Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00046

MỘT HƯỚNG TIẾP CẬN LAI GHÉP PHÂN LỚP MỐI QUAN HỆ GIỮA BỆNH VÀ THỜI GIAN VIẾT TÀI LIỆU LÂM SÀNG

Huỳnh Hữu Nghĩa, Vũ Sơn Lâm, Hồ Bảo Quốc
Khoa CNTT, Đại học Khoa học Tự nhiên, Thành phố Hồ Chí Minh
huynhnghiavn@gmail.com, lamvuson@gmail.com, hbquoc@fit.hcmus.edu.vn

TÓM TẮT *— Khi nghiên cứu tài liệu lâm sàng, các bác sĩ, các nhà nghiên cứu hay những người chăm sóc bệnh nhân muốn biết một bệnh/rối loạn xảy ra vào điểm nào (quá khứ, hiện tại, tương lai hoặc kéo dài từ quá khứ đến hiện tại,…) so với thời điểm tài liệu được viết. Những thông tin về thời gian này rất hữu ích trong việc xây dựng phác đồ điều trị cho bệnh nhân, xây dựng hệ thống hỏi đáp, tóm tắt tài liệu. Bài báo đề xuất một hướng tiếp cận lai ghép giữa luật và máy học để phân lớp mối quan hệ giữa bệnh/rối loạn và thời gian viết tài liệu lâm sàng, kết quả hướng tiếp cận đạt được độ chính xác (accuracy) là 0.5194 cao hơn hệ thống được xếp hạng nhất (0.328) trong ShARe/CLEF eHealth Evaluation Lab 2014.*

Từ khóa *— Rút trích thông tin lâm sàng; Rút trích mối quan hệ thời gian; Xử lý ngôn ngữ tự nhiên.*

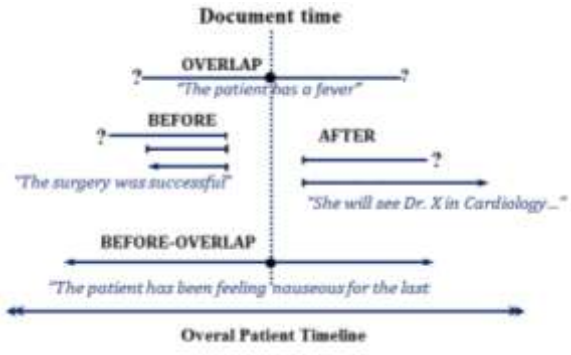
I. GIỚI THIỆU

Những tài liệu lâm sàng (clinical documents) như tóm tắt xuất viện (discharge summary), các báo cáo xét nghiệm (x-quang, siêu âm, điện tim) được viết bởi các y tá, bác sĩ hay những người chăm sóc bệnh nhân nhằm ghi lại những thông tin quan trọng trong quá trình điều trị của bệnh nhân. Đặc biệt là các tóm tắt xuất viện, nó mô tả quá trình điều trị, tình trạng bệnh nhân và kế hoạch chăm sóc. Mục đích chính của nó là hỗ trợ quá trình chăm sóc bệnh nhân cũng như là những ghi chú bàn giao giữa các bác sĩ [1]. Cùng với sự phát triển của công nghệ thông tin, các tài liệu y khoa này dần được số hóa, nguồn dữ liệu này ngày càng lớn và chứa đựng rất nhiều thông tin có giá trị. Việc rút trích thông tin cần thiết từ tài liệu lâm sàng đang được cộng đồng nghiên cứu rất quan tâm thông qua các tổ chức nghiên cứu như: I2B2[1] (Informatics for Integrating Biology and Bedside) và ShARe/CLEF eHealth[2].

Rút trích thông tin thời gian (Temporal Information Extraction – TIE) là một thách thức lớn trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) và nó là một thành phần quan trọng trong nhiều hệ thống NLP, chẳng hạn như hệ thống Hỏi – Đáp (Question – Answering), tóm tắt tài liệu (Document Summarization), dịch máy (Machine Translation) [5]. Trong lĩnh vực y khoa, việc rút trích thông tin thời gian có thể được ứng dụng để xây dựng biểu đồ thông tin quá trình điều trị bệnh nhân (timeline) hoặc tạo ra các tóm tắt điều trị, nó cũng có thể được áp dụng trong các hệ thống suy luận từ việc khai thác dữ liệu y khoa nhằm tìm ra các thông tin hữu ích, từ đó nâng cao hiệu quả điều trị bệnh nhân cũng như phục vụ cho các công tác nghiên cứu xa hơn.

Một hệ thống rút trích thông tin thời gian trong y khoa thường bao gồm các thành phần:

- Nhận diện các sự kiện (các rối loạn, các điều trị…)
- Nhận diện các biểu thức thời gian (Temporal expression)
- Phân lớp mối quan hệ giữa thời gian và các sự kiện.



Hình 1. Lược đồ thể hiện tất cả khả năng mà Bệnh/rối loạn có thể được phân lớp so với thời gian viết tài liệu (Document time).

Rút trích mối quan hệ thời gian (Temporal relation) thường bao gồm: quan hệ thời gian giữa sự kiện và sự kiện, quan hệ thời gian giữa sự kiện và các mốc thời gian, quan hệ thời gian giữa sự kiện và thời gian viết tài liệu, trong đó, loại quan hệ thứ ba giữ một vai trò khá quan trọng. Chẳng hạn, khi nghiên cứu tài liệu lâm sàng, các bác sĩ, các nhà nghiên cứu hay những người chăm sóc bệnh nhân muốn biết nhanh một rối loạn (disorder) xảy ra vào thời gian nào

[1] https://www.i2b2.org/NLP/HeartDisease
[2] http://clefehealth2014.dcu.ie/

(quá khứ, hiện tại, tương lai hoặc kéo dài từ quá khứ đến hiện tại,…) so với thời điểm tài liệu được viết. Điều này đòi hỏi những người nghiên cứu phải tìm ra giải pháp để rút trích mối quan hệ thời gian giữa rối loạn và thời gian viết tài liệu (Document time).

Theo [8] cho thấy thời gian viết tài liệu này được cho là tương đương (về mặt chức năng) với thời gian mà bệnh nhân gặp bác sĩ hoặc đến bệnh viện. Các giá trị phân lớp thời gian thể hiện mối quan hệ về mặt thời gian giữa bệnh/rối loạn và thời gian tài liệu lâm sàng được tạo ra [4]. Các giá trị phân lớp thời gian gồm có 5 giá trị: BEFORE, OVERLAP, BEFORE_OVERLAP, AFTER và UNKNOWN (giá trị mặc định). Hình 1 thể hiện các giá trị phân lớp thời gian so với thời gian tài liệu được viết như sau:

- *BEFORE* được sử dụng khi bệnh/rối loạn đã hết trước khi bệnh nhân gặp bác sĩ (tức là trước thời gian tài liệu được viết). Ví dụ: xét câu "Patient had tumor removed." bệnh/rối loạn "tumor" được gán nhãn phân lớp là "BEFORE" do động từ "removed" ở thì quá khứ.
- *OVERLAP* được sử dụng khi bệnh/rối loạn hoặc tình trạng đang xảy ra hoặc đúng ngay thời điểm gặp bệnh nhân và cùng lúc tài liệu được viết. Ví dụ: xét câu "These findings could represent ileus or early small bowel obstruction." Cả hai bệnh/rối loạn "ileus" và "small bowel obstruction" có thể được gán nhãn phân lớp "OVERLAP" bởi vì thì hiện tại đơn của động từ "represent" cho biết điều đó.
- *BEFORE-OVERLAP* được sử dụng khi một bệnh/rối loạn đã xuất hiện trước thời điểm viết tài liệu và còn tiếp tục trong hiện tại. Đơn giản chỉ cần, khi bệnh/rối loạn đã bắt đầu trước cuộc kiểm tra hoặc bệnh nhân viếng thăm và tiếp tục cho đến hiện tại, và thường (không phải luôn luôn) tương ứng với việc sử dụng thì hiện tại hoàn thành trong tiếng Anh. Ví dụ: xem câu "Patient has had a tumor for the past two months." Bệnh/rối loạn "umor" được gán nhãn phân lớp là "BEFORE-OVERLAP".
- *AFTER* được sử dụng khi bệnh/rối loạn có thể xảy ra sau thời điểm viết tài liệu, có thể được lên lịch hay có kế hoạch bắt đầu sau thời điểm tài liệu được viết. Ví dụ: xét câu "Patient needs a follow up abdomian MRI within 1 month to evaluate her renal lesions." Bệnh/rối loạn "renal lesions" được gán nhãn phân lớp "AFTER".

Mục tiêu của bài báo trình bày hướng tiếp cận lai ghép gồm dựa trên máy học và luật để phân lớp mối quan hệ giữa bệnh/rối loạn và thời điểm tài liệu được viết. Cụ thể, dựa trên kết quả của hai công trình [2, 3] nhóm tác giả đã cải tiến hiệu quả cho hệ thống bằng cách tích hợp thêm phương pháp dựa trên luật vào hệ thống sau giai đoạn máy học. Nội dung bài báo được trình bày phần tiếp theo.

II. CÁC CÔNG TRÌNH LIÊN QUAN

In the 2012 i2b2 Challenge, nhóm tác giả công trình [7] đã xây dựng một bộ phân lớp mối quan hệ thời gian giữa sự kiện và thời gian viết tài liệu lâm sàng dựa trên phương pháp máy học. Mối quan hệ này được gọi là Tlinks giữa Event và Section time (bao gồm: Discharge time và Admission time). Tlinks có ba giá trị: BEFORE, OVERLAP và AFTER. Các thuật toán SVM và CRF++ được sử dụng để xây dựng bộ phân lớp và tập đặc trưng.

In ShARe/CLEF eHealth 2014, kết quả các đội top 5 trên thuộc tính DocTime được trình bày trong bảng 1. Trong bài báo chỉ giới thiệu tóm lược ba công trình đầu.

Bảng 1. Kết các đội top 5 trên thuộc tính DocTime Class trong ShARe/CLEF eHealth 2014.

Rank	Team	Accuracy
1	TeamHITACHI	0.328
2	LIMSI	0.322
3	TeamHCMUS	0.306
4	DKI-Medical	0.179
5	TeamHPI	0.060

Team HITACHI [5] đã sử dụng mô đun DocTime trong cTAKES[3] và bổ sung một số đặc trưng mới. cTAKES (clinical Text Analysis and Knowledge Extraction System) là một hệ thống xử lý ngôn ngữ tự nhiên dùng cho việc rút trích thông tin trên các bệnh án điện tử, được phát triển bởi Apache. cTAKES được xây dựng dựa trên framework UIMA[4] (Unstructured Information Management Architecture). Các đặc trưng được sử dụng bao gồm tokens và POS tags trong của sổ [-3,3] xung quanh rối loạn, thì của động từ, tiêu đề của phân mục (section) và động từ gần nhất. Tác giả cũng bổ sung thêm đặc trưng biểu thức thời gian (temporal expression).

LIMSI [6] dùng phương pháp máy học để phân lớp. Tác giả đã phân tích trên dữ liệu huấn luyện và nhận thấy tính quan trọng của cấu trúc tài liệu. Ví dụ: phân mục "Chief Complaint" thường đề cập đến rối loạn xuất hiện trong quá khứ, phân mục "Discharge" đề cập đến các rối loạn có thể xảy ra khi xuất viện. Vì vậy, tác giả đã xây dựng một danh sách các tiêu đề phân mục thường gặp. Tài liệu được chia thành năm đoạn bằng nhau (equal-sized bins) và vị trí của rối loạn trong năm đoạn này cũng được sử dụng là đặc trưng phân lớp. Các đặc trưng được sử dụng bao gồm: Vị trí

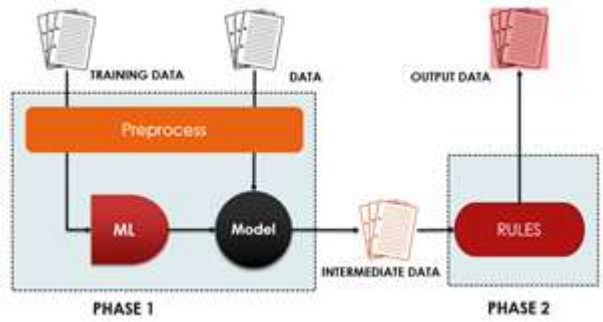
[3] http://ctakes.apache.org/
[4] https://uima.apache.org/

của rối loạn, loại tài liệu, loại phân mục và Bag-of-words. Kết quả độ chính xác là 0.322 thấp hơn rất nhiều so với khi họ tham gia thách thức I2B2 năm 2012.

TeamHCMUS [2, 3] cũng áp dụng hướng tiếp cận dựa trên máy học. Nhóm tác giả sử dụng Weka tool và LibSVM để xây dựng một tập phân lớp dựa trên máy học. Thuật toán áp dụng để phân lớp là SVM và tập đặc trưng được đề xuất. Kết quả độ chính xác chỉ đạt 0.306.

III. HƯỚNG TIẾP CẬN

Hướng tiếp cận thực hiện gồm giai đoạn 1: sử dụng phương pháp máy học để phân lớp các rối loạn vào những lớp quan hệ ngữ nghĩa với thời điểm tài liệu được viết. Giai đoạn 2: sử dụng phương pháp luật để hiệu chỉnh kết quả giai đoạn 1 nhằm tăng hiệu quả cho hướng tiếp cận, xem hình 2. Chi tiết được trình bày cụ thể ở phần tiếp theo.



Hình 2. Kiến trúc tổng quát của hướng tiếp cận

A. Tiền xử lý (Preprocess)

Do tính nhạy cảm của dữ liệu y khoa, nên để đảm bảo tính riêng tư, các tên bác sĩ, tên bệnh viện… đều được mã hóa với các ký tự đặc biệt như dấu *, [,]. Điều này dẫn đến việc xử lý (tách câu, phân tích cú pháp, gán nhãn từ loại…) sẽ không chính xác. Chính vì vậy, cần phải qua một giai đoạn tiền xử lý để giải quyết các vấn đề trên. Giai đoạn này sẽ thực hiện thay thế các tên đó bằng các tên giả. Ví dụ: Đoạn văn nguyên mẫu chưa được tiền xử lý (xem hình 3) và sau khi tiền xử lý (xem hình 4) như sau:

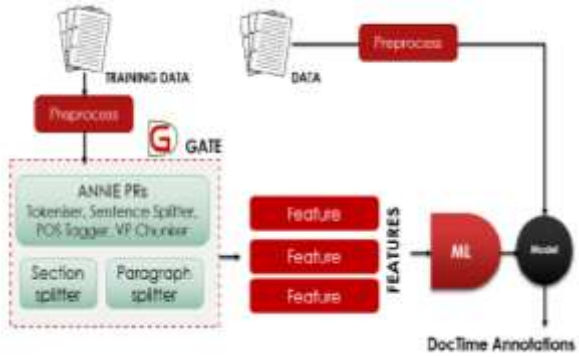
"She was transfered to [**Hospital1 27**] *per recommendation of her GI specialist Dr.* [**First Name (STitle) 5060**]*"*

Hình 3. Văn bản nguyên mẫu.

"She was transfered to Ohio Hospital per recommendation of her GI specialist Dr. Gates"

Hình 4. Văn bản sau khi tiền xử lý.

B. Giai đoạn 1 (Phase 1): Dựa trên máy học



Hình 5. Kiến trúc xử lý giai đoạn 1.

Sau khi tiền xử lý dữ liệu huấn luyện, bước tiếp theo GATE[5] (General Architecture for Text Engineering) được sử dụng để thực hiện tách section (Section splitter), tách đoạn (Paragraph splitter), cắt câu (Sentence splitter), tách token, gán nhãn từ loại (POS tagger) và xác định cụm động từ (VP Chunker). Sau đó, chương trình thực hiện rút trích các đặc trưng và đưa vào bộ phân lớp để huấn luyện xây dựng Model (xem hình 5).

Các đặc trưng được sử cho phân lớp theo phương pháp máy học như sau:

[5] https://gate.ac.uk/

- *Đặc trưng loại tài liệu:* Trong tập dữ liệu gồm có bốn loại tài liệu: Discharge summary, Radiology report, Echo report và ECG report. Mỗi loại tài liệu sẽ có xu hướng ghi nhận các thông tin liên quan đến quá trình điều trị của bệnh nhân ở một giai đoạn nhất định, điều này giúp hỗ trợ xác định mối quan hệ thời gian giữa bệnh/rối loạn và thời gian viết tài liệu. Các tài liệu thường thể hiện diễn tả tình trạng của bệnh nhân trước khi nhập viện, trong quá trình điều trị hoặc sau khi xuất viện. Ví dụ: Hầu hết các bệnh/rối loạn xuất hiện trong các báo cáo (ECG report, Radiology report và ECG report) đều rơi vào giai đoạn bệnh nhân gặp bác sĩ, do đó thường được phân lớp là OVERLAP.
- *Đặc trưng phân mục:* Phân mục là một đặc trưng quan trọng, nó giúp xác định mối quan hệ. Trong một tài liệu, mỗi phân mục sẽ được dùng để ghi các thông tin ở mỗi giai đoạn cụ thể. Ví dụ: Phân mục "History of Present Illness" thường đề cập đến các thông tin ở quá khứ (trước thời gian tài liệu được viết), trong khi phân mục "Medication" lại đề cập đến các thông tin ở hiện tại (trong khoảng thời gian tài liệu được viết) và "Discharge Instruction" sẽ đề cập đến các thông tin trong tương lai (sau khi bệnh nhân xuất viện). Trong bảng 1 thể hiện sự phân bổ phân lớp trong các phân mục:

Bảng 2. Sự phân bổ phân lớp trong các phân mục

PHÂN MỤC	BEFORE	BEFORE-OVERLAP	OVERLAP	AFTER
Chief complaint	10%	90%	0 %	0 %
Physical examination	1 %	3 %	95%	1 %
Discharge instruction	0 %	4 %	10%	2 %
Labs-studies	0 %	9 %	91%	0 %
History of present illness	39%	44%	16%	0 %

Bằng việc quan sát trên tập dữ liệu, một danh sách các tên phân mục được xây dựng thủ công. Việc phân tách nội dung tài liệu thành các phân mục được thực hiện bằng luật văn phạm JAPE trong GATE. Một ví dụ về luật để phát hiện section "chief complaint":

```
Rule: chief_complaint( {Token.lcString == "chief", Token.docType
      == "DISCHARGE_SUMMARY"} {Token.lcString == "complaint"}
      {Token.lcString == ":"} ) : match --> : match.SectionHeader
      = {kind = "chief complaint"}
```

Số lượng phân mục được xác định sau khi áp dụng tập luật trên từng loại tài liệu được thống kê trong bảng 2.

Bảng 3. Số lượng các phân mục có trong tập dữ liệu huấn luyện

LOẠI TÀI LIỆU	SỐ LƯỢNG PHÂN MỤC ĐƯỢC TRÍCH
Discharge summary	33
Echo report	3
Radiology report	11
Ecg report	1

- *Đặc trưng về thì và thể của động từ (Tense and Aspect feature):* Dựa trên bộ phân tích cú pháp xác định thì và thể của động từ trong câu. Thì bao gồm các giá trị: *past*, *present*, *future* và thể của động từ có thể nhận các giá trị: *progressive*, *perfective*, *perfective-progressive*. Chương trình sử dụng Processing Resouce ANNIE VP Chunker trong GATE để xác định thì của các động từ trong câu chứa bệnh/rối loạn.
- *Đặc trưng mối quan hệ về các mốc thời gian lâm sàng:* Trong lĩnh vực lâm sàng có một số mốc thời gian rất đặc trưng chẳng hạn như: *on postoperative*, *on physical examination*, *day of admission*, *hospital day one*, *post-discharge* được gọi là các thuật ngữ thời gian lâm sàng (clinical date time terms). Một mối quan hệ giữa bệnh/rối loạn và dòng thời gian lâm sàng được định nghĩa theo biểu thức chính quy như sau: <*Quan hệ*, *Dòng thời gian lâm sàng*> trong đó *Quan hệ* có thể nhận các giá trị như: ON, BEFORE, AFTER và *Dòng thời gian lâm sàng* là những thuật ngữ lâm sàng. Biểu thức chính qui được sử dụng để nhận dạng cả Quan hệ và Mốc thời gian lâm sàng. Ví dụ: Biểu thức chính quy để nhận dạng một mối quan hệ về mốc thời gian lâm sàng <*ON*, *POSTOPERATIVE*> trong câu "On postoperative day #1, the patient was taken to arteriogram" là: (on)?\s*postoperative\s*day\s*(#?\d+|(%NUMBER%)) trong đó, %NUMBER% là chuỗi sẽ được thay thế bởi các chuỗi chỉ số đếm như: one, two, three …
- *Đặc trưng biểu thức thời gian*: Sử dụng biểu thức chính quy để rút trích biểu thức thời gian và chuẩn hóa kết quả. Các kết quả được sử dụng như một đặc trưng phân lớp.

Đối với dữ liệu ở tập đánh giá (TEST DATA) cũng thực hiện giai đoạn tiền xử lý, sau đó áp dụng mô hình (Model) ở bước trên để phân lớp DocTime Class cho các bệnh/rối loạn (xem hình 5).

C. Giai đoạn 2: Dựa trên luật

Giai đoạn đoạn này, tập luật được đề xuất để chỉnh sửa kết quả phân lớp ở giai đoạn 1. 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 Rối loạn.

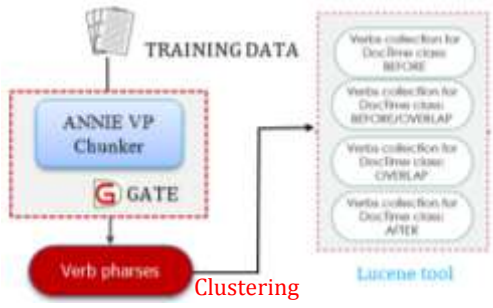
1. Đặc điểm cụm động từ

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*, v.v…

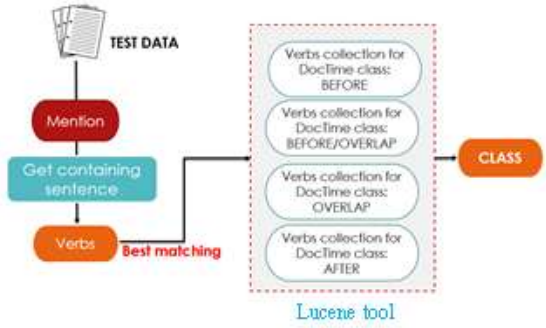
Những cụm động từ phổ biến được xác định bằng cách tính trọng số (tf-idf) của mỗi cụm động từ. Các cụm động từ đại diện cho một lớp khi nó xuất hiện nhiều trong lớp này, nhưng xuất hiện ít ở lớp khác.

Quá trình thực hiện tính trọng số cho cụm động từ như sau: Từ dữ liệu huấn luyện (train data), 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). Chương trình sử dụng Lucene[6] 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 (trừ lớp UNKNOWN, mặc định) (xem hình 6).

Đối với dữ liệu đánh giá (test data), hệ thống rút trích cụm động từ liên quan đến rối loạn, cụm động từ này được so khớp với bốn tập phân lớp (BEFORE, OVERLAP, BEFORE-OVERLAP và AFTER) thông qua công cụ Lucene 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 (xem hình 7).



Hình 6. Quá trình xử lý huấn luyện

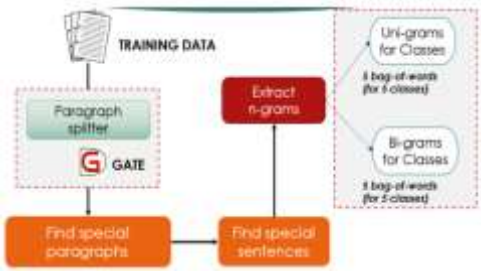


Hình 7. Quá trình xử lý đánh giá

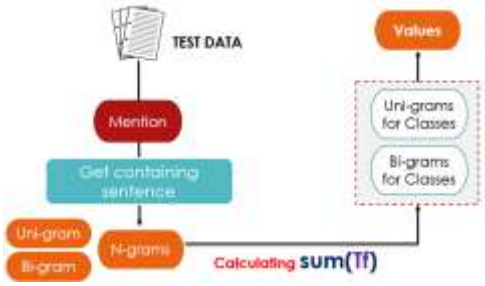
2. Đặc điểm n-grams

Đặc điểm n-grams được xây dựng dựa trên ý tưởng như sau: "Nếu một 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: Đoạn văn với phần lớn các *bệnh/rối loạn* thuộc phân lớp BEFORE_OVERLAPS, 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 (training data), mỗi tài liệu được tách thành các đoạn. Trong mỗi 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 đoạn văn đó. Trên các câu này rút trích các uni-gram, bi-gram, 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 năm túi từ uni-gram cho năm lớp và năm túi từ bi-gram cho năm lớp tương ứng (BEFORE, OVERLAP, BEFORE_OVERLAP, AFTER and UNKNOWN) (xem hình 8).



Hình 8. Quá trình xử lý training data



Hình 9. Quá trình xử lý test data

[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 (xem hình 9).

3. 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

IV. KẾT QUẢ

Tập dữ liệu huấn luyện và đánh giá được cung cấp bởi diễn đàn nghiên cứu Shared task of ShARe/CLEF eHealth 2014. Dữ liệu huấn luyện có 298 tài liệu gồm bốn loại: Discharge summary, ECHO report, Rediology report và ECG report. Dữ liệu đánh giá chỉ có 133 tài liệu chỉ một loại là discharge summary. Số lượng bệnh/rối loạn được thống kê trong bảng 3 và 4.

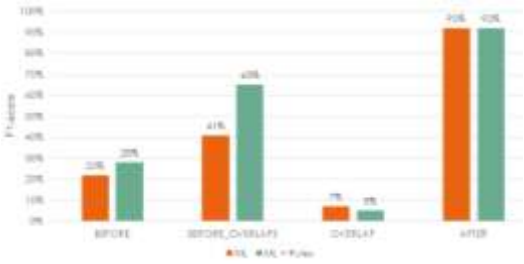
Bảng 5. Số lượng rối loạn (disorder) trên tập dữ liệu huấn luyện

DOC Types	#DOC	#Disorder	Percent %
Discharge summary	136	9098	79%
ECHO report	54	1429	12%
ECG report	54	196	2 %
Radiology report	54	831	7 %
Total	**298**	**11554**	

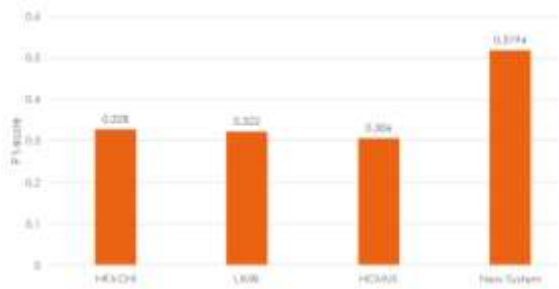
Bảng 4. Số lượng rối loạn trên tập dữ liệu đánh giá.

DOC Types	#DOC	#Disorder	Percent %
Discharge summary	133	8003	100%
ECHO report	0	0	0
ECG report	0	0	0
Radiology report	0	0	0
Total	**133**	**8003**	

Đối với phương pháp máy học, thực hiện đánh giá với 10-fold cross validation trên dữ liệu huấn luyện để chọn ra thuật toán phân lớp tốt nhất. Các thuật toán được thực nghiệm gồm NaiveBayes, C4.5, kNN và SVM. Kết quả thực nghiệm trên dữ liệu huấn luyện cho thấy thuật toán SVM cho kết quả phân lớp tốt nhất. Cho nên, thuật toán SVM được chọn dùng để huấn luyện và phân lớp trên tập dữ liệu đánh giá. Kết quả thực nghiệm cho thấy sau khi áp dụng luật thì kết quả đã cải tiến về độ chính xác (tức độ chính xác cho các phân lớp tăng lên) (xem hình 10). Vì thế, Kết quả của hướng tiếp cận mới (New System) trên tập dữ liệu đánh giá với độ đo F-score là 0.5194. Như vậy, hướng tiếp cận kết hợp máy học và luật đã cải tiến hiệu quả cho bài toán phân lớp mối quan hệ giữa bệnh/rối loạn và thời gian viết tài liệu lâm sàng (xem hình 11).



Hình 10. So sánh kết quả của các hệ thống



Hình 11. So sánh kết quả của các hệ thống

V. KẾT LUẬN

Bài báo đã đề xuất một hướng tiếp cận kết hợp giữa máy học và luật. Ở phần máy học, chương trình đã thực nghiệm một số thuật toán phân lớp như: NaiveBayes, C4.5, kNN và SVM để chọn thuật toán tốt nhất cho phân lớp. Đối với phần luật, dựa trên những đặc điểm của động từ và kết hợp với n-grams để xây dựng tập luật. Kết quả đã có cải tiến so với các hệ thống trước đó. Tuy nhiên, việc xác định mối quan hệ giữa bệnh/rối loạn và thời gian viết tài liệu là bài toán khá phức tạp, đòi hỏi phải nghiên cứu sâu hơn về dữ liệu để tìm kiếm đặc trưng mới, kết hợp các phương pháp luật và máy học. Thời gian tới, chúng tôi sẽ tiếp tục nghiên cứu và phân tích để tìm kiếm đặc trưng mới của dữ liệu với mong muốn cải tiến hiệu quả cho bài toán phân lớp mối quan hệ giữa bệnh/rối loạn và thời gian viết tài liệu lâm sàng nhằm giúp cho người dùng có những thông tin với độ tin cậy cao.

TÀI LIỆU THAM KHẢO

[1] Hanna Suominen, Tobias Schreck, Gondy Leroy, Harry Hochheiser, Lorraine Goeuriot, Liadh Kelly, Danielle L Mowery, Jaume Nualart, Gabriela Ferraro, Daniel Keim. *Task 1 of the CLEF eHealth Evaluation Lab 2014 Visual-Interactive Search and Exploration of eHealth Data*. CEUR Workshop Proceedings, ISSN 1613-0073, Vol-1180, 2014.

[2] Huu Nghia Huynh, Son Lam Vu, Bao Quoc Ho, "*ShARe/CLEFeHealth: A Hybrid Approach for Task 2*", CEUR Workshop Proceedings, ISSN 1613-0073, Vol-1180, 2014.

[3] Huỳnh Hữu Nghĩa, Vũ Sơn Lâm, Hồ Bảo Quốc. *Một Hướng Tiếp Cận Xác Định Mối Quan Hệ giữa Bệnh và Thời Gian Viết Tài Liệu Lâm Sàng*. Hội thảo quốc gia lần thứ XVII: Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông, pages 155 – 160, Đắk Lắk, 30-31/10/2014.

[4] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, David Martinez, Guido Zuccon, João Palotti (2014), "*Overview of the ShARe/CLEF eHealth Evaluation Lab 2014*", Springer International Publishing Switzerland.

[5] Nishikant Johri, Yoshiki Niwa, Veera Raghavendra Chikka, "*Optimizing Apache cTAKES for Disease/Disorder Template Filling: Team HITACHI in 2014 ShARe/CLEF eHealth Evaluation Lab*", CEUR Workshop Proceedings, ISSN 1613-0073, Vol-1180, 2014.

[6] Thierry Hamon, Cyril Grouin, Pierre Zweigenbaum. "*Disease and Disorder Template Filling Using Rule-Based and Statistical Approaches*", CEUR Workshop Proceedings, ISSN 1613-0073, Vol-1180, 2014.

[7] Tang B, Wu Y, Jiang M, et al. *A hybrid system for temporal information extraction from clinical text*. J Am Med Inform Assoc 2013.

[8] Will Styler, Guergana Savova, Martha Palmer, James Pustejovsky, Tim O'Gorman, and Piet C. de Groen (2014), "THYME Annotation Guidelines.

A COMBINED APPROACH FOR TEMPORAL RELATION CLASSIFICATION

Nghia Huynh, Lam Vu, Quoc Ho

ABSTRACT*— On reviewing clinical documents, doctors, researchers and caregivers of patients want to know the time when a disease / disorder appears (in the past, at the present, in the future or from the past until now...), compared to the time the document was written. The information about this period of time is very useful in creating treatment regimen for patients, making inquiry system and summarizing the documents. This paper proposes a hybrid approach between rules and machine learning to classify the relationship between diseases/disorders and the time for writing clinical documents, the oriented approach has the result of accuracy 0.5194, which is higher than the best ranking system (0.328) in the ShARe/Clef eHealth 2014 Evaluation Lab.*