = TP / (TP + FN)

Precision (P) = TP / (TP + FP)

TP: Số lượng do hệ thống dự đoán đúng

FP: Số lượng do hệ thống dự đoán sai

FN: Số lượng mà hệ thống không dự đoán được.

5.3 Kết quả đánh giá

Tập dữ liệu do ShARe/CLEF eHealth 2014 cung cấp. Dữ liệu đánh giá gồm 133 tài liệu (discharge summary). Kết quả của hệ thống được đánh giá trên từng thuộc tính và tổng thể trong bảng 2. Trong dữ liệu đánh giá, riêng thuộc tính thứ 7 (GC) thì không có dữ liệu đánh giá nên độ chính xác là 100% và các độ đo còn lại là 0.

BẢNG 5
KẾT QUẢ CỦA HƯỚNG TIẾP CẬN

Thuộc tính	Acc	F1	P	R
NI	0.910	0.803	0.735	0.885
SC	0.995	0.736	0.760	0.713
UI	0.877	0.385	0.274	0.646
CC	0.937	0.410	0.317	0.577
SV	0.961	0.662	0.626	0.702
CO	0.899	0.441	0.340	0.625
GC	1.000	0.000	0.000	0.000
BL	0.551	0.330	0.309	0.354
DT	0.519	0.519	0.519	0.519
TE	0.830	0.313	0.337	0.292
Tất cả	**0.849**	**0.461**	**0.422**	**0.532**

Kết quả hướng tiếp cận đề xuất của chúng tôi (TeamHCMUS) được xếp thứ hai theo đánh giá hệ số tính đúng (accuracy) trong các nhóm tham gia, xem bảng 5.

BẢNG 6
KẾT QUẢ CỦA 10 NHÓM THAM GIA SHARE/CLEF EHEALTH 2014.

Các nhóm tham gia	Acc
TeamHITACHI	0.868
TeamHCMUS	0.849
RelAgent	0.843
DFKI-Medical	0.822
LIMSI	0.804
TeamUEvora	0.802
ASNLP CLEFeHealth2014 Text_result	0.793
TeamCORAL	0.790
TeamGRIUM	0.780
HPI_2a clefehealth2014 submission_29	0.769

Một số nhận xét về phương pháp mà 3 nhóm đầu tiên sử dụng. Nhóm HITACHI sử dụng hướng tiếp dựa trên máy học và luật cho chín thuộc tính, riêng thuộc tính DocTime chỉ dựa trên máy học, bước đầu tiên tác giả dùng phương pháp máy học và bước thứ hai họ sử dụng phương pháp luật để tinh chỉnh kết quả ở bước máy học cho nên hệ thống của họ đạt kết quả cao nhất (Accuracy là 0.868). Trong khi đó, nhóm HCMUS chỉ áp dụng phương pháp máy học và luật trên thuộc tính DocTime Class, những thuộc tính còn lại sử dụng phương pháp luật, kết quả accuracy của thuộc tính DocTime là 0.519 cao nhất trong tất cả các nhóm. Còn nhóm RelAgent sử dụng hướng tiếp cận hoàn toàn dựa trên luật và đạt kết quả thấp hơn hai nhóm đầu tiên (xem bảng 6). Như vậy xét về mặt phương pháp thì hướng tiếp cận lai ghép cho kết quả tốt hơn.

6 KẾT LUẬN

Trong bài báo trình bày các hướng tiếp cận cho bài toán rút trích mối quan hệ giữa các khái

niệm y tế từ những tài liệu lâm sàng trong lĩnh vực y tế và đề xuất một hướng tiếp cận rút trích mối quan hệ y tế trên bài toán cụ thể là xác định các giá trị cho các thuộc liên quan đến khái niệm y tế (điền mẫu). Đây là bài toán khá phức tạp đòi hỏi phải kết hợp nhiều kỹ thuật nhằm giải quyết bài toán. Hướng tiếp cận đã kết hợp các phương pháp dựa trên từ điển, dựa trên luật, biểu thức chính qui và máy học. Tập luật được xây dựng dựa trên mối quan hệ ngữ nghĩa phụ thuộc giữa khái niệm và dấu hiệu nhận biết giá trị, biểu thức thời gian.

Hiệu quả của hướng tiếp cận được nêu trên là nguồn động viên khích lệ cho chúng tôi khi tham gia cùng với cộng đồng quốc tế giải quyết bài toán có ý nghĩa thực tế cao. Đây cũng là lĩnh vực nghiên cứu khá mới mẽ ở Việt Nam. Tuy nhiên, hướng tiếp cận cũng cần tập trung nghiên cứu cải tiến hiệu quả cho những thuộc tính còn thấp như thuộc tính thứ tám (BL) và thứ chín (DT). Đây là hai thuộc tính khá phức tập trong quá trình xác định giá trị cho thuộc tính.

An approach in health relation extraction

Huynh Huu Nghia, Ho Bao Quoc, Nguyen An Te

***Abstract*—Extracting relations among medical concepts is very important in the medical field. The relations denote the events or the possible relations between the concepts. Information about these relations provides users with a full view of medical problems. This helps physicians and health-care practitioners make effective decisions and minimize errors in the treatment process. This paper collects methods for relations extraction in health texts and presents an approach on one type of specific relation (i.e. template filling). The approach combines methods including rule-based and machine learning-based. The rule-based method uses the relation of semantic dependencies among the concepts to extract the rule set. The machine learning-based method uses the SVM (Support Vector Machine) algorithm and a feature set proposed. The results of the approach were estimated on an accuracy of 0.849.**

***Keywords*—Relation extraction, information extraction, clinical information extraction, text mining.**

TÀI LIỆU THAM KHẢO

[1]. Aron Culotta and Jeffrey Sorensen. (2004). Dependency tree kernels for relation extraction. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pages 423–429.

[2]. Ben Abacha, Asma; Zweigenbaum, Pierre. (2011). Automatic extraction of semantic relations between medical entities: a rule based approach. In: J Biomed Semantics 2 Suppl 5, pp. S4. – URL http://dx.doi.org/10.1186/2041-1480-2-S5-S4.

[3]. Corney, David P A.; Buxton, Bernard F.; Langdon, William B.; Jones, David T.: BioRAT: extracting biological information from full-length papers. In: Bioinformatics 20 (2004), Nov, No. 17, pp. 3206–3213. – URL http://dx.doi.org/10.1093/bioinformatics/bth386.

[4]. Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. (February 2003). Kernel methods for relation extraction. Journal of Machine Learning Research, 3:1083–1106, February 2003.

[5]. Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In Proceedings of the 5th ACM Conference on Digital Libraries, pages 85–94, 2000.

[6]. Fundel, Katrin; Küffner, Robert; Zimmer, Ralf. (2007). RelEx–relation extraction using dependency parse trees. In: Bioinformatics 23, Feb, No. 3, pp. 365–371. – URL http://dx.doi.org/10.1093/bioinformatics/btl616.

[7]. Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pages 1247–1250.

[8]. Longhua Qian, Guodong Zhou, Fang Kong, Qiaoming Zhu, and Peide Qian. (2008). Exploiting constituent dependencies for tree kernelbased semantic relation extraction. In Proceedings of the 22nd International Conference on Computational Linguistics, pages 697–704.

[9]. Morante, R.; Daelemans, W. (2009). Learning the scope of hedge cues in biomedical texts. In: Workshop on BioNLP, pp. 28–36.

[10]. Min Zhang, Jie Zhang, and Jian Su. (2006). Exploring syntactic features for relation extraction using a convolution tree kernel. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 288–295.

[11]. Michael Collins and Nigel Duffy. (2001). Convolution kernels for natural language. In Advances in Neural Information Processing Systems 13.

[12]. Min Zhang, Jie Zhang, Jian Su, and GuoDong Zhou. (2006). A composite kernel to extract relations between entities with both flat and structured features. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pages 825–832.

[13]. Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. (2009). Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011.

[14]. Marie-Catherine de Marneffe and Christopher D. Manning. (2013). Stanford typed dependencies manual, September 2008, revised for the Stanford Parser v. 3.3 in December 2013.

[15]. Nghia Huynh and Quoc Ho. (2015). A Combined Approach for Disease/Disorder Template Filling. Proceedings: 2015 IEEE International Conference on Knowledge and Systems Engineering, pages 328–331. Ho Chi Minh City, Vietnam, October 2015. ISBN 978-1-4673-8013-3/15 $31.00 © 2015 IEEE DOI 10.1109/KSE.2015.62

[16]. Razvan Bunescu and Raymond Mooney. (2005). A shortest path dependency kernel for relation extraction. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, pages 724–731.

[17]. Razvan Bunescu and Raymond Mooney. (2006). Subsequence kernels for relation extraction. In Advances in Neural Information Processing Systems 18, pages 171–178.

[18]. Shubin Zhao and Ralph Grishman. (2005). Extracting relations with integrated information using kernel methods. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pages 419–426.

[19]. Sergey Brin. (1998). Extracting patterns and relations from the World Wide Web. In Proceedings of the 1998 International Workshop on the Web and Databases.

[20]. Yee Seng Chan and Dan Roth. (2010). Exploiting background knowledge for relation extraction. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 152–160.