Vietnamese text summarization using BARTpho

Văn Kim Ngân[1], Nguyễn Thị Bảo Hân[2], Tiêu Kim Hảo[3], TS Đỗ Trọng Hợp[4]

[1] Trường Đại học Công Nghệ Thông Tin
[2] Đại học Quốc gia Thành phố Hồ Chí Minh
{[1]19520177,[2]19520071,[3]19521480}@gm.uit.edu.vn, [4]hopdt@uit.edu.vn

Abstract. Trong bài báo này, chúng tôi sẽ trình bày về kỹ thuật Text Summarization sử dụng phương pháp BARTpho, và cụ thể ở đây là phiên bản BARTpho$_{word}$, đây là pre-trained model monolingual seq2seq quy mô lớn đầu tiên được dùng cho tiếng Việt. Text Summarization đã được nghiên cứu rộng rãi và áp dụng trong nhiều lĩnh vực khác nhau bằng ngôn ngữ tiếng Anh, nhưng đối với tiếng Việt vẫn còn ở giai đoạn đầu do một số tài liệu, hệ thống bị thiếu bộ dữ liệu chuẩn. Trong bài báo cáo này, chúng tôi xây dựng một mô hình tóm tắt văn bản trên bộ dữ liệu 'VNDS' bằng phương pháp BARTpho với độ đo được sử dụng là ROUGE.

1 Introduction

Trong những năm gần đây, text summarization - tóm tắt văn bản - đã được nghiên cứu và ngày càng phát triển để trở thành ưu thế trong lĩnh vực NLP. Summarization được áp dụng nhiều trong các ứng dụng NLP, từ các công cụ tìm kiếm (ví dụ: Google hoặc Bing) trả về mô tả ngắn gọn về các trang Web tương ứng với truy vấn tìm kiếm; các nhà cung cấp tin tức trực tuyến tạo ra các điểm nổi bật của một tài liệu Web trên giao diện của nó; bác sĩ hỗ trợ tìm kiếm nhanh chóng các thông tin liên quan từ một số tạp chí đã xuất bản, các trang và cổng thông tin y tế trên Web, bệnh án điện tử,… Tóm tắt văn bản chủ yếu được phân thành hai loại: extraction và abstraction. Extraction nhằm mục đích tạo ra một bản tóm tắt văn bản ngắn cho tài liệu bằng cách chọn các cụm từ hoặc các câu nổi bật từ văn bản gốc. Trong khi đó, đối với phương pháp abstraction sẽ tạo ra một bản tóm tắt cho một tài liệu từ đầu, thậm chí sử dụng các từ không xuất hiện trong văn bản gốc. Trong báo cáo này, chúng tôi sẽ trình bày mô hình tóm tắt văn bản sử dụng phương pháp extraction trên bộ dữ liệu VNDS, bộ dữ liệu này được dùng cho summarization task và mô hình được sử dụng ở đây là BARTpho.

BARTpho được chúng tôi sử dụng trong bài toán này vì tính hiệu quả trong việc đào tạo mô hình monolingual seq2seq với quy mô lớn đối với tiếng Việt. Trong báo cáo này, chúng tôi sẽ sử dụng BARTpho với phiên bản BARTpho$_{word}$, là pre-trained model monolingual seq2seq quy mô lớn đầu tiên được dùng cho tiếng Việt, dựa trên seq2seq denoising autoencoder BART, mang lại tính hiệu quả cao. Loại văn bản đầu vào của BARTpho$_{word}$ là word level. Việc phân đoạn từ tiếng Việt có ảnh hưởng tích cực đối với việc pre-training seq2seq.

2 Related Work

PhoBERT (Nguyen and Nguyen, 2020) là pre-trained model đơn ngôn ngữ quy mô lớn đầu tiên được dùng cho ngôn ngữ tiếng Việt, giúp đạt được state-of-the-art performances trên nhiều downstream Vietnamese NLP/NLU tasks (Truong et al., 2021; Nguyen and Nguyen, 2021; Dao et al., 2021; Thin et al., 2021). PhoBERT được pre-trained trên 20GB word-level corpus bằng văn bản tiếng Việt, sử dụng RoBERTa pre-training approach (Liu et al., 2019) sẽ giúp tối ưu hóa BERT nhằm đem lại hiệu suất mạnh mẽ hơn. Bên cạnh PhoBERT còn có các mô hình đơn ngôn ngữ sử dụng cho tiếng Việt được phổ biến rộng rãi như viBERT và vELECTRA (Bui et al., 2020), cả hai đều dựa trên BERT và ELECTRA pretraining approaches (Devlin et al., 2019; Clark et al., 2020) và pre-trained trên syllable-level Vietnamese text corpora. Nguyen et al. (2021) tiến hành một nghiên cứu thực nghiệm và cho thấy rằng PhoBERT cho ra kết quả hoạt động tốt hơn so với viBERT trong cùng một nhiệm vụ cơ bản là abstractive summarization bằng ngôn ngữ tiếng Việt.

BARTpho được tạo ra dựa trên BART. Ở đây chúng tôi sử dụng BART vì nó giúp tạo ra performances mạnh nhất trên downstream tasks so với các model pre-trained seq2seq khi sử dụng số lượng tương đương nhau của các tham số mô hình và kích thước của pre-training data (Lewis et al., 2020; Raffel et al., 2020; Qi et al., 2020). BART cũng được sử dụng để pre-train cho các mô hình đơn ngôn ngữ, sử dụng các ngôn ngữ khác như Pháp (Eddine et al., 2020) và Trung Quốc (Shao et al., 2021).

Cũng giống với BART, nhưng ở đây chúng tôi sẽ giới thiệu mBART – multilingual sequence-to-sequence denoising auto-encoder, được pretrain trên large-scale monolingual corpus với nhiều ngôn ngữ và sử dụng BART objective (Lewis et al., 2019). Văn bản đầu vào đã được 'noise' bằng 'masking phrases' và 'permuting sentences', sau đó sử dụng mô hình Transformer (Vaswani et al., 2017) đã được học để khôi phục lại các văn bản.

Vì BART sử dụng standard sequence-to-sequence Transformer architecture từ (Vaswani et al., 2017). Thành phần Transformer bao gồm encoder và decoder. Đầu vào của encoder đầu tiên sẽ đi qua một lớp seft-attention, đầu ra được truyền vào một mạng nơ ron truyền thẳng (feed-forward). Decoder cũng có hai thành phần đó nhưng nằm giữa chúng là một lớp attention giúp decoder tập trung vào phần quan trọng của câu đầu vào. Ý tưởng cốt lõi đằng sau mô hình Transformer đó chính là self-attention, một lớp giúp cho encoder nhìn vào các từ khác khi đang mã hóa một từ cụ thể. Transformer tạo các ngăn xếp là các lớp self-attention để xây dựng cho cả encoder và decoder thay vì các lớp RNNs hay CNNs. Kiến trúc chung này giúp cho Transformer model tính toán một cách song song, thay vì một chuỗi như RNNs.

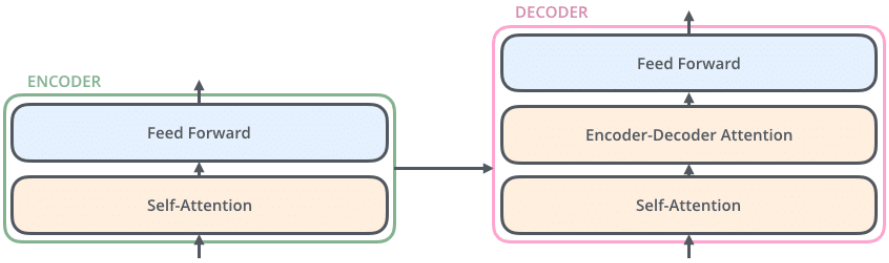

Fig. 1. Minh họa Transformer architecture.

3 Text Pre-processing

3.1 VNDS Datasets

Bộ dữ liệu chúng tôi sử dụng có tên là VNDS Nguyen et al. (2019). Bộ dữ liệu được thu thập từ tất cả các bài báo từ năm 2016 đến năm 2019 thuộc các danh mục "thế giới", "tin tức", "pháp luật" và "kinh doanh", chúng được lấy từ ba nguồn là: tuoi-tre.vn, vnexpress.net và nguoiduatin.vn, chúng tôi lựa chọn trang web trên vì các tin tức của các trang này luôn được cập nhật một cách nhanh nhất. Sau khi xác định được các danh mục, tác giả đã xây dựng một bộ phân loại đơn giản sử dụng TF-IDF và SVM để lọc ra từ các bài báo, trong đó nội dung là về bảng câu hỏi, tuyển sinh, bình luận phân tích và dự báo thời tiết. Họ chọn các chủ đề này vì trong tài liệu gốc đã có sẵn các nội dung trên. Tuy nhiên, thông tin này không quá quan trọng đối với việc tóm tắt tài liệu, vì vậy tập dữ liệu cuối cùng chỉ chứa các sự kiện tin tức đã được xử lý. Ngoài ra, các tài liệu thu thập được bao gồm cả các tài liệu ngắn, ví dụ bài báo chỉ chứa ba hoặc bốn câu. Để đảm bảo tính đa dạng của nội dung, họ chỉ giữ lại những tài liệu có ít nhất năm câu. Sau khi thu thập tài liệu, nhóm tác giả thực hiện thêm một bước nữa đó là xử lý dữ liệu. Họ đã sử dụng NLTK để phân đoạn câu và dùng vitk5 tool để phân đoạn từ. Cuối cùng, bộ dữ liệu thành ba phần theo tỷ lệ sau: 70% cho training, 15% cho development và 15% cho testing.

Table 1. Bảng thống kê của bộ dữ liệu.

	Train set	Val set	Test set
number of samples	105,418	22,642	22,644
#avg number of sentences in abstract	1.22	1.22	1.23
#avg number of words in abstract	28.48	28.54	28.59
#avg number of sentences in body	17.72	17.81	17.72
#avg number of words in body	418.37	419.66	418.74

3.2 Pre-training data

Để pre-training data, BARTpho$_{word}$ sử dụng PhoBERT pre-training corpus, chứa 20GB uncompressed text. Bộ dữ liệu này gồm hai nguồn chính: (i) đầu tiên là Vietnamese Wikipedia corpus (~1GB), và (ii) corpus thứ hai (~19GB) được tạo ra bằng cách xóa các bài viết tương tự hoặc trùng lặp từ 50GB Vietnamese news corpus. Ngoài ra, BARTpho$_{word}$ cũng sử dụng lại tokenizer của PhoBERT để áp dụng tập từ vựng có kích thước 64K loại subword và BPE (Sennrich et al., 2016) để phân đoạn các câu. BARTpho$_{word}$ có khoảng 420M parameters.

4 Our BARTpho

4.1 Architecture

BARTpho$_{word}$ sử dụng kiến trúc "lớn" với 12 lớp encoder và decoder cũng như sơ đồ pre-training của BART (Lewis et al., 2020). BART là một công cụ mã hóa tự động có nhiệm vụ ánh xạ giữa corrupted document so với tài liệu gốc. Nó được triển khai dưới dạng mô hình sequence-to-sequence với bidirectional encoder sử dụng corrupted text và left-to-right autoregressive decoder. BART sử dụng standard sequence-to-sequence Transformer architecture từ (Vaswani et al., 2017), ngoài ra, dựa theo GPT, thay vì dùng ReLU activation functions thì chúng tôi sẽ sử dụng GeLUs (Hendrycks & Gimpel, 2016) và khởi tạo các tham số từ N(0; 0,02). Dựa theo Liu et al. (2020), chúng tôi thêm một lớp layer-normalization phía trên của encoder và decoder.

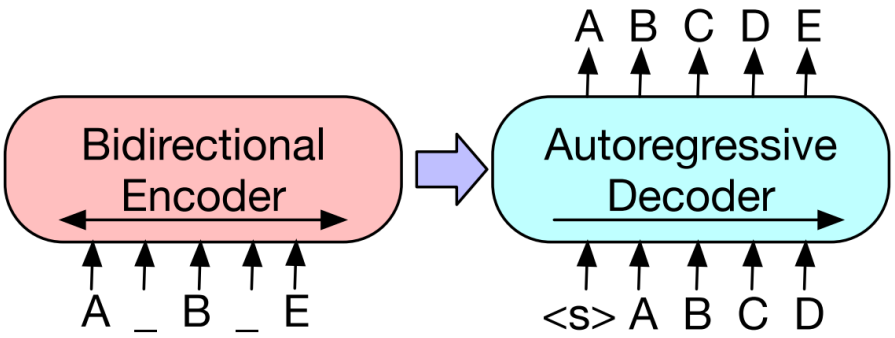

Fig. 2. BART: Inputs của encoder không cần phải căn chỉnh với outputs của decoder vì nó cho phép biến đổi nhiễu tùy ý (arbitary noise transformations). Ở đây, một document đã bị 'corrupted' do thay thế các khoảng văn bản bằng các ký hiệu <mask>. 'Corrupted' document (bên trái) được mã hóa bằng mô hình hai chiều (bidirectional model), (bên phải) là bộ giải mã tự động khôi phục (autoregressive decoder) có chức năng dự đoán văn bản, nhằm đối chiếu với tài liệu gốc. Với fine-tuning, uncorrupted document là đầu vào của encoder và decoder, từ đó chúng tôi sẽ sử dụng các biểu diễn từ kết quả đầu ra cuối cùng của decoder.

4.2 Pre-training BARTpho

Pre-training BART có hai giai đoạn: (1) corrupting văn bản đầu vào bằng hàm gây nhiễu tùy ý (arbitrary noising function) và (2) sequence-to-sequence model học cách xây dựng lại văn bản gốc. Khác với các denoising autoencoders đã có trước đây, BART cho phép áp dụng bất kỳ loại document corruption nào.

Dựa theo Lewis et al. (2020), chúng tôi sẽ sử dụng hai loại gây nhiễu trong noising function, bao gồm text infilling và sentence permutation.

Text Infilling. Chúng tôi lấy ngẫu nhiên mẫu thử một cụm văn bản với độ dài của chúng được rút ra từ phân phối Poisson (λ = 3.5), mỗi cụm được thay thế bằng duy nhất một <mask> token. Nếu độ dài của cụm bằng 0 thì tương đương với việc chèn các <mask> token vào. Text infilling được lấy cảm hứng từ SpanBERT (Joshi et al., 2019), nhưng độ dài của mẫu thử SpanBERT được tính bằng phân phối khác (clamped geometric) và thay thế các cụm lần lượt bằng <mask> token với chính xác độ dài trước đó của nó. Text infilling dạy cho mô hình dự đoán có bao nhiêu token bị thiếu trong một đoạn văn bản.

Sentence Permutation. Một tài liệu được chia thành các câu dựa trên các điểm dừng, có thể là dấu chấm câu hoặc kết thúc một đoạn văn và các câu này sẽ được xáo trộn theo một thứ tự ngẫu nhiên.

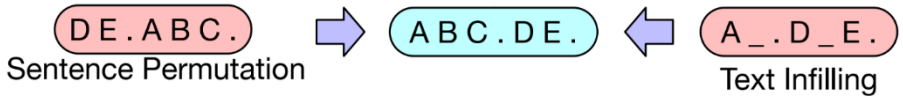

Fig. 3. Minh họa các biến đổi để làm nhiễu đầu vào, ở đây là Text Infilling và Sentence Permutation.

5 Experiments

5.1 Experimental setup

Chúng tôi sử dụng BARTpho để tiến hành pre-train large model với 12 lớp bên trong encoder và decoder, có kích thước bên trong là 1024. Văn bản được tokenized bằng byte-pair encoding giống như của GPT-2 (Radford et al., 2019). Như đã nói ở phần §4.2, ở đây chúng tôi sẽ kết hợp text infilling và sentence permutation, có 30% token trong mỗi document sẽ bị mask và tất cả các câu đều sẽ bị hoán vị một cách ngẫu nhiên. Chúng tôi sử dụng PhoBERT pre-training corpus, chứa 20GB uncompressed text.

Để tối ưu hóa, ở đây chúng tôi sẽ sử dụng Adam (Kingma and Ba, 2015) với các siêu tham số lần lượt là α = 1e-4, β_1 = 0.9, β_2 = 0.99, ε = 1e-5. Ngoài ra, nhóm sử dụng thêm Learning rate Finder, được đề xuất bởi nhà nghiên cứu (Lesie Smith, 2015). Ý tưởng của phương pháp là khởi tạo với learning rate rất rất nhỏ sao đó tính

loss trên 1 mini-batch, sau đó tăng learning rate lên (nhân 2..vv..) và tính loss cho mini-batch tiếp theo. Tiếp tục thực hiện cho đến khi loss bắt đầu tăng lên thay vì giảm xuống. Chọn learning rate thấp hơn điểm minimum (recommend minimum/10). Đối với mô hình BARTpho, chúng tôi đã chạy 3 training epoch trong vòng 10 tiếng.

5.2 Automatic Evaluation

Ở đây, nhóm chúng tôi sẽ sử dụng độ đo ROUGE (Recall-Oriented Understudy for Gisting Assessment) là độ đo đánh giá mức độ giống nhau giữa các câu đầu ra của mô hình so với câu reference.

ROUGE-N. ROUGE-N đo lường số lượng matching 'n-gram' giữa văn bản do mô hình của chúng tôi tạo ra so với câu tham chiếu (reference). Chỉ số này là thước đo dựa trên recall và sự so sánh của n-gram. Một n-gram chỉ đơn giản là một nhóm các tokens/words, Một unigram (1 gram) sẽ bao gồm một từ duy nhất. Một bigram (2 gam) bao gồm hai từ liên tiếp. Với ROUGE-N, N đại diện cho n-gram mà chúng ta đang sử dụng. Đối với ROUGE-1, chúng ta sẽ đo tỷ lệ matching của unigrams giữa đầu ra của mô hình chúng tôi với lại câu tham chiếu. tương tự với ROUGE-2 và ROUGE-3 sẽ sử dụng bigram và trigrams tương ứng.

$$\text{ROUGE-N} = \frac{\sum_{S\in\{ReferrenceSummaries\}} grams_n \in S \sum count_{match}(gram_n)}{\sum_{S\in\{ReferrenceSummaries\}} grams_n \in S \sum count(gram_n)}$$

Trong đó, n là độ dài n-grams, grams và $count_{match}$(grams) là số lượng tối đa của n-grams xuất hiện trong câu đầu ra của mô hình và câu tham chiếu.

ROUGE-L. ROUGE-L tính toán dãy các từ chung dài nhất (LCS) giữa câu của mô hình đầu ra với câu tham chiếu, nghĩa là chúng ta sẽ đếm chuỗi tokens dài nhất được chia sẻ giữa 2 câu. Ý tưởng ở đây là sequence được chia sẻ càng dài thì càng cho thấy sự giống nhau giữa hai sequences. ROUGE-L được đo bằng F-score, để tính được thì cần phải tính recall và precision.

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$
$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

Từ đó ROUGE-L được tính bằng

$$\text{ROUGE-L} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

Trong đó, LCS(X, Y) là chuỗi từ chung dài nhất giữa câu X(câu tham chiếu) và câu Y (câu đầu ra của mô hình), m là độ dài của câu tham chiếu, n là độ dài câu đầu ra mô hình, $\beta = \frac{P_{lcs}}{R_{lcs}}$ kiểm soát tầm quan trọng tương đối của precision và recall.

5.3 Experimental Results

Table 2. ROUGE scores (tính bằng %) được tính trên bộ dữ liệu VNDS. R-1, R-2 và R-L là viết tắt của ROUGE-1, ROUGE-2 và ROUGE-L. [*] là kết quả được nghiên cứu bởi Nguyen et al. (2022).

Model	Validation set			Test set		
	R-1	R-2	R-L	R-1	R-2	R-L
mBART [*]	60.06	28.69	38.85	60.03	28.51	38.74
$\text{BARTpho}_{\text{syllable}}$ [*]	60.29	29.07	39.02	60.41	29.20	39.22
$\text{BARTpho}_{\text{word}}$ [*]	60.55	29.89	39.73	60.51	29.65	39.75
$\text{BARTpho}_{\text{word}}$	**60.40**	**29.94**	**39.97**	**59.54**	**28.20**	**38.55**

Bảng 2 là bảng so sánh kết quả ROUGE scores trên validation và test set của mô hình mBART, $\text{BARTpho}_{\text{syllable}}$ và $\text{BARTpho}_{\text{word}}$. Vì chúng tôi chỉ huấn luyện mô hình $\text{BARTpho}_{\text{word}}$ nên ở đây ta sẽ so sánh kết quả nhóm đạt được so với kết quả của Nguyen et al. (2022). Đối với tập validation, R-1 vẫn còn thấp hơn so với kết quả được báo cáo của Nguyen et al. (2022). Nhưng ở R-2 và R-L thì lại cho kết quả tốt hơn. Còn ở tập test kết quả đạt được thấp hơn so với các mô hình khác. Tuy nhiên, đó chỉ là nhận xét khách quan, không cân xứng vì nhóm chỉ dùng 1300 sample cho tập valid và 1300 sample cho tập test, nên kết quả là không đáng đáng tin bằng khi dùng full valid và test set như của tác giả.

Ở đây chúng tôi sẽ phân tích kỹ hơn các yếu tố không cân xứng giữa nhóm và của tác giả Nguyen et al. (2022). Đầu tiên về quá trình huấn luyện, ở model của tác giả Nguyen et al. (2022), BARTpho model được huấn luyện trên 15 epoch trong vòng 6 ngày, trong khi đó, nhóm chúng tôi chỉ huấn luyện 3 epoch trong 10 giờ đồng hồ. Đối với bộ dữ liệu, các mô hình đều được train trên bộ dữ liệu VNDS Nguyen et al. (2019), tuy nhiên nhóm tác giả đã phát hiện ra một số bài báo bị trùng lặp trong bộ dữ liệu này và tiến hành lọc loại bỏ các bài trùng lặp đó. Về quá trình tối ưu hóa, nhóm tác giả, sử dụng Adam và fine-tuning trên 20 epoch, đồng thời còn sử dụng thêm grid search để chọn lựa Adam initial learning rate từ {1e-5, 2e-5, 3e-5, 5e-5}. Để có được ROUGE-L score tốt nhất, nhóm tác giả đã lựa chọn model checkpoint có ROUGE-L score cao nhất trên tập validation và sau đó áp dụng model checkpoint đã chọn được đó cho tập test.

Ở bảng 3, ta sẽ so sánh $\text{BARTpho}_{\text{word}}$ với các phương pháp extractive summarization khác. Có thể thấy rõ rằng, kết quả ROUGE score mà $\text{BARTpho}_{\text{word}}$ đạt được tốt hơn so với các phương pháp khác.

Từ bảng 2 và 3, ta có thể rút ra nhận xét rằng, kết quả nhóm đạt được khi so với kết quả của Nguyen et al. (2022) vẫn chưa tốt bằng vì các nguyên nhân đã được nêu trên, tuy nhiên khi nhìn vào tổng quát ở bảng 3, so với các phương pháp khác thì kết quả này đã đạt được khá tốt.

Table 3. ROUGE scores của các phương pháp extractive summarization. * là các kết quả lấy theo bản báo cáo của Nguyen et al. (2019).

Model	Original test set		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**
Lead-2*	5.86	4.77	5.80
LSA*	50.11	20.45	34.14
LexRank*	44.16	20.06	31.41
TextRank*	44.77	19.04	27.50
Luhn*	45.20	20.26	31.95
KL*	51.28	20.07	34.60
Sumbasic*	**52.65**	19.13	26.32
SVR*	50.41	**23.67**	**35.02**
CNN*	48.17	21.93	33.73
LSTM*	46.56	20.29	32.49
BARTpho$_{\text{word}}$	**59.54**	**28.20**	**38.55**

6 Conclusion

Qua bài báo cáo này, nhóm chúng tôi đã trình bày BARTphoword - pre-trained model monolingual seq2seq quy mô lớn đầu tiên được dùng cho tiếng Việt, là một công cụ mã hóa tự động có nhiệm vụ ánh xạ giữa corrupted document so với tài liệu gốc. Chúng tôi đã chứng minh tính hiệu quả của BARTpho bằng cách thể hiện rằng BARTpho hoạt động tốt hơn đối thủ cạnh tranh khác và giúp tạo ra SOTA performance trong việc text summarization task dành cho tiếng Việt. Tuy chưa đạt được kết quả cao như của nhóm nghiên cứu VinAI Research, nhưng nếu xét với các mô hình khác thì kết quả nhóm đạt được khá tốt. Nếu mô hình được huấn luyện trên máy có cấu hình tốt hơn và áp dụng các phương pháp tối ưu hóa khác thì sẽ đạt được kết quả tốt hơn nữa.

References

1. Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2021. Bartpho: Pre-trained sequence-to-sequence models for vietnamese. *arXiv preprint arXiv:2109.09701*.
2. Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of EMNLP*, pages 1037–1042.
3. Van-Hau Nguyen, Thanh-Chinh Nguyen, Minh-Tien Nguyen, and Nguyen Xuan Hoai. 2019. Vnds: A vietnamese dataset for summarization. In *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 375–380. IEEE
4. Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of NAACLHLT*, pages 2146–2153.
5. Linh The Nguyen and Dat Quoc Nguyen. 2021. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. In *Proceedings of NAACL: Demonstrations*, pages 1–7.

6. Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent Detection and Slot Filling for Vietnamese. In *Proceedings of InterSpeech*, pages 4698–4702.
7. Dang Van Thin, Lac Si Le, Vu Xuan Hoang, and Ngan Luu-Thuy Nguyen. 2021. Investigating Monolingual and Multilingual BERT Models for Vietnamese Aspect Category Detection. *arXiv preprint*, arXiv:2103.09519.
8. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*, arXiv:1907.11692.
9. The Viet Bui, Thi Oanh Tran, and Phuong Le-Hong. 2020. Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models. In *Proceedings of PACLIC*, pages 13–20.
10. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171– 4186.
11. Hieu Nguyen, Long Phan, James Anibal, Alec Peltekian, and Hieu Tran. 2021. VieSum: How Robust Are Transformer-based Models on Vietnamese Summarization? *arXiv preprint*, arXiv:2110.04257v1.
12. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of ACL*, pages 7871–7880.
13. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Textto-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.
14. Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting Future N-gram for Sequence-to-SequencePre-training. In *Findings of EMNLP*, pages 2401–2410.
15. Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. 2020. BARThez: a Skilled Pretrained French Sequence-to-Sequence Model. *arXiv preprint*, arXiv:2010.12321.
16. Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. CPT: A Pre-Trained Unbalanced Transformer for Both Chinese Language Understanding and Generation. *arXiv preprint*, arXiv:2109.05729.
17. Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL*, pages 1715–1725.
18. Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR*.