LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của cá nhân tôi. Các số liệu, kết quả trong luận án là trung thực và chưa từng công bố trong bất kỳ công trình nào khác. Các kết quả nghiên cứu của tôi cùng với các tác giả khác đã được sự nhất trí của các đồng tác giả khi đưa vào nội dung luận án. Tôi đã trích dẫn đầy đủ các tài liệu tham khảo, công trình nghiên cứu liên quan ở trong nước và quốc tế.

Tác giả

Lê Ngọc Thắng

LỜI CẢM ƠN

Luận án được thực hiện tại Viện Công nghệ thông tin – Đại học Quốc gia Hà Nội, dưới sự hướng dẫn khoa học của PGS.TS Phạm Bảo Sơn và TS. Lê Quang Minh.

Trước tiên Tôi xin bày tỏ lòng biết ơn sâu sắc tới tập thể giáo viên hướng dẫn, những người đã đưa tôi đến với lĩnh vực nghiên cứu này. Các thầy đã tận tình giảng dạy, hướng dẫn giúp tôi tiếp cận và đạt được thành công trong các nghiên cứu của mình; luôn tận tâm động viên, khuyến khích và chỉ dẫn giúp tôi hoàn thành được bản luận án này.

Tôi xin cảm ơn PGS.TS Nguyễn Minh Tiến, TS. Nguyễn Chí Thành, nhà báo Trần Lệ Thủy đã chia sẻ kinh nghiệm, tài liệu và hỗ trợ trong quá trình thực hiện luận án này.

Cuối cùng, tác giả xin chân thành cảm ơn các thành viên trong Gia đình, những người luôn dành cho tác giả những tình cảm nồng ấm và sẻ chia những lúc khó khăn trong cuộc sống, luôn động viên giúp đỡ tác giả trong quá trình nghiên cứu. Luận án cũng là món quà tinh thần mà tác giả trân trọng gửi tặng đến các thành viên trong Gia đình.

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CHỮ CÁI VIẾT TẮT

ATS	*Automatic Text Summarization – Hệ thống tóm tắt văn bản tự động*
BART	*Bidirectional and Auto-Regressive Transformers*
BERT	*Bidirectional Encoder Representations from Transformers*
D	*Văn bản tóm tắt*
LSA	*Latent Semantic Analysis - Phân tích ngữ nghĩa tiềm ẩn*
MMR	*Maximal Marginal Relevance - Mức độ liên quan cận biên tối đa*
NER	*Named Entity Recognition – Thực thể có tên*
NMF	*Non-negative Matrix Factorization – Phân tử hóa ma trận không âm*
NLP	*Natural Language Processing – Xử lý ngôn ngữ tự nhiên*
PhoBERT	*Phở BERT*
RNNs	*Recurrent Neural Network – Mạng nơ ron hồi quy*
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation - Độ đo đánh giá độ tương tự văn bản*
RST	*Rhetorical Structure Theory - Lý thuyết cấu trúc tu từ*
Pre-trained model	*Mô hình huấn luyện trước*
S	*Câu trong văn bản*
TF	*Term Frequency - Tần suất của từ*
TF.ISF	*Term frequency. Inverse sentence frequency - Tần suất của từ. Nghịch đảo tần suất câu*
Wordnet	*Mạng từ*

DANH MỤC CÁC HÌNH

DANH MỤC CÁC BẢNG

MỞ ĐẦU

1. Tình hình hoạt động phức tạp trên Internet hiện nay

Theo thống kê chưa đầy đủ đến cuối năm 2015, có khoảng 380 báo, 9 tạp chí và 60 đài phát thanh tiếng Việt trên thế giới và 400 trang web, tạp chí điện tử, các tài khoản mạng xã hội (Facebook, Twitter...) và blog cá nhân trong nước tán phát tài liệu xuyên tạc, kích động dư luận xã hội. Về báo chí, Việt Nam có 138 báo điện tử [1], 1600 trang thông tin điện tử, 420 mạng xã hội, diễn đàn. Một số báo điện tử vẫn để xảy ra tình trạng đăng tin, bài có nội dung nhạy cảm, thiếu cân nhắc trong sử dụng từ ngữ, hình ảnh; đưa tin thiếu khách quan, không đúng sự thật, phát triển theo hướng câu khách, rẻ tiền. Một số tạp chí lách luật để tự sản xuất tin tiềm ẩn nhiều nguy cơ mất an toàn, an ninh thông tin, vì đây là kênh lan truyền thông tin nhanh chóng tới người dùng, nhất là các tin đồn thất thiệt.

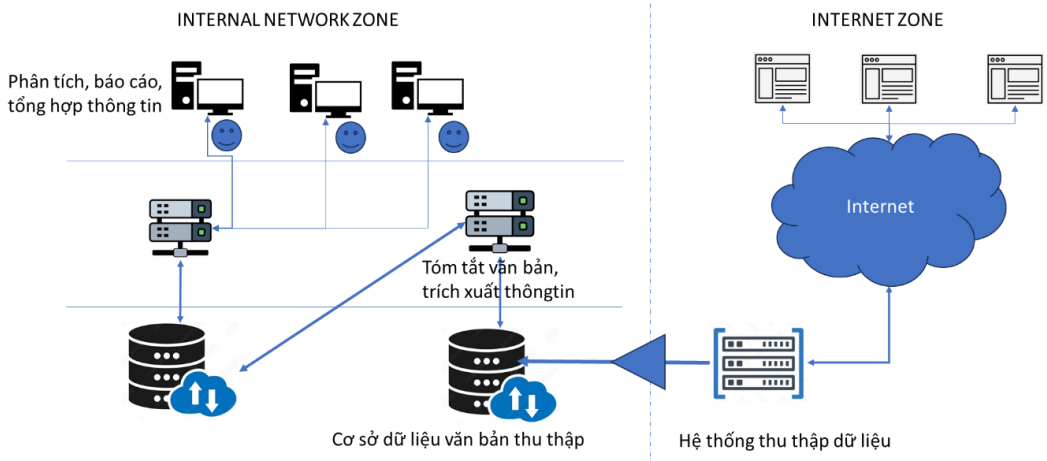
Từ thực tiễn đó, cho thấy yêu cầu xây dựng hệ thống thông tin với mục tiêu quản lý thông tin trên mạng Internet, trong đó có nhiệm vụ về quản lý dữ liệu báo mạng điện tử là cấp thiết để phục vụ công tác quản lý thông tin truyền thông.

2. Hiện trạng công tác thu thập thông tin

Với số lượng hàng nghìn trang báo điện tử, trang thông tin điện tử và các trang web tiếng Việt hiện nay, nhưng cơ quan quản lý phải theo dõi, giám sát, tổng hợp thông tin một cách thủ công do chưa có công cụ hỗ trợ nên việc theo dõi dòng thông tin chính trên báo chí và các trang thông tin điện tử rất khó khăn. Thực trạng trên cho thấy việc xây dựng hệ thống thu thập thông tin tự động trên Internet, có khả năng xử lý thông tin lớn, theo thời gian thực, có khả năng tự phân tích, tổng hợp văn bản tiếng Việt từ các nguồn khác nhau trong đó có các trang báo mạng điện tử tiếng Việt nhằm hỗ trợ công tác của cơ quan quản lý nhà nước là rất cấp thiết. Để giải quyết bài toán này, hệ thống cần đáp ứng các yêu cầu cơ bản sau:

- Tự động thu thập thông tin từ các trang thông tin tổng hợp, báo điện tử trong nước có lượng truy cập lớn, có tác động ảnh hưởng lớn tới xã.

- Xây dựng công cụ hỗ trợ cơ quan quản lý tóm tắt, trích xuất, phân tích, tổng hợp, đánh giá nội dung thông tin trên các trang thông tin tổng hợp, báo điện tử.



Hình 1. Hệ thống thu thập, phân tích và xử lý thông tin trên mạng Internet.

[1] https://vi.wikipedia.org/wiki/Danh_sách_báo_mạng_điện_tử_tiếng_Việt (số liệu tính đến năm 2022)

Do đặc thù liên quan đến công tác của cơ quan quản lý, hệ thống trên phải đảm bảo tuyệt đối an toàn và tách biệt với mạng Internet nên có những đặc điểm về mặt an toàn thông tin, an ninh mạng như sau: (1) Thông tin được thu thập trực tuyến (online) trên các trang báo mạng điện tử quan Hệ thống thu thập dữ liệu đặt ở vùng mạng ngoài (Internet). (2) Sau khi thu thập, tiền xử lý dữ liệu, văn bản sẽ được cập nhật, lưu trữ vào vùng trong (Vùng mạng riêng của cơ quan quản lý hệ thống) chỉ kết nối với hệ thống Thu thập dữ liệu thông qua kết nối 1 chiều (sử dụng data diode); không có kết nối chiều ra từ vùng mạng trong đến Internet. (3) Hệ thống tóm tắt văn bản, trích xuất thông tin được thực hiện hoàn toàn tại vùng trong, không kết nối Internet.

Xuất phát từ nhu cầu và thực tiến đó tôi đề xuất nghiên cứu đề tài "*Nghiên cứu, phát triển kỹ thuật tóm tắt văn bản tiếng Việt phục vụ công tác thu thập, xử lý thông tin lan truyền trên mạng internet*" tại Viện Công nghệ thông tin - Đại học Quốc gia Hà Nội.

3. Đối tượng, phạm vi nghiên cứu

Đối tượng nghiên cứu của Luận án: Các phương pháp tóm tắt văn bản trên thế giới; Các phương pháp tóm tắt văn bản tiếng Việt; Các đặc trưng quan trọng của văn bản báo mạng điện tử tiếng Việt; Kho ngữ liệu huấn luyện tóm tắt văn bản; Các phương pháp đánh giá tóm tắt văn bản.

Phạm vi nghiên cứu của Luận án: Luận án tập trung nghiên cứu, đề xuất phương pháp mới nâng cao độ chính xác trong bài toán tóm tắt đơn văn bản báo mạng điện tử tiếng Việt theo hướng trích rút.

4. Mục tiêu nghiên cứu

Mục tiêu của luận án là nghiên cứu các đặc trưng quan trọng của văn bản báo mạng điện tử cho bài toán tóm tắt đơn văn bản tiếng Việt. Qua đó đề xuất hai phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt: *Một là*, phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt dựa trên đồ thị và bộ hệ số đặc trưng văn bản; *Hai là*, phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt sử dụng mô hình huấn luận trước (pre-trained model).

Mục tiêu cụ thể: (1) Nghiên cứu các đặc trưng quan trọng của văn bản báo mạng điện tử tiếng Việt, qua đó đề xuất lựa chọn tập đặc trưng để đưa vào mô hình. (2)Đề xuất phương pháp tính độ tương tự câu trong văn bản báo mạng điện tử tiếng Việt dựa trên các đặc trưng quan trọng. (3) Đề xuất phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt dựa trên đồ thị và bộ hệ số đặc trưng văn bản. (4) Đề xuất phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt sử dụng mô hình huấn luận trước (pre-trained model).

5. Phương pháp nghiên cứu

Phương pháp nghiên cứu của luận án kết hợp nghiên cứ lý thuyết với nghiên cứu, kiểm chứng kết quả các phương pháp đề xuất bằng thực nghiệm.

Về lý thuyết: Nghiên cứu các công trình khoa học trong và ngoài nước liên quan đến bài toán tóm tắt văn bản gồm các phương pháp tiếp cận truyền thống và phương pháp dựa trên các mô hình học sẵn. Phân tích ưu, nhược điểm của các kỹ thuật đã có, từ đó đề xuất cải tiến kỹ thuật trên.

Về thực nghiệm: Thu thập dữ liệu các bài báo mạng điện tử, tiến hành xử lý dữ liệu để xây dựng kho ngữ liệu thử nghiệm phục vụ đánh giá các phương pháp đề xuất. Sử dụng các phương pháp đánh giá đã được cộng đồng nghiên cứu trên thế giới chấp thuận để phân tích và đánh giá kết quả các kỹ thuật đã đề xuất.

6. Nội dung nghiên cứu

(1) Nghiên cứu và đề xuất lựa chọn các đặc trưng quan trọng cho bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt bằng phương pháp khảo sát trên kho ngữ liệu văn bản báo mạng điện tử tiếng Việt. (2) Nghiên

cứu và đề xuất phương pháp tính độ tương đồng câu trong báo mạng điện tử. (3) Nghiên cứu và đề xuất hai phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt: Phương pháp dựa trên đồ thị và Phương pháp sử dụng mô hình huấn luyện trước (pre-trained model).

7. Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học: Nghiên cứu chuyên sâu và có hệ thống về văn bản báo mạng điện tử tiếng Việt và bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt. Làm rõ cơ sở toán học của các đặc trưng văn bản báo mạng điện tử tiếng Việt và phương pháp tiếp cận mới, góp phần giải quyết các bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt sau này.

Ý nghĩa thực tiễn: Nghiên cứu xây dựng tập đặc trưng văn bản quan trọng của báo mạng điện tử tiếng Việt và phương pháp tính độ tương tự câu trong văn bản báo mạng điện tử tiếng Việt. Nghiên cứu phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt dựa trên đồ thị và dựa trên mô hình huấn luyện trước và có thể áp dụng xây dựng các phần mềm tóm tắt văn bản thể loại báo mạng điện tử tiếng Việt.

8. Bố cục của luận án

Luận án gồm 04 chương và các phần mở đầu, kết luận, tài liệu tham khảo và danh mục các công trình nghiên cứu đã được công bố của tác giả.

Chương I. Tổng quan về tóm tắt văn bản và tóm tắt văn bản tiếng Việt: Nghiên cứu và trình bày tổng quan về tóm tắt văn bản tự động và các ứng dụng của tóm tắt văn bản; về các phương pháp tóm tắt văn bản tiếng Việt và các kho ngữ liệu phục vụ tóm tắt văn bản tiếng Việt; qua đó chỉ ra những hạn chế về mặt trích chọn đặc trưng của văn bản báo mạng điện tử cũng như việc hạn chế trong các kho ngữ liệu phục vụ bài toán tóm tắt văn bản tiếng Việt.

Chương II. Xây dựng kho ngữ liệu tóm tắt văn bản báo mạng điện tử tiếng Việt: Nghiên cứu và trình bày tổng quan về sự ra đời, phát triển của báo mạng điện tử tiếng Việt, những đặc trưng về cấu trúc và ngôn ngữ của báo mạng điện tử tiếng Việt và xây dựng kho ngữ liệu VNNEWS.100.2018 phục vụ cho bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt.

Chương III. Tóm tắt văn bản báo mạng điện tử dựa trên đồ thị: Nghiên cứu, đề xuất phương pháp tính độ tương đồng câu trong văn bản báo mạng điện tử tiếng Việt dựa trên đánh giá độ quan trọng của Thực thể có tên, Từ khóa và từ gán nhãn (Tags). Đề xuất phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt dựa trên LextRank và LexRank có tính đến vai trò của Thực thể có tên và Từ khóa và từ gán nhãn; thực nghiệm trên bộ dữ liệu VNNEWS.100.2018 để đánh giá kết quả.

Chương IV. Tóm tắt văn bản báo mạng điện tử dựa trên mô hình huấn luyện trước: Nghiên cứu và trình bày về tri thức có sẵn trong văn bản, các tri thức được sử dụng trong các phương pháp học không giám sát (unsupervised learning). Đề xuất phương pháp tóm tắt văn bản trích rút dựa trên pre-trained model có bổ sung tri thức cho trước; thực nghiệm mô hình đề xuất trên các kho ngữ liệu chuẩn của cả hai ngôn ngữ tiếng Anh và tiếng Việt.

CHƯƠNG I. TỔNG QUAN VỀ BÀI TOÁN TÓM TẮT VĂN BẢN VÀ TÓM TẮT VĂN BẢN TIẾNG VIỆT

Chương này trình bày cơ sở lý thuyết về bài toán tóm tắt văn bản, bao gồm các khái niệm cơ bản, các phương pháp tiếp cận, các kho ngữ liệu thường dùng trong thử nghiệm, các phương pháp đánh giá bài toán tóm tắt văn bản. Chương này cũng trình bày các đặc điểm của tiếng Việt và hiện trạng nghiên cứu về tóm tắt văn bản tiếng Việt. Trên cơ sở phân tích hiện trạng, các ưu, nhược điểm của các hướng tiếp cận hiện nay, luận án đề xuất các nội dung cần tập trung nghiên cứu trong luận án.

1.1. Tổng quan

1.1.1. Khái niệm về tóm tắt văn bản:

Tóm tắt văn bản tự động đã được nghiên cứu từ những năm 1950 của thế kỷ 20. 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. 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.

1.1.2. Các giai đoạn và tham số của hệ thống tóm tắt văn bản

Theo Sparck Jones, Hệ thống tóm tắt văn bản tự động (ATS) bao gồm 3 giai đoạn chính sau: *Phân tích (Interpretation); Biến đổi (Transformation); Tổng hợp (Generation)*. Kết quả của tóm tắt văn bản phụ thuộc bởi các tham số đầu vào, tham số mục đích và tham số đầu ra gồm: *Tham số đầu vào (Input factors); Tham số mục đích (Purpose factors)*; *Tham số đầu ra (Output factors)*.

1.1.3. Phân loại các hệ thống tóm tắt văn bản

Có rất nhiều phương pháp tiếp cận về tóm tắt văn bản nên cũng có rất nhiều cách phân loại các hệ thống tóm tắt văn bản, có thể liệt kê một số cách phân loại sau: *Theo kết quả; Theo chức năng của văn bản tóm tắt; Theo nội dung; Theo miền dữ liệu; Theo mức độ chi tiết; Theo số lượng;Theo ngôn ngữ.*

1.1.5. Ứng dụng của hệ thống tóm tắt văn bản

Các thể loại văn bản được nghiên cứu trong lĩnh vực tóm tắt văn bản như: *Tóm tắt văn bản tin tức (News Summarization); Tóm tắt định hướng quan điểm/ tình cảm (Opinion/Sentiment Summarization; Tóm tắt văn bản mạng xã hội (Blog/Tweet, Social networking Summarization); Tóm tắt sách (Books Summarization; Tóm tắt thư điện tử (Email Summarization); Tóm tắt văn bản y sinh (Biomedical Documents Summarization); Tóm tắt văn bản pháp luật (Legal Documents Summarization); Tóm tắt báo khoa học (Scientific Paper Summarization)*.

1.2. Các phương pháp nghiên cứu về tóm tắt văn bản trên thế giới

Thông thường, các phương pháp tóm tắt văn bản được tiếp cận theo 02 hướng: Tóm tắt trích rút, Tóm tắt tóm lược và Tóm tắt lai. Trong mỗi hướng tiếp cận có các phương pháp khác nhau.

1.2.1. Tóm tắt trích rút

Phương pháp trích rút không nhằm viết lại văn bản đầu vào mà sử dụng các phương pháp biểu diễn văn bản sau đó so sánh, xếp hạng và tìm ra các câu quan trọng nhất để sinh bản tóm tắt. Sau khi tiền xử lý văn bản đầu vào, hệ thống sẽ biểu diễn văn bản dưới các dạng thức khác nhau như N-gram, bag-of-word (túi từ), đồ thị… để thuận lợi cho việc xử lý dữ liệu. Việc đánh giá mức độ quan trọng của các câu trong văn bản được sử dụng phù hợp theo từng dạng thức biểu diễn của văn bản đầu vào

1.2.2. Tóm tắt tóm lược

Tóm tắt tóm lược yêu cầu phải phân tích, hiểu sâu về văn bản gốc và viết lại câu, không trích nguyên văn các câu trong văn bản gốc. Bản tóm tắt tóm lược được hình thành trên cơ sở phân tích, hiểu các ý chính của văn bản đầu vào thông qua việc sử dụng các phương pháp xử lý ngôn ngữ tự nhiên, phân tích cú pháp và diễn đạt các nội dung chính của văn bản dưới dạng bản tóm tắt có ít từ hơn với cách diễn đạt rõ ràng.

1.2.3. Tóm tắt lai

Tóm tắt lai là sự kết hợp giữa phương pháp trích rút và tóm lược. Thông thường phương pháp tóm tắt lai gồm 04 giai đoạn: 1) Tiền xử lý văn bản; 2) trích xuất câu quan trọng; 3) sinh bản tóm tắt thông qua các phương pháp tóm lược dựa trên các câu được trích xuất và 4) Xử lý hậu kỳ bằng cách kiểm tra tính đúng đắn của các câu được sinh ra trong quá trình tóm lược.

1.3. Các nghiên cứu về tóm tắt văn bản tiếng Việt

Việc nghiên cứu tóm tắt văn bản tiếng Việt bắt đầu được quan tâm từ những năm đầu thế kỷ 21. Một số sản phẩm nghiên cứu tiêu biểu có thể kể đến như

Tuy nhiên, những nghiên cứu tiêu biểu về tóm tắt văn bản tiếng Việt đã được công bố cho thấy phương pháp tiếp cận chủ yếu theo hướng trích rút câu.

1.4. Công cụ xử lý văn bản tiếng Việt

Đối với lĩnh vực xử lý văn bản tiếng Việt, các công cụ cơ bản tiền xử lý văn bản như tách câu (Sentence Segmentation), tách từ (Word Tokenization), nhận dạng thực thể có tên (Named Entity Recognition), gán nhãn từ loại (Part-Of-Speech Tagging) đã được phát triển với kết quả cho độ chính xác cao. Một số công cụ tiêu biểu cso thể kể đến như sau: *vntokenizer 4.1*, *VnCoreNLP*, *coccoc-tokenizer*, *UETsegmenter*.

1.5. Kho ngữ liệu và phương pháp đánh giá

Kho ngữ liệu phổ biến sử dụng trong tóm tắt văn bản trên thế giới có: DUC (Document Understanding Conference); TAC (Text Analysis Conference); SummBank; CNN-corpus; CNN-DailyMail; BillSum.

Về kho ngữ liệu tiếng Việt, đến thời điểm thực hiện luận án này, tác giả đã tìm hiểu có 04 kho ngữ liệu được công bố rộng rãi sau: *VNDS*; *VietnameseMDS*; *ViMs*; *VSoLSCSum*.

Phương pháp đánh giá: Để đánh giá độ chính xác của bản trích rút tự động, chúng tôi sử dụng phương pháp Precision and recall. và đánh giá dựa trên độ đo ROUGE

Suleiman, A. đã chỉ ra rằng không có bản tóm tắt vàng (Golden Summarization) cho quá trình thử nghiệm và vấn đề chính của bộ dữ liệu tóm tắt văn bản là chất lượng của bản tóm tắt tham chiếu (Tóm tắt vàng). Đối với các kho ngữ liệu tóm tắt văn bản tiếng Việt đã được công bố, VNDS cũng giống như CNN/Daily Mail sử dụng phần nổi bật (hightlight) của văn bản làm bản tóm tắt, bản tóm tắt ở đây là phần sa pô của bài báo, là một thành phần mang nhiều nội dung của báo mạng điện tử không phải là bản tóm tắt.

1.6. Các kiến thức nền tảng

1.6.1. Một số kiến thức nền tảng về tiếng Việt

Tiếng Việt là ngôn ngữ không biến hình từ và âm tiết tính, nghĩa là mỗi một tiếng (âm tiết) được phát âm tách rời nhau và được thể hiện bằng một chữ viết. Hai đặc trưng này chi phối toàn bộ tổ chức bên trong của hệ thống ngôn ngữ Việt. Tiếng Việt có những đặc điểm cơ bản sau cần lưu ý khi nghiên cứu về hệ thống tóm tắt văn bản tiếng Việt: Về cấu tạo, đơn vị cấu tạo từ của tiếng Việt là âm tiết. Về phân loại từ, tiếng Việt có hai loại từ là thực từ và hư từ. Về từ đồng nghĩa, từ đồng nghĩa được hiểu là những từ khác nhau nhưng có

nghĩa giống hoặc gần giống nhau, cùng chỉ một sự vật, một đặc tính hay một hành động nào đó. Về chính tả, trong tiếng Việt cũng có đặc điểm về chính tả cần lưu ý so với tiếng Anh như các từ đồng âm (lý/lí, kỹ/kĩ...), vị trí dấu thanh (tỏa/ toả, thúy/thuý...).

1.6.2. Độ tương tự câu trong văn bản

Đối với văn bản d gồm có n câu: $d = \{s_1, s_2, \dots, s_n\}$. Hàm mục tiêu của bài toán độ tương tự là S (s_i, s_j) trong đó S$\in$ (0,1), và i, j = 1, ..., n. Giá trị hàm S càng cao thì sự giống nhau về nghĩa của s_i, s_j càng nhiều.

1.6.3. Biểu diễn văn bản dưới dạng đồ thị

Trong biểu diễn đồ thị, các thành phần văn bản (từ hoặc câu) được biểu diễn bằng các đỉnh và các cạnh biểu diễn sự kết nối giữa các thành phần của văn bản có liên quan với nhau. Thông thường có hai phương thức biểu diễn văn bản dưới dạng đồ thị: đồ thị từ vựng (lexical graph) và đồ thị ngữ nghĩa (semantic graph).

1.6.4. Mô hình huấn luyện trước (Pre-trained Model)

Mô hình huấn luyện trước (pre-trained model) là một loại mô hình học sâu – một thể hiện của thuật toán thần kinh giống như bộ não người giúp tìm các hình mẫu hoặc đưa ra dự đoán dựa trên một tập dữ liệu lớn và đa dạng trước khi được tinh chỉnh hoặc sử dụng cho một nhiệm vụ cụ thể. Quá trình tiền huấn luyện giúp mô hình học được các biểu diễn tổng quát về ngôn ngữ, thông tin, hoặc cấu trúc dữ liệu.

1.6.5. Kỹ thuật nhúng từ (Word Embedding)

Word embedding là kỹ thuật biểu diễn từ vựng để làm đầu vào cho các mô hình học máy. Theo đó, đối với kỹ thuật Word Embedding các từ vựng (text) trong văn bản sẽ được ánh xạ sang dạng thức của vector số trong một không gian nhiều chiều nhằm xử lý dữ liệu một cách hiệu quả hơn.

1.6.6. Mô hình Transformer

Transformer được giới thiệu trong bài báo nổi tiếng "Attention is All You Need" của Vaswani và cộng sự, được trình bày tại hội nghị NeurIPS 2017. Mô hình Transformer có một kiến trúc mới sử dụng cơ chế chú ý (attention mechanism) để hiệu quả xử lý các chuỗi đầu vào và đầu ra có độ dài thay đổi, đã đạt được những thành tựu lớn trong nhiều ứng dụng trong các mô hình học máy cho dữ liệu chuỗi như dịch máy, tổng hợp tiếng nói và xử lý ngôn ngữ tự nhiên như BERT, GPT...1.6.7. Mô hình BERT và PhoBERT.

1.7. Những vấn đề luận án cần tập trung giải quyết

Trên cơ sở nhận định và phân tích các kết quả đã đạt được cũng như những hạn chế trong các công trình công bố của các tác giả đi trước, luận án đề xuất mô hình hệ thống tóm tắt văn bản báo mạng điện tử tiếng Việt dựa trên 02 phương pháp tiếp cận như sau: *Một là*, phương pháp tiếp cận dựa trên đồ thị. *Hai là*, phương pháp tiếp cận dựa trên mô hình huấn luyện trước BERT.

Theo 02 phương pháp tiếp cận trên, luận án xác định các nội dung nghiên cứu chính là: (1) Nghiên cứu các đặc trưng quan trọng của văn bản báo mạng điện tử tiếng Việt, qua đó đề xuất lựa chọn tập đặc trưng để đưa vào mô hình đồ thị. (2) Đề xuất phương pháp tính độ tương tự câu trong văn bản báo mạng điện tử tiếng Việt dựa trên các đặc trưng quan trọng. (3) Nghiên cứu phương pháp tính tri thức có sẵn trong văn bản để tinh chỉnh và đề xuất phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt sử dụng mô hình

1.8. Kết luận Chương I

Chương I đã trình bày về bài toán tóm tắt văn bản và các cách tiếp cận để phân loại, ứng dụng của tóm tắt văn bản tự động. Chương này đã nghiên cứu các phương pháp tiếp cận để giải quyết bài toán tóm tắt văn bản tự động trên thế giới và ứng dụng trong Tiếng Việt, đã nghiên cứu các kiến thức cơ bản sử dụng trong tóm

văn bản tự động. Chương này cũng đã đánh giá một số vấn đề còn hạn chế trong tóm tắt tự động văn bản tiếng Việt làm cơ sở để đề xuất 02 phương pháp tiếp cận cho bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt. Chương tiếp theo sẽ giới thiệu phương pháp về xây dựng kho ngữ liệu phục vụ bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt.

CHƯƠNG II. XÂY DỰNG KHO NGỮ LIỆU TÓM TẮT VĂN BẢN BÁO MẠNG ĐIỆN TỬ TIẾNG VIỆT

Chương này trình bày những nội dung cơ bản báo mạng điện tử tiếng Việt bao gồm sự hình thành, phát triển, đặc trưng về cấu trúc và ngôn ngữ của báo mạng điện tử tiếng Việt. Nội dung chính sẽ trình bày về đặc điểm của các cấu phần như tít, sa pô, từ gán nhãn, thực thể có tên, trên cơ sở đó đề xuất phương pháp xây dựng kho ngữ liệu phục vụ đánh giá bài toán tóm tắt báo mạng điện tử tiếng Việt.

2.1. Đặt vấn đề

Như đã trình bày tại Mục 1.5.4 Chương I, hiện nay các kho ngữ liệu phục vụ đánh giá tóm tắt văn bản tiếng Việt chưa được công bố nhiều. Đối với các kho ngữ liệu đã công bố, mỗi văn bản chỉ bao gồm văn bản gốc và bản tóm tắt tham chiếu, không có các đặc trưng khác. Đối với thể loại văn bản báo mạng điện tử là thể loại văn bản đã được phát triển đồng bộ, định hình thống nhất qua nhiều giai đoạn, có cấu trúc thông tin, đặc điểm ngôn ngữ đặc trưng riêng thì hiện nay chưa có kho ngữ liệu nào đáp ứng đầy đủ các cấu trúc đó. Do vậy, để phục vụ bài toán tóm tắt văn bản báo mạng điện tử cần thiết phải nghiên cứu về các đặc trưng về cấu trúc và ngôn ngữ của văn bản báo mạng điện tử tiếng Việt để từ đó xây dựng kho ngữ liệu đánh giá thử nghiệm riêng bao gồm tối đa nhất các đặc trưng có trong văn bản báo mạng điện tử.

2.2. Khái niệm và sự hình thành báo mạng điện tử

Báo mạng điện tử là một loại hình báo chí được xây dựng dưới hình thức của một trang web, phát hành trên mạng Internet, có ưu thế trong chuyển tải thông tin một cách nhanh chóng, tức thời, đa phương tiện và tương tác cao. Quá trình hình thành và phát triển của báo mạng điện tử Việt Nam thành 03 giai đoạn:

- Giai đoạn từ năm 1997 đến năm 2001: giai đoạn đánh dấu sự ra đời của báo mạng điện tử Việt Nam.

- Giai đoạn từ năm 2001 đến năm 2005: giai đoạn phát triển vượt bậc của các trang thông tin điện tử của các cơ quan báo chí lớn.

- Giai đoạn từ năm 2005 đến nay: giai đoạn này đánh dấu sự phát triển, trưởng thành của báo mạng điện tử Việt Nam.

2.3. Đặc trưng ngôn ngữ của báo mạng điện tử

Đặc điểm về cấu trúc, thông thường, cấu trúc thông tin của một bài báo trong báo mạng điện tử được tổ chức theo nhiều cửa, mỗi yếu tố dưới đây được gọi là một cửa gồm: Tít chính, Sa pô, Chính văn, Tít phụ, Tranh, ảnh, Đồ hình (sơ đồ, bản đồ, biểu đồ…), Video và hình ảnh động, Audio, Các box thông tin, tư liệu (hộp dữ liệu), Các đường link, Các từ khóa và từ gán nhãn (Tags).

Đặc điểm về ngôn ngữ, báo mạng điện tử có các đặc điểm ngôn ngữ là có khả năng tích hợp nhiều loại hình ngôn ngữ, có kết cấu mở, cô đọng ngắn gọn, ngôn ngữ thông báo chiếm vai trò chủ yếu, ngôn ngữ mang tính thời sự nóng hổi; tít và sa pô có tính độc lập cao và có vai trò ngôn ngữ, thông tin lớn.

2.3.1. Tít trong báo mạng điện tử

Tít báo hay còn được gọi là tiêu đề, đầu đề, nhan đề… của bài báo. Tít là thuật ngữ mượn từ tiếng Anh (title) và tiếng Pháp (titre). Mặc dù không phải là từ gốc tiếng Việt nhưng tít đã trở thành khái niệm rất quen thuộc trong đời sống báo chí, trở thành một thuật ngữ chuyên ngành. Tít là nội dung cô đọng nhất định danh thông tin, vì vậy các đối tượng (thực thể có tên) được đề cập đến trong tít sẽ là các thành phần chứa thông tin

Sa pô là phần gạch ngang ở trên có nhiệm vụ tóm tắt hoặc cho biết thông tin quan trọng, cần thiết, hấp dẫn của một sự kiện hoặc một vấn đề. Sapô là một thành phần của bài báo, có chứa nhiều thông tin quan trọng, được tác giải viết với mục đích, ý nghĩa thu hút người đọc; không phải là bản tóm tắt của văn bản báo mạng điện tử.

Từ khóa và từ gán nhãn (Tags): Mỗi tờ báo điện tử hướng theo lĩnh vực riêng, người dùng riêng, tương đương với bộ từ khoá riêng cho từng lĩnh vực. Thông thường, mỗi bài báo mạng điện tử sử dụng tối đa 5 tags, tối thiểu 3 tags. Từ khóa và từ gán nhãn có vai trò ngữ nghĩa rất quan trọng trong bài báo mạng điện tử.

Thực thể có tên (Named Entity): Ở đây, chúng tôi kế thừa các quan điểm nghiên cứu của các tác giả đi trước với kết luận các thực thể có tên được xem là quan trọng khi xuất hiện từ 2 lần trở lên trong nội dung bài báo. Đồng thời, bổ sung thêm đặc trưng qua ngôn ngữ báo chí như đã trình bày ở trên 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.

2.4. Xây dựng kho ngữ liệu

2.4.1. Phương pháp xây dựng kho ngữ liệu

Phương pháp thu thập dữ liệu: Để xây dựng kho ngữ liệu trong bài báo này chúng tôi 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 (nay là https://znews.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.

Phương pháp xây dựng bản tóm tắt: Đối với mỗi văn bản chúng tôi cũng xây dựng 01 bản trích rút giữ lại khoảng 30% số câu trong văn bản tương ứng là S30 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.

Tiền xử lý dữ liệu: Sau khi thu thập, các văn bản sẽ được chuyển qua bộ tiền xử lý từ để phân tách thành câu, từ và loại bỏ các từ dừng. Để phân tách câu và từ, chúng tôi sử dụng công cụ VnCoreNLP được phát triển và xây dựng bởi Thành Vũ và cộng sự.

2.4.2. Đặc tả kho ngữ liệu VNNEWS.100.2018

Cấu trúc kho ngữ liệu VNNEWS.100.2018 bao gồm 100 thư mục đánh số thứ tự từ 1 đến 100 tương ứng với 100 bài báo được thu thập gồm 2240 câu. Tập S30 gồm 100 văn bản rút gọn, có tổng số 624 câu. Số lượng bài trên các trang báo điện tử được phân bố như sau: http://dangcongsan.vn: 15 văn bản. https://znews.vn/: 30 văn bản. https://vnexpress.net: 65 văn bản. 12 chủ đề gồm (Xây dựng Đảng: 02 văn bản; Thời sự: 10 văn bản; Người Việt Nam ở nước ngoài: 06 văn bản; Xã hội: 10 văn bản; Công nghệ: 07 văn bản; Đối ngoại: 01 văn bản; Thế giới: 08 văn bản; Giáo dục: 20 văn bản; Pháp Luật: 08 văn bản; Đời sống - Văn hóa: 10 văn bản; Kinh doanh: 09 văn bản; Khoa học: 10 văn bản).

2.5. Kết luận Chương II

Chương II đã nghiên cứu và trình bày tổng quan về sự ra đời, phát triển của báo mạng điện tử tiếng Việt, những đặc trưng về mặt ngôn ngữ của báo mạng điện tử tiếng Việt. Chương này đã nghiên cứu và trình bày tổng quan về vai trò của cách thành phần tiêu đề, sa pô, thực thể có tên, từ khóa và từ gán nhãn trong văn bản báo mạng điện tử tiếng Việt. Theo đó, chương này đã nghiên cứu và xây dựng kho ngữ liệu VNNEWS.100.2018 bao gồm đầy đủ các đặc trưng nêu trên để phục vụ cho bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt.

CHƯƠNG III. PHƯƠNG PHÁP TÓM TẮT VĂN BẢN BÁO MẠNG ĐIỆN TỬ DỰA TRÊN MÔ HÌNH ĐỒ THỊ

Chương này trình bày nội dung đề xuất tóm tắt văn bản theo phương pháp tiếp cận dựa trên đồ thị, một trong những phương pháp tiếp cận phổ biết nhất của tóm tắt trích rút. Nội dung chính sẽ trình bày về phương pháp biểu diễn văn bản dưới dạng đồ thị, phương pháp tính độ tương đồng câu trong văn bản báo mạng điện tử và đề xuất phương pháp tóm tắt văn bản báo mạng điện tử dựa trên hai thuật toán TextRank và LexRank. Phương pháp đề xuất được cài đặt thử nghiệm trên tập ngữ liệu VNNEWS.100.2018 và so sánh kết quả với thuật toán gốc với các thuật toán cơ bản.

3.1. Đặt vấn đề

Để giải quyết các bài toán tóm tắt văn bản, chúng ta thường sử dụng các mô hình biểu diễn văn bản như mô hình túi từ (bag-of-words), không gian véc tơ, mô hình đồ thị …. Các mô hình có cấu trúc dữ liệu trực quan đóng vai trò trung gian để hệ thống tóm tắt văn bản tự động hiểu được ngôn ngữ tự hiện dưới dạng văn bản. Trong đó, mô hình đồ thị có ưu điểm là không cần các kỹ thuật xử lý ngôn ngữ tự nhiên đặc thù cho từng loại ngôn ngữ nhưng vẫn thể hiện được các thông tin đặc trưng cấu trúc quan trọng của văn bản như trật tự các từ, vị trí các từ trong câu, mối quan hệ ngữ nghĩa giữa các câu trong văn bản…, do vậy mô hình này đã được sử dụng hiệu quả trong tóm tắt trích rút câu.

3.2. Phát biểu bài toán

Biểu diễn văn bản dưới dạng đồ thị: Văn bản được biểu diễn bằng đồ thị, theo đó, mỗi đỉnh trong đồ thị tương ứng với một câu trong văn bản, mỗi cạnh nối hai đỉnh trong đồ thị biểu diễn mối liên hệ giữa hai câu. Trọng số của mỗi cạnh chính là giá trị độ tương đồng (value of similarity) giữa hai câu. Với văn bản $d = \{s_1, s_2, \dots, s_n\}$, được biểu diễn dưới dạng đồ thị vô hướng:

$$d = (\mathrm{V}, \mathrm{E})$$

Trong đó:

V: là tập các đỉnh

E: là tập các cạnh, *E* là tập con của *V* x *V.*

In(V_i): là tập các đỉnh trỏ đến V_i

Out(V_i): là tập các đỉnh mà V_i trỏ đến

w_{ij} là trọng số của cạnh nối hai đỉnh *i, j*.

Hệ thống tóm tắt văn bản theo phương pháp đồ thị gồm 03 bước cơ bản: Tiền xử lý văn bản; Biểu diễn văn bản dưới dạng đồ thị và xử lý văn bản; Chọn câu và sinh bản tóm tắt.

3.3. Đề xuất ý tưởng

Có thể thấy đối với phương pháp tiếp cận tóm tắt văn bản dưới dạng đồ thị có 02 yếu tố ảnh hưởng quyết định đến chất lượng trích câu trong văn bản là: (1) Phương pháp tính độ tương đồng câu trong văn bản. (2) Giải thuật xếp hạng câu. Chính vì vậy, luận án sẽ khai thác ưu điểm của thể loại văn bản báo mạng điện tử để tính độ tương đồng câu trong văn bản và thực nghiệm đánh giá hiệu quả của phương pháp này dựa trên 02 thuật toán xếp hạng được sử dụng rộng rãi là thuật toán TextRank được giới thiệu bởi Mihalcea, R., và Tarau, P. và thuật toán LexRank được giới thiệu bởi Erkan, G., và Radev, D.

3.4. Tính độ tương đồng câu trong văn bản báo mạng điện tử

3.4.1. Độ tương đồng ngữ nghĩa

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:

$$Sim_r(S_1, S_2) = \frac{|\{w_i | w_i \in S_1 \& w_i \in S_2\}| + 2 \times |\{w_i | w_i \in Tg\}| + |\{w_i | w_i \in Tt\}|}{|S_1| + |S_2|}$$

3.4.2. Độ tương đồng về thứ tự từ

Li và cộng sự đã đề xuất phương pháp tính toán độ giống nhau của hai câu dựa trên thứ tự của các từ. Để đơn giản, trong nghiên cứu này chúng tôi bỏ qua bước tìm từ tương tự. Thuật toán tính độ tương tự được mô tả như sau:

- $T(S_1 \cup S_2) = (w_1, w_2 \ldots w_m)$ là tập các từ được trích ra.

- $R_1 = (r_{11}, r_{12} \ldots r_{1m})$ là vector thứ tự từ trong câu S_1.

- $R_2 = (r_{21}, r_{22} \ldots r_{2m})$ là vector thứ tự từ trong câu S_2.

Với mỗi từ w_i trong T nếu w_i có trong S_1 thì r_{1i} nhận giá trị là thứ tự của w_i trong S_1, ngược lại $r_{1i} = 0$.

$$Sim_o(S_1, S_2) = 1 - \frac{|S_1 - S_2|}{|S_1 + S_2|} = 1 - \frac{\sqrt{\sum_{i=1}^{m}(r_1 i - r_2 i)^2}}{\sqrt{\sum_{i=1}^{m}(r_1 i + r_2 i)^2}}$$

3.4.3. Đề xuất phương pháp tính độ tương đồng câu

Để tính độ giống nhau của hai câu, như trong [54] chúng tôi kết hợp hai thước đo trên. Biểu thức kết hợp giữa hai thước đo như sau:

$$Sim_{ro}(S_1, S_2) = a * Sim_r(S_1, S_2) + b * Sim_o(S_1, S_2) \text{ with } a + b = 1$$

Trên thực tế chưa có công thức để xác định trọng số *a, b*. Trong trường hợp này để đảm bảo sự cân bằng của các đặc trưng chúng tôi lựa chọn $a = 0.5$ và $b = 0.5$.

3.5. Tóm tắt văn bản báo mạng điện tử dựa trên trên mô hình đồ thị

3.5.1. Mô hình đề xuất đối với thuật toán TextRank

TextRank là một kỹ thuật học không giám sát (Unsupervised Learning) sử dụng tóm tắt văn bản theo phương pháp trích rút Rada Mihalcea và Paul Tarau [66]. TextRank được phát triển từ giải thuật PageRank được giới thiệu bởi các nhà nghiên cứu của Google là Sergey Brin, Lawrence PageBrin, S. and Page, L. (1998) và Sergey Brin, Lawrence Page (2012) trong [99, 100]. TextRank không dựa trên bất kỳ dữ liệu đào tạo nào trước đó và có thể hoạt động với bất kỳ đoạn văn bản tùy ý nào.

Ở đây, văn bản sau khi được tiền xử lý sẽ chuyển sang bước tiếp theo để tính độ tương đồng câu luận án sử dụng 02 phương pháp:

(1) Độ tương đồng dựa trên của thuật toán TextRank cơ bản trong thuật toán gốc được xác định như sau:

Đối với văn bản *d*, thuật toán TextRank cơ bản xác định độ tương tự của S_i và S_j như sau:

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

(2) Độ tương đồng câu của văn bản báo mạng điện tử đề xuất tại ***mục 3.4.3*** với hàm sentence-similarity-$Sim_{so}(S_1, S_2, T_g, T_s)$.

Sau đó, thuật toán PageRank sẽ được áp dụng trên đồ thị đề tính trọng số cho các câu.

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 [11], 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. Sau khi kết thúc thuật toán, sẽ lựa chọn 30% số câu có trọng số cao nhất và sắp xếp lại thứ tự của các câu này theo thứ tự trong văn bản gốc để sinh bản tóm tắt.

3.5.2. Mô hình đề xuất đối với thuật toán LexRank

LexRank cũng là một cách tiếp cận dựa trên biểu đồ không giám sát để tóm tắt đa văn bản tự động [33]. LexRank được sử dụng để tính toán tầm quan trọng của câu dựa trên khái niệm về tính trung tâm của vector riêng trong biểu diễn biểu đồ của câu. Theo đó, văn bản sau khi được tiền xử lý sẽ chuyển sang bước tiếp theo để tính độ tương đồng câu. Ở đây, độ tương đồng câu được tính theo 02 phương pháp:

(1) Độ tương đồng của thuật toán LexRank cơ bản xác định bởi cosin giữa hai vector tương ứng:

$$idf - modified - cosine(x, y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum x_i \in x (tf_{x1,x} idf_{x1})^2} \times \sqrt{\sum y_i \in y (tf_{y1,y} idf_{y1})^2}}$$

trong đó $tf_{w,s}$ là số lần xuất hiện của từ w trong câu s.

(2) Độ tương đồng câu của văn bản báo mạng điện tử đề xuất tại ***mục 3.4.3*** với hàm sentence-similarity-$Sim_{so}(S_1, S_2, Tg, Ts)$.

Sau khi tính độ tương đồng câu, ma trận kề được tạo bằng cách đặt ngưỡng t = 0,1, mọi điểm độ tương đồng câu có giá trị dưới 0,1 không được đưa vào ma trận kề. Sau khi khởi tạo đồ thị, việc sắp xếp câu được thực hiện, mỗi giá trị của ma trận độ tương đồng được chia cho bậc của mỗi nút. Mức độ trung tâm ở đây là mức độ tương ứng của mỗi nút. Cuối cùng thông qua phương pháp Power iteration chúng ta tính được kết quả cuối cùng. Sau khi kết thúc thuật toán, sẽ lựa chọn 30% số câu có trọng số cao nhất và sắp xếp lại thứ tự của các câu này theo thứ tự trong văn bản gốc để sinh bản tóm tắt.

3.5.3. Đánh giá thử nghiệm

3.5.3.1. Môi trường thực nghiệm

Để tiến hành thực nghiệm, chúng tôi sử dụng ngôn ngữ lập trình Java để cài đặt các thuật toán tính độ tương tự câu, thuật toán TextRank, LexRank.

3.5.3.2. Kho ngữ liệu thực nghiệm

Để đánh giá hiệu quả của giải pháp đề xuất, luận án thử nghiệm 02 thuật toán TextRank, LexRank trên kho ngữ liệu VNNEWS.100.2018 đã xây dựng ở Chương II.

3.5.3.3. Kết quả thực nghiệm và so sánh

a. So sánh giữa phương pháp cơ sở TextRank và phương pháp đề xuất của luận án: Kết quả thực nghiệm được thể hiện trong **Bảng 1** qua độ đo F_1-score với kết quả cao hơn đạt 64,2% với thuật toán TextRank sử dụng phương pháp tính độ tương đồng câu được đề xuất cho văn bản báo mạng điện tử tại ***mục 3.4.3*** với hàm sentence-similarity-$Sim_{ro}(S_1, S_2, Tg, Ts)$.

Bảng 1. Kết quả thực nghiệm TextRank

Độ tương đồng câu	Kho ngữ liệu	Precision	Recall	F1-score
TextRank based Sim	VNNEWS.100.2018	64,0	60,1	62,0
sentence-similarity-Sim$_{ro}$	VNNEWS.100.2018	**66,3**	**62,2**	**64,2**

b. So sánh giữa phương pháp cơ sở LexRank và phương pháp đề xuất của luận án: Kết quả thực nghiệm được thể hiện trong **Bảng21** qua độ đo F1-score với kết quả cao hơn đạt 64,4% với thuật toán LexRank sử dụng phương pháp tính độ tương đồng câu được đề xuất cho văn bản báo mạng điện tử tại mục 3.4.3 với hàm sentence-similarity-Simso(S1, S2, Tg, Ts).

Bảng 2. Kết quả thực nghiệm LexRank

Độ tương đồng câu	Kho ngữ liệu	Precision	Recall	F_1-score
LexRank based Sim	VNNEWS.100.2018	65,0	61,0	62,9
sentence-similarity-Sim$_{ro}$	VNNEWS.100.2018	**66,4**	**62,5**	**64,4**

c. So sánh giữa TextRank và LexRank: Kết quả thực nghiệm được thể hiện trong **Bảng 3** qua độ đo F1-score với kết quả cao nhất đạt 64,4% với thuật toán LexRank sử dụng phương pháp tính độ tương đồng câu được đề xuất cho văn bản báo mạng điện tử tại mục 3.4.3 với hàm sentence-similarity-Simso(S1, S2, Tg, Ts).

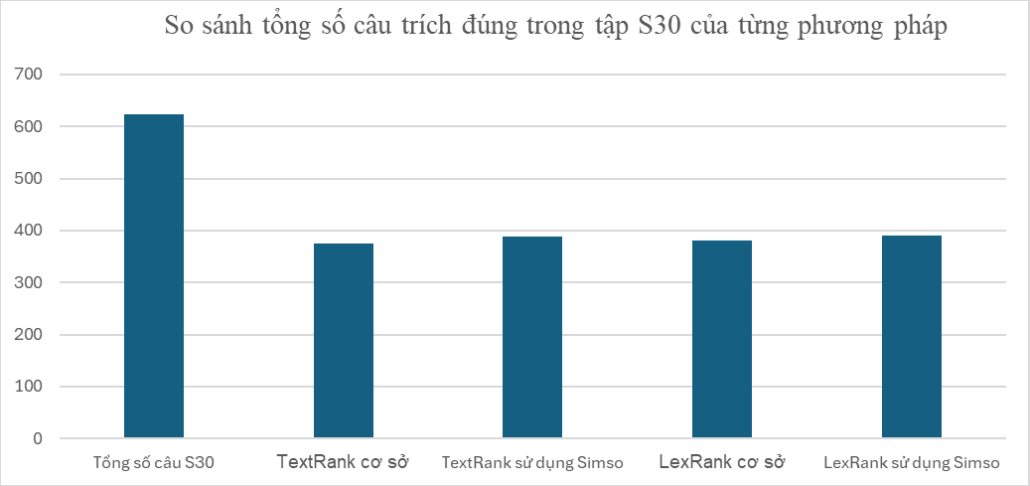
Bảng 3. Kết quả thực nghiệm trên kho ngữ liệu VNNEWS.100.2018

Thuật toán	Độ tương đồng câu	Kho ngữ liệu	Precision	Recall	F1-score
TextRank	*TextRank based Sim*	VNNEWS.100.2018	64,0	60,1	62,0
	sentence-similarity-Sim$_{ro}$	VNNEWS.100.2018	**66,3**	**62,2**	**64,2**
LexRank	*LexRank based Sim*	VNNEWS.100.2018	65,0	61,0	62,9
	sentence-similarity-Sim$_{ro}$	VNNEWS.100.2018	**66,4**	**62,5**	**64,4**

Từ **Bảng 3.** 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.

- Kết quả tóm tắt theo phương pháp đề xuất của luận án trên thuật toán LexRank cho kết quả tốt nhất.



Hình 2. So sánh tổng số câu trích đúng của từng phương pháp

Hình 2 thể hiện sự so sánh tuyệt đối số câu trích đúng trong tập S30 đối với từng phương pháp. Trong tập S30 các chuyên gia lựa chọn tổng số 624 câu của 100 văn bản. Kết quả số câu trích đúng với mỗi phương pháp như sau: Thuật toán TextRank cơ sở là 375 câu. Thuật toán TextRank sử phương pháp tính độ tương đồng theo đề xuất của luận án là 388 câu. Thuật toán LexRank cơ sở là 380 câu. Thuật toán LexRank sử phương pháp tính độ tương đồng theo đề xuất của luận án là 390 câu.

Kết quả cho thấy không có nhiều sự chênh lệnh giữa kết quả tốt nhất (390 câu) và kết quả hạn chế nhất (375 câu) là 2% tổng số câu của tập S03 cho thấy sự ổn định của phương pháp đồ thị trong tóm tắt văn bản.

3.6. Kết luận Chương III

Chương III đã nghiên cứu, phương pháp tính độ tương đồng câu trong văn bản báo mạng điện tử tiếng Việt dựa trên đánh giá độ quan trọng của Thực thể có tên, Từ khóa và từ gán nhãn (Tags). Chương này đã nghiên cứu phương pháp biểu diễn văn bản dưới dạng đồ thị và trình bày các thuật toán tóm tắt văn bản TextRank và LexRank qua đó đề xuất phương pháp tóm tắt văn bản báo mạng điện tư tiếng Việt dựa trên LextRank và LexRank dựa trên đánh giá độ quan trọng của Thực thể có tên, Từ khóa và từ gán nhãn (Tags). Chương này cũng đã thực nghiệm thuật toán TexRank, LexRank theo phương pháp đề xuất trên bộ dữ liệu VNNEWS.100.2018 để đánh giá kết quả.

CHƯƠNG IV. TÓM TẮT VĂN BẢN BÁO MẠNG ĐIỆN TỬ DỰA TRÊN MÔ HÌNH HUẤN LUYỆN TRƯỚC BERT

Chương này trình bày nội dung đề xuất theo hướng sử dụng mô hình huấn luyện trước (BERT) để giải quyết bài toán tóm tắt văn bản báo mạng điện tử. Nội dung chính sẽ trình bày vắn tắt về mô hình huấn luyện trước BERT và tri thức có sẵn (prior-knowledge) trong văn bản; đề xuất ý tưởng tinh chỉnh (fine-turning) mô hình BERT bằng việc bổ sung (injection) tri thức có sẵn để giải quyết bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt. Phương pháp đề xuất được thực nghiệm trên 05 kho ngữ liệu (03 kho ngữ liệu tiếng Anh và 02 kho ngữ liệu tiếng Việt) để so sánh và đánh giá kết quả.

4.1. Đặt vấn đề

Như đã trình bày tại 1.6.7 Chương I, mô hình BERT đã đạt được những đột phá quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên, trong đó có bài toán tóm tắt văn bản tự động. Ưu điểm của BERT là ngoài việc đã được huấn luyện trước trên các tập dữ liệu lớn, còn có thể tinh chỉnh (fine-turning) mô hình cho các tác vụ cụ thể dựa trên các ngữ cảnh, dữ liệu đầu vào của bài toán thực tế. Kết quả nghiên cứu các phương pháp tóm tắt văn bản truyền thống trước đây, nhất là đối với các phương pháp tiếp cận học không giám sát như mô hình đồ thị đã trình bày ở Chương II, cho thấy, bản thân nội tại trong mỗi văn bản đều chứa đựng các tri thức riêng, có thể nghiên cứu, sử dụng để tinh chỉnh mô hình BERT nhằm nâng cao hiệu quả cho bài toán tóm tắt văn bản. Vì vậy, trong chương này, luận án đề xuất kỹ thuật tóm tắt văn bản dựa trên mô hình BERT, được tinh chỉnh sử dụng tri thức sẵn có trong văn bản.

4.2. Phát biểu bài toán

4.2.1. Tri thức sẵn có (Prior knowledge)

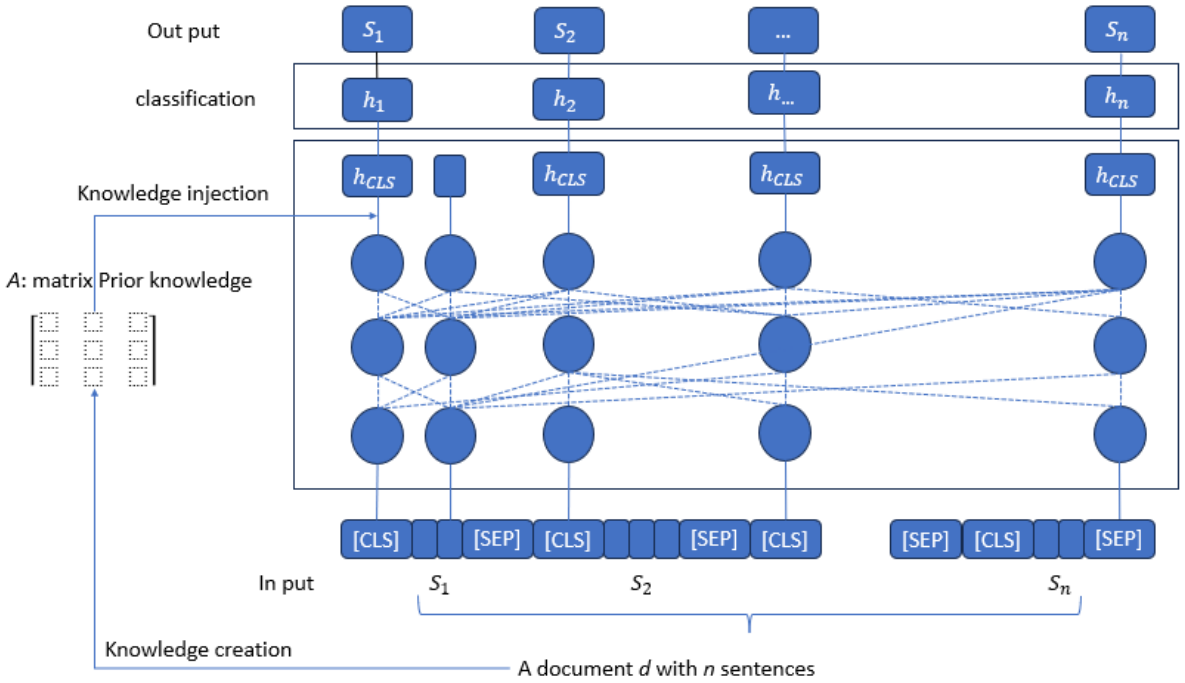
Với văn bản $d = \{s_1, s_2, \dots, s_n\}$ có *n* câu, tri thức sẵn có (tiền tri thức) của văn bản được hiểu là mức độ quan trọng của mỗi câu trong văn bản *d*. Tri thức sẵn có được định nghĩa là một ma trận $\boldsymbol{A} = [n \times n]$ được khởi tạo từ văn bản *d* qua một số phép toán tính toán độ tương đồng câu. Trong đó, mỗi giá trị $\boldsymbol{A}[i, j]$ biểu thị mối tương quan về ngữ nghĩa giữa câu s_i và câu s_j.

4.2.2. Phát biểu bài toán

Với văn bản $d = \{s_1, s_2, \dots, s_n\}$ có *n* câu, nhiệm vụ của bài toán tóm tắt văn bản theo hướng trích rút sử dụng mô hình huấn luyện trước có bổ sung tri thức sẵn có là phân loại câu trong văn bản để đánh giá đó là câu quan trọng hay không quan trọng. Giả sử $\mathfrak{D}$ là tập văn bản huấn luyện, đối với câu $s_i \in d$ xác suất của câu s_i được đưa vào bản tóm tắt là một xác xuất có điều kiện $\hat{y}_i$ được tính theo hàm $p(\hat{y}_i|\theta, A, \mathfrak{D})$. Bản tóm tắt cuối cùng sẽ bao gồm các câu s_i được dự đoán là quan trọng. Siêu tham số θ có thể được học từ tập văn bản huấn luyện $\mathfrak{D}$ với sự bổ sung hoặc không của tri thức sẵn có $\boldsymbol{A}$.

4.3. Đề xuất ý tưởng

Có sự tương quan giữa các câu trong văn bản. Sự tương quan cho phép chúng ta ước tính tầm quan trọng của một câu so với những câu khác. Dựa trên mối tương quan, có một số phương pháp học không giám sát để tóm tắt trích xuất như phương pháp đồ thị đã nghiên cứu ở Chương III. Chúng tôi lập luận rằng mối tương quan giữa các câu có thể là một chỉ báo hữu ích có thể hướng dẫn quá trình học tập trung hơn vào các câu quan trọng. Để mã hoá mối tương quan với quá trình huấn luyện, chúng tôi giới thiệu một mô hình trích rút văn bản mới. Mô hình sử dụng BERT làm cơ sở và để tinh chỉnh BERT, các chỉ báo tương quan này sẽ được đưa vào BERT thông qua các hàm Attention.



Hình 3. Mô hình BERT tóm tắt văn bản sử dụng tri thức sẵn có

Hình 3. mô tả mô hình được đề xuất để tóm tắt trích xuất bằng cách sử dụng tri thức sẵn có (tiền tri thức). Cho một tài liệu d có n câu, trước tiên mô hình tạo ra tri thức về d dưới dạng ma trận A. Mô hình sau đó bổ sung thêm Mã thông báo [CLS] và [SEP] để nối n câu để tạo thành một chuỗi đầu vào mới. Trình tự mới được đưa vào kiến trúc Tranformer để thu được các véc tơ ngữ cảnh của từng mã thông báo. Tri thức từ ma trận A được đưa vào các lớp chú ý của BERT trong quá trình tinh chỉnh có điều kiện sự. Cuối cùng, mô hình sử dụng véc tơ từ Mã thông báo [CLS] để phân loại ước tính tầm quan trọng của mỗi câu. Các câu có độ tin cậy cao về tầm quan trọng được chọn lọc để tạo thành câu cuối cùng bản tóm tắt.

Mô hình chia sẻ ý tưởng tinh chỉnh BERT để tóm tắt [91]. Tuy nhiên, thay vì chỉ sử dụng BERT, chúng tôi giới thiệu kiến thức sẵn có mã hóa mối tương quan giữa các câu và đưa kiến thức vào BERT. Việc ràng buộc tạo ra sự chú ý của BERT tập trung nhiều hơn vào một số những câu quan trọng giúp nâng cao chất lượng tóm tắt.

4.4. Mô hình bài toán tóm tắt văn bản sử dụng tri thức sẵn có

4.4.1. Quá trình tạo tri thức

Như đã đề cập, tồn tại mối tương quan giữa các câu trong một tài liệu. Các mối tương quan có lợi cho việc ước tính tầm quan trọng của câu. Phần này trình bày việc tạo ra kiến thức có sẵn trong ba thuật toán độ tương tự: Độ tương đồng Cosine, Độ tương đồng LexRank và Hệ số hóa ma trận (Matrix factorization).

a. Độ tương đồng Cosine: Cho văn bản $d = \{s_1, s_2, \dots, s_n\}$ có n câu, một ma trận $\boldsymbol{A} = [n \times n]$ là ma trận độ tương đồng của n câu trong văn bản d, theo đó mỗi giá trị $A[i,j]$ biểu thị độ tương đồng của câu s_i và câu s_j. Phép ánh xạ sử dụng SentenceBERT [90] để biểu diễn câu thành các véc tơ tương ứng. Sau đó, áp dụng công thực tính độ tương đồng sau:

$$Cos(X,Y) = \frac{\sum_i x_i \times y_i}{\sqrt{\sum_i (x_i)^2} \times \sqrt{\sum_i (y_i)^2}}$$

Trong đó, x_i và y_i là hai véc tơ của hai câu s_i và s_j có độ dài n.

b. Độ tương đồng LexRank: Độ tương đồng LexRank đã được mô tả tại mục 3.5.2 của Chương III. Sau khi tính toán ma trận độ tương đồng, một ma trận kề được tạo ra từ ma trận độ tương đồng bằng việc so sánh độ tương tự với ngưỡng 0.1 theo công thức sau:

$$A[i,j] = \begin{cases} 1 & if\ cos(A[i,j]) \geq 0.1 \\ 0 & otherwise \end{cases}$$

Sau khi tính ma trận **A**, LexRank tiếp tục tính ma trận kết nối **C**. Ma trận kết nối phản ánh xác suất chuyển đổi giữa các câu dựa trên điểm số theo cặp của chúng. Ma trận kết nối được tạo bằng cách chuẩn hóa theo hàng của ma trận kề **A**. Bước chuẩn hóa này đảm bảo rằng tổng của mỗi hàng trong ma trận kết nối bằng 1, tạo ra phân bố xác suất hợp lệ.

$$C[i,j] = \frac{A[i,j]}{\sum(A[i,:])}$$

c. Hệ số hóa ma trận không âm (Non-negative Matrix factorization): Hệ số hóa ma trận không âm (NMF) [52] là một kỹ thuật giảm kích thước được sử dụng để phân tích dữ liệu đa biến. Nó phân rã một ma trận không âm thành tích của hai ma trận không âm, điển hình là ma trận dữ liệu và một ma trận biểu diễn dựa trên các bộ phận. NMF đã được chứng minh là có hiệu quả phương pháp tóm tắt trích xuất [80, 81]. Cho X là ma trận không âm có kích thước m x n, trong đó m biểu thị số lượng tính năng (biến) và n đại diện cho số lượng mẫu (điểm dữ liệu). Hệ số hóa là được thực hiện bằng phép tính gần đúng:

$$\boldsymbol{X} \approx \boldsymbol{W} \times \boldsymbol{H}$$

Để áp dụng NNF, ma trận tương tự Cosine **X** (tương tự ma trận **A**) đầu tiên được tạo ra. Sau đó, NMF phân tích **X** thành hai ma trận **W** và **H**. Khi đó ma trận **W** được đưa vào multi-head attention của BERT.

4.4.2. Biểu diễn dữ liệu đầu vào

a. Khởi tạo đầu vào: Cho một văn bản đầu vào $d = \{s_1, s_2, \dots, s_n\}$ có n câu, bước đầu tiên là tạo một chuỗi đầu vào để nhúng ánh xạ. Để làm được điều đó, chúng ta nối hai đặc biệt mã thông báo: [CLS] và [SEP]. Trong đó mã thông báo[CLS] được chèn vào đầu mỗi câu để thể hiện ý nghĩa của câu; mã thông báo [SEP] được chèn vào cuối mỗi câu để tách hai câu. Việc nối n câu sẽ tạo ra một câu mới để nhúng vào kiến trúc Transformer. Việc tạo đầu vào được thể hiện ở phần xây dựng đầu vào như tại Hình 3.

b. Ánh xạ chuỗi đầu vào (Input mapping): Giả sử S là chuỗi mới được hình thành sau khi ghép tất cả các câu bằng mã thông báo [CLS] và [SEP]. Dãy *S* có thể được biểu diễn như sau.

$$S = [CLS]\ s_1\ [SEP][CLS]\ s_2\ [SEP]\dots\ [CLS]\ s_n\ [SEP]$$

Vector của mã thông báo [CLS] xuất ra từ BERT được sử dụng làm đại diện cho mỗi câu.

$$H = Encoder(S)$$

Chính xác hơn, vector ngữ cảnh *H* có thể được hiểu là tập hợp các vector ẩn của tất cả n mã thông báo [CLS] tương ứng với n câu trong chuỗi *S*.

$$H = \{h_1, h_2, \dots, h_n\}$$

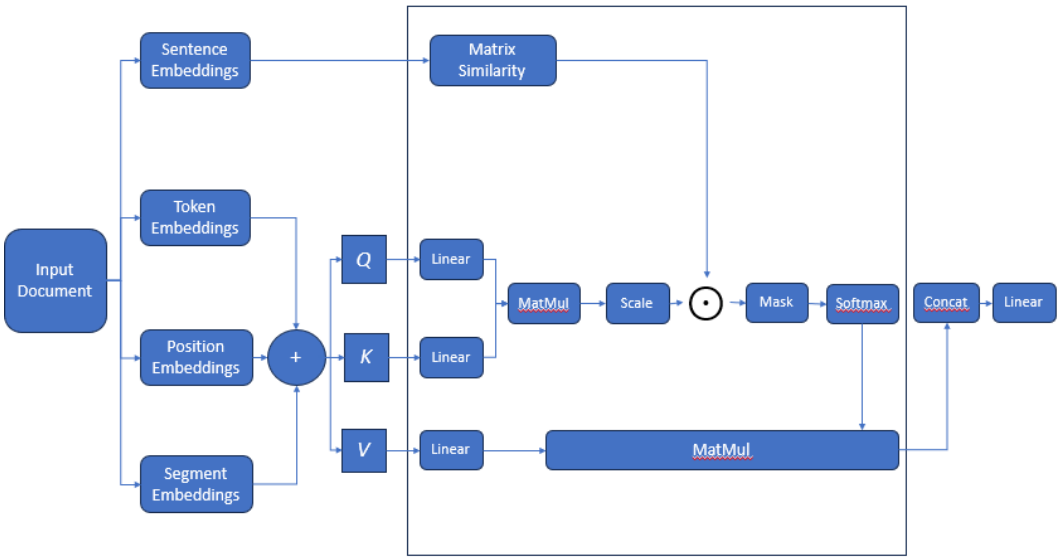
4.4.3. Bổ sung tri thức (Knowledge injection)

Để tạo ra tri thức, chúng tôi sử dụng SentenceBERT [90] để tính toán các phần nhúng câu và sau đó hình thành một trong ba ma trận (ma trận Cosine, ma trận LexRank và ma trận NMF) ***A***. Phần nhúng đầu vào là tổng của ba phần: nhúng mã thông báo, nhúng vị trí và nhúng phân đoạn. Các hàm chú ý của BERT có thể

được mô tả dưới dạng ánh xạ từ vector truy vấn Q và một tập hợp các cặp vector khóa-giá trị (K, V) tới vector đầu ra - cường độ chú ý. Tích số chấm của truy vấn với tất cả các khóa được tính như sau:

$$scores = QK^T$$

Kiến thức trước đó được đưa vào mô hình bằng cách tính tích số điểm theo từng phần tử với một trong ba ma trận (ma trận độ tương tự Cosine, ma trận LexRank, hệ số hóa ma trận không âm) $\boldsymbol{A}$ để làm cho mô hình chú ý hơn đến các cặp câu có độ tương đồng cao hơn trong tài liệu.



Hình 4. Bổ sung (chèn) tri thức cho BERT's multi-head attention.

Sau đó, phép tính được thu nhỏ lại $\sqrt{dk}$ và áp dụng hàm softmax để thu được trọng số trên các giá trị.

$$scores = QK^T \odot A + MASK$$

$$Attention(Q, K, V) = softmax\left(\frac{score}{\sqrt{dk}}\right)V$$

trong đó $\boldsymbol{A}$ biểu thị tiền tri thức dưới dạng ma trận tương quan.

Trong nghiên cứu này, tiền tri thức chỉ được đưa vào lớp chú ý đầu tiên và hai lớp đầu tiên của BERT. Sau khi bổ sung tri thức **Hình 4**, biểu diễn ẩn của chuỗi đầu vào S được biểu diễn như sau.

$$\hat{H} = \{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_n\}$$

Biểu diễn của $\hat{\boldsymbol{H}}$ là đầu vào để phân loại nhằm ước tính tầm quan trọng của các câu tương ứng với mỗi vector ẩn $\hat{\boldsymbol{h}}_{\boldsymbol{i}}$. Lớp đầu ra cuối cùng là bộ phân loại sigmoid:

$$\hat{y}_i = \sigma(W_o \hat{h}_i + b_o)$$

trong đó $\hat{y}_i$ là xác suất dự đoán của câu s_i cho thấy tầm quan trọng của câu. Xác suất được sử dụng để lựa chọn câu.

4.4.4. Chọn câu, sinh bản tóm tắt

Sau khi huấn luyện, mô hình được áp dụng vào các tập kiểm tra để trích xuất tóm tắt tài liệu. Sau khi xếp hạng các câu được dự đoán dựa trên tầm quan trọng (điểm số) của chúng, mô hình sử dụng thuật toán Mức độ liên quan cận biên tối đa (MMR) [26] để tạo thành bản tóm tắt cuối cùng. MMR lặp đi lặp lại xây dựng một bản tóm tắt bằng cách bao gồm câu có điểm cao nhất với công thức sau:

$$S_{next} = max_{s \in S - S_{sum}}(0.7 * f(s) - 0.3 * sim(s, S_{sum}))$$

trong đó *S* là tập hợp tất cả các câu trong tài liệu, S_{sum} chứa các câu hiện tại trong bản tóm tắt, *f(s)* là điểm số câu từ mô hình, *sim()* là độ tương tự Cosine của câu với S_{sum}. Khi trích xuất bản tóm tắt của một tài liệu mới, trước tiên MMR sử dụng mô hình để lấy điểm cho mỗi câu. Sau đó, nó xếp hạng các câu này theo điểm số của chúng để tạo ra một danh sách được sắp xếp (giảm dần) và chọn những câu được xếp hạng hàng đầu làm bản tóm tắt. Trong quá trình chọn câu, MMR sử dụng trigram để giảm sự dư thừa.

4.4.5. Huấn luyện và suy diễn (Training and inference)

Để huấn luyện, mô hình nhận tài liệu đầu vào và đưa kiến thức từ tài liệu đó vào quá trình huấn luyện. Việc biểu diễn các vector ngữ cảnh của mã thông báo [CLS] được sử dụng để dự đoán quyết định xem một câu có quan trọng hay không bằng cách sử dụng hàm sigmoid. Hàm mất mát (loss function) của mô hình là mất mát entropy chéo nhị phân (binary cross-entropy loss), đo lường sự khác biệt giữa xác suất dự đoán $\hat{y_i}$ và nhãn mục tiêu y_i trên N mẫu huấn luyện.

$$loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \, log\,(\hat{y_i}) + (1 - y_i)\, log\,(1 - \hat{y_i})]$$

Để suy luận, với một tài liệu đầu vào, mô hình được đào tạo ước tính tầm quan trọng của từng câu bằng cách sử dụng tri thức được đưa vào. Sau khi dự đoán, *m* câu quan trọng (có khả năng xác suất cao nhất) được chọn để tạo thành bản tóm tắt cuối cùng bằng thuật toán MMR.

4.5. Đánh giá thử nghiệm

4.5.1. Kho ngữ liệu thực nghiệm

Để đánh giá kết quả của mô hình đề xuất, luận án sử dụng 04 kho ngữ liệu gồm 02 kho ngữ liệu tiếng Anh, 02 kho ngữ liệu tiếng Việt. Đối với tiếng Anh, 02 kho ngữ liệu sử dụng phổ biến là CNN-DailyMail và BillSum (US, CA). Đối với tiếng Việt, sử dụng kho ngữ liệu VNDS và kho ngữ liệu được xây dựng tại Chương II là VNNEWS.100.2018.

4.5.2. Quy trình thực hiện

Chúng tôi đã sử dụng PyTorch và phiên bản "bert-base-uncased" BERT để triển khai mô hình. Phiên bản tiếng Anh của BERT [30] được sử dụng cho CNN/DailyMail và BillSum và phiên bản tiếng Việt của BERT [78] được sử dụng cho VNDS và VNNEWS.100.2028. Việc tiền xử lý văn bản, cả văn bản nguồn và bản tóm tắt vàng đều được tách từ bằng công cụ BERT's subwords tokenizer. Riêng đối với VNNEWS.100.2028 không phải tiền xử lý do tập ngữ liệu này đã được xử lý trước đó. SentenceBert được sử dụng để tách từ từng câu nhằm tạo ra tri thức từ độ tương tự Cosine, LexRank và các thuật toán phân tích hệ số không âm. Đối với tập ngữ liệu tiếng Việt VNDS và VNNEWS100.2018. Do kích thước đầu vào tối đa của mô hình pre-trained PhoBERT mới đạt 256 token, vì vậy, văn bản đầu vào chỉ giới hạn ở kích thước tối đa 256 từ. Để giải quyết vấn đề này, đối với 02 kho ngữ liệu tiếng Việt gồm bài toán tóm tắt báo mạng điện tử (thường có kích thước trên 500 từ), do vậy, trước hết cần chia văn bản thành các thành phần (block) gồm tập hợp các câu có kích thước dưới 256 từ theo phép phân chia tuyến tính. Mỗi thành phần sẽ được đưa vào mô hình PhoBERT để biểu diễn chuỗi đầu vào độc lập cho hệ thống tóm tắt. Sau đó, kết quả đầu ra của các thành phần sẽ được tổng hợp tuyến tính để sinh bản tóm tắt cuối cùng.

Tất cả các mô hình khai thác đều được đào tạo trong 50.000 bước với tốc độ học là 2e−3 trên GPU A100. Chúng tôi sử dụng trình tối ưu hóa Adam đã sử dụng với β1 = 0,9 và β2 = 0,999. Tốc độ học tập tuân theo phương pháp khởi động với giá trị 10000.

$$lr = 2e^{-3}.min(step^{-0.5}, step.warmup^{-1.5})$$

Điểm kiểm tra mô hình đã được lưu và đánh giá trên bộ xác thực sau mỗi 1.000 bước. 3 điểm kiểm tra hàng đầu được chọn dựa trên tổn thất đánh giá trên bộ xác thực và báo cáo kết quả trung bình trên bộ kiểm tra.

Đối với tập dữ liệu CNN/-DailyMail, sau khi xếp hạng, 3 câu có điểm cao nhất sẽ được trích xuất để tạo thành một bản tóm tắt [59, 111, 113]. Đối với tập dữ liệu BillSum, khi dự đoán tóm tắt cho một tài liệu mới, mô hình sử dụng thuật toán MMR chọn những câu quan trọng cho đến khi đạt giới hạn độ dài 2000 ký tự [71]. Đối với VNDS, mô hình chọn 2 câu trên cùng để tạo thành bản tóm tắt [84]. Đối với VNNews.100.2018, thực hiện huấn luyện mô hình trên tập ngữ liệu VNDS [84] với 02 trường hợp có và không sử dụng tri thức có sẵn theo Cosine, sau đó chạy mô hình được huấn luyện trên tập ngữ liệu VNNews.100.2018 và tạo 02 bản tóm tắt gồm: bản rút gọn 03 câu của mô hình được đối sánh với sa pô của bài báo và bản rút gọn với 10 câu được đối sánh với tập S30.

4.5.3. Phương pháp đánh giá

Để đánh giá chúng tôi sử dụng ROUGE_score với ROUGE-1.5.5 bằng cách sử dụng pyrouge4 với các tham số "-c 95 -2 -1 -U -r 1000 -n 2 -w 1.2 -a -s -f B -m". Các gói thư viện cung cấp một số ROUGE-score và F-score của ROUGE-1, ROUGE-2 và ROUGE-L được sử dụng làm thước đo chính cho việc tóm tắt văn bản.

4.5.4. Kết quả thực nghiệm

4.5.4.1. Về hiệu suất

Việc so sánh hiệu suất được thực hiện với hai kịch bản. Kịch bản đầu tiên là xác nhận sự đóng góp của việc đưa tri thức vào mô hình đề xuất. Cài đặt thứ hai là thử thách mô hình được đề xuất bằng các phương pháp tiên tiến (SOTA) để tóm tắt trích xuất ba bộ dữ liệu.

Bổ sung tri thức sẵn có và không bổ sung tri thức: Điểm ROUGE trong **Bảng 4**. cho thấy rằng việc bổ sung kiến thức mang lại lợi ích cho mô hình được đề xuất cho việc tóm tắt trích xuất. Điểm ROUGE của mô hình được đề xuất luôn tốt hơn so với điểm cơ sở không sử dụng phương pháp bổ sung tri thức. Mô hình sử dụng NMF kém hơn mô hình sử dụng Cosine và LexRank với biên độ rất nhỏ. Tồn tại những khoảng cách rất nhỏ giữa hai mô hình: sử dụng phương pháp bổ sung tri thức và không sử dụng phương pháp bổ sung tri thức. Nó cho thấy mô hình xương sống sử dụng BERT là một phương pháp mạnh mẽ để tóm tắt trích xuất [59]. Bằng cách sử dụng các mẫu đào tạo để tinh chỉnh, nó có thể đạt được kết quả cạnh tranh.

Bảng 4. Kết quả trích rút câu giá trị **In đậm** là kết quả tốt nhất với $p \leq 0.05$

Data	Method	ROUGE-1	ROUGE-2	ROUGE-L
CNN-DailyMail	No injection	42.80	19.19	39.37
	Injection (Cosine)	**43.13**	**20.19**	**39.54**
	Injection (LexRank)	42.96	20.05	39.53
	Injection (NMF)	42.87	20.08	39.45
BillSum (US)	No injection	41.56	21.23	38.35
	Injection (Cosine)	**42.09**	**21.87**	38.54
	Injection (LexRank)	41.85	21.56	**38.69**
	Injection (NMF)	41.85	21.56	38.55
BillSum (CA)	No injection	44.61	19.24	41.13
	Injection (Cosine)	44.69	19.28	41.33

	Injection (LexRank)	**44.96**	**19.84**	**41.51**
	Injection (NMF)	44.72	19.45	41.26
VNDS	No injection	52.80	23.96	37.13
	Injection (Cosine)	**53.43**	**24.35**	**37.55**
	Injection (LexRank)	52.89	24.01	37.02
	Injection (NMF)	52.91	23.99	36.46
VNNews.100.2018 (chapeau)	No injection	42.95	19.33	31.44
	Injection Cosine	**43.26**	**19.57**	**31.97**
VNNews.100.2018 (S30)	No injection	32.67	17.01	26.86
	Injection Cosine	**33.49**	**17.55**	**27.11**

So sánh với các mô hình tiên tiến - SOTA models: Kịch bản thứ hai là thách thức mô hình đề xuất bằng các phương pháp tóm tắt rút trích mạnh mẽ trên bốn bộ dữ liệu kiểm tra (test set). Kịch bản này xác nhận khoảng cách giữa mô hình đề xuất và các phương pháp SOTA bao gồm các phương pháp sau: **HIBERT**, **PNBERT**, **BERTSum**, **DiscoBERT**, **MatchSum**, **DOC**, **DOC+SUM**, **SumBasic**, **SVR**, **CNN**, **LSTM**.

Bảng 5. Kết quả trích rút câu, giá trị **In đậm** là kết quả tốt nhất.

Data	**Method**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
CNN-DailyMail	HIBERT [111]	42.37	19.95	38.83
	PNBERT [113]	42.69	19.60	38.83
	BERTSum [59]	43.85	20.34	39.90
	DiscoBERT [108]	43.77	20.85	40.67
	MatchSum [112]	**44.41**	**20.86**	**40.55**
	Our model (Cosine)	43.13	20.19	39.54
	Our model (LexRank)	42.96	20.05	39.53
	Our model (NMF)	42.87	20.08	39.45
BillSum (US)	DOC [50]	38.51	21.38	31.49
	SUM [50]	40.69	23.88	33.65
	DOC+SUM [50]	40.80	23.83	33.73
	Our model (Cosine)	**42.09**	**21.87**	38.54
	Our model (LexRank)	41.85	21.56	**38.69**
	Our model (NMF)	41.85	21.56	38.55
BillSum (CA)	DOC [50]	38.35	19.76	32.89
	SUM [50]	38.90	20.79	33.20
	DOC+SUM [50]	39.65	**21.14**	34.05
	Our model (Cosine)	44.69	19.28	41.33
	Our model (LexRank)	**44.96**	19.84	**41.51**
	Our model (NMF)	44.72	19.45	41.26
VNDS	SumBasic [84]	52.65	19.13	26.32
	SVR [84]	50.41	23.67	35.02
	CNN [84]	48.17	21.93	33.73
	LSTM [84]	46.56	20.29	32.49
	Our model (Cosine)	**53.43**	**24.35**	**37.55**
	Our model (LexRank)	52.89	24.01	37.02
	Our model (NMF)	52.91	23.99	36.46

Điểm ROUGE trong **Bảng 5**. phù hợp với điểm số trong Bảng 6. trong đó mô hình đề xuất thu được kết quả đầy hứa hẹn. Đối với tập dữ liệu CNN-DailyMail, MatchSum [112] cho kết quả tốt nhất, tiếp theo là DiscoBERT [108], BERTSum [59] và mô hình của chúng tôi. Điều này là do MatchSum sử dụng kết hợp ngữ

nghĩa bằng cách sử dụng Siamese-BERT có thể khớp chính xác bản tóm tắt vàng với các ứng cử viên. Mô hình DiscoBERT mã hóa cấu trúc diễn ngôn vào quy trình tóm tắt bằng cách sử dụng bộ mã hóa biểu đồ diễn ngôn. Ngược lại, mô hình đề xuất với kiến trúc khá đơn giản để có thể cạnh tranh được với hai phương pháp phức tạp. Tuy nhiên, khoảng cách giữa mô hình đề xuất và hai phương pháp mạnh là không nhiều. So với HIBERT và PNBERT, mô hình đề xuất cho kết quả tốt hơn. Đối với tập dữ liệu BillSum và VNDS, mô hình đề xuất thể hiện sự cải thiện đáng kể so với các phương pháp cơ sở. Ví dụ: trong tập dữ liệu BillSum, nó cho kết quả tốt hơn DOC+SUM, một phương pháp tổng hợp cũng sử dụng BERT để trích xuất các câu quan trọng.

So sánh VNDS và VNNEWS.100.2018: Điểm ROUGE trong **Bảng 6** cho thấy cũng tương đồng với các kết quả thực nghiệm trên 04 tập dữ liệu tóm tắt văn bản tại Mục 4.3, việc bổ sung tri thức có sẵn trong văn bản báo mạng điện tử cũng đã cải thiện được hiệu năng của mô hình. Kết quả thực nghiệm trên tập dữ liệu VNDS cao hơn so với thực nghiệm trên tập VNNews.100.2018 là hợp lý do VNDS là tập dữ liệu đã được xây dựng đồng nhất. Kết quả đối sánh trên tập chapeau của VNNews.100.2018 tốt hơn so với kết quả đối sánh trên tập S30 do dữ liệu được huấn luyện trên VNDS với các bản rút gọn là sa pô của các bài báo thu thập, không phải là bản tóm tắt hoàn chỉnh của văn bản. Đối với VNDS, mô hình đề xuất đạt kết quả tốt nhất. Những kết quả này xác nhận tính hiệu quả của phương pháp đề xuất của chúng tôi, sử dụng phương pháp bổ sung tri thức sẵn có để cải thiện chất lượng lựa chọn câu trong bài toán tóm tắt văn bản tiếng Việt.

Bảng 6. Kết quả VNDS và VNNEWS.100.2018

Train	Test	Method	ROUGE-1	ROUGE-2	ROUGE-L
VNDS	VNNews.100.2018 (chapeau)	No injection	42.95	19.33	31.44
		Injection Cosine	**43.26**	**19.57**	**31.97**
	VNNews.100.2018 (S30)	No injection	32.67	17.01	26.86
		Injection Cosine	**33.49**	**17.55**	**27.11**
	VNDS	No injetion	52.80	23.96	37.13
		Injection Cosine	**53.43**	**24.35**	**37.55**

4.5.4.2. Về hiệu quả các kỹ thuật

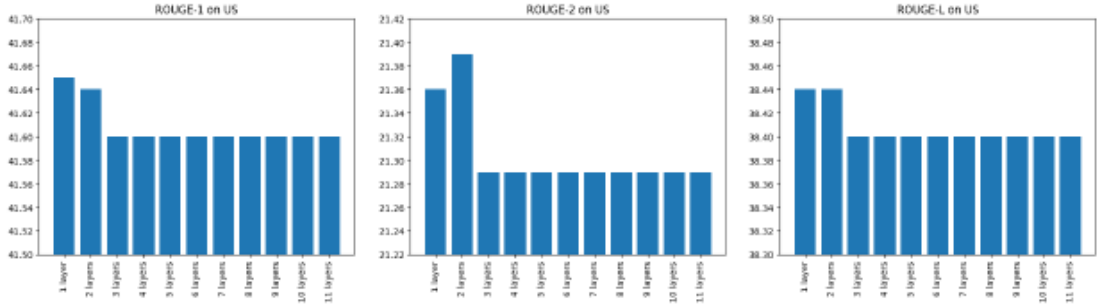
So sánh giữa Trích rút và Tóm lược: Quan sát này nhằm xác nhận sự đóng góp của việc đưa tri thức vào việc tóm tắt mang tính trích rút và tóm lược. Trong trường hợp tóm tắt trích rút, mô hình sử dụng kiến trúc tương tự trong **Hình 4**. Để tóm tắt tóm lược, chúng tôi đã sử dụng mã hóa-giải mã framework tiêu chuẩn. Bộ mã hóa là BERT được đào tạo trước với việc đưa tri thức vào **Hình 4**. và bộ giải mã sử dụng Transformer 6 lớp được khởi tạo ngẫu nhiên.

Bảng 7. Kết quả tóm tắt trích rút và tóm lược trên bộ dữ liệu CNN-DailyMail

Data	**Method**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Extraction	No injection	42.80	19.19	39.37
	Injection (Cosine)	**43.13**	**20.19**	**39.54**
	Injection (LexRank)	42.96	20.05	39.53
	Injection (NMF)	42.87	20.08	39.45
Abstraction	No injection	**40.78**	**18.75**	**37.88**
	Injection (Cosine)	40.36	18.37	37.47
	Injection (LexRank)	40.32	18.31	37.44
	Injection (NMF)	40.33	18.32	37.44

Các kết quả trong **Bảng 7** chỉ ra rằng kiến thức sẵn có ở dạng ma trận tương quan dường như thích hợp cho việc tóm tắt rút trích. Một lý do có thể là ma trận tương quan thể hiện mối quan hệ giữa các câu trong tài liệu phù hợp hơn cho việc trích rút. Do đó, việc đưa tri thức có trước vào mô hình trích rút giúp cải thiện kết quả tóm tắt. Ngược lại, tóm tắt tóm lược đòi hỏi sự hiểu biết sâu sắc hơn về toàn bộ tài liệu. Kết quả là, việc đưa kiến thức vào dưới dạng ma trận tương quan không cải thiện được điểm ROUGE. Nó gợi ý việc điều tra các cách biểu diễn kiến thức trước khác nhau để tóm tắt một cách trừu tượng.

Đánh giá sự chú ý với tri thức bổ sung: Ở đây chúng ta sẽ quan sát hành vi của mô hình khi sử dụng tính năng chèn tri thức qua việc kiểm tra để đánh giá tác động của tri thức sẵn có đó lên các lớp chú ý (Attention). Để làm được điều đó, tri thức sãn có đó được đưa vào từng lớp ($1 \leqslant L \leqslant 11$). Ví dụ: với L = 1, kiến thức từ LexRank đã được đưa vào lớp BERT đầu tiên. L = 2, kiến thức từ LexRank được đưa vào lớp thứ nhất và thứ hai của BERT. Sau đó, mô hình được đào tạo lại trên tập dữ liệu BillSum và được đánh giá trên các bộ thử nghiệm của hóa đơn Hoa Kỳ và CA. Vì các kết quả tương đối giống nhau nên chúng tôi chỉ báo cáo quan sát dựa trên kiến thức từ LexRank.



Hình 5. Tri thức được bổ sung từ LexRank vào US BillSum cho mỗi lớp.

Các kết quả được mô tả trong **Hình 5** cung cấp những hiểu biết sâu sắc về hiệu suất của mô hình được đề xuất với việc đưa kiến thức từ LexRank vào các lớp khác nhau. Các phát hiện chỉ ra rằng mô hình đạt được kết quả tốt nhất khi tri thức được đưa vào lớp đầu tiên hoặc hai lớp đầu tiên của mô hình. BERT. Một lý do có thể là mô hình cần có tri thức trước ở một số lớp đầu tiên. Sau khi học, sự đóng góp của kiến thức ở các lớp tiếp theo dường như đã bão hòa. Bằng cách tóm tắt các kết quả, chúng tôi có thể kết luận rằng việc kết hợp tri thức sẵn có vào lớp đầu tiên hoặc hai lớp đầu tiên của mô hình BERT sẽ nâng cao khả năng học và hiểu mối tương quan giữa các câu trong tài liệu.

4.6. Kết luận Chương IV

Các kết quả Chương IV đã đạt được gồm: Đã nghiên cứu và trình bày về tri thức có sẵn trong văn bản là các tri thức được sử dụng trong các phương pháp học không giám sát (unsupervised learning) và đề xuất phương pháp tóm tắt văn bản trích rút dựa trên mô hình huấn luyện trước có bổ sung tri thức có sẵn trong văn bản. Thực nghiệm mô hình đề xuất trên các kho ngữ liệu chuẩn của cả hai ngôn ngữ tiếng Anh và tiếng Việt, phân tích kết quả đầu ra cua mô hình đề xuất bao gồm cả kết quả thực nghiệm trên bộ dữ liệu văn bản báo mạng điện tử. Kết quả thực nghiệm cho thấy mô hình đề xuất có cải tiến nhất định về hiệu quả trong bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt. Đồng thời, với mô hình đề xuất trên, tri thức cho trước được tính toán riêng biệt, bổ sung độc lập trong quá trình tính toán, do vậy có thể nghiên cứu, sử dụng nhiều dạng tri thức có sẵn của văn bản để nâng cao hiệu suất bài toán trích rút câu.

KẾT LUẬN

Tóm tắt văn bản là một trong lĩnh vực quan trọng của xử lý ngôn ngữ tự nhiên. Trong đó, bài toán tóm tắt văn bản tiếng Việt, có ý nghĩa quan trọng trong việc nâng cao hiệu quả khai thác, xử lý thông tin từ các kho ngữ liệu, tài liệu văn bản tiếng Việt, nâng cao hiệu suất tìm kiếm, tổng hợp thông tin. Đối với lĩnh vực quản lý nhà nước về thông tin và truyền thông, việc quản lý thông tin, dư luận, báo mạng trên Internet đóng vai trò rất quan trọng. Mục tiêu của luận án này nghiên cứu một số phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt, có thể ứng dụng vào thực tiễn để xây dựng các phần mềm tóm tắt văn bản báo mạng điện tử tiếng Việt phục vụ công quản lý thông tin và truyền thông. Ở đó, lĩnh vực quản lý báo mạng điện tử được các cơ quan quản lý nhà nước đặc biệt quan tâm do các đặc điểm tính chất của thể loại phương tiện thông tin này. Vì vậy, việc nghiên cứu, nâng cao hiệu suất và độ chính xác của tóm tắt văn bản tiếng Việt, trong đó có văn bản báo mạng điện tử tiếng Việt cần được quan tâm về cả phương diện khoa học và thực tiễn. Chính vì vậy, mục tiêu của luận án này nghiên cứu một số phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt, có thể ứng dụng vào thực tiễn để xây dựng các phần mềm tóm tắt văn bản báo mạng điện tử tiếng Việt phục vụ công tác quản lý nhà nước về thông tin và truyền thông.

I. Các kết quả đạt được của luận án

1. Đã nghiên cứu các phương pháp tóm tắt văn bản và tóm tắt văn bản tiếng Việt. Đã nghiên cứu các đặc trưng của văn bản báo mạng điện tử tiếng Việt và qua đó xây dựng kho ngữ liệu tóm tắt văn bản báo mạng điện tử có đặc trưng riêng gồm sa pô và từ gán nhãn.

2. Đã đề xuất phương pháp tính độ tương đồng câu trong văn bản báo mạng điện tử tiếng Việt; xây dựng kho ngữ liệu phục vụ bài toán tóm tắt văn bản báo mạng điện tử tiếng Việt.

3. Đã nghiên cứu, thử nghiệm phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt dựa trên đồ thị với 02 giải thuật học không giám sát (TextRank và LexRank).

4. Đã nghiên cứu mô hình huấn luyện trước (pre-trained model), đề xuất phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt dựa trên mô hình huấn luyện trước, có bổ sung các tri thức có sẵn trong văn bản.

II. Những đóng góp mới của luận án

1. Đã chỉ ra vấn đề hạn chế trong xây dựng kho ngữ liệu văn bản tiếng Việt, sự khác biệt giữa sa pô của bài báo mạng điện tử với bản tóm tắt do con người thực hiện và đề xuất phương pháp tính độ tương đồng câu trong văn bản báo mạng điện tử tiếng Việt dựa trên đặc trưng của thể loại văn bản này. Xây dựng kho ngữ liệu tóm tắt văn bản báo mạng điện tử có đặc trưng riêng gồm sa pô và từ gán nhãn để thử nghiệm phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt dựa trên đồ thị sử dụng độ tương đồng câu theo đặc trưng của thể loại văn bản này.

2. Đề xuất phương pháp tóm tắt văn bản báo mạng điện tử tiếng Việt dựa trên mô hình huấn luyện trước, có bổ sung tri thức có sẵn phương pháp tính độ tương đồng câu trong văn bản.

III. Hướng nghiên cứu tiếp theo

- Mở rộng tập đặc trưng của văn bản báo mạng điện tử tiếng Việt.

- Xây dựng kho ngữ liệu văn bản báo mạng điện tử tiếng việt đủ lớn cho mô hình huấn luyện trước, bao gồm đầy đủ các đặc trưng riêng có của thể loại văn bản này.

- Nghiên cứu, khai thác tri thức có sẵn trong văn bản báo mạng điện tử để nâng cao hiệu suất, độ chính xác trong bài toán tóm tắt trích rút câu sử dụng mô hình học sẵn.

DANH MỤC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ

[CT1] Thắng, L., & Minh, L. (2018). Một số đặc trưng trong tóm tắt văn bản báo mạng điện tử tiếng Việt. In Kỷ yếu Hội nghị Khoa học công nghệ quốc gia lần thứ XI về Nghiên cứu có bản và ứng dụng công nghệ thông tin (FAIR), (pp. 330 - 335).

[CT2] Thắng, L., Minh, L. and Sơn, P. (2020). Tóm tắt báo mạng điện tử tiếng Việt sử dụng TextRank. In Kỷ yếu Hội nghị Khoa học công nghệ quốc gia lần thứ XIII về Nghiên cứu có bản và ứng dụng công nghệ thông tin (FAIR), (pp. 623 - 627).

[CT3] Le Ngoc Thang, Le Quang Minh (2023), Vietnamese online newspapers summarization using LexRank, Сборник научных трудов по материалам Международной научно-практической конференции 28 декабря 2023г.: Белгород, ISSN 2713-1513.

[CT4] Thang Le Ngoc, Minh Le Quang (2024), “Vietnamese Online Newspaper summarization using Pre-trained model”, Актуальные исследования: МЕЖДУНАРОДНЫЙ НАУЧНЫЙ ЖУРНАЛ (CURRENT RESEARCH: INTERNATIONAL SCIENTIFIC JOURNAL). №2 (184), 09 – 16.

[CT5] Ngoc-Thang Le, Minh-Tien Nguyen, Nhat-Minh Do, Chi-Thanh Nguyen and Quang-Minh Le. A method to utilize prior knowledge for extractive summarization based on pre-trained language models. Submitted to Vietnam Journal of Science and Technology on 01 March 2024 UTC with Submission ID 20241.