BỘ GIÁO DỤC VÀ ĐÀO TẠO **VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM**
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN THỊ THU HIỀN

NGHIÊN CỨU PHƯƠNG PHÁP CHUẨN HOÁ VĂN BẢN VÀ NHẬN DẠNG THỰC THỂ ĐỊNH DANH TRONG NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT

Chuyên ngành: Hệ thống thông tin
Mã số: 9 48 01 04

TÓM TẮT LUẬN ÁN TIẾN SỸ NGÀNH HỆ THỐNG THÔNG TIN

Hà Nội – 2023

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học 1: **PGS.TS. Lương Chi Mai**
Người hướng dẫn khoa học 2: **TS. Nguyễn Thị Minh Huyền**

Phản biện 1: PGS.TS. Ngô Xuân Bách
Phản biện 2: TS. Đỗ Văn Hải
Phản biện 3: PGS.TS. Nguyễn Phương Thái

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án tiến sĩ, họp tại Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi …. giờ ….', ngày …. tháng …. năm 2023

Có thể tìm hiểu luận án tại:
- Thư viện Học viện Khoa học và Công nghệ
- Thư viện Quốc gia Việt Nam

MỞ ĐẦU

Xử lý ngôn ngữ tự nhiên (XLNNTN) là lĩnh vực khoa học máy tính kết hợp giữa trí tuệ nhân tạo và ngôn ngữ học tính toán, nhằm xử lý tương tác giữa con người và máy tính sao cho máy tính có thể hiểu hay bắt chước được ngôn ngữ của con người. XLNNTN bao gồm hai nhánh lớn là xử lý tiếng nói và xử lý văn bản.

Một trong những bài toán quan trọng trong hiểu ngữ nghĩa văn bản viết hay nói là nhận dạng thực thể định danh (Named Entity Recognition - NER). Đây là một bài toán tiền đề cho các hệ thống về hiểu ngôn ngữ hay khai phá văn bản như trích xuất sự kiện, hỏi đáp tự động hay tìm kiếm ngữ nghĩa. Đã có nhiều nghiên cứu đạt được những kết quả rất khả quan cho bài toán NER với dữ liệu văn bản viết thông thường trong nhiều ngôn ngữ trên thế giới cũng như tiếng Việt. Trong khi đó, các nghiên cứu về nhận dạng thực thể định danh cho văn bản đầu ra của nhận dạng tiếng nói (Automatic Speech Recognition - ASR) có những khó khăn riêng so với văn bản viết, hầu như chưa có công trình nào cho tiếng Việt.

Nhận dạng tiếng nói là một quá trình chuyển đổi tín hiệu tiếng nói của một ngôn ngữ cụ thể thành một chuỗi các từ có nội dung tương ứng ở định dạng văn bản. Văn bản đầu ra của ASR thường không có cấu trúc, chẳng hạn như không có dấu câu, không viết hoa chữ cái đầu câu hoặc tên riêng, tên địa danh, ... Điều này dẫn đến khó khăn trong quá trình hiểu và hạn chế khả năng khai thác văn bản đầu ra của ASR trong hầu hết các ứng dụng. Việc nhận dạng thực thể định danh từ văn bản đầu ra của nhận dạng tiếng nói tự động do đó có những đặc trưng khác biệt vì nó luôn chứa nhiều lỗi nhận dạng, đặc biệt là các thực thể định danh nhiều khi nằm ngoài từ điển. Các lỗi ASR thường xảy ra trong các từ cấu thành nên thực thể định danh hoặc trong ngữ cảnh của những từ đó, do vậy làm ảnh hưởng trực tiếp đến hiệu suất của NER. Ngoài ra, các hệ thống NER phải đối mặt với những vấn đề về sự thiếu hụt một số dấu hiệu quan trọng như chữ viết hoa, dấu chấm câu. Bên cạnh đó, để cải thiện kết quả đầu ra của ASR, người ta cần chuẩn hóa văn bản bằng cách loại bỏ các từ vô nghĩa, chuẩn hóa dữ liệu kiểu số, ngày, tháng, khôi phục dấu câu và viết hoa, xử lý từ nước ngoài, ... Sau xử lý, văn bản cuối sẽ có cấu trúc tốt và dễ hiểu hơn so với văn bản đầu ra của ASR, đồng thời khi đưa vào triển khai trong các ứng dụng thực tế đạt hiệu quả cao hơn. Như vậy, việc phát triển các giải pháp chuẩn hoá văn bản và nhận dạng thực thể định danh từ văn bản đầu ra của ASR là cần thiết để cải thiện chất lượng tổng thể của hệ thống ASR.

Tuy nhiên, việc chuẩn hoá văn bản đầu ra của ASR, cụ thể là vấn đề khôi phục dấu câu, chữ hoa vẫn còn không ít vấn đề cần cải thiện. Bên cạnh ý nghĩa trong việc cải thiện chất lượng đầu ra của ASR thì dấu câu, chữ hoa cũng là một

trong những thông tin quan trọng, hữu ích cho bài toán nhận dạng thực thể định danh. Có thể thấy, không phải tất cả các từ viết hoa trong tiếng Việt đều được coi là thực thể định danh. Ngược lại, thực thể định danh cũng không nhất thiết là các từ/cụm từ viết hoa đầy đủ. Đặc biệt, cũng là thực thể định danh nhưng được phân loại thành các dạng thực thể khác nhau. Do đó, việc khôi phục dấu câu, chữ hoa là một trong các yếu tố quan trọng giúp tối ưu hóa hệ thống nhận dạng thực thể định danh trong văn bản đầu ra ASR.

Trong thực tế, đã có nhiều phương pháp xử lý NER cho văn bản đầu ra ASR nhưng chủ yếu tập trung ở ngôn ngữ giàu tài nguyên như tiếng Anh, tiếng Trung, tiếng Nhật. Có rất ít nghiên cứu áp dụng NER cho ASR tiếng Việt và các nghiên cứu này cũng mới chỉ tập trung cho văn bản hội thoại ngắn. Từ những thách thức đó, nghiên cứu sinh đã lựa chọn nghiên cứu đề tài *"Nghiên cứu phương pháp chuẩn hoá văn bản và nhận dạng thực thể định danh trong nhận dạng tiếng nói tiếng Việt"*.

Mục tiêu và nhiệm vụ nghiên cứu: Luận án tập trung đề xuất giải pháp và triển khai thực nghiệm cho hai mục tiêu cụ thể. *Thứ nhất* là chuẩn hóa văn bản bằng cách khôi phục dấu câu, chữ hoa, *thứ hai* là nhận dạng thực thể định danh trên văn bản đầu ra của hệ thống ASR tiếng Việt.

Nội dung nghiên cứu: Luận án nghiên cứu đặc thù dữ liệu và lỗi đầu ra của các hệ thống ASR tiếng Việt, tìm hiểu các vấn đề cơ bản của bài toán NER cũng như các thách thức của bài toán. Tiếp theo, xây dựng bộ dữ liệu phục vụ cho việc huấn luyện và đánh giá các mô hình. Trên cơ sở đó, đề xuất mô hình khôi phục dấu câu và chữ hoa phục vụ chuẩn hóa văn bản đầu ra của ASR tiếng Việt. Bài toán NER cho văn bản đầu ra của ASR tiếng Việt được nghiên cứu giải quyết theo hai hướng: hệ thống đường ống và hệ thống đầu-cuối.

Phạm vi nghiên cứu: Nghiên cứu sẽ tập trung vào hướng giải quyết các vấn đề liên quan đến xử lý văn bản đầu ra của ASR với văn bản tiếng nói dài, khó xử lý. Bên cạnh đó, với vấn đề chuẩn hóa văn bản đầu ra của ASR, nghiên cứu chỉ tập trung thiết kế mô hình dự đoán dấu câu, chữ hoa và coi hệ thống ASR có tỉ lệ lỗi từ (WER) bằng 0%. Về mô hình giải quyết bài toán NER, luận án sử dụng hệ thống ASR thực tế có WER là 4.85%.

Phương pháp nghiên cứu, triển khai: Luận án đã thực hiện nghiên cứu lý thuyết, bao gồm tổng quan về các bài toán cần giải quyết, các phương pháp, kĩ thuật đã được sử dụng để giải quyết các bài toán này và hiệu quả của chúng. Trên cơ sở đó, luận án đề xuất các giải pháp để khắc phục một số vấn đề còn tồn tại. Luận án cũng chú trọng triển khai phương pháp thực nghiệm nhằm đo lường, đánh giá các mô hình đề xuất giải quyết bài toán, so sánh với các phương pháp khác. Về dữ liệu thực nghiệm, luận án cần xây dựng các bộ dữ liệu văn bản kết hợp với tiếng nói tương ứng nhằm đáp ứng các bài toán đặt ra.

Các đóng góp của luận án: Xây dựng các bộ dữ liệu văn bản kết hợp với tiếng nói cho huấn luyện và đánh giá các mô hình chuẩn hoá và nhận dạng thực thể định danh cho văn bản đầu ra của các hệ thống ASR. Các dữ liệu này được mô tả trong các công trình [CT1, CT2, CT4, CT6]; Đề xuất và cải tiến mô hình khôi phục dấu câu và chữ hoa giúp chuẩn hoá văn bản đầu ra của ASR tiếng Việt. Mô hình này được đưa ra, đánh giá và cải tiến trong các công trình [CT2, CT3, CT5]; Đề xuất hai giải pháp nhận dạng thực thể định danh trong văn bản đầu ra của ASR tiếng Việt theo hướng tiếp cận Pipeline và E2E. Các giải pháp này được trình bày và đánh giá trong các công trình [CT4, CT6].

Bố cục luận án: Ngoài phần mở đầu và kết luận, luận án được cấu trúc thành 4 chương. Chương 1 trình bày tổng quan các vấn đề nghiên cứu. Chương này phát biểu và nêu ý nghĩa ứng dụng của các bài toán, chỉ ra các thách thức cần giải quyết và khảo sát các nghiên cứu về nhận dạng tiếng nói và nhận dạng thực thể định danh từ tiếng nói nói chung và đối với tiếng Việt nói riêng. Chương 2 - Kiến thức cơ sở, trình bày những kiến thức nền tảng được sử dụng để định hướng và là cơ sở để đề xuất mô hình chuẩn hoá và nhận dạng thực thể định danh cho văn bản đầu ra của ASR. Chương 3 giới thiệu về bài toán khôi phục dấu câu và chữ hoa cho hệ thống ASR tiếng Việt. Trong chương này, luận án trình bày mô hình đề xuất, dữ liệu và các kết quả thực nghiệm cho bài toán. Chương 4 đề xuất phương pháp nhận dạng thực thể định danh cho văn bản đầu ra của ASR tiếng Việt theo hai hướng tiếp cận đường ống và E2E, trình bày các kết quả thực nghiệm, và so sánh hai cách tiếp cận.

Chương 1. TỔNG QUAN VẤN ĐỀ NGHIÊN CỨU

Với văn bản đầu ra của ASR, các thông tin đặc trưng về dấu câu, chữ hoa cho NER không còn tồn tại, gây nhiều khó khăn cho xử lý. Do đó, việc nghiên cứu, xử lý và chuẩn hóa văn bản đầu ra của ASR, giúp cải tiến hệ thống ASR và phục vụ cho đầu vào của hệ thống NER là quan trọng và có ý nghĩa. Chương này sẽ trình bày tổng quan về XLNNTN, những khó khăn khi xử lý ngôn ngữ tiếng Việt. Tìm hiểu chung về hệ thống ASR, những đặc trưng trong văn bản đầu ra của hệ thống ASR và các nghiên cứu liên quan đến việc chuẩn hóa văn bản đầu ra của ASR giúp hỗ trợ cho mô hình NER. Tiếp theo, luận án mô tả bài toán NER, những khó khăn khi xử lý NER cho tiếng nói tiếng Việt và các nghiên cứu liên quan. Cuối chương sẽ trình bày tổng quan về dữ liệu sử dụng trong từng bài toán.

1.1. Xử lý ngôn ngữ tự nhiên

1.1.1. Giới thiệu

XLNNTN là một lĩnh vực con trong khoa học máy tính, kết hợp giữa trí tuệ nhân tạo và ngôn ngữ học tính toán. Các công cụ như phân tích, nhận dạng cảm xúc, nhận dạng thực thể định danh, phân tích cú pháp, ngữ nghĩa, ... đã giúp

XLNNTN trở thành chủ đề hay để nghiên cứu trong nhiều lĩnh vực khác nhau như dịch máy, trích xuất thông tin, tóm tắt văn bản, trả lời câu hỏi tự động, ... Nhiều ứng dụng XLNNTN trên các thiết bị thông minh xuất hiện ở khắp mọi nơi, thu hút được nhiều sự quan tâm của cộng đồng.

XLNNTN có thể được chia ra thành hai nhánh lớn, bao gồm xử lý tiếng nói và xử lý văn bản. Vấn đề xử lý văn bản sau nhận dạng tiếng nói là một thách thức cần được giải quyết. Luận án cũng đặt ra vấn đề cần chuẩn hoá văn bản đầu ra của nhận dạng tiếng nói tiếng Việt và nhận dạng thực thể định danh.

1.2. Nhận dạng tiếng nói tự động

1.2.1. Giới thiệu sơ lược về hệ thống nhận dạng tiếng nói tự động

Nhận dạng tiếng nói tự động được Yu và Deng phát biểu như sau: *"đó là một thuật ngữ được sử dụng để mô tả các quy trình, công nghệ và phương pháp cho phép tương tác giữa người và máy tính tốt hơn thông qua việc dịch tiếng nói của con người sang định dạng văn bản"* [3].

Một cách phổ biến nhất thường được sử dụng để đánh giá hiệu suất của hệ thống ASR chính là WER. Số liệu WER dựa trên khoảng cách Levenshtein, đo lường số lần chèn, xóa và thay thế trong một chuỗi.

$$WER = \frac{I + D + S}{N} * 100 \quad (1.1)$$

trong đó, *I* là số lần chèn, *D* là số lần xóa, *S* là số lần thay thế và *N* là số từ trong văn bản.

Đối với hệ thống ASR tiếng Việt, tại VLSP đã sử dụng tỷ lệ lỗi âm tiết (SyER) thay vì tỷ lệ lỗi từ để đánh giá hiệu suất của hệ thống ASR.

$$SyER = \frac{S + D + I}{N} \quad (1.2)$$

trong đó, *S* là số lần thay thế, *D* là số lần xóa, *I* là số lần chèn, *C* là số lượng âm tiết và *N* là số lượng âm tiết trong văn bản.

1.2.2. Đặc trưng văn bản đầu ra của hệ thống nhận dạng tiếng nói và các vấn đề cần xử lý

Văn bản đầu ra của ASR thường có những đặc trưng riêng, khác so với văn bản viết thông thường, đặc biệt là trong tiếng Việt: Văn bản không chứa dấu câu và chữ hoa; Các từ tên riêng nước ngoài, các chữ viết tắt không được nhận dạng chính xác; Kiểu số, kiểu tiền tệ nhận dạng thành kiểu chữ cái, địa chỉ email hoặc địa chỉ website hay các siêu liên kết thường là một cụm từ một liên tục và có quy chuẩn nhưng bị nhận dạng thành các từ, cụm từ không tuân theo quy tắc chuẩn; tiếng Việt có rất nhiều từ vay mượn từ các ngôn ngữ khác để tạo ra từ mới; chèn từ, xoá từ, thay thế từ,…

1.3. Chuẩn hoá văn bản đầu ra của nhận dạng tiếng nói

1.3.1. Vấn đề khôi phục dấu câu, chữ hoa

Viết hoa chính là việc xác định chính xác dạng của từ, phân biệt giữa bốn loại: tất cả các chữ cái viết thường, tất cả các chữ cái viết hoa, chỉ viết hoa chữ cái đầu tiên của âm tiết và chữ hoa hỗn hợp bao gồm một số chữ cái viết hoa và một số chữ cái viết. Khôi phục dấu câu là nhiệm vụ chèn chúng vào các vị trí thích hợp trong một văn bản đầu vào không có bất kỳ dấu câu nào.

Mặt khác, quy tắc viết hoa chữ cái đầu âm tiết thứ nhất của một câu hoàn chỉnh cho thấy sự liên quan giữa chữ hoa và dấu câu, nghĩa là hai nhiệm vụ này cần phải được xử lý cùng lúc. Tuy nhiên, các nghiên cứu thường tập trung giải quyết một nhiệm vụ cụ thể. Rõ ràng, kết quả xử lý đơn lẻ như vậy không thể giúp cải thiện hiệu quả đầu ra của ASR, dẫn đến gần đây xuất hiện ngày càng nhiều các hướng nghiên cứu tích hợp cả hai nhiệm vụ. Ngay cả khi xử lý tích hợp thì việc xác định khôi phục dấu câu hay chữ hoa trước cũng là một vấn đề vì thứ tự xử lý cũng có thể sẽ ảnh hưởng lẫn nhau cũng như đến kết quả cuối cùng [13].

1.3.2. Các phương pháp xử lý

Một trong những cách triển khai ban đầu cho phương pháp viết hoa tự động là dựa trên tập luật, nghĩa là sử dụng nguyên tắc xác định phần bắt đầu của một câu mới để chỉ ra kí tự được viết hoa [17]. Các nghiên cứu chỉ ra rằng, hệ thống dựa trên luật khó duy trì vì chúng có thể liên tục yêu cầu bổ sung các luật mới. Mô hình ngôn ngữ là mô hình tính xác suất giúp dự đoán từ tiếp theo trong chuỗi các từ. Mô hình ngôn ngữ tính xác suất của một từ w_k cho trước trong ngữ cảnh của $n-1$ từ trước đó $w_{k-1}, w_{k-2}, ..., w_{k-(n-1)}$. Xác suất này có thể được biểu thị bởi $P(w_k| w_{k-1}, w_{k-2}, ..., w_{k-(n-1)})$. Các nghiên cứu về khôi phục dấu câu và mô hình kết hợp [19] dựa trên mô hình ngôn ngữ n-gram đã được đề xuất. Nhược điểm của mô hình n-gram là không đánh giá được ngữ cảnh của toàn bộ câu, do đó, trong nhiều trường hợp không thể đưa ra một xác suất chính xác. Ngay cả với các tài nguyên máy tính ngày nay về khả năng lưu trữ và xử lý, các mô hình có số n cao vẫn khó xử lý do yêu cầu lưu trữ của chúng. Theo các nhà nghiên cứu, viết hoa hay dấu câu có thể được coi là một vấn đề gán nhãn tuần tự. Với một chuỗi $W=w_0w_1w_2...w_n$, mô hình dự đoán chuỗi viết hoa $C=c_0c_1c_2...c_n$ với c_i tương ứng với tất cả viết thường, viết hoa chữ đầu tiên, viết hoa tất cả và viết hoa trộn lẫn. Tương tự, dự đoán dấu câu $E=e_0e_1e_2...e_n$ trong đó e_i biểu thị một dấu câu hoặc không có dấu câu nào. Một số nghiên cứu sử dụng mô hình Entropy cực đại [21], mô hình Markov ẩn [22] và mô hình Markov Entropy cực đại [23] cho cả hai nhiệm vụ. Trường ngẫu nhiên có điều kiện cũng là mô hình xác suất được sử dụng để phân đoạn và gán nhãn dữ liệu chuỗi [24]. CRF có ưu điểm hơn so với MEMM và các mô hình Markov khác do CRF là một mô hình đồ thị vô hướng, cho phép CRF có thể định nghĩa phân phối xác suất của toàn bộ trạng thái.

Gần đây, các nghiên cứu đã sử dụng kiến trúc mạng nơ-ron cho bài toán khôi phục dấu câu, chữ hoa. Với tiếp cận mạng nơ-ron, có thể đưa ra mô hình mạng cho cả mức từ và mức ký tự. Susanto và các cộng sự [27] đã đề xuất sử dụng mạng nơ-ron hồi quy ở cấp ký tự để xử lý sai lệch trong các trường hợp viết hoa trộn lẫn (ví dụ: MacKenzie). RNN đã chứng minh sự hữu ích trong việc lập mô hình dữ liệu tuần tự. Tilk và các cộng sự [28] đã sử dụng mô hình mạng nơ-ron hồi quy hai chiều có thêm một tầng ẩn cho phép xử lý dữ liệu theo chiều ngược lại một cách linh hoạt hơn so với RNN truyền thống, kết hợp với cơ chế chú ý để khôi phục dấu chấm câu. Mô hình này có thể sử dụng các ngữ cảnh dài theo cả hai hướng và hướng sự chú ý khi cần thiết, cho phép hoạt động tốt hơn trên các tập dữ liệu về tiếng Anh và tiếng Estonia trước đây. Kể từ năm 2017, với sự ra đời của kiến trúc Transformer [29], các phiên bản khác nhau BERT [30], RoBERTa [31] đã mở ra nhiều hướng nghiên cứu mới. Rei và các cộng sự [32] đã ứng dụng khôi phục viết hoa phụ đề video được tạo bởi hệ thống ASR sử dụng mô hình BERT. Cách tiếp cận này dựa trên mã hóa từ theo ngữ cảnh được huấn luyện trước và áp dụng tinh chỉnh bằng các mô hình fine-tuning. Phương pháp này chứng minh sự vượt trội so với các phương pháp tiếp cận khác không chỉ về hiệu suất mà còn về thời gian tính toán. Nhóm nghiên cứu của Alam [33] đã thử nghiệm một số mô hình Transformer như BERT, RoBERTa, ALBERT, DistilBERT, mBERT, XLM-RoBERTa cho tiếng Anh và ngôn ngữ Bangla. Đối với tiếng Anh, các kết quả tốt nhất quan sát được trên mô hình $\text{RoBERTa}_{\text{LARGE}}$ khi khôi phục tốt dấu chấm, tuy nhiên hiệu quả xử lý dấu phẩy và dấu hỏi chấm lại tương đối thấp.

Vấn đề nghiên cứu khôi phục dấu câu, chữ hoa đối với văn bản đầu ra tiếng nói tiếng Việt vẫn còn khá mới mẻ. Gần đây, Thuy Nguyen và cộng sự [34] đã thử nghiệm mô hình mạng nơ-ron học sâu BiLSTM và Hieu Dinh cùng cộng sự [35] đã sử dụng mô hình Transformer cho khôi phục dấu câu. Kết quả nghiên cứu đầu tiên được nghiên cứu sinh và các cộng sự đề xuất đã tập trung giải quyết vấn đề về khôi phục chữ hoa riêng lẻ. Tiếp theo đó, mô hình được tích hợp để có thể khôi phục đồng thời dấu câu và chữ hoa cho tiếng nói tiếng Việt. Các kết quả nghiên cứu mới này cho tiếng Việt được coi là tiền đề cho các nghiên cứu tiếp tục phát triển nhằm chuẩn hóa văn bản đầu ra của hệ thống ASR tiếng Việt cho các mục đích cụ thể. Cụ thể, Uyen và các cộng sự [13] đã đề xuất kiến trúc mô hình JointCapPunc để khôi phục dấu câu, chữ hoa theo kiến trúc xếp tầng, nghĩa là khôi phục chữ hoa trước sau đó mới đến lớp khôi phục dấu câu. Một mô hình ngôn ngữ được huấn luyện trước Transformer như vậy sẽ có tham số lớn, gây khó khăn trong mô hình Pipeline do sự gia tăng độ trễ. Ngoài ra, dữ liệu nghiên cứu cũng được thực hiện trên các đoạn hội thoại ngắn, trong lĩnh vực y tế.

1.4. Nhận dạng thực thể định danh

1.4.1. Định nghĩa

Sundheim và Grishman giới thiệu lần đầu tiên tại hội nghị MUC-6 [36]: *"Nhận dạng thực thể định danh là một quá trình xác định tìm kiếm các từ hoặc cụm từ có nghĩa từ văn bản ngôn ngữ tự nhiên phân loại thành các nhóm duy nhất được định nghĩa trước đó như: tên người, tên tổ chức, ngày giờ, địa điểm, con số, tiền tệ…"*. Aggarwal, C. C [37] phát biểu: *"Nhận dạng thực thể định danh là bài toán xác định thực thể có tên từ các văn bản dưới dạng tự do và phân lớp chúng vào một tập các kiểu được định nghĩa trước như người, tổ chức và địa điểm"*.

1.4.2. Thách thức cho bài toán NER trong văn bản đầu ra của ASR tiếng Việt

Tiếng Việt chưa có dữ liệu văn bản đầu ra ASR có gán nhãn NER chuẩn đủ lớn phục vụ cho huấn luyện, đánh giá. Những thách thức cho bài toán NER trong văn bản đầu ra của ASR tiếng Việt bao gồm: Trong các văn bản đầu ra của ASR, việc viết hoa bị bỏ qua gây khó khăn cho hệ thống nhận dạng. Việc xác định biên của một từ trong tiếng Việt khó khăn hơn so với các ngôn ngữ khác, do tiếng Việt thuộc loại hình ngôn ngữ đơn lập, tức là, một từ có thể được tạo nên bởi một hoặc nhiều tiếng. Yêu cầu hệ thống có khả năng phân biệt loại thực thể. Do không có nhiều ràng buộc về tên riêng nên có thể khiến hệ thống bỏ qua hoặc nhầm nó với một thực thể khác. Đặc biệt, lỗi ASR làm cho các thực thể định danh bị bỏ sót và các thực thể định dạng bị nhận dạng sai. Nếu một hoặc nhiều từ cấu thành thực thể định danh bị nhận dạng sai thì rất khó để nhận ra đúng thực thể định danh. Ngược lại, ngay cả khi tất cả các từ cấu thành thực thể định danh được nhận dạng chính xác, cũng có thể không nhận ra đúng thực thể định danh do thiếu ngữ cảnh trong văn bản đầu ra của ASR. Tên nước ngoài, tên viết tắt trong văn bản đầu ra ASR cũng có thể bị nhận dạng theo nhiều phiên bản khác nhau. Hiện tượng đồng âm khác nghĩa trong tiếng Việt phổ biến hơn các ngôn ngữ Ấn- Âu.

1.4.3. Tình hình nghiên cứu NER cho văn bản đầu ra của ASR

1.4.3.1. Các nghiên cứu theo hương tiếp cận Pipeline

Trong giai đoạn đầu tiên, Kim và cộng sự [42] đã đề xuất nhận dạng thực thể định danh trên văn bản đầu ra của ASR dựa trên tập luật. Ưu điểm của phương pháp là yêu cầu lưu trữ nhỏ, có thể mở rộng các luật. Tuy nhiên, nhược điểm là các quy tắc cần được xây dựng thủ công, đặc biệt khi đầu vào là văn bản đầu ra của ASR thì thông tin viết hoa cho thực thể định danh sẽ không còn nữa, việc lấy thông tin ngôn ngữ cần thiết để xây dựng các luật sẽ khó khăn. Để khắc phục điều này, rất nhiều các nghiên cứu dựa trên học máy đã được các nhà nghiên cứu đề xuất như mô hình HMM [43], mô hình entropy cực đại [44], CRF [45], [46], HMM-CRF [47], máy véc-tơ hỗ trợ [48] và tập trung chủ yếu cho tiếng Anh, tiếng Trung, tiếng Nhật, tiếng Pháp. Các nghiên cứu cũng chỉ ra rằng cần kết hợp thêm các đặc trưng về âm tiết, kết hợp các thông tin dấu câu, chữ hoa và cải thiện lỗi trong văn bản đầu ra của ASR để tăng hiệu suất NER.

Có thể nhận thấy, với cách tiếp cận Pipeline, thành phần NER phải đối phó với một văn bản không chuẩn hóa như văn bản thông thường và chứa nhiễu [52]. Cách tiếp cận này sẽ chịu ảnh hưởng của lỗi văn bản đầu ra của ASR và sự lan truyền lỗi qua từng bước.

1.4.3.2. Các nghiên cứu theo hương tiếp cận End-to-End

Ghannay và các cộng sự [53] đã đề xuất thử nghiệm đầu tiên phương pháp nhận dạng thực thể định danh từ tiếng nói tiếng Pháp theo hướng E2E. Các tác giả đề xuất mô hình kiến trúc RNN sâu, bao gồm *nc* lớp tích chập, tiếp theo là *nr* lớp lặp lại một chiều hoặc hai chiều, một lớp tích chập tìm kiếm và một lớp được kết nối đầy đủ ngay trước lớp Softmax. Hệ thống được huấn luyện E2E bằng cách sử dụng hàm CTC-loss [10] để dự đoán chuỗi ký tự từ âm thanh đầu vào. Kết quả thực nghiệm cho thấy, mô hình E2E vẫn kém hiệu quả hơn so với Pipeline kết hợp tính năng POS được sử dụng để gán nhãn đầu ra ASR trước khi xử lý NER và cho rằng POS thực sự quan trọng đối với nhiệm vụ NER. Caubriere và cộng sự [54] đã triển khai E2E dựa trên hệ thống DeepSpeech2 với kiến trúc bao gồm một chồng hai lớp 2D-invariant convolutional, năm lớp biLSTM và một lớp softmax cuối cùng. Hệ thống cũng sử dụng hàm CTC-loss cho phép liên kết giữa âm thanh đầu vào và chuỗi ký tự đầu ra. So sánh với kết quả tốt nhất của chiến dịch đánh giá ETAPE, hệ thống E2E đề xuất đã cho thấy mức độ cải thiện tương đối là 4%, cách tiếp cận này cũng chưa đạt hiệu suất tốt hơn so với phương pháp Pipeline mà các tác giả đề xuất trong cùng nghiên cứu. Theo Chan và các cộng sự [55], khi thực nghiệm mô hình Pipeline đề xuất sử dụng BERT để huấn luyện trước vẫn đạt hiệu suất cao hơn E2E và cho rằng, mặc dù các mô-đun trong Pipeline có thể bị ảnh hưởng bởi sự lan truyền lỗi, chúng vẫn có thể tận dụng việc huấn luyện trước để tăng hiệu suất, đặc biệt khi hệ thống ASR được cải thiện tốt.

1.5. Tổng quan dữ liệu

Để phục vụ cho mục đích huấn luyện và đánh giá mô hình chuẩn hoá văn bản đầu ra của hệ thống ASR trong Chương 3, nghiên cứu cần xây dựng bộ dữ liệu lớn, tập văn bản này được xóa định dạng (bỏ dấu câu, chuyển chữ hoa thành chữ thường).

Bộ dữ liệu văn bản và âm thanh đã gán nhãn mẫu phục vụ mục đích huấn luyện và đánh giá mô hình cho bài toán NER theo hướng tiếp cận đường ống và E2E trong Chương 4 được tận dụng từ bộ dữ liệu văn bản NER VLSP 2018[1]. Tương ứng với tập văn bản chuẩn này là tập văn bản được xóa định dạng và dữ liệu thu âm với các giọng đọc khác nhau, trong môi trường khác nhau. Đồng thời, để tiết kiệm chi phí thu âm, tất cả dữ liệu văn bản của VLSP sẽ sử dụng hệ thống TTS của Google để tạo ra dữ liệu âm thanh tổng hợp. Sau đó, bộ dữ liệu

[1] Dữ liệu từ cuộc thi NER tại Hội thảo VLSP (Vietnamese Language and Speech Processing) 2018: https://vlsp.org.vn/vlsp2018/ner

âm thanh tổng hợp sẽ qua hệ thống ASR của VAIS để được bộ dữ liệu văn bản phục vụ huấn luyện mô hình NER E2E. Chi tiết về các bộ dữ liệu sẽ được mô tả cụ thể trong Chương 3, Chương 4.

1.6. Kết luận Chương 1

Chương 1 đã trình bày tổng quan về XLNNTN, các khó khăn trong xử lý ngôn ngữ tiếng Việt. Những nghiên cứu về đặc trưng văn bản đầu ra ASR, các vấn đề cần giải quyết và tổng quan các nghiên cứu liên quan giúp chuẩn hóa văn bản đầu ra ASR đã được trình bày. Bên cạnh giới thiệu cơ bản về bài toán NER, tầm quan trọng và cách thức đánh giá hệ thống, nghiên cứu cũng đưa ra những thách thức đối với bài toán NER trong văn bản đầu ra của ASR tiếng Việt và các nghiên cứu liên quan để từ đó xác định những nội dung cần giải quyết. Đồng thời, Chương 1 cũng đã giới thiệu tổng quan về các bộ dữ liệu sử dụng trong luận án.

Chương 2. KIẾN THỨC CƠ SỞ

Chương 2 trình bày chi tiết về một số mô hình học sâu cho xử lý chuỗi, mô hình biểu diễn từ và mô hình gán nhãn chuỗi. Những kiến thức nền tảng này là cơ sở quan trọng để định hướng việc đề xuất các mô hình chuẩn hoá và nhận dạng thực thể định danh cho văn bản đầu ra của ASR tiếng Việt trong Chương 3, Chương 4. Đồng thời, Chương 2 cũng giới thiệu về phương pháp học đa tác vụ, chương 4 sẽ áp dụng phương pháp này để thiết kế một mô hình nhận dạng thực thể định danh theo hướng E2E.

2.1. Mô hình xử lý chuỗi

2.1.1. GRU

Rất khó để nắm bắt sự phụ thuộc khoảng cách xa bằng cách sử dụng mô hình RNN vì các gradient có xu hướng suy biến hoặc loại bỏ với các chuỗi dài, do đó, mô hình GRU [73] đã được đề xuất để giải quyết vấn đề này. Sự khác biệt chính giữa RNN thông thường và GRU là GRU hỗ trợ việc kiểm soát trạng thái ẩn. Điều này có nghĩa là có các cơ chế để quyết định khi nào nên cập nhật và khi nào nên xóa trạng thái ẩn.

Mô hình GRU giảm tín hiệu cổng thành hai so với mô hình LSTM. Hai cổng được gọi là cổng cập nhật z_t và một cổng đặt lại r_t.

Mặc dù vậy, GRU cũng tồn tại một số hạn chế khi xử lý các chuỗi dữ liệu rất dài như: có khả năng mất mát thông tin quan trọng trong quá trình xử lý chuỗi, vẫn giới hạn về khả năng mô hình hóa mối quan hệ phức tạp trong chuỗi, cần nhiều tham số để huấn luyện, do đó làm tăng yêu cầu về lượng dữ liệu huấn luyện và tài nguyên tính toán.

Sự ra đời của mô hình Transformer đã tạo ra bước đột phá mới, giúp mô hình xử lý hiệu quả với nhiều tác vụ khác nhau, đồng thời hạn chế được một số nhược điểm của RNN và các biến thể của nó như LSTM hay GRU. Luận án đã

áp dụng mô hình Transformer trong thiết kế mô hình chuẩn hoá văn bản đầu ra của ASR tiếng Việt ở Chương 3.

2.1.2. Transformer

Transformer là mô hình học sâu, trong đó sử dụng cơ chế chú ý *(attention)* để tính toán ảnh hưởng của các biến đầu vào đến kết quả đầu ra. Mô hình này được dùng phổ biến trong lĩnh vực XLNNTN, tuy nhiên gần đây còn được phát triển cho các ứng dụng khác như thị giác máy, xử lý tiếng nói.

Giống như những mô hình dịch máy khác, kiến trúc tổng quan của mô hình Transformer bao gồm hai phần chính là bộ mã hóa *(Encoder)* và bộ giải mã *(Decoder)*. Trong mô hình Transformer, bộ mã hoá chịu trách nhiệm xử lý đầu vào và biểu diễn các từ hoặc câu thành các véc-tơ biểu diễn có ý nghĩa. Bộ giải mã có nhiệm vụ chuyển đổi biểu diễn của đầu vào thành một chuỗi đầu ra.

Mô hình Transformer sử dụng nhiều khối mã hóa và giải mã để xử lý dữ liệu. Mỗi khối bao gồm một tầng tự chú ý đa đỉnh và mạng nơ-ron truyền thẳng. Tầng tự chú ý đa đỉnh cho phép mô hình học các biểu diễn đa chiều của câu, trong khi mạng nơ-ron truyền thẳng học các biểu diễn phi tuyến của từng vị trí.

Tự chú ý: là một cơ chế quan trọng trong mô hình Transformer, cho phép mô hình xác định mức độ quan trọng của các từ trong câu bằng cách tính toán một trọng số cho mỗi từ dựa trên tương quan với các từ khác. Điều này giúp mô hình hiểu được mối quan hệ ngữ nghĩa và cú pháp trong câu.

Cơ chế chú ý đa đỉnh: Trong mô hình Transformer, mỗi tầng tự chú ý sử dụng cơ chế chú ý đa đỉnh. Cơ chế này cho phép mô hình học các biểu diễn đa chiều của câu bằng cách tính toán chú ý từ nhiều không gian biểu diễn khác nhau, giúp tăng khả năng học các mối quan hệ phức tạp trong câu. Việc sử dụng cơ chế chú ý đa đỉnh giúp mô hình học được nhiều khía cạnh khác nhau của câu và cung cấp biểu diễn phong phú hơn cho dữ liệu đầu vào.

2.2. Mô hình biểu diễn từ

2.2.1. Word2Vec

Được phát triển bởi Tomas Mikolov và các cộng sự tại Google vào năm 2013, Word2Vec là một kỹ thuật biểu diễn véc-tơ từ để giải quyết các vấn đề XLNNTN nâng cao. Nó có thể lặp lại trên một kho văn bản lớn để tìm hiểu các liên kết hoặc sự phụ thuộc giữa các từ. Word2Vec xác định mối quan hệ ngữ nghĩa giữa từ bằng cách dự đoán từ hiện tại dựa trên ngữ cảnh xung quanh nó hoặc ngược lại. Kết quả của Word2Vec là các biểu diễn véc-tơ từ, có thể được sử dụng trong các mô hình học máy khác nhau [69].

Word2Vec cung cấp hai biến thể dựa trên mạng nơ-ron: CBOW và Skip-gram. CBOW dự đoán từ hiện tại dựa trên ngữ cảnh xung quanh nó. Đầu vào của CBOW là một cửa sổ các từ xung quanh từ hiện tại và mục tiêu là dự đoán từ hiện tại. Ngược lại, skip-gram cố gắng dự đoán ngữ cảnh xung quanh từ hiện tại dựa trên từ hiện tại. Skip-gram lấy từ hiện tại và dự đoán các từ trong ngữ cảnh

xung quanh nó. Sau khi đã trích xuất các biểu diễn véc-tơ từ từ mô hình Word2Vec, chúng có thể được sử dụng để thực hiện các tác vụ trong XLNNTN.

Khi có một lượng dữ liệu lớn và cần mô hình học biểu diễn từ ngữ phức tạp, giúp nắm bắt được các mối quan hệ tương quan giữa từ trong câu, hiểu được ý nghĩa của từ trong ngữ cảnh cụ thể và tạo ra các biểu diễn phù hợp thì các mô hình học sâu trở lên phù hợp hơn. Với sự ra đời của mô hình Transformer, nhiều biến thể mới được mở rộng, luận án đã cải tiến mô hình BERT cho dữ liệu tiếng Việt khi đề xuất mô hình nhận dạng thực thể định danh.

2.2.2. BERT

BERT là một mô hình ngôn ngữ học sâu, được giới thiệu bởi Jacob Devlin và các cộng sự tại Google Research vào năm 2018.

Kiến trúc chung: Mô hình BERT có kiến trúc mạng học sâu sử dụng nhiều tầng mã hoá Transformer. Tuy nhiên, điểm đặc biệt của BERT là sử dụng hai biểu diễn từ: biểu diễn từ vào và biểu diễn từ ra [71].

BERT là một phương pháp mới để tiền huấn luyện các bộ biểu diễn véc-tơ từ. Một điểm đặc biệt ở BERT mà các mô hình biểu diễn véc-tơ từ trước đây chưa từng có đó là kết quả huấn luyện có thể tinh chỉnh được. Khi BERT được tinh chỉnh trong một nhiệm vụ nào đó, bộ Transformer tiền huấn luyện sẽ hoạt động như một bộ mã hóa và một bộ phân loại được khởi tạo ngẫu nhiên được thêm vào trên cùng. Trong trường hợp NER, trình phân loại chỉ đơn giản là một phép chiếu từ kích thước các từ đến kích thước tập nhãn, toán tử Softmax tiếp theo thực hiện chuyển điểm số thành xác suất của nhãn.

2.3. Mô hình gán nhãn chuỗi

2.3.1. Softmax

Softmax là một hàm kích hoạt thường được sử dụng trong các mô hình phân loại đa lớp để chuyển đổi đầu ra của mạng thành một phân phối xác suất. Softmax thường được áp dụng cho lớp đầu ra cuối cùng của mô hình để tính toán xác suất dự đoán cho mỗi lớp.

Hàm softmax là một hàm liên tục và khả vi, điều này rất hữu ích trong việc tính toán đạo hàm để cập nhật các trọng số trong quá trình huấn luyện mạng nơ-ron. Việc sử dụng hàm softmax không chỉ hữu ích trong các tác vụ phân loại đa lớp, mà còn có thể được áp dụng trong các bài toán khác như xác định mức độ tin cậy của dự đoán hoặc tạo ra một phân phối xác suất từ các giá trị đầu vào.

Tuy nhiên, hàm softmax cũng có một số hạn chế. Khi số lượng lớp rất lớn, việc tính toán và xử lý đồng thời các giá trị mũ có thể trở nên phức tạp và tốn nhiều thời gian tính toán. Đồng thời, hàm softmax không kháng nhiễu, có nghĩa là nếu có sự biến động mạnh trong giá trị đầu vào, các giá trị xác suất đầu ra có thể dễ dàng bị lệch và dẫn đến sai lệch trong dự đoán.

2.3.2. CRF

Conditional Random Fields (CRF) được đề xuất bởi Lafferty và đồng nghiệp vào năm 2001. Đây là một mô hình đồ thị xác suất vô hướng, kết hợp các đặc điểm của mô hình Markov ẩn và mô hình entropy tối đa. CRF là một trường hợp đặc biệt của mô hình Markov ngẫu nhiên, giải quyết vấn đề thiên vị nhãn do mô hình Markov ẩn gây ra. Ngoài ra, đặc điểm ngữ cảnh có thể được xem xét để lựa chọn đặc trưng tốt hơn. CRF được sử dụng để tính toán mật độ phân phối xác suất điều kiện của một tập hợp biến ngẫu nhiên đầu ra khác dựa trên một tập hợp biến ngẫu nhiên đầu vào.

Mục tiêu của việc huấn luyện một CRF là học các tham số của các hàm đặc trưng sao cho tối đa hóa hàm log-likelihood của dữ liệu huấn luyện. Điều này có thể được thực hiện bằng cách sử dụng ước lượng tối đa độ ảnh hưởng hoặc các phương pháp tối ưu hóa khác.

2.4. Học đa tác vụ

Con người có thể học nhiều nhiệm vụ cùng một lúc. Trong quá trình học tập, con người có thể sử dụng những kiến thức đã học trong một nhiệm vụ để học một nhiệm vụ khác. Lấy cảm hứng từ khả năng học tập của con người, học đa tác vụ có mục đích là cùng học nhiều nhiệm vụ liên quan để kiến thức chứa trong một nhiệm vụ có thể được tận dụng bởi các nhiệm vụ khác với hy vọng cải thiện hiệu suất tổng quát hóa của tất cả các nhiệm vụ [76].

Theo Zang và cộng sự, MTL được định nghĩa như sau: "Với *m* nhiệm vụ học $\{T_i\}_m^{i=1}$ trong đó tất cả các nhiệm vụ hoặc một tập hợp con của chúng có liên quan với nhau, học đa tác vụ nhằm mục đích học *m* nhiệm vụ cùng nhau để cải thiện việc học mô hình cho từng nhiệm vụ T_i bằng cách sử dụng kiến thức có trong tất cả hoặc một số nhiệm vụ." [77]. Trong học sâu thường sử dụng hai phương pháp là chia sẻ tham số cứng và chia sẻ tham số mềm [78].

Trong nhiều trường hợp, mô hình chỉ quan tâm tới hiệu suất của một tác vụ cụ thể, tuy nhiên để tận dụng được những lợi ích mà MTL mang lại, có thể thêm vào một số tác vụ liên quan với mục đích là cải thiện thêm hiệu suất trên tác vụ chính, các tác vụ này gọi là các tác vụ phụ trợ (Auxiliary task). Việc tìm kiếm một tác vụ phụ trợ phần lớn dựa trên giả định rằng tác vụ phụ trợ phải liên quan đến nhiệm vụ chính theo một cách nào đó và sẽ hữu ích cho việc dự đoán tác vụ chính.

Với giả thuyết rằng, mô hình khôi phục dấu câu, chữ hoa có thể cung cấp thêm các thông tin, hỗ trợ tốt hơn và giúp nâng cao hiệu quả nhận dạng thực thể định danh, luận án đã tận dụng tri thức về các phương pháp học tập đa tác vụ và tác vụ phụ trợ để đề xuất mô hình nhận dạng thực thể định danh cho văn bản đầu ra của ASR theo hướng E2E.

2.5. Kết luận Chương 2

Chương 2 đã trình bày những kiến thức nền tảng về các kĩ thuật biểu diễn từ như Word2Vec, GloVe, BERT. Mô tả chi tiết về đặc điểm, kiến trúc của một số mô hình xử lý chuỗi như Transformer, GRU. Đồng thời, các mô hình gán nhãn như softmax, CRF cũng được giới thiệu. Đặc biệt, phương pháp chia sẻ tham số cứng, chia sẻ tham số mềm và tác vụ phụ trợ trong học đa tác vụ cũng được trình bày. Những mô hình được giới thiệu trong chương này sẽ là cơ sở để hướng tới xây dựng mô hình cho bài toán chuẩn hoá và nhận dạng thực thể định danh cho văn bản đầu ra ASR tiếng Việt được trình bày ở Chương 3, Chương 4.

Chương 3. CHUẨN HÓA VĂN BẢN ĐẦU RA CỦA HỆ THỐNG NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT

Bên cạnh yêu cầu cải tiến hệ thống ASR để giảm thiểu lỗi từ thì chuẩn hóa văn bản đầu ra của hệ thống ASR bao gồm khôi phục dấu câu, chữ hoa cũng sẽ giúp văn bản dễ đọc, dễ hiểu và cung cấp các thông tin quan trọng cho nhiều ứng dụng. Trong phạm vi nghiên cứu luận án, nghiên cứu sinh cũng đặt giả thiết rằng việc kết hợp khôi phục, dấu câu chữ hoa sẽ hỗ trợ cho mô hình NER đạt hiệu suất cao hơn. Chương 3 này sẽ trình bày về bài toán khôi phục dấu câu, chữ hoa trong văn bản đầu ra tiếng nói tiếng Việt, những khó khăn, hạn chế khi thực hiện nhiệm vụ này và từ đó đề xuất giải pháp, cách thức xây dựng dữ liệu, thiết lập mô hình và các kết quả thực nghiệm.

3.1. Bài toán khôi phục dấu câu và chữ hoa

Đầu vào: văn bản đầu ra của hệ thống ASR tiếng Việt

Đầu ra: văn bản được khôi phục dấu câu, chữ hoa

Phạm vi nghiên cứu: Về dữ liệu: Từ các trang báo mạng chính thống của Việt Nam, với tỉ lệ lỗi từ trong văn bản là 0%. Về dấu câu: Tập trung khôi phục ba loại dấu câu là dấu chấm, dấu phẩy, dấu chấm hỏi. Về chữ hoa: Phân biệt 2 nhãn chính là chữ thường, chữ hoa, không xử lý các nhãn như chữ hoa trộn lẫn hay toàn bộ.

Hướng giải quyết: Đề xuất xử lý chuỗi đầu vào, đầu ra trong đó quan tâm tới ngữ cảnh của các từ xung quanh đoạn cắt. Đề xuất mô hình theo hướng học sâu để tăng hiệu suất khôi phục dấu câu, chữ hoa.

3.2. Đề xuất mô hình

Mô hình xử lý được tiến hành theo các bước sau:

(1) Bước một, văn bản đầu ra của ASR tiếng Việt sẽ được đưa qua mô-đun phân đoạn để cắt chuỗi đầu vào.

(2) Bước hai, mô hình khôi phục dấu câu, chữ hoa sẽ lấy các phân đoạn được cắt xử lý song song và tạo ra một danh sách nhãn dấu câu, chữ hoa đầu ra.

(3) Cuối cùng, sử dụng mô-đun hợp nhất các phân đoạn để trích xuất kết quả đầu ra được gán nhãn hợp nhất tương ứng với văn bản đầu vào.

3.2.1. Đề xuất xử lý cắt chuỗi văn bản đầu vào và hợp nhất chuỗi đầu ra

Nghiên cứu đã đề xuất một kỹ thuật mới nhằm xử lý cắt, ghép chuỗi bằng cách cắt có chồng lấn với ***ý tưởng chính là nhằm đảm bảo các đoạn cắt thu được có đủ ngữ cảnh của các từ để mô hình CaPu dự đoán tốt nhất***. Sau khi xử lý các đoạn cắt có chồng lấn, thực hiện hợp nhất các đoạn này thành chuỗi đầu ra của chuỗi ban đầu.

3.2.1.1. Phân đoạn chồng lấn

Hướng giải quyết được đề xuất là chia nhỏ chuỗi đầu vào thành các đoạn có kích thước cố định, với phần chồng lấn chiếm một nửa độ dài đoạn cắt.

Có thể mô tả hình thức cách phân đoạn chồng lấn như sau: Độ dài đoạn cắt được chọn là một số chẵn các từ. Gọi l là độ dài đoạn cắt, k là độ dài đoạn chồng lấn, khi đó ta có $l=2k$. Mỗi chuỗi từ đầu vào S chứa n từ kí hiệu là $w_1, w_2,..., w_n$ sẽ được cắt thành $\lceil n/l \rceil + \lceil (n-k)/l \rceil$ đoạn chồng lấn, trong đó, đoạn cắt thứ i là chuỗi con các từ $[w_{(i-1)k+1}, \ldots, w_{(i+1)k}]$. Trong nghiên cứu đã khảo sát các giá trị của l, k và bằng thực nghiệm đã lựa chọn các giá trị này cho phù hợp.

3.2.1.2. Hợp nhất đoạn chồng lấn

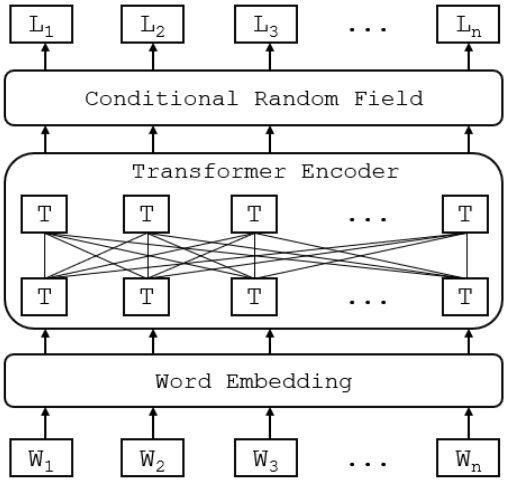
Vì câu đầu vào được phân chia thành các đoạn chồng lấn, do đó, với vấn đề hợp nhất các đoạn chồng lấn, cần phải xác định được những từ nào sẽ được bỏ đi và từ nào sẽ được giữ trong phần hợp nhất của câu cuối cùng.

Gọi c là độ dài đoạn sẽ giữ lại hay loại bỏ trong các đoạn chồng lấn. Để đơn giản cho tính toán, lấy $c = \lfloor k/2 \rfloor$. Theo quan sát, các từ cuối của đoạn chồng lấn thứ nhất và các từ đầu tiên trong đoạn chồng lấn thứ hai (các từ xung quanh đoạn cắt) sẽ không có nhiều ngữ cảnh. Do vậy, thuật toán sẽ loại bỏ đoạn c thuộc cuối đoạn chồng lấn (1) (phần gạch chéo) và giữ lại đoạn c ở đoạn chồng lấn (2) (phần chấm). Theo đó, các từ còn lại của đầu đoạn chồng lấn (1) được giữ lại và các từ còn lại ở đầu đoạn chồng lấn (2) sẽ bị loại bỏ. Điều này đảm bảo cho các từ ở phần chồng lấn được giữ lại luôn ở giữa các đoạn, sẽ có nhiều ngữ cảnh giúp cho việc khôi phục được chính xác hơn. Các đoạn loại bỏ và giữ lại của các phần chồng lấn sẽ được lặp lại cho các phân đoạn chồng lấn tiếp theo. Phần hợp nhất sau ghép nối được mô tả như sau.

$$[w_1, \ldots w_{2k-c}] + \sum_{i=2}^{n-1} [w_{(i-1)k+c}, \ldots w_{ik+c}] + [w_{n-2k+c}, \ldots w_n] \qquad (3.1)$$

3.2.2. Đề xuất mô hình học sâu cho mục đích khôi phục dấu câu, chữ hoa

Hình 3.1 giới thiệu mô hình CaPu đề xuất cho bài toán khôi phục dấu câu và chữ hoa cho văn bản đầu ra ASR tiếng Việt gồm các thành phần: bộ nhúng từ, Transformer Encoder và CRF.



Hình 3.1: Mô hình CaPu đề xuất cho văn bản đầu ra của ASR tiếng Việt

3.3. Xây dựng dữ liệu

Để có nguồn dữ liệu văn bản đầu ra của ASR tiếng Việt đủ lớn cho việc huấn luyện mô hình CaPu, bộ dữ liệu TextCaPu được thu thập từ các trang tin tức điện tử Việt Nam bao gồm vietnamnet.vn, dantri.com.vn, vnexpress.net.

Bộ dữ liệu $Text_{CaPu}$ được chia thành bộ huấn luyện $Text_{CaPu\text{-}train}$, bộ đánh giá $Text_{CaPu\text{-}vl}$ và bộ kiểm tra $Text_{CaPu\text{-}test}$. Với dữ liệu huấn luyện, bộ $Text_{CaPu\text{-}train}$ được chuyển về chữ thường và loại bỏ các dấu câu để mô phỏng giống với đầu ra của ASR, giữ nguyên dữ liệu kiểu số, ngày tháng và không có lỗi từ trong văn bản.

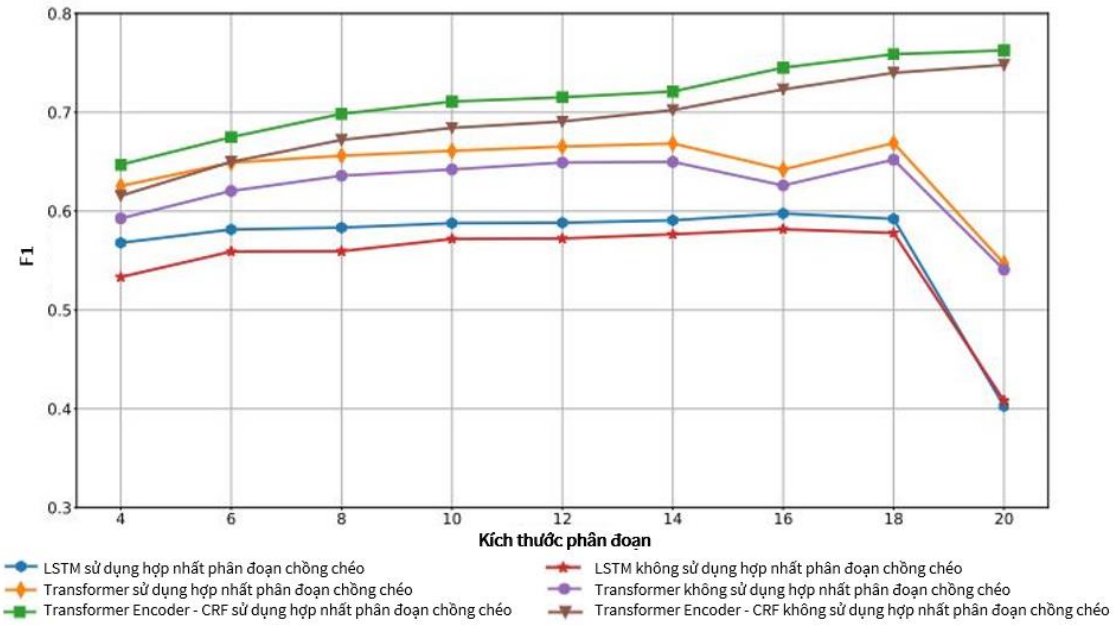
Bảng 3.1: Thông tin bộ dữ liệu

Nhãn	Bộ dữ liệu huấn luyện	Bộ dữ liệu kiểm tra
U	15.4M	74K
L	69.3M	507K
$	76.6M	525K
.	2.7M	24K
,	5.3M	30K
?	53K	2.6K

3.4. Kết quả thực nghiệm

3.4.1. Đánh giá về sử dụng hợp nhất đoạn chồng lấn

Hình 3.2 hiển thị biểu đồ so sánh với kết quả của các mô hình với các kích thước phân đoạn khác nhau, trong các trường hợp sử dụng hoặc không sử dụng hợp nhất đoạn chồng lấn. Các mô hình sử dụng hợp nhất đoạn chồng lấn luôn cho kết quả tốt hơn. Đặc biệt, ở mô hình đề xuất là Transformer Encoder – CRF, kết quả sử dụng hợp nhất có kết quả cao nhất là 0.88. Kết quả xác nhận giả thuyết của nghiên cứu rằng việc bổ sung thêm ngữ cảnh bằng cách cách xếp các đoạn chồng lấn và phân đoạn, hợp nhất các đoạn chồng lấn giúp cải thiện mô hình.



Hình 3.2: Các mô hình sử dụng và không sử dụng hợp nhất đoạn chồng lấn

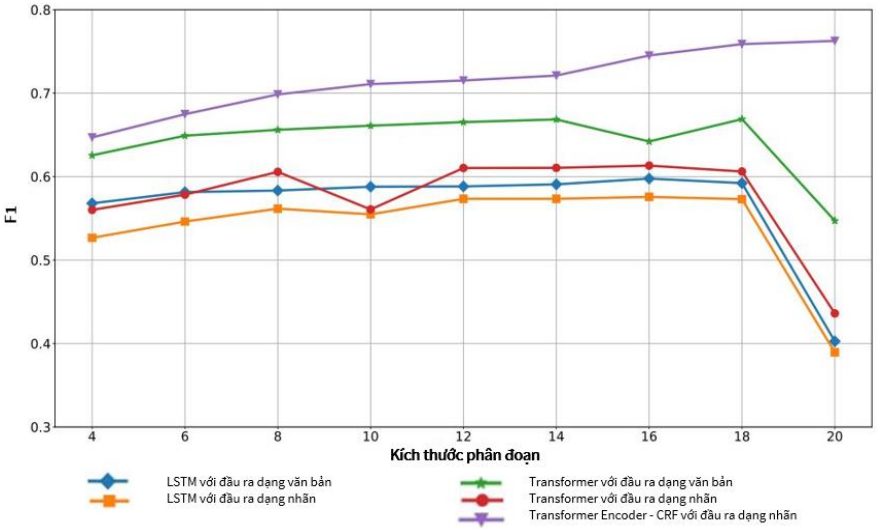
Nghiên cứu trình bày kết quả cho mô hình được đề xuất Transformer Encoder - CRF khi áp dụng hoặc không áp dụng hợp nhất đoạn chồng lấn và cũng chỉ thống kê trong các nhãn ('U' '.' ',' '?'), bỏ qua các nhãn ('L' '$'), vì số lượng chính xác nhiều, nên không cần thiết để so sánh hiệu quả. Bảng 3.2 cho thấy sự vượt trội của phương pháp hợp nhất đoạn chồng lấn so với không sử dụng khi điểm F1 trên tất cả các lớp được cải thiện đáng kể từ 0.01 đến 0.05. Kết quả cho thấy, các từ ở đoạn giữa phần xếp chồng lấn cung cấp nhiều thông tin dự đoán hơn và quá trình hợp nhất có thể chọn phần thích hợp của khu vực xếp chồng này.

Bảng 3.2: So sánh kết quả mô hình Transformer Encoder - CRF khi áp dụng và không áp dụng hợp nhất chồng lấn

Mô hình	**Nhãn**	**Precision**	**Recall**	**F1**
Transformer Encoder-CRF áp dụng hợp nhất chồng lấn	U	0.90	0.86	0.88
	.	0.71	0.57	0.63
	,	0.66	0.53	0.59
	?	0.75	0.52	0.62
Transformer Encoder-CRF không áp dụng hợp nhất chồng lấn	U	0.89	0.85	0.87
	.	0.69	0.54	0.61
	,	0.65	0.50	0.57
	?	0.74	0.47	0.58

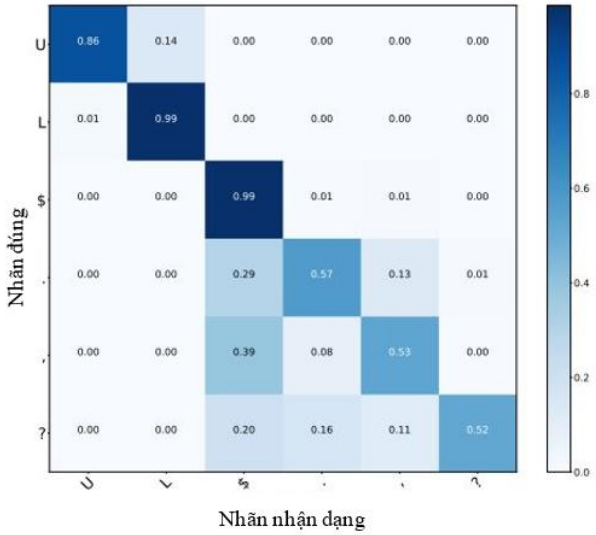
3.4.2. Đánh giá đầu ra văn bản mã hóa và văn bản thô

Kết quả cho các mô hình sử dụng đầu ra gán nhãn và văn bản thông thường được so sánh trong hình 3.3, trong đó, mô hình LSTM và mô hình Transformer với văn bản thông thường có kết quả tốt hơn so với sử dụng đầu ra gán nhãn và mô hình đề xuất cho kết quả tốt nhất.



Hình 3.3: Kết quả các mô hình với đầu ra dạng văn bản hoặc dạng nhãn

Đồng thời, ma trận lỗi trong hình 3.4 cũng cho thấy phần trăm dự đoán đúng/sai lệch các nhãn dấu câu, chữ hoa cho mô hình đề xuất Transformer Encoder - CRF. Khả năng khôi phục đúng chữ thường, chữ hoa và không dấu rất cao (0.86-0.99), sau đó giảm dần với các dấu chấm, dấu phẩy và dấu hỏi chấm.



Hình 3.4: Ma trận lỗi cho mô hình Transformer Encoder – CRF

3.4.3. Đánh giá tốc độ

Kết quả so sánh thời gian thực thi của 3 mô hình có văn bản đầu ra được mã hóa và văn bản thuần túy hiển thị trong Bảng 3.3 với 2080 ti (GPU), batch_size: 128. Với đầu ra văn bản mã hóa, các mô hình có thời gian xử lý nhanh hơn văn bản thuần túy. Đầu ra văn bản mã hóa thậm chí còn cho thấy hiệu suất vượt trội khi nó được sử dụng với mô hình được đề xuất.

Bảng 3.3: Đánh giá tốc độ (tokens/second)

Đầu ra	**Transformer**	**LSTM**	**Transformer Encoder -CRF**
Dạng gán nhãn	263*s* → 2209*t/s*	217*s* → 2678*t/s*	90*s* → 6457*t/s*
Dạng văn bản	355*s* → 1637*t/s*	230*s* → 2526*t/s*	-

Chương 4. NHẬN DẠNG THỰC THỂ ĐỊNH DANH CHO VĂN BẢN ĐẦU RA NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT

Chương 4 trình bày chi tiết về bài toán NER và đề xuất mô hình, xây dựng dữ liệu, đưa ra kết quả thực nghiệm nhằm đánh giá, so sánh các giải pháp cho NER của văn bản đầu ra của ASR tiếng Việt theo cách tiếp cận Pipeline truyền thống và cách tiếp cận E2E. Cách tiếp cận truyền thống Pipeline dựa trên giả thiết rằng việc kết hợp một mô hình khôi phục dấu câu và chữ hoa như mô hình CaPu sẽ cung cấp thông tin hữu ích làm đầu vào giúp mô hình NER đạt hiệu suất cao hơn. Cách tiếp cận E2E là một quy trình phức hợp từ đầu đến cuối, giúp hệ thống hoạt động thuận tiện hơn, tránh được những lỗi lan truyền qua các bước giải các bài toán trung gian. Giải pháp E2E cho bài toán NER đề xuất mô hình giải quyết đồng thời cả hai bài toán khôi phục dấu câu, chữ hoa và nhận dạng thực thể định danh.

4.1. Bài toán nhận dạng thực thể định danh cho hệ thống nhận dạng tiếng nói tiếng Việt

Đầu vào: Văn bản đầu ra của ASR tiếng Việt

Đầu ra: Gán nhãn thực thể định danh theo hướng tiếp cận Pipeline và E2E

Phạm vi nghiên cứu: Về dữ liệu: Văn bản dài, từ vựng lớn. Hệ thống ASR phục vụ đánh giá có WER là 4.85%. Về thực thể định danh: Nhận dạng ba loại thực thể chính là tên người, tên tổ chức và tên địa điểm.

Hướng nghiên cứu: Xây dựng bộ dữ liệu phù hợp cho mục đích huấn luyện và đánh giá mô hình. Đối với cách tiếp cận Pipeline, nghiên cứu đề xuất kết hợp mô hình CaPu vào hệ thống với mục đích nâng cao hiệu suất mô hình NER. Cách tiếp cận E2E, sử dụng tiền huấn luyện mô-đun CaPu cho mô hình. Đề xuất mô hình học sâu cho mô hình NER.

4.2. Xây dựng dữ liệu

4.2.1. Bộ dữ liệu huấn luyện

Bộ dữ liệu thứ nhất, $Text_{CaPu}$, là một bộ dữ liệu lớn bao gồm các văn bản tin tức được lấy từ các trang báo điện tử của Việt Nam. Tập văn bản này được xoá định dạng (bỏ dấu câu, chuyển chữ hoa thành chữ thường) và gán nhãn dấu câu, chữ hoa phục vụ cho mục đích huấn luyện mô hình chuẩn hoá văn bản đầu ra của hệ thống ASR; Bộ dữ liệu thứ hai, $Text_{ViBERT}$, là bộ dữ liệu huấn luyện mô hình ViBERT thu thập từ nhiều miền trên Internet bao gồm tin tức, luật, giải trí, Wikipedia,... ; Bộ dữ liệu thứ ba, $Text_{VLSP}$, là bộ dữ liệu văn bản đã gán nhãn NER của VLSP 2018. Tập văn bản chuẩn này được sử dụng để huấn luyện mô hình NER theo cách tiếp cận Pipeline; Bộ dữ liệu thứ tư, $Text_{VLSP\text{-}TTS\text{-}ASR}$, là bộ dữ liệu để huấn luyện mô hình NER theo tiếp cận E2E. Đầu tiên, dữ liệu tiếng nói được tổng hợp từ văn bản huấn luyện của bộ dữ liệu NER VLSP 2018 sử dụng hệ thống TTS của Google. Sau đó dữ liệu tiếng nói này được đưa qua hệ thống ASR của VAIS để thu được văn bản đầu ra ASR.

4.2.2. Bộ dữ liệu kiểm tra

Cả hai cách tiếp cận Pipeline và E2E đều sử dụng một bộ dữ liệu thu âm bởi bốn giọng đọc trong môi trường khác nhau từ bộ dữ liệu kiểm tra NER của VLSP 2018 với 26 giờ âm thanh. Sau đó, bộ dữ liệu âm thanh này được đưa qua hệ thống ASR của VAIS (với WER bằng 4.85%) để nhận được bộ dữ liệu văn bản đầu ra của ASR, $Text_{VLSP\text{-}Audio\text{-}ASR}$ để phục vụ cho mục đích đánh giá các mô hình đề xuất. Đồng thời, bộ dữ liệu kiểm tra VLSP chuẩn $Text_{VLSP\text{-}test}$ hay bộ dữ liệu VLSP được xoá định dạng $Text_{VLSP\text{-}UnCaPu}$, cũng được sử dụng để đánh giá và so sánh mô hình trong các điều kiện đầu vào khác nhau.

4.3. Nhận dạng thực thể định danh theo hướng tiếp cận Pipeline

4.3.1. Đề xuất mô hình

4.3.1.1. Mô hình tổng quát

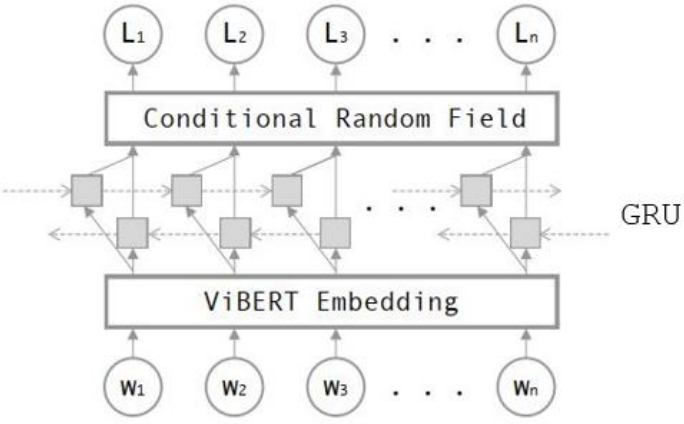
Đề xuất kiến trúc tổng quát hệ thống NER hướng Pipeline: (1) Hệ thống ASR sẽ chuyển tín hiệu tiếng nói sang dạng văn bản. (2) Tiếp theo, qua mô hình CaPu, văn bản đầu ra của ASR sẽ được khôi phục dấu câu, chữ hoa. (3) Cuối cùng, từ mô hình CaPu, thông tin của các thực thể được gán nhãn bằng cách sử dụng mô hình NER.

4.3.1.2. Mô hình khôi phục dấu câu, chữ hoa

Nghiên cứu cũng đặt giả thiết mô hình CaPu sẽ hỗ trợ tăng hiệu suất mô hình NER tiếng Việt. Mô hình đề xuất và các kết quả thực nghiệm đã được trình bày chi tiết trong Chương 3 của luận án, đồng thời được công bố trong các công trình (CT2), (CT3), (CT5) của nghiên cứu sinh và các cộng sự.

4.3.1.3. Mô hình học sâu cho nhận dạng thực thể định danh

Nghiên cứu đã đề xuất sử dụng kiến trúc RoBERTa [31] và huấn luyện trên kho ngữ liệu tiếng Việt để tạo ra một mô hình ngôn ngữ được huấn luyện trước. Do giới hạn về năng lực tính toán, mô hình huấn luyện đã giảm số lượng lớp ẩn, số đỉnh chú ý và kích thước từ nhúng từ mô hình kiến trúc cơ sở RoBERTa và được đặt tên là ViBERT. Hình 4.1 mô tả thiết kế mô hình NER, trong đó, ViBERT được sử dụng để nhúng câu đầu vào, mô hình GRU hai chiều và lớp CRF được gắn vào đầu ViBERT để phân loại nhãn thực thể của mỗi từ đầu vào.



Hình 4.1: Đề xuất mô hình NER

4.3.2. Kết quả đánh giá

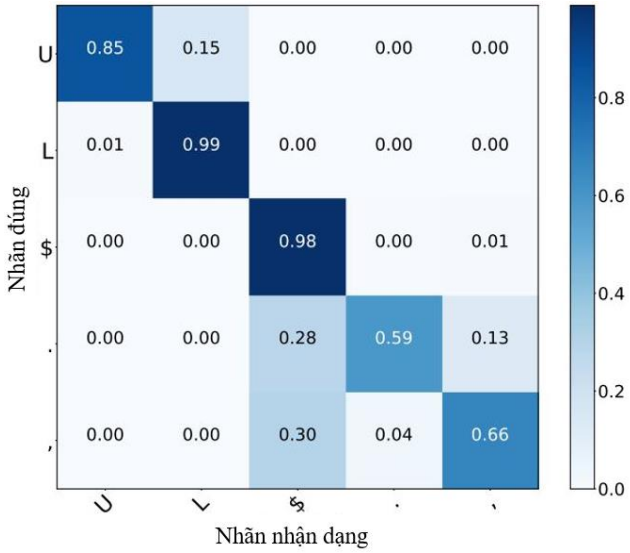
Trong mô hình NER, nghiên cứu kết hợp ViBERT với lớp GRU và lớp CRF cho thấy hiệu quả khi tạo ra kết quả F1 là 0.9018, cao hơn đáng kể khi so sánh với kết quả đã công bố trước đó (bảng 4.1). Đây là kết quả được đánh giá trực tiếp bằng cách sử dụng bộ dữ liệu $Text_{VLSP\text{-}test}$ của NER VLSP 2018.

Bảng 4.1: Đánh giá các mô hình NER dựa trên bộ dữ liệu NER VLSP 2018

Mô hình	**F1**
Vi Tokenizer + Bidirectional Inference [76]	0.8878
VNER [77]	0.7752
Multi layers LSTM [76]	0.8380
CRF/MEM + BS [76]	0.8408
ViBERT+GRU+CRF (mô hình đề xuất)	**0.9018**

Với tỷ lệ lỗi từ của hệ thống ASR là 4.85%, bảng 4.2 cho thấy rằng nếu văn bản đầu ra của ASR được đưa trực tiếp vào mô hình NER, hiệu quả nhận dạng thực thể sẽ giảm từ 0.9018 xuống 0.6389. Tầm quan trọng của chữ hoa và dấu câu cũng được quan sát thấy trong thử nghiệm chạy mô hình NER trên văn bản bỏ dấu dấu câu và chữ hoa, điểm F1 giảm từ 0.9018 xuống 0.7535.

Hình 4.2 chứng minh kết quả của mô hình CaPu trên văn bản chuẩn bỏ dấu câu và chữ hoa. Độ chính xác của khôi phục ký tự viết hoa là 0.85. Việc khôi phục dấu câu sẽ khó hơn, độ chính xác luôn duy trì ở mức gần 0.60 đối với dấu chấm (‘.’) và 0.66 đối với dấu phẩy (‘,’).



Hình 4.2: Đánh giá mô hình CaPu trên văn bản chuẩn bỏ dấu câu và chữ hoa

Bảng 4.2 cũng chứng tỏ hiệu quả của mô hình CaPu trong việc cải thiện độ chính xác của mô hình NER. Điểm F1 của mô hình NER tăng từ 0.6319 lên 0.6713 khi áp dụng mô hình này trên văn bản đầu ra của ASR và cải thiện hơn 0.06 điểm F1 của mô hình NER khi áp dụng cho văn bản bỏ dấu câu và chữ hoa.

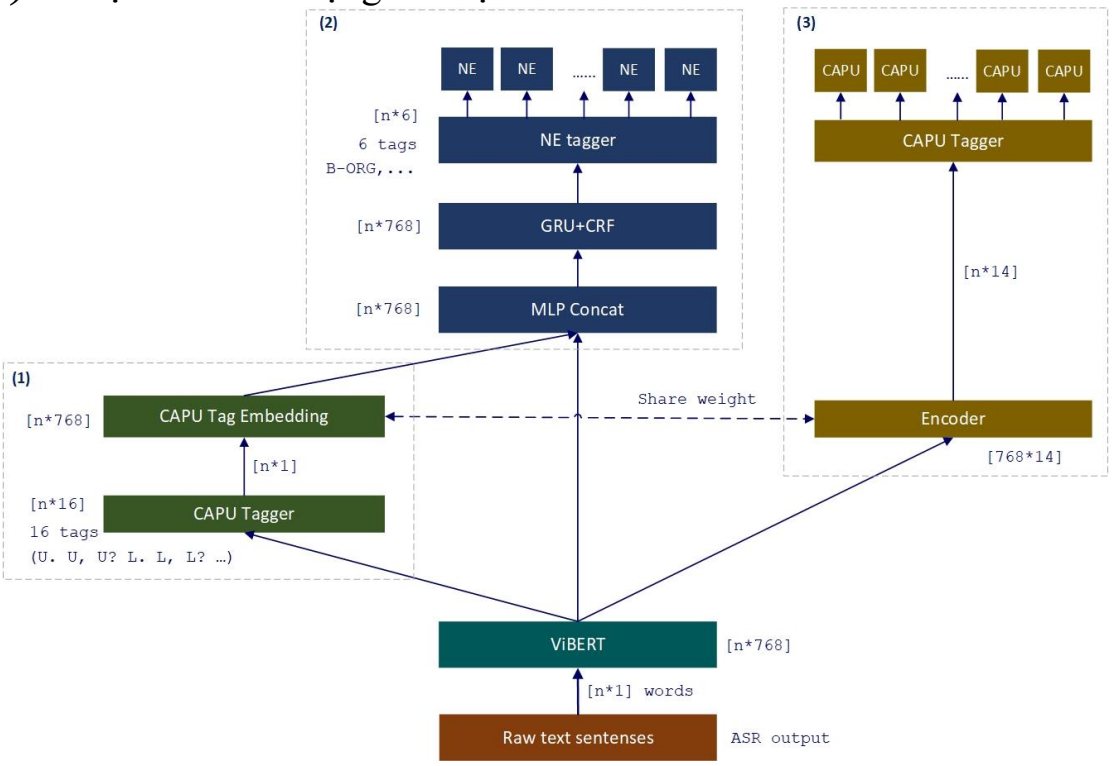
Bảng 4.2: Đánh giá mô hình NER đề xuất theo cách tiếp cận Pipeline với các kiểu văn bản đầu vào khác nhau

Kiểu đầu vào	**F1**
Văn bản chuẩn (*Text$_{VLSP\text{-}test}$*)	0.9018
Văn bản đầu ra của ASR (*Text$_{VLSP\text{-}Audio\text{-}ASR}$*)	0.6319
Văn bản đầu ra của ASR + CaPu (*Text$_{VLSP\text{-}Audio\text{-}ASR}$* + CaPu)	**0.6713**
Văn bản chuẩn bỏ dấu câu, chữ hoa (*Text$_{VLSP\text{-}UnCaPu}$*)	0.7535
Văn bản chuẩn bỏ dấu câu, chữ hoa+CaPu (*Text$_{VSP\text{-}UnCaPu}$* +CaPu)	**0.8141**

4.4. Nhận dạng thực thể định danh theo hướng End-to-End

4.4.1. Đề xuất mô hình

Hình 4.3 biểu diễn mô hình E2E đề xuất, bao gồm luồng NER chính kết hợp với một luồng nhận dạng dấu câu, chữ hoa có vai trò bổ sung thông tin. Dữ liệu đưa vào mô hình là văn bản đầu ra của ASR tiếng Việt không dấu câu, không chữ hoa có độ dài *N*. Câu đầu vào được đưa qua bộ biểu diễn ngôn ngữ tiếng Việt ViBERT. Ở nghiên cứu này, tiếp cận học chuyển giao được áp dụng với mô hình ViBERT là mô hình đã được tiền huấn luyện và được giữ nguyên trong mô hình E2E được đề xuất ở đây. Đầu ra của ViBERT là một ma trận có kích thước *(Nx768)* là một biểu diễn dạng ma trận của câu đầu vào.



Hình 4.3: Đề xuất kiến trúc NER theo tiếp cận E2E

4.4.2. Kết quả thực nghiệm

Kết quả bảng 4.3 cho thấy, nếu văn bản đầu ra của ASR được đưa trực tiếp vào mô hình NER, kết quả nhận dạng thực thể giảm từ 0.9018 xuống 0.6319.

Bảng 4.3: Đánh giá mô hình NER đề xuất theo cách tiếp cận E2E với các kiểu văn bản đầu vào khác nhau

Các kiểu dữ liệu đầu vào	**F1**
Văn bản chuẩn (*Text*$_{VLSP\text{-}test}$)	0.9018
Văn bản đầu ra của ASR (*Text*$_{VLSP\text{-}Audio\text{-}ASR}$)	0.6319
Văn bản đầu ra của ASR+CaPu E2E (*Text*$_{VLSP\text{-}Audio\text{-}ASR}$ +CaPu E2E)	**0.6780**
Văn bản chuẩn bỏ dấu câu, chữ hoa+CaPu E2E (*Text*$_{VLSP\text{-}UnCaPu}$ +CaPu E2E)	**0.8178**

Tầm quan trọng của chữ hoa và dấu câu cũng được quan sát thấy trong thử nghiệm chạy mô hình NER. Bảng kết quả chứng tỏ hiệu quả của mô hình CaPu giúp cải thiện độ chính xác của mô hình NER trên văn bản đầu ra của ASR, điểm F1 của mô hình NER tăng xấp xỉ 0.05 từ 0.6319 lên 0.6780. Mô hình này cũng giúp cải thiện rõ rệt 0.1398 điểm F1 khi áp dụng cho văn bản chuẩn bỏ dấu câu, chữ hoa so với văn bản đầu ra của ASR.

Bảng 4.4 cho thấy kết quả F1 khi so sánh mô hình E2E và mô hình Pipeline với các bộ dữ liệu kiểm tra khác nhau. Mô hình E2E đề xuất kết quả tốt hơn nhưng không đáng kể so với với Pipeline (0.0067 với văn bản đầu ra của ASR và 0.0037 đối với văn bản chuẩn bỏ dấu câu, chữ hoa).

Bảng 4.4: So sánh mô hình E2E với mô hình Pipeline

Hệ thống NER	**F1**
Văn bản đầu ra của ASR + CaPu Pipeline (*Text*$_{VLSP\text{-}Audio\text{-}ASR}$ +CaPu Pipeline)	0.6713
Văn bản đầu ra của ASR + CaPu E2E (*Text*$_{VLSP\text{-}Audio\text{-}ASR}$ +CaPu E2E)	**0.6780**
Văn bản chuẩn bỏ dấu câu, chữ hoa + CaPu Pipeline (*Text*$_{VLSP\text{-}UnCaPu}$ +CaPu Pipeline)	0.8141
Văn bản chuẩn bỏ dấu câu, chữ hoa + CaPu E2E (*Text*$_{VLSP\text{-}UnCaPu}$ +CaPu E2E)	**0.8178**

Mặc dù với kết quả chưa cải thiện tốt hơn nhiều, nhưng với mô hình Pipeline, quá trình huấn luyện các thành phần riêng biệt, đòi hỏi các thuật toán huấn luyện riêng và hàm mất mát riêng với ứng với mỗi thành phần, số lượng lớn siêu tham số dẫn đến độ phức tạp cao trong thiết kế và huấn luyện. Các sai số phát sinh trong mỗi thành phần không được tính toán khi kết hợp với các thành phần khác nên sai số tích lũy lớn. Ngược lại, với mô hình E2E, tất cả các tham số của mô hình được huấn luyện đồng thời với chỉ một hàm mất mát. Toàn bộ luồng đồ thị tính toán được tối ưu đồng thời bởi thuật toán lan truyền ngược. Các sai số phát sinh giữa các thành phần đều được tính toán do đó giảm thiểu sai số chung. Quá trình suy diễn cũng đơn giản và nhanh hơn khi không có những bước chuyển trung gian giữa các mô hình thành phần. Chính vì vậy, mô hình E2E vẫn có những lợi thế nhất định và việc tiếp tục cải tiến mô hình E2E cho bài toán NER tiếng nói tiếng Việt là cần thiết để đạt được hiệu suất cao hơn và tận dụng được tính ưu việt trong huấn luyện mô hình và trong triển khai ứng dụng vào thực tế.

4.5. Kết luận Chương 4

Chương 4 đã đề xuất mô hình NER cho hệ thống ASR tiếng Việt theo hướng tiếp cận đường ống và E2E. Thực nghiệm đã chứng minh hiệu quả của việc kết hợp mô hình CaPu giúp tăng hiệu suất mô hình NER. Luận án đã giới thiệu bộ dữ liệu đầu tiên cho nghiên cứu NER cho văn bản đầu ra của ASR tiếng Việt. Mô hình E2E kết quả tốt hơn nhưng chưa đáng kể so với mô hình đường ống. Việc kết hợp mô hình học tập đa tác vụ với mô hình khôi phục dấu chấm câu và chữ hoa đã tăng điểm F1 lên xấp xỉ 0.05 và cải thiện rõ rệt 0.14 điểm F1 của mô hình NER khi áp dụng cho văn bản chuẩn bỏ chữ hoa, dấu câu.

KẾT LUẬN

Các kết quả chính của luận án

(1) Xây dựng các bộ dữ liệu ban đầu phục vụ cho thực nghiệm các mô hình chuẩn hoá và nhận dạng thực thể định danh cho văn bản đầu ra của hệ thống ASR tiếng Việt.

(2) Thiết kế mô hình Transformer Encoder – CRF cho bài toán khôi phục viết hoa và dấu câu cho văn bản đầu ra của ASR tiếng Việt. Luận án đề xuất cách phân chia đoạn mới cho câu đầu vào sử dụng phân đoạn, hợp nhất các đoạn chồng lấn, giúp các từ xung quanh đoạn cắt có nhiều ngữ cảnh để nhận dạng được chính xác hơn. Đầu ra của mô hình là văn bản tiếng Việt có đầy đủ dấu câu, chữ hoa, giúp tăng độ chính xác của quá trình nhận dạng thực thể định danh ở bước tiếp theo.

(3) Đề xuất mô hình biểu diễn ngôn ngữ tiền huấn luyện cho văn bản tiếng Việt với tên gọi ViBERT dựa theo kiến trúc RoBERTa. Mô hình được huấn luyện dựa trên tập dữ liệu lớn văn bản tiếng Việt chính thống để biểu diễn ngôn ngữ tiếng Việt trong không gian véc-tơ giúp tăng hiệu quả áp dụng các thuật toán học sâu trong XLNNTN tiếng Việt. Mô hình được áp dụng vào các mô-đun biểu diễn véc-tơ từ cho các mô hình NER tiếp theo.

(4) Xây dựng mô hình đường ống cho bài toán NER tiếng nói tiếng Việt. Nghiên cứu cho thấy tác động hiệu quả của mô hình biểu diễn ngôn ngữ được tiền huấn luyện ViBERT để áp dụng cho nhiệm vụ NER trên văn bản đầu ra của ASR tiếng Việt và đã đạt được kết quả khả quan. Đồng thời nghiên cứu cũng chứng tỏ được tầm quan trọng của việc kết hợp mô hình CaPu vào chuẩn hóa văn bản đầu vào cho mô hình NER giúp cải thiện đáng kể hiệu suất của mô hình.

(5) Thiết kế mô hình E2E giải quyết bài toán NER cho tiếng nói tiếng Việt cùng với các đề xuất mới như kỹ thuật chia sẻ tham số, kỹ thuật huấn luyện đa tác vụ. Bên cạnh thực nghiệm cho thấy đạt hiệu suất tương đương mô hình đường ống, mô hình E2E còn cho thấy ưu thế của việc tích hợp hệ thống trên một mô hình duy nhất giúp thuận lợi cho quá trình huấn luyện, giảm thiểu sai số

phát sinh giữa các thành phần, tăng tốc độ thực thi, tăng khả năng triển khai trong thực tiễn.

Hướng phát triển nghiên cứu

(1) Nghiên cứu giải pháp giảm thiểu sự ảnh hưởng của lỗi dữ liệu trong văn bản đầu ra của ASR, đồng thời, bổ sung bộ dữ liệu từ điển NER chuẩn mực phục vụ cho mục đích huấn luyện nhằm nâng cao chất lượng mô hình NER tiếng Việt.

(2) Thực nghiệm NER cho khôi phục chữ hoa, giúp hệ thống E2E ASR được cải thiện hơn.

(3) Thực nghiệm các mô hình đề xuất trong nghiên cứu này với các bộ dữ liệu tiếng Anh, Trung Quốc,... đã công bố để đối sánh tính hiệu quả của mô hình.

(4) Áp dụng mô hình đề xuất để nhận dạng thực thể định danh cho văn bản thuộc các lĩnh vực chuyên biệt nhằm làm rõ tính khả thi của mô hình.

(5) Tiếp tục cải tiến mô hình E2E và các thuật toán huấn luyện tương ứng để đạt hiệu suất tốt hơn cho bài toán NER tiếng nói tiếng Việt.

DANH MỤC CÔNG TRÌNH CỦA TÁC GIẢ

[CT1]. Nguyen Thi Minh Huyen, Ngo The Quyen, Vu Xuan Luong, Tran Mai Vu, **Nguyen Thi Thu Hien** (2018), *"VLSP shared task: Named Entity Recognition"*, Journal of Computer Science and Cybernetics, V.34, N.4, p.283–294.

[CT2]. **Thu Hien Nguyen**, Binh Nguyen Thai, Hung Nguyen Vu Bao, Truong Do Quoc, Mai Luong Chi, Huyen Nguyen Thi Minh (2019), *"Recovering Capitalization for Automatic Speech Recognition of Vietnamese using Transformer and Chunk Merging"*, Proceedings of 2019 the 11th International conference on Knowledge and Systems Engineering (KSE), p.430-434.

[CT3]. Binh Nguyen, Vu Bao Hung Nguyen, **Hien Nguyen**, Pham Ngoc Phuong, The-Loc Nguyen, Quoc Truong Do, Luong Chi Mai (2019), *"Fast and Accurate Capitalization and Punctuation for Automatic Speech Recognition Using Transformer and Chunk Merging"*, Proceeding in the Oriental COCOSDA, Cebu, Philippines, p. 1-5, doi: 10.1109/O-COCOSDA46868.2019.9041202.

[CT4]. Nguyen, T. Binh, Nguyen, Q. Minh, Nguyen, **Nguyen, T.Hien**, Do, Q. Truong, & Luong, C. Mai (2020) *"Improving Vietnamese Named Entity Recognition from Speech Using Word Capitalization and Punctuation Recovery Models"*, Proceeding in the INTERSPEECH 2020, 4263-4267, Shanghai, China.

[CT5]. **Thu Hien Nguyen,** Thai Binh Nguyen, Ngoc Phuong Pham, Quoc Truong Do, Tu Luc Le, Cshi Mai Luong (2021), *"Toward Human-Friendly ASR Systems: Recovering Capitalization and Punctuation for Vietnamese Text"*, IEICE TRANSACTIONS on Information and Systems, Vol.E104-D, No.8, p.1195-1203 (SCIE, Q3).

[CT6]. **Thu Hien Nguyen**, Thai-Binh Nguyen, Quoc-Truong Do, Tuan-Linh Nguyen (2022), *End-to-end named entity recognition for vietnamese speech*, Proceeding in the 25th Oriental COCOSDA, pp.193-197, 979-8-3503-9855-7.