Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00098

RÚT TRÍCH VÀ CHUẨN HÓA KHÁI NIỆM Y TẾ TRÊN TÀI LIỆU LÂM SÀNG

Huỳnh Hữu Nghĩa[1], Hồ Bảo Quốc[2]

[1]Bộ môn Tin học, Đại học Lao động - Xã hội, Thành phố Hồ Chí Minh

[2] Khoa CNTT, Đại học Khoa học Tự nhiên, Thành phố Hồ Chí Minh

huynhnghiavn@gmail.com, hbquoc@fit.hcmus.edu.vn

TÓM TẮT: *Các bác sĩ và những người chăm sóc sức khỏe thường xuyên có nhu cầu cập nhật thông tin từ những kết quả nghiên cứu khoa học và tham khảo những kết quả điều trị từ các tài liệu trong hồ sơ bệnh án điện tử. Bên cạnh đó, bệnh nhân (hoặc thân nhân) cũng có quyền hiểu biết về bệnh án của mình và nhiều kiến thức liên quan, sự hiểu biết này càng giúp cho quá trình điều trị được tốt hơn. Tuy nhiên, việc cập nhật và hiểu biết về tài liệu bệnh án đối với người dùng (bác sĩ, người chăm sóc sức khỏe, bệnh nhân,...) đang gặp vấn đề quá tải về số lượng tài liệu, tài liệu y tế được tạo ra hàng ngày và lưu trữ trong các hệ thống máy tính. Vì vậy, việc rút trích những thông tin có ích (khái niệm và mối quan hệ) từ tài liệu y tế cung cấp cho người dùng là một nhu cầu rất thiết thực. Tài liệu y tế được viết theo ngôn ngữ tự nhiên và trình bày theo dạng văn bản không có cấu trúc (hoặc bán cấu trúc), cho nên việc rút trích thông tin từ tài liệu là một thách thức đối với các hệ thống xử lý ngôn ngữ tự nhiên Thời gian gần đây, cộng đồng nghiên cứu đang tập trung giải quyết những bài toán liên quan đến rút trích thông tin y tế chẳng hạn như: rút trích khái niệm, mối quan hệ và sự kiện. Trong phạm vi bài báo này, nhóm tác giả trình bày một hướng tiếp cận rút trích và chuẩn hóa khái niệm y tế xuất hiện trong các tài liệu lâm sàng, trong đó nhóm tác giả đề xuất một bộ nhãn gán cho từng token, tập đặc trưng phân lớp nhãn token và xây dựng hệ thống rút trích khái niệm dựa trên thuật toán phân lớp học máy CRF (Conditional Random Fields). Hướng tiếp cận được sử dụng tham gia diễn đàn nghiên cứu SemEval 2015, kết quả do tổ chức SemEval đánh giá và công bố, hiệu quả của hệ thống đạt độ chính xác (precision) là 0.720, độ bao phủ (recall) là 0.690 và độ hài hòa (F-score) là 0.704.*

Từ khóa: *Rút trích thông tin lâm sàng; Rút trích khái niệm y tế; Xử lý ngôn ngữ tự nhiên.*

I. GIỚI THIỆU

Những khái niệm trong lĩnh vực y học thường đề cập đến các đối tượng như: *bệnh, rối loạn, thuốc (tên thuốc, liều lượng, phương thức quản lý, tần xuất quản lý ...), điều trị (thủ tục, biện pháp điều trị, thuốc điều trị ...), các vấn đề y tế, xét nghiệm, protein, di truyền (gien)* ... Các mối quan hệ cho biết sự liên quan giữa các khái niệm như: điều trị *giải quyết được* vấn đề y tế, điều trị *làm xấu đi* vấn đề y tế, xét nghiệm *phát hiện ra* vấn đề y tế… Việc nhận diện khái niệm là tiền đề để xác định mối quan hệ giữa chúng, các khái niệm và mối quan hệ có ý nghĩa rất quan trọng đối với người dùng trong lĩnh vực y tế như: bác sĩ, nhà nghiên cứu, sinh viên y khoa, nhân viên y tế, bệnh nhân cũng như thân nhân, ... kể cả ngoài lĩnh vực như: công ty bảo hiểm, … Một số trường hợp cụ thể cho thấy ý nghĩa của khái niệm và mối quan hệ đối với người dùng như sau: các bác sĩ muốn biết mối quan hệ giữa các khái niệm *điều trị* và *vấn đề y tế* để giúp họ đưa ra quyết định điều trị hiệu quả và hạn chế những sai sót, các nhà nghiên cứu muốn tìm hiểu về mối quan hệ giữa các khái niệm *di truyền* (*gien)* và *bệnh* nhằm giải thích những căn bệnh liên quan đến yếu tố di truyền, và còn rất nhiều trường hợp khác nữa.

Nhiều khái niệm và mối quan hệ đang nằm trong các dữ liệu y tế như: các tóm tắt xuất viện, các kết quả xét nghiệm, các công trình nghiên cứu khoa học… Những dữ liệu này được tạo ra liên tục hằng ngày và đang lưu trữ với nhiều dạng khác nhau như: âm thanh, hình ảnh và văn bản. Cụ thể, văn bản tường thuật (clinical narratives) chứa nhiều khái niệm đề cập đến các điều kiện lâm sàng, các vị trí giải phẫu trên cơ thể, các loại thuốc được sử dụng trong quá trình điều trị và những thủ tục (thủ thuật). Việc rút trích các khái niệm và mối quan hệ giữa chúng là cơ sở nền tảng để phát triển các ứng dụng như: tìm kiếm thông tin, hỏi đáp, tóm tắt văn bản và hệ thống hỗ trợ ra quyết định. Nhiều hình thức mặt chữ (surface forms) biểu diễn cùng khái niệm, cho nên việc rút trích và ánh xạ những khái niệm xuất hiện trong tài liệu văn bản đến những thuật ngữ đã được định nghĩa trong các từ vựng hoặc ontology (hay gọi là chuẩn hóa) nhằm giúp cho người dùng dễ dàng nhận biết và hiểu được các khái niệm và mối quan hệ một cách dễ dàng.

Trong lĩnh vực y học có nhiều nguồn tài nguyên từ vựng và ontology phong phú, có thể được tận dụng để nhận diện các khái niệm và liên kết các khái niệm hoặc chuẩn hóa. Một trong những nguồn tài nguyên đó là UMLS (Unified Medical Language System), nó chứa trên 130 từ vựng (lexicons/thesauri) với các thuật ngữ từ nhiều ngôn ngữ khác nhau, trong đó UMLS Metathesaurus tích hợp những nguồn tài nguyên chuẩn như: SNOMED-CT, ICD9 và RxNORM được sử dụng rộng rãi trên thế giới trong chăm sóc lâm sàng, y tế cộng đồng và dịch tễ học. Ngoài ra, UMLS cũng cung cấp một mạng ngữ nghĩa, trong đó mỗi khái niệm trong Metathesaurus được biểu diễn bởi một ký hiệu nhận dạng duy nhất khái niệm (CUI - Concept Unique Identifier) và được phân loại ngữ nghĩa [11].

Trong bài báo này, nhóm tác giả sẽ trình bày kết quả nghiên cứu rút trích và chuẩn hóa khái niệm xuất hiện trong các tài liệu văn bản lâm sàng. Nội dung trình bày bao gồm: mô tả bài toán thực hiện, tóm lược các phương pháp rút trích khái niệm, sau đó đề xuất một hướng tiếp cận rút trích khái niệm và đánh giá kết quả của hướng tiếp cận so với những nhóm khác cùng tham gia giải quyết bài toán.

II. BÀI TOÁN

Khái niệm trong lĩnh vực y học thì rất rộng cho nên việc rút trích khái niệm tổng quát rất khó thực hiện, vì lý do này mà cộng đồng nghiên cứu đã thực hiện rút trích từng loại khái niệm cụ thể như: thuốc, xét nghiệm, điều trị, vấn đề y tế, ... bài toán nghiên cứu trong bài báo này cũng chỉ thực hiện mục tiêu là rút trích một loại khái niệm cụ thể, khái niệm cụ thể đó là những bệnh/rối loạn (disease/desorder). Thách thức chính của bài toán là rút trích các khái niệm với những thể hiện khác nhau, một khái niệm có thể gồm những token liên tục và không liên tục hoặc lồng nhau. Ví dụ 1, xét câu văn bản "The rhythm appears to be *atrial fibrillation*." (*Nhịp tim chứng tỏ là rung tâm nhĩ*) khái niệm cần rút trích gồm những token liên tục là "*atrial fibrillation*" (*rung tâm nhĩ*). Ví dụ 2, xét câu văn bản "The *left atrium* is moderately *dilated*." (*Tâm nhĩ trái đã bị giãn vừa phải*) khái niệm cần rút trích gồm những token không liên tục là "*left atrium ... dilated*". Ví dụ 3, xét câu văn bản sau "*Abdomen: Soft, nontender, nondistended, normal active bowel sounds.*" câu trên có 2 khái niệm lồng nhau là "*Abdomen … nontender*" (*bụng cứng*) và "*Abdomen … nondistended*" (*bụng không bị sưng to*), trong đó có token chung là "*Abdomen*"; và xét câu văn bản sau "*The aortic root and ascending aorta are moderately dilated." (Căn nguyên động mạch chủ và động mạch chủ hướng thượng đã bị giản vừa phải)*, trong câu trên có hai khái niệm lồng nhau là "*aortic root … dilated*" *(căn nguyên động mạch chủ bị giản)* và "*ascending aorta … dilated*" *(động mạch chủ hướng thượng bị giản).* Qua những ví dụ trên có thể thấy việc rút trích khái niệm là một thách thức đối với các hệ thống xử lý ngôn ngữ tự nhiên, cũng như những hệ thống khai thác thông tin y tế.

Procedures performed: Right Heart Catheterization
Pericardiocentesis

Complications: None
Medications given during procedure: None
Hemodynamic data
Height (cm): 180 Weight (kg): 74.0
Body surface area (sq. m 93): 1. Hemoglobin (gm/dl):
Heart rate: 102

Pressure (mmHg)

	Sys	Dias	Mean	Sat
RA	14	13	8	
RV	36	9	12	
PA	44	23	33	62%
PCW	25	30	21	

Conclusions: Postoperative cardiac transplant
Abnormal hemodynamics
Pericardial effusion
Successful pericardiocentesis
General comments:
1600cc of serosanguinous fluid were drained from the pericardial sac with improvement in hemodynamics.

Hình 1. Một phần của báo cáo thông tin thuốc kích thích tim

Chuẩn hóa khái niệm là yêu cầu tiếp theo trong bài toán, nghĩa là khái niệm sau khi rút trích phải được ánh xạ đến các thuật ngữ đã được chuẩn hóa trong các từ vựng hoặc ontology thuộc lĩnh vực y học (UMLS). Ví dụ, khái niệm "*atrial fibrillation*" tương ứng với thuật ngữ "atrial fibrillation" được định nghĩa trong UMLS có mã số CUI là C0004238, khái niệm "*left atrium ... dilated*" tương ứng với thuật ngữ "left atrial dilatation" (*giãn tâm nhĩ trái*) được xác định trong UMLS với mã số CUI là C0344720.

Admit 10/23
71 yo woman h/o DM, HTN, Dilated CM/CHF, Afib s/p embolic event, chronic diarrhea, admitted with SOB. CXR pulm edema. Rx'd Lasix.
All: none
Meds Lasix 40mg IVP bid, ASA, Coumadin 5, Prinivil 10, glucophage 850 bid, glipizide 10 bid, immodium prn
Hospitalist = Smith PMD = Jones Full Code, Cx >101

Hình 2. Một mẫu ghi chú của bác sĩ nội trú

Việc rút trích khái niệm từ các tài liệu văn bản lâm sàng có nhiều thách thức đối với các hệ thống xử lý ngôn ngữ tự nhiên. Một số thách thức cụ thể như sau: *dữ liệu văn bản không có cấu trúc hoặc bán cấu trúc, nhiều ký tự/chữ viết tắt, thiếu dấu chấm câu (xem hình 1 và 2), lỗi chính tả, tính đồng nghĩa của các từ hoặc cụm từ, các cụm từ thường không đúng ngữ pháp, sự đa dạng về mặt từ vựng, hình thức thể hiện của khái niệm, chuẩn hoá khái niệm và các mối quan hệ phức tạp*. Trong những thách thức nêu trên, các tác giả chỉ nghiên cứu đề xuất một cách tiếp cận giải quyết thách thức gồm "*hình thức thể hiện của khái niệm và chuẩn hóa khái niệm*".

III. CÁC PHƯƠNG PHÁP RÚT TRÍCH KHÁI NIỆM

Để rút trích những khái niệm xuất hiện trong các tài liệu văn bản, tác giả đã nghiên cứu khảo sát tổng hợp những phương pháp áp dụng rút trích khái niệm từ văn bản trong lĩnh vực y tế được trình bày như sau:

A. Phương pháp dựa trên từ điển

Hướng tiếp cận nhận diện khái niệm dựa trên từ điển là sử dụng các từ điển thuộc lĩnh vực cụ thể có sẵn để xác định tên của các khái niệm xuất hiện trong tài liệu. Các từ điển được xây dựng bằng thủ công hay tự động từ các nguồn tài nguyên thuật ngữ tiêu biểu. Hướng tiếp cận cung cấp một số lợi ích như khả năng liên kết các mục từ (entries) cơ sở dữ liệu đến các đoạn thông tin (snippets) văn bản và chuẩn hóa các khái niệm ở cấp độ ngữ nghĩa và cú pháp từ vựng (lexicosyntactic). Tuy nhiên, hiệu quả của hướng tiếp cận dựa trên từ điển phụ thuộc rất nhiều vào tính toàn diện và rõ ràng của thông tin được cung cấp từ các nguồn thuật ngữ cơ bản.

Một số ứng dụng được công bố và thương mại thực hiện nhận diện thực thể dựa trên từ điển trong văn bản y khoa với các tỷ lệ thành công và khả năng khác nhau. Hệ thống EbiMed nhận diện các tên thuốc bằng cách sử dụng từ điển thuốc được xây dựng từ MedlinePlus. Tuy nhiên, hệ thống không cung cấp thông tin đánh giá hệ thống. Một số ứng dụng khác như ProMiner và Peregrine đã chứng minh thành công trong việc xác định một số phân loại các thực thể cho cả hai lĩnh vực y sinh học và lâm sàng như Genes, thuốc và bệnh. Riêng hệ thống nhận diện thực thể như MedLEE (Medical Language Extraction and Encoding System) và cTAKES (clinical Text Analysis and Knowledge Extraction

System) được áp dụng rất thành công trong các hệ thống lâm sàng đối với việc xác định và mã hóa thông tin liên quan đến bệnh nhân gồm các chẩn đoán, bệnh và điều trị.

B. Phương pháp dựa trên luật

Hướng tiếp cận dựa trên luật áp dụng các luật được tạo bằng thủ công để xác định các thực thể định danh trong văn bản. Các hệ thống bao gồm một tập luật mô tả các mẫu tạo thành thuật ngữ sử dụng các đặc trưng về ngữ pháp (ví dụ: từ loại, cú pháp, từ vựng, hình thái học và chính tả) cũng như tri thức lĩnh vực kết hợp với các từ điển. Nó dựa trên sự kết hợp các biểu thức chính quy, heuristics và các luật được tạo bằng thủ công. Tuy nhiên, việc xây dựng và duy trì tập luật thì chi phí cao và cần đội ngũ chuyên gia. Hơn nữa, các hệ thống rút trích khái niệm dựa trên luật thiếu đi khả năng thích ứng với các lĩnh vực khác, thường phụ thuộc vào nhiệm vụ và ngôn ngữ cụ thể.

Hamon và các cộng sự xây dựng hệ thống dựa trên luật ngôn ngữ học để phân tích các tài liệu lâm sàng nhằm rút trích các tên dược phẩm và thông tin liên quan đến dược phẩm [2]. Hệ thống cũng được phát triển để rút trích các dược phẩm không có trong từ điển. Sự thành công của hệ thống được đánh giá là phần đóng góp vào quá trình xử lý ngôn ngữ tự nhiên lâm sàng. Công trình [12] xây dựng tập luật (30 luật) để gán nhãn khái niệm liên quan đến các bệnh xuất hiện trong tài liệu lâm sàng.

C. Phương pháp dựa trên học máy

Hướng tiếp cận dựa trên học máy áp dụng các thuật toán học khác nhau để huấn luyện các mô hình thống kê nhằm giải quyết bài toán. Mô hình có thể được áp dụng để nhận diện và rút trích các thực thể trong các kho ngữ liệu khác nhau. Thế mạnh của hệ thống dựa trên học máy phụ thuộc vào chất lượng và khả năng phân biệt của các đặc trưng văn bản được áp dụng cũng như việc chọn thuật toán phân lớp [4]. Hướng tiếp cận học máy phát biểu bài toán nhận diện thực thể như một vấn đề tìm ra ranh giới và phân lớp văn bản. Hướng tiếp cận này cung cấp khả năng thích nghi tốt hơn cho các lĩnh vực khác nhau so với hướng tiếp cận dựa trên luật. Tuy nhiên, để huấn luyện một mô hình nhận diện thực thể chính xác yêu cầu phải có kho ngữ liệu được gán nhãn thủ công. Điều này có thể đòi hỏi nhiều chi phí và nhân công lao động.

Các kỹ thuật dựa trên học máy được áp dụng rất thành công trong việc nhận diện các tên gene [5] và tên hóa chất [6]. Công trình [3] đã phát triển một phần mềm mã nguồn mở (ChemicalTagger) nhằm xác định các tên hóa học cụ thể trong văn bản. Trong hướng tiếp cận dựa trên học máy, kỹ thuật CRF (Conditional Random Fields) là một trong những lựa chọn phù hợp và phổ biến nhất để áp dụng thành công cho bài toán nhận diện thực thể y sinh học. Các công trình [7, 8] đã phát triển các ứng dụng dựa trên CRF là BANNER và BioEnEx. Các ứng dụng này đã chứng minh sự thành công trong việc nhận diện thực thể y sinh học cũng như thực thể y khoa (chẳng hạn như bệnh).

D. Phương pháp lai ghép

Hướng tiếp cận lai ghép cho NER áp dụng sự kết hợp các kỹ thuật của các hướng tiếp cận dựa trên từ điển, dựa trên luật hoặc hệ thống dựa trên học máy. Một số hệ thống tiếp cận theo hướng lai ghép đã cho thấy những thành công ở các cấp độ khác nhau [13, 14]. Hướng tiếp cận này đang là xu hướng được các nhóm nghiên cứu vận dụng nhằm nâng cao hiệu quả cho hệ thống.

IV. ĐỀ XUẤT GIẢI QUYẾT BÀI TOÁN

Hướng tiếp cận giải quyết bài toán được nhóm tác giả sử dụng dựa trên phương pháp học máy, áp dụng thuật toán gán nhãn chuỗi tuần tự CRF (Conditional Random Fields). Thuật toán CRF sử dụng bộ nhãn BIO để gán nhãn cho thực thể, trong đó B (Begin) gán cho token bắt đầu, I (Inside) gán cho token bên trong và O (Outside) gán cho token bên ngoài thực thể. Tuy nhiên, các khái niệm cần rút trích được đề cập trong bài báo này thì có nhiều thể hiện khác nhau trong tài liệu văn bản y tế, một khái niệm có thể gồm những token liên tục và không liên tục hoặc lồng nhau nên không thể sử dụng bộ nhãn BIO cho bài toán rút trích khái niệm, vì vậy nhóm tác giả đã đề xuất bộ gán nhãn khái niệm phù hợp cho bài toán và tập đặc trưng dùng để phân lớp và đồng thời đề xuất một hệ thống rút trích và chuẩn hóa khái niệm được giới thiệu ở phần tiếp theo.

A. Bộ gán nhãn khái niệm

Bộ nhãn được nhóm tác giả đề xuất là BIEO, việc sử dụng bộ nhãn BIEO cho bài toán rút trích khái niệm được thực hiện như sau: nhãn B (Begin) được gán cho token bắt đầu khái niệm, nhãn I (Inside) gán cho các token bên trong khái niệm, nhãn E (End) gán cho token cuối cùng của khái niệm và nhãn O (Outside) gán cho các token không thuộc khái niệm.

Một số ví dụ minh hoạ việc sử dụng bộ nhãn, ví dụ 1: xét câu văn bản "*The rhythm appears to be atrial fibrillation*." trong câu trên có xuất hiện khái niệm "*atrial fibrillation*" gồm 2 token liên tục, nhãn B được gán cho token bắt đầu khái niệm, nhãn E gán cho token cuối cùng của khái niệm và nhãn O gán cho những token còn lại không thuộc khái niệm như sau: "*The*/O *rhythm*/O *appears*/O *to*/O *be*/O ***atrial*/B *fibrillation*/E** ./O"; ví dụ 2: xét câu văn bản "*The left atrium is moderately dilated."* trong câu trên có xuất hiện khái niệm "*left atrium ... dilated*" gồm các token không liên tục, việc gán nhãn được thực hiện gán nhãn B cho token đầu tiên của khái niệm, nhãn E gán cho token cuối

của khái niệm và nhãn I gán cho token bên trong khái niệm như sau: "*The*/O *left*/**B** *atrium*/**I** *is*/O *moderately*/O *dilated*/**E** ./O*"* và việc gán nhãn này cũng được sử dụng cho trường hợp tương tự khái niệm "*Heart ... irrgular ... rhythm*" thể hiện gồm 3 token rời nhau xuất hiện trong câu văn bản như: "***Heart*/B** **:/O** *VI*/**O** *systolic*/**O** *murmur*/**O** **,/O** ***irrgular*/I** *rate*/**O** *and*/**O** ***rhythm*/E** **./O***"*; ví dụ 3: xét câu văn bản sau "***Abdomen**: Soft, **nontender**, **nondistended**, normal active bowel sounds.*" câu trên có 2 khái niệm lồng nhau là "***Abdomen*** … ***nontender***" và "***Abdomen*** … ***nondistended***", trong đó token chung "***Abdomen***" được gán nhãn B và nhãn E được cho 2 token cuối của 2 khái niệm như sau: "***Abdomen*/B** **:/O** *Soft*/**O** **,/O** ***nontender*/E** **,/O** ***nondistended*/E** **,/O** *normal*/**O** *active*/**O** *bowel*/**O** *sounds*/**O** **./O***"*; và xét câu văn bản sau "*The* ***aortic root*** *and* ***ascending aorta*** *are moderately* ***dilated****."*, trong câu trên có hai khái niệm lồng nhau là "***aortic root*** … ***dilated***" và "***ascending aorta*** … ***dilated***", cho nên nhãn E được gán cho token "***dilated***" chung của 2 khái niệm và nhãn {B, I} gán cho các token bắt đầu của 2 khái niệm như sau: "*The*/**O** ***aortic*/B** ***root*/I** *and*/**O** ***ascending*/B** ***aorta*/I** *are*/**O** *moderately*/**O** ***dilated*/E** **./O***"*

Như vậy, tác giả đã đề xuất bộ nhãn **BIEO** dùng để gán nhãn khái niệm liên quan đến những thể hiện của các khái niệm nhằm mục tiêu rút trích những khái niệm xuất hiện trong tài liệu văn bản, khi đề xuất bộ nhãn nhóm tác giả cũng quan tâm đến tính khả thi trong phần hậu xử lý rút trích khái niệm, tức là loại nhãn token nào được ghép lại với nhau để hình thành nên khái niệm.

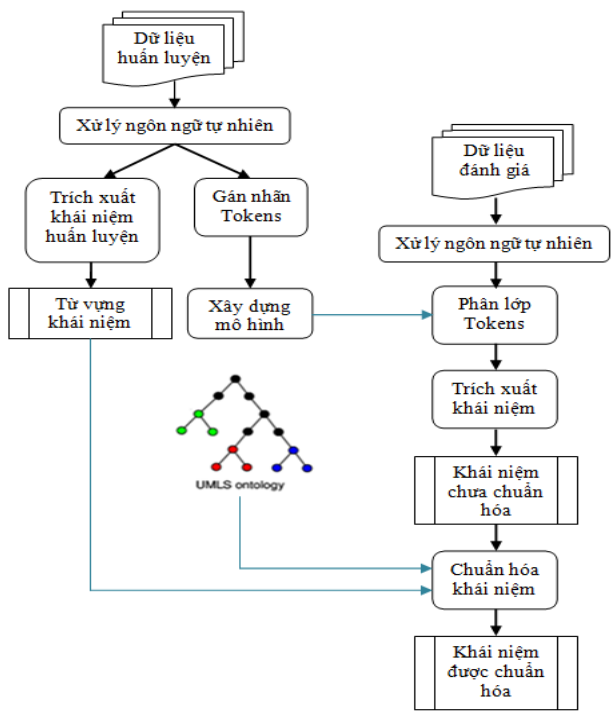
B. Tập đặc trưng phân lớp khái niệm

Trong phương pháp học máy, tập đặc trưng có vai trò quan trọng và ảnh hưởng đến hiệu quả của phương pháp, đặc trưng chính là đặc điểm để nhận diện và phân lớp, trong bài toán này chúng tôi đã nghiên cứu đặc điểm của dữ liệu y tế và đề xuất tập đặc trưng phù hợp dùng để phân lớp nhãn token cho bài toán rút trích khái niệm như sau:

- *Đặc trưng ngữ cảnh*: chỉ token hiện tại đang xét và hai token liền trước và liền sau của token đang xét. Các token xung quanh token đang xét đóng vai trò là thông tin ngữ cảnh.
- *Đặc trưng mặt chữ (Orthographic)*: token đang xét là chữ thường, in hoa, hoa ký tự đầu, …
- *Đặc trưng từ loại (Part of Speech)*: từ loại của token đang xét, các từ loại bao gồm *danh từ, động từ, tính từ, giới từ, trạng từ*, …
- *Đặc trưng thứ tự nhãn (label sequences)*: là thứ tự nhãn được gán cho từng token.

C. Hệ thống rút trích và chuẩn hóa khái niệm

Các đề xuất nêu trên được vận dụng để thiết kế hệ thống rút trích và chuẩn hoá khái niệm (xem hình 2), hệ thống được thiết kế chia là hai giai đoạn huấn luyện và rút trích chuẩn hóa



Hình 3. Kiến trúc hệ thống rút trích và chuẩn hóa khái niệm

1. Giai đoạn huấn luyện

Từ dữ liệu huấn luyện, qua thành phần xử lý ngôn ngữ tự nhiên, trích xuất danh sách khái niệm được chuyên gán nhãn, xây dựng mô hình phân lớp nhãn token dựa trên tập đặc trưng đã đề xuất; và (2) rút trích và chuẩn hoá khái niệm là bước thực hiện trích xuất khái niệm xuất hiện trong tài liệu y tế và ánh xạ những khái niệm đến các thuật ngữ đã được định trong các ontologies của lĩnh vực y tế. Chi tiết từng thành phần được trình bày phần sau.

a) Xử lý ngôn ngữ tự nhiên

Đầu vào của hệ thống là những tài liệu được viết bằng ngôn ngữ tự nhiên, dạng không có cấu trúc hoặc bán cấu trúc. Để máy tính có thể xử lý được thì cần có sự chuyển đổi tài liệu từ dạng không có cấu trúc hoặc bán cấu trúc về các dạng có cấu trúc như: tách câu, tách từ, chuẩn hóa từ, gán nhãn từ loại, …

Tách câu là công việc phân tách một tài liệu thành các câu cơ bản. Việc phát hiện ranh giới câu không phải là việc làm tầm thường khi dấu chấm câu "." không phải luôn luôn xuất hiện ở cuối câu. Trong tài liệu văn bản y tế chứa các thực thể định danh hoặc các khái niệm và các từ viết tắt có chứa các dấu câu như một thành phần thuật ngữ chuẩn chuyên ngành. Ví dụ, E. coli, W.H.O, … Việc nhận diện đúng các ranh giới câu là quan trọng cho các bài toán rút trích thông tin.

Tách từ là một quá trình phân chia một dòng văn bản (câu) thành các đơn vị nhỏ nhất được gọi là các token chẳng hạn như: các từ (word), dấu câu (punctuations) và dấu cách (separators). Tách token có thể thực hiện trực tiếp trên tài liệu hoặc các câu phức hợp cho kết quả là một trình tự các token. Quá trình này thường phụ thuộc vào các

heuristics đơn giản như sự phân cách của các token dựa trên các ký tự khoảng trắng chẳng hạn như các khoảng trống hoặc ngắt dòng và dấu câu.

Chuẩn hóa từ là quá trình làm giảm các từ biến tố về các dạng cơ bản của chúng. Chuẩn hóa có thể thực hiện thông qua xử lý từ gốc (stemming) hoặc dạng chính tắt của từ (lemmatization). Xử lý từ gốc là làm giảm các từ về từ gốc của nó như việc giảm các từ *increasing*, *increased* hay *increases* về từ gốc *increas*. Những chương trình thực hiện xử lý từ gốc được gọi là stemmer như: Porter stemmer và Snowball stemmer là các trường hợp phổ biến về những chương trình xử lý từ gốc đối với ngôn ngữ tiếng Anh. Xử lý chính tắt là quá trình giảm các từ về các dạng chính tắt của chúng như việc giảm các từ *increasing*, *increased* hay *increases* về từ gốc *increase*. Xử lý chính tắt đôi khi đòi hỏi công việc phức tạp chẳng hạn như sự hiểu biết bối cảnh hoặc gán nhãn từ loại. Xử lý chính tắt có liên quan chặt chẽ với xử lý từ gốc nhưng khác nhau ở bối cảnh xuất hiện của từ trong câu. Mục đích của chuẩn hóa từ là nhằm tăng hiệu quả cho hệ thống xử lý ngôn ngữ tự nhiên.

Gán nhãn từ loại là quá trình gán các từ trong một câu đến từ loại tương ứng dựa trên ngữ cảnh xuất hiện của nó. Ví dụ, gán nhãn từ loại có thể gán một từ đến một loại danh từ, động từ, … Gán nhãn từ loại không phải là công việc đơn giản vì một từ có thể có nhiều từ loại phụ thuộc vào ngữ cảnh xuất hiện. Trong trường hợp cụ thể, từ *antibiotic* xuất hiện như một danh từ trong cụm từ *antibiotic treats bacterial infection*, trong khi đó nó xuất hiện như một tính từ trong cụm từ *antibiotic agent*. Việc gán nhãn từ loại thường dựa trên các thuật toán học máy như mô hình Markov ẩn (Hidden Markov Models) được huấn luyện trên các kho ngữ liệu đã được gán nhãn từ loại bằng thủ công. Bên cạnh đó, việc gán nhãn từ loại dựa trên luật cũng được phát triển.

Hiện nay, một số thư viện mã nguồn mở hỗ trợ xử lý ngôn ngữ tự nhiên áp dụng cho văn bản tổng quát hay y sinh như: OpenNLP[1], Stanford NLP[2], MedPost[3], dTagger[4] hoặc GeniaSS[5]. Hệ thống đã tận dụng tích hợp một số thư viện có sẵn nêu trên nhằm hỗ trợ cho phần xử lý ngôn ngữ tự nhiên.

b) Trích xuất khái niệm huấn luyện

Trên dữ liệu huấn luyện, các khái niệm đã được gán nhãn và chuẩn hóa bởi chuyên gia, tác giả muốn sử dụng danh sách các khái niệm này như một từ vựng dùng để phục vụ giai đoạn chuẩn hóa khái niệm trong hệ thống. Cho nên, bước này hệ thống trích danh sách các khái niệm từ dữ liệu huấn luyện, danh sách khái niệm gồm hai cột: cột thứ nhất chứa giá trị chuẩn hóa là mã số định danh và cột thứ hai chứa các khái niệm tương ứng có trong dữ liệu huấn luyện.

c) Gán nhãn Tokens

Bước này, hệ thống sử dụng bộ nhãn BIEO đã đề xuất để gán nhãn token liên quan đến các khái niệm trong dữ liệu huấn luyện nhằm mục đích xây dựng mô hình phân lớp token. Thành phần gán nhãn tokens dựa trên yêu cầu như sau: xét trường hợp khái niệm gồm những token liên tục và không liên tục, token bắt đầu khái niệm thì gán nhãn B, token cuối của khái niệm gán nhãn E, token bên trong khái niệm thì gán I; xét trường hợp khái niệm lồng nhau có những token chung phía trước khái niệm, nếu token chung bắt đầu của 2 khái niệm thì gán nhãn B, token chung và bên trong của hai khái niệm thì gán nhãn I, token cuối của mỗi khái niệm được gán nhãn E; và xét trường hợp khái niệm lồng nhau có chung token phía cuối, nhãn B và I gán cho những token bắt đầu và bên trong của mỗi khái niệm, còn token cuối chung của 2 khái niệm được gán nhãn E. Cuối cùng, token không thuộc khái niệm thì gán nhãn O (xem ví dụ ở mục A phần IV).

d) Xây dựng mô hình phân lớp

Xây dựng mô hình là bước áp dụng thuật toán học máy để tạo ra mô hình học máy, mô hình sẽ học ngữ cảnh của tài liệu huấn luyện đã được gán nhãn khái niệm theo qui tắt được qui định bởi thuật toán học máy (cách thức học) và đặc điểm cần học (tập đặc trưng). Bước này còn được gọi là quá trình huấn luyện học máy, máy sẽ học cách thức phân lớp gán nhãn cho từ token trong tài liệu văn bản dựa trên thuật toán học máy và tập đặc trưng. Hệ thống đã tích hợp thuật toán CRFs được cung cấp bởi Stanford NLP dưới dạng là một thư viện mở và tập đặc trưng được đề xuất ở trên.

2. Rút trích và chuẩn hóa khái niệm

Giai đoạn này xử lý trên những tài liệu cần rút trích và chuẩn hóa khái niệm (dữ liệu đánh giá), quá trình xử lý thực hiện lần lượt các bước sau:

a) Xử lý ngôn ngữ tự nhiên

Chức năng xử lý tương tự như phần mô tả trong giai đoạn huấn luyện, kết quả đầu ra tập các câu trong mỗi tài liệu.

[1] https://opennlp.apache.org/
[2] https://stanfordnlp.github.io/CoreNLP/
[3] https://sourceforge.net/projects/medpost/
[4] https://specialist.nlm.nih.gov/dTagger/
[5] http://www.nactem.ac.uk/y-matsu/geniass/

b) Phân lớp tokens

Bước này áp dụng mô hình đã được tạo ra ở giai đoạn huấn luyện để phân lớp nhãn (BIEO) cho từng token trên tập dữ liệu cần trích xuất khái niệm. Kết quả đầu ra là tập dữ liệu với mỗi token được gán nhãn phân lớp.

c) Trích xuất khái niệm

Thành phần xử lý xét trên từng câu văn bản mà các tokens đã được gán nhãn. Thuật toán xử lý dựa trên các nhãn của token để xây dựng tổ hợp ghép các token lại để tạo thành các khái niệm cần rút trích. Việc ghép dựa trên qui tắt đã được đề xuất ở phần trên (phần IV, mục A, ví dụ). Kết quả đầu ra là danh sách các khái niệm (chưa chuẩn hóa) xuất hiện trong tài liệu.

d) Chuẩn hóa khái niệm

Chuẩn hóa là chức năng ánh xạ những khái niệm trích xuất từ tài liệu đến các thuật ngữ được định nghĩa trong các nguồn tài nguyên thuộc lĩnh vực y tế. Nguồn tài nguyên mà hệ thống sử dụng là UMLS, thuật toán xử lý được chia làm bước nhỏ, đầu tiên hệ thống sẽ tra cứu trong danh sách khái niệm đã xây dựng ở giai đoạn huấn luyện (phần IV, mục C, tiểu mục b) nếu không tìm thấy hệ thống sẽ tra cứu trong UMLS, trường hợp tìm thấy trùng khớp giữa khái niệm và thuật ngữ thì hệ thống sẽ gán mã số định danh tương ứng cho giá trị chuẩn hóa của khái niệm, ngược lại thì gán chuỗi "CUI-less" cho giá trị chuẩn hóa của khái niệm. Đầu ra của thành phần xử lý này là danh sách khái niệm xuất hiện trong từng tài liệu tương ứng đã được chuẩn hóa, kết quả này cũng là kết quả cuối cùng của hệ thống rút trích và chuẩn hóa khái niệm xuất hiện trong tài liệu văn bản lâm sàng.

V. ĐÁNH GIÁ KẾT QUẢ ĐỀ XUẤT

A. Dữ liệu thực nghiệm

Hệ thống được sử dụng tham gia diễn đàn nghiên cứu SemEval 2015, bộ dữ liệu sử dụng thực nghiệm do ban tổ chức diễn đàn cung cấp thông qua kho ngữ liệu ShARe, nó bao gồm 531 tài liệu lâm sàng được chọn từ cơ sở dữ liệu lâm sàng MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care II) phiên bản 2.5, trong đó 431 tài liệu sử dụng làm dữ liệu huấn luyện (data training) được gán nhãn khái niệm (*bệnh/rối loạn)* và 100 tài liệu dùng để đánh giá hệ thống (data test) chưa được gán nhãn khái niệm.

B. Đánh giá và kết quả

Ban tổ chức cũng đưa ra quy tắc đánh giá dựa trên độ đo hài hòa (F-score). Độ đo hài hòa là một trong những độ đo đánh giá được áp dụng rộng rãi cho hệ thống rút trích thông tin. Nó đo lường đầy đủ và toàn diện hệ thống. Độ đo hài hòa được tính bằng cách so sánh kết quả đầu ra của hệ thống đối với những phán đoán thủ công. Các yếu tố (chẳng hạn như: tài liệu hoặc khái niệm) được xác định một cách chính xác bởi hệ thống so với tiêu chuẩn vàng là "true positives". Các yếu tố được xác định bởi hệ thống nhưng không xuất hiện trong tiêu chuẩn vàng là "false positives". Trong khi đó, các yếu tố xuất hiện trong tiêu chuẩn vàng mà không được nhận diện bởi hệ thống là "false negatives". Bảng 1 cung cấp một cách nhìn tổng quát trên các độ đo cơ bản.

Bảng 1. Tổng quát trên các độ đo cơ bản cho các hệ thống rút trích thông tin.

		Tiêu chuẩn vàng (Gold Strandard)	
		Positive	Negative
Kết quả hệ thống	Positive |	True positive (TP)	False positive (FP)
	Negative |	False negative (FN)	Ture negative (TN)

Các độ đo cơ bản được sử dụng để xác định độ chính xác (precision) và độ bao phủ (recall) của hệ thống, nó được kết hợp một cách có hệ thống để sinh ra độ đo hài hòa cuối cùng. Độ chính xác đo sự chính xác của hệ thống bằng cách tính tỷ lệ chính xác các kết quả đầu ra giữa tất cả các kết quả được sinh bởi hệ thống. Độ chính xác được tính như sau:

$$Precision = \frac{TP}{TP + FP}$$

Độ bao phủ đo tính đầy đủ của hệ thống bằng cách tính tỷ lệ đúng kết quả đầu ra của hệ thống trong việc so sánh dựa trên sự chính xác với tiêu chuẩn vàng. Độ bao phủ được tính như sau:

$$Recall = \frac{TP}{TP + FN}$$

Độ đo hài hòa tính độ hài hòa giữa độ chính xác và độ bao phủ, công thức tính như sau:

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Độ chính xác, độ bao phủ và độ hài hòa có giá trị từ 0 đến 1, trong đó 1 cho biết là tốt nhất và 0 cho biết là tệ nhất.

Yêu cầu đánh giá trong bài toán này được tính như sau: khi hệ thống rút trích khái niệm và chuẩn hóa trùng khớp với tập đánh giá thì mới được tính đúng, còn nếu khái niệm rút trích bị sai hoặc giá trị chuẩn hóa bị sai so với tập đánh giá thì xem là không đúng. Tất cả các kết quả của những hệ thống tham gia đều gửi về cho ban tổ chức SemEval 2015 đánh giá và công bố kết quả. Kết quả đánh giá hệ thống của tác giả (HCMUS) đạt như sau: Độ chính xác (precision) là 0.720, độ bao phủ (recall) là 0.732 và độ hài hòa (F-score) là 0.757 [9].

C. Bàn luận kết quả

Trong phần này, các tác giả xin phép sử dụng kết quả của nhóm có hiệu quả cao nhất (ezDL) [10] và nhóm có kết quả ở nhóm khá (LIST-LUX) [1] để bàn luận về những đề xuất của mình như sau:

1. Rút trích khái niệm

Xét hướng tiếp cận giữa HCMUS và LIST_LUX: Hướng tiếp cận giữa hai hệ thống nhìn chung có nhiều điểm tương đồng chẳng hạn như: cả hai đều dựa trên phương pháp học máy, sử dụng thuật toán phân lớp CRFs và tập đặc trưng. Tuy nhiên, khi xem cụ thể từng phần thì giữa hai hệ thống có một vài điểm khác biệt như: bộ nhãn phân lớp token, HCMUS đề xuất bộ nhãn BIEO và LIST-LUX đề xuất bộ nhãn BIESTO (nhãn B gán cho token bắt đầu khái niệm, I gán cho token bên trong, E token cuối cùng, T gán cho những token ở giữa nhưng không thuộc khái niệm và khái niệm chỉ có 1 token thì gán nhãn S), HCMUS có sử dụng một đặc trưng "*đặc trưng thứ tự nhãn*" không trùng so với LIST-LUX. Có thể do những khác biệt nêu trên mà làm cho kết quả của HCMUS tốt hơn LIST-LUX.

Xét hướng tiếp cận giữa HCMUS và ezDL: Cả hai hệ thống cùng tiếp cận theo phương pháp học máy, tuy nhiên khi áp dụng thì lại khác nhau. HCMUS sử dụng thuật toán phân lớp CRFs, bộ nhãn phân lớp token (BIEO) và tập đặc trưng đề xuất thực hiện một lần để xác định tất cả các thể hiện của khái niệm trong tài liệu văn bản. Còn ezDL thực hiện xác định tất cả những thể hiện của khái niệm được chia làm hai giai đoạn: (1) ezDL sử dụng thuật CRFs, bộ nhãn truyền thống (BIO) và tập đặc trưng (*Túi từ, xử lý từ gốc, tiền tố và hậu tố có chiều dài 1 đến 5; Mặt chữ như: từ chứa số, dấu gạch chéo, ký tự đặc biệt; Ngữ pháp như: nhãn từ loại (PoS), cụm từ (chunk) và các cụm động từ và danh từ chính; Tìm kiếm từ điển so khớp với của sổ +2 đến -2; Tiêu đề mục, loại tài liệu và gom cụm câu*) để xác định những khái niệm thể hiện gồm các token liên tục, (2) ezDL sử dụng thuật toán phân lớp SVM của thư viện LibSVM để tìm ra có sự tồn tại mối quan hệ giữa hai khái niệm hay không với tập đặc trưng (*Túi từ, nhãn từ loại và các nhãn cụm từ của tất cả các token xuất hiện giữa hai khái niệm; Một số luật thể hiện mối quan hệ giữa hai khái niệm; Vị trí của giới từ, liên từ, động từ chính và các ký tự đặc biệt như dấu (:), gạch nối (-) và dấu chấm phẩy (;) trong nội dung của khái niệm đầu tiên*) để xác định những khái niệm thể hiện gồm các token không liên tục. Như vậy, có thể thấy rằng phương pháp của ezDL phức tạp hơn HCMUS và đạt kết quả tốt hơn.

2. Chuẩn hoá khái niệm

Xét phương pháp giữa HCMUS và LIST-LUX: HCMUS thực hiện gồm 2 giai đoạn: (1) Hệ thống so khớp khái niệm trong danh sách khái niệm được xây dựng từ tập dữ liệu huấn luyện (xem phần IV - C - 1 - b), nếu tìm thấy sẽ lấy mã số định danh tương ứng để gán cho khái niệm, ngược lại thực hiện giai đoạn (2) Hệ thống tích hợp thư viện MetaMap làm lớp trung gian để tìm kiếm trong UMLS, kết quả của MetaMap trả về với danh sách các khái niệm trùng khớp với độ tương đồng tương ứng, nếu khái niệm và thuật ngữ có độ tương đồng cao nhất trên 90% thì được chọn làm giá trị chuẩn hóa, ngược lại giá trị chuẩn hóa là "CUI-less". Đối với LIST-LUX sử dụng SQL truy vấn 11 nhóm ngữ nghĩa liên quan đến các rối loạn (disorders) từ nguồn tài nguyên SNOMED để xây dựng một danh sách thuật ngữ riêng với tổng cộng 348760 dòng; xây dựng thuật toán so sánh chuỗi giữa khái niệm được trích từ văn bản và các thuật ngữ trong danh sách được xây dựng; trong trường hợp không tìm ra được sự trùng khớp chính xác, sử dụng độ đo tương đồng (similarity) để tính mức độ liên quan nhất giữa khái niệm và thuật ngữ; và nếu độ tương đồng cao nhất trên ngưỡng 0.8 (được thiết lập cố định) thì được chọn là giá trị chuẩn hóa, ngược lại thì giá trị chuẩn hóa là "CUI-less".

Xét phương pháp giữa HCMUS và ezDL: ezDL tiến hành 3 bước: (1) Tìm kiếm trực tiếp trên từ điển: mỗi từ (word) trong khái niệm, họ tìm tất cả các biến thể của nó trong từ LVG[6], sau đó thực hiện tất cả các hoán vị để tìm kiếm chuỗi trong UMLS, nếu chuỗi được tìm thấy bất kỳ một khái niệm trong UMLS thì họ liên kết đến mã số CUI tương ứng với khái niệm. (2) Tìm kiếm từ điển dựa trên những khái niệm biến đổi: Thực hiện (bán tự động) tách chuỗi dựa trên các từ chức năng như giới từ và các nhóm từ (động từ, tính từ và trạng từ). Họ cho rằng hầu hết các khái niệm (rối loạn) có thể được chứa trong cụm danh từ đơn (NP) và cũng có khá nhiều chứa cụm danh từ có liên quan với các cụm giới từ (PPs), cụm động từ (VPs) và cụm tính từ (ADJPs). Họ kết hợp với UMLS để xây dựng một từ điển khái niệm biến đổi. (3) Thuật toán so chuỗi tương đồng: Nếu một khái niệm không tìm thấy qua hai bước trên, thì họ tạo ra danh sách các chuỗi con từ UMLS. Sau đó, họ sử dụng thuật toán LED (Levenshtein Edit Distance) để tìm kiếm chuỗi tốt nhất. Nếu chuỗi tìm kiếm tốt nhất lớn hơn giá trị ngưỡng xác định thì lấy mã số CUI tương ứng, ngược lại thì lấy chuỗi "CUI-less".

[6] http://lexsrv2.nlm.nih.gov/

Kết quả rút trích và chuẩn hóa khái niệm của 3 nhóm liên quan được trình bày trong bảng 2, với kết quả này cho thầy hướng tiếp cận đề xuất của các tác giả cũng mang đến hiệu quả nhất định.

Bảng 2. Kết quả đánh giá của ezDL, HCMUS và LIST-LUX.

	Nghiêm ngặt			**Không nghiêm ngặt**		
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*
ezDl	0.783	0.732	0.757	0.815	0.761	0.787
HCMUS	**0.720**	**0.690**	**0.704**	**0.782**	**0.720**	**0.749**
LIST-LUX	0.649	0.580	0.613	0.675	0.603	0.637

Các tác giả cũng đã phân tích trên dữ liệu huấn luyện và đánh giá thấy rằng chữ viết tắt trong tài liệu lâm sàng là một trong những nguyên nhân ảnh hưởng đến hiệu quả của hướng tiếp cận ở phần chuẩn hóa khái niệm, một số ví dụ minh họa được trình bày trong bảng 3 và bảng 4 như sau:

Bảng 3. Cùng khái niệm (viết tắt) nhưng có các mã số định danh khác nhau.

Mã số định danh	Chữ viết tắt
C0004247	A-V
C0729580	A-V
C1283343	LAD
C0226032	LAD

Bảng 4. Cùng mã số định danh nhưng các khái niệm (viết tắt) khác nhau.

Mã số định danh	Chữ viết tắt
C0007226	CV
C0007226	CVS
C0447720	L4 on L5
C0447720	L4-L5

VI. KẾT LUẬN

Trong bài báo đã trình bày một hướng tiếp cận dựa trên học máy nhằm rút trích các khái niệm xuất hiện trong các tài liệu văn bản y tế. Các tác giả có một số đề xuất như: bộ nhãn BIEO dùng để gán nhãn phân lớp token nhằm xác định khái niệm, tập đặc trưng phân lớp và hệ thống rút trích và chuẩn hóa khái niệm y tế. Tuy kết quả chưa phải là tốt nhất nhưng nó cũng góp phần làm phong phú các phương pháp rút trích và chuẩn hóa khái niệm, với kết quả đạt được là nguồn động viên khích lệ chúng tôi có những nghiên cứu sâu hơn trong tương lai, nhằm tìm kiếm những giải pháp nâng cao hiệu quả cho bài toán rút trích và chuẩn hóa khái niệm y tế.

TÀI LIỆU THAM KHẢO

[1] Asma Ben Abacha, Aikaterini Karanasiou, Yassine Mrabet, and Julio Cesar Dos Reis. LIST-LUX: Disorder Identification from Clinical Texts. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 427-432, Denver, Colorado, June 4-5, 2015. Association for Computational Linguistics.

[2] Hamon, Thierry; Grabar, Natalia: Linguistic approach for identification of medication names and related information in clinical narratives. *In: J Am Med Inform Assoc 17 (2010)*, No. 5, pp. 549-554. - URL http://dx.doi.org/10.1136/jamia.2010.004036.[6]

[3] Hawizy, Lezan; Jessop, David M.; Adams, Nico; Murray-Rust, Peter: ChemicalTagger: A tool for semantic text-mining in chemistry. *In: J Cheminform 3 (2011)*, No. 1, pp. 17. - URL http://dx.doi.org/10.1186/1758-2946-3-17.[7]

[4] Krauthammer, Michael; Nenadic, Goran: Term identification in the biomedical literature. *In: J Biomed Inform 37 (2004)*, Dec, No. 6, pp. 512-526. - URL http://dx.doi.org/10.1016/j.jbi.2004.08.004. [9]

[5] Klinger, Roman; Friedrich, Christoph M.; Mevissen, Heinz T.; Fluck, Juliane; Hofmann-Apitius, Martin; Furlong, Laura I.; Sanz, Ferran: Identifying gene-specific variations in biomedical text. *In: J Bioinform Comput Biol 5 (2007)*, Dec, No. 6, pp. 1277-1296. [10]

[6] Klinger, Roman; Kolárik, Corinna; Fluck, Juliane; Hofmann-Apitius, Martin; Friedrich, Christoph M.: Detection of IUPAC and IUPAC-like chemical names. *In: Bioinformatics 24 (2008)*, Jul, No. 13, pp. i268-i276. - URLhttp://dx.doi.org/10.1093/bioinformatics/btn181. [11]

[7] Leaman, Robert; Gonzalez, Graciela: BANNER: an executable survey of advances in biomedical named entity recognition. *In: Pac Symp Biocomput (2008)*, pp. 652-663.[12]

[8] Mahbub Chowdhury, Faisal; Lavelli, Alberto: Disease Mention Recognition with Specific Features. *In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP)*, ACL 2010, pp. 83-90. [13]

[9] Nghia Huynh, Quoc Ho. TeamHCMUS: Analysis of Clinical Text. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 370-374, Denver, Colorado, June 4-5, 2015. c2015 Association for Computational Linguistics.

[10] Parth Pathak, Pinal Patel, Vishal Panchal, Sagar Soni, Kinjal Dani, Narayan Choudhary, Amrish Patel. ezDI: A Supervised NLP System for Clinical Narrative Analysis. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 412-416, Denver, Colorado, June 4-5, 2015. Association for Computational Linguistics.

[11] Olivier Bodenreider and Alexa T McCray. 2003. Exploring semantic groups through visual approaches. Journal of biomedical informatics, 36(6):414-432. [M1]

[12] Ramanan S. V., Shereen Broido, and Senthil Nathan P. Performance of a multi-class biomedical tagger on clinical records. *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*. [18]

[13] Tikk, Domonkos; Solt, Illés: Improving textual medication extraction using combined conditional random fields and rule-based systems. *In: J Am Med Inform Assoc 17 (2010)*, No. 5, pp. 540-544. - URL http://dx.doi.org/10.1136/jamia.2010.004119.[20]

[14] Yaoyun Zhang, Jingqi Wang, Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, and Hua Xu. UTH_CCB: A Report for SemEval 2014 - Task 7 Analysis of Clinical Text. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 802-806, Dublin, Ireland, August 23-24, 2014.[21]

MEDICAL CONCEPTS EXTRACTION AND NORMALIZATION ON CLINICAL DOCUMENTS

Nghia Huynh, Quoc Ho

ABSTRACT: *Doctors and health care providers usually need to keep up to update information from science articles and refer to the treatment results from documents in electronic medical records. In addition, patients (or relatives) have the right to know their medical records and relevant knowledge, which makes the treatment process to be better. However, the users' updating and understanding of medical records (doctors, health care providers, patients,...) are experiencing overload problems of the numbers of documents and medical documents created daily and stored in computer systems. Therefore, extracting useful information (concepts and relationships) from medical documents provided to users is a very practical need. Medical documents are written in the natural language and presented in the unstructured (or semi-structured) text. Thus, extracting information from the documents is a challenge for natural language processing systems. In recent times, the research community is focusing on solving problems related to the extraction of medical information, such as the extraction of concepts, relations and events. In this article, the authors present an approach to the extraction and normalization of medical concepts that appear in clinical documents in which the authors propose a set of tags for each token; a set of features that is used to classify tokens; and build a conceptual extraction system based on Conditional Random Fields (CRF) algorithms. The approach was used to participate in the SemEval 2015 research forum. The assessment and publication of results were conducted by the SemEval organization. The efficiency of the system achieved the precision of 0.720, the recall of 0.690 and the F-score of 0.704.*

DOI: 10.15625/vap.2017.00098