ISSN 2354-0575

SỬ DỤNG BERT CHO TÓM TẮT VĂN BẢN TIẾNG VIỆT

Bùi Đức Thọ*, Đỗ Thị Thu Trang, Ngô Thanh Huyền
Trường Đại học Sư phạm Kỹ thuật Hưng Yên
** Tác giả liên hệ: buiducthobn@gmail.com*

Ngày tòa soạn nhận được bài báo: 27/10/2021
Ngày phản biện đánh giá và sửa chữa: 24/11/2021
Ngày bài báo được duyệt đăng: 02/12/2021

Tóm tắt:

Bài báo này giới thiệu phương pháp tóm tắt văn bản theo hai hướng trích rút và tóm lược, sử dụng mô hình ngôn ngữ huấn luyện trước. Để làm điều này, đối với bài toán trích rút, chúng tôi sử dụng mô hình BERTSum. Mô hình sử dụng BERT (Bidirectional Encoder Representations from Transformers) để mã hoá các câu đầu vào và dùng LSTM (Long Short Term Memory Networks) để biểu diễn mối quan hệ giữa các câu. Đối với bài toán tóm lược, chúng tôi sử dụng BERT để mã hóa ngữ nghĩa của văn bản đầu vào để sinh ra bản tóm tắt phù hợp. Chúng tôi thử nghiệm phương pháp trên bộ dữ liệu tiếng Việt được chia sẻ từ bài báo VNDS (A Vietnamese Dataset for Summarization) [19] và đánh giá phương pháp bằng ROUGE (Recall - Oriented Understudy for Gisting Evaluation). Kết quả thực nghiệm cho thấy giữa hai bài toán tóm tắt trích rút và tóm tắt tóm lược BERT đạt hiệu quả hơn ở bài toán tóm tắt trích rút.

Từ khóa: *Tóm tắt văn bản, xử lý ngôn ngữ, học máy, học sâu, học không giám sát.*

1. Giới thiệu

Tóm tắt văn bản tự động là phương pháp rút gọn lại một lượng lớn các thông tin thành một bài toán tóm tắt cô đọng, ngắn gọn, lựa chọn được thông tin quan trọng và bỏ thông tin dư thừa [1]. Kết quả tóm tắt đảm bảo giữ được các thông tin quan trọng, đúng ngữ nghĩa, chính tả, không làm sai lệch nội dung văn bản gốc. Tóm tắt văn bản luôn là một nhiệm vụ đầy thách thức đối với xử lý ngôn ngữ tự nhiên (NLP: Natural Language Processing). Các hướng nghiên cứu được xem qua ba cách tiếp cận: các kỹ thuật làm nổi bật câu đơn giản (từ những năm 1950), học máy cổ điển (1990), và học sâu (2015) [1]. Đầu ra của một hệ thống tóm tắt văn bản mang lại lợi ích cho nhiều ứng dụng NLP như tìm kiếm Web. Công cụ tìm kiếm Google thường trả về một đoạn mô tả ngắn về các trang Web tương ứng với truy vấn tìm kiếm, hoặc nhà cung cấp tin tức trực tuyến cung cấp các điểm nổi bật của tài liệu Web trên giao diện của nó. Điều này đòi hỏi các hệ thống tóm tắt văn bản chất lượng cao.

Bài toán tóm tắt văn bản thường sử dụng các kỹ thuật theo hai hướng sau: học có giám sát [2-5] và học không giám sát [6]. Ngoài hai kỹ thuật trên thì học sâu cũng là một kỹ thuật mới, nhiệm vụ của học sâu là tìm ra mô hình dữ liệu trừu tượng hóa ở mức cao bằng cách sử dụng tập hợp các thuật toán với nhiều lớp xử lý với cấu trúc phức tạp như: Mạng nơ ron tích chập CNN (Convolutional Neural Network); Mạng nơ ron hồi quy LSTM.

Những thành công gần đây của các mô hình ngôn ngữ huấn luyện trước là hướng tiếp cận mới cho bài toán tóm tắt văn bản. Mặc dù tóm tắt văn bản trích rút và tóm lược đã được áp dụng cho tiếng Anh với những kết quả khả quan, ứng dụng kỹ thuật này cho bài toán tóm tắt văn bản tiếng Việt vẫn còn hạn chế. Trong bài báo này chúng tôi giới thiệu mô hình tóm tắt văn bản theo hướng trích rút và hướng tóm lược dựa trên BERT (Bidirectional Encoder Representations from Transformers)[21]. Những đóng góp chính của bài báo này như sau:

- Bài báo sử dụng mô hình ngôn ngữ huấn luyện trước cho tóm tắt văn bản tự động, tập trung vào tóm tắt văn bản theo hai hướng trích rút và tóm lược dựa trên BERT.
- Bài báo so sánh kết quả của mô hình với các phương pháp khác. Kết quả cho thấy mô hình tóm tắt dựa trên BERT cho kết quả khả quan.

Phần còn lại của bài viết này được tổ chức như sau. Phần 2 cung cấp các nghiên cứu liên quan.

Phần 3 giới thiệu mô hình. Tiếp theo, Phần 4 trình bày kết quả thực nghiệm và thảo luận. Cuối cùng, Phần 5 đưa ra kết luận và hướng phát triển.

2. Các nghiên cứu liên quan

Các nghiên cứu tóm tắt văn bản đã được trình bày trong tài liệu [1, 11, 16, 18, 21]. Các nghiên cứu đã có những cách tiếp cận bài toán tóm tắt văn bản với một số dạng khác nhau: bài toán tóm tắt văn bản theo hướng trích rút [16], bài toán tóm tắt văn bản theo hướng tóm tắt [18], Một số mô hình huấn luyện trước [21].

Các kỹ thuật tóm tắt theo hướng trích rút là sinh ra các đoạn tóm tắt văn bản bằng cách chọn một tập các câu trong văn bản gốc, các đoạn tóm tắt này chứa các câu quan trọng nhất của đầu vào 16]. Các phương pháp cơ bản của bàn toán tóm tắt trích rút gồm: Phương pháp TF-IDF (Term Frequency - Inverse Document Frequency) [3]. Phương pháp TextRank: Thuật toán kế thừa sự tính toán của thuật toán PageRank [8]; Phương pháp phân lớp dùng SVM (Support Vector Machine) [7,11]; Mạng nơ-ron tích chập (Convolutional Neural Network - CNN [2]; Mạng LSTM (Long Short Term Memory networks).

Các kỹ thuật tóm tắt theo hướng tóm lược là sinh ra bản tóm tắt cho văn bản đầu vào tương ứng. Bản tóm tắt dựa vào các đại lượng đặc trưng như từ, cụm từ, thành phần quan trọng của câu để tạo ra các câu mới cho văn bản tóm tắt [18], hoặc dựa trên phương pháp rút gọn câu để tạo ra bản tóm tắt. Các phương pháp cơ bản của bàn toán tóm tắt tóm lược gồm: Phương pháp ILP (Integer Linear Programming) [14]. Phương pháp Seq2seq: Mô hình chuỗi-tới-chuỗi (Sequence to Sequence - Seq2seq) [5] là một mô hình Deep Learning được đề xuất bởi nhóm tác giả *Ilya Sutskever*, *Oriol Vinyals*, *Quoc V. Le* trên bài báo năm 2014 [20]. Mục đích là tạo ra một chuỗi đầu ra từ một chuỗi đầu vào mà độ dài của 2 chuỗi này có thể khác nhau. Theo đó khi văn bản gốc được đưa vào thì sẽ được chuyển thành một văn bản khác có độ dài ngắn hơn mà vẫn mang đầy đủ ý nghĩa.

3. Giới thiệu BERT

Biểu diễn mã hóa hai chiều từ các sự biến đổi (Bidirectional Encoder Representations from Transformers) được gọi tắt là BERT [9] là một kỹ thuật học máy dựa trên sự biến đổi (Transformer), được dùng cho việc huấn luyện trước xử lý ngôn ngữ tự nhiên (NLP) được phát triển bởi Jacob Devlin và cộng sự từ Google đã tạo ra và công bố BERT vào năm 2018 [9, 17]. BERT được thiết kế để huấn luyện các mô hình hai chiều sâu sắc từ văn bản không được gắn nhãn bằng cách điều chỉnh chung trên cả hai bối cảnh bên trái và bên phải trong tất cả các lớp. BERT khác biệt với các mô hình một chiều (Unidirectional) khi chỉ học các biểu diễn từ trái qua phải hoặc từ phải qua trái. Chính vì vậy, mô hình BERT được huấn luyện trước có thể được tinh chỉnh chỉ với một lớp đầu ra bổ sung để tạo ra các mô hình hiện đại cho một loạt các tác vụ. Chẳng hạn trả lời câu hỏi và suy luận ngôn ngữ, không có sửa đổi kiến trúc của tác vụ cụ thể.

3.1. BERT cho bài toán trích rút

Với mô hình tóm tắt trích rút văn bản dựa trên BERT, các từ trong một câu đầu vào sẽ được biến đổi bằng BERT để có được một véc tơ đầu ra. Véc tơ này là biểu diễn của câu đầu vào. Véc tơ đó sẽ là đầu vào của một mạng nơ ron truyền thẳng (Feed-forward Network). Mạng này sẽ cho ra một véc tơ cuối cùng cho phân lớp. Kết quả ở bộ phân lớp cho biết câu đó có là câu tóm tắt hay không [4]. Để sử dụng BERT cho tóm tắt trích rút, chúng tôi sử dụng mô hình BERTSUM là mô hình trích rút tốt nhất dựa trên BERT.

Mã hóa dữ liệu đầu vào: Để mã hóa được nhiều câu trong văn bản đầu vào, các mã thông báo [CLS] và [SEP] được chèn vào trước và sau mỗi câu. Các câu của văn bản sẽ được mã hóa bằng BERT sử dụng ba loại nhúng (embeddings): nhúng cho các từ, nhúng cho phân đoạn, và em nhúng cho vị trí. Sau khi mã hõa, các câu được biểu diễn dưới dạng các véc tơ ngữ nghĩa và các véc tơ này là đầu vào cho bộ phân lớp nhị phân.

Tinh chỉnh với lớp tóm tắt: Sau khi lấy các véc tơ từ BERT, BERTSUM xây dựng một số lớp cụ thể về tóm tắt xếp chồng lên nhau trên đỉnh đầu ra của BERT. BERT tinh chỉnh các lớp tóm tắt này thông qua mô hình biến đổi liên câu (Inter-sentence Transformer), mạng Nơ ron hồi quy (Recurent Neural network) và bộ phân loại đơn giản (Simple Classifier). Chúng tôi tiếp tục sử dụng Pytorch và phiên bản Bert-Base-Uncased để triển khai mô hình. BERT và các lớp tóm tắt cùng được tinh chỉnh. Bài báo sử dụng trình tối ưu hóa Adam với $\beta_1 = 0.9$, β_2

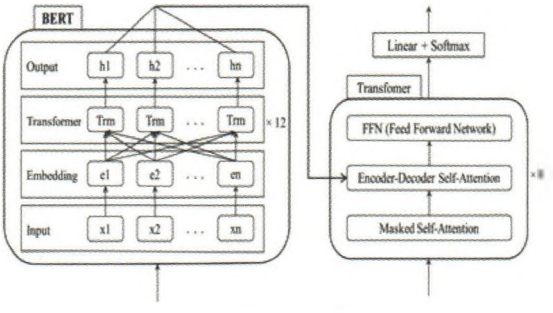
= 0.999 cho quá trình tinh chỉnh. Sử dụng Learning Rate (tốc độ học) theo [19]:

$$lr = 2e^{-3} * \min(step^{-0.5}, step * warmup^{-1.5}) \quad (1)$$

Sử dụng GPU GeForce RTX 2080 Ti để huấn luyện mô hình. Sử dụng hàm cross-entropy và Trigram Blocking để giảm bớt sự dư thừa (*Paulus và cộng sự, 2018*) [20].

3.2. BERT cho bài toán tóm tắt tóm lược

Khi ứng dụng BERT cho tóm tắt tóm lược là hình thành vấn đề dưới dạng nhiệm vụ tuần tự với hai thành phần chính: bộ mã hóa và bộ giải mã. Bộ mã hóa sẽ mã hóa chuỗi đầu vào thành các vectơ trạng thái và bộ giải mã sẽ giải mã từng mã thông báo thành tài liệu tóm tắt. BERT phù hợp cho mô hình huấn luyện trước và thích hợp cho huấn luyện nhanh với độ chính xác cao hơn các mô hình hiện có. Chúng tôi đưa ra thiết kế của một bộ giải mã dựa trên sự biến đổi làm mã hóa, nhưng chỉ sử dụng 8 khối biến đổi và một lớp tuyến tính cùng với lớp *softmax* ở trên cùng.



Hình 1. Mô hình tóm tắt tóm lược

Đầu vào được ký hiệu là $X = x_1, x_2, \ldots, x_n$ và tóm tắt tương ứng được ký hiệu là $A = a_1, a_2, \ldots, a_n$. Mô hình bắt đầu bằng cách nhập X và thu được đại diện của mỗi mã thông báo H. Mô hình đưa H vào và đầu ra của bộ giải mã ở bước thời gian thứ t. Xác suất của từ vựng ở bước thời gian thứ t có thể đạt được như thể hiện trong phương trình sau:

$$P_t(w) = f_{decoder}(w \mid H, Y < t) \quad (2)$$

Xác suất này được điều chỉnh trên đầu ra của bộ giải mã cho đến bước thời gian thứ t và đầu ra của bộ mã hóa H. Lặp lại cho đến khi tạo ra ký tự cuối dãy <*EOS*> hoặc đạt đến giới hạn ký tự. Khi huấn luyện ta sử dụng thêm hàm cross-entropy.

$$\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \cdot \log(\hat{y}_{ij}) \quad (3)$$

Trong đó C là số câu, N là mẫu huấn luyện.

3.3. Huấn luyện

Chúng tôi sử dụng mô hình BERT [19] cho quá trình huấn luyện. Mô hình BERT được huấn luyện sử dụng hai bài toán là MLM và NSP. Hàm mất mát được tính trên hai bài toán nhỏ. BERT gốc có hai phiên bản với hai kích thước mô hình khác nhau [12]. Mô hình cơ bản ($BERT_{BASE}$) sử dụng 12 lớp (khối mã hóa của Transformer) với 768 nút ẩn (kích thước ẩn) và lớp tự tập trung 12 đầu, tổng số lượng tham số là 110 triệu. Mô hình lớn ($BERT_{LARGE}$) sử dụng 24 lớp với 1024 nút ẩn và tầng tự tập trung 16 đầu, tổng số lượng tham số là 340 triệu.

4. Kết quả thực nghiệm và thảo luận

4.1. Dữ liệu

Bài báo sử dụng bộ dữ liệu VNDS (A Vietnamese Dataset for Summarization) [19] đã được sưu tầm từ trước trong các chuyên mục như: "Thế giới", "Tin tức", "Luật" và "Kinh doanh". Tiếp theo là xử lý dữ liệu, sử dụng NLTK 2 để phân đoạn câu và sử dụng công cụ vitk 3 để phân đoạn từ. Cuối cùng bộ dữ liệu được chia thành 3 phần: 70 % cho huấn luyện, 15 % cho thẩm định, 15 % cho thử nghiệm. Cụ thể ta có bảng sau:

Bảng 1. Số liệu thống kê của bộ dữ liệu

	Huấn luyện	Thẩm định	Thử nghiệm
Số lượng tài liệu	105418	22642	22644
Số câu trung bình trong phần tóm tắt	1.22	1.22	1.23
Số từ trung bình trong phần tóm tắt	28.48	28.54	28.59
Số câu trung bình trong nội dung	17.72	17.81	17.72
Số từ trung bình trong nội dung	418.37	419.66	418.74

4.2. Thiết đặt thực nghiệm

Bộ dữ liệu thành 3 phần: bộ huấn luyện, bộ thẩm định và bộ thử nghiệm. Tất cả các phương pháp xếp hạng không giám sát chỉ được áp dụng cho bộ thử nghiệm. Các phương pháp học sâu khác đã được huấn luyện trên bộ huấn luyện với việc sử dụng bộ thẩm định. Đối với phương pháp trích rút văn bản, ta thiết đặt số câu được trích rút là 2 (chọn hai câu làm giá trị trung bình của số câu trong phần tóm tắt).

Mô hình $BERT_{BASE}$ được sử dụng để huấn luyện bộ tóm tắt. Mô hình gồm 12 lớp, 110 triệu

tham số được huấn luyện sử dụng dữ liệu Wiki cho nhiều ngôn ngữ, trong đó có tiếng Việt. Kích thước véc tơ của BERT là 768. Kích thước của mạng nơ-ron truyền thẳng là 256. Mô hình được huấn luyện với thuật toán tối ưu Adam, với 12 vòng lặp (12 Epochs), sử dụng một GeForce RTX 2080 Ti.

Với tóm tắt trích rút, mô hình tóm tắt sử dụng hai câu nhãn 1 (trong số các câu mà BERTSUM dự đoán nhãn là 1) với xác suất cao nhất để làm bản tóm tắt. Với tóm tắt tóm lược, độ dài của bản tóm tắt được sinh ra bằng với độ dài của bản tóm tắt mẫu của từng văn bản. Các thiết lập này hoàn toàn trùng khớp với bài báo *VNDS* [19].

4.3. Độ đo đánh giá

Để đánh giá bản tóm tắt có 2 phương pháp đó là đánh giá thủ công và đánh giá tự động. Việc đánh giá thủ công mặc dù hiệu quả nhưng nó vẫn tồn tại 2 vấn đề lớn. Thứ nhất là đòi hỏi sự can thiệp của con người, tốn kém chi phí, tốn công sức. Thứ hai là các phán đoán mang tính chủ quan. Chính vì vậy trong bài báo này, ta sẽ sử dụng độ đo ROUGE (Recall - Oriented Understudy for Gisting Evaluation) [12] để so sánh kết quả của các mô hình. Đây cũng là phương pháp phổ biến nhất để đánh giá hệ thống tóm tắt [15]. Các câu trích chọn và tóm lược sẽ được so khớp với các câu chuẩn trong tập dữ liệu để tính điểm ROUGE. ROUGE dựa vào sự giống nhau trên các từ (n-grams) để tính toán ra điểm. Do đó bản tóm tắt có điểm ROUGE càng cao thì càng tốt.

$$ROUGE-N=\frac{\sum_{s\in S_{ref}}\sum_{gram_n\in s}Count_{match}(gram_n)}{\sum_{s\in S_{ref}}\sum_{gram_n\in s}Count(gram_n)} \quad (4)$$

Trong đó n là độ dài của n - gram. $Count_{match}(gram_n)$ là số n-gram tối đa cùng xuất hiện trong một bản tóm tắt đề xuất và các câu chuẩn. $Count(gram_n)$ là số lượng n-gram trong các câu chuẩn. Bài báo sử dụng pyrouge[1] với tham số "-c 95 -2 -1 -U -r 1000 -n 2 -w 1.2 -a -s -f B –m".

Do độ dài bản tóm tắt đề xuất và các câu chuẩn là khác nhau, nên ta sử dụng F1-Score để cân bằng giữa độ chính xác (Precision) và độ hồi tưởng (Recall).

$$Precision=\frac{TP}{TP+FP} \quad (5)$$

$$Recall=\frac{TP}{TP+FN} \quad (6)$$

$$F_1=\frac{2\times Recall\times Precision}{Recall+Precision} \quad (7)$$

Trong bài báo này, chúng tôi sử dụng ROUGE-1, ROUGE-2 và ROUGE L cho quá trình so sánh.

- ROUGE-1: tính toán sự giống nhau dựa trên các từ đơn (uni-gram).
- ROUGE-2: tính toán sự giống nhau trên 2 từ liên tục (bi-gram)
- ROUGLE L: tính toán chuỗi dài nhất có thể tính

4.4. Các mô hình dùng để so sánh

Chúng tôi so sánh các mô hình tóm tắt trích rút sử dụng học sâu với một số thuật toán mạnh cho bài toán tóm tắt văn bản. Các thuật toán đó bao gồm:

• **Lead-*m***: thuật toán này đơn giản chỉ lấy m câu đầu tiên của một văn bản làm bản tóm tắt [22]. Mặc dù thuật toán đơn giản nhưng kết quả khá tốt cho bài toán tóm tắt trích rút đơn văn bản.

• **LSA** (Latent Semantic Analysis): là phương pháp tóm tắt dựa trên phân tích ma trận [23]. Cho một văn bản, thuật toán này tạo ra một ma trận giữa từ và câu, sau đó áp dụng phân tích ma trận để sinh ra các chủ đề ẩn và ánh xạ các câu vào các chủ đề. Các giá trị ánh xạ có thể dùng để đo độ quan trọng của các câu.

• **LexRank**: là thuật toán cơ bản cho tóm tắt trích chọn câu [24]. Ý tưởng chính của LexRank là biểu diễn một văn bản dưới dạng một đồ thị với các đỉnh là các câu và ước lượng độ quan trọng của các câu trên đồ thị đó. LexRank tạo một đồ thị nhẫu nhiên thống kê (stochastic graph) và tính độ quan trọng của các câu trên đồ thị đó. Các câu quan trọng sẽ được lựa chọn dựa vào điểm số.

• **TextRank**: thuật toán TextRank [5] cho tóm tắt văn bản kế thừa từ thuật toán PageRank [25]. Thuật toán giả sử rằng một câu là quan trọng nếu nó nhận được nhiều liên kết từ các câu khác. TextRank tạo ra một đồ thị các câu trung tâm và xếp hạng các câu này để tìm ra các câu quan trọng nhất.

• **Luhn**: là thuật toán dựa trên phương pháp kinh nghiệm cho tóm tắt văn bản [1]. Thuật toán giả định rằng các câu quan trọng chứa các từ thường xuyên xuất hiện trong văn bản.

• **KL-divergence**: Thuật toán giả sử rằng một câu nếu được đưa vào bản tóm tắt nếu nó giảm sự

[1] https://github.com/andersjo/pyrouge

phân kỳ tính bằng KL [26].

• **Sumbasic**: là thuật toán cơ bản cho tóm tắt văn bản [27]. Thuật toán giả sử rằng các từ có tần suất xuất hiện nhiều thì quan trọng và có xác suất cao nằm trong các câu tóm tắt.

• **SVR (Support Vector Regression)**: thuật toán được huấn luyện với dữ liệu trong tập huấn luyện. Một số đặc trưng gồm vị trí câu, độ dài câu, ... [28].

4.5. So sánh kết quả thực nghiệm

Chúng tôi so sánh kết quả mô hình đề xuất với các mô hình khác trên bộ dữ liệu. Với hai bài toán tóm tắt tóm lược và trích rut cho kết quả như sau:

Tóm tắt trích rút.

Chúng tôi so sánh mô hình BERTSUM với các mô hình tóm tắt tóm lược trên bộ dữ liệu chuẩn VNDS của tiếng Việt. Các kết quả trong bảng được lấy từ bài báo [19].

Bảng 2. Kết quả tóm tắt trích rút
(Chữ đậm thể hiện mô hình tốt nhất)

Mô hình	ROUGE-1	ROUGE-2	ROUGE-L
Lead-2	0.0586	0.0477	0.0580
LSA	0.5011	0.2045	0.3414
LexRank	0.4416	0.2006	0.3141
TextRank	0.4477	0.1904	0.2750
Luhn	0.4520	0.2026	0.3195
KL	0.5128	0.2007	0.3460
Sumbasic	*0.5265*	0.1913	0.2632
SVR	0.5041	**0.2367**	**0.3502**
CNN	0.4817	0.2193	0.3373
LSTM	0.4656	0.2029	0.3249
BERTSUM	**0.5349**	0.1829	**0.3502**

Kết quả từ Bảng 2 cho thấy mô hình BERTSUM cho kết quả khả quan cho tóm tắt trích chọn. BERTSUM đạt kết quả tốt nhất trên ROUGE-1 và ROUGE-L. Kết quả này là do: BERTSUM sử dụng sức mạnh của BERT để biểu diễn các câu đầu vào và BERTSUM mô hình mối quan hệ giữa các câu trong một văn bản. Hai lý do trên giúp mô hình nhận ra được câu quan trọng trong văn bản, để đưa ra dự đoán chính xác. Tuy nhiên, mô hình không đạt kết quả tốt nhất với ROUGE-2. Nguyên nhân có thể do các mô hình đạt kết quả tốt như SVR sử dụng đặc trưng được định nghĩa bởi con người phù hợp hơn trong việc mô hình hoá dữ liệu. Khi đánh giá các mô hình học có giám sát khác như SVR, CNN, và LSTM cho kết quả khá tốt. Trong đó, SVR tốt nhất với ROUGE-2 và cho kết quả bằng với BERTSUM với ROUGE-L. Điều này chứng tỏ các đặc trưng sử dụng cho SVR phù hợp với bộ dữ liệu. Một điều thú vụ là CNN và LSTM không đạt kết quả tốt nhất. Điều này khá dễ hiểu khi bài báo [19] sử dụng CNN và LSTM với kiến trúc đơn giản. Kết quả của hai mô hình này có thể được cải thiện khi sử dụng kiến trúc phức tạp hơn.

Một số mô hình tóm tắt không sử dụng dữ liệu huấn luyện cũng cho kết quả khả quan. Thuật toán Sumbasic tốt thứ 2 với ROUGE-1, thuật toán LSA và KL tốt thứ 2 với ROUGE-L. Điều này chứng tỏ có thể áp dụng các thuật toán học không giám sát cho tóm tắt văn bản trong trường hợp không có dữ liệu huấn luyện.

Tóm tắt tóm lược

Khi ứng dụng BERT cho tóm tắt tự động văn bản theo hướng tóm lược, kết quả thực nghiệm cho ta bảng so sánh kết quả của các thuật toán đối với tóm tắt tự động văn bản theo hướng tóm lược như Bảng 3.

Bảng 3. Kết quả tóm tắt tóm lược

Mô hình	ROUGE-1	ROUGE-2	ROUGE-L
fastAbs	0.5452	0.2301	**0.3764**
PointerSum	0.5437	0.2285	0.3751
Bottom-up	0.5011	0.2280	0.3574
BERT	**0.5478**	**0.2381**	0.3667

Bảng kết quả thực nghiệm cho thấy BERT cho kết quả khả quan với bài toán tóm tắt tóm lược cho tiếng Việt. Mô hình cho kết quả tốt nhất với ROUGE-1 và ROUGE-2. Điều này là do BERT mã hoá ngữ nghĩa của văn bản đầu vào để sinh ra bản tóm tắt phù hợp. Tuy nhiên, fastAbs đạt kết quả tốt nhất với ROUGE-L. Điều này cho thấy kết quả mô hình BERT có thể cần được cải thiện thêm, ví dụ như tinh chỉnh với các tham số phù hợp hơn, hoặc kết hợp với các kiến trúc khác như CNN hoặc LSTM.

5. Kết luận

Bài báo này giới thiệu kết quả hai bài toán tóm tắt trích rút và tóm tắt tóm lược sử dụng BERT. Kết quả thực nghiệm trên bộ dữ liệu tiếng Việt cho thấy ứng dụng BERT cho kết quả tốt đối với bài toán tóm tắt trích rút. Đối với bài toán tóm tắt tóm lược BERT cần tinh chỉnh để đạt kết quả tốt hơn.

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.2021.69.

Tài liệu tham khảo

[1]. Nguyễn Nhật An, *Nghiên cứu phát triển các kỹ thuật tự động tóm tắt văn bản tiếng Việt*, Luận án tiến sĩ, Viện Khoa học và Công nghệ Quân sự, tr. 8-23, 2015.

[2]. Đoàn Xuân Dũng, *Tóm tắt văn bản sử dụng các kỹ thuật trong Deep Learning*, Luận văn thạc sĩ, trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, tr. 1-8, 2018.

[3]. Nguyễn Viết Hạnh, Nghiên cứu tóm tắt văn bản tự động và ứng dụng, Luận văn thạc sĩ, trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, tr. 12-16, 2018.

[4]. Đỗ Thị Thu Trang, Trịnh Thị Nhị, Ngô Thanh Huyền, "Sử dụng BERT cho tóm tắt trích rút văn bản". *Tạp chí Khoa học và Công nghệ Trường Đại học Sư phạm Kỹ thuật Hưng Yên*, 2020, **Số 26/ tháng 6 năm 2020**, tr. 74-79.

[5]. Lâm Quang Tường, Phạm Thế Phi, và Đỗ Đức Hào, "Tóm tắt văn bản tiếng Việt tự động với mô hình SEQUENCE-TO-SEQUENCE". *Tạp chí Khoa học Trường Đại học Cần Thơ*, 2017, **Số chuyên đề: Công nghệ Thông tin (2017)**, tr. 125-132.

[6]. BOSER B., GUYON I., VAPNIK V., "A Training Algorithm for Optimal Margin Classifiers". *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (ACM)*, 1992, pp. 144-152.

[7]. BURGES C., "A Tutorial on Support Vector Machines for Pattern Recognition". *Proceedings of Int Conference on Data Mining and Knowledge Discovery,* 1998, **Vol 2, No 2**, pp 121-167.

[8]. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, "*Neural Machine Translation by Jointly Learning to Align and Translate*", 2014, CoRR, abs/1409.0473. Retrieved from https://arxiv.org/abs/1409.0473/

[9]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "*BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*", 2018. In arXiv:1810.04805v2 [cs.CL].

[10]. Keras, "*A Word2Vec Keras Tutorial*", 2017, Retrieved from https://adventuresinmachinelearning.com/word2vec-keras-tutorial/

[11]. Knight, Kevin, and Daniel Marcu, "Statistics-Based Summarization - Step one: Sentence compression". *17th National Conference of Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence (AAAI-2000),* 2000, pp. 703-710.

[12]. Lin, C.-Y. and Hovy, E. H., "Automatic Evaluation of Summaries Using N-gram co-occurrence Statistics", in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, **Volume 1 (NAACL-HLT)**, pp. 71-78.

[13]. M.-T. Nguyen, H.-D. Nguyen, T.-H.-N. Nguyen, and V.-H. Nguyen, "Towards state-of-the-art baselines for vietnamese multi-document summarization", in *10th International Conference on Knowledge and Systems Engineering (KSE)*, 2018, pp. 85-90.

[14]. Martins, A.F., Smith, N.A., "Summarization with a Joint Model for Sentence Extraction and Compression", in *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing, Association for Computational Linguistics*, 2009, pp. 1–9.

[15]. Minh-Tien Nguyen, *A Study on Web Document Summarization by Exploiting Its Social Context. Doctoral Dissertation*. School of Information Science Japan Advanced Institute of Science and Technology, 2018.

[16]. Nenkova A., McKeown K., "A Survey of Text Summarization Techniques", in *Mining Text Data*, Springer, 2012, pp. 43-76.

[17]. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. In Google AI Blog. URL: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html/

[18]. Rau, Lisa F, and Paul S Jacobs, "Creating Segmented Database from Free Text for Text Retrieval" SIGIR '91 *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM*, 1991, pp. 337-346.

[19]. Van - Hau Nguyen , Minh - Tien Nguyen, Thanh - Chinh Nguyen, Xuan - Hoai Nguyen, "VNDS: A Vietnamese Dataset for Summarization". In *IEEE Conference Proceedings (IEEE Conf Proc)*, 2019, pp. 375-380.

[20]. William B Dolan and Chris Brockett, "Automatically Constructing a Corpus of Sentential Paraphrases", in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[21]. Xin Rong, "Word2vec Parameter Learning Explained", 2016 - In *arXiv*, 1411.2738v4

[22]. A. Nenkova, "Automatic text summarization of newswire: Lessons Learned from the Document Understanding Conference," in *AAAI*, 2005, pp. 1436-1441.

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

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

[25]. Brin, S.; Page, L., "The anatomy of a large-scale hypertextual Web search engine" (PDF). *Computer Networks and ISDN Systems*, 1998, **30 (1–7)**, pp. 107–117.

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

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

[28]. C. Cortes and V. Vapnik, "Support-vector networks". *Machine Learning*, 1995, **20(3)**, pp. 73-297.

VIETNAMESE TEXT SUMMARIZATION BASE BERT METHODS

Abstract:

This paper introduces the method of text summarization in the two directions of extraction and summarization, using a pre-trained language model. To do this, for the extraction problem, we use the BERTSum model. The model uses BERT (Bidirectional Encoder Representations from Transformers) to encode input sentences and uses LSTM (Long Short Term Memory Networks) to represent relationships between sentences. For the summary problem, we use BERT to encode the semantics of the input text to generate a suitable summary. We tested the method on a Vietnamese dataset shared from VNDS (A Vietnamese Dataset for Summarization) [19] and evaluated the method by ROUGE (Recall - Oriented Understudy for Gisting Evaluation). Experimental results show that between the two problems of abstraction summarization and summarization, BERT is more effective in the problem of abstraction.

Keywords: *text summarization, NLP, machine learning, deep learning, unsupervised learning.*