HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN ĐÌNH TUẤN

TÓM TẮT TIN TỨC TIẾNG VIỆT SỬ DỤNG MÔ HÌNH BERT

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH
MÃ SỐ: 8.48.01.01

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

HÀ NỘI-2024

Đề án được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TS. NGUYỄN MẠNH HÙNG**

Phản biện 1: PGS.TS. Nguyễn Long Giang

Phản biện 2: PGS.TS. Phan Xuân Hiếu

Đề án đã được bảo vệ trước Hội đồng chấm đề án thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông
Vào lúc: 08 giờ 30 ngày 20 tháng 03 năm 2024
Có thể tìm hiểu đề án tại:
Thư viện của Học viện Công nghệ Bưu chính Viễn Thông

MỞ ĐẦU

Trong kỷ nguyên kỹ thuật số hiện nay, sự phát triển nhanh chóng của các nguồn tin tức trực tuyến hay các trang báo khiến mọi người ngày càng gặp nhiều khó khăn trong việc cập nhật thông tin trước khối lượng thông tin có sẵn, và một lượng khổng lồ tin tức được tạo mới hằng ngày. Điều này nhấn mạnh tầm quan trọng đặc biệt của việc tóm tắt văn bản tự động, một lĩnh vực nhằm mục đích cô đọng các văn bản dài thành những bản tóm tắt ngắn gọn mà không làm mất đi bản chất, ý nghĩa của nội dung gốc, cũng cần đảm bảo được sự chính xác trong sử dụng câu từ và chính tả, ngữ pháp.

Sự ra đời của các mô hình nơ ron nhân tạo đã tạo ra các phương pháp mới cho việc tóm tắt văn bản tự động, Trong số này, đặc biệt là các mô hình sử dụng khung tuần tự (seq2seq), đã cho thấy thành công đáng chú ý. Các mô hình Seq2seq biến nhiệm vụ tóm tắt thành vấn đề ánh xạ một chuỗi văn bản đầu vào thành một chuỗi văn bản tóm tắt, tương tự như bài toán dịch ngôn ngữ, trong đó đầu vào và đầu ra sẽ là hai chuỗi ký tự khác nhau nhưng tương đồng về ngữ nghĩa, chỉ có điều khác là trong bài toán tóm tắt thì kết quả đầu ra sẽ ngắn gọn hơn đầu vào.

Tuy nhiên, việc áp dụng mô hình seq2seq trong việc tóm tắt tin tức tiếng Việt cũng gặp phải những thách thức đặc biệt, phần lớn là do đặc thù ngôn ngữ và ngữ nghĩa của Tiếng Việt.

Mô hình BERT (Bidirectional Encoder Representations from Transformers) là một phát triển đột phá trong xử lý ngôn ngữ tự nhiên, đã đặt ra các tiêu chuẩn mới về hiểu ngữ cảnh và ngữ nghĩa của văn bản.

Đề án này đề xuất một cách tiếp cận mới trong việc tóm tắt các bài báo Tiếng Việt bằng cách khai thác sức mạnh của BERT để giúp

cho mô hình học máy có thể hiểu rõ nội dung của các tin tức bằng Tiếng Việt.

Đề án gồm ba Chương:

● **Chương 1: Bài toán tóm tắt tin tức Tiếng Việt**

Trong chương này, đề án sẽ trình bày bài toán tóm tắt các bài báo Tiếng Việt, xem xét các giải pháp hiện có, thảo luận các phương pháp tóm tắt văn bản chung và đề xuất giải pháp sử dụng mô hình BERT trong bài toán tóm tắt Tiếng Việt.

● **Chương 2: Cơ sở lý thuyết của các mô hình sử dụng trong đề án**

Chương này sẽ cung cấp một cái nhìn tổng quan toàn diện về các mô hình làm nền tảng cho đề án này. Đề Án sẽ bắt đầu bằng việc khám phá kiến trúc mô hình biến đổi, kiến trúc này đóng vai trò là nền tảng cho mô hình BERT. Hiểu kiến trúc này là điều cần thiết để nắm bắt cách thức hoạt động của BERT, đặc biệt là cấu trúc chỉ dành cho bộ mã hóa. Sau đó, đề án sẽ tìm hiểu về kiến trúc của mô hình LSTM, là kiến trúc chính trong mô hình khung tuần tự seq2seq.

● **Chương 3: Giải pháp BERT-LSTM-LSTM và kết quả thực nghiệm, thảo luận**

Chương này đề án trình bày khung thử nghiệm cho giải pháp, từ thu thập dữ liệu các bài báo tin tức bằng Tiếng Việt, xử lý để đưa vào các mô hình đã trình bày ở Chương 2, đến đo lường đánh giá, thử nghiệm so sánh các biến thể của mô hình. Kết quả cho thấy tính hiệu quả của mô hình BERT-LSTM-LSTM trong việc tạo ra các bản tóm tắt chính xác và ngắn gọn về các bài báo tiếng Việt. Những phát hiện này không chỉ khẳng định phương pháp đề xuất mà còn mở ra hướng nghiên cứu trong tương lai về tóm tắt văn bản Tiếng Việt.

CHƯƠNG 1: BÀI TOÁN TÓM TẮT TIN TỨC TIẾNG VIỆT

1.1 Giới thiệu bài toán tóm tắt văn bản tiếng Việt

Tóm tắt văn bản nói chung được chia thành hai phương pháp chính: tóm tắt trích xuất (extractive summarization) và tóm tắt tóm lược (abstractive summarization) [5]. Tóm tắt trích xuất bao gồm việc chọn các cụm từ hoặc câu chính từ văn bản gốc và biên soạn chúng để tạo thành một bản tóm tắt. Ngược lại, tóm tắt tóm lược đòi hỏi phải hiểu các ý chính trong văn bản và tạo ra các câu mới với cùng ngữ nghĩa với văn bản gốc.

Bài toán tóm tắt văn bản [13]:

Cho văn bản nguồn $X = \{x_1, x_2, x_3, \ldots, x_l\}$.

l là độ dài của văn bản nguồn và x thuộc bộ từ vựng V_s.

Mục tiêu là tạo ra bản tóm tắt $Y' = \{y'_1, y'_2, y'_3, \ldots, y'_m\}$.

m là độ dài của bản tóm tắt y' thuộc bộ từ vựng V_t .

$m \ll l$ để đảm bảo bản tóm tắt sẽ ngắn hơn văn bản nguồn.

- Nếu $Y' \subseteq X$ bản tóm tắt được coi là dạng trích xuất, các thành phần của bản tóm tắt được lấy trực tiếp từ văn bản nguồn.
- Nếu $Y' \nsubseteq X$ bản tóm tắt là dạng tóm lược, có thành phần của bản tóm tắt không xuất hiện trong văn bản nguồn.

1.2 Các nghiên cứu liên quan

1.2.1 Thảo luận các nghiên cứu về tóm tắt văn bản trên thế giới

Cách tiếp cận dựa trên quy tắc trong những ngày khởi đầu

Bước đột phá đầu tiên trong lĩnh vực tóm tắt văn bản được đánh dấu bằng các hệ thống dựa trên quy tắc vào cuối thế kỷ 20. Một công trình quan trọng trong giai đoạn này là của Luhn (1958) [15].

Đột phá với học máy

Cuộc cách mạng học sâu

Sự xuất hiện của mã hoá ngữ cảnh (Contextual Embeddings) và các mô hình biến đổi (Transformer)

1.2.2 Thảo luận một số nghiên cứu về tóm tắt văn bản tại Việt Nam

Sử dụng các mô hình khung tuần tự

Trong nghiên cứu "Tóm tắt văn bản tiếng Việt tự động với mô hình Sequence-to-Sequence" của Lâm Quang Tường, Phạm Thế Phi và Đỗ Đức Hào, các nhà nghiên cứu đã sử dụng phương pháp học sâu để tự động hóa việc tóm tắt văn bản cho Tiếng Việt [5].

Các phương pháp tóm tắt văn bản tóm lược

Lê Thanh Hương và Lê Tiến Mạnh từ Đại học Bách khoa Hà Nội đề xuất một cách tiếp cận sáng tạo trong việc tóm tắt văn bản tóm lược [3].

Trích xuất khía cạnh bằng cách sử dụng mô hình BERT và các câu phụ trợ

Nguyễn Ngọc Điệp và Nguyễn Thị Thanh Thủy khám phá việc trích xuất khía cạnh trong văn bản tiếng Việt, một thành phần quan trọng của khai phá quan điểm theo khía cạnh [1].

Tóm tắt trích xuất sử dụng mô hình BERT

Bài viết của Đỗ Thị Thu Trang, Trịnh Thị Nhị và Ngô Thanh Huyền giới thiệu phương pháp trích xuất để tạo ra bản tóm tắt bằng mô hình BERT [6].

1.3 Kết luận chương

Chương này tạo tiền đề cho đề án bằng cách giới thiệu bài toán tóm tắt văn bản Tiếng Việt, trong phần 1.2, đề án đi sâu vào bối cảnh nghiên cứu của tóm tắt văn bản, cả trên toàn cầu và ở Việt Nam. Chương tiếp theo đề án sẽ trình bày khung cơ sở lý thuyết của các thành phần có trong giải pháp được đề xuất.

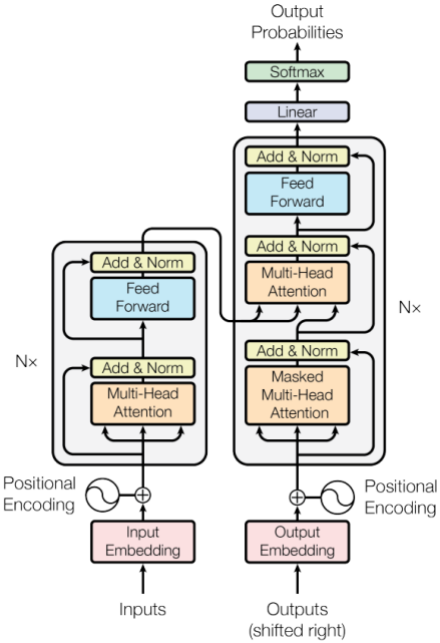
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT CỦA CÁC MÔ HÌNH SỬ DỤNG TRONG ĐỀ ÁN

2.1 Giới thiệu mô hình biến đổi (Transformer)

2.1.1 Nguồn gốc của mô hình biến đổi

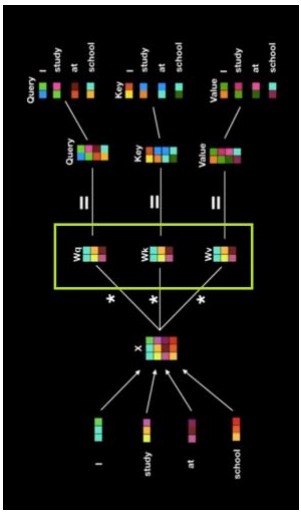
2.1.2 Kiến trúc của mô hình biến đổi: Cơ chế chú ý và mã hóa vị trí

Mô hình Transformer, kể từ khi ra đời, đã nổi bật nhờ kiến trúc độc đáo. Kiến trúc này gồm hai khái niệm cốt lõi: cơ chế chú ý và mã hóa vị trí.



Hình 2-1 Kiến trúc của mô hình Transformer [21]

Mô hình Transformer bao gồm một bộ mã hóa và một bộ giải mã, giống như cấu trúc của khung tuần tự seq2seq, mỗi lớp bao gồm nhiều lớp con thực hiện các hoạt động tự chú ý và mạng truyền thẳng.



Hình 2-2 Cơ chế tự chú ý [24]

Cơ chế tự chú ý

Khả năng tự chú ý sẽ tính toán mức độ liên quan có trọng số của tất cả các từ trong một chuỗi cho mỗi từ. Trong hình Wq, Wk và Wv là những ma trận tham số mà mô hình cần huấn luyện, để tìm ra mối liên kết giữa các từ trong câu [24].

Vectơ truy vấn, khoá và giá trị (Query, Key và Value) [24]: Mỗi từ được biểu thị bằng ba vectơ, vectơ truy vấn (Query), vectơ khóa (Key) và vectơ giá trị (Value), được tạo bằng cách nhân vectơ nhúng của chuỗi đầu vào với ma trận tham số Wq, Wk, Wv. Phương trình tính toán mức độ chú ý (Attention) như sau [24]:

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2.1)$$

Mã hóa vị trí

Do cơ chế tự chú ý vốn không xem xét thứ tự của các từ nên mã hóa vị trí (Positional Encoding) được thêm vào phần mã hoá của

từng từ để đưa thông tin về vị trí của các từ trong chuỗi [21].

2.2 Giới thiệu về Mô hình BERT

2.2.1 BERT: Một kiến trúc mới được xây dựng trên mô hình Biến Đổi (Transformer)

Mô hình Biểu diễn bộ mã hóa hai chiều từ mô hình Biến Đổi (BERT) được coi là một sự đột phá then chốt trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) [10].

Mối quan hệ của BERT với mô hình Biến Đổi

Xây dựng bối cảnh trên hai chiều: BERT chỉ tập trung vào đầu vào, áp dụng cách đào tạo hai chiều. Điều này cho phép BERT hiểu ngữ cảnh của một từ dựa trên tất cả môi trường xung quanh nó (bên trái và phải của từ) [10].

Cơ chế tự chú ý: Tính năng này rất quan trọng để hiểu được ý nghĩa sắc thái của các từ và cụm từ trong các ngữ cảnh khác nhau [10].

BERT cải tiến mô hình Transformer để phục vụ cho các tác vụ xử lý ngôn ngữ

Đào tạo trước và Tinh chỉnh (Pre-training and Fine-Tuning): BERT mở rộng khả năng của mô hình Biến Đổi thông qua phương pháp huấn luyện trước (Pre-training) và khả năng tinh chỉnh (Fine-Tuning) [10].

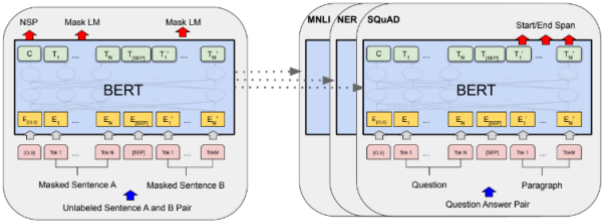
Những đổi mới của BERT: Mô hình BERT được huấn luyện bằng hai cách huấn luyện mới - Mô hình ngôn ngữ mặt nạ (Masked Language Model) và Dự đoán câu tiếp theo (Next Sentence Prediction) - trong giai đoạn tiền đào tạo.

2.2.2 Sự ra đời của BERT: Cách mạng hóa NLP

Mô hình BERT đã được các nhà nghiên cứu tại Google AI Language giới thiệu trong bài viết mang tính bước ngoặt của họ, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", xuất bản vào tháng 10 năm 2018.

2.2.3 Khám phá kiến trúc của BERT

Thành phần cốt lõi



Hình 2-3 Quá trình Pre-training và Fine-Tuning cho BERT [10]

Đào tạo hai chiều

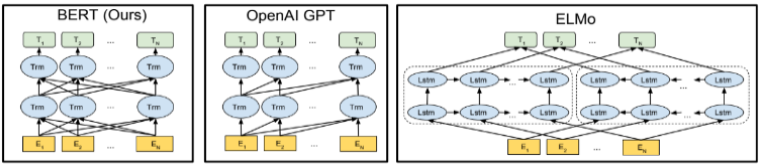
Mô hình ngôn ngữ mặt nạ (Masked Language Model) (Mask LM): Một trong những cách huấn luyện chính của BERT bao gồm việc che giấu ngẫu nhiên các từ trong dữ liệu đầu vào và sau đó dự đoán các từ bị che giấu này chỉ dựa trên ngữ cảnh của chúng [10].

Dự đoán câu tiếp theo (Next Sentence Prediction) (NSP): BERT còn được đào tạo bằng cách sử dụng một nhiệm vụ liên quan đến việc dự đoán liệu một cặp câu nhất định có liên hệ với nhau một cách tự nhiên hay không[10].

Kiến trúc cụ thể

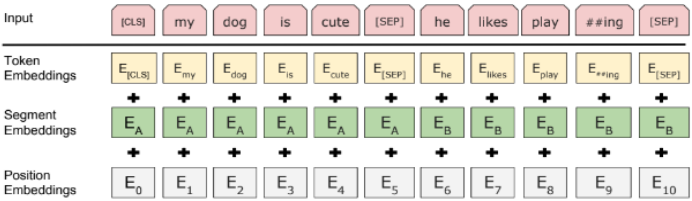
Các lớp và tham số của BERT: BERT có hai phiên bản chính: BERT-Base và BERT-Large. BERT-Base bao gồm 12 lớp (các khối

Transformer), 768 đơn vị ẩn (hidden units) và 12 lớp tự chú ý (self-attention heads), tổng cộng khoảng 110 triệu tham số. BERT-Large mở rộng hơn với 24 lớp, 1024 đơn vị ẩn và 16 lớp tự chú ý, tổng cộng khoảng 340 triệu tham số.



Hình 2-4 Kiến trúc của BERT so với OpenAI GPT và ElMo [10]

Phần nhúng (Embeddings): BERT sử dụng ba loại phần nhúng để thể hiện văn bản đầu vào: phần nhúng mã (token embeddings) (biểu diễn ở cấp độ từ), phần nhúng phân đoạn câu (segment embeddings) (phân biệt giữa các câu cho các nhiệm vụ liên quan đến cặp câu) và phần nhúng vị trí (position embeddings) (cho biết vị trí của các từ trong câu) [10].



Hình 2-4 Các phần nhúng đầu vào của BERT [10]

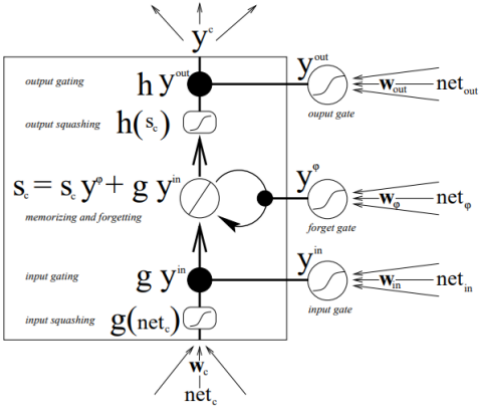
Kiến trúc của BERT đại diện cho một cột mốc quan trọng trong công nghệ NLP, kết hợp sức mạnh của đào tạo hai chiều, học sâu và việc sử dụng sáng tạo các cơ chế tự chú ý để đạt được mức độ hiểu ngôn ngữ chưa từng có.

2.3 Giới thiệu về mạng bộ nhớ dài ngắn hạn LSTM

2.3.1 Sự ra đời của mạng LSTM

2.3.2 Kiến trúc của LSTM

Phần này đi sâu vào các thành phần chính của kiến trúc LSTM, bao gồm đầu vào, đầu ra và ba cổng riêng biệt xử lý luồng thông tin: cổng đầu vào (input gate), cổng quên (forget gate) và cổng đầu ra (output gate) [19].



Hình 2-5 Kiến trúc Ô trạng thái của LSTM với ba cổng [19]

Thành phần chính

Ô trạng thái (Cell State).

Cổng đầu vào (Input Gate)

Cổng quên (Forget Gate)

Cổng đầu ra (Output Gate)

Chức năng của các cổng

Cổng đầu vào (Input Gate): Xác định thông tin mới sẽ được thêm vào ô trạng thái.

Cổng quên (Forget Gate): Quyết định thông tin nào bị loại bỏ

khỏi ô trạng thái.

Cổng đầu ra (Output Gate): Điều khiển đầu ra dựa trên ô trạng thái và đầu vào.

Bằng cách kết hợp các cơ chế để ghi nhớ và quên thông tin có chọn lọc, LSTM có thể duy trì thông tin liên quan của các chuỗi dài, khiến chúng trở nên lý tưởng cho nhiều ứng dụng trong xử lý ngôn ngữ tự nhiên, phân tích dữ liệu theo thời gian (time series) và hơn thế nữa.

2.3.3 Ứng dụng rộng rãi của mạng LSTM

2.4 Kết luận chương

Tóm lại, chương này đã cung cấp một cái nhìn tổng quan toàn diện về các mô hình làm nền tảng cho đề án này. Bắt đầu bằng việc khám phá kiến trúc mô hình Transformer, kiến trúc này đóng vai trò là nền tảng cho mô hình BERT. Hiểu kiến trúc này là điều cần thiết để nắm bắt cách thức hoạt động của BERT, đặc biệt là cấu trúc chỉ dành cho bộ mã hóa. Tiếp theo đề án tìm hiểu sâu về BERT, từ mối liên hệ với mô hình biến đổi, tới thiết kế các lớp trong BERT, đây sẽ là thành phần chính tạo nên sự khác biệt trong giải pháp cho bài toán tóm tắt văn tin tức Tiếng Việt. Sau đó, đề án đã tìm hiểu về kiến trúc của mô hình LSTM, là nền tảng chính cho bộ giải mã của giải pháp. Chương tiếp theo sẽ đi sâu vào việc triển khai và đánh giá phương pháp này, bao gồm cả cách tinh chỉnh từng tham số để giải quyết những thách thức cụ thể trong bài toán tóm tắt tin tức Việt Nam.

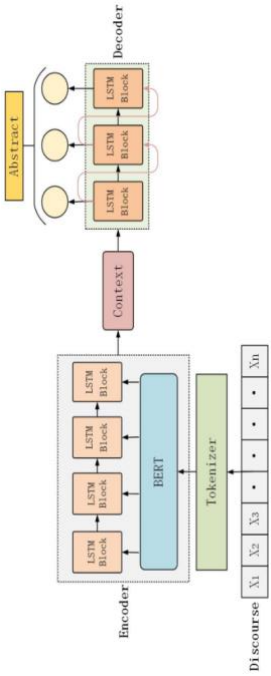
CHƯƠNG 3: GIẢI PHÁP BERT-LSTM-LSTM VỚI CƠ CHẾ TỰ CHÚ Ý VÀ KẾT QUẢ THỰC NGHIỆM, THẢO LUẬN

3.1 Giải pháp đề xuất của đề án

3.1.1 Sử dụng mô hình BERT trong bài toán tóm tắt tin tức Tiếng Việt với phương pháp tóm tắt tóm lược

Đề án đề xuất một mô hình mới, BERT-LSTM-LSTM với cơ chế chú ý (Attention) (gọi tắt là BLLA) [11], kết hợp các điểm mạnh của mô hình BERT để hiểu sâu về văn bản theo ngữ cảnh với khả năng học theo khung tuần tự (seq2seq) của LSTM (Long Short-Term Memory), và cơ chế chú ý (Attention).

Kiến trúc của mô hình BLLA (BERT-LSTM-LSTM)



Hình 3-1 Mô hình BLLA [11]

Mô hình có cấu trúc như sau:

BERT để tạo ra mã hoá theo ngữ cảnh: Mô hình bắt đầu với BERT để xử lý các bài báo tiếng Việt đầu vào, tạo ra các mã hoá cho từng từ trong văn bản theo ngữ cảnh nhằm nắm bắt các sắc thái ngữ nghĩa của văn bản, mỗi từ sẽ được mã hoá là một vectơ 768 chiều, không gian vectơ lớn này cho phép mô hình nắm bắt được sự hiểu biết phong phú về ngữ nghĩa và cú pháp của ngôn ngữ, bao gồm cả việc xử lý độ phức tạp về âm điệu và hình thái của tiếng Việt, vì mỗi chiều có thể biểu thị các đặc điểm khác nhau của ngôn ngữ.

Lớp LSTM đầu tiên trong khung tuần tự (Bộ mã hóa): Sau đó, kết quả mã hoá của từng từ theo ngữ cảnh từ BERT sẽ được đưa vào lớp LSTM đầu tiên, đóng vai trò là bộ mã hóa trong khung tuần tự (seq2seq).

Lớp LSTM thứ hai (Bộ giải mã): Biểu diễn văn bản đã được nén được chuyển đến lớp LSTM thứ hai, đóng vai trò là bộ giải mã. Lớp LSTM thứ hai sẽ được khởi tạo trạng thái từ trạng thái cuối của lớp LSTM đầu tiên, điều này đảm bảo bộ giải mã bắt đầu nhiệm vụ của mình với toàn bộ hiểu biết có được từ bộ mã hoá cho văn bản đầu vào. Trong quá trình giải mã, đối với mỗi bước, chuỗi hiện tại của bản tóm tắt được tạo ra cho đến nay sẽ được cung cấp làm đầu vào để dự đoán từ tiếp theo xuất hiện trong bản tóm tắt.

Sự tăng cường của lớp chú ý (Attention): Sau khi xử lý bởi lớp LSTM thứ hai (bộ giải mã), một cơ chế chú ý được áp dụng để tinh chỉnh việc dự đoán từ tiếp theo cho bản tóm tắt. Lớp này hướng sự tập trung của mô hình đến các phần có liên quan của văn bản, nâng cao hiệu quả dự đoán của bộ giải mã. Lớp chú ý đảm bảo rằng

các từ trong bản tóm tắt được tạo vừa phong phú về ngữ cảnh vừa tập trung vào thông tin chính, có sự liên kết chặt chẽ với văn bản nguồn.

3.1.2 Ưu điểm của giải pháp so với các phương pháp hiện có

Tăng cường hiểu biết theo ngữ cảnh

Xử lý đặc điểm ngôn ngữ của Tiếng Việt

Giảm lượng dữ liệu cần đào tạo để hiểu ngôn ngữ Tiếng Việt

Việc kết hợp BERT vào khung tuần tự để giải quyết bài toán tóm tắt tin tức Tiếng Việt, tận dụng kiến thức của BERT có được trong việc đào tạo trước để hiểu và xử lý các mẫu ngôn ngữ phức tạp, cải thiện đáng kể tính chính xác, nhưng giảm đáng kể lượng dữ liệu để huấn luyện mô hình.

3.1.3 Các công cụ và nền tảng sẽ được sử dụng để triển khai và thử nghiệm

Ngôn ngữ lập trình và thư viện học máy

Python & Keras

Mô hình được đào tạo trước cho Tiếng Việt

PhoW2V

PhoBERT

Mô hình đa ngôn ngữ BERT (BERT Multilingual Base Model)

Bộ công cụ Xử Lý Ngôn Ngữ Tự Nhiên cho Tiếng Việt

VnCoreNLP

Môi trường phát triển, thực thi và thư viện

Google Colab, TensorFlow, Transformers

3.2 Thiết kế giải pháp BERT là bộ mã hóa và LSTM là bộ giải mã trong mô hình BLLA

3.2.1 Khai thác thông tin chi tiết theo ngữ cảnh của BERT và bước huấn luyện đầu tiên

Đề án sử dụng BERT trong kiến trúc của mô hình BLLA (BERT-LSTM-LSTM với cơ chế chú ý).

Mã hoá đầu ra 768 chiều: Đầu ra của BERT bao gồm các vectơ có 768 chiều cho mỗi từ trong chuỗi đầu vào [10]. Các phần mã hoá này là công cụ giúp nắm bắt các sắc thái cần thiết để tóm tắt chính xác, cung cấp đầu vào phong phú cho các thành phần khung tuần tự (seq2seq) để học hỏi và xử lý.

Tích hợp BERT với khung tuần tự (Seq2Seq)

Quá trình chuyển đổi từ bộ mã hóa của BERT sang kiến trúc seq2seq diễn ra liền mạch, với các vectơ 768 chiều đóng vai trò là đầu vào cho lớp mã hóa LSTM hai chiều (Bidirectional LSTM) (BiLSTM).

Chiến lược đào tạo và các siêu tham số

Với vai trò quan trọng của BERT, chiến lược đào tạo giai đoạn một được thiết kế cẩn thận để tối ưu hóa các thành phần khung tuần tự:

Giữ lại kiến thức được đào tạo trước của BERT: Giai đoạn đào tạo ban đầu, vì sự khác biệt giữa hai phần của mô hình (một mô hình lớn được đào tạo trước và khung tuần tự chưa được đào tạo) đề án sẽ cố định BERT (frozen layers) để đảm bảo tận dụng được các kiến thức mà mô hình có được trong giai đoạn đào tạo trước [14].

Trình tối ưu (optimizers) và hàm mất mát (loss function): **Adam** được sử dụng là trình tối ưu cho BLLA cùng với hàm mất mát là **sparse categorical crossentropy**. Trọng tâm là tối ưu hóa các lớp khung tuần tự seq2seq để hoạt động liền mạch với BERT đã

được đào tạo trước. Tốc độ học tập trong giai đoạn này là **0,001**.

3.2.2 Thiết kế LSTM làm bộ giải mã và quá trình huấn luyện bước hai

Cấu hình bộ giải mã LSTM

Bộ giải mã LSTM có 256 đơn vị hay 256 blocks cho các thành phần LSTM. Sử dụng kĩ thuật bỏ qua các đơn vị (droput) với tỉ lệ là 40% tại các lớp LSTM.

Bộ mã hóa LSTM hai chiều (Bidirectional LSTM Encoder): Phần đầu tiên của khung tuần tự là lớp LSTM hai chiều (BiLSTM), nâng cao khả năng của mô hình trong việc nắm bắt ngữ nghĩa từ cả hai hướng của chuỗi đầu vào. Lớp LSTM hai chiều này sẽ học từ dữ liệu đầu vào để tạo ra bản nén lại thông tin quan trọng của văn bản để sử dụng cho bộ giải mã.

Bộ giải mã LSTM: Theo sau bộ mã hóa BiLSTM, bộ giải mã bao gồm một lớp LSTM, cũng có 256 đơn vị, có nhiệm vụ tạo chuỗi của bản tóm tắt. Lớp này nhận các phần nén thông tin văn bản đầu vào từ lớp BiLSTM, sau đó dùng thông tin đó, cùng với thông tin bản tóm tắt thời điểm hiện tại để đưa ra dự đoán cho từ tiếp theo trong chuỗi tóm tắt. Quá trình này sẽ lặp lại cho đến khi ta có một bản tóm tắt đầy đủ.

Tích hợp cơ chế chú ý

Lớp đa chú ý (Multi-Head Attention): Đề án sử dụng lớp đa chú ý, đề án sử dụng ba lớp chú ý, mô hình sẽ chú ý vào các phần khác nhau của đầu vào đã được mã hóa khi tạo từng từ trong bản tóm tắt.

Cách lớp chú ý hoạt động

Chiến thuật huấn luyện bước hai

Với sự hiểu biết nền tảng đã được thiết lập ở giai đoạn một thông qua việc cố định BERT, giai đoạn hai của quá trình huấn luyện mô hình BLLA bắt đầu với một quy trình tinh chỉnh toàn diện cho mô

hình ở tất cả các phần. Giai đoạn này đề án sẽ giải phóng các lớp BERT (unfreeze layers) [14], cho phép toàn bộ mô hình, bao gồm cả BERT và bộ giải mã LSTM, được huấn luyện đồng thời.

Giải phóng BERT để tối ưu hóa

Điều chỉnh để huấn luyện BERT: Trong giai đoạn này, thuộc tính có thể huấn luyện của mô hình BERT ở tất cả các lớp được đặt thành **True**, biểu thị rằng trọng số của bộ BERT sẽ được tham gia vào quá trình học tập.

Điều chỉnh tốc độ học tập: Tốc độ học tập sẽ giảm xuống là **2e-5** để phù hợp cho việc tinh chỉnh BERT, do BERT có rất nhiều tham số đã được học từ trước, điều này giúp ngăn chặn những nhiễu loạn đáng kể trong các phần của BERT.

3.3 Đánh giá và thảo luận mô hình BLLA

3.3.1 Xây dựng bộ dữ liệu tin tức Tiếng Việt cho bài toán tóm tắt

Nguồn dữ liệu

Bộ dữ liệu được lấy từ ba bộ dữ liệu chính, mỗi bộ đóng góp các khía cạnh riêng biệt cho kho dữ liệu tổng thể:

VNDS (Bộ dữ liệu tóm tắt tin tức Tiếng Việt) [2]: Được tạo ra bởi Nguyễn Văn Hậu và các cộng sự. Với 105,418 dữ liệu tin tức cho việc đào tạo.

Bộ dữ liệu tin tức Tiếng Việt từ Báo Lao Động [28]: Được tạo ra bởi PHẠM ĐỨC và lưu trữ trên Kaggle, bộ dữ liệu này bao gồm 290,282 bài báo, được thu thập trực tiếp từ Báo Lao Động vào ngày 19 tháng 5 năm 2022.

Bộ dữ liệu Tin tức Trực tuyến Tiếng Việt [29]: Được tạo ra bởi HAITRANQUANG, cũng có trên Kaggle, bộ dữ liệu này làm phong phú thêm kho dữ liệu với 171,135 bài viết từ tháng 7 năm 2022, lấy từ 25 trang tin tức trực tuyến nổi tiếng ở Việt Nam.

Các bước tiền xử lý chi tiết

Xử lý các tin tức lớn thành các cặp dữ liệu Tiêu Đề - Nội Dung:

Phần nội dung của tin tức sẽ được tách thành các câu, đề án sử dụng Tiêu Đề của tin tức để làm bản tóm tắt ngắn gọn cho phần nội dung. Ngoài ra để giảm hơn nữa những phần nội dung thừa đề án sử dụng một thuật toán sử dụng chính BERT để tìm Cosine Similarity giữa các câu trong nội dung của tin tức và Tiêu Đề, từ đó tìm ra các câu có sự liên quan nhất với Tiêu Đề [7].

Công thức của cosine similarity

Cosine Similarity đo cosine của góc giữa hai vectơ khác 0 trong không gian đa chiều, cung cấp độ đo về sự tương tự về hướng của chúng. Công thức được sử dụng là [7]:

$$\text{Cosine Similarity} = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3.1)$$

Lựa chọn câu để đưa vào phần nội dung trong cặp Tiêu Đề - Nội Dung: Dựa trên điểm số có được, 3 câu có điểm cao nhất sẽ được chọn. Những câu này, được cho là chứa thông tin quan trọng nhất liên quan đến tiêu đề, sau đó được kết hợp với tiêu đề để tạo thành một cặp Tiêu Đề - Nội Dung hay Bản Tóm Tắt - Nội Dung, dùng làm đầu vào cho quá trình đào tạo mô hình BLLA.

Làm sạch và chuẩn hóa dữ liệu:

Các cặp dữ liệu sẽ được làm sạch và chuẩn hóa, bao gồm loại bỏ các đề cập và tên không liên quan, và chuẩn hóa Unicode. Ngoài ra đề án cũng xóa ký tự đặc biệt như các dấu phẩy, ngoặc đơn v.v và xoá các chữ số để giúp mô hình tập trung hơn vào từ, tránh nhiễu trong quá trình huấn luyện.

Chuyển từ thành mã (tokenizer):

Bộ chuyển đổi BERT tokenizer được sử dụng để chuyển văn bản thành các mã đại diện. Đối với các bản tóm tắt, bộ mã hoá của TensorFlow đã được sử dụng để chuyển đổi văn bản. Sau đó dữ liệu sẽ được xử lý bằng padding và truncating để thống nhật độ dài của Bản Tóm Tắt và Nội Dung. Bản Tóm Tắt sẽ có độ dài tối đa 20 tokens và Nội Dung sẽ là 200 tokens.

Phân chia dữ liệu thành tập huấn luyện và tập kiểm tra:

Tập dữ liệu được chia thành các tập huấn luyện và kiểm tra với tỉ lệ 90% cho tập huấn luyện và 10% cho tập kiểm tra.

3.3.2 Cài đặt thử nghiệm

Cài đặt mô hình

Mô hình BLLA được đánh giá với ba phiên bản riêng biệt để xác định cấu hình hiệu quả nhất cho bài toán tóm tắt tin tức Tiếng Việt:

Phiên bản sử dụng BERT-base-multilingual-cased: Phiên bản này sử dụng mô hình bert-base-multilingual-cased làm bộ mã hóa, tích hợp nó với bộ giải mã LSTM. Bộ giải mã LSTM được định cấu hình với 256 đơn vị và phần mã hoá cho bộ giải mã được khởi tạo từ PhoW2V tại phiên bản 100 chiều.

Phiên bản sử dụng PhoBERT: Có cấu trúc tương tự phiên bản đầu tiên, phiên bản này thay thế PhoBERT làm bộ mã hóa, vẫn cấu hình bộ giải mã LSTM với 256 đơn vị và phần mã hoá lấy dữ liệu từ PhoW2V, 100 chiều.

Phiên bản chỉ sử dụng LSTM-LSTM trong khung tuần tự Seq2Seq: BLLA cũng được so sánh với mô hình seq2seq truyền thống sử dụng bộ mã hóa LSTM và bộ giải mã LSTM, cả hai đều được cài đặt 128 đơn vị. Phần mã hoá cho cả bộ mã hóa và bộ giải mã được khởi tạo từ PhoW2V, 100 chiều.

Dữ liệu sử dụng:

Các mô hình được huấn luyện trên bộ dữ liệu gồm các cặp Tiêu Đề - Nội Dung, và được tiền xử lý như đã mô tả ở mục 3.2.1. Bộ dữ liệu được sử dụng để huấn luyện gồm 120,000 mẫu, chia thành hai bộ huấn luyện (train) và kiểm tra (validation) với tỉ lệ là 90-10. Như vậy các mô hình sẽ huấn luyện trên bộ gồm 108,000 mẫu và kiểm tra tại mỗi bước với 12,000 mẫu.

Các mô hình đều trải qua quá trình đào tạo hai giai đoạn:

Giai đoạn 1: Tiến hành trong **10 epochs** với tốc độ học tập là **0,001**, các lớp BERT được đóng băng để ưu tiên học ở phía bộ giải mã LSTM và các lớp mã hoá.

Giai đoạn 2: Cũng kéo dài **10 epochs** nhưng với tốc độ học giảm xuống còn **2e-5**, giai đoạn này sẽ mở cho các lớp BERT được học tập để tinh chỉnh toàn diện trên toàn mô hình. Cơ chế Early Stopping đã được triển khai để tạm dừng quá trình đào tạo nếu giá trị validation loss bắt đầu tăng lên, ngăn chặn vấn đề quá khớp (overfitting) để đảm bảo tính tổng quát của mô hình.

Tài nguyên tính toán

Các mô hình trong đề án được huấn luyện trên TPU của Google Colab với 35 GB RAM.

3.3.3 Biện pháp đánh giá

Để đánh giá hiệu suất của mô hình BLLA và các biến thể của nó, đề án sử dụng điểm BLEU (Bilingual Evaluation Understudy) làm thước đo đánh giá [20].

Triển khai BLEU trong đánh giá mô hình

Việc triển khai sử dụng điểm BLEU để đánh giá mô hình BLLA bao gồm các bước sau:

Chuẩn bị bản tóm tắt tham chiều: Bộ dữ liệu cho kiểm tra

điểm số (khác với bộ dữ liệu dùng để kiểm tra trong quá trình đào tạo), được lấy từ tập kiểm tra của bộ VNDS (khoảng 20,000 dữ liệu).

Tạo bản tóm tắt: Dữ liệu nội dung của bộ dữ liệu kiểm tra được mô hình BLLA và các biến thể của nó xử lý để sinh ra bản tóm tắt.

Tính toán điểm BLEU: Sử dụng thư viện xử lý ngôn ngữ tự nhiên NLTK, điểm BLEU được tính cho từng bản tóm tắt do mô hình tạo dựa trên các văn bản tham chiếu tương ứng. Điểm BLEU trung bình trên tất cả các mẫu thử nghiệm cung cấp thước đo toàn diện về chất lượng tóm tắt của từng mô hình.

Phân tích so sánh: Bằng cách so sánh điểm BLEU của các mô hình khác nhau ta sẽ có cái nhìn tổng quát về hiệu quả của từng mô hình.

3.3.4 Phân tích so sánh hiệu suất dựa trên điểm BLEU và độ phức tạp của các mô hình

Để đánh giá hiệu quả của các mô hình BLLA trong bài toán tóm tắt tin tức Tiếng Việt, đề án sẽ xem xét hai biến thể chính là: phoBERT và BERT. Mô hình phoBERT là một mô hình đơn ngữ được đào tạo trước cụ thể cho Tiếng Việt, trong khi BERT (bert-base-multilingual-cased) là mô hình đa ngôn ngữ bao gồm tiếng Việt trong kho dữ liệu được đào tạo của nó.

Dữ liệu đánh giá

Đề án sử dụng 200 mẫu không nằm trong bộ huấn luyện và bộ kiểm tra để làm dữ liệu đánh giá khả năng sinh những bản tóm tắt của các mô hình.

Độ phức tạp của các mô hình

Để bổ sung cho phân tích hiệu suất của các mô hình BLLA trong bài toán tóm tắt, đề án sẽ xem xét độ phức tạp trong kiến trúc của mỗi mô hình, như được trình bày chi tiết trong bảng bên dưới.

Bảng 3-1 Tham số của các mô hình

Tên các lớp (Layer Name)	**BLLA - phoBERT**	**BLLA - BERT base**	**LSTM - LSTM**
Mô hình mã hoá			
phoBERT / BERT / Encoder Embedding	134,998,272	177,853,440	14,555,600
Decoder Embedding	3,495,900	3,495,900	3,495,900
Lớp hồi quy			
Bidirectional LSTM	918,528	918,528	234,496
Decoder LSTM	117,248	117,248	117,248
Lớp chú ý (Attention)			
Multi Head Attention	98,816	98,816	98,816
Lớp chuẩn hóa và các lớp bổ sung			
Lớp chuẩn hoá	256	256	256
Lớp Dense (Fully-connected layer)	4,509,711	4,509,711	4,509,711
Tổng tham số	144,138,731	186,993,899	23,012,027

Số liệu đánh giá

Các mô hình được đánh giá bằng cách sử dụng điểm BLEU. Đề án sử dụng bốn loại điểm BLEU là: BLEU-1, BLEU-2, BLEU-3 và BLEU-4 (tương ứng với unigram tới 4-gram) để cung cấp cái nhìn rõ ràng về khả năng của các mô hình. Trong phần này đề án sử dụng dạng BLEU theo phần trăm để so sánh.

Hiệu suất của từng mô hình được trình bày chi tiết trong bảng sau:

Bảng 3-2 Hiệu suất các mô hình

Model	BLEU-1 Score	BLEU-2 Score	BLEU-3 Score	BLEU-4 Score
BLLA - phoBERT	68,08	58,53	50,06	41,89
BLLA - BERT base	48,7	41,26	35,28	30,30
LSTM-LSTM	23,05	16,49	11,38	7,74

Từ kết quả này, điểm BLEU chỉ ra rằng mô hình BLLA sử dụng phoBERT vượt trội hơn mô hình sử dụng BERT base trên tất cả các cấp độ n-gram. Điều này cho thấy rằng việc đào tạo trước và tập trung vào ngôn ngữ Tiếng Việt của phoBERT, mang lại hiệu quả đáng kể trong việc nắm bắt các sắc thái, ngữ nghĩa của tiếng Việt.

Mô hình BLLA sử dụng BERT base cho các điểm số cao hơn đáng kể so với mô hình khung tuần tự LSTM truyền thống. Điều này củng cố giá trị của các mô hình ngôn ngữ được đào tạo trước trong việc nâng cao khả năng nắm bắt theo ngữ cảnh và cho ra kết quả tốt hơn cho các tác vụ xử lý ngôn ngữ tự nhiên.

Một số ví dụ kết quả được sinh từ mô hình BLLA - phoBERT trên bộ dữ liệu đánh giá.

Nội dung:

“Trước đó , như Tuổi_Trẻ đã thông_tin , em Hoàng_Long_Nhật , học_sinh lớp 6.2 Trường THCS Duy_Ninh , đã phải nhập_viện điều_trị bốn ngày vì bị cô_giáo bắt cả lớp tát 230 cái . Nguyên_nhân em Nhật bị tát là vì một bạn ngồi cạnh " tố " với cô_giáo là em nói_tục . Cô_giáo Thuỷ , người vừa bị khởi_tố về tội hành_hạ người khác”

Bản tóm tắt do người tạo:

“Khởi tố cô giáo chỉ đạo cả lớp tát học sinh 231 cái”

Bản tóm tắt do mô hình sinh ra:

“Khởi tố cô giáo bắt cả lớp tát học sinh 231 cái”

Qua các phân tích so sánh này cho thấy việc sử dụng các mô hình ngôn ngữ được đào tạo trước như BERT và phoBERT, kết hợp với các lớp LSTM và cơ chế chú ý, giúp tăng đáng kể chất lượng kết quả trong bài toán tóm tắt tin tức Tiếng Việt. Đặc biệt, mô hình BLLA - phoBERT nổi lên nhờ sự cân bằng giữa độ phức tạp của mô hình và hiệu quả hoạt động, giúp nó trở thành một cách tiếp cận đầy hứa hẹn cho nhiệm vụ này.

3.4 Kết luận chương

Kết quả cho thấy tính hiệu quả của mô hình BLLA hay BERT-LSMT-LSTM với cơ chế Chý ý, trong việc tạo ra các bản tóm tắt chính xác và ngắn gọn cho các bài báo tiếng Việt.

KẾT LUẬN

Bài toán tóm tắt văn bản nói chung và tóm tắt tin tức Tiếng Việt nói riêng, là một lĩnh vực nghiên cứu sôi động trong xử lý ngôn ngữ tự nhiên, với rất nhiều ứng dụng trong thực tế. Đề án này nhằm nghiên cứu để sử dụng những khả năng của BERT trong bài toán tóm tắt các bài báo Tiếng Việt theo phương pháp tóm lược. Thông qua việc ứng dụng các kỹ thuật NLP, đề án đã đạt được một số kết quả trong lĩnh vực tóm tắt văn bản tự động, ngoài ra cũng đối mặt với những thách thức vốn có và mở ra con đường nghiên cứu trong tương lai.

Một trong những đạt được của đề án là triển khai thành công mô hình BLLA kết hợp BERT với khung tuần tự LSTM. Mô hình đã được thiết kế và tinh chỉnh để tạo ra những bản tóm tắt ngắn gọn và giàu thông tin cho các bài tin tức Tiếng Việt.

Một khía cạnh quan trọng của đề án này là việc tạo ra một bộ dữ liệu đáng kể bao gồm 500,000 mẫu tin tức tiếng Việt và phần tóm tắt của chúng. Bộ dữ liệu này không chỉ tạo điều kiện thuận lợi cho việc đào tạo và tinh chỉnh mô hình BERT-LSTM mà còn đóng vai trò là nguồn tài nguyên quý giá cho cộng đồng nghiên cứu, cung cấp nền tảng cho các nghiên cứu và phát triển mô hình trong tương lai.

Việc đánh giá toàn diện hiệu suất của mô hình, sử dụng điểm BLEU, đã giúp đề án tìm ra được kiến trúc hiệu quả nhất cho bài toán trong việc tạo ra các bản tóm tắt mạch lạc và phù hợp với ngữ cảnh.

Đề án cũng gặp phải nhiều thách thức nhấn mạnh sự phức tạp của học máy và NLP. Với việc chỉ được đào tạo trên một tập dữ liệu cụ thể, mô hình vẫn gặp phải các vấn đề về khả năng khái quát hóa trên các loại nội dung khác nhau, như gặp tình trạng không nắm bắt

được chính xác nội dung chính cần đưa vào bản tóm tắt. Do đó nhấn mạnh sự cần thiết của các bộ dữ liệu rộng hơn nữa để đảm bảo khả năng thích ứng và hiệu suất của mô hình trên đa dạng các dữ liệu đầu vào.

Chi phí tính toán liên quan đến các mô hình BERT và LSTM cũng là một thách thức lớn khác, đặt ra những hạn chế đối với các ứng dụng thời gian thực hoặc triển khai trong môi trường hạn chế về tài nguyên.

Trong tương lai đề án có thể tập trung vào việc phát triển mô hình BERT-LSTM thông qua cải tiến kiến trúc, huấn luyện trên các bộ dữ liệu đa dạng hơn. Những điều này có thể nâng cao hơn nữa chất lượng tóm tắt và tính linh hoạt của mô hình.

Ngoài việc tóm tắt tin tức, mô hình có thể thử nghiệm trên các nhiệm vụ tóm tắt ở các dạng văn bản khác tin tức, từ việc tóm tắt các văn bản pháp luật đến các bài báo học thuật và các nội dung trên mạng xã hội.

Tóm lại, đề án này đã đóng góp một phần nhỏ trong bài toán tóm tắt tin tức Tiếng Việt, kiểm chứng được khả năng sử dụng các mô hình được đào tạo trước cho các bài toán của ngôn ngữ Việt Nam. Những thách thức gặp phải trong suốt quá trình nghiên cứu đã giúp mang lại những hiểu biết sâu sắc trong lĩnh vực xử lý ngôn ngữ tự nhiên.