**Sự Tổng Hòa Sâu Sắc giữa Âm Thanh, Ngữ Cảnh và Khả Năng Chịu Lỗi: Các Kiến Trúc Học Sâu Tiên Tiến cho Xử Lý Ngôn Ngữ Tự Nhiên và Nhận Dạng Giọng Nói Tiếng Việt**

Báo cáo này trình bày phân tích chuyên sâu về mối quan hệ cộng sinh giữa xử lý âm thanh (Sound Processing) và Xử lý Ngôn ngữ Tự nhiên (NLP), đặc biệt tập trung vào các mô hình nhúng từ có ngữ cảnh (contextual word vectors) và vai trò của chúng trong việc giải quyết các thách thức ngôn ngữ đặc thù của tiếng Việt, bao gồm cả khả năng xử lý các văn bản thiếu dấu hoặc mắc lỗi chính tả.

**I. Khung Cơ Sở: Bối Cảnh Ngôn Ngữ Việt Nam trong Trí Tuệ Nhân Tạo**

Phần này thiết lập sự cần thiết về mặt kỹ thuật đối với các phương pháp nâng cao, dựa trên những khó khăn duy nhất mà ngữ âm và chính tả tiếng Việt đặt ra cho các hệ thống AI hiện đại.

**1.1. Những Thách Thức Đặc Thù của Tiếng Việt trong Xử Lý Ngôn Ngữ và Giọng Nói**

Tiếng Việt là một ngôn ngữ đơn âm tiết với hệ thống thanh điệu phức tạp, nơi ý nghĩa của một từ có thể thay đổi hoàn toàn chỉ dựa vào thanh điệu. Đặc điểm này tạo ra một rào cản cơ bản cho Nhận dạng Giọng nói Tự động (ASR), bởi vì những thay đổi nhỏ về âm học có thể dẫn đến những thay đổi ngữ nghĩa lớn.1

Ngoài ra, tiếng Việt còn đối diện với sự đa dạng về mặt hình thái và phương ngữ. Sự khác biệt đáng kể giữa các phương ngữ khu vực (Bắc, Trung, Nam) cùng với tốc độ nói và cách phát âm khác nhau đòi hỏi các mô hình phải mạnh mẽ và được đào tạo trên các bộ dữ liệu quy mô lớn, đa dạng.1

Trước đây, tiếng Việt được coi là một "ngôn ngữ thiếu tài nguyên" (*under-resourced language*), dẫn đến nhu cầu cao về các kho ngữ liệu giọng nói chất lượng cao và quy mô lớn cho các tác vụ ASR. Nỗ lực gần đây đã tạo ra các tập dữ liệu mới, chẳng hạn như một kho ngữ liệu giọng nói lớn mới với thời lượng 100.5 giờ được thu thập từ nhiều nguồn âm thanh khác nhau trên Internet, sau đó được xử lý và phiên âm thủ công. Việc kiểm tra cho thấy tập dữ liệu này có sự đa dạng tốt về giới tính, chủ đề và phương ngữ.2

Sự phức tạp về mặt ngôn ngữ này là nguyên nhân trực tiếp dẫn đến tỷ lệ lỗi từ (Word Error Rate - WER) cao trong các hệ thống ASR truyền thống. Chẳng hạn, một mô hình học sâu lai (hybrid deep learning) kết hợp Mạng nơ-ron Tích chập (CNN) và Bộ nhớ Dài-Ngắn Hai chiều (BiLSTM) đạt WER là 14.7% trên tập dữ liệu chuẩn VIVOS của tiếng Việt.1 Mức WER này không chỉ là một số liệu hiệu suất mà còn là bằng chứng cho thấy sự cần thiết phải có một lớp mạnh mẽ trong NLP để bù đắp những sai sót không thể tránh khỏi của mô hình âm học. Do đó, mối liên kết giữa "âm thanh" và "NLP" không phải là một sự tối ưu hóa tùy chọn mà là một tầng bắt buộc để đảm bảo độ tin cậy.

Sự khan hiếm dữ liệu chất lượng đã đẩy mạnh việc áp dụng các kiến trúc tiên tiến như mô hình LAS (Listener, Attention, Speller) và Speech-Transformer cho ASR tiếng Việt. Tuy nhiên, hiệu suất của các mô hình này hoàn toàn phụ thuộc vào việc sử dụng các bộ dữ liệu mới, phản ánh chính xác những khó khăn trong việc nhận dạng giọng nói từ các phương ngữ và miền chủ đề khác nhau.2

**II. Sự Phát Triển của Biểu Diễn Ngữ Cảnh: Cầu Nối Giữa Âm Thanh và Ngữ Nghĩa**

Phần này giải thích tại sao các mô hình nhúng từ có ngữ cảnh (contextual word embeddings) lại trở nên thiết yếu, vượt qua giới hạn của các mô hình nhúng từ tĩnh (non-contextual) như Word2Vec.

**2.1. Từ Vector Tĩnh đến Nhúng Ngữ Cảnh Động**

Các mô hình nhúng từ truyền thống, như Word2Vec hay FastText, gán cho mỗi từ một vector cố định, bất kể ngữ cảnh nó xuất hiện. Điều này khiến chúng không thể nắm bắt hiện tượng đa nghĩa (polysemy) và các mối quan hệ phụ thuộc sắc thái, vốn rất quan trọng trong các ngôn ngữ có cấu trúc đơn âm tiết như tiếng Việt.

Sự thay đổi mô hình đã đến với kiến trúc Transformer. Các mô hình nhúng ngữ cảnh (ví dụ: các biến thể của BERT, RoBERTa, XLM-R) tạo ra các vector từ động dựa trên toàn bộ chuỗi đầu vào. Điều này cho phép chúng đạt được sự hiểu biết ngữ nghĩa và cú pháp thực sự, vốn là nền tảng cho các nhiệm vụ NLP hiện đại.

**2.2. Các Mô hình Ngôn ngữ Được Huấn luyện Trước (PLMs) trong NLP Tiếng Việt: Kiến Trúc và Tác Động**

Các PLM đã tạo ra một cuộc cách mạng trong xử lý tiếng Việt:

**2.2.1. PhoBERT (Kiến trúc RoBERTa)**

PhoBERT là mô hình nền tảng hiệu suất cao cho NLP tiếng Việt. Được xây dựng trên kiến trúc RoBERTa và huấn luyện trên các kho ngữ liệu tiếng Việt lớn, PhoBERT cung cấp khả năng mã hóa đơn ngữ mạnh mẽ.3 Mô hình này đã chứng minh tính hiệu quả của mình khi được sử dụng làm nền tảng cho các mô hình chuyên biệt khác và đạt hiệu suất cạnh tranh trong các nhiệm vụ nhúng câu, với điểm tương quan Pearson cao (84.87) trên STS Benchmark.3

**2.2.2. ViSoBERT (Kiến trúc XLM-R)**

ViSoBERT đại diện cho một bước tiến chuyên biệt hóa cần thiết. Được xây dựng bằng kiến trúc XLM-R, ViSoBERT được thiết kế rõ ràng để xử lý văn bản mạng xã hội tiếng Việt.4 Văn bản mạng xã hội nổi tiếng là nhiễu, không chính thức và thường xuyên thiếu các quy tắc chính tả nghiêm ngặt (liên quan trực tiếp đến vấn đề "văn bản không dấu" của người dùng). ViSoBERT đạt được kết quả tiên tiến trong các nhiệm vụ cụ thể như nhận dạng cảm xúc và phát hiện ngôn ngữ thù địch 4, thường với ít tham số hơn so với các đối thủ cạnh tranh, làm cho nó trở thành một lựa chọn hiệu quả cho miền dữ liệu này.5

Sự tồn tại và thành công của ViSoBERT chứng minh một xu hướng quan trọng: các PLM tổng quát như PhoBERT có thể không đủ cho các nguồn dữ liệu có tính đồng nhất thấp. Khả năng chịu lỗi trong NLP tiếng Việt đòi hỏi các mô hình chuyên biệt được huấn luyện trước trên chính độ nhiễu và từ vựng của miền mục tiêu (ví dụ: tiếng lóng mạng xã hội), nhấn mạnh rằng việc xử lý văn bản bị lỗi chính tả hoặc thiếu dấu cần các mô hình nhận biết ngữ cảnh domain.

**2.3. Các Kỹ Thuật Ngữ Cảnh Hóa Nâng Cao: Tích Hợp Mạng Nơ-ron Đồ thị (GAT)**

Sự tinh chỉnh kiến trúc đang vượt ra ngoài mô hình hóa tuần tự. Mô hình TextGraphFuseGAT là một ví dụ đột phá, tích hợp bộ mã hóa Transformer được huấn luyện trước (PhoBERT) với Mạng Lưới Tập trung Đồ thị (GAT).6

Cơ chế hoạt động là xây dựng một đồ thị kết nối đầy đủ trên các vector nhúng token do PhoBERT tạo ra, cho phép tầng GAT nắm bắt các **phụ thuộc liên token phong phú** vượt ra ngoài ngữ cảnh tuần tự được mô hình hóa bởi Transformer.6 Điều này dẫn đến sự cải thiện hiệu suất trong các tác vụ phân loại cấp token đầy thách thức, chẳng hạn như Nhận dạng Thực thể Có tên (NER), đặc biệt là trong các miền chuyên biệt phức tạp như bản ghi lời nói y tế (VietMed-NER).6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình** | **Kiến trúc Cơ sở** | **Trọng tâm Dữ liệu Huấn luyện Trước** | **Điểm mạnh Nhiệm vụ Chính** | **Cơ chế Ngữ cảnh/Đổi mới Chính** |
| PhoBERT 3 | RoBERTa (Bộ mã hóa Transformer) | Văn bản Tiếng Việt Tiêu chuẩn | NER, Phân loại Văn bản, Nhúng câu Tổng quát | Mã hóa đơn ngữ mạnh mẽ của ngôn ngữ viết tiêu chuẩn. |
| ViSoBERT 4 | XLM-R (Bộ mã hóa Transformer) | Văn bản Mạng xã hội Việt Nam | Nhận dạng Cảm xúc, Phát hiện Ngôn ngữ Thù địch | Tối ưu hóa cho ngôn ngữ mạng xã hội, ồn ào và thiếu dấu. |
| TextGraphFuseGAT 6 | PhoBERT + GAT (Lai) | Đặc thù Miền (ví dụ: NER Y tế/COVID-19) | Phân loại Cấp Token, Nhận dạng Thực thể Nâng cao | Tích hợp GAT để mô hình hóa các phụ thuộc quan hệ phi tuần tự giữa các token. |

**III. Sự Phụ Thuộc Lẫn Nhau giữa ASR và NLP: Sửa Lỗi và Nâng Cao Ngữ Nghĩa**

Phần này liên kết trực tiếp các thách thức xử lý "âm thanh" với các giải pháp "nhúng ngữ cảnh", mô tả các kỹ thuật đột phá trong việc sửa lỗi ASR dựa trên ngữ cảnh.

**3.1. Thách Thức và SOTA trong Nhận dạng Giọng nói Tự động Tiếng Việt (ASR)**

Các nền tảng kiến trúc ASR tiên tiến truyền thống cho tiếng Việt thường sử dụng các phương pháp học sâu lai, ví dụ: kết hợp CNN để trích xuất các đặc trưng không gian từ biểu đồ tần phổ âm thanh với BiLSTM để nắm bắt các phụ thuộc thời gian.1

Những đột phá gần đây hơn đã áp dụng các kiến trúc SOTA tổng quát, chẳng hạn như LAS và Speech-Transformer, vào các tác vụ tiếng Việt, tận dụng các kho ngữ liệu chất lượng cao, đa dạng cho việc đào tạo.2

**3.2. NLP là Cơ chế Hậu Sửa Lỗi cho ASR**

Các hệ thống ASR thường tạo ra nhiều bản phiên âm khả thi (N-best hypotheses). Đóng góp quan trọng của NLP là sắp xếp lại các giả thuyết này bằng cách sử dụng thông tin ngữ cảnh có nguồn gốc từ trạng thái hội thoại hoặc lượt lời trước đó của người dùng.7 Điều này được thực hiện bằng cách tính toán sự tương đồng từ vựng và ngữ nghĩa giữa các giả thuyết và thông tin ngữ cảnh.7

Ngay cả các mô hình ASR tiên tiến cũng mắc lỗi, cản trở các cuộc đối thoại trôi chảy với AI đàm thoại. Các phương pháp tiên tiến sử dụng Mô hình Ngôn ngữ Lớn (LLMs) để tăng cường ngữ cảnh và sửa lỗi ASR, đặc biệt trong AI đàm thoại theo định hướng mục tiêu, nơi lỗi thể hiện tính linh hoạt ngôn ngữ đáng kể (các biến thể từ vựng, cú pháp và bỏ sót từ).7

Sự phụ thuộc này vào các mô hình nhúng ngữ cảnh là tuyệt đối cần thiết. Các mô hình không ngữ cảnh không thể phân biệt giữa các tùy chọn đầu ra ASR có âm học tương tự nhưng ngữ nghĩa khác biệt. Các mô hình ngữ cảnh như PhoBERT hoặc ViSoBERT cung cấp không gian vector ngữ nghĩa cần thiết để LLM đo lường mức độ hợp lý so với ngữ cảnh đối thoại, khiến chúng trở thành điều kiện tiên quyết cho các tác vụ hậu sửa lỗi.

Một quy trình sửa lỗi tinh vi được áp dụng: nếu việc sắp xếp lại ngữ nghĩa (sử dụng nhúng ngữ cảnh) không tìm được giả thuyết hợp lý nào, một lớp phòng thủ thứ hai sẽ được triển khai. Nó sẽ sắp xếp ngữ cảnh bằng cách sử dụng **sự tương ứng ngữ âm** với giả thuyết ASR tốt nhất để sửa lỗi.7 Chiến lược kép này chứng tỏ sự pha trộn phức tạp giữa thông tin âm thanh/ngữ âm và ngữ cảnh ngữ nghĩa.

Việc sử dụng các mô hình ngữ cảnh trong hậu sửa lỗi ASR mang lại lợi ích trực tiếp và đáng kể cho trải nghiệm người dùng, cải thiện điểm đánh giá của người dùng lên 0.8–1 điểm (trên thang 5) khi phương pháp sửa lỗi hoạt động hiệu quả.7 Điều này cho thấy sự hiểu biết ngôn ngữ tự nhiên (NLU) được cải thiện từ ngữ cảnh có thể giảm thiểu những thiếu sót cơ bản của mô hình âm học, khẳng định rằng sự tương tác liền mạch giữa con người và AI trong lời nói phụ thuộc nhiều hơn vào ngữ cảnh ngữ nghĩa hơn là độ chính xác âm học thô.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cơ chế/Tầng** | **Vai trò Chính** | **Yêu cầu Mô hình NLP** | **Chức năng trong Sửa lỗi ASR** |
| Đầu ra Mô hình Âm học (ASR) | Tạo ra N-best hypotheses. | CNN-BiLSTM, LAS, Transformer [1, 2] | Chuyển đổi âm thanh phức tạp (thanh điệu) thành các giả thuyết văn bản mơ hồ. |
| Sắp xếp lại Ngữ cảnh | Chọn giả thuyết hợp lý nhất về mặt ngữ nghĩa. | Nhúng Ngữ cảnh (ví dụ: PhoBERT) + LLM Tăng cường 7 | Giảm thiểu sự nhầm lẫn thanh điệu/phương ngữ bằng cách áp đặt các ràng buộc ngữ nghĩa. |
| Xếp hạng Ngữ cảnh Ngữ âm | Sửa lỗi khi sắp xếp lại ngữ nghĩa thất bại. | Các chỉ số khoảng cách ngữ âm | Khắc phục các lỗi còn lại dựa trên sự tương đồng âm thanh với các thuật ngữ ngữ cảnh đã biết. |

**IV. Kỹ Thuật Chịu Lỗi: Xử Lý Văn Bản Tiếng Việt Thiếu Dấu và Có Nhiễu**

Phần này tập trung vào câu hỏi quan trọng của người dùng về việc xử lý văn bản tiếng Việt không có dấu ("không dấu") và các lỗi chính tả phổ biến.

**4.1. Vấn đề Quan trọng của Phục hồi Dấu (Diacritics Restoration)**

Văn bản tiếng Việt không có dấu là văn bản có tính mơ hồ cao, vì thanh điệu và dấu phụ xác định ý nghĩa của từ. Phục hồi dấu là một bước tiền xử lý bắt buộc đối với hầu hết các ứng dụng NLP nghiêm túc, bao gồm hệ thống chatbot hoặc trả lời câu hỏi.9

Về cơ bản, nhiệm vụ này được coi là một **bài toán sequence-to-sequence** (chuỗi đầu vào thành chuỗi đầu ra), thường được ví như Dịch máy Thống kê (SMT), trong đó ngôn ngữ nguồn là văn bản không dấu và ngôn ngữ đích là văn bản có đầy đủ dấu.10

**4.2. Các Phương pháp Học Sâu để Phục hồi Dấu**

**4.2.1. Kiến Trúc Lặp (LSTM/Bi-GRU)**

Các mô hình đầu tiên và hiệu quả sử dụng Mạng LSTM dựa trên ký tự làm mô hình ngôn ngữ.10 Một phương pháp lai tiên tiến hơn, kết hợp Mạng nơ-ron Tích chập (CNN) với Đơn vị Lặp cổng Hai chiều (Bi-GRU), đã đạt được độ chính xác cao. Mô hình này kết hợp một khối dư (*residual block*) để giải quyết vấn đề gradient biến mất (*vanishing gradient*).12 Phương pháp CNN-BiGRU này đạt **Độ chính xác Từ 94.77%** và Độ chính xác Ký tự 98.63%.12

**4.2.2. Các Giải pháp Dựa trên Transformer**

SOTA hiện tại dựa vào kiến trúc Transformer nhờ khả năng nắm bắt ngữ cảnh toàn cục vượt trội và song song hóa.

* **TDP (Transformer Decoder Penalty):** Mô hình này sử dụng một phương pháp lai, giữ lại Bộ giải mã Transformer ở cấp độ ký tự và thêm một tầng Penalty (Hình phạt), hạn chế tập hợp các chữ cái có dấu phụ khả thi cho chuỗi đầu ra, tăng cường cả tốc độ và độ chính xác.9 Đối với các ứng dụng công nghiệp, tốc độ khôi phục là "cực kỳ quan trọng".9 Sự chuyển đổi từ kiến trúc Bi-GRU/LSTM sang Transformer là do yêu cầu về độ trễ thấp và khả năng song song hóa của Transformer, giúp xử lý khối lượng lớn văn bản tiền xử lý hiệu quả.
* **Mô hình Transformer Hợp nhất (ví dụ: ByT5):** Các phương pháp tiên tiến, sử dụng các biến thể mô hình Transformer như ByT5, giải quyết **việc phục hồi dấu và sửa lỗi chính tả đồng thời**. Phương pháp kết hợp này đã cho thấy tiềm năng ứng dụng thực tế lớn, đạt độ chính xác từ-alpha lớn hơn 94% trên nhiều ngôn ngữ, bao gồm cả tiếng Việt, vượt trội hơn đáng kể so với các phương pháp kiểm tra chính tả hoặc từ điển cổ điển.13

Việc xử lý đồng thời khôi phục dấu và sửa lỗi chính tả cho thấy các giải pháp chịu lỗi mạnh mẽ nhất coi việc chuẩn hóa chính tả (đặt dấu) và sửa lỗi từ vựng (chính tả) là các nhiệm vụ sequence-to-sequence hợp nhất. Bằng cách đóng khung vấn đề là một phép biến đổi thống nhất từ *đầu vào nhiễu* sang *đầu ra sạch*, mô hình có thể tận dụng ngữ cảnh toàn cục để giải quyết sự mơ hồ đồng thời, mang lại một biểu diễn văn bản cuối cùng sạch hơn so với một quy trình xử lý tuần tự, vốn có thể khuếch đại lỗi.

**4.3. Xử lý Lỗi Chính tả Phụ âm và Lỗi Từ vựng trong Dữ liệu Tiếng Việt Nhiễu**

Ngoài việc thiếu dấu phụ, văn bản nhiễu thường chứa các lỗi nhầm lẫn phụ âm phổ biến đặc trưng cho cách phát âm khu vực (ví dụ: *ch-tr, s-x, d-r-gi, l-n*).14

Các mô hình học sâu giải quyết vấn đề này bằng cách huấn luyện trên các tập dữ liệu được tăng cường cẩn thận, trong đó các câu đúng được sửa đổi để bao gồm tất cả các khả năng nhầm lẫn phụ âm.14 Quy trình này bao gồm việc phát hiện âm tiết bị lỗi, tạo ra các từ ứng cử viên dựa trên sự tương đồng, và sử dụng mô hình ngôn ngữ (thường dựa trên học sâu) để chọn cách viết đúng.14

So sánh Hiệu suất của Các Kiến trúc Phục hồi Dấu Tiếng Việt

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Kiến trúc Mô hình** | **Phương pháp Tiếp cận** | **Thành phần Chính** | **Độ chính xác Từ (Xấp xỉ)** | **Hiệu suất/Đổi mới** |
| CNN-BiGRU (w/ Residual Block) 12 | Sequence-to-Sequence (Dựa trên RNN) | CNN, Bi-GRU, Residual Block | 94.77% | Độ chính xác cao; giải quyết vấn đề vanishing gradient. |
| Transformer Decoder Penalty (TDP) 9 | Transformer Decoder Cấp Ký tự | Transformer Decoder, Tầng Penalty | Hiệu suất hứa hẹn | Tốc độ và hiệu quả cao, quan trọng cho ứng dụng công nghiệp. |
| Transformer Hợp nhất (ví dụ: ByT5 variant) 13 | Sequence-to-Sequence (Dựa trên Transformer) | Transformer/Encoder-Decoder đầy đủ | Độ chính xác từ-alpha > 94% | Xử lý đồng thời phục hồi dấu và sửa lỗi chính tả tổng quát (SOTA về chịu lỗi). |

**V. Đột Phá, SOTA và Tương lai AI Tạo Sinh tại Việt Nam**

Phần này liên kết sự tiến bộ kiến trúc cụ thể với các xu hướng lớn hơn, tập trung vào chiến lược phát triển AI tại Việt Nam và việc áp dụng Mô hình Ngôn ngữ Lớn (LLMs).

**5.1. Các Đóng góp Nghiên cứu và Bộ công cụ Nổi bật**

**VnCoreNLP** được công nhận là nền tảng của NLP tiếng Việt, cung cấp một quy trình chú thích Java nhanh chóng và dễ sử dụng, đạt được kết quả SOTA trong các tác vụ cơ bản như tách từ, gán nhãn POS, nhận dạng thực thể có tên (NER) và phân tích phụ thuộc.15 Đây là bộ công cụ được thiết lập để chú thích ngôn ngữ nền tảng.

**5.2. Sự Tích hợp của Mô hình Ngôn ngữ Lớn (LLMs) vào Hệ sinh thái AI Việt Nam**

Việt Nam đang có sự đầu tư chiến lược vào phát triển AI. Chính phủ đã ưu tiên phát triển cơ sở hạ tầng AI quốc gia và các Mô hình Ngôn ngữ Lớn (LLMs) dành riêng cho tiếng Việt thông qua các chỉ thị như Quyết định 38/2020/QĐ-TTg của Thủ tướng Chính phủ.16 Điều này báo hiệu một sự chuyển đổi chiến lược lớn sang các giải pháp AI tạo sinh tự phát triển, thường được gọi là "ChatGPT đặc thù cho Việt Nam."

LLMs đã được triển khai một cách cấu trúc để hậu sửa lỗi ASR, tận dụng khả năng mô hình hóa ngữ cảnh ngôn ngữ phức tạp (đã thảo luận trong Mục III) và tăng cường danh sách nhiệm vụ cho AI đàm thoại.7

Chiến lược quốc gia này nhằm mục đích tích hợp AI, dữ liệu và công nghệ đám mây, ngụ ý rằng các đột phá trong tương lai sẽ ít tập trung vào việc cải tiến độ chính xác gia tăng trên các điểm chuẩn truyền thống mà sẽ tập trung vào **mở rộng quy mô mô hình và cơ sở hạ tầng tính toán** (trung tâm dữ liệu, điện toán đám mây) để hỗ trợ các khả năng LLM thực sự.16

**5.3. Ý nghĩa của AI Tạo Sinh đối với Nghiên cứu ASR/NLP Tiếng Việt**

Mặc dù ngành công nghiệp đang thúc đẩy việc áp dụng LLM, các thách thức đánh giá học thuật (như VLSP 2025) thường hạn chế việc sử dụng các tập dữ liệu bên ngoài hoặc các mô hình được huấn luyện trước lớn cho các nhiệm vụ cốt lõi như ASR/TTS.17

Sự ràng buộc này là rất quan trọng: nó buộc các nhà nghiên cứu phải tập trung vào việc tạo ra các kiến trúc mới, hiệu quả và các bộ dữ liệu nội địa vượt trội, ngăn ngừa sự phụ thuộc quá mức vào các mô hình đa ngôn ngữ lớn (có thể không nắm bắt được sự phức tạp tinh tế về chính tả và thanh điệu của tiếng Việt). Điều này đảm bảo sự phát triển liên tục của nghiên cứu SOTA đơn ngữ.

Các mô hình được huấn luyện trước (PLMs) như PhoBERT và ViSoBERT sẽ không bị lỗi thời. Thay vào đó, chúng sẽ đóng vai trò là các **bộ mã hóa và tokenizer chuyên biệt** mạnh mẽ, cung cấp dữ liệu đầu vào hoặc được tinh chỉnh trong các LLMs lớn hơn, đang nổi lên của Việt Nam. LLMs rất tốn kém về mặt tính toán để chạy cho mọi tác vụ nhỏ. Các PLM nội bộ, đặc biệt là những mô hình chuyên biệt cho văn bản nhiễu (ViSoBERT), sẽ vẫn cần thiết cho các tác vụ nhanh chóng, đặc thù miền và trích xuất đặc trưng hiệu quả, hình thành một thành phần quan trọng của kiến trúc AI phân cấp.

**VI. Kết luận Chiến lược và Khuyến nghị**

Phân tích này cho thấy một sự tổng hòa mạnh mẽ đang diễn ra giữa xử lý âm thanh và xử lý ngôn ngữ dựa trên các mô hình ngữ cảnh tiên tiến.

**6.1. Tổng hợp Điểm mạnh và Khoảng trống Công nghệ Hiện tại**

* **Điểm mạnh:** NLP tiếng Việt đã chuyển dịch rõ ràng từ các mô hình tĩnh sang các kiến trúc Transformer tinh vi (PhoBERT, ViSoBERT, mô hình tăng cường GAT), đạt hiệu suất cao trong cả miền văn bản sạch và nhiễu.3
* **Điểm mạnh:** Mối liên hệ nhân quả giữa xử lý âm thanh và ngôn ngữ được củng cố bởi các quy trình hậu sửa lỗi ASR mạnh mẽ, nhận biết ngữ cảnh, tận dụng LLMs để giảm đáng kể tác động của lỗi âm học.7
* **Khoảng trống:** Mặc dù có tiến bộ đáng kể (ví dụ: độ chính xác từ 94.77% trong phục hồi dấu 12), thách thức cốt lõi vẫn là sự mơ hồ vốn có cao của văn bản không dấu/nhiễu và sự khác biệt khu vực trong ASR.

**6.2. Hướng Nghiên cứu Chiến lược để Tăng cường Khả năng Chịu lỗi và Hiểu Ngữ cảnh**

Dựa trên những phát hiện trên, các hướng nghiên cứu và phát triển chiến lược nên tập trung vào:

1. **Ưu tiên Mô hình Chịu lỗi Hợp nhất:** Nghiên cứu trong tương lai cần tập trung vào việc tối ưu hóa các mô hình dựa trên Transformer để thực hiện đồng thời phục hồi dấu, sửa lỗi chính tả và chuẩn hóa tiếng lóng (dựa trên thành công của các mô hình như ByT5 và ViSoBERT).5 Việc này đảm bảo đầu ra văn bản sạch ngay lập tức phù hợp cho các tác vụ NLU cấp cao.
2. **Đầu tư vào Dữ liệu Đa phương thức và Giàu Ngữ cảnh:** Để vượt qua các thách thức ASR, cần tiếp tục đầu tư vào việc tạo ra các kho ngữ liệu quy mô lớn, được phiên âm, **đặc thù cho phương ngữ** 2 và phát triển các mô hình tăng cường LLM tích hợp sâu trạng thái hội thoại và kiến thức nhiệm vụ để xử lý lỗi ngữ cảnh tốt hơn.7
3. **Tập trung vào GAT và Ngữ cảnh Phi tuần tự:** Các đột phá kiến trúc nên tiếp tục khám phá việc kết hợp mô hình hóa quan hệ phi tuần tự (như GAT) 6 vào các PLM tổng quát để vượt qua các cơ chế chú ý tuyến tính và đạt được sự hiểu biết giống như đồ thị sâu hơn về cú pháp và ngữ nghĩa tiếng Việt, điều này đặc biệt quan trọng đối với các phụ thuộc tầm xa trong các câu phức tạp.

**Nguồn trích dẫn**

1. A study on the application of deep learning in Vietnamese speech recognition - | World Journal of Advanced Engineering Technology and Sciences, truy cập vào tháng 10 30, 2025, <https://journalwjaets.com/sites/default/files/fulltext_pdf/WJAETS-2025-0877.pdf>
2. Automatic Speech Recognition of Vietnamese for a New Large-Scale Corpus - MDPI, truy cập vào tháng 10 30, 2025, <https://www.mdpi.com/2079-9292/13/5/977>
3. Vietnamese Embedding · Models - Dataloop, truy cập vào tháng 10 30, 2025, <https://dataloop.ai/library/model/dangvantuan_vietnamese-embedding/>
4. Visobert · Models - Dataloop, truy cập vào tháng 10 30, 2025, <https://dataloop.ai/library/model/uitnlp_visobert/>
5. ViSoBERT: A Pre-Trained Language Model for Vietnamese Social Media Text Processing (EMNLP'2023) - GitHub, truy cập vào tháng 10 30, 2025, <https://github.com/uitnlp/ViSoBERT>
6. An Encoder-Integrated PhoBERT with Graph Attention for Vietnamese Token-Level Classification - arXiv, truy cập vào tháng 10 30, 2025, [https://arxiv.org/pdf/2510.11537?](https://arxiv.org/pdf/2510.11537)
7. Contextual ASR Error Handling with LLMs Augmentation for Goal-Oriented Conversational AI - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/html/2501.06129v1>
8. Contextual ASR Error Handling with LLMs Augmentation for Goal-Oriented Conversational AI - ACL Anthology, truy cập vào tháng 10 30, 2025, <https://aclanthology.org/2025.coling-industry.32.pdf>
9. TDP – A Hybrid Diacritic Restoration with Transformer Decoder - ACL Anthology, truy cập vào tháng 10 30, 2025, <https://aclanthology.org/2020.paclic-1.9.pdf>
10. Vietnamese Diacritics Restoration Using Deep Learning Approach - Semantic Scholar, truy cập vào tháng 10 30, 2025, <https://www.semanticscholar.org/paper/Vietnamese-Diacritics-Restoration-Using-Deep-Hung/b726244caff59dcb9cbd4814b8827fc8e5d0b0a3>
11. duongntbk/restore\_vietnamese\_diacritics: A Transformer based NLP solution to restore diacritics for Vietnamese text with 94.05% accuracy on the test dataset. You don't need to understand Vietnamese to use this, I promise :). - GitHub, truy cập vào tháng 10 30, 2025, <https://github.com/duongntbk/restore_vietnamese_diacritics>
12. Deep Learning Based Vietnamese Diacritics Restoration - National ..., truy cập vào tháng 10 30, 2025, <https://scholars.ncu.edu.tw/en/publications/deep-learning-based-vietnamese-diacritics-restoration/>
13. Correcting Diacritics and Typos with a ByT5 Transformer Model - MDPI, truy cập vào tháng 10 30, 2025, <https://www.mdpi.com/2076-3417/12/5/2636>
14. Deep Learning Approach for Vietnamese Consonant Misspell Correction - ResearchGate, truy cập vào tháng 10 30, 2025, <https://www.researchgate.net/publication/342620453_Deep_Learning_Approach_for_Vietnamese_Consonant_Misspell_Correction>
15. [1801.01331] VnCoreNLP: A Vietnamese Natural Language Processing Toolkit - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/abs/1801.01331>
16. Vietnam Generative AI Market trend - TechValley, truy cập vào tháng 10 30, 2025, <https://techvalley.biz/vietnam-generative-ai-market-trend-1295/>
17. VLSP 2025 Challenge on Vietnamese Voice Conversion, truy cập vào tháng 10 30, 2025, <https://vlsp.org.vn/vlsp2025/eval/vvc>