DOI: 10.15625/vap.2020.00220

TÓM TẮT VĂN BẢN BÁO MẠNG ĐIỆN TỬ TIẾNG VIỆT SỬ DỤNG TEXTRANK

Lê Ngọc Thắng[1,3], Phạm Bảo Sơn[2], Lê Quang Minh[3]

[1]Văn phòng Bộ Công an

[2]Đại học Quốc gia Hà Nội

[3]Viện Công nghệ thông tin, Đại học Quốc gia Hà Nội

lengocthang@gmail.com,sownpb@vnu.edu.vn, quangminh@vnu.edu.vn

TÓM TẮT: *Trong bài báo này chúng tôi đề xuất mô hình tóm tắt tự động văn bản tiếng Việt thể loại báo mạng điện tử. Văn bản được biểu diễn dưới dạng đồ thị, mỗi đỉnh của đồ thị biểu diễn một câu trong văn bản, trọng số các cạnh nối giữa các đỉnh biểu diễn sự tương tự về ngữ nghĩa giữa hai câu (đỉnh). Độ quan trọng của câu được xác định qua thuật toán TextRank, trong đó có bổ sung một số đặc trưng riêng của thể loại báo mạng điện tử. Hệ thống sẽ trích rút ra những câu quan trọng để đưa vào bản tóm tắt (mặc định 30 % số câu của văn bản). Để kiểm chứng mô hình đề xuất chúng tôi so sánh kết quả với kết quả tóm tắt của chuyên gia và kết quả của thuật toán TextRank cơ sở.*

Từ khóa: *Tóm tắt văn bản tiếng Việt, báo mạng điện tử, TextRank, tags.*

I. GIỚI THIỆU

Tóm tắt văn bản tự động đã được nghiên cứu từ những năm 1950 của thế kỷ XX. 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 [12] 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. Cách tiếp cận phân loại phổ biến nhất là theo kết quả (output). Theo cách phân loại này có tóm tắt theo phương pháp trích rút (Extract) và tóm tắt theo phương pháp tóm lược (Abstract). Trong đó tóm tắt theo phương pháp trích rút là bản tóm tắt bao gồm các đơn vị quan trọng như câu, đoạn được trích rút, chọn ra từ văn bản gốc; tóm tắt theo phương pháp tóm lược là bản tóm tắt bao gồm những khái niệm, nội dung được tóm lược từ văn bản gốc.

Hiện nay trên thế giới có nhiều công trình nghiên cứu về tóm tắt tự động văn bản cho nhiều ngôn ngữ khác nhau, tập trung mạnh nhất là đối với tiếng Anh, tiếng Nhật và tiếng Hoa. Về phương pháp tóm tắt phần lớn vẫn tập trung vào phương pháp trích rút với các mô hình đề xuất đa dạng và phong phú như: phương pháp sử dụng đặc trưng về tần suất từ TF×IDF, phương pháp phân cụm (cluster based), phương pháp phân tích ngữ nghĩa tiềm ẩn (LSA), phương pháp học máy (machine learning), mạng nơron (neural networks), dựa trên truy vấn (query based), hồi quy toán học (mathematical regression) hay mô hình đồ thị (graphical models).

Về lĩnh vực tóm tắt tự động văn bản tiếng Việt, hiện nay các nghiên cứu chủ yếu tập trung vào hướng trích rút với các mô hình sử dụng đặc trưng chung của văn bản tiếng Anh. Một số công trình tiêu biểu như 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 [1], 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 [1] 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 xuất thực từ, thực thể có tên, dữ liệu số, tương tự với tiêu đề và câu trung tâm để tính trọng số câu. Các nghiên cứu trên chủ yếu sử dụng trên tập dữ liệu là các văn bản báo mạng điện tử tiếng Việt nhưng chưa sử dụng các đặc trưng riêng của thể loại văn bản này như [10] đề cập.

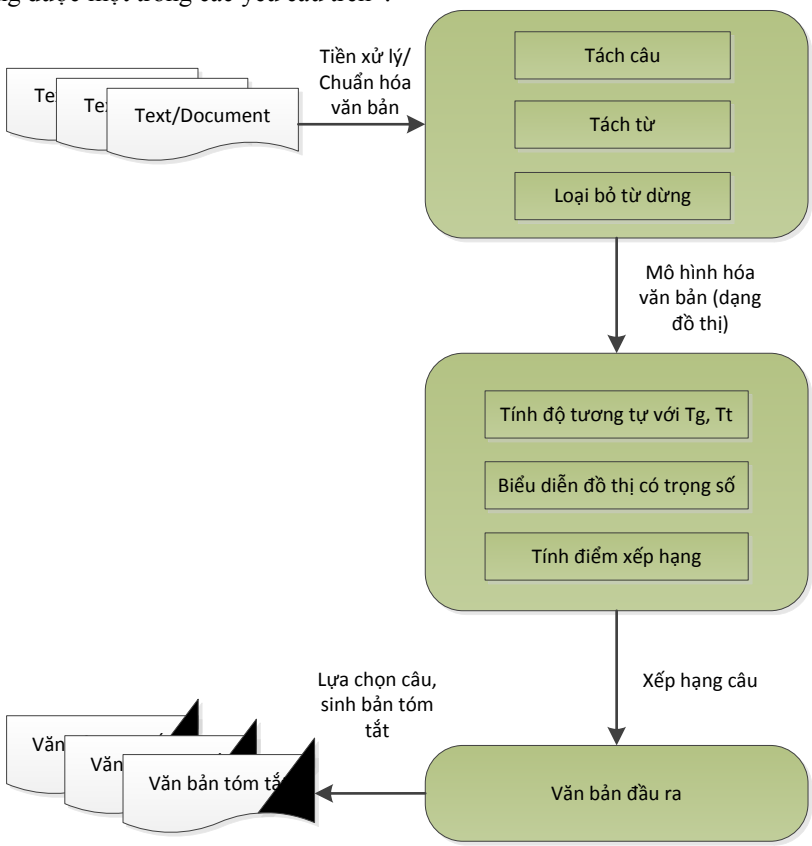
Vì vậy, trong nghiên cứu này chúng tôi đề xuất phương pháp tóm tắt tự động dựa trên phương pháp TextRank và bổ sung đặc trưng riêng của văn bản báo mạng điện tử tiếng Việt. Trong phần II của bài báo chúng tôi sẽ trình bày mô hình tóm tắt văn bản gồm các nội dung: vai trò từ khóa, từ gán nhãn (tags), mô hình TextRank được đề xuất trong bài báo này. Dữ liệu thực nghiệm, phương pháp đánh giá và kết quả sẽ được trình bày ở phần III. Phần IV sẽ trình bày kết luận và kiến nghị.

II. MÔ HÌNH TÓM TẮT

Báo mạng điện tử tiếng Việt đã phát triển qua ba giai đoạn. Hiện nay cấu trúc thông tin trong một bài báo mạng điện tử thường gồm 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. Sa pô là câu mào đầu của báo, có xu hướng càng ngắn gọn càng tốt, mục đích là để tạo sự hấp dẫn cho người đọc.

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 tít chính, trong sa pô là những thành phần mang nhiều thông tin trong văn bản. 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. Kết quả nghiên cứu tại [10] cũng đã chỉ rõ vấn đề này.

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 tít chính 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*.



Hình 1. Mô hình tóm tắt với TextRank được chúng tôi đề xuất

A. Tiền xử lý văn bản

Văn bản đầu vào có định dạng file *.txt. Văn bản sẽ được đưa qua bộ tiền xử lý văn bản để tách câu, tách từ và loại bỏ các từ dừng.

Để tách câu, tách từ chúng tôi sử dụng công cụ VnCoreNLP do nhóm tác giả Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras and Mark Johnson phát triển và xây dựng. Chúng tôi sử dụng công cụ này vì ngoài khả năng tách câu, tách từ mà còn cung cấp công cụ gán nhãn từ loại để phân biệt từ đơn, từ ghép và nhận biết các danh từ riêng (thực thể có tên) với độ chính xác khá cao.

Từ dừng (stopwords) được định nghĩa là những từ xuất hiện phổ biến trong văn bản nhưng không mang nhiều ngữ nghĩa trong phân tích ngôn ngữ học, hoặc xuất hiện rất ít trong tập ngữ liệu nên không đóng góp nhiều về mặt ý nghĩa. Vì vậy, việc loại bỏ từ dừng sẽ làm giảm độ nhiễu về ngữ nghĩa của các từ này của văn bản. Để loại bỏ từ dùng chúng tôi xây dựng một module so sánh các từ trong câu với danh sách từ dừng trong từ điển từ dừng tại https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords.txt. Nếu từ nào xuất hiện trong từ điển từ dừng thì loại khỏi câu trong văn bản.

B. Mô hình hóa văn bản dưới dạng đồ thị

Tính độ tương tự

Trong mô hình của chúng tôi, văn bản sau khi tiền xử lý sẽ được biểu diễn dưới dạng đồ thị vô hướng có trọng số. Mỗi đỉnh đồ thị tương ứng với một câu trong văn bản, mỗi cạnh nối hai đỉnh biểu thị mối quan hệ giữa hai câu. Trọng số cạnh là giá trị độ tương tự giữa hai câu. Đối với TextRank, phương pháp tính độ tương tự giữa câu là yếu tố căn bản ảnh hưởng đến kết quả của đầu ra. Những câu quan trọng nhất là những câu có độ tương tự đối với phần còn lại cao nhất. Phương pháp tính độ tương tự trong thuật toán gốc được xác định như sau:

Đối với văn bản D:

Gọi:

- $S = S_1, S_2, \ldots, S_n$, trong đó S_i là câu thứ i trong văn bản có n câu.

Với hai câu S_i và S_j sau khi đã được tiền xử lý, loại bỏ từ dừng, câu S_i được biểu diễn bởi tập n từ $w_1, w_2, \ldots, w_n$ thuật toán TextRank cơ bản xác định độ tương tự của S_i và S_j như sau:

$$Sim(S_i, S_j) = \begin{cases} \frac{|\{w_i | w_i \in S_i \ \& \ w_i \in S_j\}|}{\log(|S_i|)+\log(|S_j|)} \ nếu \ i \neq j \\ 0 \ nếu \ i = j \end{cases}$$

Để bổ sung ngữ nghĩa của từ gán nhãn và thực thể có thể có tên trong phương pháp tính độ tương đồng giữa hai câu, ta gọi:

- Tg là tập từ gán nhãn: $Tg = \{Tg_1, Tg_2 \ldots., Tg_m\}$.

- Tt là tập các thực thể có tên: $Tt = \{Tt_1, Tt_2, \ldots, Tt_k\}$.

Các tập Tg, Tt, sẽ được chuẩn hóa đảm bảo $Tg \cap Tt = \emptyset$, 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. Bằng việc gán trọng số ngữ nghĩa cho các từ khóa và thực thể có tên chúng tôi đề xuất công thức sau:

$$Sim'(S_i, S_j) = \begin{cases} \frac{|\{w_i | w_i \in S_i \ \& \ w_i \in S_i\}|+2*|\{w_i | w_i \in Tg\}|+|\{ w_i | \ w_i \in Tt\}|}{\log(|S_i|)+\log(|S_j|)} \ nếu \ i \neq j \\ 0 \ nếu \ i = j \end{cases}$$

Để đạt được hiệu quả cao khi sử dụng các hệ số này cần phải có một quá trình thực nghiệm trên nhiều bộ dữ liệu khác nhau hoặc qua quá trình học máy để xác định giá trị phù hợp của chúng. Do thời gian thực nghiệm chưa nhiều đồng thời việc hình thành bộ dữ liệu thực nghiệm cũng chiếm nhiều thời gian nên qua quá trình kiểm thử trên tập 50 văn bản chúng tôi chọn giá trị hệ số ngữ nghĩa cho từ gán nhãn là 3, cho thực thể có tên là 2.

Xếp hạng các câu quan trọng

Sau khi biển diễn văn bản dưới dạng đồ thị và tính toán ma trận độ tương tự thuật toán PageRank sẽ được áp dụng để tính toán giá trị mỗi đỉnh.

Giả sử với mỗi đỉnh V_i gọi $S(V_i)$ là trọng số của nó, phương trình quan hệ giữa đỉnh V_i và các đỉnh kề của nó được tính theo đồ thị vô hướng như sau:

$$S(v_i) = (1-d) + d * \sum_{v_j \in C(v_i)} \frac{Sim(v_i, v_j)}{\sum_{v_k \in C(v_i)} Sim(v_j, v_k)} * S(v_j)$$

Thuật toán khởi tạo giá trị trọng số ban đầu của mỗi đỉnh là 1, vòng lặp sẽ được thực hiện cho đến khi hội tụ, tức là sự thay đổi về trọng số của mỗi định nhỏ hơn một ngưỡng ε rất nhỏ, hoặc sau số lần lặp xác định. Điều kiện hội tụ được xác định thông qua quá trình thực nghiệm với ε = 0,001. Theo Lê Thanh Hương [1], đối với mô hình tóm tắt văn bản chúng tôi sử dụng hệ số d (DAMPING_FACTOR) của giải thuật PageRank là 0,85. Giá trị của mỗi đỉnh sau thuật toán PageRank biểu thị mức độ quan trọng của câu.

C. Chọn câu, sinh tóm tắt

Các câu sẽ được sắp xếp theo mức độ quan trọng giảm dần, sau đó sắp xếp lại theo thứ tự trong văn bản để sinh văn bản đầu ra. Ở đây chúng tôi sẽ lấy lần lượt các câu có trọng số từ cao xuống thấp trong đó số lượng câu được xác định thông quan tỉ lệ nén của văn bản tóm tắt, mặc định là 30 %. Các câu sau khi được đưa vào bản tóm tắt sẽ được sắp xếp lại theo thứ tự trong văn bản để có kết quả cuối cùng.

III. DỮ LIỆU THỰC NGHIỆM, ĐÁNH GIÁ KẾT QUẢ TÓM TẮT

A. Xây dựng kho ngữ liệu

Như đã trình bày ở trên, đối với bài toán tóm tắt văn bản tiếng Việt hiện có một số kho ngữ liệu chia sẻ trên mạng internet tuy nhiên kho những ngữ liệu hiện nay chưa có từ gán nhãn (tags) của văn bản nên không sử dụng được trong bài toán này. Do vậy, chúng tôi sử dụng kho ngữ liệu thử nghiệm của riêng mình đã được xây dựng tại [10]. Kho dữ liệu thử nghiệm này bao gồm 100 văn bản được 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.

Bản tóm tắt của mỗi văn bản được trích rút giữ lại 30 % số câu trong văn bản tạo thành tập kết quả chuyên gia. Chúng tôi phối hợp với chuyên gia là nhà báo có kinh nghiệm để lựa chọn câu trong bản tóm tắt.

B. Đánh giá thực nghiệm

Để đánh giá độ chính xác của bản trích rút tự động, chúng tôi sử dụng phương pháp đánh giá đồng chọn. Phương pháp đánh giá này 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 dựa trên ba đặc trưng là độ đo chính xác (precision), độ đo triệu hồi (recall) và độ đo f- score.

Độ đo chính xác (precision): Được tính dựa trên tỉ lệ giữa tổng số câu trùng nhau của văn bản tóm tắt thủ công và văn bản tóm tắt của hệ thống với tổng số câu văn bản tóm tắt của hệ thống.

Độ đo triệu hồi (recall): Được tính dựa trên tỉ lệ tổng số câu trùng nhau của văn bản tóm tắt thủ công và văn bản tóm tắt của hệ thống với tổng số câu của văn bản tóm tắt thủ công.

Độ đo F-score là một độ đo kết hợp giữa precision và recall. Người ta gọi F_1-score là một hàm điều hòa của của độ đo chính xác và độ đo triệu hồi. Các giá trị F_1-score nhận giá trị trong đoạn [0, 1], trong đó giá trị tốt nhất là 1.

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

trong đó: *SM* là tập câu trích rút từ hệ thống, *SH* là tập câu trích rút thủ công, |*SM*| là số phần tử của tập *SM*.

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

	Precision	Recall	F_1-score
Sim	0,640	0,601	0,620
Sim'	0,663	0,622	0,642

Từ Bảng 1, chúng tôi có một số nhận xét sau đối với kết quả trên tập dữ liệu thử nghiệm:

- Việc tính đến trọng số ngữ nghĩa của từ gán nhãn và thực thể có tên trong phương pháp tính độ tương đồng câu cho kết quả khả quan hơn tuy không nhiều.

- So sánh với tại [10] cho kết quả thấp hơn cho thấy việc áp dụng phương pháp TextRank vào tóm tắt văn bản báo mạng điện tử cần phải nghiên cứu để tiếp tục có phương pháp cải tiến.

Khi xem xét cụ thể từng bản trích rút do chuyên gia và do hệ thống thực hiện chúng tôi nhận thấy cũng giống như trong [10] các câu trong bản trích rút theo phương pháp TextRank cũng phân bố không đồng đều trong văn bản.

IV. KẾT LUẬN VÀ KIẾN NGHỊ

Bài báo đã đưa ra phương pháp tiếp cận tóm tắt trích rút đối với văn bản báo mạng điện tử dựa trên phương pháp TextRank có bổ sung một số đặc trưng riêng của báo mạng điện tử là từ gán nhãn và thực thể có tên. Kết quả thu được từ thực nghiệm cho thấy vai trò quan trọng về ngữ nghĩa của từ gãn nhãn và thực thể có tên trong bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt.

Trong thời gian tới chúng tôi sẽ tiếp tục thử nghiệm trên các tập dữ liệu khác nhau nhằm tối ưu hóa phương pháp tính độ tương đồng câu với từ gán nhãn và thực thể có tên, nâng cao hiệu quả của phương pháp này. Đồng thời chúng tôi cũng sẽ bổ sung giải pháp loại bỏ câu tương đồng nhằm hạn chế số lượng các câu có sự tương đồng cao nhưng có trọng số lớn cùng được lựa chọn vào bản tóm tắt.

V. LỜI CẢM ƠN

Chúng tôi chân thành gửi lời cám ơn tới nhà báo Trần Lệ Thủy - phóng viên báo Phụ Nữ Việt Nam, câu lạc bộ ngôn ngữ EQ 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.

TÀI LIỆU THAM KHẢO

[1] 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.

[2] 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.

[3] 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/.

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

[5] 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.

[6] 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.

[7] 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.

[8] Rada Mihalcea, Paul Tarau, "TextRank: Bringing Order into Texts", Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004.

[9] Lê Thị Ngọc Thơ, "Rút trích từ khóa từ văn bản pháp luật tiếng Việt bằng thuật toán TextRank", Hội nghị khoa học Đại học Công nghệ Tp. Hồ Chí Minh, 2019.

[10] Lê Ngọc Thắng, Lê Minh Quang, Kỷ yếu Hội nghị Quốc gia lần thứ XI về nghiên cứu cơ bản và ứng dụng công nghệ thông tin (FAIR), 2018.

[11] Trương Quốc Định, Nguyễn Quang Dũng, "Một giải pháp tóm tắt văn bản tiếng Việt", Hội thảo quốc gia lần thứ XV: Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông, 2012.

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

[13] Federico Barrios, Federico López, Luis Argerich, Rosita Wachenchauzer, "Variations of the Similiraty Function of TextRank for Automated Summarization", 44 JAIIO - ASAI 2015 - ISSN: 2451-7585, pages 65-72, 2016.

[14] 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.

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

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

VIETNAMESE ONLINE NEWSPAPERS SUMMURIZATION USING TEXTRANK

Le Ngoc Thang, Pham Bao Son, Le Quang Minh

ABSTRACT: *In this article we propose the model for summarizing automatically Vietnamese online newspapers. The text is represented graphically, each vertex represents one sentence in the text, the weight of the edges connecting two vertices represents the semantic similarity between these two sentences (vertices). The importance of the sentence is determined through the TextRank algorithm, which has added some specific features of the online newspapers. The system will extract important sentences to make the summary (default 30 % number of sentences in the documents). To verify the proposed model, we compare the results with the summurizations of the expert and the results of the basic TextRank algorithm.*