Kỷ yếu Hội nghị Quốc gia lần thứ XI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 09-10/08/2018
DOI: 10.15625/vap.2018.00044

MỘT SỐ ĐẶC TRƯNG TRONG TÓM TẮT VĂN BẢN BÁO MẠNG ĐIỆN TỬ TIẾNG VIỆT

Lê Ngọc Thắng[1,2], Lê Quang Minh[2]
[1] Cục Tham mưu An ninh, Bộ Công an
[2] Viện Công nghệ thông tin, Đại học Quốc gia Hà Nội
lengocthang@gmail.com, quangminh@vnu.edu.vn

TÓM TẮT: *Tóm tắt văn bản tự động đã được nghiên cứu từ những năm 1950 của thế kỷ 20. Tóm tắt tự động văn bản tiếng Việt mới chỉ được tập trung nghiên cứu từ những năm đầu của thế kỷ 21. Về cơ bản những nghiên cứu này là ngắn hạn, đơn lẻ và tập trung vào hướng trích rút qua việc sử dụng những đặc trưng của ngôn ngữ tiếng Anh để áp dụng vào mô hình tóm tắt tự động văn bản tiếng Việt. Phần lớn các kết quả thử nghiệm đều được thực hiện trên thể loại văn bản báo mạng điện tử. Tuy nhiên, cho đến nay, chưa có nhiều nghiên cứu về đặc trưng ngôn ngữ của thể loại văn bản báo mạng điện tử tiếng Việt phục vụ cho bài toán trích rút câu. Bài báo này sẽ nghiên cứu một số đặc trưng riêng, trên cơ sở đó áp dụng đánh giá các đặc trưng đó trong trích rút câu phục vụ tóm tắt tự động văn bản tiếng Việt thể loại báo mạng điện tử.*

Từ khóa: *tóm tắt văn bản tự động, tóm tắt văn bản tiếng Việt, báo mạng điện tử, từ khóa.*

I. GIỚI THIỆU

Theo quan điểm của các nhà nghiên cứu về tóm tắt văn bản thì bản tóm tắt là một bản rút gọn của một hay nhiều văn bản gốc thông qua việc lựa chọn và tổng quát hóa các khái niệm quan trọng. Theo Mani và cộng sự [1] thì tóm tắt văn bản là quá trình trích lược chắt lọc những thông tin quan trọng nhất từ văn bản gốc để tạo ra một phiên bản giản lược sử dụng cho các mục đích hoặc nhiệm vụ khác nhau. Thông thường một văn bản tóm tắt có độ dài không quá nửa so với văn bản gốc.

Có rất nhiều phương pháp tiếp cận về tóm tắt văn bản, qua đó cũng có rất nhiều cách phân loại các hệ thống tóm tắt văn bản, tuy nhiên, thông thường người ta hay sử dụng cách phân loại theo kết quả đầu ra (output). Đối với cách phân loại này thì có 02 phương pháp tóm tắt văn bản đó là tóm tắt theo phương pháp trích rút (Extract) và tóm tắt theo phương pháp tóm lược (Abstract).

Phương pháp tóm tắt trích rút là phương pháp tìm ra các đơn vị quan trọng nhất của văn bản đầu vào (đơn vị thường sử dụng là câu) sau đó lựa chọn các câu có liên quan đến các đơn vị quan trọng này để tạo ra văn bản tóm tắt. Đặc trưng của phương pháp này là xác định xem một câu của văn bản đầu vào có thuộc văn bản tóm tắt hay không, do vậy văn bản tóm tắt cũng thường tuân theo thứ tự nội dung của văn bản đầu vào. Đối với phương pháp này đã có một số hướng tiếp cận: Hướng tiếp cận tiên phong; hướng tiếp cận theo thống kê; hướng tiếp cận dựa trên kết nối văn bản; hướng tiếp cận dựa trên lý thuyết đồ thị; hướng tiếp cận dựa vào học máy và hướng tiếp cận đại số.

Phương pháp tóm lược xuất phát từ mục tiêu hiểu đầy đủ nội dung văn bản tóm tắt, sau đó tạo ra các câu mới cho bản tóm tắt theo tỉ lệ yêu cầu của người dùng. Phương pháp này rất giống với cách tóm tắt của con người nhưng về mặt thực tế rất khó để đạt được kết quả như tóm tắt thủ công. Một số hướng nghiên cứu đã dựa vào các đơn vị đặc trưng như từ, cụm từ, thành phần câu quan trọng để sinh ra các câu mới cho bản tóm tắt. Một số hướng tiếp cận của phương pháp này như sau: Dựa vào các từ hoặc cụm từ quan trọng; dựa trên kỹ thuật cô đọng văn bản; dựa trên kỹ thuật rút gọn văn bản, nối câu; dựa trên kỹ thuật rút gọn câu.

Về lĩnh vực tóm tắt tự động văn bản tiếng Việt, với hướng tiếp cận tóm tắt trích rút có một số công trình như của Nguyễn Lê Minh và cộng sự [2], Hà Thành Lê và cộng sự [3], Đỗ Phúc và Hoàng Kiếm [4], Lê Thanh Hương và cộng sự [5], Nguyễn Thị Thu Hà [6], Nguyễn Nhật An [7]. Nguyễn Lê Minh và cộng sự [2] trích rút sử dụng phương pháp SVM với các đặc trưng gồm vị trí câu, chiều dài câu, độ liên quan chủ đề, tần suất từ, cụm từ chính và khoảng cách từ. Hà Thành Lê và cộng sự [3] kết hợp một số phương pháp trích rút đặc trưng trong trích rút văn bản tiếng Việt như đặc trưng về tần suất từ TF×IDF, vị trí, từ tiêu đề, từ liên quan. Các đặc trưng được kết hợp tuyến tính với nhau để tính trọng số mỗi câu trong văn bản gốc. Lê Thanh Hương và cộng sự [5] sử dụng giải thuật PageRank cải tiến với hệ số nhân cho các từ xuất hiện trong tiêu đề văn bản để trích rút câu. Nguyễn Thị Thu Hà [6] sử dụng đặc trưng tần suất từ, vị trí câu và đặc trưng tiêu đề để trích rút câu quan trọng. Nguyễn Nhật An [7] trích rút câu dựa trên các đặc trưng vị trí câu, tần suất từ, độ dài câu, xác suất thực từ, thực thể có tên, dữ liệu số, tương tự với tiêu đề và câu trung tâm để tính trọng số câu.

Có thể nhận thấy các đề tài trên có chung một số đặc điểm là phần lớn vẫn sử dụng đặc trưng chung của ngôn ngữ trong nội tại văn bản. Có một số nghiên cứu bước đầu đề cập tới tiêu đề văn bản như [5], [7], còn lại nhìn chung chưa khai thác được nhiều các thông tin liên quan khác của văn bản. Trong khi đó kết quả thực nghiệm phần lớn dựa trên dữ liệu là thể loại văn bản báo mạng điện tử, đây là một thể loại văn bản có tính đặc thù, có nhiều đặc trưng riêng, có nhiều thành phần mang thông tin khác nhau.

Xuất phát từ thực tế đó, để nghiên cứu bài toán trích rút câu cho thể loại văn bản báo mạng điện tử tiếng Việt chúng tôi đã xây dựng một mô hình để giải quyết bài toán này dựa trên các đặc trưng riêng của thể loại báo mạng điện tử. Đối tượng nghiên cứu, xử lý của bài báo này là văn bản báo mạng điện tử tiếng Việt.

II. ĐẶC TRƯNG CỦA VĂN BẢN BÁO MẠNG ĐIỆN TỬ TIẾNG VIỆT

Theo Nguyễn Thị Trường Giang [9] cho đến nay, báo mạng điện tử Việt Nam đã phát triển qua 3 giai đoạn. Giai đoạn từ năm 1997-2001 là giai đoạn hình thành, nội dung chủ yếu lấy từ báo in đưa lên, thường chưa có thông tin do phóng viên báo mạng điện tử tự làm. Giai đoạn từ 2001-2005 là giai đoạn phát triển nở rộ của những báo lớn, tuy nhiên, đội ngũ những người làm báo mạng điện tử chưa được phát triển bài bản, cơ bản vẫn còn thiếu chuyên nghiệp. Giai đoạn 2005 đến nay đánh dấu sự trưởng thành về số lượng và chất lượng của báo mạng điện tử Việt Nam. Giai đoạn này các báo mạng điện tử đã đi vào chuyên nghiệp, chú trọng nhiều đến nội dung và hình thức, một số báo cũng đã có được những thương hiệu và phong cách riêng. Các ưu điểm vượt trội của báo mạng điện tử như khả năng đa phương tiện, tính tương tác cao, tìm kiếm nhanh ngày càng được quan tâm và tận dụng khai thác một các có hiệu quả.

Về ngôn ngữ, báo mạng điện tử ở Việt Nam đã hình thành lên những đặc điểm chính về mặt ngôn ngữ sau: Ngôn ngữ báo mạng điện tử là ngôn ngữ đa phương tiên; ngắn gọn, rõ ràng, dễ hiểu; mang tính thời sự cao nhất trong các loại hình báo chí; các thành tố được trình bày linh hoạt, phục vụ cho liên kết đa chiều; ngôn ngữ thể hiện tính hội nhập cao.

Về cơ bản, hiện nay cấu trúc thông tin trong một bài báo mạng điện tử gồm 11 phần: Tít chính, Sa pô, Chính Văn, Tít phụ, Tranh - ảnh, Đồ hình, Video và ảnh động, Âm thanh, Các box thông tin và tư liệu, các đường link, Từ khóa và Tags. Trong đó Sa pô là câu mào đầu của báo, mục đích là để tạo sự hấp dẫn cho người đọc. Theo Hoàng Anh [10] Sa pô có thể bao gồm một câu, vài câu hoặc nhiều câu. Trong báo chí hiện đại lời mào đầu thường có xu hướng càng ngắn gọn càng tốt. Hiện nay có một số nghiên cứu của các nhà công nghệ thông tin thường nhầm lẫn Sa pô là phần tóm tắt bài báo của tác giả. Trong [12] tác giả nêu đoạn văn bản cần tóm tắt là cả bài báo và phần tóm tắt là đoạn mô tả phía dưới tiêu đề. Nguyễn Nhật An [7] cũng đề cập tới việc sử dụng bản tóm tắt của tác giả dưới tiêu đề bài báo để làm cơ sở đánh giá kết quả nghiên cứu. Những sự nhầm lẫn này có thể dẫn đến những kết quả không chính xác đối với những nghiên cứu trên.

Qua nghiên cứu về đặc điểm của báo mạng điện tử, chúng tôi nhận thấy các từ khóa, từ gán nhãn (Tags) và các thực thể có tên, các cụm từ có trong câu tiêu đề, trong sa pô là những thành phần mang nhiều thông tin trong văn bản. Để xử lý thông tin trên mạng internet, những nghiên cứu hiện nay không chỉ tập trung vào nội dung của đối tượng mà đã đề cập tới các thành phần mang thông tin khác. Trong [8] các tác giả cũng đã sử dụng hashtags của ảnh facebook được cung cấp bởi người dùng để nhận dạng ảnh qua học sâu. Do vậy để trích xuất câu trong văn bản, chúng tôi thấy rằng cần phải nghiên cứu, đánh giá vai trò về mặt ngữ nghĩa của các đặc trưng trên đối với văn bản báo mạng điện tử.

2.1. Từ khóa và nhãn trong báo mạng điện tử

Theo từ điển "The Oxford English Dictionary" [14] thì từ khóa (keywords) là một từ dùng để nói đến một chìa khóa hay là một loại mật mã nào đó và sử dụng để giải quyết, giới thiệu về một sự vật hiện tượng cụ thể. Đó là một từ đóng vai trò quan trọng và có ý nghĩa trong việc thể hiện nội dung của một văn bản. Theo Lê Thanh Hà [11] "Từ khoá trên báo điện tử là một cụm từ gồm 4 đến 8 chữ tóm tắt nội dung chủ đề của bài viết, được các tờ báo sử dụng nhiều nhất và là cụm từ thông dụng nhất trong việc trực tiếp dùng để tìm kiếm tin tức hằng ngày, về những vấn đề mang tính thời sự, xã hội, kinh tế, đời sống, giải trí, công nghệ… trong và ngoài nước. Mỗi tờ báo điện tử hướng theo lĩnh vực riêng, người dùng riêng, tương đương với bộ từ khoá riêng cho từng lĩnh vực. Nếu muốn trang web của bạn để xếp hạng cao trong kết quả tìm kiếm và có thể thu hút nhiều người đến trang web – bạn phải chọn từ khóa một cách cẩn thận".

Từ khóa chính thích hợp sẽ giúp tác phẩm báo mạng điện tử nằm ở đầu trên bảng kết quả của các công cụ tìm kiếm như Google, Bing. Từ khóa chính sẽ được nhà báo lựa chọn từ nội dung bài viết. Trong trường hợp có nhiều từ khóa khác nhau, để lựa chọn từ khóa được tìm kiếm nhiều nhất song vẫn phù hợp với nội dung bài viết, nhà báo có thể lựa chọn bằng kinh nghiệm, độ nhạy cảm, thói quen hoặc bằng các công cụ hỗ trợ. Từ khóa được lựa chọn phải đáp ứng yêu cầu phù hợp với nội dung bài viết song phải là từ khóa được nhiều người tìm kiếm qua các công cụ tìm kiếm trên mạng. Một số công cụ gợi ý từ khóa thường được sử dụng là keywordtool.io, google trends.

Sau khi xác định được từ khóa chính, biên tập viên, nhà báo có thể xác định thêm các từ khóa liên quan – được gọi là Tag (từ gán nhãn). Tag được định nghĩa là từ khoá liên quan đến bài viết, không phải là từ khoá chính. Tag là các từ khóa gần nghĩa với từ khóa chính hoặc là từ khoá đơn lẻ theo công thức: Who (ai, cái gì) – Where (ở đâu, xảy ra ở đâu) – What (vấn đề gì) – When (xảy ra khi nào). Thông thường, mỗi bài báo mạng điện tử sử dụng tối đa 5 tags, tối thiểu 3 tags. Trong đó, phần lớn được các bài báo mạng điện tử phân bố theo cơ cấu tag 1 – 2 là từ khoá gần nghĩa (có thể nhiều hơn), tag 3 – 4 – 5 là từ khoá theo công thức.

Như vậy rõ ràng từ khóa và từ gán nhãn có vai trò ngữ nghĩa rất quan trọng trong bài báo mạng điện tử.

2.2. Thực thể có tên

Theo Nguyễn Trí Nhiệm, Nguyễn Thị Trường Giang [13] cấu trúc theo hình tháp ngược "Cái gì – Ai – Ở đâu – Khi nào – Vì sao – Như thế nào" là cấu trúc hiện đại phù hợp với yêu cầu viết cho báo mạng điện tử là phải đưa mọi thông tin quan trọng lên đầu. Với cấu trúc này, về phương diện ngôn ngữ, hai yếu tố Ai – Cái gì trở thành chủ ngữ và vị ngữ của câu, những yếu tố còn lại trở thành trạng ngữ. Do vậy, các thực thể có tên người, địa danh, tổ chức,… sẽ đóng một vai trò quan trọng về ngữ nghĩa trong văn bản báo mạng điện tử. Nguyễn Nhật An [7] cũng đã chỉ ra vai trò quan trọng của thực thể có tên trong văn bản tiếng Việt thuộc thể loại tin tức. Trong [15] tác giả cũng đã sử dụng tiêu chí thực thể có tên xuất hiện 02 lần trở lên trong văn bản là thực thể có ngữ nghĩa quan trọng để trích rút câu.

Ở đây, các thực thể có tên được xem là quan trọng khi xuất hiện từ 2 lần trở lên trong nội dung bài báo, hoặc là các thực thể có tên trong tiêu đề hoặc trong sa pô. Sau đây khi đề cập đến các thực thể có tên chúng ta hiểu là các thực thể có tên đáp ứng được một trong các yêu cầu trên*.

III. TRÍCH RÚT CÂU VĂN BẢN BÁO MẠNG ĐIỆN TỬ TIẾNG VIỆT DỰA TRÊN TỪ KHÓA VÀ THỰC THỂ CÓ TÊN

Có hai vấn đề cần được xem xét đối với phương pháp tóm tắt văn bản theo hướng trích rút câu. Một là, xem xét sự phù hợp của từng đặc trưng trong bài toán tóm tắt văn bản tiếng Việt và lựa chọn tập đặc trưng phù hợp đối với văn bản tiếng Việt cần nghiên cứu. Hai là, mỗi giá trị đặc trưng sử dụng phải được xác định hệ số sao cho thích hợp nhất đối với bài toán.

Trong bài báo này, để tính độ quan trọng câu chúng tôi dựa trên 03 đặc trưng là từ khóa chính, từ khóa nhãn (tags) và thực thể có tên, sau đó xác định các hệ số đặc trưng phù hợp. Bài toán được mô hình hóa như sau:

Đối với văn bản V:

Gọi:

- $S = S_1, S_2 \dots , S_l$, trong đó S_i là câu thứ i trong văn bản có l câu.

- $T_i = t_1, t_2, \dots, t_q$, trong đó t_j là từ thứ j trong câu S_i có q từ.

- $X = x_1, x_2 \dots , x_n$ là tập các từ khóa. Giá trị ngữ nghĩa của từ khóa x_i trong câu được tính là α nếu từ khóa đó có trong câu, là 0 nếu từ khóa không có trong câu.

- $Y = y_1, y_2, \dots , y_m$ là tập các từ gán nhãn. Giá trị ngữ nghĩa của từ gán nhãn y_i trong câu được tính là β nếu từ khóa đó có trong câu, là 0 nếu từ gán nhãn không có trong câu.

- $Z = z_1, z_2, \dots , z_k$ là tập các thực thể có tên. Giá trị ngữ nghĩa của thực thể có tên z_i trong câu được tính là γ nếu từ khóa đó có trong câu, là 0 nếu thực thể có tên không có trong câu.

Các tập X, Y, Z sẽ được chuẩn hóa đảm bảo $X \cap Y = \emptyset$; $Y \cap Z = \emptyset$; $Z \cap X = \emptyset$, nghĩa là nếu một từ thuộc nhiều tập thì sẽ được chuẩn hóa chỉ giữ lại ở tập có trọng số ngữ nghĩa cao nhất.

Độ quan trọng câu được xác định bằng công thức: $w(V) = \alpha \times |X \cap T| + \beta \times |Y \cap T| + \gamma \times |Z \cap T|$, với |X| là số phần tử của X.

Như vậy ta có thể coi α, β, γ là các hệ số đặc trưng về ngữ nghĩa của từ khóa, từ khóa nhãn và thực thể có tên trong văn bản V. Thực tiễn ngữ nghĩa chúng ta có thể nhận thấy giá trị ngữ nghĩa của từ khóa cao hơn từ khóa nhãn và ngữ nghĩa của từ khóa nhãn cao hơn ngữ nghĩa của thực thể có tên. Do vậy ta về tương quan ngữ nghĩa ta có $\alpha > \beta > \gamma$. Để xác định giá trị phù hợp và đạt được hiệu quả cao khi sử dụng các hệ số này, ta cần phải có quá trình thực nghiệm với các kết quả của giải thuật hoặc áp dụng các phương pháp học máy. Do thời gian thực nghiệm chưa được nhiều nên chúng tôi tạm thời lấy giá trị cho các hệ số này sau một số lần đối sánh kết quả với các hệ số khác nhau trên 50 văn bản mẫu là $\alpha = 2$, $\beta = 1.5$, $\gamma = 1$.

Với mỗi từ x_i trong câu đều có thể đóng vai trò vừa là từ khóa, là từ gán nhãn và thực thể có tên, do vậy, chúng tôi chỉ lựa chọn trường hợp có trọng số cao nhất.

Ví dụ 02 câu: *Ngày 13/5, Tổng bí thư Nguyễn Phú Trọng và các đại biểu Quốc hội ứng cử tại Đơn vị bầu cử số 1 Đoàn đại biểu Quốc hội thành phố Hà Nội đã dành cả ngày tiếp xúc cử tri các quận Ba Đình, Hoàn Kiếm và Tây Hồ, chuẩn bị cho Kỳ họp thứ 5, Quốc hội khóa XIV. Cuộc tiếp xúc giữa Tổng Bí thư và các cử tri diễn ra ngay sau ngày bế mạc Hội nghị Trung ương 7 Ban Chấp hành Trung ương Đảng khóa XII, thông điệp lớn nhất của cử tri là bày tỏ niềm tin vào kết quả cuộc đấu tranh phòng chống tham nhũng trong thời gian vừa qua.*

Trong bài báo có tập các từ khóa "Tổng bí thư", "tiếp xúc", "cử tri";

Tập từ gán nhãn "phòng", "chống", "tham nhũng", "cán bộ", "Hà Nội".

Tập thực thể có tên của 02 câu này được xác định bao gồm: "Nguyễn Phú Trọng".

Trọng số w của câu 1 được tính như sau:

$w_1 = 2 \times 3 + 1.5 \times 1 + 1 \times 1 = 8.5$ do có chưa 03 từ khóa, 01 từ gán nhãn và 01 thực thể có tên.

Trọng số w của câu 2 được tính:

$w_2 = 2 \times 3 + 1.5 \times 3 + 1 \times 0 = 10.5$ do có chứa 03 từ khóa, 03 từ gán nhãn và 0 có thực thể có tên.

Sau khi tính độ quan trọng các câu, chúng tôi sẽ sắp xếp thứ tự các câu theo thứ tự giảm dần của trọng số w. Căn cứ vào tỉ lệ trích rút của người dùng hệ thống sẽ chọn từ trên xuống để sinh bản trích rút tương ứng.

Để thực hiện tiền xử lý tiếng Việt chúng tôi sử dụng thư viện VnCoreNLP [16] của Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras and Mark Johnson (2018) cho việc tách câu, tách từ, gán nhãn từ loại và nhận dạng thực thể có tên. Sau khi nhận dạng thực thể có tên, chúng tôi xây dựng tập thực thể có tên theo tiêu chí*.

Lưu đồ thuật toán trích rút câu sẽ được tiến hành như sau:

Với mỗi văn bản V:

1. Xác định tập từ khóa X.
    a. X = Words(Keywords) // sử dụng công cụ tách từ cho từ khóa chính thu thập được trong văn bản.
2. Xác định tập từ gán nhãn Y
    a. Y = Words(Tags) // sử dụng công cụ tách từ cho tập từ khóa gán nhãn thu thập được trong văn bản.
    b. $\forall\ y_i \in Y$ nếu $y_i \in X \rightarrow Y = Y \setminus \{y_i\}$
3. Xác định tập thực thể có tên Z
    a. Z = NER(V) // sử dụng công cụ nhận dạng thực thể có tên cho văn bản V
    b. Z = Z* // xây dựng Z* từ Z
    c. $\forall\ z_i \in Z$ nếu $z_i \in X \rightarrow Z = Z \setminus \{z_i\}$
    d. $\forall\ z_i \in Z$ nếu $z_i \in Y \rightarrow Z = Z \setminus \{z_i\}$
4. Đối với mỗi văn bản xác định tập câu S
    a. S = Sentences(V) // Sử dụng công cụ tách câu đối với văn bản V
5. Đối với mỗi câu $s_i \in S$
    a. T = Words(s_i) //Tách câu thành tập từ T
    b. Tính trọng số w_i của câu s_i : $w(s_i) = 2 \times |X \cap T| + 1.5 \times |Y \cap T| + |Z \cap T|$
6. Sắp xếp tập câu S theo trọng số w_i.
7. Lựa chọn số câu theo tỉ lệ người dùng cần trích rút để sinh bản trích rút.

IV. ĐÁNH GIÁ KẾT QUẢ TÓM TẮT

4.1. Xây dựng kho ngữ liệu

Như đã trình bày ở trên, hiện nay kho ngữ liệu dành cho tóm tắt văn bản còn khá hạn chế, ít được chia sẻ trong cộng đồng. Có một số kho ngữ liệu chia sẻ trên mạng Internet tuy nhiên kho những ngữ liệu hiện nay chưa có từ khóa của văn bản nên không sử dụng được trong bài toán này. Do vậy, chúng tôi bắt buộc phải xây dựng kho ngữ liệu thử nghiệm của riêng mình. Để xây dựng kho ngữ liệu trong bài báo này chúng tôi xác định phương pháp sau:

Lựa chọn ngẫu nhiên các bài báo từ các trang báo mạng điện tử Việt Nam gồm các trang http://dangcongsan.vn, https://news.zing.vn, https://vnexpress.net , đảm bảo mỗi bài báo có khoảng 500 từ trở lên. Mỗi bài báo sẽ được thu thập 04 nội dung gồm: tiêu đề, sa pô, nội dung, từ khóa và từ gán nhãn. Mỗi nội dung được lưu vào một file .txt tương ứng.

Đối với mỗi văn bản chúng tôi cũng xây dựng 01 bản trích rút giữ lại khoảng 30%, 01 bản trích rút giữ lại khoảng 60% số câu trong văn bản tương ứng là S30 và S60 để làm kết quả so sánh. Chúng tôi sử dụng chuyên gia là một nhà báo có kinh nghiệm để lựa chọn số câu trong mỗi văn bản. Để đảm bảo tính độc lập của kết quả, chuyên gia chỉ được cung cấp tiêu đề và nội dung văn bản, không được cung cấp thông tin về các từ khóa hay từ gán nhãn. Do việc xây dựng tập văn bản tóm tắt của chuyên gia mất nhiều thời gian, công sức nên trong bài báo này chúng tôi sử dụng trên tập 100 văn bản.

4.2. Phương pháp đánh giá thực nghiệm

Để đánh giá độ chính xác của bản trích rút tự động, chúng tôi sử dụng phương pháp Precision and recall. Phương pháp đánh giá này được sử dụng phù hợp với các bản tóm tắt theo hướng trích rút câu qua việc so sánh giữa bản tóm tắt do hệ thống trích rút với bản tóm tắt do con người trích rút sử dụng độ đo chính xác (precision), triệu hồi (recall), các giá trị f- score.

Độ đo chính xác (precision) là tỉ số giữa số lượng các câu được cả hệ thống và con người trích rút trên số các câu được hệ thống trích rút.

Độ đo triệu hồi (recall) là tỉ số giữa số lượng các câu được trích rút bởi hệ thống trùng với số các câu mà con người trích rút trên số các câu chỉ được lựa chọn bởi con người.

Độ đo F-score là một độ đo kết hợp giữa precision và recall. Ở đây chúng tôi quan tâm đến độ đo F_1-score được định nghĩa là trung bình hàm điều hòa của recision và recall. Các giá trị F-score nhận giá trị trong đoạn [0, 1], trong đó giá trị tốt nhất là 1.

$$Precision = \frac{|M \cap H|}{|M|}; \qquad Recall = \frac{|M \cap H|}{|H|}; \qquad F_1\text{-}score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

Trong đó: M là tập câu trích rút từ hệ thống, H là tập câu trích rút bởi chuyên gia, |M| là số phần tử của tập M.

Bảng 1. Đánh giá độ chính xác trên tập gồm 100 văn bản

	S30	S60
Precision	76,67%	76,79%
Recall	69,70%	71,67%
F_1-score	73,02%	74,14%

Từ **Bảng 1**, chúng tôi có một số nhận xét sau:

- Độ chính xác có kết quả khá tốt, trên 75%.

- Độ triệu hồi cũng cho kết quả khả quan, xấp xỉ 70%.

- Độ đo F_1-score có kết quả khá tốt trong cả hai trường hợp, chứng tỏ vai trò ngữ nghĩa của các đặc trưng trên trong văn bản báo mạng điện tử.

- Độ chính xác và độ triệu hồi giữa tập S30 và S60 là tương đối giống nhau, cho thấy tỉ lệ lựa chọn đúng không phụ thuộc quá nhiều vào tỉ lệ câu được trích xuất. Tuy nhiên, kết quả cũng cho thấy đối với tỉ lệ trích rút cao cho kết quả chính xác cao hơn.

Khi xem xét cụ thể từng bản trích rút do chuyên gia và do hệ thống thực hiện chúng tôi nhận thấy:

- Có sự khác biệt như sau: Bản trích rút của chuyên gia được lựa chọn đồng đều trong văn bản (các câu được chọn phân bố khá đều trong toàn bộ văn bản); bản trích rút do hệ thống lựa chọn có một số bài báo phân bố không đều, phần cuối nội dung bài báo thường ít được chọn.

- Số lượng câu không chứa ít nhất một trong ba đặc trưng là từ khóa, từ gán nhãn, thực thể có tên (câu có trọng số bằng 0) tương đối ít (14,3%).

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã đưa ra phương pháp tiếp cận tóm tắt trích rút đối với văn bản báo mạng điện tử dựa trên đánh giá độ quan trọng của từ khóa chính, từ khóa gán nhãn và thực thể có tên. Kết quả thu được từ thực nghiệm cho thấy vai trò của các đặc trưng này trong văn bản báo mạng điện tử và khẳng định đây là các tiếp cận có triển vọng trong việc trích rút câu đối với văn bản báo mạng điện tử.

Trong thời gian tới chúng tôi sẽ nâng cao hiệu quả của phương pháp này bằng cách mở rộng tập văn bản thử nghiệm và xác định các tham số ngữ nghĩa α, β và γ qua học máy. Đồng thời chúng tôi cũng sẽ nghiên cứu việc sử dụng 3 đặc trưng này của văn bản báo mạng điện tử kết hợp với các đặc trưng chung của văn bản tiếng Việt đã được nghiên cứu trước đây.

VI. LỜI CẢM ƠN

Chúng tôi chân thành gửi lời cám ơn tới nhà báo Trần Lệ Thủy phóng viên báo Phụ Nữ Việt Nam đã hỗ trợ chúng tôi trong quá trình nghiên cứu và xây dựng kho ngữ liệu cho bài báo này, chúng tôi cũng trân trọng gửi lời cám ơn nhóm tác giả thư viện VnCoreNLP.

VII. TÀI LIỆU THAM KHẢO

[1] Mani, I., House, D., Klein, G., et al. The TIPSTER SUMMAC Text Summarization Evaluation. In Proceedings of EACL, 1999.

[2] M.L. Nguyen, A. Shimazu, X.H. Phan, T.B. Ho, S. Horiguchi, Sentence Extraction with Support Vector Machine Ensemble. In Proceedings of the First World Congress of the International Federation for Systems Research : The New Roles of Systems Sciences For a Knowledge-based Society, 2005.

[3] Thanh Le Ha, Quyet Thang Huynh, Chi Mai Luong, A Primary Study on Summarization of Documents in Vietnamese, Proceedings of the First World Congress of the International Federation for Systems Research: The New Roles of Systems Sciences For a Knowledge-based Society, 2005.

[4] Đỗ Phúc, Hoàng Kiếm, Rút trích ý chính từ văn bản tiếng Việt. Tạp chí Công nghệ Thông tin và Truyền thông, 2006.

[5] Lê Thanh Hương, Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho tiếng Việt, Báo cáo tổng kết đề tài cấp KH và CN cấp bộ, Đại học Bách khoa Hà Nội, 2014.

[6] Nguyễn Thị Thu Hà, Phát triển một số thuật toán tóm tắt văn bản tiếng Việt sử dụng phương pháp học bán giám sát, Luận án Tiến sỹ, Học viện Kỹ thuật quân sự, 2012.

[7] Nguyễn Nhật An, Nghiên cứu, phát triển các kỹ thuật tự động tóm tắt văn bản tiếng Việt, Luận án Tiến sỹ Toán học, Viện Khoa học và Công nghệ Quân sự, 2015.

[8] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Manohar Paluri, Laurens van der Maaten, "Advancing state-of-the-art image recognition with deep learning on hashtags", https://code.facebook.com/posts/1700437286678763/advancing-state-of-the-art-image-recognition-with-deep-learning-on-hashtags/.

[9] Nguyễn Thị Trường Giang, Báo mạng điện tử - những vấn đề cơ bản, Nhà xuất bản Chính trị Quốc gia, 2014.

[10] Hoàng Anh, Những kỹ năng về sử dụng ngôn ngữ trong truyền thông đại chúng, Nhà xuất bản Đại học Quốc gia Hà Nội, 2008.

[11] Lê Thanh Hà, Cách thức tạo từ khóa (Keyword) trên báo điện tử Việt Nam, Luận văn Thạc sỹ chuyên ngành Báo chí học, Trường Đại học Khoa học xã hội và Nhân văn, 2016.

[12] Lâm Quang Tường, Phạm Thế Phi, Đỗ Đức Hào, Tóm tắt văn bản tiếng Việt tự động với mô hình sequence-to-sequence, Tạp chí Khoa học Trường Đại học Cần Thơ, pp.125-132, 2017.

[13] Nguyễn Trí Nhiệm, Nguyễn Thị Trường Giang, Báo mạng điện tử - đặc trưng và phương pháp sáng tạo, Nhà xuất bản Chính trị Quốc gia, 2014.

[14] https://en.oxforddictionaries.com/

[15] Nguyễn Ngọc Duy, Phan Thị Tươi, Tóm tắt văn bản trên cơ sở phân loại ý kiến độc giả của báo mạng tiếng Việt, Tạp chí Phát triển KH&CN, Tập 19, số K5-2016, 2016.

[16] https://github.com/vncorenlp.

THE FEATURES OF THE VIETNAMESE ONLINE NEWSPAPER IN TEXT SUMMARIZATION

Le Ngoc Thang, Le Quang Minh

ABSTRACT: *Text auto summarization have been studied since the 1950s of the 20th century. In Vietnamese language, it has only been focused on in the early years of the 21st century. At this time, these studies are short, single and focused on the direction of extracting using the features of the English language. Most of the results are tested in the online newspaper document. However, up to now, there have not been many studies on the linguistic features of Vietnamese online newspaper document for the sentence extraction. This article will study the features of Vietnamese online newspaper document and how to apply them in the pharse of automatic sentences extraction of text summarization.*