HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Viengnakhone Seesamoud

NGHIÊN CỨU CÁC PHƯƠNG PHÁP TÓM TẮT VĂN BẢN VÀ THỬ NGHIỆM VỚI DỮ LIỆU TIẾNG LÀO

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(*Theo định hướng ứng dụng*)

HÀ NỘI – NĂM 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Viengnakhone Seesamoud

NGHIÊN CỨU PHƯƠNG PHÁP TÓM TẮT VĂN BẢN VÀ THỬ NGHIỆM VỚI DỮ LIỆU TIẾNG LÀO

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.TSKH. HOÀNG ĐĂNG HẢI

HÀ NỘI - NĂM 2025

LỜI CAM ĐOAN

Tôi tên là Viengnakhone Seesamoud, là học viên chuyên ngành Khoa học máy tính, khóa 23. Tôi xin cam đoan đây là công trình nghiên cứu khoa học của cá nhân tôi, được sự hướng dẫn của PGS.TSKH. Hoàng Đăng Hải.

Các thông tin được sử dụng tham khảo trong đề án tốt nghiệp được thu thập từ các nguồn có độ tin cậy, đã được kiểm chứng, được công bố rộng rãi và tất cả các thông tin trích dẫn đã được tôi bổ sung nguồn gốc rõ ràng ở phần Tài liệu tham khảo. Các số liệu, kết quả nghiên cứu được trình bày trong đề án tốt nghiệp này là do chính tôi thực hiện một cách nghiêm túc, trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin lấy danh dự và uy tín của bản thân để đảm bảo cho lời cam đoan này.

Hà Nội, ngày 29 tháng 07 năm 2025

Người hướng dẫn	Tác giả thực hiện
(ký tên)	*(ký tên)*
PGS.TSKH. Hoàng Đăng Hải	Viengnakhone Seesamoud

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
AI	Artificial Intelligence	Trí tuệ nhân tạo
ANN	Artificial Neural netmork	Mạng Nơ-ron nhân tạo
BART	Bidirectional and Auto-Regressive Transformer	
BERT	Bidirectional Encoder Representations from Transformers	
CNNs	Convolutional Neural Networks - CNNs	Mạng nơ-ron tích chập
DL	Deep Learning	Học sâu
GPT	Generative Pre-trained Transformer	
IR	Information Retrieval	Tìm kiếm thông tin
LLMs	Large Language Models	Mô hình ngôn ngữ lớn
ML	Machine Learning	Học máy
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
PEGASUS	Pre-training with Gap-sentences for Abstractive Summarization	
T5	Text-to-Text Transfer Transformer	
TTVB		Tóm tắt văn bản

DANH MỤC BẢNG

DANH MỤC HÌNH

MỞ ĐẦU

Trong bối cảnh hội nhập khu vực Đông Nam Á, Việt Nam và Lào luôn duy trì mối quan hệ hữu nghị, hợp tác toàn diện trên nhiều lĩnh vực, bao gồm kinh tế, văn hóa, giáo dục và chính trị. Tuy nhiên, dù có chung đường biên giới và mối quan hệ truyền thống lâu đời, rào cản ngôn ngữ giữa hai quốc gia vẫn là một trở ngại lớn trong việc giao lưu và trao đổi. Việc thiếu các công cụ hỗ trợ trong xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) giữa tiếng Việt và tiếng Lào đã gây khó khăn cho cả việc giao tiếp hằng ngày, nghiên cứu học thuật và phát triển kinh tế - xã hội chung.

Một trong số những ứng dụng NLP có ý nghĩa thiết thực đối với nhiều lĩnh vực trong đời sống là tóm tắt văn bản (TTVB). Với lượng dữ liệu văn bản ngày càng nhiều trên mạng Internet, việc truy xuất thông tin từ lượng dữ liệu khổng lồ này đặt ra những yêu cầu cấp thiết về việc nghiên cứu và xây dựng các giải pháp TTVB, giúp người dùng nhanh chóng nắm bắt được thông tin kịp thời. Các nghiên cứu gần đây tập trung vào ứng dụng phương pháp NLP trong nhiều lĩnh vực, góp phần nâng cao hiệu quả trong TTVB. Tuy nhiên, việc nghiên cứu áp dụng các phương pháp này vào bài toán tóm tắt tiếng Việt và tiếng Lào còn hạn chế. Các tập dữ liệu mẫu dùng cho huấn luyện còn chưa đầy đủ. Còn khá ít các nghiên cứu đánh giá về tính hiệu quả của các phương pháp áp dụng cho các đặc thù của ngôn ngữ, điển hình như tiếng Việt và tiếng Lào.

Số lượng các nghiên cứu về lĩnh vực NLP tăng đáng kể, chứng tỏ chủ đề này thực sự rất có tiềm năng và đáng được quan tâm nghiên cứu. Hơn nữa, khi thực hiện học tập nghiên cứu tại Việt Nam bằng tiếng Việt, tôi nhận thấy các phương pháp TTVB rất hữu ích trong việc hỗ trợ tìm kiếm thông tin cô đọng từ các văn bản một cách nhanh chóng, hiệu quả. Chính vì vậy, tôi chọn đề tài ***"Nghiên cứu các phương pháp tóm tắt văn bản và thử nghiệm với dữ liệu tiếng Lào"*** làm đề án tốt nghiệp.

Mục tiêu đề án tốt nghiệp mong muốn đạt được là tìm hiểu, khảo sát các phương pháp NLP hiện đại ứng dụng trong bài toán tóm tắt văn bản, thực hiện một số thử nghiệm với các tập dữ liệu hiện có về tiếng Việt và tiếng Lào. Qua đó có thể đánh giá mức độ thực hiện của các phương pháp đối với các đặc thù của ngôn ngữ tiếng Việt

và đặc biệt là ngôn ngữ tiếng Lào của quê hương tôi. Kết quả này có thể được áp dụng trong tương lai là hướng tới việc phát triển một hệ thống dịch tự động giữa tiếng Việt và tiếng Lào, từ đó hỗ trợ việc giao tiếp, hợp tác và trao đổi thông tin giữa hai quốc gia.

Bố cục đề án tốt nghiệp ngoài phần mở đầu và kết luận có 03 chương chính, như sau:

Chương 1. Tổng quan về vấn đề tóm tắt văn bản: cơ sở lý thuyết về các vấn đề Xử lý ngôn ngữ tự nhiên và các ứng dụng của chúng; Bài toán tóm tắt văn bản và ý nghĩa của tóm tắt văn bản; đặc điểm của ngôn ngữ tiếng Việt và tiếng Lào trong bài toán tóm tắt văn bản; Trình bày vấn đề nghiên cứu đặt ra trong bài.

Chương 2. Khảo sát, đánh giá các phương pháp tóm tắt văn bản sử dụng NLP: phân tích một số mô hình NLP hiện đại cho tóm tắt văn bản; Khảo sát và đánh giá các phương pháp tóm tắt văn bản sử dụng NLP; Phân tích đánh giá một số mô hình hỗ trợ tóm tắt văn bản đa ngôn ngữ sử dụng vào tóm tắt văn bản; Phân tích các phương pháp tạo lập tập dữ liệu (Dataset) cho bài toán tóm tắt văn bản; Trình bày một số phương pháp đánh giá cho hệ thống tóm tắt văn bản; Đề xuất mô hình thử nghiệm cho bài toán tóm tắt văn bản tiếng Việt và tiếng Lào.

Chương 3. Thử nghiệm mô hình tóm tắt văn bản với tiếng Việt và tiếng Lào: đề xuất môi trường thử nghiệm tóm tắt văn bản tiếng Việt và tiếng Lào bao gồm các nội dung: Thiết lập môi trường thử nghiệm với công cụ phần mềm, thư viện hỗ trợ, thiết bị phần cứng, quy trình thử nghiệm; Tạo lập các tập dữ liệu thử nghiệm từ các bộ dữ liệu *VietNews* và *LaoNews Classification*; xây dựng mô hình và huấn luyện mô hình; Đánh giá hiệu năng mô hình; Thảo luận và đánh giá kết quả thử nghiệm.

CHƯƠNG 1. TỔNG QUAN VỀ VẤN ĐỀ TÓM TẮT VĂN BẢN

1.1 Xử lý ngôn ngữ tự nhiên và các ứng dụng

1.1.1 Khái niệm về xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) là một lĩnh vực của trí tuệ nhân tạo (Artificial Intelligence - AI) tập trung vào việc giúp máy tính hiểu, diễn giải và tạo ra ngôn ngữ của con người một cách tự nhiên [1]. Nhờ sự phát triển của các mô hình học sâu và dữ liệu lớn, NLP đã đạt được những bước tiến vượt bậc, tạo ra nhiều ứng dụng hữu ích trong đời sống và công việc. NLP kết hợp nhiều kỹ thuật từ lĩnh vực ngôn ngữ học tính toán, học máy (Machine Learning - ML) và học sâu (Deep Learning - DL) để xử lý văn bản và giọng nói, cho phép máy tính thực hiện các tác vụ như phân loại văn bản, phân tích cảm xúc, dịch máy và trích xuất thông tin [1].

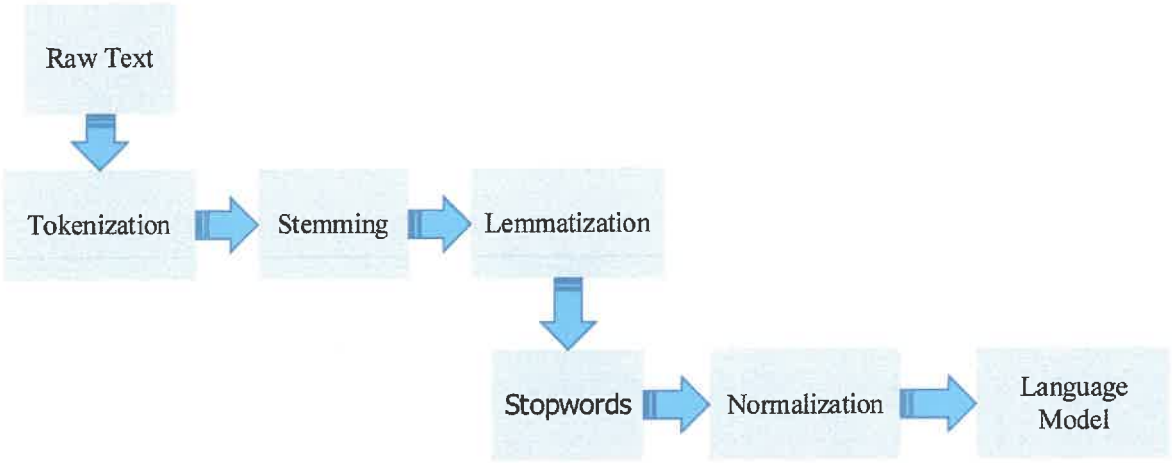
Về mặt kỹ thuật, NLP bao gồm hai giai đoạn chính: tiền xử lý dữ liệu và mô hình hóa ngôn ngữ. Tiền xử lý dữ liệu bao gồm các bước như tokenization (tách từ), stemming (rút gọn từ), lemmatization (chuẩn hóa từ) và loại bỏ stopwords nhằm chuẩn hóa dữ liệu đầu vào, giúp mô hình dễ dàng xử lý hơn [2]. Giai đoạn mô hình hóa ngôn ngữ sử dụng các phương pháp từ truyền thống như mô hình n-gram, Hidden Markov Models (HMMs) đến các mô hình hiện đại dựa trên mạng nơ-ron như Recurrent Neural Networks (RNNs) và Transformers [3].

Dữ liệu thô (Raw text) được thu thập từ các nguồn khác nhau, ví dụ như mạng xã hội, email, trang Web,... Dữ liệu thô thường không có cấu trúc và không sắp xếp. Bước tiền xử lý thực hiện làm sạch, chuẩn hóa dữ liệu để cung cấp cho mô hình ngôn ngữ.

Tokenization là phân tách văn bản thành những đơn vị nhỏ - gọi là thẻ (token). Ví dụ văn bản "Hà Nội có từ bao giờ" được tách thành từng từ như sau: ["Hà", "Nội", "ơi", "có", "từ", "bao", "giờ"].

Stemming là bước rút gọn một từ về dạng gốc của nó. Ví dụ "Studies" → "Studi", "Giving" → "Giv", "Buying" → "Buy". Stemming có thể rút gọn dẫn đến sai như ví dụ trên. Do đó, cần thêm bước Lemmatization.

Lemmatization là quá trình rút gọn một từ thành dạng cơ sở có xét đến ngữ cảnh và ngữ nghĩa của nó. Ví dụ: "Studies" → "Study", "Giving" → "Give", "Buying" → "Buy".



Hình 1.1: Các giai đoạn chính của NLP

Stopwords là bước xóa bỏ các từ dừng không quan trọng, ít ý nghĩa, ví dụ các từ thông dụng như "the", "and".

Normalization là bước chuẩn hóa văn bản, nhằm loại bỏ nhiễu và đồng nhất từ ngữ trước khi đưa vào mô hình ngôn ngữ. Một số kỹ thuật chuẩn hóa thường sử dụng như: chuyển tất cả văn bản về chữ thường, xóa dấu câu, các biểu tượng, các đường link; xóa khoảng trăng dư thừa, loại bỏ ký tự lặp do lỗi.

Nhờ những tiến bộ trong NLP, các ứng dụng thực tế như chatbot, trợ lý ảo, hệ thống hỏi đáp và phân tích văn bản tự động ngày càng trở nên phổ biến. Các nghiên cứu gần đây tập trung vào việc cải thiện độ chính xác của mô hình NLP bằng cách sử dụng các mô hình ngôn ngữ lớn (Large Language Models - LLMs) như BERT (Bidirectional Encoder Representations from Transformers) và GPT (Generative Pre-trained Transformer) [4], [5]. Tuy nhiên, NLP vẫn đối mặt với nhiều thách thức như xử lý ngữ cảnh phức tạp, tính mơ hồ của ngôn ngữ tự nhiên và khả năng tổng quát hóa trên nhiều ngôn ngữ khác nhau [6].

1.1.2 Vai trò của NLP trong khai thác thông tin

NLP đóng vai trò quan trọng trong khai thác thông tin (Information Retrieval - IR) bằng cách cải thiện khả năng truy vấn, trích xuất và phân tích dữ liệu văn bản từ nhiều nguồn khác nhau [7]. Trong bối cảnh dữ liệu phi cấu trúc chiếm phần lớn nội dung trên Internet, các mô hình NLP hiện đại giúp tăng cường khả năng tìm kiếm thông tin có liên quan bằng cách xử lý ngữ cảnh, đồng nghĩa và quan hệ ngữ nghĩa giữa các từ [8].

Một trong những ứng dụng quan trọng của NLP trong IR là tìm kiếm ngữ nghĩa, trong đó các mô hình nhúng từ như Word2Vec [9], GloVe [10] và BERT [4] giúp cải thiện độ chính xác của kết quả tìm kiếm bằng cách hiểu ngữ cảnh thay vì chỉ so khớp từ khóa đơn thuần. Ngoài ra, các mô hình NLP còn hỗ trợ TTVB tự động để rút gọn thông tin từ các tài liệu lớn, giúp người dùng nhanh chóng tiếp cận nội dung quan trọng [11].

Bên cạnh đó, NLP còn được ứng dụng trong trích xuất thực thể có tên (Named Entity Recognition - NER) nhằm xác định các thực thể như tên người, tổ chức, địa điểm từ văn bản, hỗ trợ hệ thống tìm kiếm và phân loại thông tin hiệu quả hơn [12]. Ngoài ra, công nghệ hỏi đáp tự động sử dụng NLP để cung cấp câu trả lời chính xác từ tập dữ liệu lớn, đóng vai trò quan trọng trong các trợ lý ảo và hệ thống hỗ trợ khách hàng [13].

Nhìn chung, NLP không chỉ giúp tối ưu hóa quá trình khai thác thông tin mà còn thúc đẩy khả năng hiểu và tương tác với dữ liệu văn bản theo cách thông minh hơn, đóng góp lớn vào sự phát triển của các hệ thống thông tin hiện đại.

1.1.3 Ứng dụng của NLP

NLP đóng vai trò quan trọng trong nhiều lĩnh vực của công nghệ thông tin, từ hệ thống tìm kiếm thông tin, phân tích cảm xúc, đến dịch máy và chatbot [8]. Sự phát triển của các mô hình học sâu đã thúc đẩy NLP đạt được những bước tiến lớn trong việc hiểu và sinh ngôn ngữ tự nhiên, qua đó nâng cao hiệu quả của các hệ thống xử lý dữ liệu văn bản. Một số ứng dụng tiêu biểu của NLP gồm:

- **Dịch máy tự động:** Một trong những ứng dụng quan trọng nhất của NLP là dịch máy tự động, điển hình như Google Translate và DeepL. Các mô hình tiên tiến như Transformer và BERT [4] đã giúp cải thiện đáng kể chất lượng dịch thuật bằng cách hiểu ngữ cảnh thay vì chỉ dịch theo từng từ đơn lẻ.
- **Phân tích phản hồi của khách hàng**: NLP đóng vai trò then chốt trong phân tích cảm xúc, hỗ trợ doanh nghiệp đánh giá phản hồi của khách hàng thông qua việc phân tích các bình luận, đánh giá sản phẩm trên mạng xã hội và các nền tảng thương mại điện tử [14]. Những mô hình học sâu như RNNs và Long Short-Term Memory (LSTM) được ứng dụng rộng rãi trong lĩnh vực này [15].
- **Trợ lý ảo và Chatbot:** NLP còn được ứng dụng trong phát triển trợ lý ảo và chatbot, tiêu biểu như Siri, Google Assistant và ChatGPT, giúp tự động hóa tương tác với người dùng bằng ngôn ngữ tự nhiên. Đặc biệt, các mô hình GPT đã nâng cao đáng kể khả năng sinh ngôn ngữ và hiểu ngữ cảnh trong hội thoại [5].
- **Nhận dạng thực thể, tìm kiếm, phân loại thông tin:** NLP cũng đóng vai trò quan trọng trong việc trích xuất thông tin và NER, phục vụ cho các bài toán tìm kiếm và phân loại dữ liệu. NLP giúp xác định các thực thể như tên người, địa điểm, tổ chức từ các văn bản, hỗ trợ tìm kiếm và phân loại thông tin hiệu quả hơn [12].
- **Trích xuất thông tin y tế:** Trong lĩnh vực y tế, NLP hỗ trợ phân tích hồ sơ bệnh án bằng cách tự động trích xuất thông tin y khoa từ văn bản phi cấu trúc, giúp các chuyên gia y tế đưa ra quyết định nhanh hơn và chính xác hơn [16].

Nhìn chung, NLP không chỉ giúp cải thiện khả năng xử lý ngôn ngữ trong các hệ thống máy tính mà còn đóng góp lớn vào nhiều ngành công nghiệp, từ tài chính, y tế, giáo dục đến thương mại điện tử, thúc đẩy quá trình tự động hóa và tối ưu hóa hiệu suất làm việc.

1.1.4 Xu hướng phát triển của NLP

Các xu hướng hiện tại của NLP tập trung vào việc cải thiện khả năng hiểu ngôn ngữ tự nhiên, giảm thiểu sự phụ thuộc vào dữ liệu gán nhãn và mở rộng ứng dụng trong nhiều lĩnh vực khác nhau [8].

Một trong những bước ngoặt quan trọng của NLP là sự phát triển của các LLMs. Các mô hình như GPT- 4 [17] và PaLM [18] sử dụng hàng trăm tỷ tham số để nâng cao khả năng sinh văn bản, trả lời câu hỏi và thực hiện các tác vụ NLP phức tạp. Những mô hình này đã cải thiện đáng kể độ chính xác của NLP trong nhiều ứng dụng thực tế, từ chatbot thông minh đến hệ thống tìm kiếm nâng cao.

Bên cạnh đó, học ít dữ liệu và học không cần dữ liệu đang trở thành hướng tiếp cận quan trọng nhằm giảm thiểu sự phụ thuộc vào dữ liệu gán nhãn. Các mô hình như GPT- 4 và T5 [19] có khả năng thực hiện các tác vụ NLP mà không cần đào tạo lại trên tập dữ liệu cụ thể, giúp tiết kiệm tài nguyên tính toán và tăng tính linh hoạt.

Một xu hướng khác là tích hợp đa phương thức, trong đó NLP được kết hợp với xử lý hình ảnh và âm thanh để hiểu nội dung đa phương tiện tốt hơn. Các mô hình như CLIP và DALL·E [20] có thể phân tích cả văn bản và hình ảnh, mở ra tiềm năng mới cho các ứng dụng trong thương mại điện tử, giáo dục và y tế.

Cuối cùng, AI đạo đức và NLP có trách nhiệm ngày càng được quan tâm để giảm thiểu thiên vị trong các mô hình AI và đảm bảo rằng NLP được sử dụng một cách công bằng, minh bạch [21]. Các nghiên cứu đang tập trung vào việc phát triển các phương pháp làm sạch dữ liệu huấn luyện và điều chỉnh thuật toán để giảm thiên vị trong các hệ thống NLP.

Nhìn chung, NLP đang không ngừng phát triển theo hướng ngày càng mạnh mẽ, thông minh và được ứng dụng rộng rãi trong nhiều lĩnh vực. Các tiến bộ trong LLMs, học với lượng dữ liệu hạn chế, NLP đa phương thức và cá nhân hóa được kỳ vọng sẽ định hình tương lai của công nghệ này.

1.2 Bài toán tóm tắt văn bản

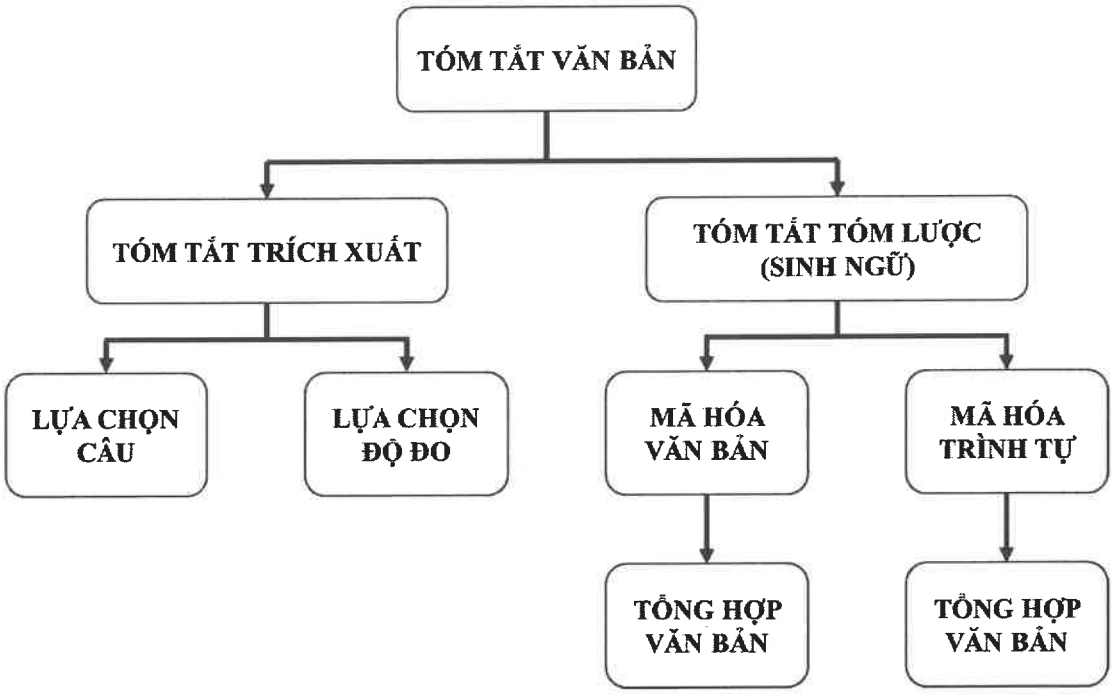
1.2.1 Khái quát về vấn đề tóm tắt văn bản

Tóm tắt văn bản (TTVB) là một trong những nhiệm vụ quan trọng của NLP, nhằm tạo ra một phiên bản rút gọn của văn bản gốc mà vẫn duy trì được nội dung cốt lõi và tính nhất quán ngữ nghĩa [11]. Với sự bùng nổ của dữ liệu văn bản trong các lĩnh vực như tin tức, y tế, pháp luật và giáo dục, việc phát triển các phương pháp tóm tắt hiệu quả đóng vai trò quan trọng trong việc tối ưu hóa quy trình xử lý và truy xuất thông tin [22].

TTVB có thể được chia thành hai loại chính: tóm tắt trích xuất và tóm tắt tóm lược hay còn gọi là tóm tắt sinh ngữ [23].

Tóm tắt trích xuất hoạt động bằng cách chọn lọc những câu quan trọng nhất từ văn bản gốc dựa trên các tiêu chí như tần suất từ vựng, trọng số TF - IDF (Term Frequency – Inverse Document Frequency) hoặc điểm số PageRank [24]. Một số thuật toán tiêu biểu cho phương pháp này bao gồm LexRank và TextRank [25].

Trong khi đó, tóm tắt tóm lược sử dụng các mô hình học sâu để tạo ra nội dung tóm tắt mới bằng cách diễn đạt lại thông tin thay vì chỉ trích xuất từ văn bản gốc [26], trong đó các mô hình dựa trên kiến trúc Seq2Seq chú ý [27] hay Transformer [28] đã đạt được kết quả rất tốt. Các hệ thống hiện đại như BART [29] và PEGASUS [30] đã cải thiện đáng kể chất lượng của tóm tắt sinh ngữ bằng cách học cách diễn đạt thông tin từ dữ liệu huấn luyện lớn.



Hình 1.2: Phân loại tóm tắt văn bản

Một thách thức lớn trong TTVB là đảm bảo tính chính xác ngữ nghĩa và tránh lỗi thông tin sai lệch. Các nghiên cứu gần đây tập trung vào việc phát triển các cơ chế kiểm tra tính nhất quán của nội dung tóm tắt bằng cách sử dụng các mô hình kiểm chứng sự thật và kiểm tra chất lượng ngữ nghĩa của đầu ra [31].

Nhìn chung, TTVB là một lĩnh vực quan trọng trong NLP, với nhiều ứng dụng trong tìm kiếm thông tin, báo chí tự động, và trợ lý ảo. Các tiến bộ trong LLMs và DL đang tiếp tục nâng cao hiệu suất của các hệ thống tóm tắt, giúp cải thiện khả năng xử lý thông tin trên quy mô lớn.

1.2.2 Vai trò, ý nghĩa của việc tóm tắt văn bản

TTVB đóng vai trò quan trọng trong hệ thống IR, giúp người dùng nhanh chóng xác định mức độ liên quan của tài liệu mà không cần đọc toàn bộ nội dung [7]. Trong lĩnh vực báo chí, các mô hình tóm tắt hỗ trợ tạo ra tiêu đề và bản tin ngắn gọn từ các bài báo dài, giúp cải thiện trải nghiệm người đọc [32].

Ngoài ra, trong hệ thống hỗ trợ ra quyết định, TTVB giúp rút gọn nội dung từ các báo cáo y khoa, tài liệu pháp lý và nghiên cứu khoa học, hỗ trợ các chuyên gia đưa ra quyết định nhanh chóng và chính xác hơn [33].

1.2.3 Ứng dụng của TTVB trong thực tiễn

Các ứng dụng của TTVB ngày càng đa dạng và mở rộng trong nhiều lĩnh vực. Một số ứng dụng tiêu biểu gồm:

- **Công cụ tìm kiếm và tối ưu hóa công cụ tìm kiếm:** Tóm tắt nội dung các trang web để hiển thị dưới dạng đoạn trích trong kết quả tìm kiếm, giúp người dùng nhanh chóng nắm bắt nội dung chính.
- **Ngành truyền thông và báo chí:** Tự động tạo các bản tin ngắn, hỗ trợ người đọc tiếp cận và hiểu nhanh những thông tin cốt lõi.
- **Trí tuệ nhân tạo và NLP:** TTVB là một trong những ứng dụng trọng tâm của NLP, đặc biệt được khai thác trong các chatbot và hệ thống hỗ trợ khách hàng nhằm nâng cao hiệu quả tương tác và truy xuất thông tin.
- **Lĩnh vực giáo dục**: Hỗ trợ học sinh, sinh viên tóm tắt tài liệu, báo cáo nghiên cứu để nâng cao hiệu quả học tập.
- **Tài chính và phân tích dữ liệu**: Tóm tắt báo cáo tài chính hoặc tin tức kinh tế để cung cấp thông tin ra quyết định nhanh chóng.

1.2.4 Những khó khăn, thách thức đặt ra trong bài toán tóm tắt văn bản

TTVB là một trong những bài toán quan trọng trong NLP, có nhiều thách thức kỹ thuật liên quan đến tính chính xác, khả năng tổng quát hóa, độ phức tạp và hiệu quả tính toán. Mặc dù các mô hình học sâu như Transformer [3] đã đạt được những bước tiến đáng kể trong lĩnh vực này, vẫn còn nhiều vấn đề cần giải quyết để cải thiện chất lượng TTVB.

a) Đảm bảo tính chính xác ngữ nghĩa và nhất quán nội dung

Một trong những thách thức lớn nhất trong TTVB là đảm bảo tính chính xác ngữ nghĩa và nhất quán nội dung [31]. Các mô hình tóm tắt tóm lược có xu hướng tạo ra thông tin sai lệch, nghĩa là chúng có thể thêm vào nội dung không có trong văn bản gốc hoặc làm sai lệch thông tin [34]. Điều này càng trở nên nguy hiểm hơn trong các lĩnh vực như y tế, pháp luật và tin tức, nơi độ chính xác của thông tin đóng vai trò then chốt.

b) Xử lý ngữ cảnh và mối quan hệ giữa các câu

TTVB đòi hỏi mô hình phải hiểu được ngữ cảnh tổng thể của tài liệu và mối quan hệ giữa các câu để tạo ra bản tóm tắt hợp lý [11]. Tuy nhiên, nhiều mô hình hiện tại vẫn gặp khó khăn khi xử lý các tài liệu dài do giới hạn về kích thước đầu vào, chẳng hạn như BERT chỉ có thể xử lý tối đa 512 token [4]. Các nghiên cứu mới như Longformer [35] đang cố gắng khắc phục vấn đề này bằng cách tối ưu hóa kiến trúc Transformer để xử lý các văn bản dài hơn.

c) Giới hạn của dữ liệu huấn luyện và tính tổng quát hóa

Việc huấn luyện mô hình tóm tắt yêu cầu lượng lớn dữ liệu gán nhãn chất lượng cao, nhưng dữ liệu song song rất khó thu thập [30]. Điều này khiến mô hình dễ bị thiên vị theo đặc điểm của tập dữ liệu huấn luyện và kém hiệu quả khi áp dụng cho các lĩnh vực mới hoặc ngôn ngữ ít phổ biến [36].

d) Chi phí tính toán cao và hiệu suất thực thi

Các mô hình tóm tắt hiện đại như BART [29] và PEGASUS [30] yêu cầu tài nguyên tính toán lớn để huấn luyện và suy luận, gây khó khăn cho việc triển khai trong môi trường thực tế với hạn chế về tài nguyên [37]. Điều này đặt ra bài toán tối ưu hóa mô hình để đảm bảo cân bằng giữa độ chính xác và chi phí tính toán.

e) Đánh giá chất lượng tóm tắt

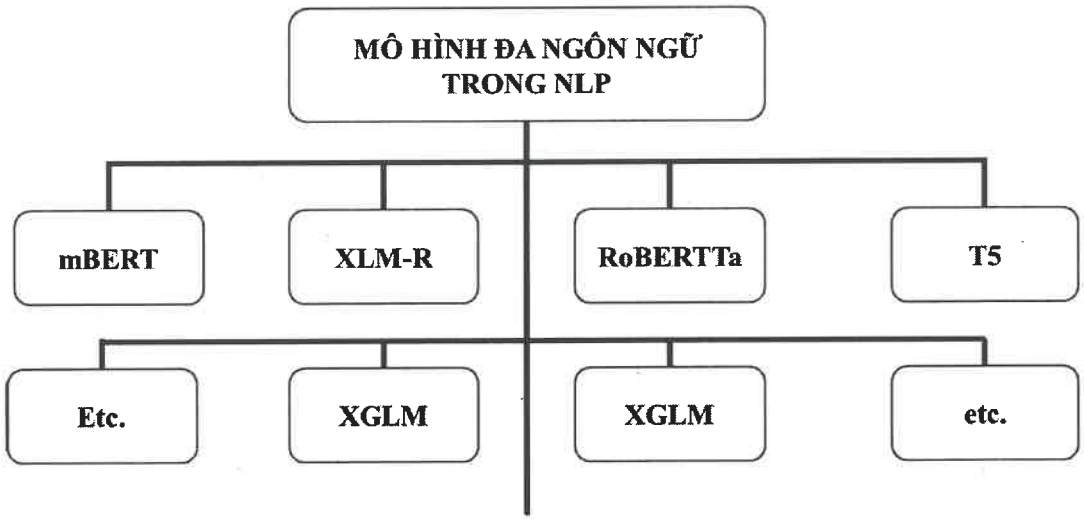
Một thách thức khác là thiếu các phương pháp đánh giá chất lượng tóm tắt hiệu quả. Các chỉ số truyền thống như ROUGE [38] chủ yếu dựa trên sự trùng lặp từ vựng giữa bản tóm tắt tóm lược và bản tóm tắt tham chiếu, nhưng không đo lường được tính chính xác ngữ nghĩa và mức độ mạch lạc của văn bản [39]. Các nghiên cứu gần đây đang đề xuất sử dụng đánh giá dựa trên mô hình ngôn ngữ như BERTScore [30] để cải thiện độ tin cậy của đánh giá.

1.3 Đặc trưng của ngôn ngữ trong bài toán tóm tắt văn bản

1.3.1 Khái quát mô hình đa ngôn ngữ của NLP

Trong lĩnh vực NLP, mô hình đa ngôn ngữ được định nghĩa là các mô hình học máy được huấn luyện trên tập dữ liệu bao gồm nhiều ngôn ngữ khác nhau, nhằm xây

dựng các biểu diễn ngôn ngữ có khả năng khái quát tốt trên phạm vi liên. Theo cách tiếp truyền thống, các mô hình NLP được phát triển theo hướng đơn ngữ, yêu cầu phải xây dựng và huấn luyện mô hình riêng biệt cho từng ngôn ngữ mục tiêu. Cách tiếp cận này dẫn đến những hạn chế lớn về mặt tài nguyên, đặc biệt đối với các ngôn ngữ ít tài nguyên.



Hình 1.3: Mô hình đa ngôn ngữ của NLP

Sự xuất hiện của các mô hình đa ngôn ngữ như mBERT (multilingual BERT) [4] và XLM-R (Cross-lingual Language Model - RoBERTa) [40] đã đánh dấu một bước ngoặt trong việc xây dựng các hệ thống NLP có khả năng tổng quát hóa trên hàng trăm ngôn ngữ. Các mô hình này được huấn luyện theo phương pháp học không giám sát trên tập dữ liệu đa ngôn ngữ lớn như Wikipedia hoặc CommonCrawl, thông qua bài toán học mô hình ngôn ngữ mà không cần nhãn dữ liệu. Trong mBERT, mô hình sử dụng kỹ thuật mô hình ngôn ngữ che từ để học biểu diễn ngữ nghĩa bằng cách dự đoán các token bị che giấu ngẫu nhiên trong câu. XLM-R cải tiến phương pháp này bằng cách mở rộng quy mô dữ liệu huấn luyện lên hàng terabyte và áp dụng kiến trúc RoBERTa để cải thiện khả năng mô hình hóa ngữ nghĩa xuyên ngôn ngữ.

Một đặc điểm quan trọng của mô hình đa ngôn ngữ là khả năng học các biểu diễn ngữ nghĩa liên ngữ, cho phép mô hình ánh xạ các câu từ các ngôn ngữ khác nhau vào cùng một không gian đặc trưng. Điều này hỗ trợ mạnh mẽ cho các tác vụ như dịch máy tự động, phân loại văn bản đa ngôn ngữ và truy vấn tìm kiếm đa ngôn ngữ.

Ngoài ra, các nghiên cứu như của Pires và cộng sự (2019) [41] chỉ ra rằng mBERT, mặc dù không sử dụng phương pháp huấn luyện song ngữ, vẫn tự động học được sự tương đồng về mặt hình thái và cú pháp giữa các ngôn ngữ.

Đáng chú ý, mô hình đa ngôn ngữ góp phần quan trọng vào việc mở rộng nghiên cứu và ứng dụng NLP cho các ngôn ngữ ít tài nguyên, vốn còn nhiều hạn chế về dữ liệu. Kỹ thuật chuyển giao không mẫu và học với ít mẫu được hỗ trợ bởi các mô hình này giúp giảm đáng kể yêu cầu về dữ liệu gán nhãn, cho phép mô hình huấn luyện trên ngôn ngữ nguồn và áp dụng trực tiếp cho ngôn ngữ đích mà không cần tái huấn luyện.

Tóm lại, sự phát triển của mô hình đa ngôn ngữ đã và đang đóng góp lớn vào việc dân chủ hóa công nghệ NLP trên toàn cầu, mở rộng khả năng tiếp cận của các công nghệ ngôn ngữ tới cộng đồng người dùng nói những ngôn ngữ kém phổ biến, đồng thời tối ưu hóa chi phí và hiệu suất triển khai các hệ thống NLP đa ngôn ngữ.

1.3.2 Đặc điểm tóm tắt văn bản tiếng Việt

Việc TTVB Tiếng Việt gặp phải nhiều thách thức do các đặc điểm ngôn ngữ riêng biệt, bao gồm:

- **Cấu trúc câu phức tạp**: Tiếng Việt sử dụng nhiều cách diễn đạt đa nghĩa, gây khó khăn trong việc xác định nội dung chính.
- **Đặc điểm ngôn ngữ học**: Sự phong phú về từ vựng, cách sử dụng từ đồng âm, từ láy và ngữ cảnh làm tăng độ phức tạp khi xử lý tự động.
- **Thiếu bộ dữ liệu chất lượng cao**: Số lượng và chất lượng các bộ dữ liệu phục vụ nghiên cứu tóm tắt văn bản Tiếng Việt còn hạn chế.
- **Mô hình ngôn ngữ chưa tối ưu**: Nhiều mô hình NLP tiên tiến được phát triển dựa trên ngôn ngữ tiếng Anh, dẫn đến việc áp dụng cho Tiếng Việt chưa đạt hiệu quả cao.

1.3.3 Đặc điểm tóm tắt văn bản tiếng Lào

TTVB tiếng Lào là một nhiệm vụ phức tạp trong NLP, do đặc điểm ngôn ngữ và hạn chế về tài nguyên dữ liệu. Một số thách thức chính bao gồm:

- **Vấn đề phân đoạn từ:** Tiếng Lào không sử dụng dấu cách giữa các từ, khiến việc phân đoạn từ trở thành một bước tiền xử lý quan trọng nhưng đầy thách thức. Không có dấu hiệu rõ ràng để xác định ranh giới từ, điều này đòi hỏi mô hình tách từ phải chính xác để cải thiện hiệu suất tóm tắt văn bản.
- **Hệ thống chữ viết và xử lý ký tự:** Bảng chữ cái tiếng Lào có hơn 50 ký tự, với nhiều nguyên âm và phụ âm kết hợp theo vị trí khác nhau, gây khó khăn cho việc nhận dạng và mã hóa văn bản. Các mô hình tóm tắt hiện đại thường dựa vào vector hóa văn bản, nhưng vì tiếng Lào có ít tập dữ liệu huấn luyện lớn, hiệu suất của các mô hình này bị hạn chế.
- **Độ phức tạp về ngữ nghĩa và cú pháp:** Tiếng Lào có cấu trúc câu linh hoạt, với việc lược bỏ chủ ngữ, động từ bị rút gọn và các dạng câu ghép phức tạp. Điều này làm cho các thuật toán tóm tắt gặp khó khăn trong việc xác định thông tin quan trọng. Ngoài ra, tiếng Lào là một ngôn ngữ đơn lập, tức là không có biến tố từ vựng, làm cho việc phân tích cú pháp trở nên khó khăn.
- **Thiếu tài nguyên ngôn ngữ số hóa:** So với các ngôn ngữ như tiếng Anh, Trung, Nhật hay Hàn, tiếng Lào có rất ít tập dữ liệu số hóa để huấn luyện mô hình học sâu. Không có nhiều bộ dữ liệu chuẩn về TTVB tiếng Lào, dẫn đến việc khó khăn trong đánh giá và so sánh chất lượng mô hình.
- **Ảnh hưởng của phương ngữ và vay mượn từ tiếng Thái:** Tiếng Lào có nhiều phương ngữ khác nhau và trong một số khu vực, nó bị ảnh hưởng mạnh bởi từ mượn từ tiếng Thái. Điều này làm cho các mô hình ngôn ngữ gặp khó khăn trong việc nhận diện nghĩa chính xác của từ, ảnh hưởng đến chất lượng TTVB.

1.4 Vấn đề nghiên cứu đặt ra trong đề án

Vấn đề nghiên cứu được đặt ra trong đề án tốt nghiệp là khảo sát các phương pháp TTVB và các mô hình NLP ứng dụng trong bài toán tóm tắt; tìm hiểu một số bộ dữ liệu hiện có phục vụ cho tóm tắt văn bản tiếng Việt và tiếng Lào; triển khai thử nghiệm các thuật toán tóm tắt trích xuất và tóm tắt tóm lược; đồng thời đánh giá hiệu

quả của các phương pháp thông qua kết quả thử nghiệm trên bộ dữ liệu tiếng Việt và tiếng Lào.

1.5 Kết luận chương

Chương 1 của đề án tốt nghiệp đã trình bày cơ sở lý thuyết liên quan đến Xử lý ngôn ngữ tự nhiên và các ứng dụng; giới thiệu bài toán tóm tắt văn bản, nêu bật ý nghĩa và tầm quan trọng của bài toán; phân tích đặc điểm ngôn ngữ tiếng Việt và tiếng Lào trong ngữ cảnh tóm tắt văn bản; đồng thời xác định rõ vấn đề nghiên cứu được đặt ra trong đề tài.

CHƯƠNG 2. KHẢO SÁT, ĐÁNH GIÁ CÁC PHƯƠNG PHÁP TÓM TẮT VĂN BẢN SỬ DỤNG NLP

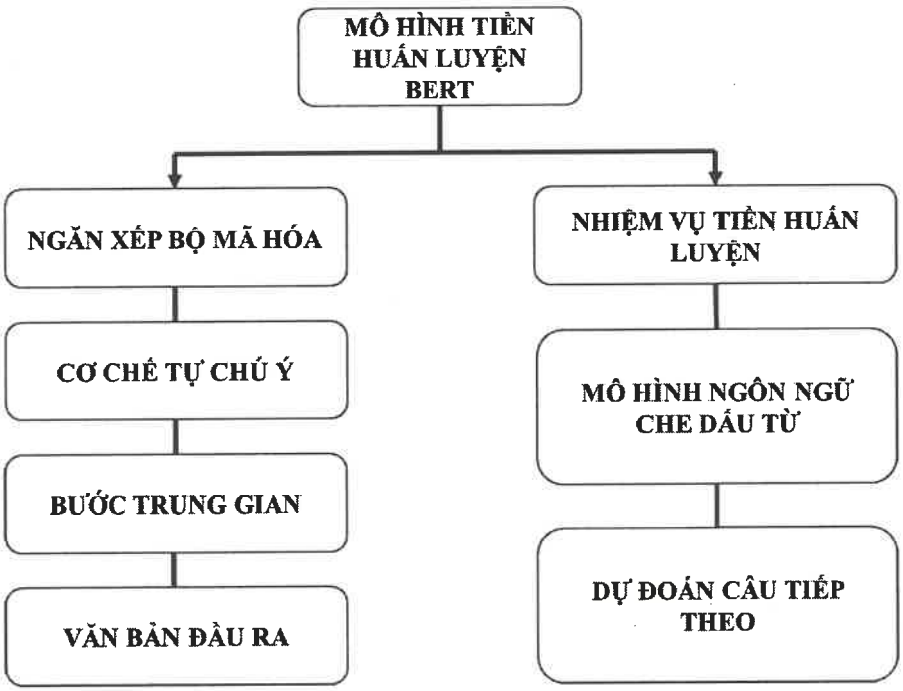
Chương này tập trung phân tích một số mô hình NLP hiện đại cho TTVB; Khảo sát và đánh giá các phương pháp TTVB sử dụng NLP; Phân tích đánh giá một số mô hình hỗ trợ TTVB đa ngôn ngữ; đề xuất mô hình thử nghiệm cho TTVB tiếng Việt và tiếng Lào sử dụng NLP; phương pháp tạo lập Dataset cho TTVB; Trình bày một số phương pháp đánh giá cho hệ thống TTVB sử dụng NLP.

2.1 Một số mô hình NLP hiện đại cho tóm tắt văn bản

TTVB tự động là một trong những nhiệm vụ cốt lõi của NLP, với hai hướng tiếp cận chính: tóm tắt trích xuất và tóm tắt tóm lược. Các mô hình hiện đại, đặc biệt là các LLMs, đã đạt được những tiến bộ đáng kể trong cả hai hướng tiếp cận này.

2.1.1 Mô hình tiền huấn luyện

Mô hình tiền huấn luyện đã trở thành trụ cột trong NLP hiện đại. Thay vì huấn luyện từ đầu trên mỗi nhiệm vụ cụ thể, các mô hình như BERT [4] và RoBERTa [42] được tiền huấn luyện trên tập dữ liệu cực lớn theo bài toán mô hình ngôn ngữ, sau đó được tinh chỉnh cho các tác vụ như TTVB.



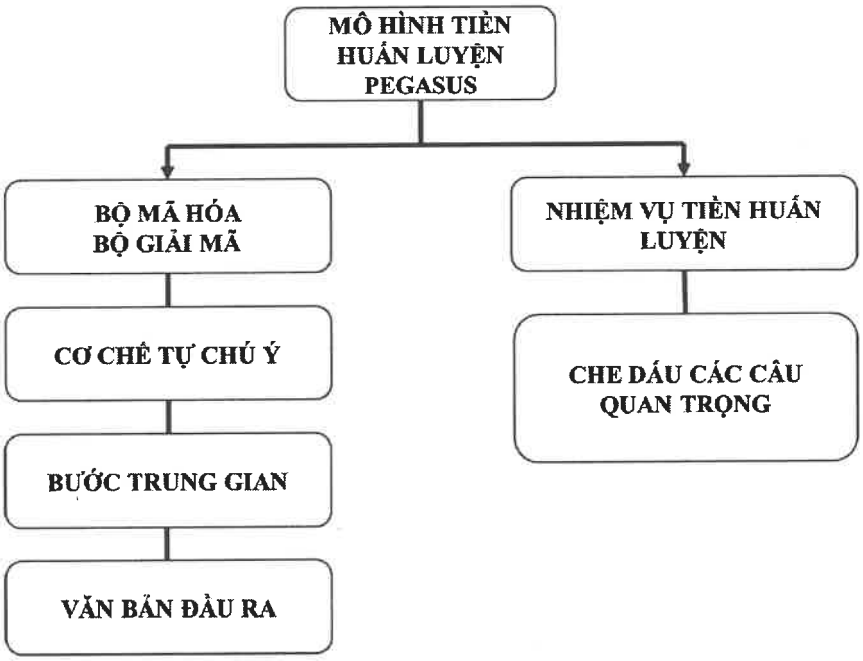
Hình 2.1: Kiến trúc mô hình tiền huấn luyện với BERT

BERTSUM, một biến thể của BERT, đã được đề xuất cho bài toán tóm tắt trích xuất bằng cách thêm các token đặc biệt vào mỗi câu và điều chỉnh kỹ thuật fine-tuning [42]. Việc sử dụng mô hình tiền huấn luyện đã cho phép các hệ thống tóm tắt học được các biểu diễn ngữ nghĩa sâu sắc và giàu thông tin ngữ cảnh, đồng thời giảm thiểu chi phí tính toán và yêu cầu dữ liệu nhãn khổng lồ như trong các phương pháp truyền thống [30].

2.1.2 Mô hình Transformers

Kiến trúc Transformer, lần đầu tiên được giới thiệu bởi Vaswani và cộng sự (2017) [28], đã tạo nên bước ngoặt cho NLP, thay thế hoàn toàn các mạng tuần tự truyền thống (RNNs, LSTM) trong nhiều tác vụ, bao gồm TTVB. Các mô hình như BART [29] đã tận dụng các kỹ thuật chèn nhiễu trong quá trình tiền huấn luyện để học khả năng sinh bản tóm tắt mạch lạc, giàu tính ngữ nghĩa.

PEGASUS tiếp tục đẩy mạnh hiệu suất bằng cách đề xuất phương pháp tiền huấn luyện dựa trên việc che giấu các câu quan trọng, giúp mô hình học trực tiếp cấu trúc của bản tóm tắt [30]. Nghiên cứu gần đây trên Nature Communications chỉ ra rằng mô hình Transformer, với khả năng xử lý ngữ cảnh dài và chú ý cơ bản, đã đạt được hiệu quả vượt trội trong việc tóm tắt các báo cáo khoa học dài dòng [43].

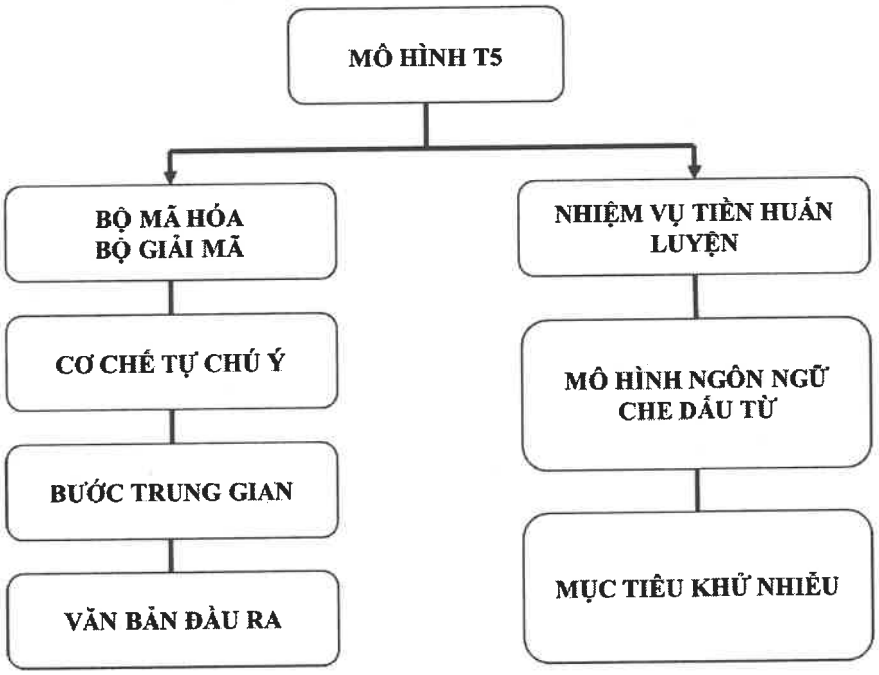


Hình 2.2: Kiến trúc mô hình tiền huấn luyện với PEGASUS

2.1.3 Mô hình Encoder-Decoder

Cấu trúc encoder - decoder hay được hiểu là cấu trúc bộ mã hóa – bộ giải mã là nền tảng của nhiều mô hình tóm tắt văn bản tóm lược hiện đại. Trong đó, encoder (bộ mã hóa) nén thông tin đầu vào thành biểu diễn vector không gian ngữ nghĩa, và decoder (bộ giải mã) sử dụng biểu diễn này để sinh bản tóm tắt.

T5 (Text-to-Text Transfer Transformer), một mô hình encoder-decoder tổng quát, chuyển đổi mọi bài toán NLP thành bài toán sinh văn bản, đã chứng minh hiệu suất mạnh mẽ trong tóm tắt các bài báo khoa học và các văn bản kỹ thuật phức tạp. Đặc biệt, T5 đã được thử nghiệm trên các tập dữ liệu chuyên biệt như CNN/DailyMail và PubMed với kết quả ROUGE score ấn tượng [19].



Hình 2.3: Kiến trúc mô hình tiền huấn luyện với Encoder - Decoder T5

Bên cạnh đó, các mô hình như ProphetNet [44], với kỹ thuật dự đoán n bước tiếp theo thay vì chỉ một bước, đã cải thiện chất lượng sinh bản tóm tắt nhờ khả năng dự đoán các đơn vị ngữ nghĩa lớn hơn.

2.1.4 Một số mô hình khác

Ngoài các kiến trúc nêu trên, nhiều mô hình chuyên dụng khác cũng đóng góp đáng kể vào tiến bộ trong TTVB:

+ **Longformer**: Được thiết kế để xử lý các tài liệu dài với kỹ thuật chú ý phân đoạn, Longformer giúp giảm chi phí tính toán từ $O(n^2)$ xuống $O(n)$, phù hợp cho tóm tắt các báo cáo nghiên cứu dài [35]. Trong các nghiên cứu trên Scientific Reports, Longformer đã cho thấy tiềm năng vượt trội trong việc TTVB học thuật dài [43].

+ **Z-Code++**: Đây là mô hình đa ngôn ngữ dựa trên kiến trúc encoder - decoder với cơ chế chú ý hiệu quả và mã hóa phân tầng, đạt hiệu suất vượt trội trong tóm tắt đa ngôn ngữ [45].

+ **ChatGPT** và **LLMs**: Gần đây, các mô hình như ChatGPT đã được thử nghiệm trong tóm tắt hội thoại y tế. Một nghiên cứu được công bố trên Nature Scientific Reports cho thấy ChatGPT đạt điểm số cao về độ mạch lạc và tính đầy đủ nội dung khi so sánh với các mô hình truyền thống và các phiên bản LLMs nhỏ hơn [46].

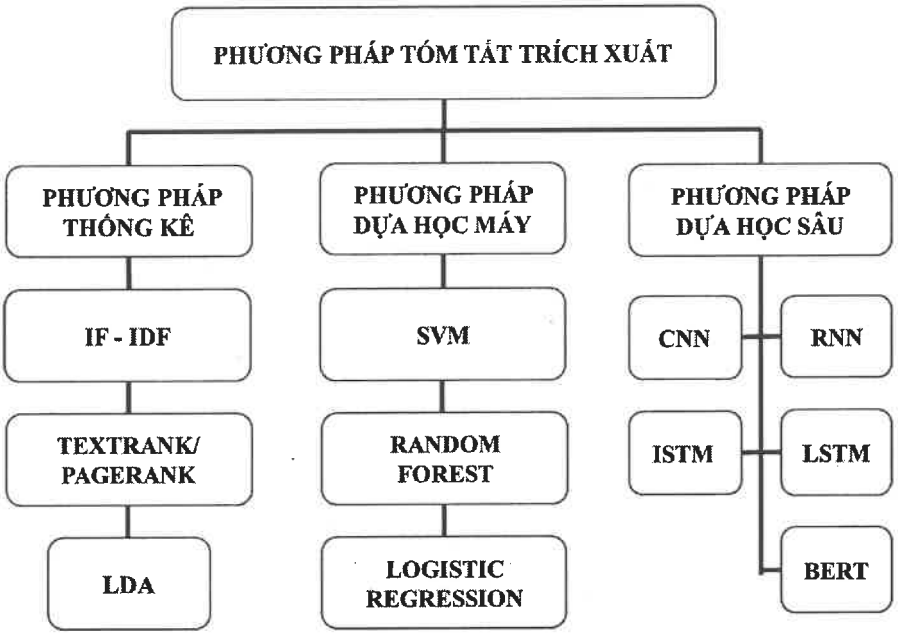
+ **LED (Longformer Encoder - Decoder)**: Một mô hình khác phát triển từ Longformer, sử dụng encoder-decoder cho các bài toán tóm tắt tài liệu dài, đặc biệt hiệu quả trên các tập dữ liệu như ArXiv và PubMed [35].

2.2 Các phương pháp tóm tắt văn bản sử dụng NLP

Hiện nay, có nhiều phương pháp tóm tắt văn bản khác nhau, có thể chia thành hai nhóm chính: Tóm tắt trích xuất và Tóm tắt tóm lược (Tóm tắt sinh ngữ).

2.2.1 Tóm tắt trích xuất

Tóm tắt trích xuất là phương pháp chọn lọc các câu quan trọng từ văn bản gốc, giữ nguyên nội dung và cấu trúc của các câu được lựa chọn. Các thuật toán phổ biến trong nhóm này bao gồm: Phương pháp thống kê, Phương pháp dựa học máy và Phương pháp dựa học sâu.



Hình 2.4: Phân loại các phương pháp tóm tắt trích xuất văn bản

a. Phương pháp thống kê

Các thuật toán sử dụng các chỉ số thống kê để đánh giá mức độ quan trọng của câu, bao gồm:

+ TF - IDF (Term Frequency - Inverse Document Frequency): Một kỹ thuật đánh giá tầm quan trọng của từ trong một văn bản, bằng cách kết hợp tần suất xuất hiện của từ trong văn bản với mức độ hiếm của từ trong toàn bộ tập dữ liệu [7].

+ TextRank/PageRank: Sử dụng đồ thị để xác định các câu quan trọng dựa trên mức độ liên kết giữa các từ khóa và câu trong văn bản [25].

+ LDA (Latent Dirichlet Allocation): Phân tích chủ đề của văn bản và chọn ra các câu phù hợp với chủ đề chính.

b. Phương pháp tóm tắt dựa trên mô hình học máy [47]

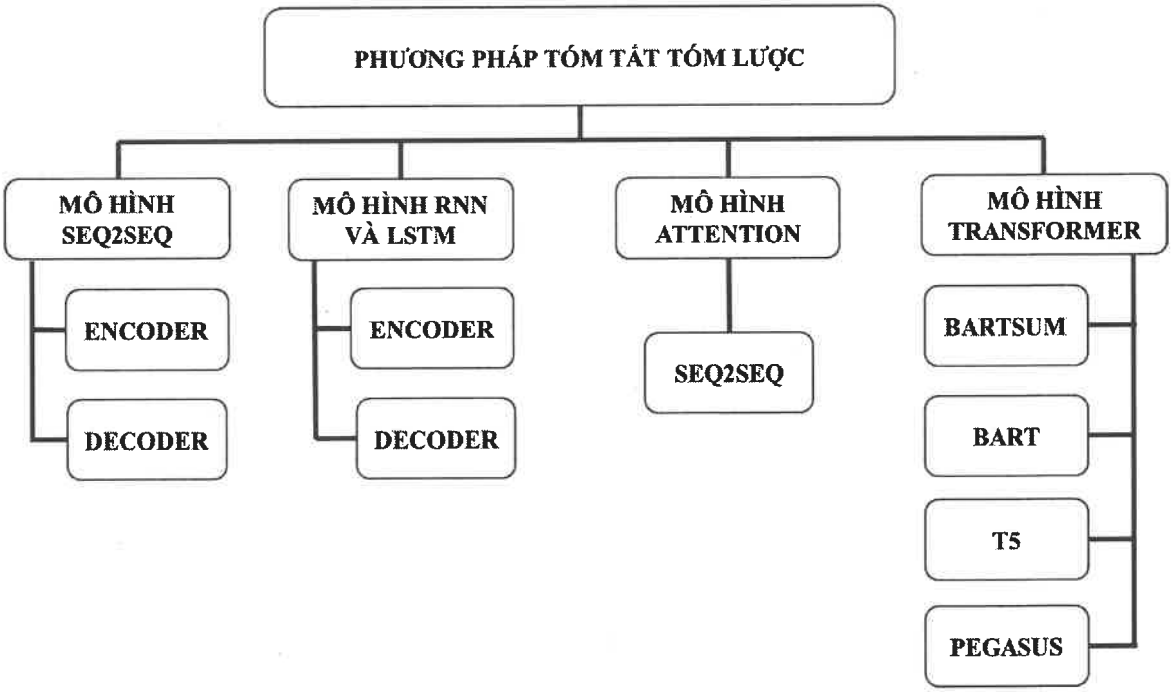
Mô hình phân loại để xác định mức độ quan trọng của từng câu: Sử dụng các đặc trưng của câu (độ dài, vị trí, từ khóa quan trọng) để huấn luyện một mô hình học máy (SVM, Random Forest, Logistic Regression) nhằm chọn ra các câu quan trọng.

c. Phương pháp tóm tắt dựa trên mô hình học sâu

Các mô hình mạng nơ-ron như CNNs, RNNs, LSTM hoặc BERT có thể được huấn luyện để học cách chọn câu quan trọng từ dữ liệu huấn luyện [47].

2.2.2 Tóm tắt tóm lược

Khác với tóm tắt trích xuất, tóm tắt tóm lược không chỉ lấy nguyên văn các câu trong tài liệu gốc mà còn diễn giải lại nội dung bằng ngôn ngữ mới. Phương pháp này khai thác các kỹ thuật hiện đại của NLP kết hợp với AI để nâng cao hiệu quả xử lý.



Hình 2.5: Phân loại các phương pháp tóm tắt tóm lược văn bản

Các phương pháp theo tóm tắt tóm lược bao gồm:

a) Mô hình seq2seq (Sequence-to-Sequence)

Áp dụng kiến trúc encoder-decoder, trong đó bộ mã hóa (encoder) chuyển đổi văn bản đầu vào thành vector ngữ nghĩa, sau đó bộ giải mã (decoder) tạo ra tóm tắt ngắn gọn [26].

b) RNNs và LSTM

Mô hình Seq2Seq sử dụng bộ mã hóa để trích xuất thông tin từ văn bản đầu vào và bộ giải mã để tạo ra bản tóm tắt mới.

c) Cơ chế chú ý

Cải thiện hiệu suất của Seq2Seq bằng cách giúp mô hình tập trung vào các phần quan trọng của văn bản thay vì xử lý toàn bộ văn bản cùng lúc.

d) Mô hình Transformer

Sử dụng cơ chế tự chú ý để tạo tóm tắt chính xác hơn. Các mô hình như BART, T5 và PEGASUS đã đạt hiệu suất cao trong nhiệm vụ TTVB [32]:

+ BERTSUM (BERT for Summarization): Mở rộng mô hình BERT để hỗ trợ tóm tắt trích xuất và tóm lược.

+ T5: Xử lý nhiều tác vụ NLP, trong đó có TTVB.

+ BART: Tối ưu hóa TTVB bằng cách sử dụng phương pháp huấn luyện không giám sát với nhiễu ngẫu nhiên.

+ PEGASUS: Một trong những mô hình mạnh nhất hiện nay, được thiết kế đặc biệt để TTVB tự nhiên.

2.2.3 Phương pháp lai

Một số nghiên cứu đề xuất mô hình lai (kết hợp), trong đó các câu quan trọng trước tiên được trích xuất, sau đó được diễn giải lại bằng mô hình tóm lược nhằm cải thiện chất lượng tóm tắt [39].

Bước 1: Trích xuất câu quan trọng bằng phương pháp Tóm tắt trích xuất.

Bước 2: Sử dụng Tóm tắt tóm lược (sinh ngữ) để diễn đạt lại nội dung trích xuất được.

Ví dụ, mô hình Pointer-Generator Network kết hợp khả năng sao chép thông tin từ văn bản gốc với khả năng tạo ra nội dung mới.

TTVB hiện đại đang được ứng dụng rộng rãi trong nhiều lĩnh vực như tìm kiếm thông tin, hỗ trợ trợ lý ảo, tóm tắt tin tức và xử lý tài liệu pháp lý [31]. Tuy nhiên, các thách thức như giữ nguyên tính chính xác của thông tin, giảm thiểu lỗi "hallucination" và tối ưu hóa tóm tắt cho văn bản dài vẫn đang là chủ đề nghiên cứu quan trọng trong cộng đồng NLP [5].

2.3 Đề xuất mô hình thử nghiệm cho tóm tắt văn bản trên dữ liệu tiếng Việt và tiếng Lào

Qua quá trình tổng quan tài liệu và đánh giá các vấn đề nghiên cứu trong bài toán TTVB áp dụng với dữ liệu ngôn ngữ tiếng Lào – một ngôn ngữ có tài nguyên huấn luyện hạn chế, có thể thấy việc lựa chọn một mô hình tóm tắt văn bản phù hợp cần dựa trên khả năng tổng quát hóa tốt, thích ứng linh hoạt với dữ liệu không chuẩn hóa, đồng thời tối ưu hóa hiệu quả huấn luyện trên tập dữ liệu nhỏ. Dựa trên các tiêu chí này, đề án đề xuất lựa chọn hai mô hình kiến trúc tiên tiến là T5 và BART vì cả hai mô hình này đều được thiết kế theo kiến trúc Encoder - Decoder, giúp tối ưu hóa quá trình hiểu và sinh ngôn ngữ, đặc biệt phù hợp với các nhiệm vụ sinh văn bản như tóm tắt, cụ thể:

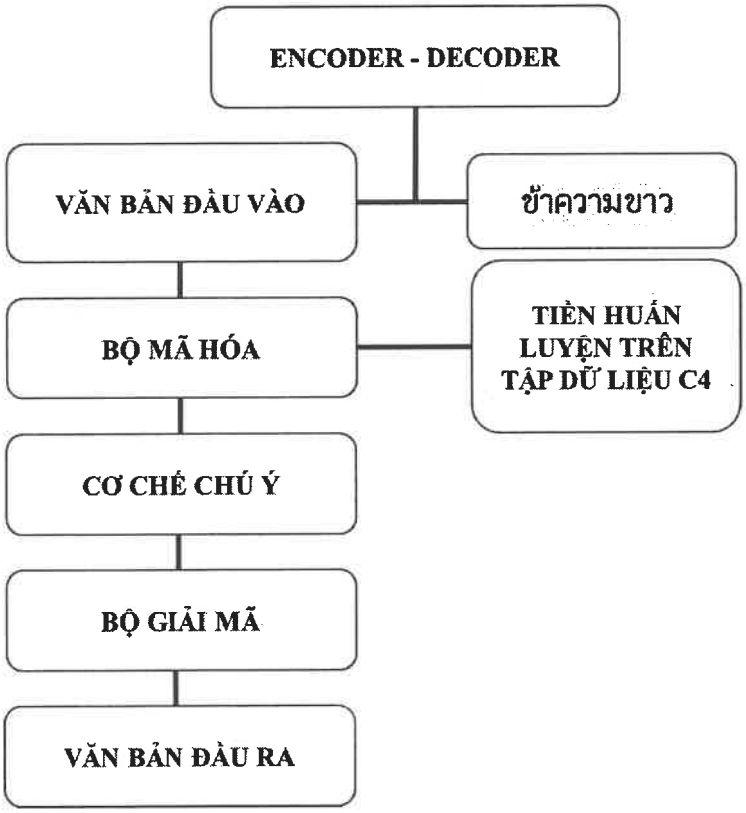
T5 nổi bật với cách tiếp cận "text-to-text" thống nhất, nơi mọi tác vụ – bao gồm tóm tắt được chuyển thành bài toán sinh văn bản. Cách tiếp cận này không chỉ nhất quán trong xử lý mà còn đặc biệt hiệu quả trong việc fine-tune trên các tập dữ liệu chuyên biệt như tiếng Lào, nhờ khả năng tận dụng tri thức ngôn ngữ đã học từ tiền huấn luyện. T5 còn hỗ trợ điều chỉnh mục tiêu như "summarize:", giúp định hướng mô hình đúng mục đích mà không cần thay đổi kiến trúc.

Trong khi đó, BART là mô hình lai, kết hợp hiệu quả giữa BERT (mã hóa hai chiều) và GPT (giải mã tự hồi tiếp), cho phép khôi phục nội dung từ văn bản bị nhiễu – một đặc điểm hữu ích khi xử lý dữ liệu ngôn ngữ tiếng Lào có thể bị lỗi cú pháp, thiếu dấu câu, không chuẩn hóa hoặc chứa từ ngữ địa phương. Do đó, BART thể hiện khả năng kết hợp giữa hiểu ngữ nghĩa toàn cục và khả năng sinh ngôn ngữ mạch lạc. Ngoài ra, cả hai mô hình đều được hỗ trợ mạnh mẽ trong các thư viện như Hugging Face Transformers, cho phép triển khai, tinh chỉnh và đánh giá linh hoạt với nhiều thiết lập khác nhau. Quan trọng hơn, T5 và BART đều đã được chứng minh hiệu quả vượt trội trên nhiều tiêu chuẩn đánh giá tóm tắt như CNN/DailyMail, XSum và cả các ngôn ngữ ít phổ biến khi tinh chỉnh đúng cách – điều này mở ra tiềm năng lớn khi áp dụng cho tiếng Lào.

Việc đề án đề xuất lựa chọn T5 và BART không chỉ dựa trên hiệu suất mô hình mà còn phản ánh định hướng chiến lược trong khai thác sức mạnh của các mô hình tiền huấn luyện hiện đại để thích ứng với những ngôn ngữ ít tài nguyên, góp phần thúc đẩy khả năng ứng dụng NLP đa ngữ trong khu vực Đông Nam Á.

2.3.1 Mô hình T5

T5 là một mô hình ngôn ngữ lớn được phát triển bởi Google Research nhằm thống nhất các bài toán NLP dưới dạng bài toán chuyển đổi văn bản thành văn bản. Kiến trúc của T5 dựa trên bộ khung Transformer encoder-decoder với khả năng mở rộng linh hoạt, được tiền huấn luyện trên tập dữ liệu C4 (Colossal Clean Crawled Corpus) có quy mô lớn, bao gồm nhiều ngôn ngữ tự nhiên trong đó có tiếng Việt và tiếng Lào [19].



Hình 2.6: Mô hình T5 cho thử nghiệm tóm tắt văn bản

Ưu điểm nổi bật của mô hình T5 là khả năng học biểu diễn ngữ nghĩa sâu rộng nhờ vào phương pháp tiền huấn luyện dạng dự đoán chuỗi từ bị che giấu, cho phép

nắm bắt tốt các ngữ cảnh dài, đồng thời thích ứng hiệu quả với các bài toán tóm tắt tóm lược.

a) Tiền xử lý dữ liệu

+ **Đối với tiếng Việt**: tập dữ liệu VLSP-Sum hoặc VieSum sẽ được sử dụng làm nguồn dữ liệu huấn luyện. Trước khi đưa vào huấn luyện, dữ liệu sẽ được chuẩn hóa bằng các công cụ NLP như VnCoreNLP nhằm thực hiện các bước tách từ, chuẩn hóa văn bản và loại bỏ các ký tự không cần thiết.

+ **Đối với tiếng Lào**:

Các nghiên cứu gần đây cho thấy mô hình LLM như GPT - 3 và ChatGPT có khả năng tạo dữ liệu huấn luyện chất lượng cao cho các ngôn ngữ ít tài nguyên, đặc biệt nếu có sự kiểm soát câu gợi ý và đánh giá đầu ra. Các bản tóm tắt do ChatGPT sinh ra thường đạt tính mạch lạc, đúng ngữ pháp và bảo toàn thông tin cốt lõi, nhờ vào kiến trúc huấn luyện dựa trên hàng tỷ câu song ngữ và các tác vụ ngôn ngữ đa nhiệm.

Do sự hạn chế về tập dữ liệu tiếng Lào, đề án đề xuất sử dụng các LLM như OpenAI GPT để tự động tạo dữ liệu huấn luyện thông qua kỹ thuật tóm tắt. Bên cạnh đó, đề án không sử dụng đầu ra mô hình một cách thô sơ, mà áp dụng quy trình kiểm duyệt bán thủ công, trong đó các bản tóm tắt được rà soát, chỉnh sửa hoặc loại bỏ nếu không đạt yêu cầu về ngữ nghĩa, độ súc tích hoặc văn phong tiếng Lào. Cách tiếp cận kết hợp giữa LLM và quy trình có con người tham gia giám sát này đã được chứng minh là hiệu quả trong việc xây dựng tập dữ liệu tin cậy cho các tác vụ NLP không có nguồn dữ liệu lớn sẵn có.

Cụ thể, các văn bản tiếng Lào có độ dài lớn sẽ được đưa vào ChatGPT để tạo ra các bản tóm tắt ngắn gọn nhưng vẫn giữ được nội dung chính, sau đó qua kiểm duyệt bán thủ công bởi tác giả đề án. Các cặp văn bản gốc và bản tóm tắt này sau đó sẽ được sử dụng như dữ liệu huấn luyện cho bài toán tóm tắt tiếng Lào. Phương pháp này không chỉ giúp mở rộng quy mô tập dữ liệu mà còn cải thiện tính đa dạng và độ bao phủ về ngữ nghĩa. Đồng thời, bộ tách từ SentencePiece với từ điển từ chung được

huấn luyện trên tập mC4 cũng được áp dụng nhằm đảm bảo tính nhất quán trong việc biểu diễn đầu vào.

b) Phương pháp huấn luyện mô hình

Quá trình huấn luyện mô hình T5 cho bài toán TTVB sẽ được tiến hành theo các bước như sau:

Bước 1: Khởi tạo mô hình T5 với trọng số tiền huấn luyện từ checkpoint t5-small.

Bước 2: Tiền xử lý dữ liệu với việc thêm tiền tố "summarize:" vào đầu mỗi văn bản đầu vào nhằm định hướng mô hình nhận diện tác vụ cần thực hiện.

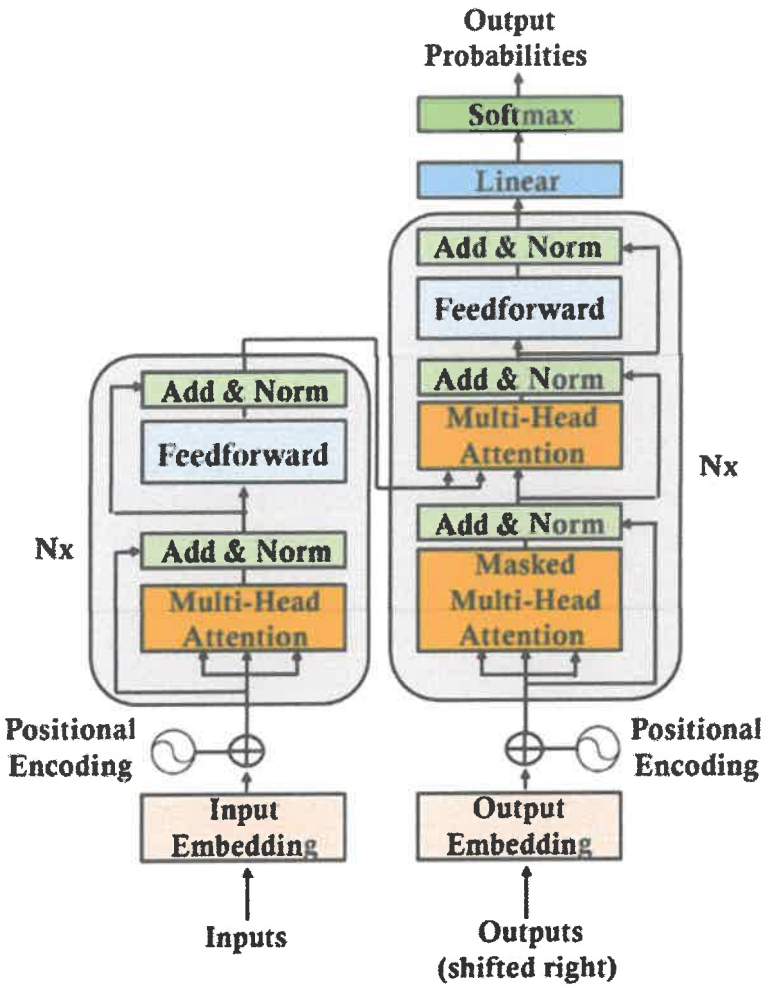
Bước 3: Thiết lập tham số huấn luyện, bao gồm:

- Learning rate: $2e^{-5}$.
- Batch size: 16.
- Optimizer: Adam với thông số betas=(0.9, 0.999) và epsilon=$1e^{-8}$.
- Scheduler: Linear warm-up followed by linear decay.
- Số epoch: 70.

Để tối ưu hóa quá trình huấn luyện và tránh hiện tượng quá khớp, kỹ thuật dừng sớm sẽ được áp dụng dựa trên chỉ số ROUGE-L trên tập kiểm thử. Đánh giá mô hình sẽ dựa trên bộ chỉ số ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), nhằm đảm bảo chất lượng tóm tắt cả về mức độ bao phủ và tính ngữ nghĩa.

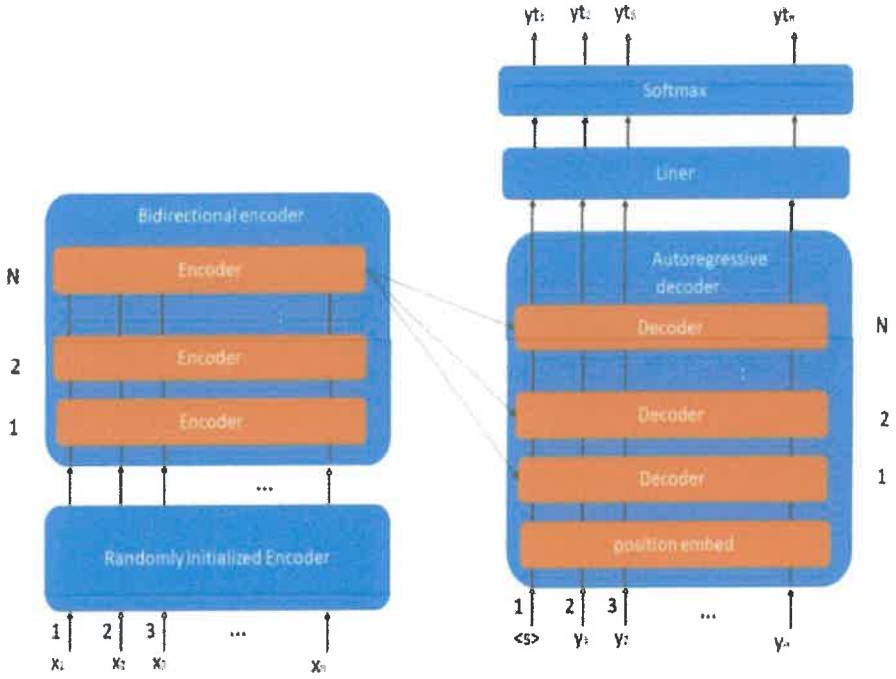
2.3.2 Mô hình BART

Được giới thiệu bởi Vaswani và các cộng sự vào năm 2017, Transformer là một kiến trúc mạng nơ-ron mang tính cách mạng, góp phần thay đổi căn bản các phương pháp NLP. Với cơ chế tự chú ý, Transformer giúp mô hình hiểu ngữ cảnh một cách toàn diện hơn và xử lý hiệu quả các chuỗi dữ liệu dài [3]. Trên cơ sở đó, nhiều mô hình tóm tắt tóm lược hiện đại đã được phát triển, trong đó đáng chú ý là BART và BARTpho.



Hình 2.7: Kiến trúc mô hình Transformer [19]

BART là một mô hình tóm tắt mạnh mẽ, kết hợp khả năng mã hóa hai chiều của BERT và khả năng sinh văn bản tự động hồi quy của GPT. BART sử dụng một kiến trúc seq2seq, trong đó bộ mã hóa được huấn luyện để tái tạo văn bản gốc từ dữ liệu bị nhiễu, còn bộ giải mã được huấn luyện để sinh ra văn bản tóm tắt từ vector ngữ cảnh. BART đã chứng minh hiệu quả vượt trội trên nhiều nhiệm vụ tóm tắt tóm lược và là lựa chọn hàng đầu cho các ngôn ngữ có tài nguyên phong phú như tiếng Anh.



Hình 2.8: Mô hình BART [19]

BARTpho là một mô hình ngôn ngữ tiền huấn luyện được thiết kế dành riêng cho tiếng Việt, phát triển dựa trên kiến trúc BART do nhóm nghiên cứu Facebook AI Research đề xuất [29]. Mô hình này được nhóm nghiên cứu của Đại học Công nghệ Sydney (UTS) xây dựng với mục tiêu khai thác đầy đủ các đặc điểm ngôn ngữ học của tiếng Việt, đồng thời giải quyết các hạn chế khi áp dụng trực tiếp các mô hình tiền huấn luyện tiếng Anh cho tiếng Việt. Về mặt kiến trúc, BARTpho kết hợp hai thành phần chính:

- **Encoder** hai chiều tương tự như BERT để học biểu diễn ngữ nghĩa toàn bộ văn bản đầu vào.
- **Decoder** sinh tự hồi tương tự như GPT để tạo ra văn bản đầu ra một cách tuần tự từ trái sang phải.

Các mô hình dựa trên kiến trúc Transformer nổi bật với khả năng hiểu ngữ cảnh phức tạp và tạo ra văn bản tự nhiên, mạch lạc; tuy nhiên, chúng yêu cầu tài nguyên tính toán cao, hạn chế khả năng ứng dụng trong môi trường phần cứng giới hạn.

a. Dữ liệu tiền huấn luyện

BARTpho được huấn luyện trên tập dữ liệu PhoCorpus, một tập hợp lớn bao gồm:

- Văn bản tin tức;
- Văn bản Wikipedia tiếng Việt;
- Văn bản đa lĩnh vực thu thập từ Common Crawl.

Tổng dung lượng dữ liệu đạt khoảng 20GB văn bản thô, tương đương với khoảng 1 tỷ tokens.

Bên cạnh đó, để phù hợp với cấu trúc từ đơn lập của tiếng Việt, BARTpho sử dụng tách từ SentencePiece với kích thước từ điển 50.000 tokens, giúp tối ưu hóa việc phân tách từ ngữ trong tiếng Việt, vốn không dựa trên dấu cách như tiếng Anh.

b. Phương pháp tiền huấn luyện

Quá trình tiền huấn luyện BARTpho dựa trên bài toán mạng autoencoder khử nhiễn, với các kỹ thuật gây nhiễu gồm:

- Che token: Che ngẫu nhiên các token trong văn bản.
- Xóa token: Xóa bỏ một số token ngẫu nhiên.
- Điền từ bị thiếu: Thay thế các chuỗi liên tiếp bằng một token mask duy nhất.
- Hoán vị câu: Hoán đổi thứ tự các câu trong văn bản.

Mục tiêu là tái tạo văn bản gốc từ phiên bản đã bị gây nhiễu, qua đó mô hình học được khả năng hiểu sâu ngữ cảnh và cấu trúc ngôn ngữ tự nhiên.

c. Ứng dụng

BARTpho đã chứng minh hiệu quả vượt trội trong nhiều tác vụ NLP tiếng Việt, chẳng hạn như:

- Tóm tắt văn bản;
- Trả lời câu hỏi;
- Sinh văn bản tự động;
- Phân loại văn bản.

Đặc biệt, trong bài toán TTVB tiếng Việt, BARTpho đạt được chỉ số ROUGE-1 và ROUGE-L cao hơn so với các mô hình trước đó như PhoBERT và mBERT fine-tuning [49] . Trong các bài kiểm tra chuẩn trên tập dữ liệu VieSum và VLSP-Sum:

- ROUGE-1: 44,5;
- ROUGE-2: 21,3;
- ROUGE-L: 40,7.

Các chỉ số này thể hiện năng lực vượt trội của BARTpho trong việc nắm bắt thông tin trọng tâm và sinh văn bản ngắn gọn, tự nhiên bằng tiếng Việt.

2.4 Các phương pháp tạo lập Dataset cho tóm tắt văn bản

2.4.1 Các nguồn dữ liệu

Việc xây dựng tập dữ liệu (dataset) cho bài toán TTVB đòi hỏi nguồn dữ liệu phải đảm bảo tính phong phú, đa dạng và phù hợp với mục tiêu nghiên cứu. Các nguồn dữ liệu thường được khai thác bao gồm:

- **Nguồn dữ liệu báo chí**: Các bài báo điện tử từ các cổng thông tin lớn như CNN, DailyMail (tiếng Anh), VNExpress, Thanh Niên (tiếng Việt), và Vientiane Times (tiếng Lào).
- **Nguồn dữ liệu học thuật**: Bao gồm các bài báo khoa học và báo cáo kỹ thuật được thu thập từ các cơ sở dữ liệu như arXiv và PubMed.
- **Nguồn dữ liệu văn bản hành chính**: Các báo cáo, nghị quyết, văn bản hành chính công khai từ các cơ quan nhà nước.
- **Nguồn dữ liệu cộng đồng**: Các tập dữ liệu mở do cộng đồng nghiên cứu xây dựng và công bố như CNN/DailyMail, XSum, Gigaword, VieSum (cho tiếng Việt).
- **Nguồn dữ liệu tổng hợp từ web**: Dữ liệu crawl từ các website, blogs, diễn đàn thông qua các công cụ như Common Crawl, Scrapy.

Tùy thuộc vào mục tiêu cụ thể (ví dụ: tóm tắt tin tức, tóm tắt báo cáo, tóm tắt hội thoại), nguồn dữ liệu sẽ được lựa chọn phù hợp nhằm đảm bảo tính đại diện và khả năng tổng quát hóa của mô hình.

2.4.2 Các kỹ thuật thu thập dữ liệu điển hình

Quá trình thu thập dữ liệu cho bài toán tóm tắt văn bản có thể được thực hiện theo một số kỹ thuật sau:

- **Web Crawling**: Sử dụng trình thu thập dữ liệu tự động để lấy nội dung văn bản từ các trang web. Các công cụ phổ biến: Scrapy, BeautifulSoup, Selenium.
- **API khai thác dữ liệu**: Truy cập vào các API cung cấp nội dung tin tức như News API, New York Times API để thu thập dữ liệu theo thời gian thực.
- **Khai thác dữ liệu mở**: Tận dụng các tập dữ liệu công khai sẵn có từ Kaggle, Hugging Face hoặc các tổ chức nghiên cứu quốc tế.
- **Thu thập bán tự động**: Kết hợp thu thập tự động với quá trình lọc và xác thực thủ công nhằm đảm bảo chất lượng dữ liệu.
- **Tập dữ liệu song ngữ**: Với các ngôn ngữ ít tài nguyên như tiếng Lào, phương pháp dịch ngược từ tiếng Anh có thể được áp dụng để tạo ra tập dữ liệu song ngữ phục vụ cho các mô hình tóm tắt đa ngôn ngữ.

2.4.3 Tiền xử lý dữ liệu: loại bỏ nhiễu, chuẩn hóa

Do dữ liệu thu thập ban đầu thường tồn tại nhiều nhiễu và sự không đồng nhất, các bước tiền xử lý sau đây là cần thiết để đảm bảo chất lượng dữ liệu:

- **Loại bỏ nhiễu**:

+ Loại bỏ các thẻ HTML, ký tự đặc biệt, đoạn quảng cáo.

+ Loại bỏ các văn bản rỗng, không đủ độ dài tối thiểu.

- **Chuẩn hóa văn bản**:

+ Chuẩn hóa bảng mã Unicode, chuyển về dạng chuẩn NFC.

+ Chuẩn hóa dấu câu, viết hoa/viết thường đồng nhất.

+ Thực hiện tách từ cho các ngôn ngữ như tiếng Việt, tiếng Lào, vốn không phân tách từ bằng dấu cách rõ ràng.

- **Loại bỏ stopwords**: Áp dụng bộ từ dừng phù hợp với từng ngôn ngữ nhằm giảm nhiễu và tăng tính hiệu quả cho các mô hình học máy.
- **Chuẩn hóa ngôn ngữ**: Với các văn bản song ngữ hoặc đa ngôn ngữ, cần thực hiện phát hiện và lọc theo ngôn ngữ để đảm bảo tính nhất quán ngôn ngữ trong tập dữ liệu.

2.4.4 Phân chia tập dữ liệu: tập huấn luyện, kiểm thử và đánh giá

Để huấn luyện và đánh giá mô hình hiệu quả, dữ liệu cần được phân chia hợp lý theo các tỷ lệ phổ biến:

- **Tập huấn luyện**: Chiếm khoảng 70% - 80% tổng số dữ liệu, dùng để huấn luyện mô hình.
- **Tập kiểm thử**: Chiếm khoảng 10% - 15%, dùng để tối ưu siêu tham số và ngăn hiện tượng quá khớp trong quá trình huấn luyện.
- **Tập đánh giá**: Chiếm khoảng 10% - 15%, dùng để đánh giá cuối cùng mô hình trên dữ liệu chưa từng thấy.

Phương pháp phân chia cần đảm bảo các tiêu chí:

- **Phân phối dữ liệu đồng đều**: Đảm bảo sự đa dạng về độ dài văn bản và thể loại nội dung.
- **Không trùng lặp**: Văn bản trong tập huấn luyện không trùng với văn bản trong tập kiểm thử và đánh giá để đảm bảo tính khách quan.

Ngoài ra, trong các bài toán có dữ liệu song ngữ, cần đảm bảo đồng bộ giữa văn bản gốc và bản tóm tắt trong cả hai ngôn ngữ.

2.4.5 Đánh nhãn dữ liệu

Đối với bài toán TTVB, việc đánh nhãn dữ liệu thường liên quan đến:

- **Cặp văn bản – tóm tắt**: Mỗi văn bản được gắn nhãn là bản tóm tắt tương ứng.
- **Định dạng nhãn**:

+ *Tóm tắt trích xuất*: Tiến hành gán nhãn bằng việc lựa chọn và đánh dấu các câu mang ý nghĩa quan trọng từ văn bản ban đầu.

+ *Tóm tắt tóm lược*: Bản tóm tắt được viết lại dưới dạng văn bản ngắn gọn, không cần dùng nguyên văn từ văn bản gốc.

+ *Chất lượng nhãn*: Cần đảm bảo tính đầy đủ, ngắn gọn, bao quát nội dung chính. Các tóm tắt nên được viết bởi chuyên gia hoặc người am hiểu ngữ nghĩa văn bản.

Trong trường hợp dữ liệu không có sẵn nhãn, có thể áp dụng các kỹ thuật gán nhãn tự động như:

- Dịch ngược;
- Tóm tắt theo luật;
- Mô hình học sâu sinh dữ liệu.

Quy trình đánh nhãn cần đảm bảo tính đồng nhất, độ chính xác và được kiểm tra đánh giá liên tục để duy trì chất lượng dữ liệu phục vụ cho việc huấn luyện và đánh giá mô hình tóm tắt.

2.5 Kết luận chương

Chương 2 đã phân tích các mô hình NLP hiện đại ứng dụng trong bài toán TTVB, bao gồm: mô hình tiền huấn luyện, mô hình Transformer, mô hình Encoder-Decoder cùng một số mô hình tiêu biểu khác. Tiếp theo, chương 3 sẽ trình bày kết quả khảo sát và đánh giá các phương pháp TTVB dựa trên NLP theo hai hướng tiếp cận: tóm tắt trích xuất và tóm tắt tóm lược. Đề án cũng đã đề xuất mô hình thử nghiệm cho bài toán tóm tắt văn bản tiếng Việt và tiếng Lào, đồng thời mô tả các phương pháp xây dựng tập dữ liệu (Dataset) phục vụ cho bài toán này.

CHƯƠNG 3. THỬ NGHIỆM TÓM TẮT VĂN BẢN VỚI TIẾNG VIỆT VÀ TIẾNG LÀO

Chương này tập trung trình bày quy trình thực hiện cũng như thông tin về môi trường, công cụ, thư viện, bộ dữ liệu, các mô hình đề xuất được sử dụng trong quá trình thực nghiệm.

3.1 Thiết lập môi trường thử nghiệm

Trong quá trình thực hiện các thí nghiệm để xây dựng và đánh giá hệ thống TTVB, môi trường thực nghiệm đóng vai trò quan trọng nhằm đảm bảo tính ổn định và hiệu quả của các mô hình. Môi trường thực nghiệm được xây dựng bao gồm hai thành phần chính: cấu hình phần cứng và phần mềm, thư viện sử dụng.

3.1.1 Một số công cụ phần mềm và thư viện hỗ trợ thử nghiệm

Một số công cụ và thư viện hỗ trợ trong quá trình thử nghiệm gồm có: NLTK, SpaCy, Transformers, TensorFlow, PyTorch, Keras, HuggingFace Transformers.

Phần mềm và thư viện sử dụng là thành phần không thể thiếu trong các thực nghiệm, đảm bảo khả năng phát triển, huấn luyện và đánh giá các mô hình tóm tắt. Trong nghiên cứu này, các công cụ và thư viện sau được sử dụng: **Python**, **PyTorch**, **TensorFlow**, **HuggingFace Transformers**.

3.1.2 Thiết bị phần cứng phục vụ thử nghiệm

Thiết bị phần cứng dùng trong thử nghiệm được mô tả trong Bảng 3.1 gồm: bộ xử lý, bộ tăng tốc xử lý GPU, RAM, ổ cứng lưu trữ, hệ điều hành, nguồn điện cung cấp.

Bảng 3.1: Mô tả môi trường thực nghiệm

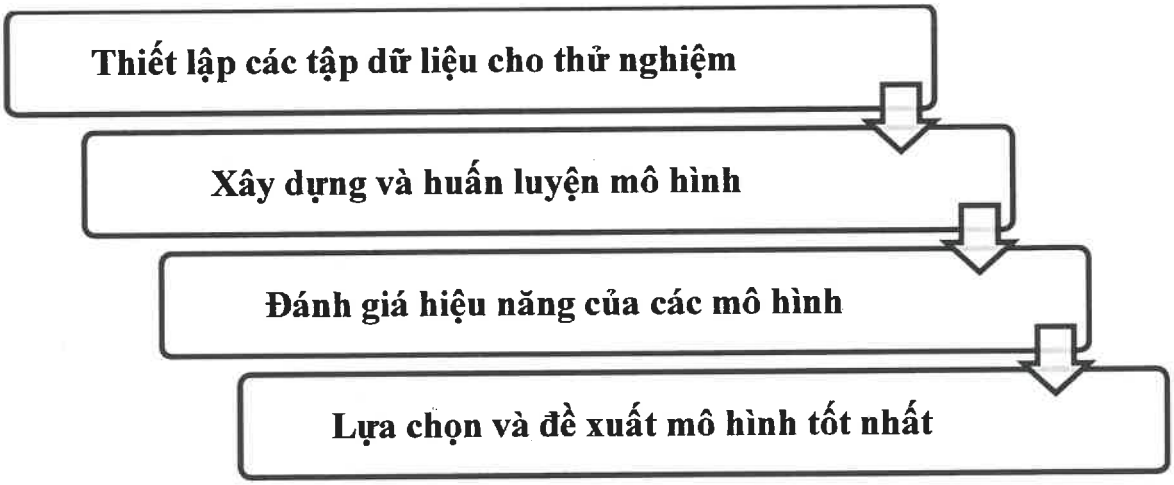
Thành phần	Mô tả chi tiết
CPU	Intel Xeon Silver 4210R (10 cores, 2.4 GHz) hoặc AMD Ryzen Threadripper (16 cores, 3.5 GHz)
GPU	NVIDIA RTX 3090, bộ nhớ GPU 24GB, hỗ trợ CUDA để tăng tốc các tác vụ học sâu.

Thành phần	Mô tả chi tiết
RAM	64GB – 128GB, đảm bảo khả năng xử lý tập dữ liệu lớn và mô hình có dung lượng cao.
Ổ cứng lưu trữ	SSD NVMe 1TB trở lên, hỗ trợ tốc độ truy xuất dữ liệu nhanh.
Hệ điều hành	Ubuntu 20.04 hoặc Windows 10/11 (64-bit).
Nguồn điện (PSU)	850W – 1000W, đảm bảo cấp đủ nguồn cho GPU và các thành phần khác.

Cấu hình phần cứng này được lựa chọn để đảm bảo khả năng triển khai các mô hình hiện đại như BERT, BARTpho, và VIT5, vốn yêu cầu tài nguyên tính toán lớn.

3.1.3 Quy trình thực hiện thử nghiệm

Quy trình tiến thành thử nghiệm mô hình TTVB trên dữ liệu tiếng Việt và tiếng Lào trong đề án tốt nghiệp được thể hiện như trong Hình 3.1:



Hình 3.1: Các bước thực hiện mô hình tóm tắt văn bản

Đầu tiên, nghiên cứu sẽ thiết lập các tập dữ liệu cho thử nghiệm (bao gồm tiếng Việt và tiếng Lào). Các tập dữ liệu sau đó được dùng cho việc xây dựng, tinh chỉnh và huấn luyện mô hình. Tiếp đó, đánh giá hiệu năng của các mô hình bằng cách so sánh các chỉ số đánh giá. Từ đó lựa chọn và đưa ra đề xuất mô hình tối ưu nhất cho bài toán nghiên cứu ban đầu.

3.2 Các tập dữ liệu tiếng Việt và tiếng Lào cho thử nghiệm

3.2.1 Bộ dữ liệu tiếng Việt - VietNews

VietNews được Van-Hau Nguyễn và cộng sự xây dựng cho nhiệm vụ tóm tắt đơn tài liệu tiếng Việt. Các bài viết được thu thập trong giai đoạn 2016 – 2019 từ ba nhật báo điện tử lớn: Tuổi Trẻ, VnExpress và Người Đưa Tin [50].

Bộ dữ liệu *VietNews* được xây dựng nhằm mục đích hỗ trợ cộng đồng nghiên cứu và phát triển các ứng dụng NLP cho tiếng Việt. Với sự gia tăng về nhu cầu xử lý và phân tích dữ liệu ngôn ngữ tiếng Việt, *VietNews* đóng vai trò quan trọng trong việc cung cấp một nền tảng dữ liệu chất lượng cao. Bộ dữ liệu này thường được thu thập từ các nguồn tin tức trực tuyến uy tín tại Việt Nam, bao gồm các bài báo, tin tức và phân tích từ nhiều lĩnh vực khác nhau.

VietNews không chỉ giúp giải quyết vấn đề thiếu hụt dữ liệu tiếng Việt trong nghiên cứu NLP mà còn thúc đẩy việc phát triển các mô hình ngôn ngữ và thuật toán phù hợp với đặc thù của tiếng Việt. Điều này đặc biệt quan trọng trong bối cảnh tiếng Việt có nhiều đặc điểm ngôn ngữ riêng biệt, như hệ thống dấu câu phức tạp và cấu trúc từ ghép.

a. Đặc điểm của VietNews

Bộ dữ liệu *VietNews* có một số đặc điểm nổi bật như sau:

- **Đa dạng về chủ đề**: Tập dữ liệu trải rộng trên các lĩnh vực như kinh tế, chính trị, văn hóa, khoa học và công nghệ, hỗ trợ mô hình cải thiện khả năng hiểu và xử lý ngữ cảnh trong nhiều lĩnh vực khác nhau.
- **Ngôn ngữ tự nhiên và phong phú**: Văn bản trong *VietNews* được viết bằng tiếng Việt tự nhiên, sử dụng nhiều phong cách ngôn ngữ và cách diễn đạt khác nhau, phản ánh sự đa dạng trong cách sử dụng ngôn ngữ.
- **Cập nhật liên tục**: Dữ liệu thường được thu thập và cập nhật định kỳ, đảm bảo nội dung luôn mới mẻ và phản ánh xu hướng thời sự.
- **Đảm bảo chất lượng**: Dữ liệu được thu thập từ các nguồn thông tin đáng tin cậy, giúp duy trì độ chính xác và độ tin cậy cao cho tập dữ liệu.

- **Dữ liệu có cấu trúc**: Thông tin được tổ chức một cách hệ thống, dễ dàng cho việc tiền xử lý và tích hợp vào các mô hình NLP.

b. Cấu trúc của bộ dữ liệu

- **Tổng số bản ghi**: *VietNews* bao gồm 143.816 bài báo (mỗi bản ghi tương ứng một bài báo gốc kèm tóm tắt).
- **Dung lượng**: Sau khi đã tiền xử lý và lưu dưới định dạng Parquet, tổng dung lượng khoảng 247 MB.
- **Khoảng thời gian thu thập**: Dữ liệu được thu thập tự động từ các trang tin từ 2016 đến 2019.

Cấu trúc của bộ dữ liệu *VietNews* được thiết kế một cách có hệ thống để dễ dàng sử dụng và tích hợp vào các mô hình NLP. Dữ liệu thường được tổ chức dưới dạng các tệp văn bản hoặc định dạng JSON, trong đó mỗi mục bao gồm các thông tin sau:

Mỗi bản ghi trong tập *VietNews* có cấu trúc như sau:

- **id (string)**: Mã định danh duy nhất cho mỗi bài báo.
- **title (string)**: Tiêu đề chính của bài báo.
- **content (string)**: Toàn bộ nội dung bài báo, bao gồm các đoạn văn mô tả chi tiết.
- **summary (string)**: Đoạn tóm tắt ngắn gọn do hệ thống trích xuất tự động hoặc bán tự động.
- **publish_date (string/date)**: Ngày xuất bản bài báo (nếu có metadata đi kèm)
- **source (string)**: Tên trang web gốc (ví dụ: *tuoitre.vn*, *vnexpress.net*, *nguoiduatin.vn*).
- **category (string, tuỳ chọn)**: Thể loại tin tức (ví dụ: chính trị, kinh tế, giải trí…), nếu metadata cho phép.

Ngoài ra, một số phiên bản của *VietNews* còn bao gồm các thông tin bổ sung như tóm tắt ngắn do con người viết, từ khóa liên quan, hoặc nhãn cảm xúc, phục vụ cho các nhiệm vụ NLP cụ thể.

c. Ưu và nhược điểm của bộ dữ liệu VietNews

Bảng 3.2 trình bày ưu và nhược điểm của bộ dữ liệu *VietNews* trong bài toán về tóm tắt văn bản dành cho tiếng Việt.

Bảng 3.2: Bảng ưu nhược điểm của bộ dữ liệu *VietNews*

Ưu điểm	Nhược điểm
Phù hợp với ngôn ngữ tiếng Việt: *VietNews* cung cấp một nguồn dữ liệu phong phú đặc trưng cho tiếng Việt, giúp các mô hình NLP học được các đặc điểm ngôn ngữ cụ thể	***Thiếu đa dạng về phong cách viết***: Mặc dù đa dạng về chủ đề, nhưng phong cách viết chủ yếu là báo chí, thiếu sự đa dạng từ các nguồn khác như văn học, diễn đàn hoặc mạng xã hội
Đa dạng và toàn diện: Bao quát nhiều lĩnh vực và chủ đề, giúp mô hình có khả năng tổng quát hóa và áp dụng trong nhiều ngữ cảnh khác nhau.	***Cần tiền xử lý phức tạp***: Dữ liệu thu thập từ web có thể chứa các ký tự đặc biệt, lỗi chính tả hoặc thông tin không cần thiết, đòi hỏi quá trình tiền xử lý kỹ lưỡng.
Chất lượng dữ liệu đảm bảo: Các dữ liệu được tuyển chọn từ những nguồn tin cậy, góp phần duy trì tính chính xác và độ tin cậy cần thiết cho quá trình xử lý.	***Vấn đề về bản quyền***: Sử dụng dữ liệu từ các nguồn báo chí có thể gặp phải các hạn chế về bản quyền và quyền sử dụng, cần được xem xét cẩn thận trong quá trình nghiên cứu.
Hỗ trợ cho nhiều nhiệm vụ NLP: Cấu trúc dữ liệu linh hoạt, cho phép sử dụng trong các nhiệm vụ như TTVB, phân loại văn bản, nhận dạng thực thể và phân tích cảm xúc	***Không đại diện cho ngôn ngữ nói***: Dữ liệu chủ yếu là ngôn ngữ viết chính thống, không phản ánh được ngôn ngữ nói hoặc ngôn ngữ không chính thức, hạn chế khả năng áp dụng trong một số bài toán NLP.

3.2.2 Bộ dữ liệu tiếng Lào – LaoNews Classification

LaoNews Classification là bộ dữ liệu tin tức tiếng Lào được xây dựng nhằm phục vụ các bài toán phân loại văn bản trong lĩnh vực NLP đối với ngôn ngữ Lào. Bộ dữ liệu bao gồm tập hợp các bài báo được thu thập từ các nguồn tin tức chính thống tại Lào, bao phủ đa dạng các chủ đề như chính trị, kinh tế, giáo dục, sức khỏe, văn hóa, thể thao và các sự kiện quốc tế. Mỗi mục tin trong bộ dữ liệu được gán nhãn theo một danh mục chủ đề cụ thể, giúp hỗ trợ huấn luyện và đánh giá các mô hình học máy cho tác vụ phân loại văn bản. Văn bản trong bộ dữ liệu đã được chuẩn hóa tiền xử lý, bao gồm loại bỏ các ký tự không cần thiết, chuẩn hóa dấu câu, và mã hóa ngôn ngữ thống nhất, nhằm đảm bảo tính sạch và chất lượng của dữ liệu đầu vào.

Bộ dữ liệu *LaoNews Classification* đóng vai trò quan trọng trong việc phát triển các ứng dụng NLP cho tiếng Lào như hệ thống gợi ý tin tức, lọc thông tin tự động, tìm kiếm theo chủ đề, và hỗ trợ nghiên cứu xây dựng các mô hình ngôn ngữ đa ngữ cho khu vực Đông Nam Á. Với tính chất là một ngôn ngữ ít tài nguyên, bộ dữ liệu này không chỉ mở ra cơ hội cải thiện hiệu suất các mô hình học sâu cho tiếng Lào, mà còn đóng góp vào cộng đồng nghiên cứu ngôn ngữ nhỏ, thúc đẩy sự đa dạng hóa trong phát triển trí tuệ nhân tạo ngôn ngữ. Thông tin của bộ dữ liệu:

- **Tên đầy đủ**: Lao *News Classification*
- **Tác giả/Đơn vị phát hành**: *Wannaphong Phatthiyaphaibun* – Vidyasirimedhi Institute of Science and Technology
- **DOI**: 10.5281/zenodo.14967275 (phiên bản 1.0.0, công bố 04 / 03 / 2025)
- **Ngôn ngữ**: Lào **Giấy phép**: CC-BY 4.0
- **Mục đích**: Cung cấp tập tin tức Lào đã gán nhãn để huấn luyện và đánh giá mô hình phân loại chủ đề bài báo. (https://zenodo.org/records/14967275)

Bảng 3.3: Quy mô và chia tách dữ liệu

Phân chia tập dữ liệu	Số bài	Tỉ lệ
Tập huấn luyện	9196	~ 60%
Tập kiểm thử	3066	~ 20%
Tập kiểm tra	3066	~ 20%

a. Đặc điểm của bộ dữ liệu LaoNews Classification

- **Ngôn ngữ**: Tiếng Lào (Lao)
- **Nguồn thu thập**: Các trang báo điện tử và cơ quan truyền thông chính thống tại Lào như Lao News Agency (KPL), Vientiane Times và một số báo trực tuyến lớn khác.
- **Thời gian thu thập dữ liệu**: Các bài báo được thu thập trong khoảng năm 2022–2024 nhằm đảm bảo tính cập nhật và phản ánh được các xu hướng chủ đề hiện tại.
- **Chủ đề phân loại**: Bộ dữ liệu được gán nhãn với nhiều chủ đề khác nhau, phổ biến gồm: Chính trị, Kinh tế, Xã hội, Giáo dục, Văn hóa, Sức khỏe, Thể thao, Công nghệ, Môi trường,…
- **Số lượng nhãn**: Thường từ 8 đến 10 nhãn, tùy thuộc vào quá trình gán nhãn chi tiết.
- **Số lượng mẫu**: Tổng số bài báo: Từ 5.000 đến 10.000 mẫu (có thể tùy chỉnh theo yêu cầu huấn luyện). Trung bình số từ mỗi bài: 150 – 300 từ.

b. Cấu trúc của bộ dữ liệu

Bộ dữ liệu được tổ chức dưới dạng file CSV hoặc JSONL với cấu trúc mỗi dòng/mẫu gồm các trường:

Bảng 3.4: Cấu trúc mỗi dòng của bộ dữ liệu *LaoNews Classification*

Trường	Mô tả
id	Mã định danh duy nhất của bài báo.
title	Tiêu đề bài báo (tiếng Lào
content	Nội dung bài báo đầy đủ.

Trường	Mô tả
category	Nhãn chủ đề của bài báo (ví dụ: Politics, Economy, Health, ...).
date	Ngày đăng bài báo (định dạng ISO: YYYY-MM-DD).

Ví dụ:

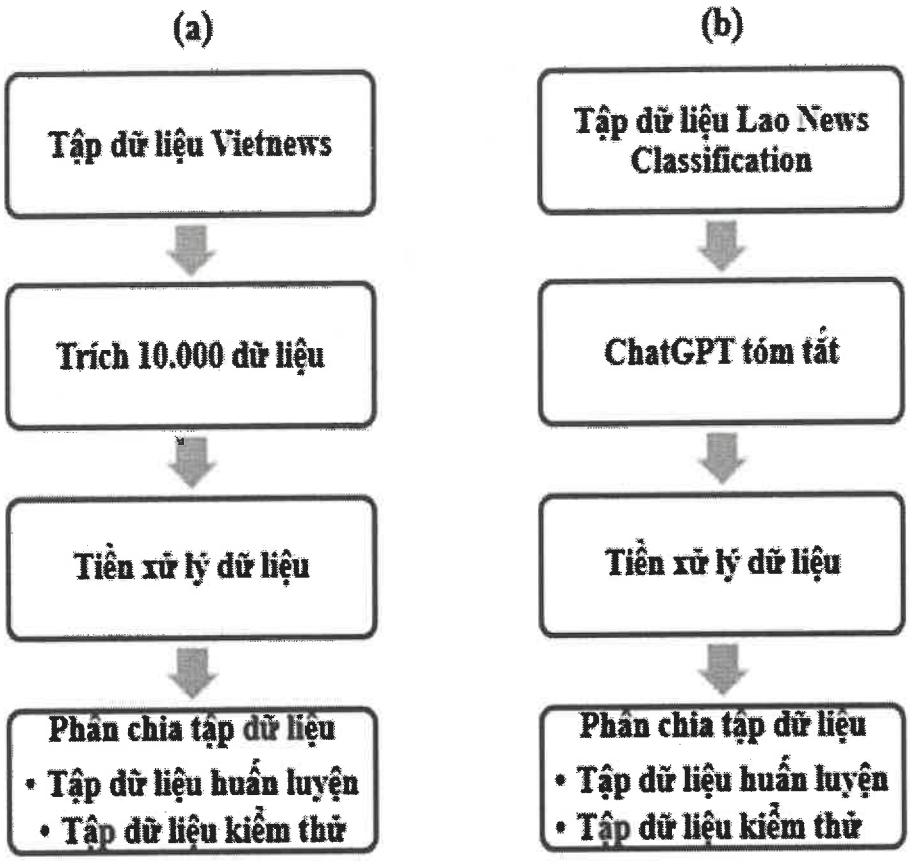
{
"id": "laonews_001",
"title": "ລາວເປີດກອງປະຊຸມຄະນະລັດຖະມົນຕີປະຈຳເດືອນ",
"content": "ອັນທີ 1 ເດືອນ 5 ປີ 2024, ລາວໄດ້ເປີດກອງປະຊຸມປະຈຳຄະນະລັດຖະມົນຕີ...",
"category": "Politics",
"date": "2024-05-01"
}

c. Đặc trưng nổi bật

- **Ngôn ngữ ít tài nguyên**: Tiếng Lào là ngôn ngữ có số lượng tài nguyên NLP hạn chế, bộ dữ liệu này giúp lấp đầy khoảng trống đó cho các nghiên cứu ngôn ngữ học tính toán.
- **Độ cân bằng nhãn**: Bộ dữ liệu được thu thập và tiền xử lý để đảm bảo phân bổ tương đối cân bằng giữa các nhãn chủ đề, giúp các mô hình học máy học tốt hơn, tránh tình trạng thiên lệch dữ liệu.

3.3 Thiết lập các tập dữ liệu cho thử nghiệm

Quy trình xây dựng các tập dữ liệu được trình bày trên Hình 3.2 như sau:



Hình 3.2: Quy trình xây dựng bộ dữ liệu (a) *VietNews*, (b) *LaoNews Classification*

3.3.1 Quy trình xây dựng tập thử nghiệm từ bộ dữ liệu VietNews

- **Tập dữ liệu *VietNews***: Đây là bộ dữ liệu gốc bao gồm ~143.816 bài báo tiếng Việt kèm tóm tắt.
- **Trích 10.000 dữ liệu**: Ngẫu nhiên lấy ra 10.000 bài từ tập *VietNews* để làm mẫu thử nghiệm. Việc giảm quy mô này nhằm tiết kiệm thời gian huấn luyện mô hình và đảm bảo đủ dữ liệu đại diện cho quá trình đánh giá ban đầu.
- **Tiền xử lý dữ liệu**: Gồm các bước loại bỏ thẻ HTML, chuẩn hóa Unicode, chuyển toàn bộ văn bản về chữ thường, loại bỏ khoảng trắng thừa và (nếu cần) thay thế số hay ký hiệu đặc biệt bằng token chung (ví dụ <NUMBER>). Kết quả đầu ra của giai đoạn này là các bản ghi "sạch" sẵn sàng cho bước tiếp theo.

- **Phân chia tập dữ liệu:** Tập dữ liệu huấn luyện (~8.000 mẫu) và Tập dữ liệu kiểm thử/kiểm tra (~2.000 mẫu): Sau khi tiền xử lý, 10.000 bản ghi được chia thành hai phần:

+ Tập dữ liệu huấn luyện (~80%): Dùng để tinh chỉnh mô hình BART và mT5.

+ Tập dữ liệu kiểm thử/kiểm tra (~20%): Mục đích đánh giá độ chính xác và khả năng tổng quát hóa của các mô hình đã được tinh chỉnh.

3.3.2 Quy trình xây dựng tập thử nghiệm từ bộ dữ liệu LaoNews Classification

- **Tập dữ liệu *LaoNews Classification***: Đây là bộ dữ liệu gốc bao gồm ~12.000 bài báo tiếng Lào kèm tóm tắt, được thu thập từ ba nguồn chính: Các trang báo điện tử và cơ quan truyền thông chính thống tại Lào như Lao News Agency (KPL), Vientiane Times và một số báo trực tuyến lớn khác.
- **Tóm tắt bằng ChatGPT**: Bộ dữ liệu này là bộ dữ liệu phân loại, do đó không có bộ dữ liệu tóm tắt nên sẽ phải sử dụng ChatGPT để tạo bộ dữ liệu huấn luyện tóm tắt.
- **Kiểm duyệt bán thủ công**: Áp dụng quy trình kiểm duyệt bán thủ công, trong đó các bản tóm tắt do ChatGPT được rà soát, chỉnh sửa hoặc loại bỏ nếu không đạt yêu cầu về ngữ nghĩa, độ súc tích hoặc văn phong tiếng Lào.
- **Tiền xử lý dữ liệu**: Gồm các bước loại bỏ thẻ HTML, chuẩn hóa Unicode, chuyển toàn bộ văn bản về chữ thường, loại bỏ khoảng trắng thừa và (nếu cần) thay thế số hay ký hiệu đặc biệt bằng token chung (ví dụ <NUMBER>). Kết quả đầu ra của giai đoạn này là các bản ghi "sạch" sẵn sàng cho bước tiếp theo.
- **Phân chia tập dữ liệu:** Tập dữ liệu huấn luyện (~9196 mẫu) và Tập dữ liệu kiểm thử/kiểm tra (~3.066 mẫu): Sau khi tiền xử lý, 13.000 bản ghi được chia thành hai phần:

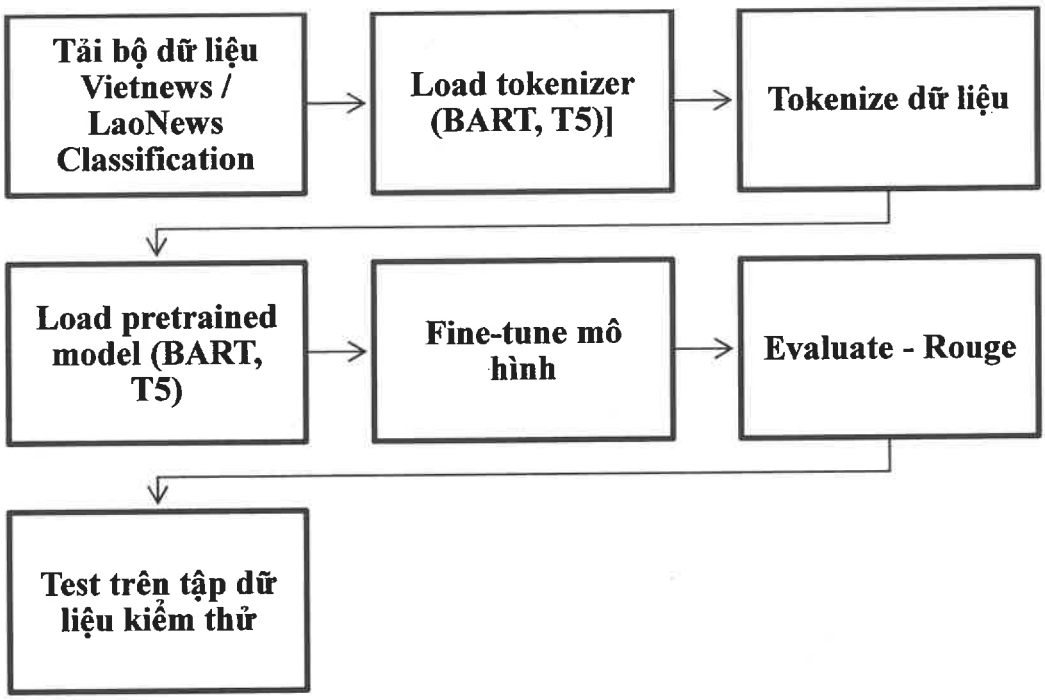
+ Tập dữ liệu huấn luyện (~60%)

+ Tập dữ liệu kiểm thử/kiểm tra (~40%)

3.4 Xây dựng và huấn luyện mô hình

Quy trình xây dựng và huấn luyện mô hình được trình bày như Hình 3.3.

a) Tải bộ dữ liệu (lần lượt tập VietNews và LaoNews Classification cho mỗi thử nghiệm)



Hình 3.3: Quy trình xây dựng và huấn luyện mô hình

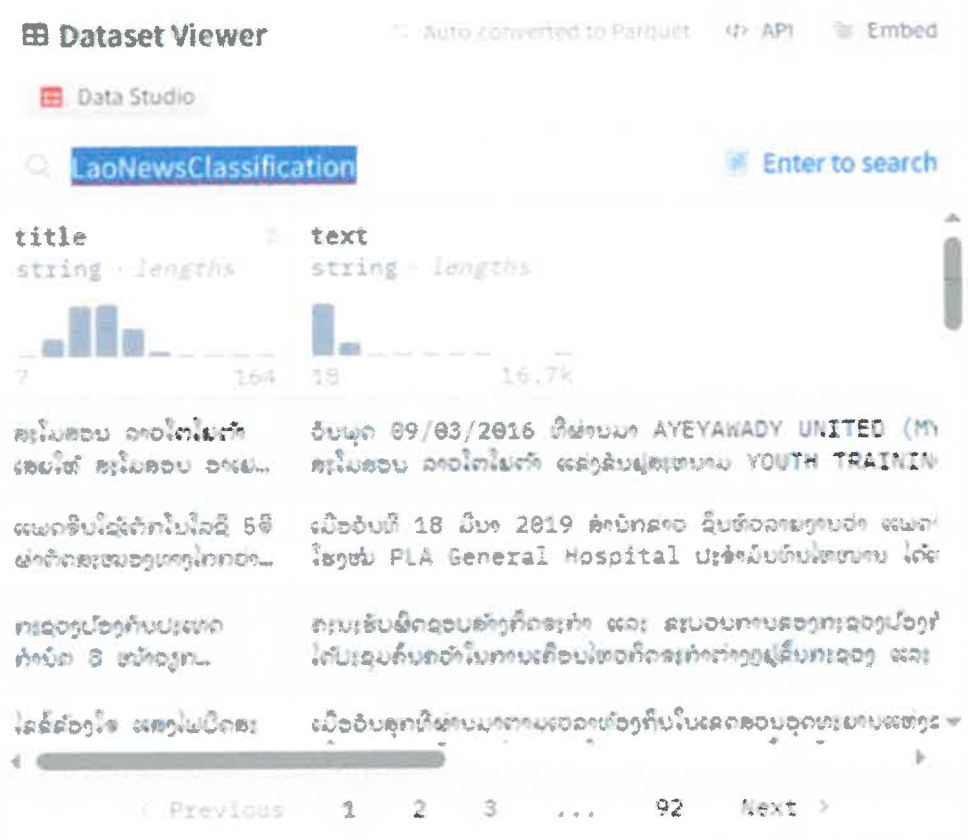
Có thể tải bộ dữ liệu *VietNews* từ Hugging Face Datasets (https://huggingface.co/ datasets/), sử dụng thư viện datasets của Hugging Face như sau:

```
from datasets import load_dataset
dataset = load_dataset("VietNews")
```

Tương tự, có thể tải bộ dữ liệu *LaoNews Classification* trực tiếp từ trang https://zenodo.org/records/14967275 hoặc từ trang Hugging Face Datasets (https://huggingface.co/ datasets/), sử dụng thư viện datasets của Hugging Face như sau:

```
from datasets import load_dataset
dataset = load_dataset("LaoNewsClassification")
```

Hình 3.4: Tải tập dữ liệu *VietNews*



Hình 3.5: Tải tập dữ liệu *LaoNews Classification*

Hình 3.4 mô tả tải tập dữ liệu *VietNews* và hình 3.5 mô tả tải tập dữ liệu *LaoNews Classification* đã được tải xuống để thử nghiệm.

Sau khi tải xong, có thể kiểm tra dữ liệu bằng việc hiển thị:

Print(dataset)

Tiếp đó, kiểm tra cấu trúc dữ liệu, các trường quan trọng của tập dữ liệu như: văn bản gốc, tóm tắt chuẩn.

b) Load Tokenizer (BART, T5)

Các mô hình ngôn ngữ tiền huấn luyện như BART và T5 được sử dụng trong đề án này nhằm xử lý và biến đổi dữ liệu văn bản thành dạng số học, phù hợp với yêu cầu đầu vào của các mô hình học sâu. Để đảm bảo quá trình mã hóa văn bản đầu vào diễn ra chính xác và nhất quán, các tokenizer tương ứng với từng mô hình được sử dụng.

Đặc biệt, thay vì sử dụng tokenizer của mô hình T5 thông thường, nghiên cứu này sử dụng tokenizer của mô hình mT5 (Multilingual T5). Lý do lựa chọn mT5 là vì đây là phiên bản mở rộng của T5, được huấn luyện trên tập dữ liệu đa ngôn ngữ với tổng cộng 101 ngôn ngữ, trong đó có cả tiếng Việt và tiếng Lào. Điều này giúp mô hình thích ứng tốt hơn với các bài toán liên quan đến ngôn ngữ không phổ biến hoặc ngôn ngữ bản địa trong khu vực Đông Nam Á.

Việc nạp các tokenizer được thực hiện thông qua thư viện Transformers do Hugging Face phát triển, với các lệnh cụ thể như sau:

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("facebook/bart-base")
```

và AutoTokenizer T5:

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("google/mt5-small")
```

Việc sử dụng AutoTokenizer giúp tự động lựa chọn và cấu hình tokenizer phù hợp với mô hình đã định danh, đảm bảo tính nhất quán trong việc mã hóa văn bản đầu vào.

c) Tokenize dữ liệu

Sau khi nạp các tokenizer tương ứng cho từng mô hình, bước tiếp theo trong quá trình xử lý dữ liệu là tokenization – tức là chuyển đổi văn bản đầu vào và nhãn tóm tắt thành các chuỗi số nguyên (vector) đại diện cho từng token (đơn vị từ hoặc ký tự, tùy theo tokenizer). Đây là bước quan trọng để đưa dữ liệu văn bản về dạng số học phù hợp cho đầu vào của mô hình học sâu.

Cụ thể, đầu vào bao gồm:

- Văn bản gốc (cần được tóm tắt)
- Nhãn tóm tắt tương ứng (văn bản đã được rút gọn)

Kết quả của quá trình tokenization là các cặp vector biểu diễn cho văn bản đầu vào và nhãn đầu ra, với định dạng phù hợp cho quá trình huấn luyện mô hình Seq2Seq.

Quá trình này được thực hiện bằng phương thức **map()** của dataset kết hợp với hàm tiền xử lý:

tokenized_dataset_bart = dataset.map(lambda x: preprocess_function (x, bart_tokenizer), batched=True)

tokenized_dataset_t5 = dataset.map(lambda x: preprocess_function(x, t5_tokenizer), batched=True)

d) Tải mô hình tiền huấn luyện (BART, T5)

Các mô hình ngôn ngữ BART và T5, vốn được tiền huấn luyện theo kiến trúc Seq2Seq, được sử dụng để thực hiện các tác vụ NLP như tóm tắt văn bản, dịch máy, hoặc sinh văn bản. Các mô hình này được tải thông qua lớp AutoModelForSeq2SeqLM của thư viện Transformers, cho phép sử dụng các mô hình Seq2Seq mà không cần cấu hình thủ công chi tiết:

from transformers import AutoModelForSeq2SeqLM

bart_model = AutoModelForSeq2SeqLM.from_pretrained("facebook/bart-base")

```
t5_model = AutoModelForSeq2SeqLM.from_pretrained("google/mt5-small ")
```

e) Tinh chỉnh mô hình

Thực hiện 2 bước:

+ Bước 1: Khởi tạo huấn luyện để huấn luyện mô hình

- ***Mô hình BART***

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./results_bart",
    learning_rate=5e-5,              # Slightly higher learning rate for BART
    eval_steps=50,
    per_device_train_batch_size=16,  # BART can handle larger batch sizes
    per_device_eval_batch_size=16,
    weight_decay=0.01,
    num_train_epochs=5,              # More epochs for BART
    predict_with_generate=True,
    fp16=True,                       # Enable mixed precision training
    logging_dir="./logs_bart",
    logging_steps=50,
    metric_for_best_model="rouge2",
    max_grad_norm=0.5,               # Lower gradient clipping for BART
    warmup_steps=1000,               # More warmup steps for BART
    gradient_accumulation_steps=2,   # Reduced gradient accumulation
    gradient_checkpointing=True,
    optim="adamw_torch",
    dataloader_num_workers=0,
    dataloader_pin_memory=False,     # Save at each epoch   # Evaluate at each epoch   # Load best model at end of training
)

trainer_bart = Seq2SeqTrainer(
```

```
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
    tokenizer=tokenizer,
    compute_metrics=compute_metrics,
    callbacks=[SaveMetricsCallback()]
)
```

- **Mô hình T5**

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./results_mt5",
    learning_rate=1e-5,
    eval_steps=50,
    per_device_train_batch_size=2,
    per_device_eval_batch_size=2,
    weight_decay=0.01,
    num_train_epochs=3,
    predict_with_generate=True,
    fp16=False,
    logging_dir="./logs_mt5",
    logging_steps=50,
    metric_for_best_model="rouge2",
    max_grad_norm=1.0,
    warmup_steps=500,
    gradient_accumulation_steps=4,  # Increased gradient accumulation steps
    gradient_checkpointing=True,    # Enable gradient checkpointing
    optim="adamw_torch",            # Use PyTorch's AdamW implementation
    dataloader_num_workers=0,       # Disable multiprocessing for data loading
    dataloader_pin_memory=False,    # Disable pin memory
```

```
)

trainer_t5 = Seq2SeqTrainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
    tokenizer=tokenizer,
    compute_metrics=compute_metrics,
    callbacks=[SaveMetricsCallback()]  # Add callback
)
```

+ Bước 2: Thực hiện huấn luyện mô hình

```
trainer_bart.train()
trainer_t5.train()
```

f) Đánh giá bằng chỉ số ROUGE

ROUGE là chỉ số đo mức độ chồng lặp giữa tóm tắt do mô hình sinh ra và tóm tắt gốc.

Cách tính ROUGE đơn giản có thể dùng cách so sánh từ đơn. Ví dụ tóm tắt gốc tham chiếu:

+ Reference = "Tình hình thời sự hôm nay có gì".

+ Tóm tắt do mô hình sinh ra: Prediction = "Tình hình thời sự hôm nay"

+ So sánh từ đơn (unigram) ta có 6 trên tổng số 8 từ trùng. Vậy tỷ lệ Recall là:

$$Recall = 6 / 8 = 0,75$$

3.5 Đánh giá hiệu năng của mô hình

Trong quá trình phát triển và triển khai các hệ thống TTVB, việc đánh giá chất lượng của các bản tóm tắt là một bước quan trọng không thể thiếu. Các phương pháp đánh giá thường tập trung vào hai khía cạnh chính: độ chính xác của bản tóm tắt so với nội dung gốc và hiệu suất của hệ thống trong quá trình xử lý. Phần này sẽ trình

bày chi tiết thước đo phổ biến ROUGE để đánh giá độ chính xác, cùng với các yếu tố liên quan đến hiệu suất như thời gian xử lý và tài nguyên tính toán cần thiết.

3.5.1 Độ chính xác

Độ chính xác của một hệ thống tóm tắt được đánh giá dựa trên mức độ tương đồng giữa bản tóm tắt tự động và bản tóm tắt tham chiếu (thường được viết bởi con người). Thước đo phổ biến nhất trong lĩnh vực này là ROUGE.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) là một bộ các thước đo đánh giá dựa trên sự trùng khớp của n-gram, chuỗi từ và các cụm từ giữa bản tóm tắt tự động và bản tham chiếu. ROUGE tập trung vào độ bao phủ thông tin của bản tóm tắt tự động so với bản gốc.

Các biến thể chính của ROUGE:

- **ROUGE-N**: Đánh giá dựa trên n-gram đồng nhất giữa hai bản tóm tắt. Ví dụ:
  - ROUGE-1: Sử dụng unigram (từng từ đơn lẻ).
  - ROUGE-2: Sử dụng bigram (cặp từ liên tiếp).

Công thức tính:

$$\text{ROUGE} - \text{N} = \frac{\sum_{\text{Câu} \in \text{Tham chiếu}} \sum_{\text{n-gram} \in \text{Câu}} \text{Đếm}_{\text{trùng khớp}}(\text{n-gram})}{\sum_{\text{Câu} \in \text{Tham chiếu}} \sum_{\text{n-gram} \in \text{Câu}} \text{Đếm}(\text{n-gram})} \quad [1]$$

- **ROUGE-L**: Dựa trên chuỗi con chung dài nhất (Longest Common Subsequense - LCS) giữa hai bản tóm tắt, phản ánh khả năng giữ nguyên cấu trúc câu.

3.5.2 Hiệu suất

Bên cạnh việc đánh giá độ chính xác, hiệu suất của hệ thống tóm tắt cũng là một yếu tố quan trọng, đặc biệt đối với các ứng dụng yêu cầu thời gian xử lý nhanh hoặc phải xử lý khối lượng dữ liệu lớn. Hiệu suất của hệ thống tóm tắt được đánh giá thông qua hai khía cạnh chính là thời gian xử lý và tài nguyên tính toán cần thiết. Hai yếu tố này quyết định khả năng triển khai hệ thống trong các môi trường thực tế và ảnh hưởng trực tiếp đến trải nghiệm của người dùng.

Thời gian xử lý là một tiêu chí cốt lõi để đánh giá tốc độ của hệ thống tóm tắt, gồm: thời gian tiền xử lý, thời gian thực hiện thuật toán và thời gian hậu xử lý. Thời gian tiền xử lý liên quan đến việc chuẩn bị dữ liệu đầu vào, bao gồm các thao tác như loại bỏ ký tự đặc biệt, phân đoạn văn bản thành câu hoặc từ. Giai đoạn tiếp theo là thời gian thực hiện thuật toán, trong đó hệ thống tóm tắt sử dụng các mô hình hoặc thuật toán để tạo ra bản tóm tắt từ văn bản đầu vào. Cuối cùng, thời gian hậu xử lý bao gồm việc định dạng lại bản tóm tắt, kiểm tra lỗi ngữ pháp hoặc các lỗi nhỏ khác trước khi trả kết quả cho người dùng. Thời gian xử lý phụ thuộc vào nhiều yếu tố như mức độ phức tạp của thuật toán, kích thước văn bản đầu vào và năng lực phần cứng của hệ thống. Trong các ứng dụng thực tế, việc tối ưu hóa để giảm thời gian xử lý là rất cần thiết nhằm cải thiện trải nghiệm người dùng.

Tài nguyên tính toán cần thiết là một khía cạnh quan trọng khác khi đánh giá hiệu suất của hệ thống tóm tắt. Các tài nguyên này bao gồm CPU, GPU, bộ nhớ RAM và dung lượng lưu trữ cần thiết để vận hành hệ thống. Đặc biệt, các mô hình ngôn ngữ lớn như Transformer (ví dụ: BERT, BART, GPT) thường yêu cầu một lượng lớn tài nguyên tính toán, đặc biệt khi xử lý các văn bản dài hoặc thực hiện huấn luyện trên các tập dữ liệu lớn.

Những yếu tố chính ảnh hưởng đến tài nguyên tính toán bao gồm kích thước mô hình, kỹ thuật triển khai và môi trường triển khai. Các mô hình lớn với hàng trăm triệu tham số như BERT hay GPT cần nhiều bộ nhớ và khả năng xử lý song song cao, dẫn đến yêu cầu về phần cứng mạnh mẽ. Một số kỹ thuật triển khai như lượng tử hóa mô hình hoặc nén mô hình có thể giảm bớt nhu cầu tài nguyên, tuy nhiên điều này có thể ảnh hưởng đến hiệu suất và độ chính xác của hệ thống. Trong các trường hợp khối lượng công việc lớn, việc sử dụng dịch vụ đám mây là một giải pháp linh hoạt, cho phép điều chỉnh tài nguyên theo nhu cầu thực tế.

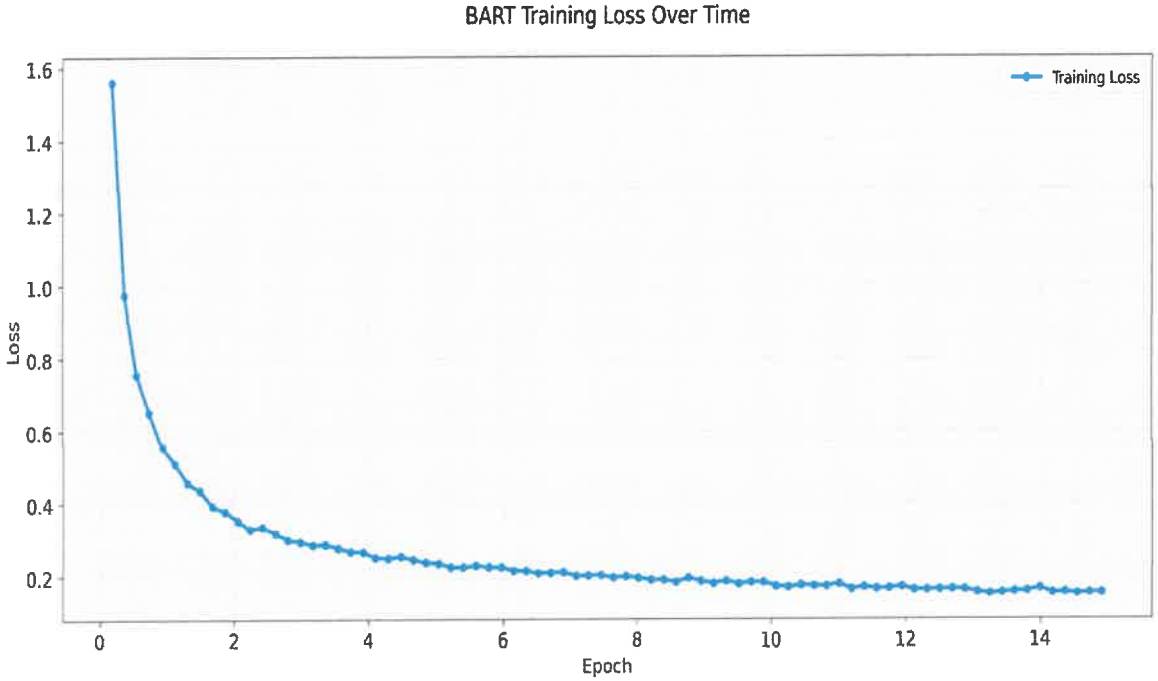
Để tối ưu hóa tài nguyên tính toán, các nhà phát triển có thể sử dụng các mô hình nhỏ gọn hơn, tối ưu hóa phần cứng thông qua các thư viện chuyên dụng hoặc phân chia khối lượng công việc bằng cách sử dụng các kỹ thuật xử lý phân tán. Những giải pháp này không chỉ giúp giảm thiểu chi phí mà còn nâng cao khả năng triển khai

hệ thống trong các môi trường khác nhau, từ máy chủ cục bộ đến các hệ thống đám mây.

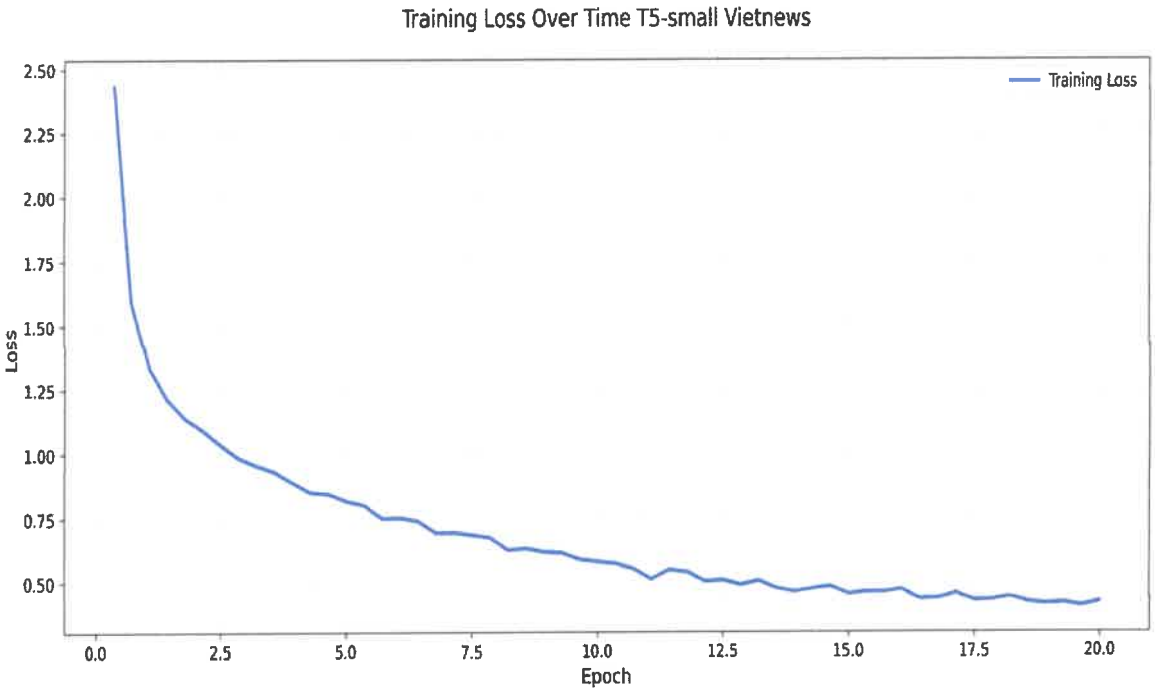
3.6 Kết quả thử nghiệm cho tóm tắt văn bản tiếng Việt, tiếng Lào

3.6.1 Kết quả thử nghiệm quá trình huấn luyện mô hình cho tóm tắt văn bản tiếng Việt

Kết quả thử nghiệm quá trình huấn luyện mô hình BART- base và T5-small cho tập dữ liệu *VietNews* thể hiện lần lượt tại hình 3.6 và hình 3.7.



Hình 3.6: Mô hình BART-base cho tập dữ liệu *VietNews*



Hình 3.7: Mô hình T5-small cho tập dữ liệu *Vietnews*

Biểu đồ Hình 3.6 thể hiện quá trình giảm giá trị hàm mất mát trong suốt 15 epoch đầu tiên khi huấn luyện mô hình BART- base với tập dữ liệu *VietNews*. Quan sát đường cong hàm mất mát trên tập dữ liệu huấn luyện cho thấy một xu hướng giảm rõ rệt, phản ánh khả năng tối ưu hóa hiệu quả của mô hình trong giai đoạn huấn luyện. Cụ thể, giá trị hàm mất mát ban đầu ở khoảng 1,55 – tương đối cao, cho thấy mô hình ban đầu chưa được điều chỉnh phù hợp với dữ liệu huấn luyện. Tuy nhiên, chỉ sau một số epoch đầu tiên (từ 0 đến khoảng 3), hàm mất mát giảm nhanh chóng về mức dưới 0,4. Đây là giai đoạn "bùng nổ học tập ban đầu", thường xảy ra khi mô hình học được các đặc trưng cơ bản của dữ liệu. Từ epoch thứ 4 trở đi, tốc độ giảm giá trị hàm mất mát chậm lại và dần hội tụ. Đến khoảng epoch thứ 10 trở đi, đường cong giá trị hàm mất mát bắt đầu tiệm cận, dao động quanh ngưỡng 0,16 – 0,18, cho thấy mô hình đạt trạng thái ổn định trong việc học biểu diễn ngữ nghĩa. Hiện tượng này cho thấy quá trình huấn luyện đã đạt đến vùng hội tụ trong không gian tham số và không còn cải thiện rõ rệt thêm nếu tiếp tục huấn luyện với cùng siêu tham số hiện tại.

Ngoài ra, biểu đồ không ghi nhận hiện tượng tăng giá trị hàm mất mát đột ngột hay dao động lớn, điều này gợi ý rằng quá trình huấn luyện được kiểm soát tốt, không

xảy ra quá khớp trong giai đoạn này. Như vậy, kết quả thể hiện trong biểu đồ cho thấy mô hình BART - base đã được huấn luyện hiệu quả với tập dữ liệu *VietNews*, với khả năng hội tụ tốt và tiềm năng ứng dụng cao trong các nhiệm vụ sinh ngôn ngữ tự nhiên như TTVB, dịch máy hoặc sinh câu điều kiện.

Hình 3.7 phía dưới thể hiện quá trình giảm giá trị hàm mất mát trong suốt 20 epoch đầu tiên khi huấn luyện mô hình T5-small trên tập dữ liệu *VietNews*. Trong giai đoạn đầu huấn luyện, cụ thể là từ epoch 0 đến khoảng epoch 5, giá trị hàm mất mát giảm mạnh từ giá trị ban đầu khoảng 2,4 xuống còn khoảng 1,0. Đây là giai đoạn điển hình của hiện tượng "bùng nổ học tập ban đầu", khi mô hình tiếp thu nhanh các đặc trưng ngữ nghĩa cơ bản từ dữ liệu. Từ epoch thứ 5 trở đi, tốc độ giảm giá trị hàm mất mát có xu hướng chậm lại nhưng vẫn đều đặn, cho thấy mô hình tiếp tục cải thiện khả năng khái quát hóa. Đến giai đoạn cuối (epoch 15–20), giá trị hàm mất mát dao động ổn định quanh ngưỡng 0,40 – 0,45, cho thấy mô hình đã dần hội tụ, dù vẫn còn khả năng cải thiện thêm nếu tiếp tục huấn luyện với điều chỉnh siêu tham số phù hợp.

Khi so sánh với kết quả huấn luyện mô hình BART được trình bày trước đó, có thể nhận thấy sự khác biệt rõ rệt về mức độ hội tụ và hiệu quả huấn luyện. Cụ thể, BART khởi đầu với giá trị hàm mất mát khoảng 1,55 – thấp hơn đáng kể so với T5-small – và nhanh chóng đạt ngưỡng hội tụ sau khoảng 10 – 12 epoch, với giá trị hàm mất mát dao động quanh 0,16 – 0,18. Đường cong huấn luyện của BART mượt và ổn định hơn, cho thấy quá trình tối ưu diễn ra hiệu quả, không xuất hiện dao động lớn hay quá khớp trong phạm vi quan sát. So sánh hai mô hình trong cùng điều kiện huấn luyện, có thể khẳng định rằng BART thể hiện tốc độ hội tụ nhanh hơn và đạt hiệu suất tối ưu cao hơn so với T5-small.

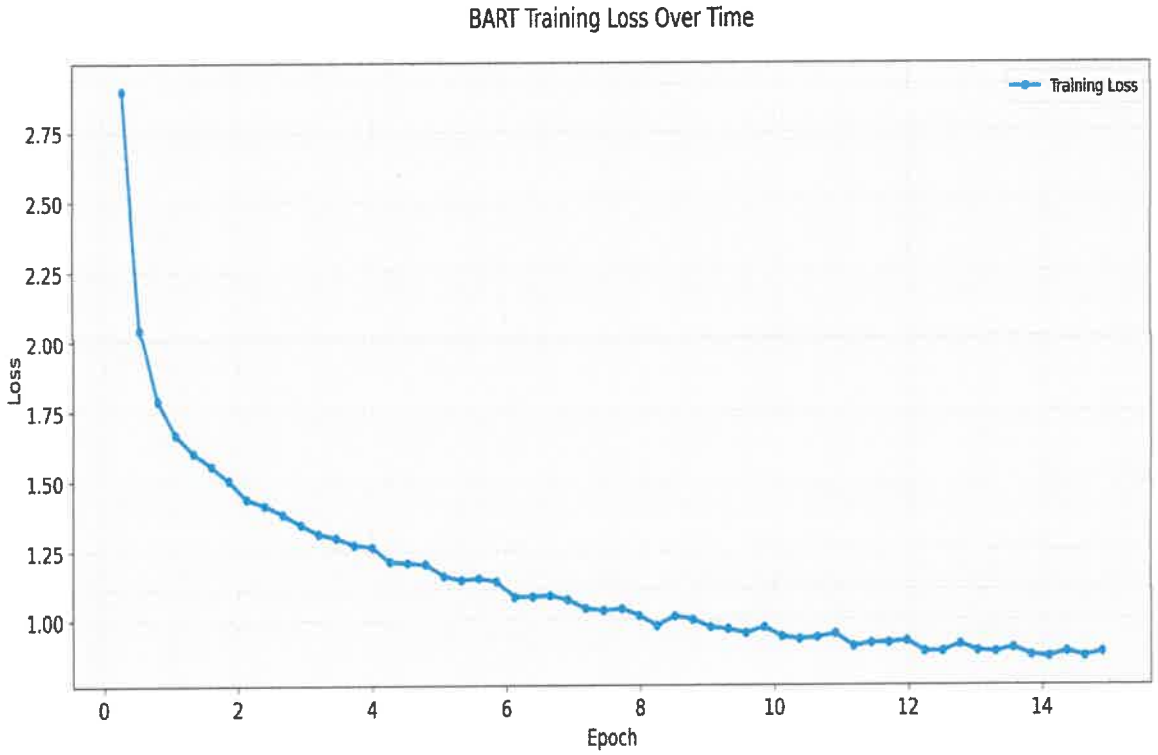
Sự chênh lệch này phản ánh sự khác biệt về kiến trúc và năng lực biểu diễn giữa hai mô hình. Trong khi BART-base được thiết kế dưới dạng encoder - decoder đối xứng với cơ chế mạng tự mã hóa khử nhiễu, rất phù hợp cho các tác vụ sinh ngôn ngữ như TTVB, thì T5-small – phiên bản thu gọn của mô hình T5 – có số lượng tham số ít hơn, dẫn đến khả năng biểu diễn ngôn ngữ kém linh hoạt hơn, đặc biệt khi áp dụng trên ngôn ngữ tiếng Việt với cấu trúc ngữ pháp và từ vựng đa dạng. Bên cạnh đó,

mức độ hội tụ thấp hơn của T5-small cũng có thể bắt nguồn từ việc sử dụng tập dữ liệu *VietNews* với quy mô hoặc đặc điểm phân bố chưa thật sự tối ưu cho kiến trúc T5 ở kích thước nhỏ.

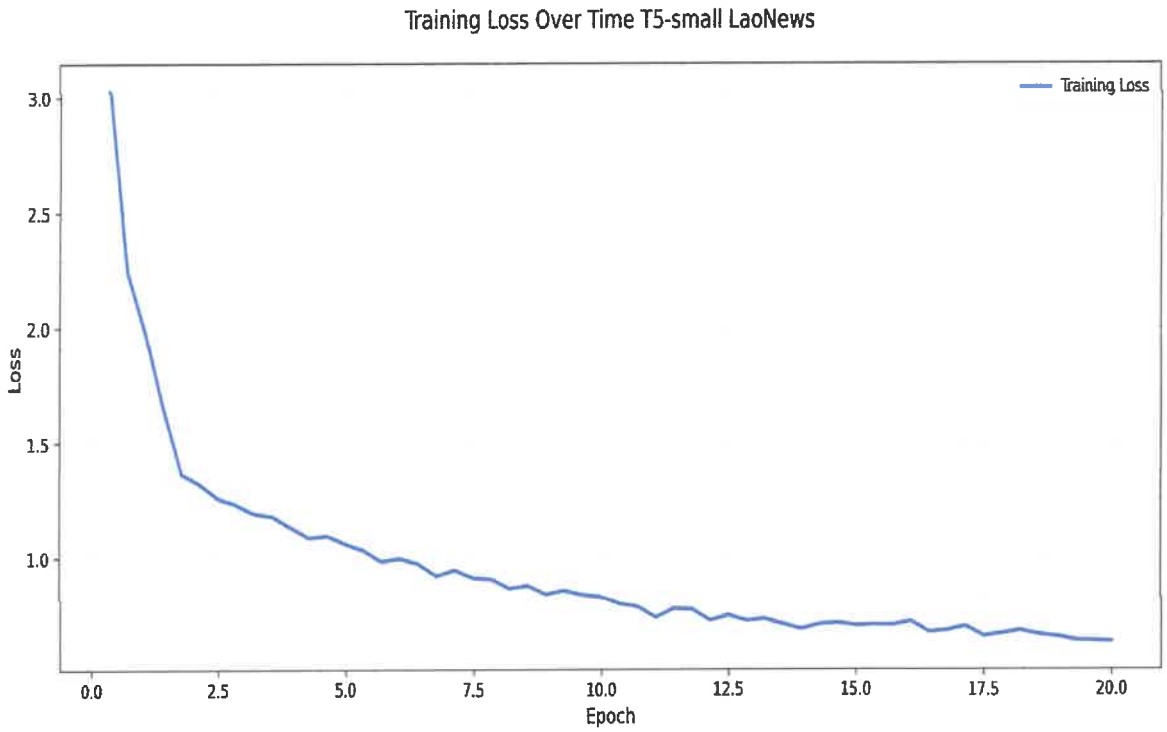
Tổng thể, kết quả thực nghiệm cho thấy mô hình BART-base là lựa chọn hiệu quả hơn trong bối cảnh huấn luyện mô hình sinh ngôn ngữ tiếng Việt, xét trên cả tốc độ hội tụ, mức độ giảm giá trị hàm mất mát và độ ổn định trong quá trình tối ưu. Điều này mở ra hướng triển khai mô hình BART-base (hoặc các biến thể như BARTpho) cho các bài toán thực tế đòi hỏi chất lượng sinh văn bản cao như tóm tắt tin tức, sinh tiêu đề, hoặc dịch máy trong ngôn ngữ tiếng Việt.

3.6.2 Kết quả thử nghiệm quá trình huấn luyện mô hình cho tóm tắt văn bản tiếng tiếng Lào

Kết quả thử nghiệm quá trình huấn luyện mô hình cho tóm tắt văn bản tiếng Lào thể hiện trên hình 3.8 và hình 3.9.



Hình 3.8: Mô hình BART-base cho tập dữ liệu *LaoNews Classification*



Hình 3.9: Mô hình T5-small cho tập dữ liệu *LaoNews Classification*

Hình 3.8 mô tả tiến trình huấn luyện của mô hình BART-base trong nhiệm vụ tóm tắt văn bản tiếng Lào với tập dữ liệu *LaoNews Classification* qua 15 epoch. Nhìn tổng thể, đường cong thể hiện giá trị hàm mất mát trên tập dữ liệu huấn luyện giảm đều theo thời gian huấn luyện, cho thấy mô hình học hiệu quả từ dữ liệu và dần hội tụ về một nghiệm tối ưu. Ở giai đoạn đầu tiên (epoch 0–3), giá trị hàm mất mát giảm rất mạnh – từ khoảng 2,85 xuống dưới 1,6. Đây là biểu hiện điển hình của hiện tượng "bùng nổ học tập ban đầu" – tương tự như với tập dữ liệu *VietNews* đã trình bày trên, khi mô hình nhanh chóng tiếp thu các đặc trưng ngữ nghĩa chính trong tập huấn luyện. Trong giai đoạn này, mô hình được điều chỉnh mạnh mẽ theo gradient lớn, dẫn đến cải thiện nhanh chóng về mặt tối ưu hóa. Từ epoch thứ 4 trở đi, tốc độ giảm giá trị hàm mất mát bắt đầu chậm lại nhưng vẫn duy trì xu hướng giảm ổn định, phản ánh khả năng học sâu các biểu diễn phân biệt liên quan đến các nhãn phân loại. Sau khoảng epoch 10, giá trị hàm mất mát dao động quanh mức 0,90 – 1,00, cho thấy mô hình đã tiến gần trạng thái hội tụ. Đường cong không có sự dao động bất thường, cũng không xuất hiện hiện tượng tăng ngược trở lại, điều này cho thấy không có dấu hiệu quá khớp rõ rệt trong giai đoạn huấn luyện.

Việc BART-base – một mô hình được tiền huấn luyện dưới dạng mạng tự mã hóa khử nhiễu – thể hiện khả năng huấn luyện tốt trong nhiệm vụ tóm tắt văn bản tiếng Lào là một điểm đáng chú ý. BART-base được thiết kế chuyên biệt cho các tác vụ sinh ngôn ngữ tự nhiên như tóm tắt, sinh văn bản và dịch máy, nên việc mô hình thể hiện xu hướng giảm loss ổn định và hội tụ nhanh chóng trên tập dữ liệu tiếng Lào – một ngôn ngữ có tài nguyên hạn chế – cho thấy tính khả chuyển mạnh mẽ của kiến trúc này. Biểu đồ giá trị hàm mất mát trên tập dữ liệu huấn luyện giảm đều từ hơn 2,8 xuống dưới 1,0 sau 15 epoch phản ánh năng lực học biểu diễn ngữ nghĩa hiệu quả của BART - base trong điều kiện dữ liệu không phong phú. Đặc biệt, kiến trúc kết hợp giữa bộ mã hóa (encoder) và giải mã (decoder) cùng cơ chế chú ý hai chiều cho phép mô hình xử lý linh hoạt các cấu trúc cú pháp và ngữ nghĩa phức tạp trong văn bản tiếng Lào. Quá trình khử nhiễu đầu vào giúp mô hình học được cách phục hồi nội dung nguyên bản từ dữ liệu đã bị che khuất hoặc xáo trộn – một đặc trưng quan trọng trong việc huấn luyện tóm tắt văn bản. Điều này rất phù hợp với mục tiêu tạo ra các bản tóm tắt ngắn gọn, chính xác nhưng vẫn bảo toàn ý nghĩa cốt lõi của văn bản gốc.

Biểu đồ thể hiện quá trình huấn luyện mô hình T5-small trên tập dữ liệu *LaoNews Classification* được thể hiện trong hình 3.9. Qua 20 epoch cho thấy xu hướng giảm giá trị hàm mất mát tương đối đều và ổn định. Mô hình bắt đầu với giá trị hàm mất mát xấp xỉ 3,05 – phản ánh độ khó nhất định trong việc mô hình hóa ngôn ngữ tiếng Lào từ trạng thái khởi tạo ban đầu. Tuy nhiên, trong khoảng 5 epoch đầu tiên, giá trị hàm mất mát giảm nhanh chóng từ 3,05 xuống còn khoảng 1,3, thể hiện hiệu ứng "bùng nổ học tập ban đầu", khi mô hình nhanh chóng thu nhận được các cấu trúc ngôn ngữ cơ bản. Sau đó, đường cong hàm giá trị mất mát tiếp tục giảm đều đặn qua các epoch, dần hội tụ về ngưỡng ~ 0,65 – 0,70 vào cuối quá trình huấn luyện (epoch 20). Tốc độ giảm giá trị hàm mất mát có xu hướng chậm lại từ khoảng epoch 12, nhưng vẫn duy trì được sự ổn định và không có hiện tượng dao động mạnh, cho thấy mô hình không gặp vấn đề quá khớp trong giai đoạn này.

Khi đối chiếu với kết quả huấn luyện mô hình BART-base trong cùng nhiệm vụ, có thể thấy một số khác biệt đáng lưu ý. Mặc dù cả hai mô hình đều đạt được xu

hướng hội tụ ổn định, BART-base hội tụ nhanh hơn và đạt giá trị hàm mất mát thấp hơn: chỉ sau 15 epoch, mô hình đã đạt giá trị hàm mất mát ổn định ở mức ~ 0,90, trong khi T5-small cần đến 20 epoch mới giảm xuống được ~ 0,65, và khởi đầu ở mức giá trị hàm mất mát cao hơn. Tuy nhiên, cần lưu ý rằng giá trị hàm mất mát của T5-small giảm nhanh hơn tương đối so với BART-base trong giai đoạn đầu, phần nào bù đắp cho quy mô mô hình nhỏ hơn và khả năng biểu diễn hạn chế.

Về mặt kiến trúc, BART - base là mô hình encoder–decoder đối xứng với chiến lược tiền huấn luyện theo kiểu khử nhiễu, trong khi T5-small được huấn luyện theo hướng đa nhiệm, với mục tiêu ánh xạ mọi bài toán NLP về dạng chuỗi đầu vào – chuỗi đầu ra. Trong ngữ cảnh tiếng Lào – một ngôn ngữ có tài nguyên hạn chế – việc T5-small vẫn học tốt và hội tụ tương đối hiệu quả là một kết quả tích cực, cho thấy tiềm năng sử dụng các mô hình nhẹ trong bối cảnh hạn chế về hạ tầng hoặc yêu cầu triển khai trên thiết bị tài nguyên thấp.Tóm lại, mô hình T5-small cho thấy khả năng học ngôn ngữ tiếng Lào ổn định và hiệu quả trong tác vụ TTVB, mặc dù không đạt mức hội tụ nhanh và giá trị hàm mất mát thấp như BART-base. Tuy nhiên, với ưu thế về hiệu năng và khả năng mở rộng, T5 - small vẫn là một lựa chọn tiềm năng trong các kịch bản ứng dụng thực tế. Việc lựa chọn giữa hai mô hình nên dựa trên sự cân đối giữa yêu cầu chất lượng đầu ra và khả năng triển khai trong môi trường cụ thể.

3.6.3 Kết quả đánh giá hiệu năng của các mô hình

Bảng 3.5 biểu thị kết quả một số chỉ số đánh giá hiệu năng các mô hình thử nghiệm TTVB trên dữ liệu tiếng Việt và tiếng Lào.

Bảng 3.5. Một số chỉ số đánh giá hiệu năng các mô hình với tiếng Việt và tiếng Lào

Bộ dữ liệu	Huấn luyện	Kiểm thử	Mô hình	ROUGE 1	ROUGE 2	ROUGE-L
VietNews	8000	2000	Bart-base	33,07	21,06	33,91
			T5-small	32,95	18,04	31,75
LaoNews Classification	12262	3066	Bart-base	27,62	14,2	27,1
			T5-small	25,5	14,15	24,64

Các chỉ số đánh giá hiệu năng của hai mô hình BART-base và T5-small trong tác vụ TTVB đối với hai ngôn ngữ: tiếng Việt (*VietNews*) và tiếng Lào (*LaoNews Classification*), thông qua ba chỉ số ROUGE phổ biến: ROUGE-1, ROUGE-2 và ROUGE-L được trình bày trên Bảng 3.5. Kết quả cho thấy mô hình BART-base luôn thể hiện hiệu suất vượt trội hơn so với T5-small trên cả hai ngôn ngữ và toàn bộ các chỉ số đánh giá, phản ánh ưu thế rõ rệt của mô hình này trong việc học và sinh ngôn ngữ tự nhiên trong bối cảnh đa ngữ.

Cụ thể, trên tập *VietNews*, BART-base đạt ROUGE-1 là 33,07, cao hơn nhẹ so với T5-small (32,95), nhưng sự chênh lệch rõ ràng hơn ở chỉ số ROUGE-2 (21,06 so với 18,04) và ROUGE-L (33,91 so với 31,75). Điều này phản ánh khả năng nắm bắt tốt hơn các chuỗi từ liên tục và cấu trúc ngữ nghĩa toàn cục của Bart-base so với T5-small. Cả hai mô hình đều có hiệu suất tốt với tiếng Việt, nhờ đặc điểm ngôn ngữ có sẵn tài nguyên và tập dữ liêuj huấn luyện đầy đủ. Đối với tập *LaoNews Classification*, hiệu suất của cả hai mô hình đều giảm đáng kể – điều này phù hợp với kỳ vọng do tiếng Lào là ngôn ngữ có tài nguyên thấp, dữ liệu huấn luyện hạn chế hơn. BART-base đạt ROUGE-1 là 27,62, ROUGE-2 là 14,2 và ROUGE-L là 27,1; trong khi T5-small chỉ đạt lần lượt là 25,5, 14,15 và 24,64. Mặc dù khoảng cách ROUGE-2 giữa hai mô hình là rất nhỏ (0,05 điểm), sự khác biệt ở ROUGE-1 (2,12 điểm) và ROUGE-L (2,46 điểm) cho thấy BART - base vượt trội hơn trong việc tái hiện lại thông tin chính xác và giữ nguyên cấu trúc logic của văn bản gốc.

3.7 Thảo luận, đánh giá kết quả thử nghiệm

Từ phân tích kết quả đạt được tại mục 3.6, có thể đưa ra một số nhận xét, đánh giá sau:

- Thứ nhất, BART-base cho hiệu năng ổn định và cao hơn T5-small, bất kể ngôn ngữ hay quy mô tập dữ liệu, nhờ kiến trúc encoder–decoder đối xứng và cơ chế tiền huấn luyện khử nhiễu mạnh mẽ.
- Thứ hai, ngôn ngữ có tài nguyên thấp như tiếng Lào gây thách thức đáng kể cho cả hai mô hình, đòi hỏi chiến lược bổ sung như fine-tuning thêm với dữ liệu song ngữ, hoặc áp dụng các mô hình đa ngữ như mBART hay mT5.

- Thứ ba, khoảng cách hiệu năng giữa hai mô hình càng rõ nét hơn trong ngôn ngữ yếu, khẳng định vai trò quan trọng của quy mô mô hình và chiến lược tiền huấn luyện trong xử lý ngôn ngữ đa ngữ.

Như vậy, đề án đưa ra nhận xét rằng mô hình BART-base phù hợp hơn cho nhiệm vụ TTVB đòi hỏi độ chính xác và chất lượng cao. Trong khi đó, mô hình T5-small, mặc dù nhẹ và tiêu tốn ít tài nguyên tính toán hơn, nhưng lại có những hạn chế rõ rệt về độ chính xác của kết quả tóm tắt. Bên cạnh đó, để nâng cao chất lượng tóm tắt, đặc biệt với bộ dữ liệu *LaoNews Classification*, cần tiếp tục cải tiến về phương pháp tiền xử lý dữ liệu cũng như tối ưu hóa mô hình phù hợp hơn với đặc điểm ngôn ngữ Lào.

3.8 Kết luận chương

Nội dung Chương 3 đã trình bày phần thiết lập môi trường thử nghiệm tóm tắt văn bản tiếng Việt và tiếng Lào bao gồm: Thiết lập môi trường thử nghiệm với công cụ phần mềm, thư viện hỗ trợ, thiết bị phần cứng, quy trình thử nghiệm; Tạo lập các tập dữ liệu thử nghiệm từ các bộ dữ liệu *VietNews* và *LaoNews Classification*; xây dựng mô hình và huấn luyện mô hình; Đánh giá hiệu năng mô hình; Thảo luận và đánh giá kết quả thử nghiệm.

Các thử nghiệm cho thấy: các chỉ số hiệu năng ROGUE (ROUGE-1, ROUGE-2) đạt kết quả cao với bộ dữ liệu *VietNews* so với bộ dữ liệu *LaoNews Classification* sử dụng hai mô hình Bart-base và T5-small. Mô hình Bart-base cho thấy ưu thế rõ rệt so với T5-small trên cả hai bộ dữ liệu. Mô hình T5 có ưu điểm về mức độ tính toán song độ chính xác không cao so với mô hình BART-base.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN TIẾP

Kết luận

Bài toán tóm tắt văn bản là chủ đề nghiên cứu vô cùng thiết thực có tính chất ứng dụng thực tế cao. Ứng dụng xử lý ngôn ngữ tự nhiện vào tóm tắt văn bản có ý nghĩa thiết thực đối với nhiều lĩnh vực trong đời sống. Với lượng dữ liệu văn bản ngày càng nhiều trên mạng Internet, việc truy xuất thông tin từ lượng dữ liệu khổng lồ này đặt ra những yêu cầu cấp thiết về việc nghiên cứu và xây dựng các giải pháp tóm tắt văn bản, giúp người dùng nhanh chóng nắm bắt được thông tin kịp thời. Tuy nhiên, việc thiếu các công cụ hỗ trợ trong xử lý ngôn ngữ tự nhiên giữa tiếng Việt và tiếng Lào đã gây khó khăn cho cả việc giao tiếp hằng ngày, nghiên cứu học thuật và phát triển kinh tế - xã hội chung.

Việc nghiên cứu áp dụng các phương pháp này vào bài toán tóm tắt tiếng Việt và tiếng Lào còn hạn chế. Các tập dữ liệu mẫu dùng cho huấn luyện còn chưa đầy đủ. Còn khá ít các nghiên cứu đánh giá về tính hiệu quả của các phương pháp áp dụng cho các đặc thù của ngôn ngữ, điển hình như tiếng Việt và tiếng Lào.

Vì vậy, đề tài của đề án tốt nghiệp này là nhằm mục đích tìm hiểu, khảo sát các phương pháp NLP hiện đại ứng dụng trong bài toán tóm tắt văn bản, thực hiện một số thử nghiệm với các tập dữ liệu hiện có về tiếng Việt và tiếng Lào. Qua đó có thể đánh giá mức độ thực hiện của các phương pháp đối với các đặc thù của ngôn ngữ tiếng Việt và đặc biệt là ngôn ngữ tiếng Lào của quê hương tôi.

Các kết quả đã thực hiện của đề án tốt nghiệp này gồm:

- Nghiên cứu, khảo sát cơ sở lý thuyết về xử lý ngôn ngữ tự nhiên và các ứng dụng, cụ thể áp dụng vào bài toán tóm tắt văn bản với các đặc điểm của ngôn ngữ tiếng Việt và tiếng Lào.
- Nghiên cứu, khảo sát, đánh giá các mô hình, phương pháp tóm tắt văn bản sử dụng NLP; Phân tích đánh giá một số mô hình hỗ trợ tóm tắt văn bản đa ngôn ngữ sử dụng vào tóm tắt văn bản; Phân tích các phương pháp tạo lập Dataset cho tóm tắt văn bản.

- Đề xuất một mô hình thử nghiệm cho bài toán tóm tắt văn bản tiếng Việt và tiếng Lào. Thiết lập môi trường thử nghiệm với công cụ phần mềm, thư viện hỗ trợ, thiết bị phần cứng, quy trình thử nghiệm; Tạo lập các tập dữ liệu thử nghiệm từ các bộ dữ liệu *VietNews* và *LaoNews Classification*; xây dựng mô hình và huấn luyện mô hình; Đánh giá hiệu năng mô hình; Thảo luận và đánh giá kết quả thử nghiệm.

Trong đề án này, các thực nghiệm đã làm rõ một số điểm nổi bật về hiệu suất giữa hai bộ dữ liệu *Vietnews* và *LaoNews Classification* cũng như hiệu quả của hai mô hình Bart-base và T5-small. Khi đánh giá theo các chỉ số ROUGE, bộ dữ liệu *VietNews* cho kết quả vượt trội so với *LaoNews Classification*. Cụ thể, chỉ số ROUGE-1 cao nhất trên VietNews đạt 33,07, trong khi *LaoNews Classification* chỉ đạt 27,62. Sự khác biệt này phần nào phản ánh các yếu tố như đặc trưng ngôn ngữ, chất lượng dữ liệu thu thập hoặc độ phức tạp của tác vụ tóm tắt giữa tiếng Việt và tiếng Lào.

Tiếp theo, khi so sánh hiệu suất giữa hai mô hình, Bart-base cho thấy ưu thế rõ rệt so với T5-small trên cả hai bộ dữ liệu. Trên *VietNews*, chỉ số ROUGE-2 của Bart-base đạt 21,06, cao hơn đáng kể so với 18,04 của T5-small, cho thấy khả năng của Bart-base trong việc bảo toàn cấu trúc ngôn ngữ và thông tin cốt lõi của văn bản gốc.

Từ những kết quả đạt được, đề án kết luận rằng mô hình Bart-base là lựa chọn phù hợp hơn cho các bài toán tóm tắt văn bản yêu cầu độ chính xác và chất lượng cao. Ngược lại, mặc dù T5-small có ưu điểm về mặt tiết kiệm tài nguyên tính toán, nhưng lại bị hạn chế về độ chính xác của kết quả tóm tắt.

Hướng phát triển tiếp

Từ kết quả đã thực hiện trong đề án, có thể thấy chất lượng tóm tắt văn bản tiếng Lào vẫn còn nhiều hạn chế. Do đó, trong các nghiên cứu tiếp theo, nhằm nâng cao chất lượng tóm tắt, đặc biệt với bộ dữ liệu *LaoNews Classification*, cần tiếp tục nghiên cứu các phương pháp cải tiến tiền xử lý dữ liệu và phát triển những mô hình được tối ưu hóa tốt hơn cho đặc thù ngôn ngữ tiếng Việt, tiếng Lào:

- **Mở rộng và làm giàu tập dữ liệu**

+ Thu thập thêm các tập dữ liệu tiếng Việt/ tiếng Lào đa dạng, bao gồm các lĩnh vực khác nhau như giáo dục, y tế, kinh tế, và xã hội để đảm bảo mô hình có khả năng tổng quát hóa tốt hơn.

+ Chú trọng vào các văn bản có ngữ cảnh phức tạp, giàu thông tin để kiểm tra khả năng xử lý ngữ nghĩa sâu của mô hình.

- **Nghiên cứu và cải tiến kiến trúc mô hình**

Nghiên cứu các kỹ thuật học chuyển giao và học liên tục để tăng cường hiệu suất của các mô hình hiện tại.

- **Đánh giá chi tiết hơn về chất lượng bản tóm tắt**

Việc kết hợp các thước đo bổ sung như BLEU, METEOR và các kỹ thuật đánh giá ngữ nghĩa bên cạnh ROUGE giúp đánh giá hiệu quả mô hình một cách toàn diện và sâu sắc hơn.

Thực hiện khảo sát ý kiến từ người dùng cuối để đánh giá chất lượng bản tóm tắt dựa trên trải nghiệm thực tế.

TÀI LIỆU THAM KHẢO

1. Hirschberg, J. and C.D. Manning, *Advances in natural language processing.* Science, 2015. **349**(6245): p. 261-266.
2. Bird, S., E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. 2009: " O'Reilly Media, Inc.".
3. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**(3): p. 5.
4. Devlin, J., et al. *Bert: Pre-training of deep bidirectional transformers for language understanding*. in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019.
5. Brown, T., et al., *Language models are few-shot learners.* Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
6. Young, T., et al., *Recent trends in deep learning based natural language processing.* ieee Computational intelligenCe magazine, 2018. **13**(3): p. 55-75.
7. Schütze, H., C.D. Manning, and P. Raghavan, *Introduction to information retrieval*. Vol. 39. 2008: Cambridge University Press Cambridge.
8. Jurafsky, D., *Speech & language processing*. 2000: Pearson Education India.
9. Mikolov, T., et al., *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781, 2013.
10. Pennington, J., R. Socher, and C.D. Manning. *Glove: Global vectors for word representation.* in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
11. Nenkova, A. and K. McKeown, *Automatic summarization.* Foundations and Trends® in Information Retrieval, 2011. **5**(2–3): p. 103-233.
12. Lample, G., et al., *Neural architectures for named entity recognition.* arXiv preprint arXiv:1603.01360, 2016.
13. Rajpurkar, P., et al., *Squad: 100,000+ questions for machine comprehension of text.* arXiv preprint arXiv:1606.05250, 2016.
14. Pang, B. and L. Lee, *Opinion mining and sentiment analysis.* Foundations and Trends® in information retrieval, 2008. **2**(1–2): p. 1-135.
15. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.
16. Wang, Y., et al., *Clinical information extraction applications: a literature review.* Journal of biomedical informatics, 2018. **77**: p. 34-49.
17. Achiam, J., et al., *Gpt-4 technical report.* arXiv preprint arXiv:2303.08774, 2023.

18. Chowdhery, A., et al., *Palm: Scaling language modeling with pathways.* Journal of Machine Learning Research, 2023. **24**(240): p. 1-113.
19. Raffel, C., et al., *Exploring the limits of transfer learning with a unified text-to-text transformer.* Journal of machine learning research, 2020. **21**(140): p. 1-67.
20. Radford, A., et al. *Learning transferable visual models from natural language supervision.* in *International conference on machine learning*. 2021. PmLR.
21. Bender, E.M., et al. *On the dangers of stochastic parrots: Can language models be too big??* in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021.
22. Gupta, V. and G.S. Lehal, *A survey of text summarization extractive techniques.* Journal of emerging technologies in web intelligence, 2010. **2**(3): p. 258-268.
23. Luhn, H.P., *The automatic creation of literature abstracts.* IBM Journal of research and development, 1958. **2**(2): p. 159-165.
24. Erkan, G. and D.R. Radev, *Lexrank: Graph-based lexical centrality as salience in text summarization.* Journal of artificial intelligence research, 2004. **22**: p. 457-479.
25. Mihalcea, R. and P. Tarau. *Textrank: Bringing order into text.* in *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.
26. Rush, A.M., S. Chopra, and J. Weston, *A neural attention model for abstractive sentence summarization.* arXiv preprint arXiv:1509.00685, 2015.
27. Bahdanau, D., K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate.* arXiv preprint arXiv:1409.0473, 2014.
28. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.
29. Lewis, M., et al., *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.* arXiv preprint arXiv:1910.13461, 2019.
30. Zhang, J., et al. *Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.* in *International conference on machine learning*. 2020. PMLR.
31. Maynez, J., et al., *On faithfulness and factuality in abstractive summarization.* arXiv preprint arXiv:2005.00661, 2020.
32. See, A., P.J. Liu, and C.D. Manning, *Get to the point: Summarization with pointer-generator networks.* arXiv preprint arXiv:1704.04368, 2017.

33. Demner-Fushman, D., W.W. Chapman, and C.J. McDonald, *What can natural language processing do for clinical decision support?* Journal of biomedical informatics, 2009. **42**(5): p. 760-772.
34. Kryściński, W., et al., *Evaluating the factual consistency of abstractive text summarization.* arXiv preprint arXiv:1910.12840, 2019.
35. Beltagy, I., M.E. Peters, and A. Cohan, *Longformer: The long-document transformer.* arXiv preprint arXiv:2004.05150, 2020.
36. Perez-Beltrachini, L. and M. Lapata, *Bootstrapping generators from noisy data.* arXiv preprint arXiv:1804.06385, 2018.
37. Strubell, E., A. Ganesh, and A. McCallum. *Energy and policy considerations for modern deep learning research.* in *Proceedings of the AAAI conference on artificial intelligence*. 2020.
38. Lin, C.-Y. *Rouge: A package for automatic evaluation of summaries.* in *Text summarization branches out*. 2004.
39. Liu, Y. and M. Lapata, *Text summarization with pretrained encoders.* arXiv preprint arXiv:1908.08345, 2019.
40. Conneau, A., et al., *Unsupervised cross-lingual representation learning at scale.* arXiv preprint arXiv:1911.02116, 2019.
41. Pires, T., E. Schlinger, and D. Garrette, *How multilingual is multilingual BERT?* arXiv preprint arXiv:1906.01502, 2019.
42. Liu, Y., et al., *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692, 2019.
43. Bao, T., H. Zhang, and C. Zhang, *Enhancing abstractive summarization of scientific papers using structure information.* Expert Systems with Applications, 2025. **261**: p. 125529.
44. Qi, W., et al., *Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training.* arXiv preprint arXiv:2001.04063, 2020.
45. He, P., et al., *Z-code++: A pre-trained language model optimized for abstractive summarization.* arXiv preprint arXiv:2208.09770, 2022.
46. Fraile Navarro, D., et al., *Expert evaluation of large language models for clinical dialogue summarization.* Scientific Reports, 2025. **15**(1): p. 1195.
47. Kedzie, C., K. McKeown, and H. Daume III, *Content selection in deep learning models of summarization.* arXiv preprint arXiv:1810.12343, 2018.
48. Xue, L., et al., *mT5: A massively multilingual pre-trained text-to-text transformer.* arXiv preprint arXiv:2010.11934, 2020.
49. Tran, N.L., D.M. Le, and D.Q. Nguyen, *BARTpho: pre-trained sequence-to-sequence models for Vietnamese.* arXiv preprint arXiv:2109.09701, 2021.

50. Nguyen, V.-H., et al. *Vnds: A vietnamese dataset for summarization*. in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*. 2019. IEEE.

Người hướng dẫn	Tác giả thực hiện
(ký tên)	*(ký tên)*
PGS.TSKH. Hoàng Đăng Hải	Viengnakhone Seesamoud

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP

Họ và tên học viên: Viengnakhone Seesamoud

Chuyên ngành: KHMT

Khóa: 2023 đợt 2

Tên đề tài: Nghiên cứu các phương pháp tóm tắt văn bản và thử nghiệm với dữ liệu tiếng Lào

Người hướng dẫn khoa học: PGS.TSKH. Hoàng Đăng Hải

Ngày bảo vệ: 19/07/2025

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Rà soát trình bày	Tiếp thu góp ý của Hội đồng, học viên đã rà soát trình bày toàn bộ luận văn, chỉnh sửa: - Lỗi soạn thảo - Lỗi chính tả - Các lỗi dùng từ (sửa lại thuật ngữ "tóm tắt trừu tượng" thành "tóm tắt tóm lược", "huấn luyện trước" thành "tiền huấn luyện" như góp ý của phản biện). - Một số hình vẽ đã được Việt hóa phù hợp (Hình 2.1, 2.2, 2.3, 2.5, 2.6).
2	Tổ chức nội dung thông tin và rà soát logic, viết gắn kết lại	Tiếp thu góp ý của Hội đồng, học viên đã rà soát, tổ chức nội dung các đầu mục, viết lại nội dung đảm bảo tính logic và có sự gắn kết, trong đó có một số bổ sung quan trọng: - Mục 2.3 đã bổ sung lý do lựa chọn mô hình T5 và BART cho bài toán tóm tắt văn bản (trang 23,24). - Mục 2.3.1, tại bước tiền xử lý, đã bổ sung thêm lập luận về mức độ tin cậy khi dùng ChatGPT để tóm tắt dữ liệu tiếng

		Lào (trang 25,26). - Mục 3.1 lược bớt giới thiệu về phần mềm và các thư viện hỗ trợ.
3	Thử nghiệm và phân tích kỹ hơn	Tiếp thu góp ý của Hội đồng, học viên đã bổ sung thêm phần phân tích các kết quả thu được tại các mục 3.6 và 3.7 của Chương 3 (trang 53-60).

Hà Nội, ngày tháng năm 2025

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
GS.TS. Từ Minh Phương	TS. Đỗ Thị Liên	PGS.TSKH Hoàng Đăng Hải	Viengnakhone Seesamoud

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 1098/QĐ-HV ngày 26 tháng 06 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi 14 giờ 25 phút, ngày 19 tháng 07 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Viengnakhone Seesamoud**

Tên đề án tốt nghiệp: **Nghiên cứu các phương pháp tóm tắt văn bản và thử nghiệm với dữ liệu tiếng Lào**

Chuyên ngành: **Khoa học máy tính** Mã số: **8480101**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ...../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**GS.TS. Từ Minh Phương**	Chủ tịch	
2	**TS. Đỗ Thị Liên**	Thư ký	
3	**PGS.TS. Trần Đăng Hưng**	Phản biện 1	
4	**TS. Nguyễn Văn Vinh**	Phản biện 2	
5	**PGS.TS. Nguyễn Mạnh Hùng**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên. (có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

- Lấy phần nào từ tiếng Lào
- Lý do lựa chọn mô hình BART và T5 để so sánh với nhau
- Lý do áp dụng cho bài toán tóm tắt văn bản tiếng Lào giữa 2 mô hình này là gì về mặt kiến trúc

7. Trả lời của học viên:

Học viên giải trình và tiếp thu chỉnh sửa theo ý

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: Trần Đăng Hưng
  2. Uỷ viên Ban kiểm phiếu: Đỗ Thị Liên
  3. Uỷ viên Ban kiểm phiếu: Nguyễn Mạnh Hưng
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 9,5

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

Bổ sung bổ trình bày
Rà soát nội dung đảm bảo tính logic, mạch lạc hợp lý
Kiểm tra lỗi chính tả

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên:

Đề nghị Học viện công nhận và cấp bằng thạc sĩ cho học viên

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS..... Không.....

Buổi làm việc kết thúc vào..... 21h10..... cùng ngày.

Chủ tịch	Thư ký
GS.TS. Từ Minh Phương	TS. Đỗ Thị Liên

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Về đề tài: Nghiên cứu phương pháp tóm tắt văn bản và thử nghiệm với dữ liệu tiếng Lào.

Chuyên ngành: Khoa học máy tính

Của học viên: Viengakhone Seesamound

Họ và tên cán bộ phản biện: PGS.TS Trần Đăng Hưng

Cơ quan công tác: Trường CNTT&TT, Trường Đại học Công nghiệp Hà Nội

NỘI DUNG NHẬN XÉT

Đề án nghiên cứu bài toán tóm tắt văn bản (tiếng Việt và tiếng Lào), đây là một trong những bài toán quan trọng trong xử lý ngôn ngữ tự nhiên hiện nay. Mặc dù có nhiều mô hình và phương pháp tóm tắt văn bản, tuy nhiên hiệu quả của các mô hình này vẫn cần cải tiến thêm. Vì vậy, đề án của học viên nghiên cứu các bài toán tóm tắt văn bản tiếng Việt và tiếng Lào là bài toán có ý nghĩa và phù hợp với chuyên ngành đào tạo.

Cụ thể nội dung đề án bao gồm:

- Nghiên cứu tổng quan về bài toán tám tắt văn bản và một số khái niệm cơ bản trong lĩnh vực xử lý ngôn ngữ tự nhiên.
- Trình bày một số phương phương tóm tắt văn bản như: Mô hình pre-trained, mô hình Transformers, Mô hình Encoder-Decoder,...
- Lựa chọn mô hình T5 và BART cho bài toán tóm tắt văn bản tiếng Việt và tiếng Lào. Đồng thời giới thiệu một vài tập dữ liệu điển hình.

- Thực nghiệm bài toán tóm tắt văn bản với các tập dữ liệu Vietnews và Laonews. Đánh giá kết quả của các mô hình dựa trên độ chính xác và hiệu suất.

Nhận xét và góp ý:

- Về cơ bản đề án được tổ chức khá tốt, có trình bày kiến thức cơ bản về lý thuyết, bài toán và dữ liệu và triển khai thực nghiệm.

Tuy nhiên có một vài điểm cần cải tiến thêm:

- Mặc dù tiêu đề của đề án là chỉ tóm tắt văn bản tiếng Lào, tuy nhiên nội dung bao gồm cả tóm tắt văn bản tiếng Việt và tiếng Lào, cần điều chỉnh tiêu đề đề án cho phù hợp.
- Một số sơ đồ/hình vẽ cần được Việt hóa cho đồng bộ với các hình vẽ khác.
- Mục 2.3 cần nói rõ hơn về lý do lựa chọn mô hình T5 và BART cho bài toán tóm tắt văn bản.
- Một số đoạn code được dán trong đề án dường như được copy dưới dạng ảnh, bị mờ và vỡ, cần xem lại.
- Mục 3.1 nên lược bớt giới thiệu về phần mềm và các thư viên hỗ trợ, chỉ cần nêu tên là đủ.
- Tập dữ liệu tin tức tiếng Việt khá cũ, thu thập từ năm 2016-219, liệu gần đây có tập dữ liệu cập nhật hơn?
- Tập dữ liệu tiếng Lào sử dụng ChatGPT để lấy tóm tắt, vậy tập dữ liệu huấn luyện có đủ độ tin cậy.

Kết luận: Đồng ý cho học viên bảo vệ trước hội đồng chấm đề án.

Hà Nội, ngày 15 tháng 7 năm 2025

CÁN BỘ PHẢN BIỆN

Phạm Đức Thắng

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: NGHIÊN CỨU PHƯƠNG PHÁP TÓM TẮT VĂN BẢN VÀ THỬ NGHIỆM VỚI DỮ LIỆU TIẾNG LÀO

Chuyên ngành: Hệ thống thông tin

Mã chuyên ngành: .....

Họ và tên học viên: Viengnakhone Seeamound

Họ và tên người nhận xét: Nguyễn Văn Vinh

Học hàm, học vị: TS

Chuyên ngành: KHMT

Cơ quan công tác: Trường ĐH Công Nghệ, ĐHQG Hà Nội

Số điện thoại:.....E-mail:.....

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Bài toán tóm tắt văn bản là bài toán quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên. Vì vậy đề tài nghiên cứu các phương pháp tóm tắt văn bản và ứng dụng cho ngôn ngữ Lào là có ý nghĩa về mặt khoa học cũng như thực tiễn.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

- Nghiên cứu tổng quan về bài toán tóm tắt văn bản và đặc điểm tiếng Việt và tiếng Lào
- Nghiên cứu về các mô hình học sâu và mô hình huấn luyện trước (Pre-training) cho bài toán tóm tắt văn bản.
- Áp dụng mô hình T5 và BART cho bài toán tóm tắt văn bản và thử nghiệm với dữ liệu tóm tắt tiếng Việt VietNewsvà tiếng Lào (tự xây dựng) cho kết quả khả quan.

III/ Những vấn đề cần giải thích thêm:

- Một số hình vẽ nên vẽ lại (không nên cắt dán) để cho đẹp và rõ ràng hơn, ví dụ hình 3.2 (trang 43)
- Mục 1.1 có thể bỏ (vì không liên quan trực tiếp đến đề án tốt nghiệp).
- Nên có ví dụ về đầu vào và đầu ra của bài toán tóm tắt văn bản trong mục 1.1.2
- Mục 2.2 chỉ nên tập trung vào phương pháp tóm tắt theo kiểu tóm lược vì đề án tập trung vào phương pháp này.

- Đề án cũng nên nêu rõ tại sao chọn mô hình T5 và BART áp dụng cho bài toán tóm tắt văn bản và vẽ ra kiến trúc (hoặc pipeline) của việc áp dụng mô hình này cho bài toán tóm tắt văn bản: đầu vào là gì, đầu ra là gì và các thành phần của T5 và BART
- Phần đánh giá cũng nên sử dụng thêm cách đánh giá khác ngoài (ROUGE) như Bertscore và ChatGPT vì ROUGE không tốt cho tóm tắt dựa vào tóm lược.
- Các mã code (trang 49, 50) nên chuyển sang phần phụ lục, ở đây ta có bảng Hyeperparameter của mô hình. Phần thử nghiệm vẫn còn đơn giản, nên thử nghiệm và phân tích kỹ hơn với cỡ dữ liệu huấn luyện khác nhau và chọn bộ siêu tham số tốt nhất.
- Một thuật ngữ nên viết cho chính xác hơn như **tóm tắt trừu tượng → tóm tắt tóm lược**. Một số lỗi chính tả cần được chỉnh sửa.

IV/ Kết luận:

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày 04 tháng 07 năm 2025

NGƯỜI NHẬN XÉT

NGUYỄN VĂN VINH