HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN ĐÌNH TUẤN

TÓM TẮT TIN TỨC TIẾNG VIỆT SỬ DỤNG MÔ HÌNH BERT

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI-2024

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

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC:

PGS.TS NGUYỄN MẠNH HÙNG

HÀ NỘI-2024

LỜI CAM ĐOAN

Tôi xin cam đoan mọi nội dung nghiên cứu, số liệu, kết quả trong đề án "Tóm tắt tin tức tiếng Việt sử dụng mô hình BERT" của tôi là công trình nghiên cứu của cá nhân tôi, mọi nội dung đều là trung thực và không sao chép từ bất kì báo cáo, công trình nào có trước.

Ký và ghi rõ họ tên

NGUYỄN ĐÌNH TUẤN

LỜI CẢM ƠN

Tôi xin gửi lời cảm ơn sâu sắc đến Phó Giáo Sư, Tiến Sĩ Nguyễn Mạnh Hùng, Khoa Công Nghệ Thông Tin, vì sự hỗ trợ và hướng dẫn liên tục trong suốt quá trình thực hiện đề án của tôi. Sự đóng góp của Thầy trong việc giảng dạy và hướng dẫn, từ việc lựa chọn đề tài đề án cho đến việc thực hiện và kiểm tra quá trình soạn thảo báo cáo này một cách tỉ mỉ, đều là rất quý giá. Kiến thức chuyên môn và lời khuyên sâu sắc của Thầy đã góp phần quan trọng giúp tôi hoàn thành thành công đề án của mình.

Ngoài ra, tôi cũng xin gửi lời cảm ơn chân thành tới toàn thể các thầy cô trong Khoa Công Nghệ Thông Tin, và các Thầy/Cô ở khoa Sau Đại Học, những người đã tận tâm giảng dạy và hướng dẫn tôi trong suốt hai năm học tập. Trí tuệ và sự động viên của mọi người là nền tảng cho sự phát triển của tôi.

NGUYỄN ĐÌNH TUẤN

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
BERT	Bidirectional Encoder Representations from Transformers	Bộ mã hoá hai chiều từ Transformer
Seq2seq	Sequence-to-sequence	Khung tuần tự
LSTM	Long Short-Term Memory	Bộ nhớ dài ngắn hạn
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
SVM	Support Vector Machine	Máy vectơ hỗ trợ
GPT	Generative Pre-trained Transformer	Bộ biến đổi tạo sinh được huấn luyện trước
BLLA	BERT LSTM LSTM with Attention	BERT với LSTM và cơ chế chú ý
Mask LM	Masked Language Model	Mô hình ngôn ngữ mặt nạ
NSP	Next Sentence Prediction	Dự đoán câu tiếp theo
RNN	Recurrent neural network	Mạng nơ ron hồi quy
GRU	Gated Recurrent Unit	Mạng nơ ron hồi quy với nút có cổng

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

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. Các phương pháp tiếp cận truyền thống đối với bài toán này chủ yếu dựa vào các phương pháp trích xuất, trong đó chọn ra một tập con các cụm từ hoặc câu hiện có từ văn bản nguồn để tạo thành một bản tóm tắt. Mặc dù giải quyết được ở một mặt nào đó, nhưng những phương pháp này thường cho ra kết quả là một bản tóm tắt rời rạc các câu từ và điểm quan trọng là không giống một bản tóm tắt do con người viết do các câu từ không được liên kết chặt chẽ.

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, các mô hình học máy sẽ học để hiểu được bối cảnh, ngữ nghĩa của văn bản nguồn, ở mức độ sâu nhất, sau đó sử dụng các phương pháp và kỹ thuật của xử lý ngôn ngữ tự nhiên để cho ra một bản tóm tắt giống như cách con người thực hiện. 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. Tiếng Việt, với cấu trúc cú pháp phức tạp và hình thái phong phú, đòi hỏi một mô hình có thể hiểu sâu sắc ngữ cảnh và sắc thái điều này ngoài những gì mà khung seq2seq truyền thống cung cấp.

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. Kiến trúc biến đổi và học trên hai chiều của văn bản đã giúp cho BERT có khả năng nắm bắt được sự phức tạp của ngôn ngữ, khiến nó trở thành ứng cử viên lý tưởng cho các nhiệm vụ đòi hỏi sự hiểu biết sâu sắc về ngữ cảnh, chẳng hạn như tóm tắt 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. Cách tiếp cận này không chỉ hứa hẹn nâng cao hiệu quả và độ chính xác của việc tóm tắt tin tức Tiếng Việt mà còn góp phần mở rộng lĩnh vực xử lý ngôn ngữ tự nhiên bằng cách thể hiện khả năng thích ứng và hiệu quả của BERT trong việc xử lý các thách thức cụ thể về ngôn ngữ của Việt Nam.

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. Các chương sắp tiếp theo sẽ đi sâu vào các khía cạnh cơ sở lý thuyết, kỹ thuật và thực nghiệm của giải pháp này, với mục đích đưa ra một cách tiếp cận toàn diện và hiệu quả cho vấn đề.

- **Chương 2: Tóm tắt tin tức Tiếng Việt sử dụng mô hình BERT**

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 tiếp theo sẽ đi sâu vào việc

triển khai và đánh giá giải pháp này, bao gồm cả cách tinh chỉnh từng phần của mô hình để giải quyết những thách thức cụ thể trong quá trình nghiên cứu của đề án.

● **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

Sự ra đời của internet đã kéo theo sự bùng nổ về dữ liệu, đặc biệt là các tin tức, bài báo trực tuyến, nhờ vậy mọi người đều có thể cập nhật thông tin từ khắp nơi trên thế giới với tốc độ tính bằng giây. Lượng thông tin khổng lồ này, mặc dù mang lại nhiều lợi ích, nhưng lại là thách thức đối với những cá nhân muốn cập nhật thông tin mà không tốn quá nhiều thời gian để đọc các tài liệu dài, như những nhà nghiên cứu thị trường, những nhà đầu tư, ngoài ra cũng là bài toán lớn cho các hệ thống thông tin khác như hệ thống phân tích hoặc dự đoán thói quen của người dùng, cần xử lý lượng lớn dữ liệu để nắm bắt được các thay đổi của thế giới. Lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP) giải quyết thách thức này thông qua việc phát triển các kỹ thuật tóm tắt văn bản, nhằm mục đích cô đọng các văn bản dài thành những bản tóm tắt ngắn gọn. Những bản tóm tắt này cố gắng duy trì thông điệp cốt lõi và mục đích của văn bản gốc đồng thời giảm đáng kể thời gian dành cho người đọc.

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, đòi hỏi phải hiểu sâu hơn về ngôn ngữ và mối liên hệ của các từ.

Ví dụ một đoạn tin tức về thị trường cổ phiếu:

“Thị trường chứng khoán Việt Nam hôm nay chứng kiến sự sôi động không ngừng với hàng loạt cổ phiếu tăng giá mạnh. Trong số đó, cổ phiếu của các công ty công nghệ và bất động sản được giao dịch nhiều nhất, thu hút sự quan tâm lớn từ phía nhà đầu tư. Các chuyên gia kinh tế nhận định, sự tăng trưởng này phản ánh niềm tin mạnh mẽ vào tiềm năng tăng trưởng kinh tế của Việt Nam trong thời gian tới. Họ cũng lưu ý rằng, bên cạnh cơ hội, thị trường cũng đang đối mặt với không ít thách thức, bao gồm lạm phát và sự biến động của thị trường quốc tế.”

Tóm tắt trích xuất:

"Thị trường chứng khoán Việt Nam hôm nay chứng kiến sự sôi động với cổ phiếu công nghệ và bất động sản được giao dịch nhiều nhất. Các chuyên gia kinh tế nhận định, sự tăng trưởng phản ánh niềm tin vào tiềm năng kinh tế của Việt Nam, họ cũng lưu ý những thách thức bao gồm lạm phát và sự biến động của thị trường quốc tế."

Trong bản tóm tắt trích xuất này, các câu chính từ văn bản gốc được chọn lọc và biên soạn để tạo thành một bản tóm tắt ngắn gọn, nắm bắt được các điểm chính: hoạt động của thị trường chứng khoán, trọng tâm là cổ phiếu công nghệ và bất động sản, cũng như quan điểm của các chuyên gia kinh tế về tiềm năng tăng trưởng và những thách thức.

Tóm tắt tóm lược:

"Hôm nay, thị trường chứng khoán Việt Nam sôi động với sự nổi bật của cổ phiếu công nghệ và bất động sản, cho thấy niềm tin vào tăng trưởng kinh tế dù đối mặt thách thức từ lạm phát và ảnh hưởng từ thị trường quốc tế."

Ở mặt khác, tóm tắt tóm lược diễn đạt lại và cô đọng văn bản gốc thành một dạng mới, tóm tắt những điểm chính mà không sao chép trực tiếp cả câu. Nó đề cập đến sự sôi động của thị trường chứng khoán, các lĩnh vực được quan tâm, triển vọng lạc quan về tăng trưởng kinh tế và thừa nhận những thách thức tiềm ẩn, tất cả đều được trình bày ngắn gọn, tổng hợp.

Ví dụ này minh họa sự khác biệt giữa tóm tắt trích xuất và tóm lược: cách đầu tiên chọn các phần của văn bản gốc, trong khi cách sau dùng một cách diễn đạt khác cho nội dung thành một dạng mới, ngắn hơn, cả hai đều nhằm mục đích chuyển văn bản gốc thành dạng mới đơn giản hơn nhưng không làm mất nội dung và ý nghĩa.

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 $Y' \not\subseteq X$ bản tóm tắt là dạng tóm lược, có thành phần của bản tóm tắt không xuất hiện trong văn bản nguồn.

Áp dụng bài toán tóm tắt văn bản vào tóm tắt tin tức tiếng Việt đặt ra những thách thức đặc biệt do tính chất ngữ điệu, cấu trúc cú pháp phức tạp và hình thái phong phú của ngôn ngữ Tiếng Việt. Những yếu tố này đòi hỏi sự hiểu biết nâng cao về ngữ nghĩa và sắc thái Tiếng Việt để đảm bảo rằng các bản tóm tắt vừa chính xác đầy đủ vừa mạch lạc về mặt ngôn ngữ, chính tả. Các mô hình NLP truyền thống, thường được phát triển tập trung vào tiếng Anh, có thể không hoạt động hiệu quả với văn bản tiếng Việt, điều này làm nổi bật sự cần thiết của các phương pháp tiếp cận chuyên biệt.

Sự khan hiếm các bộ dữ liệu toàn diện được xử lý dành cho tiếng Việt càng làm phức tạp thêm việc phát triển và đánh giá các mô hình tóm tắt. Đề án này nhằm mục đích khám phá và ứng dụng các kỹ thuật NLP tiên tiến, đặc biệt là mô hình BERT, nhằm giải quyết bài toán tóm tắt văn bản tin tức Tiếng Việt, nâng cao hiệu quả và độ chính xác của các công cụ tóm tắt tiếng Việt.

1.2 Các nghiên cứu liên quan

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

Tóm tắt văn bản là một nhiệm vụ quan trọng trong lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP), đã chứng kiến những tiến bộ đáng kể nhờ nỗ lực của các nhà nghiên cứu trên toàn thế giới. Những nỗ lực nghiên cứu này kéo dài trong nhiều thập kỷ, các phương pháp đã chuyển đổi từ các hệ thống dựa trên quy tắc sang các phương pháp học máy và học sâu phức tạp. Phần này cung cấp cái nhìn tổng quan về những phát triển này, nhấn mạnh các nghiên cứu then chốt, những đột phá về công nghệ và bối cảnh phát triển của lĩnh vực 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. Các hệ thống này dựa trên các quy tắc

ngôn ngữ được xác định trước để xác định các câu hoặc cụm từ chính để đưa vào bản tóm tắt. Một công trình quan trọng trong giai đoạn này là của Luhn (1958), người đã đề xuất một phương pháp dựa trên tần suất của các từ quan trọng [15], đặt nền móng cho việc tóm tắt văn bản tự động.

Đột phá với học máy

Sự ra đời của học máy đã mang lại sự thay đổi lớn trong nghiên cứu ở lĩnh vực tóm tắt văn bản. Các kỹ thuật như Máy vectơ hỗ trợ (SVM) và cây quyết định đã được sử dụng để phân loại các câu dựa trên khả năng chúng được đưa vào bản tóm tắt. Công trình của Kupiec, Pedersen và Chen (1995) về một thuật toán tóm tắt văn bản có thể huấn luyện được [12] đã đánh dấu cột mốc quan trọng, cho thấy tiềm năng của học máy trong việc tự động hóa quá trình tóm tắt.

Cuộc cách mạng học sâu

Thập kỷ vừa qua đã có nhiều thay đổi nhờ cuộc cách mạng học sâu. Sự ra đời của các mô hình khung tuần tự (seq2seq) của Sutskever, Vinyals và Le (2014) [18] cũng như sự phát triển của các cơ chế chú ý (attention) và mô hình biến đổi (Transformer) sau đó đã cải thiện đáng kể chất lượng trong cả bài toán tóm tắt trích xuất và tóm tắt tóm lược. Đáng chú ý là mô hình BERT của Devlin et al. (2018) [10] và GPT của OpenAI đã thiết lập ra các tiêu chuẩn mới trong việc tạo các bản tóm tắt mạch lạc và phù hợp với ngữ cảnh.

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

Khái niệm mã hoá ngữ cảnh, được giới thiệu bởi các mô hình như BERT (Devlin và cộng sự, 2018) [10] và ELMo (Peters và cộng sự, 2018) [16], đã thay đổi cách tiếp cận trong việc hiểu các sắc thái ngữ nghĩa của văn bản, mang lại lợi ích đáng kể cho các nhiệm vụ tóm tắt tóm lược nói riêng và Xử lý ngôn ngữ tự Nhiên nói chung. Các mô hình biến đổi (Transformer), với cơ chế tự chú ý (self-attention), đã cải tiến hơn nữa khả năng nắm bắt bản chất của văn bản qua các chuỗi dài, khiến chúng đặc biệt phù hợp để ứng dụng trong tóm tắt các văn bản phức tạp.

Lĩnh vực tóm tắt văn bản đã được nghiên cứu từ lâu trên thế giới, đi từ việc sử

dụng các cách xử lý dựa trên quy tắc, tới sự phát triển các mô hình học máy tiên tiến ngày nay, đã cho thấy sự phát triển mạnh mẽ và tầm quan trọng của lĩnh vực này.

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

Tại Việt Nam, lĩnh vực tóm tắt văn bản cũng nhận được nhiều chú ý trong thời gian gần đây. Các nghiên cứu mới nhất tập trung vào việc áp dụng các kỹ thuật NLP tiên tiến vào giải quyết bài toán. Phần này nêu bật những nghiên cứu chính ở Việt Nam đã góp phần đáng kể vào sự tiến bộ của công nghệ tóm tắt văn bản Tiếng Việt.

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]. Bằng cách sử dụng mô hình Word2vec để trích xuất và biểu diễn các từ Tiếng Việt trong văn bản, họ đã sử dụng mô hình khung tuần tự (Sequence-to-Sequence) và cơ chế chú ý (Attention) để tạo ra các bản tóm tắt ngắn gọn, kết quả được lấy ra bằng thuật toán Beam Search. Công trình của họ, đã được xuất bản vào năm 2017, đánh dấu một bước quan trọng trong việc ứng dụng học sâu vào bài toán với Tiếng Việt và cho thấy tiềm năng của các mô hình khung tuần tự trong việc hiểu và tóm tắt các văn bản Tiếng Việt phức tạp.

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]. Họ đã giới thiệu một phương pháp dựa trên các quy tắc diễn ngôn, các ràng buộc cú pháp và biểu đồ từ để tạo ra các bản tóm tắt từ các ý chính của văn bản. Cách tiếp cận này nhấn mạnh sự phức tạp của việc tạo ra các bản tóm tắt mạch lạc với đầy đủ thông tin mà không cần trích xuất trực tiếp các câu, cho thấy việc giải quyết bài toán tóm tắt tóm lược với Tiếng Việt là rất khả thi.

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]. Nghiên cứu của họ chứng minh tính hiệu quả của việc sử dụng các mô hình ngôn ngữ được đào tạo trước như BERT, được tăng cường bằng các câu phụ trợ được tạo từ các từ khóa khía cạnh. Phương pháp này cho phép hiểu rõ hơn các ý kiến trong văn bản Tiếng Việt, góp phần mở rộng lĩnh vực phân tích quan điểm và khai phá khía cạnh trong Tiếng Việt.

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]. Bằng cách biểu diễn các câu dưới dạng vectơ đặc trưng thông qua BERT và phân loại chúng để xác định các câu quan trọng nhất cho bản tóm tắt, cách tiếp cận của họ kết hợp các điểm mạnh của học sâu với hiệu quả của tóm tắt trích xuất. Nghiên cứu này đã nhấn mạnh khả năng hoạt động hiệu quả của mô hình BERT với Tiếng Việt, cũng cho chúng ta thấy tiềm năng lớn của sử dụng các mô hình huấn luyện trước như BERT trong bài toán với ngôn ngữ Tiếng Việt.

Các nghiên cứu nói trên đã cho thấy sự phát triển mạnh mẽ của nghiên cứu tóm tắt văn bản ở Việt Nam. Thông qua việc tích hợp các công nghệ NLP như học sâu, mô hình khung tuần tự và kiến trúc dựa trên mô hình biến đổi, các nhà nghiên cứu đang có những bước tiến đáng kể trong việc vượt qua những thách thức mà ngôn ngữ Tiếng Việt đặt ra. Những đóng góp này không chỉ nâng cao nền tảng kiến thức về NLP mà còn là tiền đề để tạo ra các công cụ tóm tắt văn bản phức tạp hơn phù hợp với nhu cầu của người Việt.

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 sẽ sử dụng trong giải pháp của đề án.

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

Mô hình sẽ cách mạng hóa lĩnh vực Xử Lý ngôn ngữ tự nhiên (NLP) lần đầu tiên được đề xuất trong bài báo có tiêu đề "Attention is All You Need" của Ashish Vaswani và cộng sự, xuất bản năm 2017 [21]. Công trình quan trọng này đã giới thiệu mô hình biến đổi (Transformer), đánh dấu sự phát triển từ các mô hình dựa trên trình tự trước đó như mạng hồi quy (Recurrent Neural Network) (RNN) và mạng bộ nhớ dài ngắn hạn (Long short-term memory) (LSTM).

Bối cảnh lịch sử

Trước khi các mô hình Transformer ra đời, các tác vụ NLP chủ yếu được xử lý bởi RNN và LSTM, được thiết kế để xử lý dữ liệu tuần tự bằng cách nắm bắt các phần phụ thuộc ở các quy mô khác nhau. Tuy nhiên, những mô hình này gặp phải nhiều thách thức, chẳng hạn như khó khăn trong việc xử lý song song và các vấn đề về ghi nhớ cần thành phần trong chuỗi dài, điều này thường dẫn đến cản trở việc đạt hiệu suất tối ưu đối với các nhiệm vụ hiểu ngôn ngữ và văn bản phức tạp.

Giới thiệu mô hình biến đổi

Mô hình Transformer được đề xuất như một giải pháp cho những vấn đề này, giới thiệu một kiến trúc mới chỉ dựa trên cơ chế chú ý (Attention) mà không phụ thuộc vào xử lý hồi quy hoặc tích chập. Bằng cách sử dụng khả năng tự chú ý (self-attention), mô hình Transformer có thể cân nhắc tầm quan trọng của các phần khác nhau của dữ liệu đầu vào, cho phép nó xử lý đồng thời toàn bộ chuỗi dữ liệu và khả năng ghi nhớ các thành phần trong những chuỗi dài hiệu quả hơn.

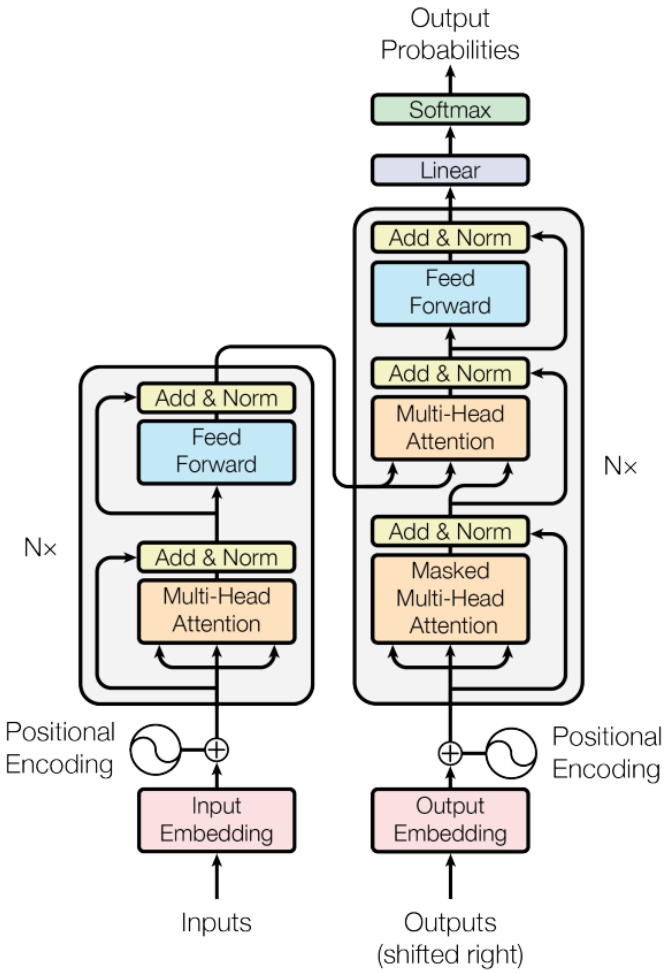
Tác động đến NLP

Sự ra đời của mô hình Transformer thể hiện sự phát triển trong NLP. Nó không chỉ giải quyết những hạn chế cố hữu của các mô hình trước đó mà còn cải thiện đáng kể hiệu quả và hiệu suất xử lý dữ liệu là ngôn ngữ. Khả năng tính toán song song của mô hình và khả năng vượt trội của nó trong việc mô hình hóa các phụ thuộc phức tạp

trong dữ liệu văn bản đã dẫn đến những tiến bộ đáng chú ý trong một loạt nhiệm vụ NLP, từ dịch máy và tóm tắt văn bản đến hệ thống trả lời câu hỏi.

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, khác biệt về cơ bản so với các mô hình ra đời trước. Kiến trúc này được xây dựng xung quanh hai khái niệm cốt lõi: cơ chế chú ý và mã hóa vị trí, cùng nhau chúng mang lại hiệu suất vượt trội trong việc xử lý dữ liệu tuần tự như văn bản. Phần này sẽ tìm hiểu các thành phần quan trọng này và luồng hoạt động của mô hình Biến Đổi.



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

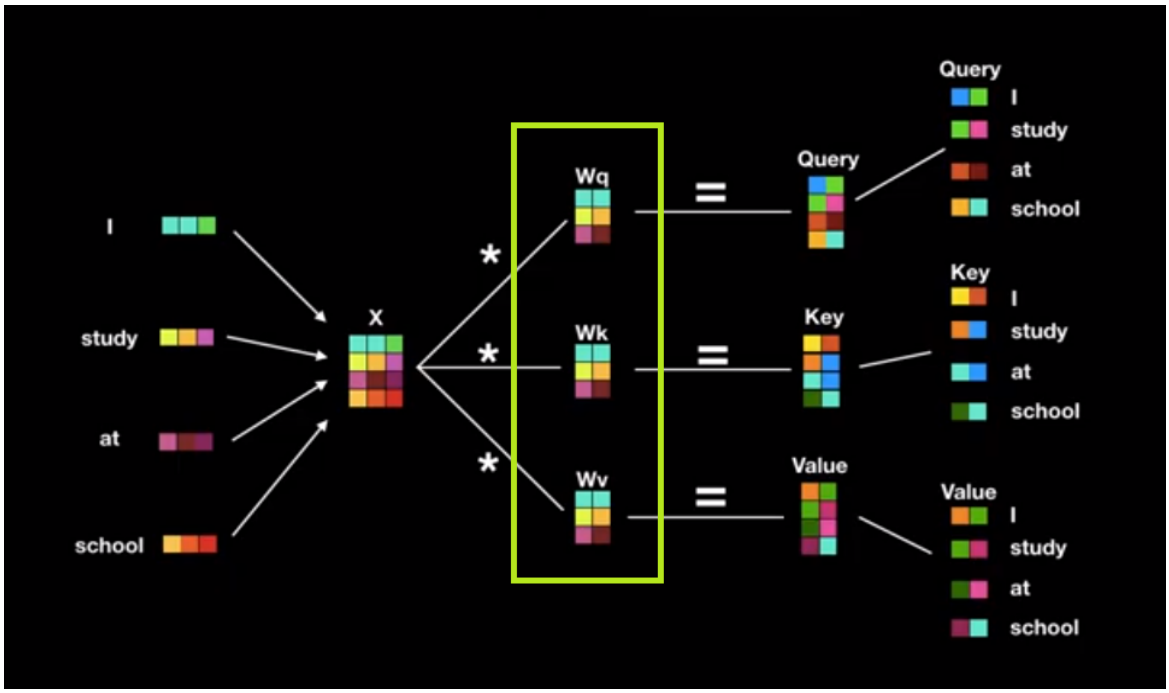
Mô hình Transformer bao gồm một bộ mã hóa và một bộ giải mã, giống như cấu trúc của khung tuần tự seq2seq, mỗi lớp bao gồm nhiều lớp con thực hiện các hoạt động tự chú ý và mạng truyền thẳng.

Bộ mã hóa: Bộ mã hóa xử lý chuỗi đầu vào thông qua 6 lớp con của nó, mỗi

lớp con bao gồm một lớp tự chú ý và một mạng truyền thẳng (fully connected feed-forward network), cùng với một lớp chuẩn hóa (layer normalization) và một kết nối dư (residual connections) ngay sau lớp chuẩn hóa. Đầu ra là một tập hợp các vectơ biểu thị chuỗi đầu vào trong không gian đa chiều (512 chiều), được tăng cường thông tin theo ngữ cảnh [21].

Bộ giải mã: Bộ giải mã cũng có kiến trúc tương tự với 6 lớp con nhưng trong mỗi lớp con thêm một lớp chú ý mặt nạ (Masked Multi-Head Attention) bổ sung tại đầu vào của bộ mã hóa. Lớp này được điều chỉnh để không đưa các từ của các vị trí tiếp theo vào attention, khi đang thực hiện dự đoán tại vị trí hiện tại. Ngoài ra đầu vào của bộ giải mã khi đưa vào bộ mã hóa (Output Embedding) cũng đã được dịch qua phải một vị trí, kết hợp điều này vào lớp chú ý mặt nạ, sẽ đảm bảo dự đoán cho vị trí thứ i chỉ phụ thuộc vào đặc điểm ngữ nghĩa của các vị trí trước i [21].

Lớp phân phối xác suất: Sau bộ giải mã là 2 lớp Linear và Softmax để tìm ra phân phối xác suất cho các từ dự đoán [21].



Hình 2-2 Cơ chế tự chú ý [24]

Cơ chế tự chú ý

Trọng tâm của kiến trúc Transformer là cơ chế tự chú ý (self-attention), một

cách tiếp cận cho phép khi thực hiện mã hóa cho mỗi từ mô hình sẽ tìm liên kết ở mọi từ khác, để biểu diễn ngữ nghĩa. Không giống như các mô hình truyền thống xử lý các chuỗi một cách tuần tự, 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, cho phép mô hình hiểu được sự liên quan giữa các từ trong câu bất kể khoảng cách của chúng trong văn bản. 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. Mức độ liên quan (sự chú ý) của từng từ với các từ khác được tính toán bằng cách lấy tích vô hướng của vectơ truy vấn (Query) và vectơ khóa (Key) của chúng, sau đó chuẩn hóa bằng một hàm softmax để đưa về một phân phối xác suất mà độ lớn sẽ đại diện cho mức độ chú ý (attention) của từ query tới từ key.

Phương trình tính toán mức độ chú ý (Attention) như sau [24]:

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

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. Điều này đảm bảo rằng mô hình có thể nhận dạng và sử dụng thứ tự của các từ [21].

Kiến trúc Transformer, với việc sử dụng sáng tạo các cơ chế tự chú ý và mã hóa vị trí, thể hiện một tiến bộ lớn. Bằng cách cho phép lập mô hình trực tiếp từ các mối quan hệ giữa tất cả các phần của dữ liệu đầu vào và duy trì thứ tự trình tự, Transformer đặt nền tảng cho các mô hình như BERT đạt được nhiều thành công trong các nhiệm vụ về hiểu ngôn ngữ.

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 (The Bidirectional Encoder Representations) 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), BERT xây dựng dựa trên các nguyên tắc nền tảng của mô hình Biến Đổi [10]. BERT đã cách mạng hóa cách máy móc hiểu ngôn ngữ của con người. Kiến trúc của nó bắt nguồn từ mô hình Transformer, đặc biệt sử dụng cơ chế mã hóa của Transformer để xử lý các từ liên quan đến tất cả các từ khác trong câu, từ đó nắm bắt ngữ cảnh hiệu quả hơn.

Mối quan hệ của BERT với mô hình Biến Đổi

Xây dựng bối cảnh trên hai chiều: Trong khi mô hình Transformer ban đầu xử lý dữ liệu văn bản bằng cách xem xét mối quan hệ giữa các cặp từ trong cả chuỗi đầu vào và đầu ra, 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), không giống như các mô hình truyền thống thường đọc văn bản theo thứ tự [10].

Cơ chế tự chú ý: Trọng tâm của BERT, giống như Transformer, là cơ chế tự chú ý. Cơ chế này cho phép BERT cân nhắc mức độ quan trọng của từng từ trong câu với mọi từ khác, từ đó tạo ra các mối quan hệ ngữ cảnh phong phú trên toàn bộ văn bản. 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). BERT được đào tạo trước trên một kho dữ liệu văn bản lớn chưa được gắn nhãn trên nhiều ngôn ngữ và chủ đề, mục tiêu là học cách biểu diễn ngôn ngữ chung. Sau đó, BERT có thể được tinh chỉnh chỉ với một lớp đầu ra bổ sung để tạo ra các mô hình mới cho nhiều nhiệm vụ khác nhau, chẳng

hạn như trả lời câu hỏi, phân tích cảm xúc và quan trọng là tóm tắt văn bản [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. Những cách đào tạo mới này giúp BERT hiểu được bối cảnh và mối quan hệ của ngôn ngữ, khiến nó trở nên khác biệt so với các mô hình dựa trên Transformer trước đây [10].

Kiến trúc của BERT thể hiện một bước tiến đáng kể trong NLP, được xây dựng dựa trên nền tảng biến đổi do mô hình Transformer đưa ra. Bằng cách khai thác sức mạnh của cơ chế xử lý hai chiều và tự chú ý, BERT đạt được sự hiểu biết sâu sắc về sắc thái và ngữ cảnh của ngôn ngữ. Tác động của nó đối với NLP là rất lớn, đưa ra các phương pháp mới để xây dựng các mô hình xử lý ngôn ngữ nhận biết theo ngữ cảnh phức tạp hơn. Kế thừa từ mô hình Transformer, BERT không chỉ được hưởng lợi từ những hiểu biết sâu sắc của mô hình Biến Đổi mà còn mở rộng đáng kể, mang lại khả năng hiểu ngôn ngữ hiệu quả hơn rất nhiều so với các mô hình trước đây.

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. Sự ra đời của BERT đánh dấu một bước cột mốc quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), đưa ra các tiêu chuẩn mới cho sự phát triển của công nghệ hiểu ngôn ngữ.

Nguồn gốc của BERT

Giới thiệu và phát triển: Được phát triển bởi Jacob Devlin và nhóm của ông tại Google, BERT xuất hiện từ nhu cầu cải thiện khả năng hiểu ngôn ngữ của máy tính theo cách có nhiều sắc thái và nhận biết ngữ cảnh hơn. Tận dụng cơ chế tự chú ý của kiến trúc Transformer, BERT được thiết kế để huấn luyện trước nhằm tạo ra các mã hóa hai chiều bằng cách sử dụng kết hợp cả ngữ cảnh bên trái và bên phải trong tất cả các lớp của nó.

Tác động đến NLP

Tiêu chuẩn mới cho các mô hình ngôn ngữ: sau khi được giới thiệu BERT đã nhanh chóng thiết lập các tiêu chuẩn mới trong một loạt các nhiệm vụ NLP, bao gồm nhưng không giới hạn ở việc trả lời câu hỏi, hiểu ngôn ngữ tự nhiên và nhận dạng thực thể có tên (named entity recognition). Thành công của BERT đã chứng minh tiềm năng sâu sắc của các mô hình ngôn ngữ được đào tạo trước trong việc cải thiện đáng kể hiệu suất đối với các nhiệm vụ NLP chỉ với những điều chỉnh tối thiểu theo từng nhiệm vụ cụ thể.

Chuyển hướng sang các mô hình được đào tạo trước: Một trong những tác động đáng kể nhất của BERT là sự thay đổi trong nghiên cứu và ứng dụng NLP theo hướng tận dụng các mô hình được đào tạo trước. Bằng cách chứng minh rằng một mô hình được đào tạo trước có thể được tinh chỉnh để đạt được hiệu suất cao trong nhiều nhiệm vụ, BERT đã truyền cảm hứng cho sự phát triển của các mô hình tiếp theo như GPT-2, GPT-3, RoBERTa và các mô hình khác, đồng thời thúc đẩy hơn nữa lĩnh vực này.

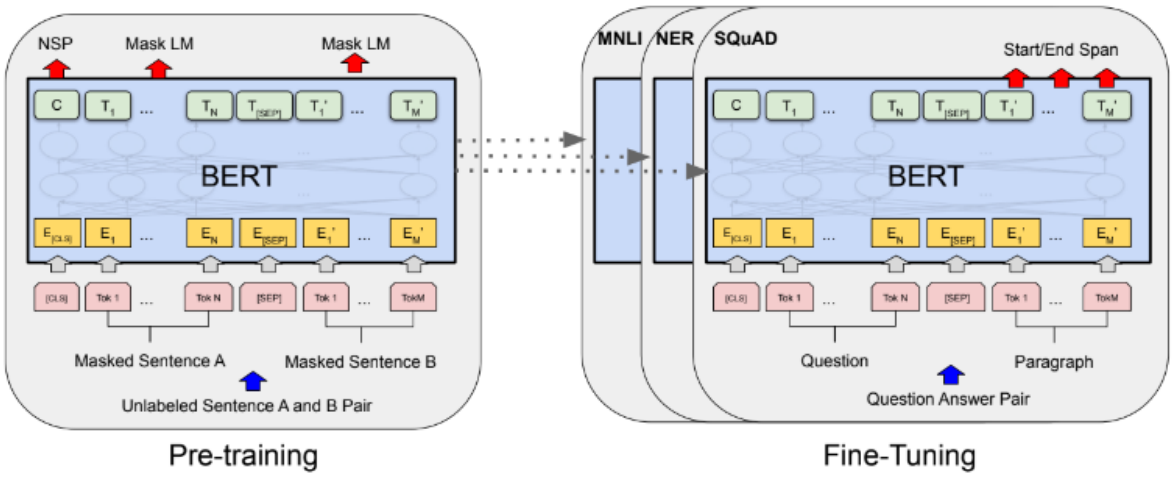
Nâng cao khả năng hiểu ngôn ngữ: Ngoài việc đạt được điểm cao trong các bài kiểm tra mô hình ngôn ngữ, BERT về cơ bản đã thay đổi cách máy móc hiểu ngôn ngữ của con người. Phương pháp đào tạo hai chiều của nó cho phép hiểu biết toàn diện hơn về ngữ cảnh và ngữ nghĩa, cho phép khả năng xử lý ngôn ngữ chính xác và đa sắc thái hơn.

Sự ra đời của BERT tại Google AI Language vào năm 2018 đã có tác động thay đổi lĩnh vực NLP, mở ra thời kì phát triển của các mô hình ngôn ngữ được đào tạo trước tiếp tục vượt qua ranh giới về những gì có thể có trong việc hiểu ngôn ngữ. Cách tiếp cận sáng tạo của nó đối với đào tạo hai chiều và xử lý nhận biết ngữ cảnh không chỉ đặt ra các tiêu chuẩn mới cho các nhiệm vụ NLP mà còn truyền cảm hứng cho một làn sóng nghiên cứu và phát triển nhằm tìm cách khai thác hơn nữa sức mạnh của việc học sâu trong việc hiểu và tạo ra ngôn ngữ của con người. Do đó, BERT được coi là mô hình nền tảng trong quá trình phát triển không ngừng của công nghệ NLP, biểu thị một bước nhảy vọt trong nỗ lực thu hẹp khoảng cách giao tiếp giữa con người và máy móc.

2.2.3 Khám phá kiến trúc của BERT

Khi được giới thiệu, kiến trúc của BERT là một sự sáng tạo lớn trong Xử lý ngôn ngữ tự nhiên (NLP), thể hiện một bước nhảy vọt đáng kể trong nỗ lực giúp máy móc hiểu và xử lý ngôn ngữ của con người. Được xây dựng dựa trên mô hình Transformer, kiến trúc của BERT được thiết kế để hiểu sâu sắc ngữ cảnh của các từ trong câu từ cả hai hướng (bên trái và bên phải), khác với các mô hình xử lý văn bản theo một hướng trước đó.

Thành phần cốt lõi



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

Đào tạo hai chiều: Không giống như các mô hình xử lý văn bản truyền thống đọc văn bản một cách tuần tự, BERT đọc toàn bộ chuỗi từ cùng một lúc, cho phép nó nắm bắt được ngữ cảnh của một từ dựa trên tất cả các từ xung quanh nó. Tính hai chiều này đạt được thông qua cơ chế tự chú ý của Transformer, cơ chế này cân nhắc mức độ ảnh hưởng của từng từ trong câu đối với từng từ khác [10].

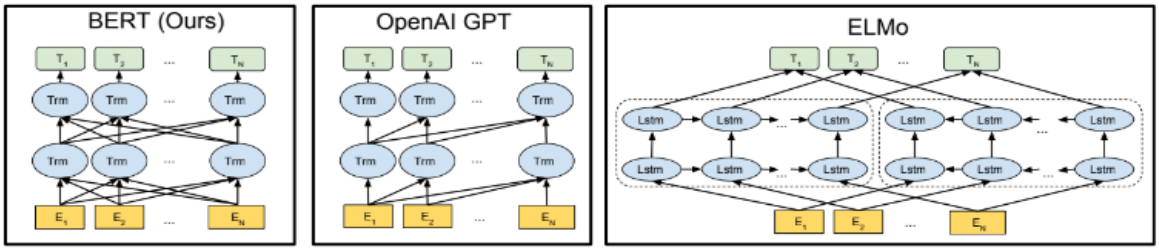
Mô hình ngôn ngữ mặt nạ (Masked Language Model) (Mask LM): Một trong những cách huấn luyện chính của BERT bao gồm việc che giấu ngẫu nhiên các từ trong dữ liệu đầu vào và sau đó dự đoán các từ bị che giấu này chỉ dựa trên ngữ cảnh của chúng. Cách tiếp cận này này cho phép BERT học được sự hiểu biết sâu sắc về ngữ cảnh của ngôn ngữ và các mối quan hệ trên các từ [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. Nhiệm vụ này giúp BERT hiểu được mối quan hệ giữa các câu, nâng cao hơn nữa khả năng hiểu cấu trúc và tính mạch lạc của văn bản [10].

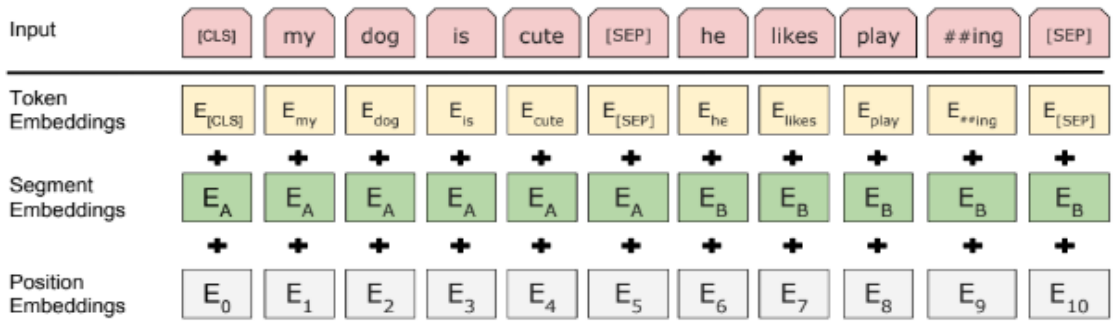
Kiến trúc cụ thể

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). Sự kết hợp của các phần này được đưa vào các lớp Transformer để tạo ra các biểu diễn phong phú theo ngữ cảnh của 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. BERT đặt ra hướng đi mới cho nghiên cứu và phát triển trong lĩnh vực xử lý ngôn ngữ tự nhiên.

2.3 Giới thiệu về mạng bộ nhớ dài ngắn hạn LSTM

2.3.1 Sự ra đời của mạng LSTM

Mạng bộ nhớ dài ngắn hạn (Long Short-Term Memory) (LSTM), một mạng nơ ron hồi quy (recurrent neural network) (RNN), đã được giới thiệu để khắc phục những hạn chế vốn có trong RNN truyền thống, đặc biệt là vấn đề mạng hồi quy không có khả năng nắm bắt các phụ thuộc dài hạn trong chuỗi dữ liệu một cách hiệu quả. Mô hình LSTM lần đầu tiên được đề xuất bởi Sepp Hochreiter và Jürgen Schmidhuber trong bài báo chuyên đề năm 1997 của họ, "Long Short-Term Memory" [18], sau đó được phát triển thêm với nghiên cứu của Felix A. Gers, và cộng sự vào năm 1999 trong bài nghiên cứu "Learning to Forget: Continual Prediction with LSTM" [19], đánh dấu một tiến bộ đáng kể trong lĩnh vực mạng nơ ron nhân tạo và học máy.

Bối cảnh lịch sử

Những thách thức ban đầu với RNN: Trước khi LSTM ra đời, RNN được biết đến với tiềm năng xử lý dữ liệu tuần tự, khiến chúng phù hợp với các tác vụ như dự đoán chuỗi theo thời gian, nhận dạng giọng nói, v.v. Tuy nhiên, RNN gặp khó khăn trong việc học các phần phụ thuộc dài hạn do các vấn đề như biến mất và bùng nổ độ dốc (gradients), trong đó sự đóng góp của thông tin giảm dần theo thời gian, khiến mô hình khó giữ lại và học hỏi từ các đầu vào trước đó trong một chuỗi dài.

LSTM ra đời

Giải pháp cho vấn đề biến mất độ dốc (Vanishing Gradients): Hochreiter và Schmidhuber đã đề xuất kiến trúc LSTM như một giải pháp cho vấn đề độ dốc biến mất [18]. LSTM được thiết kế với khả năng ghi nhớ thông tin trong thời gian dài nhờ cấu trúc độc đáo của chúng, bao gồm các ô nhớ và cổng thông tin.

Mục đích ban đầu

Được thiết kế để dự đoán tuần tự: Mục đích chính của mạng LSTM là cải thiện khả năng đưa ra dự đoán của mô hình dựa trên chuỗi dữ liệu dài. Bằng cách giữ lại thông tin liên quan và quên dữ liệu không cần thiết thông qua các ô nhớ và cổng, LSTM có thể duy trì độ dốc ổn định hơn trong quá trình huấn luyện, cho phép chúng học từ các điểm dữ liệu đã xảy ra từ lâu trong chuỗi [18].

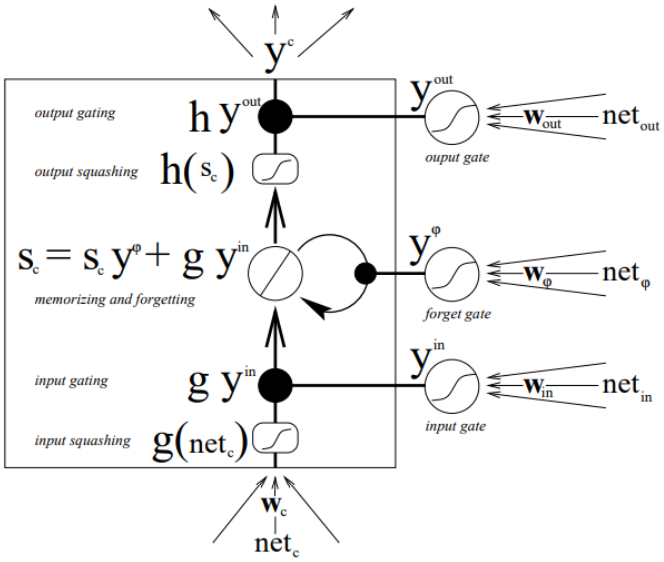
Tác động và phát triển

Áp dụng rộng rãi: Kể từ khi được giới thiệu, LSTM đã được áp dụng rộng rãi trên nhiều lĩnh vực khác nhau. Các nhà nghiên cứu đã phát triển và tối ưu hóa hơn nữa kiến trúc ban đầu, dẫn đến các biến thể như LSTM hai chiều (Bi-directional LSTMs) và mạng nơ ron hồi quy với nút có cổng (GRU), mang lại những cải tiến trong các bài toán cụ thể.

Sự ra đời của mạng LSTM thể hiện một thời điểm quan trọng trong nỗ lực mô hình hóa dữ liệu tuần tự một cách hiệu quả. Bằng cách giải quyết vấn đề quan trọng của việc học các phụ thuộc lâu dài, LSTM đã mở ra những con đường mới trong nghiên cứu và ứng dụng học máy, mở đường cho những tiến bộ trong các lĩnh vực như xử lý ngôn ngữ tự nhiên, nhận dạng giọng nói, v.v. Khả năng ghi nhớ và sử dụng thông tin dài hạn của LSTM đã khiến nó trở thành nền tảng trong việc phát triển các mô hình phức tạp đòi hỏi sự hiểu biết về bối cảnh theo thời gian.

2.3.2 Kiến trúc của LSTM

Mạng bộ nhớ dài ngắn hạn LSTM, một mạng nơ ron hồi quy đặc biệt, được thiết kế để khắc phục những hạn chế của RNN truyền thống trong việc nắm bắt các phụ thuộc dài hạn trong dữ liệu chuỗi. Kiến trúc của LSTM được thiết kế khéo léo để điều chỉnh luồng thông tin, cho phép nó ghi nhớ và quên thông tin trong khoảng thời gian dài một cách hiệu quả. 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): Thành phần quan trọng nhất của LSTM là ô trạng thái, một loại "băng tải" chạy thẳng xuống toàn bộ chuỗi. Nó cho phép thông tin được truyền đi tương đối không thay đổi nếu cần thiết, đảm bảo rằng mạng có thể duy trì sự phụ thuộc lâu dài. Trạng thái ô được sửa đổi bằng các cổng ở mỗi bước, thêm hoặc bớt thông tin nếu cần [19].

Cổng đầu vào (Input Gate): Cổng đầu vào kiểm soát mức độ thông tin mới truyền vào ô trạng thái. Nó bao gồm lớp kích hoạt sigmoid quyết định giá trị nào sẽ cập nhật và lớp tanh tạo ra một vectơ các giá trị mới có thể được thêm vào ô trạng thái [19].

Cổng quên (Forget Gate): Cải tiến quan trọng nhất của LSTM, cổng quên quyết định thông tin nào sẽ bị loại bỏ khỏi ô trạng thái. Nó xem xét trạng thái trước đó và đầu vào hiện tại, chuyển nó qua hàm sigmoid để xác định phần nào của ô trạng thái nên được giữ lại hoặc loại bỏ [19].

Cổng đầu ra (Output Gate): Cổng đầu ra kiểm soát thông tin được đưa ra từ ô trạng thái. Cổng này lấy đầu vào hiện tại và đầu ra trước đó, xử lý chúng thông qua hàm sigmoid để quyết định phần nào của ô trạng thái sẽ được xuất ra, sau đó áp dụng hàm tanh cho ô trạng thái (làm cho các giá trị nằm trong khoảng từ -1 đến 1) và nhân nó

với đầu ra của cổng sigmoid, sao cho chỉ những phần được chọn mới là đầu ra [19].

Chức năng của các cổng

Mỗi đơn vị LSTM lấy đầu vào từ đơn vị trước và chuyển đầu ra của nó sang đơn vị tiếp theo, các cổng trong mỗi đơn vị sẽ đưa ra các quyết định quan trọng về những gì cần giữ lại và những gì cần loại bỏ. Thiết kế này giải quyết vấn đề độ dốc biến mất (vanishing gradient) của RNN bằng cách cho phép thông tin đi qua một cách có chọn lọc, giúp LSTM có khả năng học và ghi nhớ trong các chuỗi dài.

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.

Kiến trúc của mạng LSTM là một tiến bộ đáng kể trong thiết kế mạng nơ ron nhân tạo, đặc biệt đối với các bài toán liên quan đến dữ liệu tuần tự. 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. Khả năng giải quyết những thách thức về sự phụ thuộc dài hạn vào dữ liệu tuần tự đã củng cố vị trí của LSTM như một nền tảng chính của kiến trúc học sâu ngày nay.

2.3.3 Ứng dụng rộng rãi của mạng LSTM

Mạng Bộ nhớ dài ngắn hạn (LSTM), đã trở thành nền tảng trong lĩnh vực học sâu, đặc biệt đối với các nhiệm vụ liên quan đến dữ liệu tuần tự. Kiến trúc độc đáo của nó đã giúp LSTM đạt được hiệu quả cao trên nhiều ứng dụng đa dạng. Phần này khám phá vị trí và lý do LSTM thường được sử dụng, thể hiện tính linh hoạt và tầm quan trọng của LSTM.

Xử lý ngôn ngữ tự nhiên (NLP)

Sinh văn bản (Text Generation): LSTM đóng vai trò then chốt trong việc tạo ra các văn bản mạch lạc và phù hợp với ngữ cảnh, cho phép tạo ra mọi văn bản từ thơ đến các bài báo.

Dịch tự động (Machine Translation): Bằng cách học và hiểu ngữ cảnh câu

và từ, LSTM cải thiện đáng kể chất lượng dịch giữa các ngôn ngữ, duy trì ý nghĩa và sự trôi chảy của văn bản gốc.

Phân tích cảm xúc (Sentiment Analysis): Khả năng hiểu ngữ cảnh qua các chuỗi văn bản dài khiến LSTM trở nên lý tưởng để phân tích và dự đoán cảm xúc của văn bản, từ các bài đăng trên mạng xã hội đến đánh giá sản phẩm.

Dự đoán dữ liệu chuỗi theo thời gian

Phân tích thị trường tài chính: LSTM cũng có thể được sử dụng để dự đoán giá cổ phiếu và xu hướng thị trường bằng cách phân tích chuỗi dữ liệu lịch sử của thị trường.

Dự báo thời tiết: Tính chất tuần tự của dữ liệu thời tiết phù hợp để sử dụng với LSTM để dự đoán điều kiện thời tiết trong tương lai, tận dụng các mô hình trong quá khứ để dự báo nhiệt độ, lượng mưa và các số liệu liên quan đến thời tiết khác.

Nhận dạng giọng nói

Ứng dụng chuyển giọng nói thành văn bản (Voice-to-Text): LSTM đã đóng góp đáng kể vào sự tiến bộ của công nghệ nhận dạng giọng nói, cho phép phiên âm chính xác hơn ngôn ngữ nói thành văn bản.

Tạo nhạc và video

Sáng tác nhạc: Bằng cách học từ chuỗi các nốt nhạc, LSTM có thể tạo ra những bản nhạc mới phù hợp về mặt phong cách với dữ liệu đào tạo.

Dự đoán khung hình video: LSTM có thể dự đoán các khung hình tiếp theo trong tương lai của video dựa trên các khung hình trong quá khứ, hữu ích trong các ứng dụng từ nén video đến nâng cao đồ họa trong trò chơi điện tử.

Tại sao LSTM được sử dụng rộng rãi

Việc áp dụng rộng rãi LSTM trên các lĩnh vực này có thể là do một số yếu tố chính:

Khả năng ghi nhớ: Kiến trúc của LSTM cho phép khả năng ghi nhớ và sử dụng thông tin theo chuỗi dài, một điểm quan trọng đối với các nhiệm vụ phụ thuộc vào thời gian.

Tính linh hoạt: LSTM đã được chứng minh là có hiệu quả cho cả việc xử lý các chuỗi một chiều và hai chiều, khiến LSTM có khả năng thích ứng với nhiều nhiệm vụ.

Hiệu suất được cải thiện: So với RNN truyền thống, LSTM đã cho thấy hiệu

suất vượt trội hơn hẳn trong việc nắm bắt các phụ thuộc dài hạn, dẫn đến mô hình chính xác và đáng tin cậy hơn.

Từ việc nâng cao khả năng hiểu ngôn ngữ tự nhiên cho đến khả năng dự đoán chuỗi thời gian phức tạp, các ứng dụng của mạng LSTM nhấn mạnh tầm quan trọng của nó trong việc vượt qua ranh giới của những gì có thể đạt được bằng trí tuệ nhân tạo. Khả năng xử lý và đưa ra dự đoán dựa trên chuỗi dữ liệu dài của LSTM không chỉ giải quyết được các vấn đề thực tế trên nhiều lĩnh vực khác nhau mà còn mở ra nhiều hướng mới cho nghiên cứu và đổi mới trong phân tích dữ liệu tuần tự.

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.

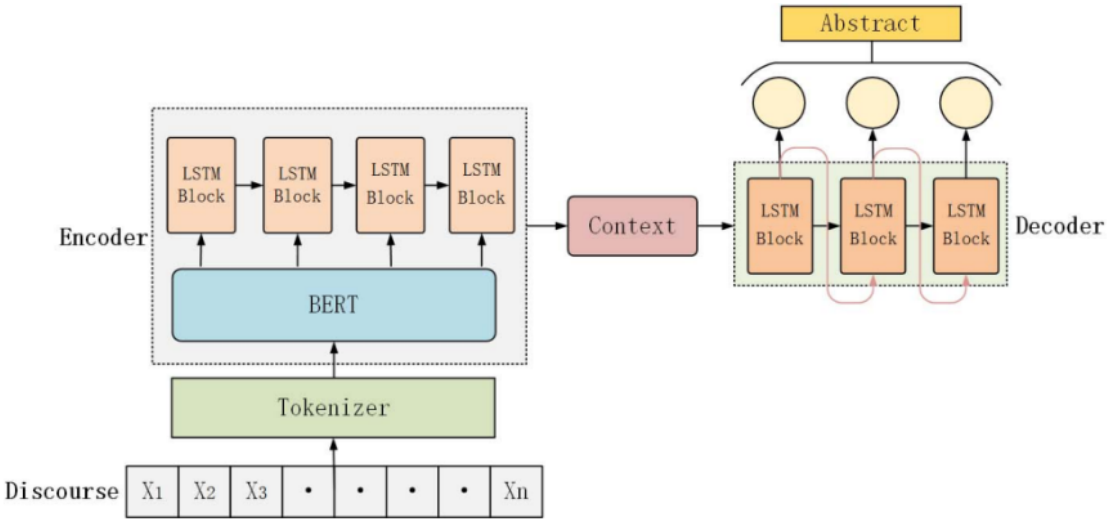
CHƯƠNG 3: GIẢI PHÁP BERT-LSTM-LSTM VỚI CƠ CHẾ TỰ CHÚ Ý VÀ KẾT QUẢ THỰC NGHIỆM, THẢO LUẬN

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

Sự tiến bộ nhanh chóng của các công nghệ Xử lý ngôn ngữ tự nhiên (NLP) mang đến cơ hội giải quyết các thách thức trong bài toán tóm tắt tin tức Tiếng Việt bằng các phương pháp tiếp cận sáng tạo. Đề án đề xuất sử dụng 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 (Bidirectional Encoder Representations from Transformers) 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 sức mạnh của cơ chế chú ý (Attention) nhằm tập trung vào các phần quan trọng của văn bản đầu vào khi tạo ra bản tóm tắt. Mô hình kết hợp này nhằm mục đích giải quyết sự phức tạp của Tiếng Việt và các sắc thái đa dạng trong tin tức để tạo ra những bản tóm tắt chính xác và mạch lạc.

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



Hình 3-1 Mô hình BLLA [11]

Kiến trúc mô hình BLLA được thiết kế để tận dụng kết quả mã hoá văn bản

đầu vào theo ngữ cảnh do BERT tạo ra, mỗi từ trong văn bản đầu vào sau khi được BERT xử lý sẽ cho ra mỗi vector thể hiện các khía cạnh của từ đó theo ngữ cảnh trong văn bản, kết quả này sẽ là đầu vào cho mô hình seq2seq bao gồm các lớp LSTM để tạo ra bản tóm tắt. 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). Bộ mã hóa LSTM này xử lý để tạo ra một bản nén cô đọng của văn bản đầu vào, để giúp cho mô hình nắm bắt thông tin và nội dung thiết yếu của nó.

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. Quá trình lặp lại này tận dụng khả năng ghi nhớ của LSTM, cho phép bộ giải mã duy trì sự logic và tính mạch lạc 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. Bằng cách tự động xem xét mức độ quan trọng của từng phần ở đầu vào, 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

Cách tiếp cận mới trong việc tích hợp BERT với khung tuần tự seq2seq của mô hình BLLA để tóm tắt các bài báo tiếng Việt mang lại lợi thế khác biệt so với các phương pháp truyền thống và thậm chí so với các mô hình khung tuần tự tiêu chuẩn không có BERT. Sự kết hợp này khai thác các khả năng của BERT để nâng cao hiệu suất và hiệu quả của việc tóm tắt văn bản.

Tăng cường hiểu biết theo ngữ cảnh

Khả năng tạo ra mã hoá theo ngữ cảnh sâu sắc của BERT làm cải tiến đáng kể sự hiểu biết của mô hình khung tuần tự phía sau về ngôn ngữ Tiếng Việt. Điều này dẫn đến các bản tóm tắt nắm bắt được ý nghĩa đa dạng của văn bản gốc. Các mô hình khung tuần tự hiện có thường sử dụng mô hình Word2Vec để tạo ra mã hoá cho văn bản đầu vào ở mức độ từ. Trong BLLA, BERT cũng được sử dụng ở vị trí tương tự, nhưng mô hình có được hiểu biết sâu sắc hơn nhờ khả năng hiểu ngôn ngữ đã được đào tạo trước của BERT, mỗi từ trong văn bản đầu vào sau khi được BERT xử lý sẽ cho ra một vectơ với 768 chiều, cao hơn rất nhiều so với các mô hình Word2Vec hiện tại, từ 100 đến 300 chiều. Nhờ đó BERT cung cấp cách trình bày toàn diện các từ trong ngữ cảnh, cải thiện khả năng của mô hình trong việc phân biệt và ưu tiên thông tin chính từ các bài báo để thực hiện tóm tắt.

Xử lý đặc điểm ngôn ngữ của Tiếng Việt

Tiếng Việt, với cấu trúc cú pháp phức tạp và hình thái từ ngữ phong phú, đặt ra những thách thức đặc biệt cho việc tóm tắt văn bản. Việc đào tạo trước của BERT về dữ liệu ngôn ngữ giúp mô hình xử lý tốt hơn những thách thức này, khiến nó đặc biệt phù hợp để tóm tắt tin tức tiếng Việt, trong đó việc hiểu chính xác các từ tới tận đặc điểm ngôn ngữ như vậy là rất quan trọng.

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

Bằng cách tích hợp BERT, mô hình tận dụng kho kiến thức được đào tạo trước khổng lồ, giảm nhu cầu về dữ liệu để đào tạo. Điều này không chỉ đơn giản hóa quá

trình đào tạo của mô hình mà còn nâng cao khả năng thích ứng của mô hình với các phong cách và lĩnh vực khác nhau của tin tức Việt Nam, giúp mô hình trở nên khác biệt so với các phương pháp đòi hỏi sự tùy biến đáng kể.

Việc kết hợp BERT vào khung tuần tự để giải quyết bài toán tóm tắt tin tức Tiếng Việt cho thấy tiềm năng lớn so với các phương pháp hiện có. Cách tiếp cận này tận dụng kiến thức của BERT có được trong việc đào tạo trước để hiểu và xử lý các mẫu ngôn ngữ phức tạp, nhờ vậy cải thiện đáng kể tính chính xác, mạch lạc và phù hợp của các bản tóm tắt tự động, 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

Việc triển khai và thử nghiệm mô hình BLLA để tóm tắt tin tức Tiếng Việt được hỗ trợ bởi một loạt công cụ, khung và mô hình đào tạo trước để mang lại hiệu quả trong việc xử lý độ phức tạp của Tiếng Việt và các nhiệm vụ học sâu, cũng như so sánh đánh giá hiệu quả của mô hình với các phương pháp hiện có, và với các biến thể của chính mô hình.

Ngôn ngữ lập trình và thư viện học máy

Python & Keras: Python là ngôn ngữ lập trình chính do nó hỗ trợ rộng rãi cho các tác vụ học máy và Xử lý ngôn ngữ tự nhiên. Keras, một thư viện mạng nơ ron được viết bằng Python chạy trên TensorFlow, được sử dụng vì tính đơn giản trong việc tạo các mô hình học máy, cho phép dễ dàng xây dựng, đào tạo và triển khai các mô hình học sâu.

Mô hình được đào tạo trước cho Tiếng Việt

PhoW2V: Mô hình Word2Vec được đào tạo trước này được huấn luyện dành riêng cho Tiếng Việt, bao gồm cả phần mã hoá âm tiết và từ, rất quan trọng để nắm bắt các sắc thái ngôn ngữ của Tiếng Việt.

PhoBERT: Tận dụng PhoBERT, một biến thể của BERT được đào tạo sẵn cho Tiếng Việt, giúp nâng cao hơn nữa khả năng hiểu và xử lý văn bản Tiếng Việt của mô hình. Khả năng hiểu Tiếng Việt sâu sắc của PhoBERT có được nhờ đào tạo trên một lượng văn bản Tiếng Việt lớn.

Mô hình đa ngôn ngữ BERT (BERT Multilingual Base Model): Mô hình này là một biến thể của BERT có hỗ trợ tiếng Việt, ngoài ra mô hình còn được đào tạo với lượng dữ liệu của 104 ngôn ngữ khác nhau trên thế giới.

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

VnCoreNLP: Bộ công cụ xử lý ngôn ngữ tự nhiên cho Tiếng Việt cung cấp các tiện ích cần thiết cho việc xử lý văn bản như phân đoạn từ, nhận dạng tên. Những công cụ này rất quan trọng trong việc chuẩn bị và xử lý dữ liệu văn bản Tiếng Việt phục vụ cho công việc đào tạo mô hình.

Môi trường phát triển, thực thi và thư viện

Google Colab, TensorFlow, Transformers: Sự kết hợp giữa Google Colab với TensorFlow và thư viện Transformers của Hugging Face cung cấp một môi trường mạnh mẽ, dễ sử dụng, dựa trên đám mây để đào tạo và thử nghiệm các mô hình học sâu. Khả năng truy cập của Colab vào TPU và GPU giúp tăng tốc tính toán, trong khi thư viện Transformers cung cấp một cách đơn giản để triển khai và tinh chỉnh các mô hình dựa trên BERT.

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) để tận dụng các nội dung mã hoá theo ngữ cảnh sâu sắc được tạo ra từ BERT nhờ đó có được nhiều thông tin hơn về ngôn ngữ nhằm phục vụ cho nhiệm vụ tóm tắt văn bản Tiếng Việt. Phần này đi sâu vào việc sử dụng BERT làm bộ mã hóa và vai trò quan trọng của nó trong việc cung cấp đầu ra là vectơ 768 chiều, đóng vai trò là đầu vào phong phú cho phần giải mã sử dụng LSTM trong mô hình khung tuần tự (seq2seq) tiếp theo.

BERT với tư cách là Bộ mã hóa:

Trọng tâm của giai đoạn đào tạo ban đầu của mô hình BLLA, BERT hoạt động như một bộ mã hóa, xử lý các bài báo tin tức Tiếng Việt để tạo ra các phần mã hoá

nắm bắt được nhiều sắc thái ngôn ngữ và mối quan hệ ngữ cảnh. Mỗi từ đầu vào được chuyển đổi thành một vectơ đa chiều, cung cấp một biểu diễn toàn diện của từ đó trong ngữ cảnh của văn bản đầu vào, một biểu diễn với nhiều thông tin hơn rất nhiều so với ban đầu.

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]. Kết quả mã hoá đầu ra với vectơ có rất nhiều chiều này giúp biến đổi đầu vào thành một bản mã hoá chứa đầy thông tin về ngữ cảnh xung quanh mỗi từ, tạo điều kiện cho mô hình có được hiểu biết sâu sắc về văn bản mà các phương pháp mã hoá khác không thể so sánh được. 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). Việc đưa trực tiếp các phần mã hoá theo ngữ cảnh của BERT vào khung tuần tự đảm bảo rằng mô hình có được nền tảng hiểu ngôn ngữ sâu sắc.

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 duy trì tính toàn vẹn của các kết quả đầu ra trong khi 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]. Cách tiếp cận này cho phép mô hình sẽ đào tạo phần khung tuần tự trước theo nhu cầu tóm tắt cụ thể của nhiệm vụ mà không ảnh hưởng đến thông tin ngữ cảnh phong phú mà BERT có được.

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** nhằm mục đích tinh chỉnh khả năng của mô hình nhằm tận dụng những hiểu biết sâu sắc về ngữ cảnh của BERT một cách hiệu quả. 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** giúp cho khung tuần tự có thể học nhanh hơn trên cơ sở đầu ra với cực kì nhiều thông tin có được từ BERT.

Việc sử dụng BERT trong mô hình BLLA làm bộ mã hóa trong giai đoạn huấn luyện ban đầu với các lớp bị đóng băng, nhằm mục tiêu khai thác các vectơ 768 chiều có được từ việc huấn luyện trước để cung cấp thông tin với nhiều hơn dữ liệu về ngữ cảnh và ngôn ngữ cho các phần xử lý khung tuần tự seq2seq. Chiến lược này là một cách tiếp cận đổi mới của mô hình để phù hợp với việc kết hợp và sử dụng một mô hình ngôn ngữ lớn đã được huấn luyện trước như BERT.

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

Trong thiết kế của BLLA, mô hình sử dụng khung tuần tự với LSTM là thành phần chính của bộ giải mã ngoài ra còn tích hợp cơ chế chú ý để tinh chỉnh đầu ra tóm tắt. Phần này đi sâu vào cấu hình bộ giải mã sử dụng LSTM và vai trò quan trọng của lớp chú ý trong việc cải thiện đầu ra bản tóm tắt.

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. Cấu hình này có vai trò then chốt để xử lý các vectơ mã hóa nhận được từ BERT, biến đổi các phần mã hoá theo ngữ cảnh phong phú thành các bản tóm tắt mạch lạc, ngắn gọn hơn.

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. Quá trình xử lý hai chiều này đảm bảo sự hiểu biết toàn diện về văn bản, thiết lập nền tảng vững chắc cho giai đoạn giải mã tiếp theo. 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 nhưng ở phiên bản đầu tiên của BLLA sử dụng ba lớp chú ý (để kiểm nghiệm khả năng của 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. Khả năng này cho phép bộ giải mã xem xét toàn bộ chuỗi đầu vào, cải thiện đáng kể mức độ liên quan và mạch lạc của các bản tóm tắt được tạo ra.

Cách lớp chú ý hoạt động: Lớp chú ý lấy đầu ra của bộ giải mã (decoder) làm truy vấn và đầu ra của bộ mã hóa (encoder) làm cả khóa và giá trị, tính toán điểm số để xác định trọng tâm trên các phần khác nhau của văn bản đầu vào. Quá trình này đảm bảo rằng mỗi từ trong phần tóm tắt được căn chỉnh theo ngữ cảnh với các phần phù hợp nhất của bài viết gốc. Lớp này sẽ sử dụng đầu ra của lớp BiLSTM đầu tiên thay vì của BERT để đảm bảo khớp về dữ liệu giữa bộ mã hoá và bộ giải mã.

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. Chiến lược này được sử dụng để tinh chỉnh sự tích hợp giữa hiểu biết theo ngữ cảnh của BERT và khả năng tạo ra các bản tóm tắt mạch lạc của mô hình khung tuần tự.

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. Thay đổi này cho phép các gradient chảy qua toàn bộ mô hình, bao gồm các lớp BERT, trong quá trình lan truyền ngược, cho phép tinh chỉnh các phần được đào tạo trước kết hợp với các thành phần của khung tuần tự với LSMT.

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 đã chứng tỏ giá trị trong việc hiểu ngữ cảnh tiếng Việt. Sự điều chỉnh này đảm bảo rằng mô hình được huấn luyện toàn bộ và tập trung duy nhất cho nhiệm vụ tóm tắt.

Trong mô hình BLLA, bộ giải mã LSTM, cùng với cơ chế chú ý và sự hiệu quả của BERT, đã kết hợp cùng việc huấn luyện với tốc độ học tập nhỏ trên toàn bộ mô hình ở giai đoạn hai, giúp cho mô hình có được khả năng tạo ra các bản tóm tắt hợp lý, ngắn gọn, mạch lạc chỉ với một lượng nhỏ tập dữ liệu học tập.

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

Để đánh giá mô hình, đề án đã xây dựng một bộ dữ liệu bao gồm nhiều bài báo Tiếng Việt, bằng việc sử dụng ba bộ dữ liệu về tin tức Tiếng Việt. Việc hợp nhất bộ ba tập dữ liệu này nhằm mục đích đem lại tính đa dạng và phức tạp của dữ liệu tin tức trong thế giới thực, cung cấp nền tảng vững chắc để đánh giá khả năng tóm tắt của mô hình.

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, VNDS đóng vai trò là bộ dữ liệu đầu tiên chuẩn mực cho bài toán tóm tắt tin tức Tiếng Việt. 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. Chưa qua quá trình xử lý

trước, nó là một bộ dữ liệu thô, chân thực về nội dung tin tức, trải dài trên nhiều chủ đề và thể loại khác nhau, phù hợp để thử thách khả năng thích ứng và tính chính xác trong kết quả tóm tắt của mô hình BLLA.

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. Nó tiếp tục đa dạng hóa tài liệu đào tạo và thử nghiệm của mô hình BLLA.

Các bước tiền xử lý chi tiết

Bộ ba tập dữ liệu về tin tức sẽ được xử lý để tạo ra tập dữ liệu huấn luyện và kiểm tra cho BLLA:

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 (Chữ ký, tiêu đề ảnh, v.v), 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]:

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

Tính toán Cosine Similarity với BERT

Tạo mã hoá cho câu: Đối với mỗi câu và tiêu đề, các đề án sử dụng BERT để tạo ra vectơ mã hoá cho câu và tiêu đề.

Tính toán độ tương tự của các câu với tiêu đề: Độ tương tự cosine giữa vectơ của tiêu đề và vectơ của mỗi câu được tính toán. Điểm này cho biết mức độ liên quan của câu với tiêu đề, giả định rằng các câu có độ tương tự cao hơn có nhiều

khả năng chứa thông tin chính liên quan đến chủ đề chính của bài viết. Điều này là khả thi vì khả năng hiểu và tạo ra bản mã hoá chứa thông tin ngữ cảnh sâu sắc của BERT.

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 nhất 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. Sự phân chia này rất quan trọng để đánh giá khả năng khái quát hóa của mô hình trên dữ liệu chưa được nhìn thấy trong quá trình học tập.

3.3.2 Cài đặt thử nghiệm

Thiết lập thử nghiệm để đánh giá mô hình BLLA bao gồm cấu hình chi tiết của các thành phần trong mô hình, cài đặt chiến lược để đào tạo và thử nghiệm cũng như tận dụng các tài nguyên tính toán hiệu suất cao. Thiết lập này được thiết kế để đánh giá khả năng tóm tắt của mô hình trên các cấu hình khác nhau.

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. Dưới đây là một mẫu được dùng trong bộ huấn luyện.

Nội dung:

“Để lan_toả tinh_thần uplift (tạm dịch là tinh_thần hứng_khởi) trong giờ nghỉ , Coca-Cola và VNG đã phối_hợp cùng nhau tổ_chức sự_kiện Watch_Party livestream trận chung_kết giải đấu Icons_Global_Championships_Tốc Chiến 2022 tại khách_sạn Kim Đô_Linh_Ngọc Đàm và Miss_Thy chọn Coca-Cola cho giờ chơi game thêm phần hứng_khởi Tại sự_kiện , người tham_dự đã có dịp gặp_gỡ Team_Flash - một đội_tuyển thể_thao điện_tử chuyên_nghiệp - về hành_trình của họ tại hai giải đấu lớn là SEA_Games và ICONS Trong thời_gian diễn ra sự_kiện , ba

thành_viên của Team_Flash là Coyote , Elly , Shy đã có màn giao_lưu và trả_lời các câu hỏi vô_cùng thú_vị từ các fan hâm_mộ bộ_môn Liên_Minh Huyền_Thoại.”

Bản tóm tắt:

“<START> Coca-Cola cùng fan esport hào_hứng cổ_vũ chung_kết thế_giới Tốc_Chiến_Icons_Global Championship 2022 <END>”

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 35GB RAM, cung cấp tài nguyên cần thiết để đào tạo một cách hiệu quả, trong thời gian tốt nhất. Việc sử dụng TPU cho phép lặp lại và thử nghiệm mô hình nhanh hơn, giảm đáng kể thời gian huấn luyện bất chấp độ phức tạp của mô hình và kích thước của tập dữ liệu.

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]. Thước đo được công nhận rộng rãi này trong lĩnh vực xử lý ngôn ngữ tự nhiên cung cấp một phương tiện định lượng để đánh giá chất lượng văn bản được tạo ra bởi các mô hình tóm tắt và dịch máy so với một tập hợp các văn bản tham chiếu.

Tìm hiểu về điểm BLEU

Điểm BLEU đánh giá sự giống nhau giữa văn bản do máy tạo ra và văn bản tham chiếu do con người viết, tập trung vào sự khớp nhau của n-gram trong văn bản được tạo

ra. Điểm số nằm trong khoảng từ 0 đến 1 (hoặc 0 đến 100 khi được biểu thị bằng phần trăm), trong đó điểm càng gần 1 cho thấy mức độ tương đồng cao giữa văn bản do máy tạo ra với văn bản tham chiếu [20]. Điểm BLEU xem xét một số khía cạnh:

Độ trùng lặp của N-gram: BLEU kiểm tra sự trùng khớp của n-gram (từ một đến bốn) giữa văn bản được tạo và văn bản tham chiếu. Tính toán tỷ lệ so khớp n-gram trong văn bản được tạo với tổng số n-gram, điều chỉnh độ khớp unigram để xử lý vấn đề khi đầu ra quá ngắn [20].

Điều chỉnh với đầu ra quá ngắn: Để xử lý vấn đề kết quả tạo ra quá ngắn dẫn đến đánh giá không chính xác, BLEU điều chỉnh điểm số với văn bản đầu ra ngắn. Điều chỉnh này được áp dụng nếu văn bản được tạo ra ngắn hơn văn bản tham chiếu, đảm bảo rằng tính ngắn gọn không ảnh hưởng quá mức đến điểm số [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). Bộ dữ liệu này cũng được xử lý các bước như với tập huấn luyện, mỗi Tiêu Đề trong bộ dữ liệu này chính là bản tóm tắt tham chiếu do con người tạo ra dùng để đánh giá với bản do mô hình sinh ra.

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, cả hai đều cùng kiến trúc chung, nhưng khác nhau ở phiên bản 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. Việc đánh giá này nhằm mục đích phân biệt tác động của mô hình đào tạo trước tập trung cho một ngôn ngữ và đa ngôn ngữ đối với bài toán tóm tắt.

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. Các bản tóm tắt trong 200 mẫu này sẽ được dùng là bản tóm tắt tham chiếu do con người tạo ra, để so sánh với bản do mô hình sinh ra.

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 này liệt kê tổng số tham số cho từng mô hình, cho thấy tài nguyên tính toán và độ sâu của kiến trúc.

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

Mô hình BLLA - BERT Base có số lượng tham số cao nhất, nằm ở BERT base, điều này là do BERT base được huấn luyện trên 104 ngôn ngữ khác nhau. Biến thể phoBERT đơn giản hơn nhưng được tối ưu hóa đặc biệt cho Tiếng Việt, mang lại sự cân bằng giữa độ phức tạp của mô hình và hiệu năng. Mô hình khung tuần tự LSTM-LSTM truyền thống nhẹ hơn đáng kể, dù hiệu năng chưa được tốt như hai phiên bản BLLA sử dụng BERT nhưng sẽ phù hợp cho các giải pháp yêu cầu bộ nhớ thấp và tốc độ xử lý nhanh, ngoài ra nó cũng giúp cho mô hình dễ đào tạo hơn, không đòi hỏi nhiều phần cứng.

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.

Điểm BLEU sẽ là từ 0 đến 1 hoặc từ 0 đến 100 nếu quy đổi theo phần trăm. 100% nghĩa là bản tóm tắt do mô hình tạo ra sẽ chính xác nhất với bản tham chiếu do con người tạo ra [20]. 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, từ đó giúp kết quả tóm tắt chính xác và đúng đắn hơn hẳn so với phiên bản đào tạo trên 104 ngôn ngữ chung như BERT base. Điểm số giảm dần từ BLEU-1 xuống BLEU-4 cho cả hai mô hình, điều này là hợp lý vì khi độ dài n-gram tăng lên, sẽ khiến cho khả năng trùng khớp ít hơn, các mô hình đã sử dụng nhiều từ mới hơn cho các bản tóm tắt của mình thay vì sử dụng các từ trích chọn từ nội dung.

Mô hình BLLA sử dụng BERT base, mặc dù không tốt như phiên bản sử dụng phoBERT, nhưng vẫn 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.

Đánh giá vai trò các khối trong mô hình BLLA

Từ những kết quả phân tích hiệu suất các biến thể của mô hình BLLA, và với mô hình khung tuần tự truyền thống, thấy rằng sự khác biệt về hiệu suất của các mô hình nằm chính quan trọng ở khối BERT, đại diện cho khối mã hoá đầu tiên để chuyển đổi các từ đầu vào thành các mã hoá 768 chiều, các mã hoá này càng chính xác hay mô hình càng hiểu và nắm bắt được ngữ nghĩa của văn bản đầu vào thì kết quả sinh ra ở bộ giải mã phía sau càng tốt.

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

Nội dung 1:

“Cục Hàng không VN (CAAV) vừa có công_văn yêu_cầu các cơ_quan , đơn_vị trong ngành hàng_không dân_dụng nắm_bắt thông_tin , theo_dõi chặt_chẽ diễn_biến của áp_thấp_nhiệt_đới / bão để chủ_động có biện_pháp ứng_phó kịp_thời , đảm_bảo an_toàn tuyệt_đối hoạt_động bay . + Triển_khai phương_án đối_phó với mọi dự_báo diễn_biến của áp_thấp_nhiệt_đới / bão . Đồng_thời , yêu_cầu các hãng hàng_không theo_dõi chặt_chẽ diễn_biến của áp_thấp_nhiệt_đới / bão để điều_chỉnh kế_hoạch bay , thay_đổi lịch bay cho phù_hợp và đảm_bảo an_toàn tuyệt_đối hoạt_động bay.”

Bản tóm tắt do người tạo:

“Cục Hàng không triển khai đối phó áp thấp nhiệt đới”

Bản tóm tắt do mô hình sinh ra:

“Cục Hàng không không chủ quan với áp thấp nhiệt đới”

Nội dung 2:

“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

Chương này đề án trình bày giải pháp và khung thử nghiệm, từ thu thập dữ liệu đến đo lường đánh giá các phiên bản khác nhau của mô hình. Kết quả cho thấy tính hiệu quả của mô hình BLLA hay BERT-LSMT-LSTM với cơ chế chú ý, trong việc tạo ra các bản tóm tắt chính xác và ngắn gọn cho các bài báo tiếng Việt. Những phát hiện này không chỉ khẳng định phương pháp đề xuất là đúng hướng, mà còn mở ra nhiều hướng nghiên cứu trong tương lai về tóm tắt văn bản 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.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Nguyễn Ngọc Điệp, Nguyễn Thị Thanh Thủy (2022), "Sử dụng BERT và câu phụ trợ cho trích xuất khía cạnh trong văn bản tiếng Việt." *Học Viện Công Nghệ Bưu Chính Viễn Thông.* Vol. 1 No. 4 (2022): *Journal of Science and Technology on Information and Communications*, Computer Science section.

[2] Nguyễn Văn Hậu, Nguyễn Thành Chinh và Nguyễn Minh Tiến, Nguyễn Xuân Hoài. (2019), "VNDS: A Vietnamese Dataset for Summarization." 375-380. 10.1109/NICS48868.2019.9023886.

[3] Lê Thanh Hương và Lê Mạnh Tiến (2013), "An approach to abstractive text summarization." *2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, 371-376.

[4] Nguyễn Diệu Linh (2021), *Phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT*, Học Viện Công Nghệ Bưu Chính Viễn thông, Hà Nội 74.

[5] Lâm Quang Tường, Phạm Thế Phi và Đỗ Đức Hào (2017), "Tóm tắt văn bản tiếng Việt tự động với mô hình Sequence-to-Sequence." *Tạp chí Khoa học Đại học Cần Thơ. CĐ Công nghệ TT* (tháng 10 2017), 125-132. DOI:https://doi.org/10.22144/ctu.jsi.2017.017.

[6] Đỗ Thị Thu Trang, Trịnh Thị Nhị, Ngô Thanh Huyền (2020), "Sử dụng BERT cho tóm tắt trích rút văn bản." *UTEHY Journal of Science and Technology*, 26, (Sep. 2020), 74-79.

Tiếng Anh

[7] Apallius de Vos, Isa M., Ghislaine L. van den Boogerd, Mara D. Fennema, and Adriana D. Correia (2022), "Comparing in context: Improving cosine similarity measures with a metric tensor." arXiv:2203.14996 [cs.CL].

[8] Bengio, Samy, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer (2015), "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks." arXiv:1506.03099 [cs.LG].

[9] Das, Dipanjan, and André Martins (2007), "A survey on automatic text summarization."

[10] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019),

"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv:1810.04805 [cs.CL].

[11] Jiang, Dan, Shaozhong Cao, and Shulin Yang (2021), "Abstractive summarization of long texts based on BERT and sequence-to-sequence model." *In 2021 2nd International Conference on Information Science and Education (ICISE-IE)*, 460-466. DOI: 10.1109/ICISE-IE53922.2021.00112.

[12] Kupiec, Julian, Jan O. Pedersen, and Francine R. Chen (1995), "A trainable document summarizer." *In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[13] Li, Yuanyuan, Yuan Huang, Weijian Huang, Junhao Yu, and Zheng Huang (2023), "An Abstractive Summarization Model Based on Joint-Attention Mechanism and a Priori Knowledge." *Applied Sciences*, 13(7): 4610. DOI: 10.3390/app13074610.

[14] Liu, Yang, and Mirella Lapata (2019), "Text Summarization with Pretrained Encoders." arXiv:1908.08345 [cs.CL].

[15] Luhn, H. P. (1958), "A Business Intelligence System." *IBM Journal of Research and Development*, 2(4): 314-319. DOI: 10.1147/rd.24.0314.

[16] Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, và Luke Zettlemoyer (2018), "Deep contextualized word representations." arXiv:1802.05365 [cs.CL].

[17] Nallapati, Ramesh, Bowen Zhou, Cicero Nogueira dos Santos, Caglar Gulcehre, and Bing Xiang (2016), "Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond." arXiv:1602.06023 [cs.CL].

[18] Hochreiter, Sepp and Schmidhuber, Jürgen (1997), "Long Short-term Memory." *Neural Computation*, vol. 9, no. 8, pp. 1735-80. DOI: 10.1162/neco.1997.9.8.1735.

[19] Gers, Felix, Schmidhuber, Jürgen, and Cummins, Fred (2000), "Learning to Forget: Continual Prediction with LSTM." *Neural Computation*, vol. 12, pp. 2451-71. DOI: 10.1162/089976600300015015.

[20] Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei Jing (2002), "BLEU: a Method for Automatic Evaluation of Machine Translation." DOI:

10.3115/1073083.1073135.

[21] Santhanam, Sivasurya (2020), "Context based Text-generation using LSTM networks." arXiv:2005.00048 [cs.CL].

[22] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014), "Sequence to Sequence Learning with Neural Networks." arXiv:1409.3215 [cs.CL].

[23] Sun, Xiaofei, Yuxian Meng, Xiang Ao, Fei Wu, Tianwei Zhang, Jiwei Li, and Chun Fan (2022), "Sentence Similarity Based on Contexts." arXiv:2105.07623 [cs.CL].

[24] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2023), "Attention Is All You Need." arXiv:1706.03762 [cs.CL].

[25] Zhang, Haoyu, Jingjing Cai, Jianjun Xu, and Ji Wang (2019), "Pretraining-Based Natural Language Generation for Text Summarization." *In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 789-797. Hong Kong, China: Association for Computational Linguistics. DOI: 10.18653/v1/K19-1074.

[26] Zhou, Qingyu, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao (2018), "Neural Document Summarization by Jointly Learning to Score and Select Sentences." arXiv:1807.02305 [cs.CL].

Tham khảo từ Internet

[27] https://phamdinhkhanh.github.io/2020/05/23/BERTModel.html, truy nhập ngày 20/12/2023

[28] https://trituenhantao.io/cac-dataset-tieng-viet/, truy nhập ngày 22/12/2023

[29] https://www.kaggle.com/datasets/phamtheds/news-dataset-vietnameses, truy nhập ngày 24/12/2023

[30] https://www.kaggle.com/datasets/haitranquangofficial/vietnamese-online-news-dataset, truy nhập ngày 24/12/2023