ISSN 2354-0575

TÓM TẮT VĂN BẢN TIẾNG VIỆT DỰA TRÊN PHƯƠNG PHÁP HỌC KHÔNG GIÁM SÁT

Nguyễn Hoàng Điệp*, Nguyễn Thị Hải Năng, Đỗ Thị Thu Trang, Ngô Thanh Huyền, Trịnh Thị Nhị

Trường Đại học Sư phạm Kỹ thuật Hưng Yên

** Diep82003@gmail.com, 0923 848 008*

Ngày tòa soạn nhận được bài báo: 16/11/2019

Ngày phản biện đánh giá và sửa chữa: 26/12/2019

Ngày bài báo được duyệt đăng: 29/12/2019

Tóm tắt:

Trong khi bài toán tóm tắt văn bản tiếng anh đã và đang được nghiên cứu rộng rãi với những kết quả đáng kinh ngạc thì bài toán tóm tắt văn bản tiếng việt vẫn chỉ đang ở giai đoạn đầu với kết quả nghiên cứu còn hạn chế. Bài này đề xuất một hướng giải quyết bài toán tóm tắt văn bản tiếng Việt tự động bằng cách mở rộng các phương pháp tóm tắt băn bản không giám sát kết hợp với điểm đánh giá mức độ quan trọng của câu cùng với mức độ liên quan của các câu.

Bài báo cung cấp kết quả thử nghiệm của việc mở rộng các phương pháp tóm tắt băn bản không giám sát kết hợp với điểm đánh giá mức độ quan trọng của câu bằng cách trích xuất các câu có xếp hạng hàng đầu, trong đó tránh chọn các câu trùng lặp về nội dung. Để kiểm chứng tính hiệu quả của phương pháp đề xuất nhóm thực hiện so sánh kết quả của nhóm với kết quả của phương pháp tóm tắt văn bản bằng học tập sâu là mạng nơ ron tích chập và mạng nơ ron hồi quy.

Sự mở rộng thay đổi của nhóm cho kết quả tốt nổi trội vì các lý do sau: Thứ nhất, việc nhặt các câu có điểm cao đảm bảo lựa chọn được các câu quan trọng. Thứ hai, việc lựa các câu có độ tương quan thấp đảm bảo các câu có nội dung giống câu đã lấy sẽ không được lấy lại, điều này đảm bảo nội dung bản tóm tắt không trùng lặp nội dung, trải rộng và bao quát được nội dung của bản gốc.

Từ khóa: *Tóm tắt văn bản, học máy, học không giám sát, xử lý ngôn ngữ tự nhiên, mạng nơ ron hồi quy, học sâu, mạng nơ ron tích chập.*

Chữ viết tắt

TT	Chữ viết tắt	Ý nghĩa
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
CNN	Convolutional Neural Network	Mạng nơ ron tích chập
LSTM	Long Short Term Memory	Mạng nơ ron hồi quy

1. Giới thiệu

Có một lượng thông tin khổng lồ có sẵn trên Internet và các tài nguyên khác như: sách, Twiter, Facebook, Youtube. Những nguồn thông tin này sẽ mang đến cho con người lượng kiến thức vô cùng quý báu nếu sử dụng được chúng. Một cơ chế trích xuất thông tin nhanh chóng và hiệu quả sẽ giúp con người chuyển những thông tin tồn tại thành thông tin hữu ích.

Tóm tắt văn bản tự động là một nhiệm vụ đầy thách thức nhưng thú vị của xử lý ngôn ngữ tự nhiên. Nhiệm vụ là tạo ra một bản tóm tắt súc tích từ một hoặc nhiều tài liệu. Đầu ra của một bản tóm tắt hệ thống mang lại lợi ích cho nhiều ứng dụng NLP như tìm kiếm trên web. Google thường trả về một mô tả ngắn về các trang web tương ứng cho một truy vấn tìm kiếm, hoặc các nhà cung cấp tin tức trực tuyến cung cấp các điểm nổi bật của một tài liệu Web trên giao diện của nó.

Hầu hết các cách tiếp cận cho bài toán tóm tắt văn bản tiếng việt là các phương pháp bán giám sát hoặc giám sát dựa trên các mô hình đồ thị

[2,3,4] hoặc xếp hạng dựa trên học giám sát dựa vào các thuốc tính [4, 20, 21].

Hình 1 Nguồn thông tin

Bài báo này tiếp cận theo hướng tóm tắt văn bản bằng cách lựa chọn các câu trong tài liệu nguồn để tạo nên bản tóm tắt mới [5,6,7,11]. Bài báo cung cấp kết quả thử nghiệm của việc mở rộng các phương pháp tóm tắt băn bản không giám sát kết hợp [10-15] với điểm đánh giá mức độ quan trọng của câu bằng cách trích xuất các câu có xếp hạng hàng đầu, trong đó tránh chọn các câu trùng lặp về nội dung.

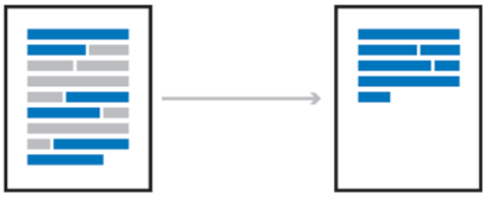
Phần còn lại của bài viết này được tổ chức như sau. Phần 2 cung cấp một số lý thuyết liên quan và dữ liệu chuẩn bị, cuối phần 2 trình bày cách thức nhóm tác giả đã thực nghiệm và đánh giá. Các kết quả và thảo luận được báo cáo trong phần 3. Cuối cùng, phần 4 rút ra kết luận và định hướng trong tương lai.

2. Cơ sở lý thuyết

2.1. Khái niệm cơ bản

2.1.1. Tóm tắt là một văn bản

Được tạo từ một hoặc nhiều văn bản, chứa một phần thông tin quan trọng trong các văn bản gốc và không dài hơn một nửa văn bản gốc.



Hình 2 Minh họa tóm tắt văn bản

2.1.2. Kỹ thuật trong tóm tắt văn bản

Học tập không giám sát: nhiệm vụ của học không có giám sát là tìm các mẫu chưa biết trước đó trong tập dữ liệu mà không có nhãn trước (tức là đầu ra đúng tương ứng cho mỗi đầu vào là không biết trước).

Học có giám sát: nhiệm vụ của học có giám sát là tìm ra một hàm ánh xạ dựa trên bộ dữ liệu huấn luyện, là các cặp dữ liệu (đầu vào-đầu ra mong muốn).

Học sâu: nhiệm vụ của học sâu là để tìm ra mô hình dữ liệu trừu tượng hóa ở mức cao bằng cách sử dụng một tập hợp các thuật toán với nhiều lớp xử lý với cấu trúc phức tạp.

Mạng nơ ron tích chập CNN: là một trong những mô hình học sâu tiên tiến, gồm có một hoặc nhiều hơn các lớp tích chập với các lớp đầy đủ kết nối (đáp ứng phù hợp với những mạng neuron nhân tạo tiêu biểu) trên đỉnh.

Mạng nơ ron hồi quy LSTM: là một trong những mô hình học sâu tiên tiến, một mạng cải tiến của RNN (Recurrent Neural Network) nhằm giải quyết vấn đề nhớ các bước dài của RNN.

2.1.3. Độ tương tự cosine

$$\cos(s_i, s_j) = \frac{s_i . s_j}{\|s_i\| . \|s_j\|}$$

Trong đó: s_i là vecto tương ứng với câu văn i

s_j là vecto tương ứng với câu văn j

$\|s_i\| . \|s_j\|$ là độ dài chuẩn của vecto s_i s_j

$\cos(s_i, s_j)$ là độ tương tự giữa 2 câu thứ i và j

Mức độ tương tự của hai câu được tính bằng cosine, cosine có giá trị thực từ 0 đến 1, cosine nhỏ tương ứng trường hợp nội dung của hai câu ít trùng lặp, hia câu có nội dung trùng lặp nhiều tương ứng với giá trị cosine lớn.

2.2. Dữ liệu

Nhóm tác giả chuẩn bị hai bộ dữ liệu cho nghiên cứu của mình để so sánh các phương pháp tóm tắt khai thác bằng tiếng Việt là VN-MDS và ViMs.

2.2.1. Bộ dữ liệu VN-MDS

Bộ dữ liệu được tạo bởi Trần và cộng sự, tại Phòng thí nghiệm của Đại học Quốc gia Hà Nội [20]. Bộ dữ liệu bao gồm các tài liệu về 200 chủ đề khác nhau được thu thập từ các nhà cung cấp tin tức trực tuyến Việt Nam. Mỗi chủ đề có hai đến năm bài viết khác nhau, thường là ba bài khác nhau. Cùng với các bản tóm tắt (gồm các câu quan trọng) được lựa chọn ra từ tài liệu gốc bởi các chuyên gia.

2.2.2. Bộ dữ liệu ViMs

Bộ dữ liệu được tạo bởi thạc sĩ Nghiêm tại Đại học Khoa học Tự nhiên, Đại học Quốc gia Hồ Chí Minh [21]. Bộ dữ liệu chứa tài liệu về 300 chủ đề khác nhau được thu thập từ Google News. Mỗi chủ đề có năm đến mười bài viết khác nhau. Cùng với các bản tóm tắt (gồm các câu quan trọng) được lựa chọn ra từ tài liệu gốc bởi các chuyên gia.

2.2.3. Thống kê quan sát dữ liệu

Tên bộ dữ liệu	Số chủ đề	Số văn bản	Tổng số câu	Số bản tóm tắt	Độ dài trung bình câu
VN-MDS	200	600	9802	400	49.2
ViMs	300	1945	25100	600	83.6

Hình 3. Thống kê quan sát 2 bộ dữ liệu

Có thể thấy rằng số lượng tài liệu và câu trong ViMs lớn hơn nhiều so với VN-MDS. Ngoài ra, chiều dài các tài liệu trong ViMs dài hơn gần hai lần so với trong các tài liệu của VN-MDS.

2.3. Phương pháp học tập không giám sát

Nhóm tác giả sử dụng sáu phương pháp xếp hạng nổi tiếng của bộ công cụ sumy để thực hiện việc tóm tắt văn bản.

***2.3.1.* Thuật toán LSA**

Ứng dụng sự phân rã của ma trận từ-câu bằng cách sử dụng Phân tách giá trị số ít để tóm tắt. Bằng cách này, chúng ta có thể có được các chủ đề ẩn và hình chiếu của mỗi câu theo chủ đề [16]. Thuật toán sử dụng giá trị tham chiếu là điểm số để phản ánh tầm quan trọng của câu.

2.3.2. **Thuật toán LexRank**

Thuật toán xây dựng một đồ thị ngẫu nhiên để tính toán tầm quan trọng tương ứng của các câu quan trọng [10]. Trong phương pháp này, các câu quan trọng được xác định bằng cách sử dụng mô hình.

2.3.3. **Thuật toán TextRank**

Thuật toán kế thừa sự tính toán của thuật toán PageRank, trong đó mà một câu văn bản là quan trọng nếu nó nhận được nhiều liên kết (tương tự điểm số) từ những người khác [17]. TextRank sử dụng cấu trúc văn bản bên trong các tài liệu và tạo ra một cụm đồ thị từ khóa trung tâm để xếp hạng các câu, sau đó trích các câu có trọng số cao để tạo thành một bản tóm tắt.

2.3.4. **Thuật toán Luhn**

Thuật toán trích các câu quan trọng bằng cách đo các thành phần quan trọng, thành phần quan trọng có chứa các từ xuất hiện nhiều hoặc thuộc câu ở vị trí quan trọng như câu đầu hay cuối [5].

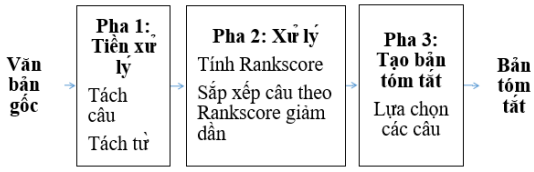
2.3.5. **Thuật toán KL**

Thuật toán đo lường sự khác biệt của phân phối xác suất unigram đã học được từ các tài liệu gốc và bản tóm tắt dựa trên KL Divergence [18].

2.3.6. **Thuật toán SumBasic**

Thuật toán sử dụng sự đơn giản hóa câu và chọn lựa từ vựng để tóm tắt [19].

2.4. Quá trình thực nghiệm



Hình 4 Tổng quan phương pháp

Pha 1: Tiền xử lý dữ liệu

Nhóm tác giả tiền xử lý dữ liệu bằng cách bóc tách lấy nội dung từ các tệp dữ liệu trong bộ dữ liệu, tách mỗi câu trên một dòng. Sau đó sử dụng bộ công cụ vitk của TS Lê Hồng Phương để thực hiện tách từ trong nội dung văn bản.

Pha 2: Xử lý dữ liệu

Lấy điểm đánh giá mức độ quan trọng rankscore của các câu, bằng cách thay đổi mở rộng mã nguồn của bộ công cụ mã nguồn mở sumy. Sau đó sắp xếp lại các câu trong văn bản theo độ quan trọng của các câu giảm dần dựa vào điểm đánh giá giảm dần. Công việc này được thực hiện với từng thuật toán trình bày trong phần 2.3.

Pha 3: Tạo ra bản tóm tắt

Các câu đưa vào bản tóm tắt dựa vào *rankscore* - điểm đánh giá mức độ quan trọng của câu, *cosine* - mức độ liên quan của câu với các câu đã lựa chọn và tham số *threshold* - ngưỡng. Cụ thể như sau:

Bước 1: Lấy kết quả từ pha 2, chọn câu có điểm rankscore cao nhất (câu quan trọng nhất).

Bước 2: Lặp lại công việc như sau

Lần lượt xét các câu còn lại, nếu câu xét s_i có

độ dài trên 5 từ và không trùng lặp nội dung với các câu đã được chọn đưa vào bản tóm tắt, thì đưa câu này vào bản tóm tắt. Các câu có nội dung không trùng lặp, nếu nó thỏa mãn ràng buộc: *max* (*cosine(*s_i , s_k)) < *threshold,* với s_k là câu đã được chọn đưa vào bản tóm tắt.

Thuật toán dừng khi tóm tắt đạt đến một ràng buộc chiều dài.

Threshold được xác định bằng thực nghiệm trên hai bộ dữ liệu VN-MDS và ViMs. Nhóm đã thử nghiệm *Threshold* trong tập giá trị trong khoảng (0,1) bước nhảy là 0..05 là {0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45,…0.95}, từ kết quả cho thấy *Threshold=0.4* cho kết quả ổn định nhất

Về chiều dài của bản tóm tắt, nhóm thử nghiệm với chiều dài khoảng 10 câu tức là 100 từ (khoảng 10 câu), và thử nghiệm với trường hợp không giới hạn độ dài bản tóm tắt, đều cho kết quả khá tốt (bảng 1,2,3 và 4).

Pha 4: Đánh giá

Mỗi bộ dữ liệu được chia làm 5 phần một cách hoàn toán ngẫu nhiên. Sử dụng bộ công cụ *ROUGE_1.5.5,* các câu được trích xuất vào bản tóm tắt được so sánh với các câu trong bản tóm tắt của các chuyên gia (các câu trong gold files) theo ROUGE_N (N = 1,2 và ROUGE_SU4)

Huấn luyện CNN và LSTM

Huấn luyện CNN, nhóm tác giả sử dụng ba nhân, hai tầng ẩn kết nối đầy đủ với kích thước là 20 và 1. Huấn luyện LSTM, nhóm sử dụng mô hình LTSM cell đơn giản mặc đạnh, kích thước của vecto đầu ra là 100.

Trên hai bộ dữ liệu tiếng việt, nhóm lấy độ dài của câu tiếng việt dài nhất là 30 từ. Huấn luyện cả hai mô hình với kích thước dữ liệu chia lô batch size là 32, số lần lặp khi huấn luyện trên lô là epochs là 25.

3. Kết quả nghiên cứu và thảo luận

Chương trình thực nghiệm được viết bằng ngôn ngữ lập trình python trên siêu máy tính UTEHY 1 đặt tại cơ sở Mỹ Hào trường ĐHSP Kỹ thuật Hưng Yên.

Nhóm đã thử nghiệm trên hai bộ dữ liệu tiếng việt ViMs và VN-MDS, so sánh kết quả và lựa chọn ra giá trị ngưỡng thích hợp là 0.4.

Có vài điểm nổi bật từ các kết quả (bảng 1). Đầu tiên, với bản tóm tắt khoảng 10 câu (100 từ) trên bộ dữ liệu VN-MDS. Thứ 2, các thuật toán học không giám sát mở rộng cho kết quả tốt hơn so với học sâu. Thứ ba, thuật toán học Sumbasic-một trong những thuật toán học không giám sát sau khi mở cho kết quả tốt hơn các thuật toán khác trên bộ dữ liệu VN-MDS.

Một điểm nổi bật từ các kết quả bảng 2 là kết quả tương tự trong bảng 1, điều này nói nên rằng thuật toán cho kết quả tốt với dữ liệu tiếng viện với bản tóm tắt khoảng 100. Tiếp theo, các thuật toán học không giám sát mở rộng vẫn cho kết quả tốt hơn so với học sâu.

Bảng 1: So sánh kết quả với độ dài bản tóm tắt 100 từ trên bộ dữ liệu VN-MDS

Phương pháp	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**
LSA	0.629	0.370	0.558
LexRank	0.643	**0.406**	0.581
TextRank	0.629	0.398	0.565
Luhn	0.612	0.368	0.550
KL	0.651	0.380	0.571
Sumbasic	**0.665**	0.394	**0.585**
CNN	0.614	0.366	0.528
LSTM	0.616	0.355	0.535

Bảng 2: So sánh kết quả với độ dài bản tóm tắt 100 từ trên bộ dữ liệu ViMs

Phương pháp	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**
LSA	0.625	0.360	0.538
LexRank	0.641	0.394	**0.564**
TextRank	0.627	0.388	0.544
Luhn	0.614	0.376	0.534
KL	0.651	0.378	0.559
Sumbasic	**0.677**	**0.390**	0.572
CNN	0.591	0.342	0.491
LSTM	0.624	0.351	0.529

Bảng 3: So sánh kết quả ROURGE-scores độ dài không giới hạn trên bộ dữ liệu VN-MDS

Phương pháp	ROUGE -1	ROUGE -2	ROUGE-SU4
LSA	0.492	0.392	0.208
LexRank	0.482	0.392	0.198
TextRank	0.447	0.374	0.166
Luhn	0.439	0.372	0.159
KL	**0.602**	0.404	0.343
Sumbasic	0.574	**0.409**	**0.305**
CNN	0.528	0.400	0.248
LSTM	0.525	0.396	0.244

Kết quả trong bảng 3 cho kết quả tương tự bảng 1 và 2, các thuật toán học không giám sát thể hiện kết quả tốt so với học sâu trên cả hai bộ dữ liệu trong trường hợp không giới hạn số từ trong bản tóm tắt.

Bảng 4: So sánh kết quả ROURGE-scores độ dài không giới hạn trên bộ dữ liệu ViMs

Phương pháp	ROUGE -1	ROUGE -2	ROUGE-SU4
LSA	**0.711**	0.445	**0.503**
LexRank	0.695	**0.464**	0.477
TextRank	0.664	**0.464**	0.433
Luhn	0.636	0.454	0.393
KL	0.697	0.411	0.474
Sumbasic	0.697	0.426	0.469
CNN	0.561	0.421	0.296
LSTM	0.707	0.431	0.495

Trên bộ dữ liệu ViMs và không giới hạn số từ trong bản tóm tắt thì LSA thể hiện kết quả tốt nổi trội so với các thuật toán học không giám át khác cũng như học sâu. Sau LSA thì Học sâu với mạng hồi quy LSTM tuy chưa cho kết quả tốt như LSA nhưng cũng cho kết quả tốt hơn so với cá thuật toán khác.

Theo kết quả thực nghiệm (tử cả 4 bảng dữ liệu), nhóm tác giả tìm thấy hai điểm nổi bật như sau: Thứ nhất, với mở rộng bộ công cụ sumy cho một số phương pháp học không giám sát sẽ mang lại kết quả tốt trong nhiều trường hợp. Tất nhiên, không có phương pháp nào đạt được kết quả tốt nhất trong mọi trường hợp. Điểm nổi bật thứ hai là độ dài của bản tóm tắt (bảng 1-2 với giới hạn 100 từ và bảng 3-4 không giới hạn số từ trong bản tóm tắt) cho thấy có mối quan hệ giữa độ dài câu và điểm đánh giá ROUGE.

4. Kết luận

Nói chung, các thuật toán và dữ liệu tốt là rất quan trọng. Bài viết này thực hiện các thí nghiệm tóm tắt văn bản tiếng việt. Nhóm tác giả khai thác mở rộng cải tiến dựa trên các phương pháp học không giám, để kiểm nghiệm hiệu quả của sự cải tiến phương pháp, nhóm so sánh với hai phương pháp học sâu.

Lời cảm ơn Nghiên cứu này được tài trợ bởi Trường Đại học Sư phạm kỹ thuật Hưng Yên trong đề tài mã số UTEHY.L.2019.53.

Tài liệu tham khảo

[1] 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 sĩ, 2012.

[2] Đỗ Phúc, Hoàng Kiếm, "Rút trích ý chính từ văn bản tiếng Việt hỗ trợ tạo tóm tắt nội dung".

[3] Nguyễn Thị Ngọc Tú , Nguyễn Thị Thu Hà , Lê Thanh Hương , Hồ Ngọc Vinh , Đào Thanh Tĩnh, Nguyễn Ngọc Cương, "Ứng dụng mô hình độ thị trong tóm tắt đa văn bản tiếng việt, (FAIR) 2015.

[4] Trương Quốc Định, Nguyễn Quang Dũng, Một giải pháp tóm tắt văn bản tiếng Việt tự động, FAIR 2012.

[5] H. P. Luhn, "The automatic creation of literature abstracts," IBM Journal of Research Development, 2(2): 159-165, 1958.

[6] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," in IJCAI: 2862-2867, 2007.

[7] T.-A. Nguyen-Hoang, K. Nguyen, and Q.-V. Tran, "Tsgvi: a graph-based summarization system for vietnamese documents," Journal of Ambient Intelligence and Humanized Computing, 3(4), pp.305-312, 2012.

[8] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, "Ranking with recursive neural networks and its application to multi-document summarization," in AAAI: 2153-2159, 2015.

[9] Nguyễn Minh Tiến, Nguyễn Thị Hải Năng, Nguyễn Hoàng Điệp, Nguyễn Văn Hậu "Learning to Estimate the Importance of Sentences for Multi-Document Summarization", International Conference on Knowledge and Systems Engineering (KSE), in IEEE, 31-36, 2018.

[10] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, 22: 457-479, 2004.

[11] K. Woodsend and M. Lapata, "Automatic generation of story highlights," in ACL: 565-574, 2010.

[12] J. A. B. Hui Lin, "A class of submodular functions for document summarization," in ACL: 510-520, 2011, June.

[13] K. Woodsend and M. Lapata, "Multiple aspect summarization using integer linear programming," in EMNLP-CoNLL: 233-243, 2012.

[14] S. Banerjee, P. Mitra, and K. Sugiyama, "Multi-document abstractive summarization using ilp based multi-sentence compression," in IJCAI: 1208-1214, 2015.

[15] M.-T. Nguyen, T. V. Cuong, N. X. Hoai, and M.-L. Nguyen, "Utilizing user posts to enrich web document summarization with matrix cofactorization," in SoICT, pp. 70-77, 2017.

[16] Y. Gong and X. Liu, "Generic text summarization using relevant measure and latent semantic analysis," in SIGIR: 19-25, 2001.

[17] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in Association for Computational Linguistics, 2004.

[18] S. Sripada and J. Jagarlamudi, "Summarization approaches based on document probability distributions," in PACLIC: 521-529, 2009.

[19] L. Vanderwendea, H. Suzukia, C. Brocketta, and A. Nenkova, "Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion," Information Processing & Management 43, 6 (2007), pp. 1606-1618. Elsevier, 2007.

[20] V.-G. Ung, A.-V. Luong, N.-T. Tran, and M.-Q. Nghiem, "Combination of features for vietnamese news multi-document summarization," in The Seventh International Conference on Knowledge and Systems Engineering (KSE), pp. 186-191, 2015.

[21] H. Nguyen, T. Le, V.-T. Luong, M.-Q. Nghiem, and D. Dinh, "The combination of similarity measures for extractive summarization," in Proceedings of the Seventh Symposium on Information and Communication Technology (SoICT): 66-72, 2016.

VIETNAMESE MULTI-DOCUMENT SUMMARIZATION BASE UNSUPERVISED LEARNING METHODS

Abstract:

Recently, English summarization has been amazing results, while Vietnamese summarization has been being at an early stage with limited results. This paper proposes a solution to summarize Vietnamese text by utilizing unsupervised learning.

The article shows the results of employing unsupervised learning methods to summarize a document. To do that, the authors compared results of unsupervised learning methods for summarization to supervised learning ones, including CNN and LSTM. The comparison can demonstrate the effectiveness of unsupervised learning methods for summarization.

Unsupervised learning methods give promising empirical results because of some reasons. Firstly, based on ranking mechanisms, they pick up high-scoring sentences, which ensure the selection of important sentences. Secondly, the selection of sentences with low correlation shows that a summary text does not overlap with remaining sentences, which are not included in the summary.

Keywords: *Text summary, machine learning, learning to rank, unsupervised learning method, NLP, CNN, LSTM.*