TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC ĐÀ NẴNG

KHOA THƯƠNG MẠI ĐIỆN TỬ

LUẬN VĂN TỐT NGHIỆP

TÊN ĐỀ TÀI:

NHẬN DẠNG GIỌNG NÓI VÀ ỨNG DỤNG MÔ HÌNH NGÔN NGỮ LỚN TRONG DỊCH TỰ ĐỘNG NGÔN NGỮ KÝ HIỆU VIỆT NAM NHẰM HỖ TRỢ GIAO TIẾP CHO NGƯỜI KHIẾM THÍNH

GVHD	: Th.S Nguyễn Văn Chức
Họ tên sinh viên	: Lê Thị Hà Vy
Mã số sinh viên	: 221124029153
Lớp	: 48K29.1
Khoa	: Thương mại điện tử
Chuyên ngành	: Khoa học dữ liệu

Đà Nẵng, tháng 11, năm 2025

TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC ĐÀ NẴNG

KHOA THƯƠNG MẠI ĐIỆN TỬ

LUẬN VĂN TỐT NGHIỆP

TÊN ĐỀ TÀI:

NHẬN DẠNG GIỌNG NÓI VÀ ỨNG DỤNG MÔ HÌNH NGÔN NGỮ LỚN TRONG DỊCH TỰ ĐỘNG NGÔN NGỮ KÝ HIỆU VIỆT NAM NHẰM HỖ TRỢ GIAO TIẾP CHO NGƯỜI KHIẾM THÍNH

GVHD	: ThS Nguyễn Văn Chức
Họ tên sinh viên	: Lê Thị Hà Vy
Mã số sinh viên	: 221124029153
Lớp	: 48K29.1
Khoa	: Thương mại điện tử
Chuyên ngành	: Khoa học dữ liệu

Đà Nẵng, tháng 11, năm 2025

TÓM TẮT

Theo thống kê của Tổng cục thống kê và báo cáo liên quan, Việt Nam hiện có một cộng đồng khoảng 2.5 triệu người điếc và nghe kém (khiếm thính). Trong những năm gần đây, với nỗ lực của toàn xã hội, ngày càng có nhiều người khiếm thính tham gia vào thị trường lao động, đặc biệt trong ngành dịch vụ, bán lẻ và F&B. Điển hình là các mô hình nhà hàng do người khiếm thính phục vụ ở Hà Nội (Vnexpress, 2025) hay quán Angel Coffee tại Đà Nẵng – nơi hỗ trợ việc làm cho thanh niên khiếm thính (Đà Nẵng, 2023).

Sự hòa nhập là tín hiệu xã hội tích cực, tuy nhiên, nó cũng đặt ra thách thức lớn về rào cản giao tiếp nhân viên khiếm thính và khách hàng không sử dụng ngôn ngữ ký hiệu. Rào cản này không chỉ gây bất tiện lớn trong các hoạt động, mà đôi khi còn ảnh hưởng trải nghiệm khách hàng.

Bên cạnh đó, các hoạt động giáo dục (khóa học online,…) và truyền thông xã hội (tin tức thời tiết,…) chủ yếu dựa trên lời nói cũng tạo ra sự bất bình đẳng trong việc tiếp cận tri thức với cộng đồng người khiếm thính.

Nhằm giải quyết vấn đề trên, bài nghiên cứu này đề xuất một giải pháp nhằm hỗ trợ giao tiếp tự động. Quy trình thực hiện qua hai giai đoạn chính: Đầu tiên, xây dựng bộ dữ liệu về văn bản tiếng Việt – văn bản đúng cú pháp ngôn ngữ ký hiệu về lĩnh vực quán cà phê thông qua phương pháp tăng cường dữ liệu để khắc phục tình trạng khan hiếm. Thứ hai, phát triển hệ thống gồm 2 module***: (1) Module nhận dạng giọng nói tự động*** sử dụng các mô hình tiên tiến như PhoWhisper để chuyển đổi lời nói thành văn bản và ***(2) module dịch máy chuyển đổi văn bản tiếng Việt sang văn bản đúng cú pháp ngôn ngữ ký hiệu Việt Nam***.

LỜI CAM ĐOAN

Em xin cam kết luận văn tốt nghiệp với đề tài "***Nhận dạng giọng nói và ứng dụng ngôn ngữ lớn trong dịch tự động ngôn ngữ ký hiệu Việt Nam hỗ trợ giao tiếp người khiếm thính***" cùng với các dữ liệu thu thập được và kết quả nghiên cứu mà em thực hiện là công trình nghiên cứu của riêng em, được tiến hành một cách công khai với sự hướng dẫn, tận tình chỉ bảo từ giảng viên Th.S Nguyễn Văn Chức.

Tất cả số liệu, thông tin trong bài nghiên cứu là trung thực. Các tài liệu tham khảo được sử dụng đã được trích dẫn đầy đủ và tuân thủ đúng quy định về học thuật. Nếu phát hiện bất kỳ sự sao chép, gian dối nào trong kết quả nghiên cứu, em xin hoàn toàn chịu trách nhiệm và chấp nhận quyết định kỷ luật của khoa và trường.

Em xin chịu hoàn toàn trách nhiệm về nội dung lời cam đoan này.

Đà Nẵng, tháng 11 năm 2025

Người thực hiện luận văn

(Ký và ghi rõ họ tên)

Lê Thị Hà Vy

LỜI CẢM ƠN

Lời đầu tiên, em xin cảm ơn sâu sắc đến thầy Nguyễn Văn Chức, người đã vui vẻ tận tình hướng dẫn, chỉ bảo, và truyền đạt kiến thức chuyên môn quý báu để hướng dẫn em trong suốt quá trình thực hiện luận văn tốt nghiệp này. Những buổi trao đổi, đề xuất, cải thiện từ gợi ý của thầy đã hướng dẫn em dần luận văn này. Em cảm ơn thầy vì sự tâm huyết, nhiệt tình của thầy.

Em cũng xin gửi lời cảm ơn chân thành đến các thầy cô trong khoa ***Thương mại điện tử*** trường *Đại học Kinh tế - Đại học Đà Nẵng*, người lái đò, trang bị cho em những kiến thức nền tảng vững chắc trong suốt năm học vừa qua. Những bài giảng của thầy cô đã khơi gợi cho em tham gia đăng ký làm luận văn trong năm học cuối này.

Em cũng xin cảm ơn đến Angel Coffee ở 123 Nguyễn Đức Trung Đà Nẵng đã gợi cho em nguồn cảm hứng để làm đề tài này. Em cũng cảm ơn sự nhiệt tình của các anh chị khiếm thính đang làm việc tại quán.

Ngoài ra, em xin cảm ơn đến gia đình, mẹ Linh, bố Quỳnh, bạn bè, những người luôn động viên, thấu hiểu những tháng ngày em tập trung vào làm luận văn. Sự tin tưởng và ủng hộ của mọi người là nguồn động lực to lớn giúp em vượt qua khó khăn trong quá trình nghiên cứu.

Cuối cùng, em xin cảm ơn tất cả những ai quan tâm, góp ý dù là nhỏ nhất trong quá trình em thực hiện luận văn cuối cấp này.

Dù đã cố gắng hết mình, luận văn này chắc chắc vẫn còn nhiều thiếu sót. Em cũng mong nhận được sự thông cảm, những ý kiến xây dựng từ quý thầy cô, bạn đọc để hoàn thiện tốt hơn trong tương lai.

MỤC LỤC

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

DANH MỤC CHỮ VIẾT TẮT

STT	Chữ viết tắt	Tiếng anh	Ý nghĩa
1	DA	Data Augmentation	Tăng cường dữ liệu
2	EDA	Easy Data Augmention	Tăng cường dữ liệu đơn giản
3	LLM	Large Language Model	Mô hình ngôn ngữ lớn
4	MT	Machine Translation	Dịch máy
5	NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
6	NNKH		Ngôn ngữ ký hiệu
7	SR	Synonym Replacement	Thay thế từ đồng nghĩa
8	RD	Random deletion	Xóa từ ngẫu nhiên
9	RI	Random insertion	Chèn từ ngẫu nhiên
10	VSL	Vietnamese Sign Language	Ngôn ngữ ký hiệu Việt Nam
11	WER	Word Error Rate	Tỷ lỗi lỗi từ

PHẦN MỞ ĐẦU

1. Câu chuyện khởi đầu

Suốt bốn năm học đại học, bản thân thường xuyên tham gia các hoạt động thiện nguyện. Những trải nghiệm liên tục đó đã giúp bản thân hình thành một tư duy luôn hướng đến cộng đồng, đặc biệt là thấu cảm và mong muốn được đóng góp cho nhóm người yếu thế trong xã hội.

Câu chuyện nghiên cứu này bắt đầu một cách rất tình cờ. Trong một lần bản thân được biết đến nhóm bạn khiếm thính đang làm việc tại quán cà phê Angel ở Đà Nẵng. Tại đây, các bạn chủ yếu đảm nhận các công việc như phục vụ nước, dọn dẹp bàn ghế... Trong khi đó, các công việc đòi hỏi giao tiếp trực tiếp với khách hàng (như gọi món) đều do các nhân viên bình thường phụ trách.

Tuy nhiên, trong quá trình phục vụ, bản thân quan sát thấy vẫn có những tình huống khách hàng cần trao đổi trực tiếp với các bạn khiếm thính, ví dụ như khi họ muốn yêu cầu thêm nước lọc hoặc giấy ăn. Những lúc như vậy, việc giao tiếp ngay lập tức gặp nhiều khó khăn và lúng túng.

Từ quan sát thực tế đó, bản thân bắt đầu suy nghĩ nghiêm túc về việc ứng dụng công nghệ để hỗ trợ giao tiếp giữa người khiếm thính và khách hàng. Là một sinh viên theo học một ngành mang tính khoa học và đổi mới cao, bản thân thấy mình có cơ hội và cả trách nhiệm áp dụng kiến thức chuyên môn để góp phần giải quyết một vấn đề xã hội cụ thể.

Xuất phát từ mong muốn đó, bản thân quyết định thực hiện đề tài này. Mục tiêu không chỉ dừng lại ở môi trường dịch vụ, mà còn mở rộng ra với mong muốn hỗ trợ người khiếm thính tiếp cận các nền tảng tri thức (như video giáo dục, thời sự) một cách dễ dàng hơn, giúp họ hòa nhập tốt hơn trong công việc và đời sống.

2. Bối cảnh và động lực nghiên cứu

Với cộng đồng ước tính có trên 2,5 triệu người điếc và khiếm thính tại Việt Nam, việc xây dựng một môi trường sống và làm việc hòa nhập là một mục tiêu xã hội cấp thiết. Trong những năm gần đây, xu hướng người khiếm thính tham gia vào thị trường lao động ngày càng tăng, điều này có thể quan sát rõ tại các thành phố lớn như Đà Nẵng.

Tuy nhiên, xu hướng tích cực này cũng làm bộc lộ một thách thức cốt lõi: **rào cản giao tiếp** giữa nhân viên khiếm thính (sử dụng Ngôn ngữ Ký hiệu) và khách hàng (sử dụng ngôn ngữ nói).

Để giải quyết rào cản này, các công nghệ tiên tiến trong Xử lý Ngôn ngữ Tự nhiên, đặc biệt là các Mô hình Ngôn ngữ Lớn (LLMs), mang lại tiềm năng to lớn. Nghiên cứu này hướng đến việc xây dựng một hệ thống hỗ trợ giao tiếp, có khả năng tiếp nhận đầu vào là giọng nói Tiếng Việt và chuyển đổi nó thành một định dạng đầu ra tương thích với Ngôn ngữ Ký hiệu. Về mặt học thuật, bài toán này nằm trong lĩnh vực Dịch máy Ngôn ngữ Ký hiệu (Sign Language Machine Translation – SLMT).

Tuy nhiên, một thách thức khoa học căn bản là **sự khác biệt về cấu trúc ngữ pháp** giữa Tiếng Việt (ngôn ngữ nói) và Ngôn ngữ Ký hiệu Việt Nam (VSL). VSL có trật tự từ, cú pháp, và các đặc điểm ngôn ngữ riêng biệt. Việc dịch máy trực tiếp từ Tiếng Việt có thể tạo ra kết quả sai lệch về ngữ nghĩa hoặc "ký hiệu bằng ngữ pháp Tiếng Việt" (Sign-Supported Vietnamese), thay vì VSL bản địa.

Do đó, một bước trung gian then chốt, và cũng là **động lực nghiên cứu chính** của đề tài, là xây dựng một mô hình dịch máy (MT) có khả năng chuyển đổi văn bản Tiếng Việt chuẩn sang văn bản tuân thủ đúng cú pháp VSL (text-to-VSL-text). Dạng biểu diễn văn bản VSL này không chỉ giúp truyền tải đúng ngữ nghĩa, mà còn đóng vai trò là một đầu vào (input) chuẩn hóa, có cấu trúc, hỗ trợ trực tiếp cho các nghiên cứu sâu hơn về tổng hợp và biểu diễn avatar 3D trong tương lai.

3. Mục tiêu nghiên cứu

Để giải quyết bài toán này, nghiên cứu đặt ra các mục tiêu cụ thể như sau:

Một là, xây dựng tài nguyên dữ liệu chuyên biệt. Phát triển một bộ dữ liệu song ngữ (Tiếng Việt – VSL Gloss) chuyên biệt cho lĩnh vực dịch vụ cà phê. Bộ dữ liệu này được xây dựng dựa trên phương pháp tăng cường dữ liệu (data augmentation) từ một tập dữ liệu lõi ban đầu. Mục tiêu là tạo ra một tập dữ liệu có độ bao phủ đủ rộng, bao quát các tình huống giao tiếp phổ biến, làm tài nguyên cốt lõi cho việc huấn luyện và đánh giá mô hình dịch máy (Mô-đun B).

Hai là, đánh giá và lựa chọn mô hình Nhận dạng Giọng nói (Module A). Thực hiện đánh giá và so sánh hiệu năng của các mô hình Nhận dạng Tiếng nói Tự động (ASR)

tiên tiến cho Tiếng Việt. Mục tiêu là xác định và lựa chọn mô hình có chỉ số Lỗi Từ (Word Error Rate - WER) thấp nhất, nhằm đảm bảo độ chính xác tối đa cho văn bản đầu vào (output của Mô-đun A) trước khi đưa vào mô hình dịch máy.

Ba là, phát triển và tối ưu hóa mô hình dịch cú pháp Vi-to-VSL (Module B). Thiết kế, huấn luyện, và tinh chỉnh các mô hình dịch máy (bao gồm ViT5, BARTpho) dựa trên kiến trúc Transformer cho bài toán chuyển đổi từ văn bản Tiếng Việt sang biểu diễn **Gloss** của Ngôn ngữ Ký hiệu Việt Nam. Mô hình phải đạt được hiệu suất cao và ổn định, với **mục tiêu cụ thể là đạt chỉ số WER trên tập kiểm thử (test set) thấp hơn 10%**, đảm bảo bản dịch có độ chính xác và tin cậy cao để có thể ứng dụng trong thực tế.

4. Phương pháp nghiên cứu

Nghiên cứu áp dụng phương pháp có cấu trúc gồm 2 module nối tiếp: Chuyển đổi giọng nói sang văn bản và dịch máy. Quy trình thực hiện trải qua ba giai đoạn chính:

Giai đoạn 1: Chuẩn bị dữ liệu. Kế thừa dữ liệu nền tảng (Vie-VSL-10k) và xây dựng dữ liệu chuyên biệt cho dịch vụ quán cà phê qua thu thập thực tế và tăng cường dữ liệu.

Giai đoạn 2: Xây dựng mô hình. Thực hiện huấn luyện và tinh chỉnh các mô hình ứng viên.

Giai đoạn 3: Đánh giá và tích hợp. Đánh giá độc lập từng module bằng các chỉ số đánh giá.

5. Phạm vi và giới hạn nghiên cứu

Xét về phạm vi, đề tại giới hạn trong bối cảnh giao tiếp tại quán cà phê, nơi có nhân viên là người khiếm thính, và khách hàng bình thường, nói tiếng Việt. Đầu vào của mô hình là giọng nói tiếng Việt, đầu ra là văn bản tiếng Việt đúng cú pháp của ngôn ngữ ký hiệu. Các vấn đề về diễn họa hình ảnh, video hoặc avatar nằm ngoài phạm vi nghiên cứu lúc này.

Xét về giới hạn, nghiên cứu phải tìm ra giải pháp cho sự khan hiếm dữ liệu chuyên biệt tính tổng quát hóa của mô hình khi áp dụng sang lĩnh vực khác.

6. Ý nghĩa khoa học và thực tiễn

Về ý nghĩa khoa học, luận văn đóng góp một bộ dữ liệu song ngữ chuyên mới cho động cồng nghiên cứu, cung cấp các kết quả thực nghiệm về việc ứng dụng kiến trúc Transformer cho bài toán dịch sang văn bản đúng theo cú pháp của ngôn ngữ ký hiệu Việt Nam.

Về ý nghĩa thực tiễn, kết quả nghiên cứu là bước đầu để giúp người khiếm thính hòa nhập vào môi trường làm việc, tạo cơ sở dữ liệu cho các ứng dụng diễn họa 3D trong tương lai.

7. Cấu trúc luận văn

Nội dung luận văn được trình bày như sau:

Mở đầu: Giới thiệu bối cảnh, mục tiêu, phương pháp tiếp cận và ý nghĩa đề tài

Chương 1: Tổng quan tình hình nghiên cứu, phân tích các công trình liên quan trong và ngoài nước.

Chương 2: Cơ sở lý thuyết, mô tả các mô hình nền tảng và các độ đo đánh giá

Chương 3: Quy trình triển khai nghiên cứu

Chương 4: Thực nghiệm, kết quả, đánh giá của các mô hình huấn luyện

Kết luận: Tổng kết đóng góp, hạn chế và định hướng phát triển

CHƯƠNG 1. GIỚI THIỆU BÀI TOÁN DỊCH MÁY NGÔN NGỮ KÝ HIỆU

Chương này cung cấp một cái nhìn tổng quan và nền tảng về bài toán dịch máy ngôn ngữ ký hiệu (NNKH). Mục tiêu của chương là phân tích bối cảnh lịch sử, đặc điểm ngôn ngữ học và hiện trạng nghiên cứu, từ đó làm rõ tầm quan trọng chiến lược của công nghệ này. Dịch myas NNKH không chỉ là một thách thức kỹ thuật mà còn là một công cụ mạnh mẽ để phá vỡ rào cản giao tiếp, tăng cường khả năng tiếp cận thông tin và thúc đẩy hòa nhập xã hội cho cộng đồng người khiếm thính trên toàn cầu nói chung và Việt Nam nói riêng.

1.1. Tổng quan về ngôn ngữ ký hiệu trên thế giới

Để xây dựng một hệ thống dịch thuật hiệu quả cho ngôn ngữ ký hiệu Việt Nam (VSL), trước hết cần phải hiểu rõ bản chất, lịch sử và cấu trúc của NNKH trên phương diện toàn cầu. Việc nắm bắt bối cảnh chung này không chỉ cung cấp nền tảng lý thuyết vững chắc mà còn giúp nhận diện đặc điểm chung và riêng, từ đó định hướng cho việc phát triển các mô hình dịch thuật phù hợp.

1.1.1. Lịch sử và sự phát triển

Lịch sử của NNKH hiện đại được đánh dấu bằng nỗ lực tiên phong trong việc giáo dục người điếc. Năm 1755, Cha Charles Michel de L'Eppe đã thành lập trường học miễn phí đầu tiên cho người điếc tại Paris, khai sinh ra hệ thống NNKH Pháp – một trong những hệ thống NNKH hoàn thiện sớm nhất trên thế giới (xuân Mỹ, 2019)

Mô hình giáo dục này đã lan toản và tạo ảnh hưởng sâu rộng. Thomas Hopkins Gallaudet, một trong những nhà giáo dục người Mỹ, sau khi sang châu Âu nghiên cứu, đã thuyết phục Laurent Clerc, một giáo viên người điếc tài năng ở Paris, cùng ông trở về Hoa Kỳ. Năm 1817, họ thành lập trường công đầu tiên cho người điếc tại Hartford, Connecticut, đặt nền tảng cho sự ra đời và phát triển của NNKH Mỹ (xuân Mỹ, 2019) .Sự kiện này đã tạo ra một dòng chảy kế thừa, khi ngôn ngữ ký hiệu Pháp trở thành nền tảng quan trọng cho sự hình thành của ngôn ngữ ký hiệu Mỹ. Đỉnh cao của những nỗ lực này là sự thành lập Đại học Gallaudet vào năm 1864, với sắc lệnh được ký bởi Tổng thống Abraham Lincoln, trở thành trường đại học đầu tiên và duy nhất trên thế giới dành cho người điếc lúc bấy giờ. Ngày nay, có hơn 200 loại NNKH khác nhau đang được sử dụng

trên toàn cầu, phản ánh sự đa dạng văn hóa và ngôn ngữ của cộng đồng người điếc (Zhang et al., 2025).

1.1.2. Bản chất của ngôn ngữ ký hiệu

Ngôn ngữ ký hiệu là một ngôn ngữ đa kênh, dựa trên thị giác truyền tải ý nghĩa thông qua sự kết hợp phức tạp của nhiều yếu tố (Saunders, Camgoz, & Bowden, 2021). Cách yếu tố này có thể chia thành hai nhóm chính:

+ Yếu tố thủ công (Manual Markers): các yếu tố được tạo ra bởi tay, bao gồm:

+ Hình dạng bàn tay: Hình dạng cụ thể của bàn tay khi thực hiện ký hiệu, ví dụ như nắm tay, xòe tay,...

+ Vị trí: Vị trí của bàn tay so với cơ thể, ví dụ như ký hiệu được thực hiện ở trước ngực, trên đầu hay bên cạnh má,...

+ Chuyển động: Hướng và quỹ đạo chuyển động của bàn tay

+ Hướng: Hướng của lòng bàn tay, ví dụ như lòng bàn tay hướng lên trên, hướng vào trong,...

+ Yếu tố phi thủ công (Non-Manual Markers): Các biểu đạt không dùng tay nhưng mang ý nghĩa ngữ pháp và ngữ nghĩa quan trọng:

+ Nét mặt: Ví dụ như việc nhướng mày có thể là biến một câu trần thuật và câu hỏi có/không (Zhang et al., 2025).

+ Chuyển động đầu và tư thế cơ thể: Góp phần nào làm rõ ngữ ngĩa và nhấn mạnh hoặc thể hiện được các vai trò ngữ pháp khác nhau.

Điều quan trọng cần nhấ mạnh là NNKH không phải là sự mô phỏng hay dịch từng từ của ngôn ngữ nói. Nó sở hưu một hệ thống cấu trúc ngữ pháp và cú pháp hoàn toàn riêng biệt, và được công nhận là bởi các nhà ngôn ngữ học (Saunders et al., 2021).

1.1.3. Ngôn ngữ ký hiệu Việt Nam

Tương tự như nhiều NNKH khác trên thế giới, ngôn ngữ ký hiệu Việt Nam cũng có nguồn gốc lịch sử liên quan đến NNKH Pháp. Lịch sử ghi nhận vào năm 1886, Cha Azemar, sau khi gửi một thanh niên câm điếc tên là Nguyễn Văn Trường sang Pháp học

tập vè phương pháp dùng ký hiệu, đã trở về và chính thức thành lập trường Lái Thiêu (xuân Mỹ, 2019).

Từ việc tìm hiểu bối cảnh chung, nghiên cứu sẽ đi sâu vào phân tích các đặc điểm cụ thể về từ vựng và ngữ pháp của NNKH Việt Nam, những yếu tố quan trọng quyết định sự thành công của bài toán dịch máy.

1.2. Đặc điểm về cấu trúc ngữ pháp ngôn ngữ ký hiệu Việt Nam

Để xây dựng một mô hình dịch máy có khả năng chuyển ngữ từ tiếng Việt nói sang NNKH một cách chính xác và tự nhiên, việc phân tích và hiểu các đặc trưng về từ vựng và cú pháp của NNKH là yêu cầu đầu tiên. Những khác biệt này chính là thách thức quan trọng.

1.2.1. Đặc điểm về từ vựng trong ngôn ngữ ký hiệu Việt Nam

Kho từ vựng của NNKH Việt Nam thường tập trung vào các khái niệm và từ ngữ thông dụng trong đời sống hàng ngày, và có phân giới hạn so với kho từ vựng phong phú của tiếng Việt nói.

Nghiên cứ của (Thu, 2018) đã xây dựng một bộ từ điển gồm khoảng 3000 từ thường dùng để phục vụ cho hệ thống diễn họa avatar 3D. Gần đây hơn, luận án tiến sĩ của Nguyễn Thị Bích Diệp (Diep, 2023) đã đóng góp một bộ từ điển quan trọng VSL-Lexicon, với khoảng 6000 đơn vị từ vựng (bao gồm từ đơn, số, và cụm từ), cung cấp nguồn tài liệu quý cho nghiên cứu.

Theo nghiên cứu (Uyên, 2020), từ vựng trong NNKH Việt Nam được hình thành chủ yếu qua 6 phương thức:

1. Trực chỉ: Đây là phương thức đơn giản nhất, người dùng chỉ trực tiếp vào đối tượng hoặc vị trí của đối tượng muốn nói đến.

2. Mô phỏng: Ký hiệu được tạo bằng cách tái hiện lại hình dáng hoặc hành động đặc trưng của sự vật, hiện tượng. Ví dụ, ký hiệu “sách” mô phỏng hành động mở sách, hay ký hiệu “núi” mô phỏng hình dáng ngọn núi.

3. Phản ánh và phân tích đặc trưng: Người dùng chọn một đặc điểm nổi bật nhất của sự vật (một bộ phận, tính chất, hoặc hành động đặc thù) để tạo thành ký hiệu.

4. Phát sinh: Một ký hiệu mới được tạo ra bằng cách kết hợp một ký hiệu gốc và một ký hiệu khác để mở rộng ý nghĩa. Ví dụ, từ ký hiệu gốc "nữ", người ta có thể tạo ra các ký hiệu phát sinh như "chị gái".

5. Vay mượn: NNKH Việt Nam vay mượn một số ký hiệu từ NNKH khác, đặc biệt là từ NNKH Pháp do ảnh hưởng từ lịch sử hình thành.

6. Bảng chữ cái và chữ số ngón tay: Phương thức này sử dụng các ký hiệu tay tương ứng với từng chữ cái trong bảng chữ cái tiếng Việt để đánh vần các từ, thường được dùng cho tên riêng, thuật ngữ hoặc những từ chưa có ký hiệu thống nhất.

Từ việc phân tích các đơn vị từ vựng, thấy răng NNKH Việt Nam có hệ thống từ vựng riêng. Bước tiếp theo là tìm hiểu các từ này được sắp xếp và kết hợp để tạo thành các câu có nghĩa hoàn chỉnh, tức là đặc điểm về cú pháp.

1.2.2. Đặc điểm cấu trúc ngữ pháp NNKH Việt Nam

Cú pháp của NNKH Việt Nam có nhiều điểm khác biệt cơ bản so với tiếng Việt nói, đòi hỏi các mô hình dịch máy có khả năng tái cấu trúc một cách linh hoạt.

Dưới đây là các quy tắc ngữ pháp đặc trưng của NNKH Việt Nam (Diep, 2023):

1. **Trật tự từ trong câu:** NNKH Việt Nam thường thay đổi trật từ từ để nhấn mạnh chủ thể hoặc đối tượng của hành động. Cấu trúc phổ biến là Chủ ngữ - Đối tượng – Động từ (SOV).

Ví dụ: Câu tiếng Việt "Tôi rất yêu động vật" khi chuyển sang cấu trúc NNKH Việt Nam sẽ trở thành "Tôi động vật yêu". Ví dụ này đồng thời minh họa cả hai quy tắc là trật từ từ SOV và việc lược bỏ từ chỉ tình thái ("rất").

2. **Lược bỏ từ:** NNKH Việt Nam có xu hướng rút gọn câu bằng cách lược bỏ các thành phân ngữ pháp được xem là không thiết yếu để truyền tải ý nghĩa cốt lõi.

Các từ thường bị lược bỏ bao gồm: giới từ (ví dụ: ở, trong, trên,..), liên từ (ví dụ: và, vì, nhưng,..) và các từ chỉ tình thái (ví dụ: rất, quá, lắm,..)

3. **Biểu đạt thì:** Trong NNKH Việt Nam, thì của các động từ không được thể hiện bằng cách biến đổi hình thái động từ như trong nhiều ngôn ngữ khác. Thay vào đó, thời gian của hành động được xác định bằng cách đặt các ký hệu chỉ thời gian (ví dụ: hôm qua, ngày mai, tuần trước, ...) ở đầu câu.

4. **Câu hỏi:** Các từ để hỏi như Ai, cái gì, khi nào, ở đâu, tại sao,...thường được đặt ở cuối câu, khác với tiếng Việt nơi chúng thườn đứng ở đầu câu.

Do đó, việc mô hình hóa chính xác các quy tắc cú pháp đặc thù này không chỉ là một thách thức, mà còn là yêu cầu tiên quyết để một hệ thống dịch máy NNKH Việt Nam đạt được sự tự nhiên và chính xác về mặt ngữ nghĩa.

1.3. Các nghiên cứu liên quan

Dịch máy ngôn ngữ ký hiệu (NNKH) là một lĩnh vực nghiên cứu liên cứu đang phát triển mạnh mẽ, kế thừa từ dịch máy nói chung, nhận dạng NNKH (Sign Language Recognition) và sản xuất NNKH (Sign Language Production). Mục này sẽ tổng hợp các công trình nghiên cứu tiêu biểu trong và ngoài nước để làm rõ các phương pháp tiếp cận chính và thách thức hiện có.

Các hướng tiếp cận chính trong dịch máy NNKH, có thể được chia thành hai hướng chính:

Sản xuất NNKH: Dịch từ văn bản hoặc giọng nói sang chuỗi ký hiệu (dưới dạng video hoặc chuỗi tư thế 3D). Đây là hướng tiếp cận được coi là hữu ích trực tiếp hơn cho cộng đồng người khiếm thính, vì nó giúp chuyển đổi khối lượng lớn nội dung từ ngôn ngữ nói/viết của thế giới người nghe sang ngôn ngữ mẹ đẻ của họ, từ đố tăng cường khả năng tiếp cận thông tin.

Dịch NNKN: Dịch từ video ký hiệu sang văn bản hoặc giọng nói, giúp người nghe hiểu được người khiếm thính.

Và trong bài nghiên cứu này sẽ tập trung vào hướng thứ nhất, sản xuất NNKH, đây là bài toán hữu ích để chuyển giao kiến thức đến người khiếm thính.

1.3.1. Các hướng tiếp cận ban đầu

Các hệ thống dịch NNKH sơ khai thường dựa trên hai hướng tiếp cận chính:

Thứ nhất, hệ thống dựa trên quy tắc và từ điển. Các hệ thống đời đầu như TESA hay VisiCast hoạt động bằng cách tra cứu một từ điển các ký hiệu đã được định nghĩa trước và áp dụng các quy tắc ngữ pháp được lập trình để sắp xếp lại trật từ từ (Patel, Patel, Khanvilkar, Patel, & Akilan, 2020). Hạn chế của phương pháp này là tính cứng nhắc, khó mở rộng và không thể xử lý được sự đa dạng của ngôn ngữ tự nhiên.

Thứ hai, sử dụng định dạng trung gian diễn họa. Để biểu diễn các ký hiệu một cách tự động bằng nhân vật ảo (avatar), các nhà nghiên cứu đã phát triển những hệ thống ký hiệu trung gian như HamNoSys (Hamburg Notation System) và ngôn ngữ đánh dấu SiGML (Signing Gesture Markup Language). Các hệ thống này mã hóa chi tiết các yếu tố của một ký hiệu thành một định dạng mà máy tính có thể đọc và tái tạo lại dưới dạng hoạt ảnh 3D (Patel et al., 2020) (Tauqeer, Muhammad, Babar, & Muhammad, 2021).

1.3.2. Mô hình Transformer

Sự bùng nổ của học sâu đã tạo ra một cuộc cách mạng trong lĩnh vực NLP và MT, chuyển từ các phương pháp lập trình sang các mô hình có khả năng tự học từ dữ liệu.

1.3.2.1. Sự ra đời của kiến trúc Transformer

Vào năm 2017, kiến trúc Transformer được giới thiệu đã thay đổi hoàn toàn bộ mặt của NLP (Vaswani et al., 2017). Khác với mô hình tuần tự như RNN hay LSTM, Transformer sử dụng cơ chế "chú ý" (attention), cho phép mô hình xem xét toàn bộ câu đầu vào cùng một lúc, xác định mối quan hệ giữa các từ, dù chúng ở cách xa nhau. Cơ chế "chú ý" đặc biệt hữu ích trong việc xử lý sự khác biệt về trật từ giữa tiếng Việt và NNKH, chẳng hạn như cấu trúc "Chủ ngữ - Tân ngữ - Động từ", vì nó cho phép mô hình xác định mối quan hệ giữa các yếu tố trong câu dù chúng năm ở những vị trí cách xa nhau trong câu gốc (Gruetzemacher & Paradice, 2022).

1.3.2.2. Học chuyển tiếp và các mô hình ngôn ngữ lớn

Transformer đã mở đường cho khái niệm Học chuyển tiếp (Transfer Learning). Thay vì huấn luyện một mô hình từ đầu cho mỗi tác vụ, các nhà nghiên cứu tiền huấn luyện (pretrain) các mô hình ngôn ngữ lớn trên một kho dữ liệu văn bản khổng lồ. Sau đó mô hình này có thể được tinh chỉnh (fine-tuning) cho các tác vụ cụ thể với một lượng dữ liệu nhỏ hơn nhiều (Gruetzemacher & Paradice, 2022).

Các LLM dựa trên kiến trúc Transformer có thể được phân thành hai loại chính:

Bảng 1-1. Mô rả các mô hình ngôn ngữ lớn dựa trên kiến trúc Transformer

Loại mô hình	Mô tả ví dụ
Mô hình mã hóa (Encoder – only)	Được thiết kế để hiểu sâu sắc ngữ cảnh văn bản. Tiêu biểu là BERT (Thu, 2018). Các nghiên cứu chỉ ra rằng việc phát triển các mô hình đơn ngữ như PhoBERT cho tiếng Việt (Nguyen & Nguyen, 2020) thường mang lại hiệu quả cao hơn các mô hình đa ngữ.
Mô hình Mã hóa – Giải mã (Encoder – Decoder)	Phù hợp cho tác vụ sinh văn bản (sequence-to-sequence), ví dụ như dịch thuật và tóm tắt. Các mô hình BARTpho (N. L. Tran, Le, & Nguyen, 2021) và ViT5 (Phan, Tran, Nguyen, & Trinh, 2022) là những mô hình tiên tiến nhất cho tác vụ sinh văn bản tiếng Việt.

1.3.3. Ứng dụng LLMs vào bài toán dịch NNKH Việt Nam

Các nghiên cứu mới nhất tại Việt Nam đã bắt đầu áp dụng những tiến bộ của LLMs để giải quyết bài toán dịch từ tiếng Việt nói sang cú pháp của NNKH.

1.3.3.1. Mô hình pipeline hiện đại

Hướng tiếp cận phổ biến hiện nay là xây dựng một hệ thống chuỗi gồm nhiều mô-đun: Âm thanh → Mô hình Speech-to-text → Văn bản tiếng Việt → Mô hình dịch thuật ngữ pháp → Văn bản đúng cú pháp NNKH Việt Nam. Quy trình này bắt đầu từ âm thanh đầu vào, được chuyển đổi thành văn bản tiếng Việt, sau đó văn bản này được đưa qua mộ mô hình dịch thuật ngữ pháp để tái cấu trúc thành văn bản theo đúng cú pháp của NNKH (Diep, 2023), (Trần Vũ Hoàng, 2025).

1.3.3.2. Các công nghệ được áp dụng

Speech-to-Text: Để chuyển đổi giọng nói tiếng Việt thành văn bản, các mô hình nhận dạng giọng nói tự động tiên tiến như PhoWhisper đã được sử dụng, cho thấy hiệu quả cao (Trần Vũ Hoàng, 2025).

Dịch thuật ngữ pháp: Giai đoạn cốt lõi của quy trình, chuyển đổi từ cú pháp tiếng Việt sang cú pháp NNKH, được đảm nhiệm bởi các mô hình mã hóa-giải mã mạnh mẽ

như ViT5. Các mô hình này được tinh chỉnh để học các quy tắc tái cấu trúc và lược bỏ đặc thù của NNKH Việt Nam (Diep, 2023)

Thách thức về dữ liệu: Một trong những rào cản lớn về việc NNKH được xem lài một trong những ngôn ngữ tài nguyên thấp, tức là có rất ít dữ liệu song ngữ để huấn luyện các mô hình lớn. Để khắc phcuj vấn đề này, các nhà nghiên cứu đang khám phá kỹ thuật tăng cường dữ liệu, một phương pháp sử dụng các mô hình Gen-AI để tự động tạo ra các câu mới dựa trên ngôn ngữ hiện có (Stefanovič, Radvilaitė, Pliuskuvienė, & Ramanauskaitė, 2025).

Những nghiên cứu tiên phong này đã đặt nền móng cho việc giải quyết bài toán dịch máy NNKH Việt Nam.

Tóm lại, chương 1 đã cung cấp cái nhìn tổng quan về bài toán dịch máy ngôn ngữ ký hiệu. NNKH là một hệ thống ngôn ngữ tự nhiên, hoàn chỉnh với lịch sử phát triển lâu và cấu trúc phức tạp. Qua nghiên cứu, NNKH Việt Nam cho thấy những đặc trưng răng biệt về từ vựng, cú pháp, tạo ra thách thức cho bài toán dịch máy tự động. Tuy nhiên, những sự tiến bộ trong lĩnh vực trí tuệ nhân tạo, đặc biệt là kiến trúc Transformer, các mô hình ngôn ngữ lớn, sẽ mở ra cơ hội tốt để giải quyết bài toán này.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

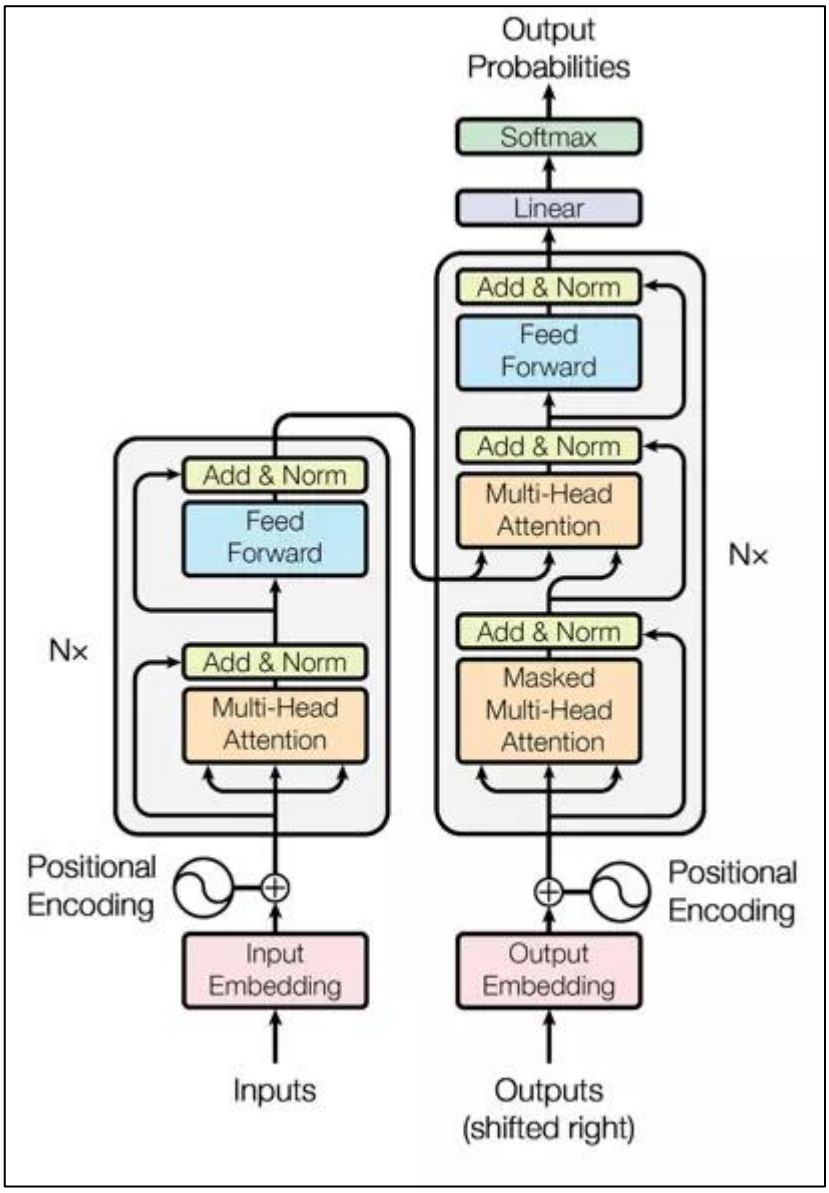
Chương này trình bày nền tảng lý thuyết cho các công nghệ cột lõi được ứng dụng trong nghiên cứu. Quy trình tổng thể của hệ thống được chia thành hai mô-đun chính và chương này sẽ đi sâu vào cơ sở khoa học từng phần. Mô đun đầu tiên là **Nhận dạng giọng nói**, có nhiệm vụ chuyển đổi âm thanh đầu vào thành văn bản tiếng Việt. Mô-đun thứ hai là Dịch máy, thực hiện chuyển đổi văn bản tiếng Việt đã nhận dạng sang cấu trúc cú pháp của ngôn ngữ ký hiệu Việt Nam.Việc hiểu rõ cơ sở lý thuyết là quan trọng để xây dựng một hệ thống phiên dịch hiệu quả, chính xác, với mục tiêu cuối cùng là tạo ra đầu vào chuẩn hóa cho các mô hình diễn họa 3D trong tương lai.

2.1. Mô hình Transformer

Kiến trúc Transformer được đề xuất lần đầu tiên bởi Vaswani và cộng sự vào năm 2017 trong bài báo có tiêu đề nổi tiếng “Attention is all you need.” Mô hình này đạt được nhiều hiệu suất cao trong nhiều tác vụ dịch máy (Vaswani et al., 2017).

Điểm khác biệt của Transformer so với các mô hình trước đó như Mạng nơ-rôn hồi quy (RNN) hay Bộ nhớ dài ngắn hạn (LSTM) là nó không sử dụng tính đệ quy hay bất kỳ phép tính tuần tự nào (L. T. T. Tran, Kim, La, & Van Pham, 2024). Điều này cho phép mô hình Transformer có khả năng xử lý song song, và tránh được vấn đề suy giảm gradient (Gruetzemacher & Paradice, 2022).

Mô hình Transformer là một mô hình sequence-to-sequence, gồm hai thành phần chính: Bộ mã hóa (Encoder) và bộ giải mã (Decoder) (Vaswani et al., 2017)



Hình 2.1. Kiến trúc mô hình Transformer

Nguồn: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. J. A. i. n. i. p. s. (2017). Attention is all you need. 30.

2.1.1. Kiến trúc tổng thể của Transformer

2.1.1.1. Đầu vào và mã hóa vị trí (Input and positional encoding)

Đầu tiên, các token đầu vào như từ, ký tự được nhúng thành các vector biểu diễn liên tục (Saunders et al., 2021).

Sau đó, lớp mã hóa vị trí (Positional Encoding) được thêm vào nhúng của chuỗi đầu vào để cung cấp thông tin về vị trí cho mạng. Các vector mã hóa vị trí này thường

được tạo ra bằng cách sử dụng hàm sin và cos (Vaswani et al., 2017). Vector này là đầu vào cho khối Encoder đầu tiên.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right)$$

Trong đó,

+ p là vị trí của token trong chuỗi đầu vào

+ i là chỉ số index dùng để tham chiếu đến chiều của vector nhúng

+ d_{model} là chiều của vector nhúng, thường là hằng số cho tất cả các lớp trong mô hình Transformer.

2.1.1.2. Bộ mã hóa (Encoder)

Encoder có nhiệm vụ mã hóa chuỗi đầu vào thành một không gian đặc trưng chiều cao. Nó bao gồm một chồng (stack) gồm L lớp giống hệt nhau. Mỗi lớp Encoder thường có hai lớp con chính (Vaswani et al., 2017):

+ Multi - Head Self - Attention: tạo ra một biểu diễn ngữ cảnh có trọng số

+ Feed – Forward Network: là lớp kết nối hoàn toàn

Xung quanh mỗi lớp con này đều áp dụng một kết nối dư (residual connection) và sau đó là chuẩn hóa lớp (layer normalization) để hỗ trợ việc truyền tải thông tin và ổn định quá trình huấn luyện.

2.1.1.3. Bộ giải mã (decoder)

Decoder chuyển đổi các biểu diễn chiều cao từ encoder trở lại thành chuỗi đầu ra. Mỗi lớp Decoder thường bao gồm 3 lớp:

+ Multi – Head Self – Attention có mặt nạ: đảm bảo rằng việc dự đoán ở một bước thời gian chỉ phụ thuộc vào các thông báo đầu ra trước đó, không nhìn thấy các từ trong tương lai.

+ Multi – Head Encoder – Decoder Attention (cross attention): lớp này tập trung vào đầu ra của encoder và nắm thông tin liên quan trong quá trình tạo chuỗi đầu ra.

+ Feed – Forward Network

2.1.2. Cơ chế Self – Attention

Self – Attention là cơ chế nền tảng của Transformer, cho phép mô hình có sự phụ thuộc toàn cục (Saunders et al., 2021). Cơ chế này hoạt động bằng cách học mối quan hệ giữa các token trong chuỗi và mức độ liên quan của mỗi bước thời gian đối với ngữ cảnh. Một khía cạnh quan trọng của self - attention là nó cho phép mô hình tập trung vào các từ quan trọng trong câu bằng cách thay đổi ma trận trọng số self – attention. Các từ quan trọng sẽ có giá trị trọng số cao hơn, trong khi các từ không quna trọng sẽ có giá trị trọng số thấp hơn. Điều này giúp mô hình hiểu được mối quan hệ và sự phụ thuộc giữa các từ trong câu.

Self – attention được tính toán bằng cách sử dụng cơ chế Scaled Dot- Product attention. Đầu tiên, ba ma trận chiếu riêng biệt được tạo ra từ vector đầu vào: Queries (Q), Keys (K) và Values (V) (Vaswani et al., 2017).

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

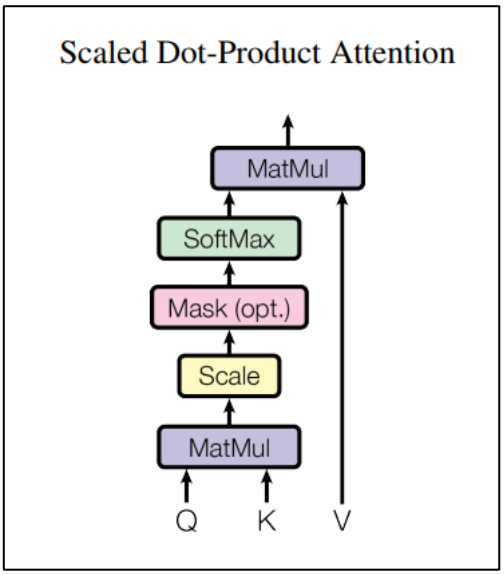
Trong đó,

+ Q, K, V là ma trận tương ứng được chiếu từ đầu vào

+ QK^T tính toán độ tương đồng giữa Queries và Keys

+ d_k là kích thước của vector key, được dùng để chia tỷ lệ nhằm tránh giá trị quá lớn đẩy hàm softmax vào các khu vực có gradient cực tiểu.

Hàm softmax chuẩn hóa kết quả thành ma trận trọng số. Ma trận trong số này sau đó nhân V để tạo thành một tổ hợp vector các giá trị, trong đó mỗi hàng là biểu diễn ngữ cảnh của token đầu vào tương ứng.



Hình 2.2 Cơ chế Scaled Dot – Product Attention

Nguồn: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. J. A. i. n. i. p. s. (2017). Attention is all you need. 30.

2.1.3. Cơ chế Multi – Head Attention

Multi head attention là kỹ thuật mở rộng của Self attention được sử dụng trong Transformer. Cơ chế này được tích hợp vào các lớp Encoder và Decoder. Multi head attention thực hiện tính toán sự chú ý h song song nhiều lần, sử dụng các bộ ma trận trọng số khác nhau (Vaswani et al., 2017). Điều này cho phép mô hình:

+ Mô hình hóa các sự kết hợp trọng số khác nhau của mỗi chuỗi, cải thiện sức mạnh biểu diễn mô hình.

+ Tạo ra các biểu diễn khác nhau của chuỗi đầu vào, học được thông tin bổ sung trong các không gian con khác nhau.

Đầu ra của mỗi head được tính bằng công thức scaled dot product attention với ma trận trọng số riêng biệt W_i^Q, W_i^K, W_i^V cho từng đầu.

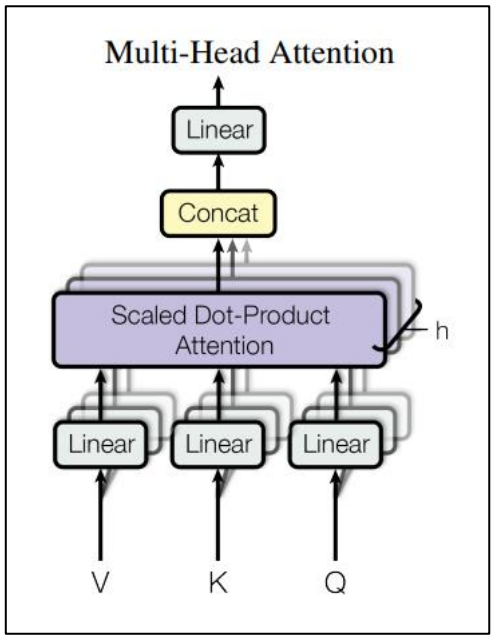
Công thức tính Multi – Head Attention:

$$MultiHead\ (Q, K, V) = Concat(head_1, \dots, head_h).W^O$$

Trong đó:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

W^O là ma trận trọng số của lớp tuyến tính cuối cùng, chiếu kết quả nối về lại kích thước ban đầu của vector nhúng (Vaswani et al., 2017).



Hình 2.3 Cơ chế Multi – Head Attention

Nguồn: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. J. A. i. n. i. p. s. (2017). Attention is all you need. *30*.

2.2. Nhận dạng giọng nói (Speech to Text – Module 1)

Module nhận dạng giọng nói đóng vai trò quan trọng trong quy trình phiên dịch, là bước đầu quyết định chất lượng của toàn hệ thống. Nhiệm vụ của module này là chuyển đổi tín hiệu âm thanh giọng nói tiếng Việt thành dạng văn bản một cách chính xác. Đây là bước đầu vào tiên quyết, do bất kỳ sai sót nào ở giai đoạn này dù nhỏ cũng sẽ được khuếch đại và ảnh hưởng trực tiếp đến chất lượng bản dịch ngôn ngữ ký hiệu cuối cùng. Một văn bản đầu đầu vào thiếu chính xác sẽ dẫn đến một chuỗi ký hiệu sai về ngữ nghĩa, làm mất tính hiệu quả của hệ thống. Để đảm bảo độ tin cậy cao nhất, các phần dưới đây sẽ làm rõ các khía cạnh lý thuyết và cơ sở lựa chọn.

2.2.1. Tổng quan về nhận dạng giọng nói

Nhận dạng giọng nói tự động (Automatic Speech Recognition), hay chuyển đổi lời nói thành văn bản, là một lĩnh vực công nghệ cho phép máy tính nhận diện và chuyển đổi ngôn ngữ nói của con người thành văn bản.

Lịch sử phát triển của bài toán này trải qua nhiều giai đoạn, từ những hệ thống ban đầu dựa trên phương pháp thống kê truyền thống. Tuy nhiên, một cách mạng đã diễn ra khi học sâu (Deep Learning) và kiến trúc mạng nơ rôn sâu ra đơi. Các mô hình LSTM đã cải thiện khả năng nắm bắt tuần tự ngữ cảnh của giọng nói. Gần đây kiến trúc Transformer đã tiếp tục tạo đột phá mới, trở thành nền tảng cho hệ thống nhận dạng giọng nói tự động với hiệu suất cao.

Đối với tiếng Việt, bài toán nhận dạng giọng nói mang những thách thức riêng. Theo nghiên cứu của Nguyễn Kết Đoàn và cộng sự (Nguyễn, Nguyễn, Trần, & Võ, 2024), tiếng Việt là một ngôn ngữ phức tạp, nơi ý nghĩa của từ không chỉ phụ thuộc vào các âm vị mà còn được quyết định bởi thanh điệu. Sự đa dạng về thanh điệu (6 thanh) và ngữ pháp phong phú đòi hỏi mô hình phải có khả năng phân biệt sự khác biệt trong âm thanh để chuyển đổi chính xác.

2.2.2. Các mô hình Transformer trong nhận dạng giọng nói

Sự ra đời của kiến trúc Transformer được giới thiệu lần đầu trong bài báo "Attention is All you need" của (Vaswani et al., 2017), tạo ra bước ngoặt trong lĩnh vực xử lý chuỗi, gồm xử lý ngôn ngữ tự nhiên và nhận dạng giọng nói.

Trong bối cảnh các mô hình ngôn ngữ và giọng nói hiện đại, "Học chuyển giao" đóng vai trò trung tâm. Học chuyển giao là kỹ thuật trong học máy, trong đó kiến thức thu được từ việc giải quyết một bài toán tổng quát trên tập dữ liệu lớn được chuyển giao và tái sự dụng để cải thiện hiệu suất trên một bài toán cụ thể, thường là có tập dữ liệu nhỏ hơn. Theo (Gruetzemacher & Paradice, 2022), quy trình này thường gồm hai bước chính:

Thứ nhất, tiền huấn luyện. Mô hình được huấn luyện trên một tập dữ liệu khổng lồ, thường là không gán nhãn (ví dụ: hàng nghìn giờ âm thanh, hoặc hàng tỷ văn bản). Mục tiêu của giai đoạn này là để mô hình học các biểu diễ tổng quát về ngôn ngữ hoặc đặc trưng âm thanh.

Thứ hai, tinh chỉnh. Sau khi đã học được kiến thức nền tảng, mô hình sẽ được tinh chỉnh trên một tập dữ liệu nhỏ hơn, có gán nhãn và dành riêng cho một tác vụ cụ thể. Để minh họa, một mô hình nhận dạng giọng nói có thể được tiền huấn luyện trên hàng nghìn giờ dữ liệu âm thanh đa ngôn ngữ để học các biểu diễn âm thanh tổng quát.

Sau đó, mô hình này được tinh chỉnh trên một tập dữ liệu nhỏ hơn chứa giọng nói tiếng Việt có gán nhãn để chuyên môn hóa.

2.2.3. Giới thiệu mô hình phoWhisper

Dựa trên các phân tích và khảo sát, mô hình phoWhipser đã được lựa chọn để triển khai cho nhận dạng giọng nói. phoWhisper là một mô hình nhận dạng giọng nói được phát triển dựa trên kiến trúc Whisper của OpenAI và đã được chuyên biệt hóa cho tiếng Việt (Le, Nguyen, & Nguyen, 2024).

Kiến trúc của phoWhisper cũng kế thừa nền tảng Transformer Encoder-Decoder mạnh của mô hình gốc. Điều này cho phép phoWhisper thừa hưởng những ưu điểm vượt trội của kiến trúc Transformer, bao gồm khả năng xử lý hiệu quả các chuỗi âm thanh dài và học các mối quan hệ phụ thuộc ngữ cảnh phức tạp trong giọng nói. Nhờ tối ưu hóa cho tiếng Việt, phoWhisper được đề xuất mang lại hiệu quả cao cho bài toán nhận dạng giọng nói trong bối cảnh đặc thù tiếng Việt (Le et al., 2024).

2.3. Tăng cường dữ liệu dạng văn bản

Tăng cường dữ liệu (Data augmentation – DA), hay gọi là làm giàu dữ liệu, là một kỹ thuật được sử dụng để mở rộng tập dữ liệu hiện có một cách nhân tạo (Stefanovič et al., 2025).

DA tạo ra các phiên bản khác nhau của dữ liệu gốc mà không cần thu thập dữ liệu. Mục tiêu chính của tăng cường dữ liệu là tăng cường hiệu suất mô hình học máy bằng cách cung cấp cho chúng lượng dữ liệu huấn luyện lớn hơn (GeeksforGeeks, 2025).

Điều này đặc biệt quan trọng trong phạm vi nghiên cứu có dữ liệu khan hiếm hoặc dữ liệu ban đầu có sự mất cân bằng giữa các lớp. Trong các bài toán dịch máy, tăng cường dữ liệu là quá trình thêm dữ liệu vào kho ngữ liệu huấn luyện nhằm nâng cao hiệu suất của mô hình.

Một số nguyên tắc trong quá trình tăng cường dữ liệu:

+ Dữ liệu tạo ra không nên quá giống dữ liệu gốc vì điều này có thể dẫn đến hiện tượng học vẹt (overfitting) (GeeksforGeeks, 2025).

+ Dữ liệu tạo ra không nên sai lệch quá nhiều so với dữ liệu ban đầu vì điều này có thể dẫn đến hiệu suất kém của mô hình (GeeksforGeeks, 2025).

2.3.1. Các phương pháp tăng cường dữ liệu

2.3.1.1. Phương pháp Easy Data Augmentation (EDA)

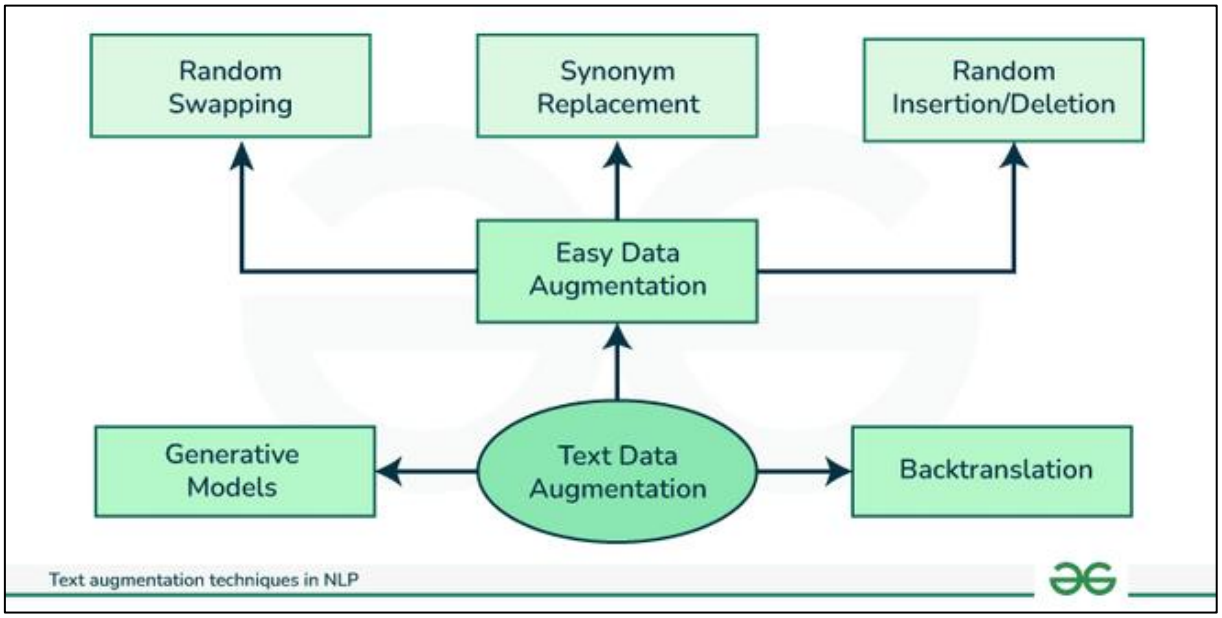
Phương pháp EDA là một phương pháp đơn gian nhưng hiệu quả để tăng cường dữ liệu (Luu, Van Nguyen, Nguyen, & Applications, 2024). Các kỹ thuật chính của EDA gồm:

+ Thay thế từ đồng nghĩa (Synonym replacement – SR): Thay thế ngẫu nhiên một hoặc nhiều từ trong câu bằng từ đồng nghĩa mà không thay đổi nghĩa tổng thể (GeeksforGeeks, 2025).

+ Chèn ngẫu nhiên (Random insertion – RI): Chèn ngẫu nhiên các từ dư thừa hoặc tương đồng về nghĩa vào câu (GeeksforGeeks, 2025).

+ Xóa ngẫu nhiên (Random Deletion – RD): Loại bỏ ngẫu nhiên các từ dư thừa và vẫn giữ nguyên ý nghĩa ngữ nghĩa của câu (GeeksforGeeks, 2025).

+ Hoán đổi ngẫu nhiên (Random Swapping – RS): Trao đổi vị trí hai từ ngẫu nhiên trong câu n lần (GeeksforGeeks, 2025).



Hình 2.4 Các phương pháp tăng cường dữ liệu

Nguồn: GeeksforGeeks. (2025). *Text Augmentation Techniques in NLP*. Retrieved from https://www.geeksforgeeks.org/nlp/text-augmentation-techniques-in-nlp/

2.3.1.2. Phương pháp dựa trên mô hình sinh (Generative models)

Dịch ngược (Back – translation): tài liệu nguồn được dịch sang một ngôn ngữ đích bằng mô hình dịch máy, sau đó dịch ngược lại từ ngôn ngữ đích sang nguồn ngữ nguồn. Quan trình này tạo câu mới để huấn luyện (GeeksforGeeks, 2025).

Sử dụng mô hình ngôn ngữ lớn: Các công cụ Gen – AI như OpenAI Chat GPT, Google Gemini, Microsoft Copilot,… có thể được sử dụng để tăng cường tập dữ liệu văn bản bằng cách tạo ra văn bản mới tương tự hoặc viết lại dựa trên mẫu gốc (Stefanovič et al., 2025).

2.4. Dịch máy (Translation machine – Module 2)

Dịch máy đóng vai trò quan trong trong kiến trúc tổng thể của hệ thống, là cầu nối then chốt giữa module nhận dạng giọng nói và module diễn họa avatar 3D. Nhiệm vụ quan trọng của module này là chuyển đổi văn bản tiếng Việt, vốn là đầu ra của quá trình nhận dạng giọng nới sang một dạng văn bản mới, biểu diễn chính xác cấu trúc ngữ pháp của NNKH Việt Nam. Quá trình chuyển đổi này không chỉ là một bước dịch đơn thuần, mà còn là tác vụ tái cấu trúc ngữ pháp phức tạp. Sự khác biệt cơ bản về cấu trúc giữa ngôn ngữ nói tiếng Việt và ngôn ngữ ký hiệu đòi hỏi một bước trung gian để chuẩn hóa văn bản trước khi có thể tiến hành diễn họa (Thu, 2018). Do đó, việc xây dựng thành công một mô hình dịch có khả năng chuyển đổi cú pháp chính xác là bắt buộc, đảm bảo ngữ nghĩa được truyền tải trọn vẹn và tạo đầu ra chuẩn hóa cho nghiên cứu sâu hơn về biểu diễn avatar 3D trong tương lai.

2.4.1. Tổng quan về bài toán dịch máy

Trong bối cảnh đề tài, bài toán dịch được định nghĩa là một tác vụ chuyển đổi văn bản sang văn bản (Text to text). Cụ thể, mô hình sẽ tiếp nhận một chuỗi văn bản đầu vào tuân theo ngữ pháp tiếng Việt và tạo ra một chuỗi văn bản đầu ra biểu diễ cấu trúc ngữ pháp đặc thù của NNKH Việt Nam. Thay vì tập trung dịch giữa hai ngôn ngữ khác nhau như thông thường, thì bài nghiên cứu này tập trung nhiều vào việc tái cấu trúc ngữ pháp.

Bài toán này đối mặt với hai thách thức chính. Thứ nhất, sự hiếm nghiêm trọng về tài nguyên dữ liệu song ngữ chất lượng cao giữa tiếng Việt và NNKH Việt Nam, đặc biệt là trong lĩnh vực chuyên biệt như dịch vụ quán cà phê. Thứ hai, sự khác biệt lớn về

cấu trúc ngữ pháp giữa hai ngôn ngữ, khi NNKH Việt Nam thường có trật từ từ khác biệt và lược bỏ các từ chức năng như giới từ, liên từ. Những thách thức này khiến các phương pháp dịch máy thống kê hoặc dựa trên luật truyền thống khó có thể đạt hiệu quả cao.

Trong những khó khăn đó, sự phát triển mô hình ngôn ngữ lớn dựa trên kiến trúc Transformer đã mở ra một hướng tiếp cận mới, hứa hẹn mang hiệu quả cao.

2.4.2. Các mô hình ngôn ngữ lớn cho ngôn ngữ tiếng Việt

Kiến trúc Transformer, được giới thiệu bởi (Vaswani et al., 2017), đã tạo ra một cuộc cách mạng trong lĩnh vực xử lý ngôn ngữ tự nhiên, đặt nền móng cho thế hệ mô hình ngôn ngữ lớn hiện đại (Gruetzemacher & Paradice, 2022). Các mô hình này được phân loại thành ba kiến trúc chính dựa trên cơ chế tiền huấn luyện:

Tự mã hóa: Các mô hình như BERT (Nguyen & Nguyen, 2020) học biểu diễn ngôn ngữ từ ngữ cảnh hai chiều (trái, phải) rất mạnh trong tác vụ hiểu ngôn ngữ như phân loại văn bản.

Tự hồi quy: Các mô hình như GPT dự đoán từ tiếp theo dựa trên các từ đứng trước, phù hợp cho tác vụ sinh văn bản.

Chuỗi sang chuỗi: Các mô hình như T5 và BART (N. L. Tran et al., 2021) được thiết kế với bộ mã hóa và giải mã, khiến chúng trở nên đặc biệt phù hợp cho bài toán yêu cầu chuyển đổi từ một chuỗi đầu vào sang một chuỗi đầu ra, ví dụ như dịch máy hoặc tóm tắt văn bản (Gruetzemacher & Paradice, 2022).

Trong nghiên cứu này, việc lựa chọn các mô hình được tiền huấn luyện riêng cho tiếng Việt, thay vì mô hình đa ngôn ngữ là một quyết định chiến lược. Các nghiên cứu chỉ ra rằng mô hình chuyên biệt cho một ngôn ngữ thường mang lại hiệu suất cao do được huấn luyện trên kho dữ liệu lớn và đa dạng cho ngôn ngữ đó, giúp mô hình nắm bắt sâu sắc hơn đặc trưng ngữ nghĩa và cú pháp (Kowsher et al., 2022).

Dựa trên phân tích trên, nghiên cứu này lựa chọn ba mô hình ứng viên thuộc kiến trúc chuỗi sang chuỗi được tiền huấn luyện chuyên sâu cho tiếng Việt để tiến hành so sánh và đánh giá.

2.4.2.1. BARTpho

Đây là mô hình dựa trên kiến trúc BART (Bidirectional and Auto-Regressive Transformers). Kiến trúc này kết hợp một bộ mã hóa hai chiều theo kiểu BERT (Bidirectional Encoder) để hiểu sâu ngữ cảnh đầu vào, và một bộ giải mã tự hồi quy theo kiểu GPT (Auto Regressive Decoder) để sinh văn bản đầu ra một cách mạch lạc. Sự kết hợp này giúp BARTpho có khả năng hiểu sâu ngữ cảnh đầu vào và sinh văn bản đầu ra một cách mạch lạc và phù hợp cho nhiệm vụ sinh văn bản (N. L. Tran et al., 2021).

2.4.2.2. ViT5 (phiên bản base và large)

ViT5 là phiên bản của mô hình T5 (Text-to-Text Transfer Transformer) được xây dựng dành riêng cho tiếng Việt. Triết lý cốt lõi của T5 là coi mọi bài toán NLP là một bài toán chuyển đổi văn bản sang văn bản". Cách tiếp cận này rất linh hoạt và hiệu quả chp nhiệm vụ đề tài, vì nó biến bài toán chuyển đổi cấu trúc ngữ pháp thành tác vụ dịch text to text (Phan et al., 2022).

Dưới đây là bảng tóm tắt siêu tham số chính của mô hình ViT5-base được lấy thông tin khi khởi tạo mô hình trên Google Colab.

Bảng 2-1. Tóm tắt siêu tham số chính của mô hình ViT5-base

Siêu tham số	Giá trị	Mô tả
d_model	768	Kích thước của vector ẩn (hidden size).
num_layers	12	Tổng số lớp của Encoder.
num_decoder_layers	12	Tổng số lớp của Decoder.
num_heads	12	Số lượng "đầu" chú ý (attention heads).
d_ff	3072	Kích thước của lớp feed-forward trung gian.
n_positions	36.096	Kích thước bộ từ vựng, được tối ưu hóa cho tiếng Việt.
pad_token_id	0	ID của token dùng để đệm (padding) chuỗi.

Với mô hình ứng viên đã được xác định, bước tiếp theo là áp dụng một phương pháp huấn luyện hiệu quả để chuyên môn hóa cho tác vụ dịch sang cú pháp NNKH Việt Nam.

2.4.3. Phương pháp huấn luyện

Phương pháp cốt lõi được áp dụng trong nghiên cứu này là Học chuyển giao (Transfer Learning). Học chuyển giao là một kỹ thuật trong học máy, cho phép tận dụng tri thức mà một mô hình đã học được từ một tác vụ trên bộ dữ liệu lớn và áp dụng tri thức đó để giải quyết một tác vụ khác, thường có liên quan, trên bộ dữ liệu nhỏ hơn (Gruetzemacher & Paradice, 2022). Trong bối cảnh tài nguyên dữ liệu song ngữ Việt – NNKH Việt Nam còn hạn chế, học chuyển giao là một lựa chọn tối ưu.

Cụ thể, nghiên cứu triển khai chiến lược thông qua hai kỹ thuật chính:

Học chuyển giao: Đây là chiến lược tổng thể, trong đó tri thức ngôn ngữ tiếng Việt tổng quát đã được các mô hình BARTpho và ViT5 học trong quá trình tiền huấn luyện, được chuyển giao để giải quyết bài toán cụ thể là dịch sang cú pháp NNKH Việt Nam. Thay vì huấn luyện một mô hình từ đầu, bắt đầu với nền tảng tri thức vững chắc, giúp quá trình hội tụ nhanh hơn, hiệu suất cao hơn.

Tinh chỉnh: Đây là quá trình huấn luyện tiếp các mô hình tiền huấn luyện trên bộ dữ liệu chuyên biệt của bài toán. Dựa trên mục tiêu đề tài, một chiến lược tinh chỉnh hai giai đoạn được áp dụng.

Giai đoạn 1: Các mô hình ứng viên được tinh chỉnh trên một bộ dữ liệu song ngữ Việt – NNKH Việt Nam tổng quát. Mục tiêu là trang bị cho mô hình khả năng dịch và tái cấu trúc ngữ pháp cơ bản.

Giai đoạn 2: Mô hình có hiệu suất tốt nhất từ giai đoạn1 sẽ được lựa chọn để tiếp tục tinh chỉnh trên bộ dữ liệu chuyên sâu về lĩnh vực dịch vụ cà phê. Giai đoạn này nhằm chuyên môn hóa kiến thức của mô hình, giúp nó xử lý chính xác các tình hướng giao tiếp dặc thù trong bối cảnh cà phê,

Quá trình xây dựng bộ dữ liệu chuyên biệt cho lĩnh vực cà phê được thự hiện một cách cẩn thận, bao gồm thu thập các mẫu câu thực tế và áp dụng các kỹ thuật tăng cường dữ liệu (data augmentation).

Tóm lại, để giải quyết bài toán chuyển đổi văn bản tiếng Việt sang cú pháp NNKH Việt Nam, nghiên cứ đã lựa chọn một phương pháp tiếp cận hiện dại. Việc sử dụng các mô hình Transformer tiên tiến được tiền huấn luyện chuyên cho tiếng Việt (BARTpho, ViT5) làm nền tảng đảm bảo mô hình có sự am hiểu sâu về ngôn ngữ nguồn.

2.5. Chỉ số đánh giá kết quả đầu ra

Trong nghiên cứu, sẽ có 2 kết quả đầu ra chính cần đánh giá. Thứ nhất, là đánh giá dữ liệu tăng cường dùng chỉ số TTR và cosine similarity. Thứ hai, là đánh giá kết quả văn bản dịch trong mô hình dịch máy gồm chỉ số BLEU và WER,

2.5.1. Các chỉ số đánh giá dữ liệu tăng cường

Chất lượng dữ liệu tăng cường được đánh giá qua hai chỉ số: Tỷ lệ từ loại (Tye Token Ratio, TTR) và độ tương đồng ngữ nghĩa (cosine similarity) giữa câu gốc và câu mới.

2.5.1.1. Chỉ số đánh giá TTR – Type Token Ratio

TTR đo độ phong phú từ vựng của tập dữ liệu bằng tỉ lệ số từ loại (unique tokens) trên tổng số từ (Reviriego, Conde, Merino-Gómez, Martínez, & Hernández, 2024). Tính bằng công thức:

$$TTR = \frac{t}{n}$$

Với t là số từ duy nhất, n là tổng số từ trong văn bản. Một giá trị TTR cao hơn sẽ cho thấy tập dữ liệu có vốn từ đa dạng hơn. Trong thực nghiệm, chỉ số TTR sẽ được tính cho tập câu gốc và tập câu tăng cường, so sánh xem vốn từ có được mở rộng hay không.

2.5.1.2. Chỉ số đánh giá Cosine Similarity

Chỉ số này đánh giá mức độ tương đồng về ngữ nghĩa giữa câu gốc và câu được sinh ra. Cách tính: trước hết chuyển mỗi câu thành vector đặc trưng (chẳng hạn lấy embedding của token từ mô hình BERT được huấn luyện sẵn), ký hiệu là V_0 (vector câu gốc), và V_a (vector câu tăng cường). Cosine similarity được tính theo công thức (Uda, Matsumoto, & Yoshida, 2024):

$$\cos(V_0, V_a) = \frac{V_0 . V_a}{\| V_0 \| . \| V_a \|}$$

Giá trị cosine càng gần 1 nghĩa là hai vector càng đồng hướng, tức câu mới giữ nội dung tương tự như câu gốc. Trong nghiên cứu của (Uda et al., 2024) về văn bản Nhật Bản, cosine similarity giữa các embedding của câu gốc và câu tăng cường được sử dụng để đánh giá mức độ bảo toàn ngữ nghĩa.

2.5.2. Các chỉ số đánh giá bản dịch

Để đánh giá và so sánh hiệu suất của các mô hình dịch máy một cách khách quan, việc sử dụng các độ đo định lượng là cần thiết. Các chỉ số này cung cấp một thước đó chuẩn hóa, cho phép xác định mô hình nào hoạt động hiệu quả nhất và đưa ra quyết định lựa chọn.

2.5.2.1. Chỉ số đánh giá BLEU - Bilingual Evaluation Understudy

BLEU là một độ đo phổ biến, được sử dụng để đánh giá sự tương đồng về từ vựng và cụm từ (n-gram) giữa bản dịch máy và bản dịch tham khảo. Chỉ số này dựa trên việc đo lường chính xác n-gram, trong đó các nghiên cứu như của (Beidas, Ghaddar, Mohi, Ahmad, & Abed, 2025) đã sử dụng nó để so sánh hiệu suất dịch của mô hình ngôn ngữ lớn dựa trên sự chồng chéo về từ vựng. Điểm BLEU càng cao cho thấy bản dịch của máy càng giống bản dịch của người về mặt từ vựng và cấu trúc cụm từ.

BLEU là một trong những chỉ số được sử dụng rộng rãi trong lĩnh vực dịch máy và thường có độ tương quan tốt với đánh giá con người. Tuy nhien, nhược điểm của nó là có xu hướng ưu tiên sự chính xác về mặt từ vựng hơn là tính mạch lạc hay ngữ nghĩa của câu.

2.5.2.2. Chỉ số đánh giá WER – Word Error Ratio

WER tính toán số lượng thao tác chỉnh sửa tối thiểu – bao gồm thay thế (Substitution), xóa (Deletion), và chèn (Insertion), cần thiết để biến câu do máy dịch thành câu tham khảo. Tỷ lệ này được tính bằng tổng số lỗi chia cho trong số từ trong câu tham khảo. Điểm WER càng thấp, bản dịch càng chính xác.

Công thức tính WER như sau (Trần Vũ Hoàng, 2025):

$$WER = \frac{S + D + I}{N}$$

Trong đó:

+ S (substitutions): số từ thay thế – số từ đầu ra bị thay thế so với bản tham chiếu

+ D (deletions): số từ xóa – số từ bị bỏ sót

+ I (insertions): số từ chèn – số bị thêm vào nhưng không có trong bản tham chiếu

+ N (total words in reference): tổng số từ trong bản tham chiếu

Giá trị WER càng thấp thì hiệu suất mô hình càng tốt.

WER là chỉ số đánh giá được đề cập trong đề tài này, đặc biệt hữu ích trong việc đo lường độ chính xác của tác vụ chuyển đổi cấu trúc ngữ pháp. Một nghiên cứu liên quan chứng minh sức mạnh của phương pháp tiếp cận trên Transformer khi mô hình ViT5 đạt được tỷ lệ WER chỉ 2%, một sự cải thiện vượt bậc so với tỷ lệ 60,68% của phương pháp truyền thống (Trần Vũ Hoàng, 2025). Sự chênh lệch là đến từ phương pháp truyền thống "chứa quá nhiều bước xử lý trung gian,.. gây ra việc tích lũy và tăng dần sai số theo từng bước" (Trần Vũ Hoàng, 2025)

Bảng 2-2. Chỉ số đánh giá bản dịch được sử dụng

Chỉ số	Mô tả	Ứng dụng trong nghiên cứu
BLEU	Đo lường độ chính xác của các cụm từ (n-gram) trong bản dịch so với bản tham khảo	Đánh giá sự lưu loát và tương đồng từ vựng của bản dịch.
WER	Đo lường tỷ lệ lỗi từ, tính bằng tổng số lần thay thế, xóa, chèn từ	Chỉ số đánh giá độ chính xác của việc chuyển đổi cấu trúc câu.

Tóm lại, với việc xác định rõ ràng cơ sở lý thuyết cho module dịch máy và công cụ đánh giá hiệu suất, chương 2 đã trình bày rõ ràng. Các chương tiếp theo sẽ trình bày chi tiết về phương pháp thực nghiệm, kết quả đạt được và phân tích sâu hơn về hiệu quả.

CHƯƠNG 3. QUY TRÌNH TRIỂN KHAI

3.1. Tổng quan kiến trúc hệ thống

Nghiên cứu này nhằm xây dựng một hệ thống hỗ trợ giao tiếp để giai quyết rào cản trong giao tiếp ở người nói tiếng Việt và người khiếm thính, đặc biệt trong lĩnh vực cà phê. Kiến trúc hệ thống nghiên cứu được triển khai theo một quy trình, gồm hai modules chính, hoạt động tuần tự.

3.1.1. Kiến trúc hệ thống

Module 1: Nhận dạng giọng nói

Mục tiêu: Chuyển đổi chính xác đầu vào là giọng nói tiếng Việt thành văn bản thông thường.

Vai trò: Đầu ra của Module này đóng vai trò là đầu vào cho Module Dịch máy (Module 2)

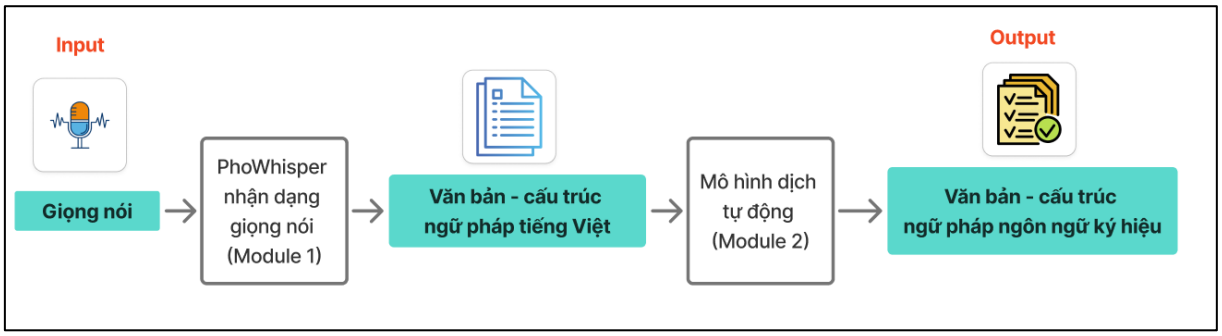
Mô hình tiềm năng được lựa chọn là phoWhisper, một mô hình mã nguồn mở dựa trên kiến trúc Whisper của OpenAI, được tinh chỉnh cho tiếng Việt.

Module 2: Dịch máy

Mục tiêu: Dịch văn bản từ đầu ra của Module 1 sang biểu diễn văn bản của NNKH Việt Nam một cách chính xác về mặt cú pháp và ngữ nghĩa.

Luận văn sẽ so sánh hiệu năng của cả ba mô hình ngôn ngữ lớn tiên tiến cho tiếng Việt, được huấn luyện trên bộ dữ liệu nền tảng Vie-VSL-10k và tinh chỉnh trên dữ liệu chuyên biệt về dịch vụ cà phê. Ba mô hình tiềm năng: ViT5 – base, ViT5 – large, BARTpho.

Đầu ra của hệ thống trong nghiên cứu này là văn bản biểu diễn đúng cú pháp của NNKH Việt Nam. Dạng biểu diễn này không chỉ giúp truyền tải đúng ngữ nghĩa mà còn là đầu vào chuẩn hóa cho các mô hình diễn họa 3D trong lương lai. Mục tiêu tổng thể của hệ thống là chuyển đổi ***giọng nói tiếng Việt thành văn bản biểu diễn đúng cú pháp của NNKH Việt Nam***.



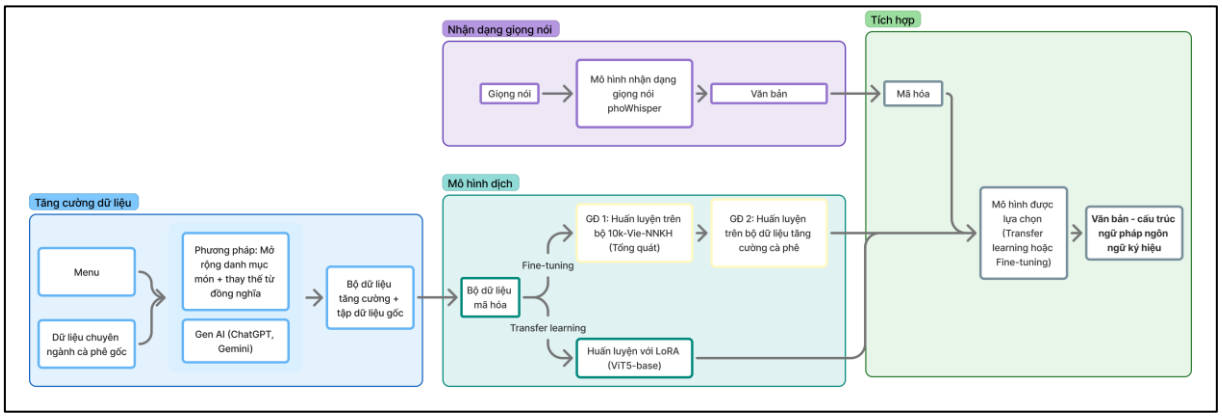
Hình 3.1. Sơ đồ tổng quan hệ thống. Hệ thống chuyển đổi từ giọng nói sang văn bản tiếng Việt (module 1) và cuối cùng là văn bản cú pháp ngôn ngữ ký hiệu (module 2)

Về tổ chức thư mục:

+ Thư mục code: Bao gồm script cho tăng cường dữ liệu, transfer learning với ViT5, fine-tuning với ViT5 base, ViT5 large, BARTpho, finetuning lần 2 với ViT5 base

+ Dữ liệu: dữ liệu kế thừa 10000 cặp câu, dữ liệu liên quan đến Coffee gốc, và dữ liệu sinh ra nhờ tăng cường.

Tất cả công việc xử lý, huấn luyện, đánh giá và triển khai được thực hiện trên Google Colab, tận dụng GPU miễn phí để xử lý mô hình ngôn ngữ lớn.



Hình 3.2. Chi tiết quy trình nghiên cứu

3.1.2. Thu thập dữ liệu

Để xây dựng mô hình dịch tiếng Việt sang NNKH, cầncặp tập dữ liệu song ngữ đủ lớn. Phương pháp thu thập như sau:

+ Ban đầu, bộ dữ liệu gồm 10.000 cặp câu về chủ đề đời sống bình thường được kế thừa từ nghiên cứu trước đó của tiến sĩ Nguyễn Thị Bích Diệp.

+ Thu thập bổ sung dữ liệu chuyên ngành dịch vụ quán cà phê. Các mẫu ngôn ngữ gốc được ghi lại thực tế tại quán cà phê Angle có sự tham gia của người khiếm thính.

Kết quả thu được 146 câu tiếng Việt và NNKH mô tả các tình huống gọi món, phục vụ tại quán. Bên cạnh đó, menu của quán gồm 7 danh mục (cà phê, trà, sữa chua,..) với khoảng 50 món thông dụng được tổng hợp lại.

Trong nghiên cứu này, menu đồ uống được lưu dưới dạng file json (khóa "danh mục": "các món") và tập câu gốc được lưu trong file CSV. Tuy nhiên, với chỉ 146 câu gốc, lượng dữ liệu này rất hạn chế cho huấn luyện mô hình dịch. Trên thực tế, các nghiên cứu về NLP cho thấy tăng cường dữ liệu giúp cải thiện đáng kể hiệu suất, đặc biệt kích thước tập dữ liệu gốc nhỏ. Do đó, giai đoạn tiếp theo là tăng cường tập dữ liệu gốc bằng kỹ thuật xử lý văn bản và sinh văn bản mới.

3.2. Nhận dạng giọng nói và chuyển thành văn bản (Speech to Text – Module 1)

3.2.1. Mô tả bài toán

Trong giao tiếp hàng ngày, giọng nói là phương thức truyền tải thông tin cơ bản và nhanh chóng nhất. Tuy nhiên, đối với cộng đồng người khiếm thính, việc tiếp nhận thông tin qua kênh thính giác là một rào cản, gây khó khăn trong việc tương tác người bình thường.

Để giải quyết vấn đền này, hệ thống cần một chức năng đóng vai trò cầu nối ghi nhận tín hiệu âm thanh từ lời nói của khách hàng và chuyển đổi chúng thành văn bản hiện thị trên màn hình.

3.2.2. Đề xuất giải pháp - áp dụng mô hình phoWhisper vào nhận diện giọng nói

Để tối ưu hóa khả năng nhận diện tiếng Việt, nghiên cứu này đề xuất sử dụng PhoWhiper – một mô hình nhận dạng giọng nói tiên tiến được phát triển dựa trên kiến trúc Transformer (Whisper của Open AI) được tinh chỉnh chuyên biệt cho dữ liệu tiếng Việt.

Thay vì huấn luyện lại mô hình từ đầu, nghiên cứu này sẽ kế thừa các kết quả đã được chứng minh từ các công trình nghiên cứu trước đó của PhoWhisper. Việc này giúp tận dụng khả năng xử lý ngôn ngữ, đồng thời tiết kiệm tài nguyên.

3.2.3. Đánh giá hiệu năng mô hình PhoWhisper

Hiệu suất của mô hình được đánh giá trên chỉ số Tỷ lệ lỗi từ (Word error rate – WER). Đây là một trong những tiêu chuẩn trong bài toán nhận diện giọng nói, giá trị WER càng thấp đồng nghĩa với độ chính xác mô hình càng cao.

Bảng dưới đây trình bày kết quả so sánh thực nghiệm giữ PhoWhisper và các mô hình phổ biến khác tại thời điểm thực hiện nghiên cứu đó (như wav2vec2) trên bộ dữ liệu tiếng Việt (CMV-Vi, VIVOS, VLSP)

Bảng 3-1. So sánh tỷ lệ lỗi từ của các mô hình nhận dạng giọng nói tiếng Việt

Mô hình	**WER (%)**
Whisper-small	31,89
wav2vec2-base-vietnamese-250h	24,13
PhoWhisper-small (chọn)	**23,61**

Nguồn: Trần Vũ Hoàng, L. Q. Đ., Huỳnh Đình Hiệp, Đoàn Mạnh Cường. (2025). Thiết kế xây dựng phần mềm phiên dịch ngôn ngữ ký hiệu tiếng Việt. *TNU Journal of Science and Technology*

LSVSC (Large-scale Vietnamese Speech Corpus) là kho dữ liệu tiếng nói Tiếng Việt chuẩn. Có quy mô 100.5 giờ âm thanh sạch (đã gỡ băng thủ công). Với chủ đề đa dạng, nhưng Tin tức (78.5%) chiếm đa số.

Từ phân tích kết quả cho thấy có một sự khác biệt đáng kể về hiệu suất giữa các kiến trúc. PhoWhisper-small đạt được tỷ lệ lỗi thấp nhất trên bộ dữ liệu. Việc mô hình này duy trì hiệu suất hàng đầu trên bộ dữ liệu đa dạng chứng tỏ độ chính xác cao mà còn khả năng tổng quát hóa mạnh. Đây là một yêu cầu quan trọng để xây dựng một hệ thống hoạt động ổn định.

```
# Mô hình nhận diện giọng nói:
model_name = 'vinai/PhoWhisper-small'
stt_processor = WhisperProcessor.from_pretrained(model_name)
```

Hình 3.3. Code gọi mô hình phoWhisper-small nhằm nhận diện giọng nói và chuyển thành dạng văn bản

Do đó, việc lựa chọn PhoWhisper làm module nhận dạng giọng nói là phù hợp, đảm bảo độ chính xác cho ứng dụng hỗ trợ người khiếm thính.

3.3. Tăng cường dữ liệu về dịch vụ quán cà phê

3.3.1. Mô tả bài toán và phương pháp tiếp cận

Do phạm vi nghiên cứu tập trung vào ngữ cảnh quán cà phê, bộ dữ liệu gốc gồm 146 câu về gọi món, có khích thước rất nhỏ. Lượng mẫu này chưa đủ để đảm bảo độ đa dạng cần thiết cho mô hình. Vì vậy, bước tăng cường dữ liệu là cần thiết.

Để giải quyết vấn đề thiếu hụt dữ liệu, hai phương pháp chính được sử dụng trong bài nghiên cứu này là:

(i) Thay thế từ đồng nghĩa (synonym replacement): Thay thế các đồ uống trong câu gốc bằng các món khác trong thực đơn để tạo ra các tổ hợp gọi món

(ii) Viết lại câu (paraphrasing): Thay đổi cấu trúc ngữ pháp hoặc sử dụng từ đồng nghĩa để diễn đạt lại câu lệnh nhưng vẫn giữ nguyên ý ban đầu.

3.3.2. Quy trình thực hiện tăng cường dữ liệu

Quá trình tăng cường dữ liệu được chuẩn bị và thực hiện với các thành phần sau:

- Dữ liệu đầu vào:
    - Bộ dữ liệu gốc: 146 câu lệnh về dịch vụ quán cà phê
    - Danh mục món: Gồm 7 nhóm đồ uống với hơn 50 món cụ thể được thu thập tại quán
- Công cụ sinh dữ liệu:
    - Nghiên cứu sử dụng các mô hình lớn để tự động sinh văn bản. Hai mô hình được lựa chọn ChatGPT (phiên bản -4o mini) và Gemini (phiên bản 2.5 Flash). Đây là phiên bản miễn phí, giúp tối ưu hóa chi phí nhưng vẫn đảm bảo hiểu và sinh dữ liệu ổn.
- Kỹ thuật prompt:

- Các câu lệnh gửi đến mô hình được thiết kế chi tiết, gồm các thành phần: bối cảnh (quán cà phê), vài trò của AI (tạo dữ liệu mới), cung cấp dữ liệu mẫu và danh sách menu.

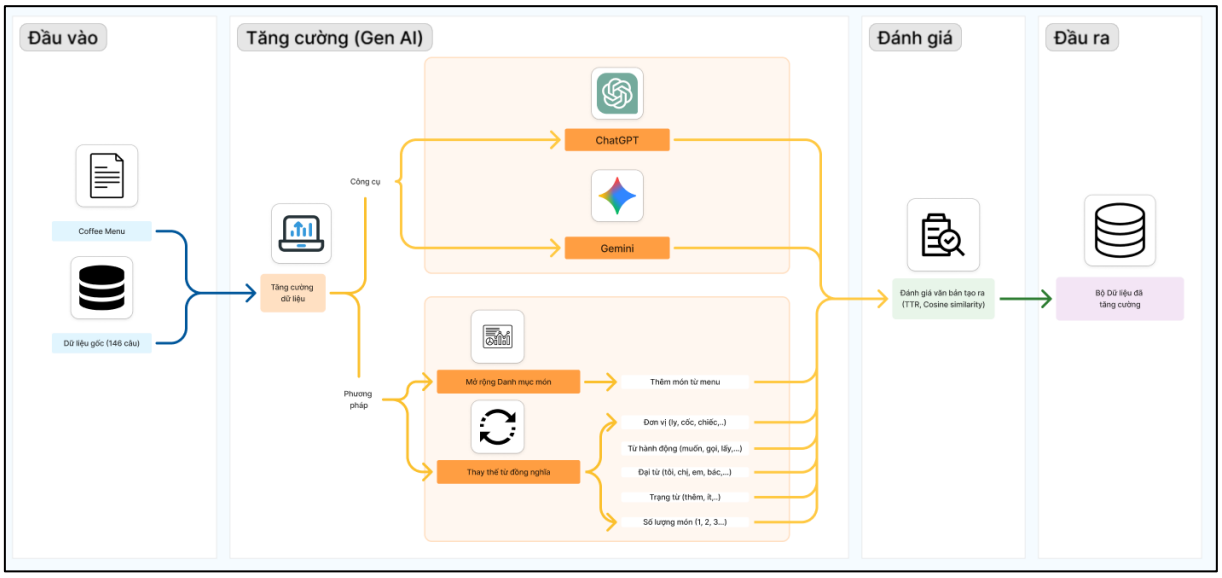
- Phương pháp tăng cường dữ liệu:
  - Mở rộng danh mục món: Thêm các món ngẫu nhiên từ menu vào trong câu
  - Thay thế từ đồng nghĩa: biến đổi từ vựng để tăng sự đa dạng, gồm thay đổi đơn vị, từ hành động, đại từ, trạng từ, số lượng.

Bảng 3-2. Một số câu gọi món của khách hàng được thu thập tại quán cà phê

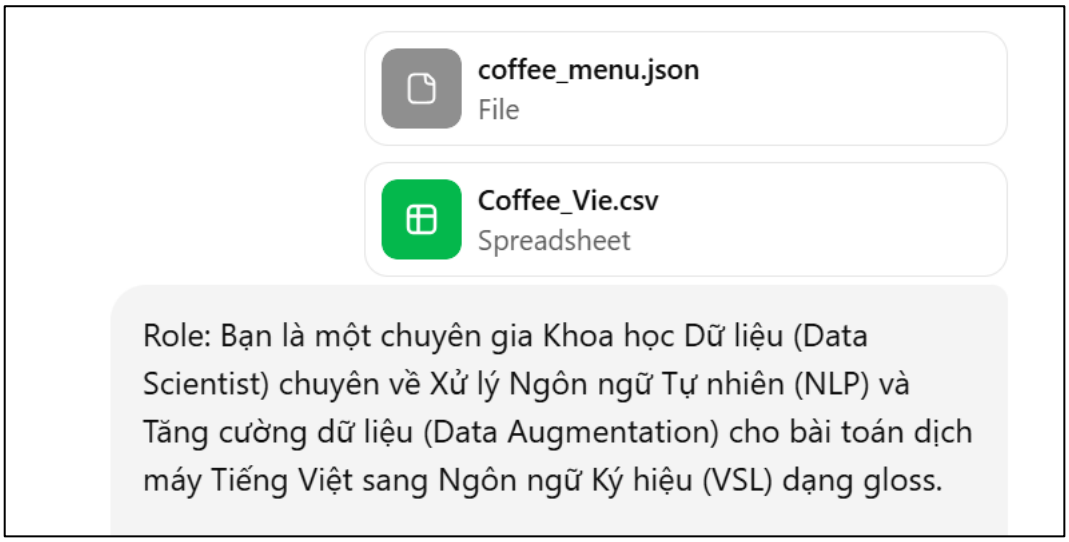
STT	vietnamese
1	Tôi muốn một ly sữa chua dâu
2	Cho mình một nước ép thơm mát lạnh.
3	Tôi muốn một ly cà phê sữa, ít đá
4	Cho tôi một ly cà phê sữa, một ly nước ép cóc
5	Lấy cho tôi một ly trà gừng nóng.
6	Lấy giúp mình một ly trà xoài ít đường.

(còn tiếp)

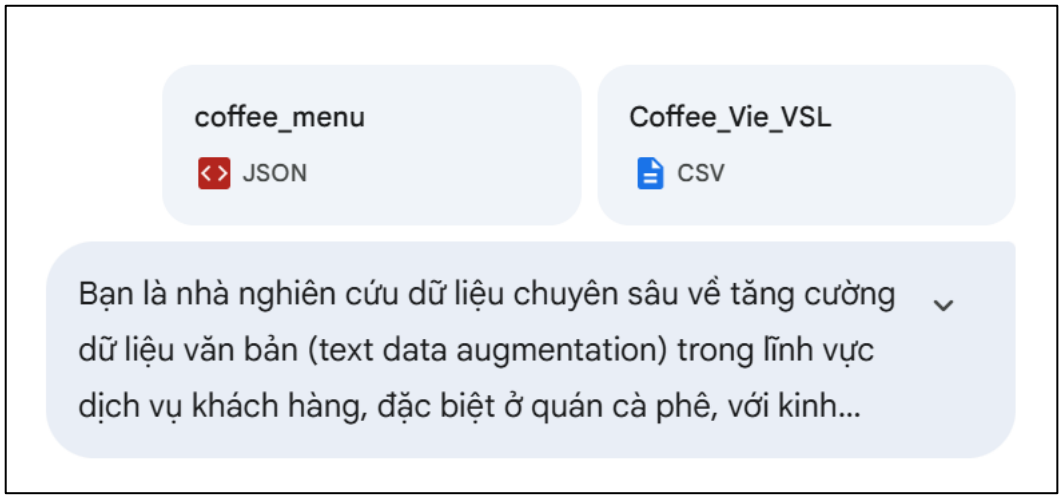
Hình 3.4. Menu tại quán cà phê được sử dụng nhằm hỗ trợ tăng cường dữ liệu



Hình 3.5. Quy trình tăng cường dữ liệu văn bản



Hình 3.6. Câu lệnh tăng cường dữ liệu, gửi file dữ liệu gốc, danh mục đồ uống để GPT-4o mini tạo dữ liệu mới



Hình 3.7. Câu lệnh tăng cường dữ liệu, gửi file dữ liệu gốc, danh mục đồ uống để Gemini 2.5 Flash tạo dữ liệu mới

Bảng 3-3. Phiên bản ChatGPT và Gemini được sử để tăng cường dữ liệu

	ChatGPT	**Gemini**
Phiên bản	GPT-4o mini (miễn phí)	2.5 Flash (miễn phí)

Bạn là một chuyên gia Khoa học Dữ liệu (Data Scientist) chuyên về Xử lý Ngôn ngữ Tự nhiên (NLP) và Tăng cường dữ liệu (Data Augmentation) cho bài toán dịch máy Tiếng Việt sang Ngôn ngữ Ký hiệu (VSL) dạng gloss.

Tôi có một tập dữ liệu gốc gồm 146 cặp câu về giao tiếp gọi món tại quán cà phê.
- Cột input: Câu tiếng Việt nói thường.

Hãy tạo ra 500 cặp câu mới bằng cách áp dụng CHỈ MỘT phương pháp thay đổi cụ thể lên dữ liệu gốc.

Task Instructions (Chọn 1 trong 4 nhiệm vụ dưới đây để thực hiện):
[TASK 1: Thay thế Từ đồng nghĩa]
- Thay thế các từ chỉ vật dụng/đơn vị trong câu gốc bằng từ đồng nghĩa (ví dụ: "ly" <-> "cốc", "dùng" <-> "uống", "muốn" <-> "lấy").
- Lưu ý: Giữ nguyên cấu trúc câu và các thành phần khác.

[TASK 2: Thay đổi Số lượng]
- Thay đổi các con số trong câu gốc (ví dụ: "1 ly" thành "2 ly", "3 cốc", "5 phần").

[TASK 3: Thêm món (Câu ghép)]
- Mở rộng câu đơn thành câu gọi nhiều món (ví dụ: "Cho 1 cà phê đen" -> "Cho 1 cà phê đen và 1 nước ép cam").

[TASK 4: Thay đổi Chủ ngữ/Đại từ]
- Thay đổi đại từ nhân xưng trong câu tiếng Việt (ví dụ: "Tôi" -> "Em", "Cháu", "Bác", "Chú", "Mình").

Input Data: Coffee.csv
Output Format: Trả về kết quả dưới dạng CSV

Hình 3.8. Câu lệnh dùng để tăng cường dữ liệu văn bản

3.3.3. Đánh giá chất lượng dữ liệu tăng cường

Chất lượng dữ liệu tăng cường được đánh giá qua hai chỉ số hai chỉ số: Tỷ lệ từ loại (Type Token Ratio, TTR) và độ tương đồng ngữ nghĩa (cosine similarity) giữa câu gốc và câu mới.

Về chỉ số TTR, chỉ số này đo độ đa dạng từ vựng của tập dữ liệu. Nếu TTR của bộ dữ liệu tăng cường cao hơn bộ gốc, điều này chứng tỏ vốn từ vựng được mở rộng, tránh lặp từ.

Về chỉ số cosine similarity, sử dụng thư viện xử lý ngôn ngữ tự nhiên, trích xuất vector embedding, tính toán độ tương đồng cosine giữa câu gốc và câu sinh ra. Chỉ số này ở mức cao, gần bằng 1 thì đảm bảo câu mới sinh ra bám sát câu gốc, không bị lệch nghĩa.

```
def calculate_ttr(sentences):
  if not sentences: return 0

  all_text = " ".join(sentences)
  tokens = all_text.split()

  if len(tokens) == 0: return 0

  types = set(tokens)
  return len(types) / len(tokens)
```

Hình 3.9. Code về xây dựng chỉ số đánh giá TTR cho dữ liệu tăng cường

```
def calculate_cosine_sim(original_docs, augmented_docs):
  if not original_docs or not augmented_docs:
    return 0

  combined_corpus = original_docs + augmented_docs
  vectorizer = TfidfVectorizer().fit(combined_corpus)

  # Biến đổi sang vector
  tfidf_original = vectorizer.transform(original_docs)
  tfidf_augmented = vectorizer.transform(augmented_docs)

  # Tính similarity của các câu Augmented so với TẤT CẢ câu Original
  # Kết quả là ma trận (n_aug x n_orig)
  cosine_sim_matrix = cosine_similarity(tfidf_augmented, tfidf_original)
```

Hình 3.10. Code về xây dựng chỉ số đánh giá Cosine Similarity cho dữ liệu tăng cường

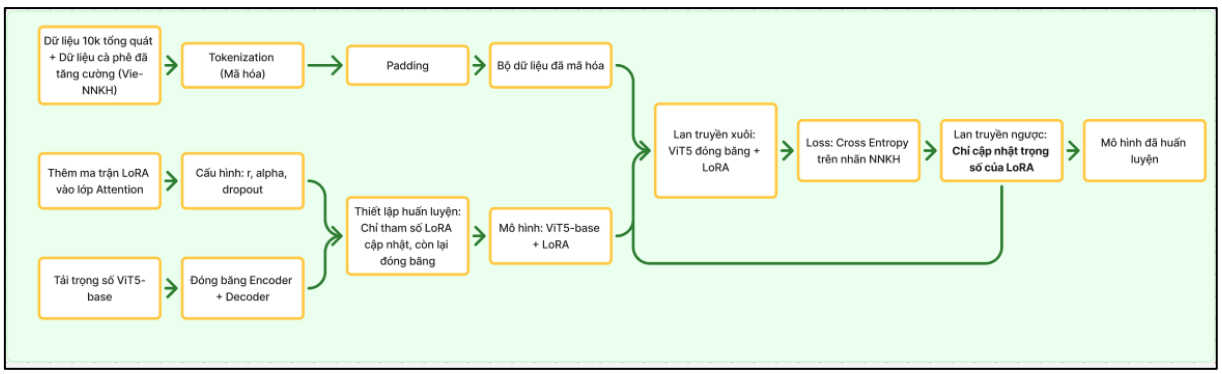
Tóm lại, sự kết hợp giữa hai chỉ số này giúp đảm bảo dữ liệu tăng cường có sự đa dạng về từ vựng, và vừa đảm bảo sự nhất quán nội dung so với dữ liệu thực tế.

3.4. Mô hình chuyển đổi văn bản tiếng Việt sang văn bản đúng cú pháp NNKH Việt Nam (Translation Machine – Module 2)

3.4.1. Sử dụng phương pháp transfer learning

Để giải quyết bài toán dịch với nguồn dữ liệu hạn chế, nghiên cứu này áp dụng phương áp Transfer Learning. Thay vì huấn luyện mô hình từ đầu (vốn đòi hỏi hàng triệu cặp câu), mô hình ngôn ngữ lớn được huấn luyện từ trước sẽ được sửu dụng, cụ thể là mô hình ViT5-base.

Quy trình thực hiện cụ thể như sau:



Hình 3.11. Quy trình huấn luyện mô hình ViT5-base thêm ma trận LoRA trong phương pháp transfer learning

3.4.1.1. Chuẩn bị và chia dữ liệu

Dữ liệu đầu vào: Tổng hợp bộ dữ liệu gồm 11.691 cặp câu song ngữ (theo cú pháp tiếng Việt và cú pháp NNKH). Trong đó, gồm 10.000 dòng dữ liệu nền tảng (kế thừa từ nghiên cứu trước về chủ đề tổng quát) và khoảng 1.500 câu chuyên về lĩnh vực dịch vụ cà phê (được thu thập và tăng cường trong nhiệm vụ 2).

Dữ liệu được chia ngẫu nhiên theo tỷ lệ 80:10:10 cho ba tập huấn luyện, kiểm thử và đánh giá. Tỷ lệ này được lựa chọn nhằm đảm bảo mô hình có đủ dữ liệu để học các quy luật ngữ pháp, đồng thời có tập dữ liệu riêng để cập nhật tham số và đánh giá, tránh hiện tượng overfitting.

Số lượng được phân bổ như sau:

Bảng 3-4. Phân chia dữ thành thành các tập train, val, test với phương pháp transfer learning

Tập	Số lượng	Tỷ lệ	Nhiệm vụ
Train	9351	80%	Huấn luyện trọng số
Val	1170	10%	Dùng để theo dõi quá trình học và early stopping
Test	1170	10%	Đánh giá sau khi huấn luyện xong

3.4.1.2. Tiền xử lý dữ liệu

Mô hình ViT5 coi mọi bài toán NLP là bài toán "Text-to-Text". Do đó, dữ liệu đầu vào được chuẩn hóa bằng cách thêm tiền tố để định hướng tác vụ cho mô hình

Input: "dịch tiếng Việt sang VSL:" + [câu tiếng Việt]

Max length: thiết lập độ dài tố đa cho chuỗi đầu vào và đầu ra là 128 token, phù hợp với độ dài trung bình của các câu giao tiếp trong bối cảnh trong dịch vụ cà phê.

```
# Tiền tố cho tác vụ dịch (rất quan trọng với T5)
prefix = "dịch tiếng Việt sang VSL: "
max_input_length = 128
max_target_length = 128
```

Hình 3.12. Thêm tiền tố trong quá trình huấn luyện mô hình ViT5

3.4.1.3. Cấu hình mô hình và kỹ thuật LoRA

Nghiên cứu sử dụng ViT5-base (khoảng 226 triệu tham số) làm nền tảng. Để tối ưu hóa tài nguyên tính toán, nghiên cứu này áp dụng kỹ thuật LoRA (Low-Rank Adaption).

Thay vì tinh chỉnh toàn bộ 226 triệu tham số của mô hình gốc, LoRA cho phép chỉ huấn luyện ma trận hạng thấp được chèn vào các lớp Attention.

Bảng 3-5. Cấu hình LoRA

Tên	**Giá trị**	**Mục đích**
Rank	16	Cân bằng khả năng biểu diễn và chi phí
Alpha	32	
Target modules	Q, v	Áp dụng LoRA vào lớp query và Value của cơ chế Attention, nơi chứa thông tin nhiều nhất

```
# Cấu hình LoRA
lora_config = LoraConfig(
    r=16,
    lora_alpha=32,
    target_modules=["q", "v"],  # Chỉ áp dụng LoRA cho lớp Query và Value
    lora_dropout=0.05,
    bias="none",
    task_type=TaskType.SEQ_2_SEQ_LM # Cực kỳ quan trọng cho T5
)
```

Hình 3.13. Code thiết lập cấu hình LoRA

Kết quả từ thiết lập này cho thấy số lượng tham số cần huấn luyện chỉ là 1.769.472, chiếm 0,77% tổng tham số mô hình.

3.4.1.4. Cấu hình tham số huấn luyện

Quá trình huấn luyện được thiết lập với các siêu tham số như sau:

Bảng 3-6. Giá trị siêu tham số trong quá trình huấn luyện mô hình ViT5 với phương pháp Transfer Learning

Siêu tham số	Giá trị	Ý nghĩa
Batch size	16	Kích thước ổn để chạy trên Google Colab
Learning rate	0.0001	Tốc độ học thường dùng cho việc cập nhật trọng số của mô hình ViT5, đảm bảo mô hình hội tự ổn định.
Epoch	10	Tuy nhiên, số vòng lặp có thể ít hơn vì thiết lập cơ chế dừng sớm
Early stopping patience	2	Trong quá trình huấn luyện, sau mỗi epoch, sẽ đánh giá Loss trên tập validation. Nếu sau hai lần liên tiếp, chỉ số Validation Loss không giảm thì huấn luyện dừng lại.

(còn tiếp)

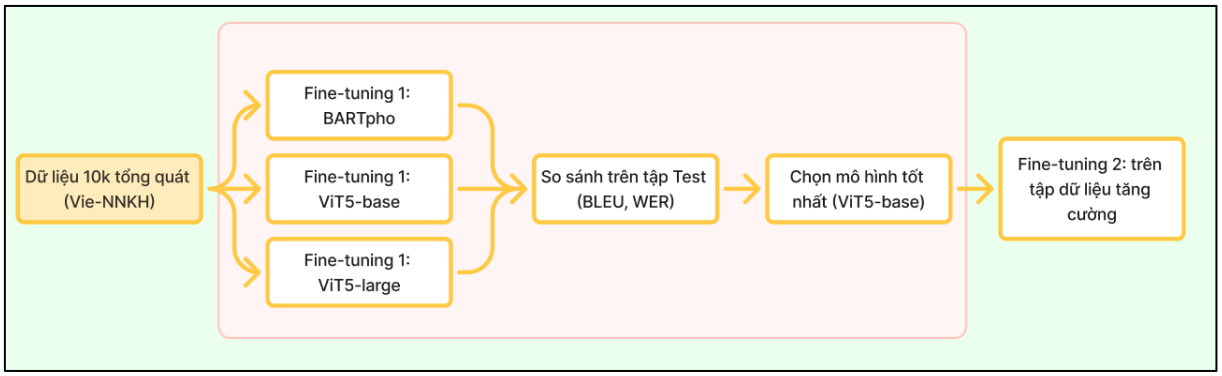
```
training_args = Seq2SeqTrainingArguments(
    output_dir="./vit5-vsl-translator",
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs= 10, # Đặt số epochs tối đa (ví dụ: 10),
    learning_rate=1e-4,
    weight_decay=0.01,
    predict_with_generate=True, # BẮT BUỘC để sinh text khi đánh giá
    logging_dir="./logs",
    logging_steps=100,

    # Cấu hình quan trọng cho Early Stopping
    eval_strategy="epoch",  # Đánh giá sau mỗi epoch
    save_strategy="epoch",        # Lưu model sau mỗi epoch
    load_best_model_at_end=True,  # Tự động tải model tốt nhất khi kết thúc
    metric_for_best_model="eval_loss",  # Theo dõi chỉ số WER
    greater_is_better=False,     # Vì eval_loss càng thấp càng tốt
    report_to = 'none',
    fp16 = True,
    push_to_hub=False
)
```

Hình 3.14. Code thiết lập cấu hình huấn luyện với mô hình ViT5-base với phương pháp transfer learning

3.4.2. Sử dụng phương pháp fine-tuning

Trong nghiên cứu này, phương pháp Fine-tuning được áp dụng, tức là cập nhật toàn bộ trọng số của các mô hình ngữ tiền huấn luyện thay vì chỉ đóng băng các lớp đặc trưng. Quá trình huấn luyện được chia làm giai đoạn nhằm tố ưu hóa khả năng thích nghi mô hình từ dữ liệu tổng quát sang dữ liệu chuyên ngành.



Hình 3.15. Quy trình huấn luyện 3 mô hình BARTpho, ViT5-base, ViT5-large với phương pháp fine-tuning

3.4.2.1. Giai đoạn fine-tuning lần 1

Mục tiêu của giai đoạn này là thiết lập nền tảng tri thức ngữ pháp và từ vựng cho các mô hình trước khi đi vvi vào dữ liệu cụ thể. Ba biến thể mô hình được sử dụng là: **BARTpho, ViT5-base và ViT5-large**.

a. Chuẩn bị và chia dữ liệu

Dữ liệu đầu vào là tập hợp 10.000 cặp câu song ngữ (văn bản theo cú pháp tiếng Việt – văn bản theo đúng cú pháp của ngôn ngữ ký hiệu), kế thừa từ công trình nghiên cứu của TS. Nguyễn Thị Bích Diệp. Để đảm bảo tính khách quan trong đánh giá, bộ dữ liệu được chia ngẫu nhiên theo tỷ lệ 80:10:10.

Bảng 3-7. Phân bố trong tập huấn luyện, kiểm thử, kiểm tra trong giai doạn Fine-tuning lần 1

Tập dữ liệu	**Số lượng mẫu**	**Tỷ lệ**

Tập huấn luyện	8.000	80%
Tập kiểm thử	1.000	10%
Tập kiểm tra	1.000	10%

b. Tiền xử lý

Do sự khác biệt về kiến trúc, quy trình tiền xử đầu vào được điều chỉnh cho phù hợp.

Với ViT5, mô hình coi mọi bài toán NLP là bài toán "Text-to-Text", vì vậy cần phải thêm tiền tố vào chuỗi đầu vào để mô hình diện diện tác vụ.

Với BARTpho, có khả năng nhận diện ngữ cảnh trực tiếp nên không yêu cầu thêm tiền tố đặc biệt.

```
# Tiền tố cho tác vụ dịch (rất quan trọng với T5)
prefix = "dịch tiếng Việt sang VSL: "
max_input_length = 128
max_target_length = 128
```

Hình 3.16. Thêm tiền tố trong quá trình huấn luyện mô hình ViT5

c. Cấu hình siêu tham số

Việc thiết lập siêu tham số đóng vai trò quyết định sự hội tụ của mô hình. Dựa vào thực nghiệm, siêu tham số được thiếp lập như sau:

Bảng 3-8. Cấu hình siêu tham số của ba mô hình (BARTpho, ViT5-base, ViT5-large) khi huấn luyện theo phương pháp fine-tuning

Siêu tham số	**BARTpho**	**ViT5-base**	**ViT5-large**
Learning rate	0.00001 (1e-5)	0.00005 (5e-5)	0.00005 (5e-5)
Batch size	16	16	4
Gradient accumulation	1	1	4

Effective batch size	16	16	16
Optimizer	AdamW	AdamW	AdamW_8bit
Metric	Evaluation loss	Evaluation loss	Evaluation loss

Xét về tốc độ học, có sự khác biệt giữa mô hình BARTpho và ViT5. Các mô hình dựa trên kiến trúc BART thường nhạy cảm hơn tốt độ học khi fine-tuning. Mức 0.00001 là an toàn để tránh hiện tượng mô hình quên kiến thức đã học trước đó. Đối với mô hình của ViT5 (0.00005), thường có khả năng chịu đựng tốt đọc lớn, thoát điểm cực tiểu địa phương, và khả năng học tổng quát hóa tốt.

Về batch size, batch size của mô hình ViT5-large (4) nhỏ hơn so với ViT5-base hoặc BARTpho (16). Vì ViT5-large có kích thước tham số lớn hơn rất nhiều (khoảng 770 triệu tham số) so với ViT5-base (220 triệu) và BARTpho, nên khi triển khai, bản thân trọng số của ViT5-large chiếm phần lớn dung lượng. Nên để mô hình có thể được huấn luyện, phải thiết lập một batch size khiêm tốn là 4.

Tuy nhiên, sự khác biệt giữa batch size giữa ViT5-large và ViT5-base, BARTpho có thể đem kết quả sai khi so sánh. Do đó, để đảm bảo sự hợp lý khi so sánh hiệu suất giữa các mô hình, cần thiết lập Effective batch size tương đương.

$$Effective\ batch\ size = Batch\ per\ device\ x\ Gradient\ accumulation$$

Khi đó:

+ ViT5-base/BARTpho: 16.1 =16

+ ViT5-large: 4.4 = 16

Dù ViT5-large chỉ xử lý 4 câu mỗi lần, nhưng nó sẽ tích lũy đạo hàm của 4 lần chạy rồi mới cập nhật trọng số một lần, tương đương khi chạy batch size = 16, đảm bảo sự công bằng giữa các mô hình.

```
model_ = [
    "VietAI/vit5-base",
    "VietAI/vit5-large",
    "vinai/bartpho-word"
]
```

Hình 3.17. Gọi ba mô hình ViT5-base, ViT5-large, BARTpho-word trong quá trình fine-tuning lần 1

```
per_device_train_batch_size = 4,
per_device_eval_batch_size = 4,
gradient_accumulation_steps = 4,
optim = 'adamw_8bit',
```

Hình 3.18. Thiết lập siêu tham số gradient accumlation ở mô hình ViT5-large trong finetung để đảm bảo sự công bằng khi so sánh hiệu suất với 2 mô hình còn lại

3.4.2.2. Giai đoạn fine-tuning lần 2: Tinh chỉnh trên dữ liệu cà phê đã tăng cường

Sau khi hoàn tất giai đoạn fine-tuning lần 1, sẽ tiến hành đánh giá của ba mô hình BARTpho, ViT5-base, ViT5-large dựa trên các chỉ số BLEU, WER. Kết quả cho thấy mô hình ViT5-base đạt kết quả tốt (sẽ thảo luận ở chương 4). Do đó, ViT5-base được chọn làm mô hình để tiếp tục giai đoạn fine-tuning lần 2.

Mục tiêu giai đoạn 2 là mô hình cần học được các từ ngữ đồ uống, cấu trúc trong bối cảnh quán cà phê mà không quên kiến thức từ giai đoạn 1.

a. Chuẩn bị dữ liệu

Dữ liệu đầu vào cho giai đoạn này là bộ dữ liệu dịch vụ cà phê, đã qua bước tăng cường, gồm 1.691 cặp câu. Mặc dù kích thước dữ liệu nhỏ hơn nhiều so dữ liệu 10.000 lúc đầu, nhưng bù lại tri thức chuyên ngành cao.

Vẫn tiếp tục chiến lược phân chia 80:10:10 cho tập huấn luyện, kiểm thử và kiểm tra.

Bảng 3-9. Phân bố tỷ lệ trong tập huấn luyện, kiểm thử, kiểm tra cho quá trình fine-tuning lần 2

Tập dữ liệu	Số lượng mẫu	Tỷ lệ
Tập huấn luyện	1352	80%
Tập kiểm thử	169	10%
Tập kiểm tra	170	10%

b. Cấu hình huấn luyện

Trong giai đoạn này, các bộ trọng số tốt nhất từ mô hình ViT5 ở giai đoạn fine-tuning lần 1 được khởi tạo lại.

Bảng 3-10. Cấu hình siêu tham số

Siêu tham số	Giá trị
Batch size	16
Tốc độ học	0.00001
Epoch	10

So với lúc fine-tuning ở giai đoạn 1, có sự thay đổi nhẹ về siêu tham số. Tốc độ học giảm xuống 0.00001. Vì khi huấn luyện trên một tập dữ liệu nhỏ, nếu vẫn tiếp tục giữ tốc độ học lớn như 0.00005 thì mô hình có xu hướng học nhanh để khớp với dữ liệu. Mức 0.00001 sẽ giúp mô hình tinh chỉnh nhỏ hơn, học tổng quát.

Tóm lại, mục tiêu ở giai đoạn này là đảm bảo mô hình cuối cùng hiểu được tiếng Việt tổng quát và hiểu được dữ liệu về dịch vụ quán cà phê.

```
model_path = "/content/drive/MyDrive/DUE - HỌC TẬP/YEAR 4 (2025-2026)/TTTN/Models/ViT5_base_time1/trained_model"
tokenizer = AutoTokenizer.from_pretrained(model_path)
model = AutoModelForSeq2SeqLM.from_pretrained(model_path)
```

Hình 3.19. Gọi mô hình ViT5-base tốt nhất để tiếp tục fine-tuning lần 2 trên bộ dữ liệu coffee đã tăng cường

```python
# Định nghĩa Training Arguments
training_args = Seq2SeqTrainingArguments(
    output_dir = '/content/drive/MyDrive/DUE - HỌC TẬP/YEAR 4 (2025-2026)/TTTN/Models/ViT5_finetune_augmented_coffee/model_checkpoints',
    eval_strategy = 'epoch',
    learning_rate = 1e-5,
    per_device_train_batch_size = 16,
    per_device_eval_batch_size = 16,
    weight_decay = 0.01,
    save_total_limit = 3,
    num_train_epochs = 10,
    predict_with_generate = True, # to generate text when evaluate
    fp16 = True,
    logging_dir = '/content/drive/MyDrive/DUE - HỌC TẬP/YEAR 4 (2025-2026)/TTTN/Models/ViT5_finetune_augmented_coffee/logs',
    logging_strategy = 'steps',
    logging_steps= 10,
    save_strategy = 'epoch',
    load_best_model_at_end = True,
    report_to = 'none',
    metric_for_best_model = 'eval_loss',
    greater_is_better = False
)
```

Hình 3.20. Thiết lập siêu tham số cho giai đoạn fine-tuning lần 2 với tốc độ học giảm so với fine-tuning lần 1

CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1. Kết quả Nhận dạng giọng nói (Speech to Text - Module 1)

Trong module đầu tiên, nghiên cứu sử dụng mô hình PhoWhisper-small để thực hiện chuyển đổi giọng nói thành văn bản. Đây là mô hình được kế thừa trực tiếp từ các nghiên cứu nền tảng trước đó mà không qua tinh chỉnh thêm, nhằm tận dụng khả năng tổng quát của nó trên bộ dữ liệu tiếng Việt.

Đặc điểm đầu ra:

+ Là chuỗi văn bản tiếng Việt (text)

Quá trình thực nghiệm được tiến hành bằng cách thu nhận tín hiệu âm thanh trực tiếp qua microphone và đánh giá khả năng xử lý theo thời gian thực. Dưới đây là các phân tích chi tiết.

4.1.1. Đánh giá độ chính xác qua kịch bản giao tiếp

Để kiểm chứng độ tin cậy mô hình, nghiên cứu đã thực hiện thử nghiệm trên các mẫu câu có độ dài khác nhau.

4.1.1.1. Kịch bản 1

Đầu vào: “Tôi học ở trường Đại học Kinh tế - Đại học Đà Nẵng”

Kết quả: Mô hình nhận diện chính xác 100% các từ trong câu, bao gồm các cụm danh từ riêng.

Mặc dù câu có độ dài tương đối, phoWhisper vẫn nắm bắt tốt và trả đầy đủ nội dung

Hình 4.1. Thử nghiệm nhận dạng giọng nói và chuyển thành văn bản bằng mô hình PhoWhisper với đầu vào "Tôi học ở trường Đại học kinh tế - Đại học Đà Nẵng"

4.1.1.2. Kịch bản 2

Đầu vào: "Thầy Chức rất tuyệt vời và rất nhiệt tình"

Kết quả: Hệ thống ghi nhận chính xác từng từ

Khả năng nhận dạng tốt, khẳng định sự phù hợp mô hình

Hình 4.2. Thử nghiệm nhận dạng giọng nói và chuyển thành văn bản bằng mô hình PhoWhisper với đầu vào "Thầy Chức rất tuyệt vời và rất nhiệt tình"

4.1.1.3. Kịch bản 3

Đầu vào: "Tôi gọi một ly cà phê đen"

Kết quả: Nhận diện chính xác yêu cầu gọi món

Đây là kịch bản quan trọng đối với mục tiêu đề tài. Kết quả cho thấy mô hình hoạt động hiệu quả với từ vựng cà phê, đảm bảo đầu vào chính xác cho module dịch.

Hình 4.3. Thử nghiệm nhận dạng giọng nói và chuyển thành văn bản bằng mô hình PhoWhisper với đầu vào "Tôi gọi một ly cà phê đen"

4.1.2. Đánh giá hiệu suất

Về tốc độ xử lý, kết quả thực nghiệm cho thấy thời gian xử lý trung bình để chuyển đổi 1 câu nói, độ dài trung bình sang văn bản dao động từ 1-2s. Trong bối cảnh giao tiếp, độ trễ này nằm trong chấp nhận được.

Về định dạng văn bản, cụ thể là viết hoa, dấu câu. Một hạn chế là đầu ra mô hình chưa thực hiện chuẩn hóa viết hoa đầu câu hoặc tên riêng. Tuy nhiên, trong phạm vi nghiên cứu, mục tiêu cốt lõi là chuyển đổi yêu cầu của người nói sang NNKH. Trong cấu trúc ngữ pháp NNKH, các yếu tố về hình thức như viết hoa chưa mang giá trị quyết định bằng nhận diện đúng từ vựng. Do đó, ưu tiên tốc độ và độ chính xác từ vựng quan trọng hơn về hình thức.

4.2. Kết quả tăng cường dữ liệu

Kết quả tăng cường dữ liệu sẽ xét trên hai khía cạnh, số lượng và chất lượng dữ liệu. Đầu tiên, sau một lệnh prompt, xét về số lượng câu tạo thành.

4.2.1. Đánh giá về số lượng và phân bố dữ liệu

4.2.1.1. Tổng quan về số lượng câu sinh ra

Thử nghiệm được tiến hành cùng một câu lệnh yêu cầu tạo 1.000 cặp câu với hai mô hình ChatGPT (phiên bản GPT-4o mini) và Gemini (phiên bản 2.5 Flash).

Bảng 4-1. Kết quả số lượng câu tạo thành của ChatGPT và Gemini với prompt tạo 1000 câu

	ChatGPT	**Gemini**
Số lượng cặp câu tạo thành	1.000	**545**

ChatGPT tạo đủ số lượng 1.000 với yêu cầu, Gemini tạo 545 câu, bé hơn so với yêu cầu. Có thể do Gemini 2.5 Flash, dùng miễn phí nên giới hạn trong số lượng câu tạo thành.

Bảng 4-2. Một số câu liên quan đến chủ đề cà phê do ChatGPT, Gemini tạo (có đồ uống theo menu, thay đổi số lượng món, gọi nhiều món cùng một lúc)

STT	Một số câu do ChatGPT tạo	Một số câu do Gemini tạo
1	Lấy mình một phần **bạc xỉu đá** được không	Mình cần **hai ly socola đá**
2	Lấy mình gọi một **sữa chua xoài** với ạ	Làm cho tôi **ba** ly **socala nóng**
3	Mình order một tách cà phê trứng nóng nha bạn	Tôi muốn **một matcha latte**
4	Làm ơn chuẩn bị lấy mình **một cà phê dừa**	Cho mình một ly **cà phê moca**
5	Lấy giúp mình **hai** ly **matcha latte** và một **cacao**	Mình gọi một ly cà phê **trứng đá**

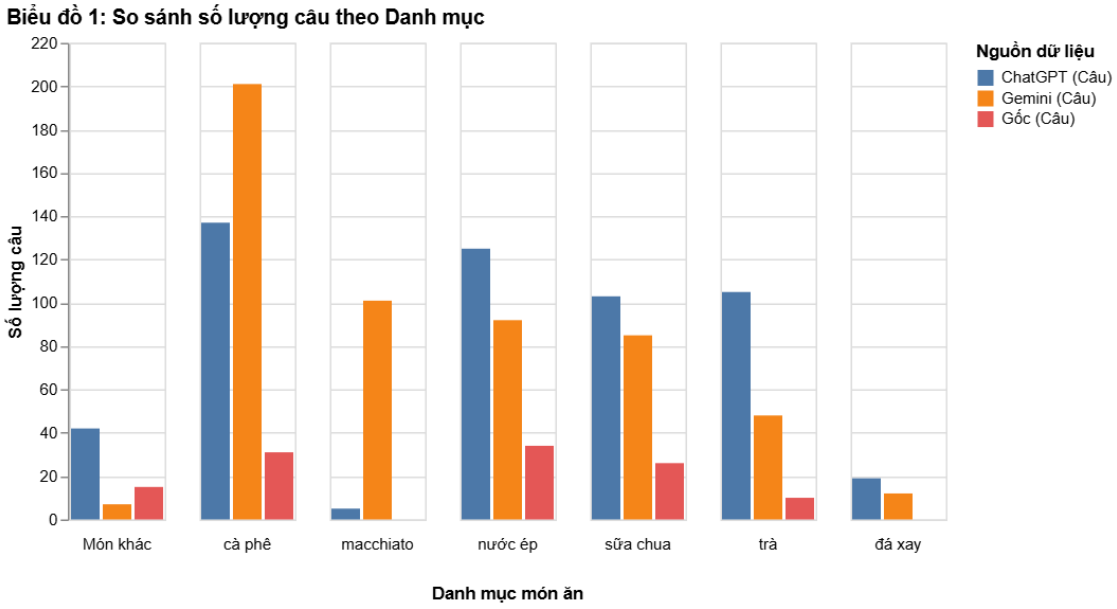
Các câu ChatGPT, Gemini tạo ra cơ bản đạt yêu cầu hãy phân tích sâu hơn về dữ liệu 2 công cụ này tạo ra.

4.2.1.2. Phân tích sự phân bố theo danh mục

Bảng 4-3. Số lượng câu theo Danh mục của ba bộ dữ liệu (Gốc, Gemini, ChatGPT)

STT	Danh mục	Gốc (Câu)	Gemini (Câu)	ChatGPT (Câu)
0	cà phê	31	201	135
1	macchiato	0	101	5
2	nước ép	34	92	125
3	sữa chua	26	85	103
4	trà	10	48	105
5	Món khác	15	7	42
6	đá xay	0	12	19

STT	Danh mục	Gốc (Câu)	Gemini (Câu)	ChatGPT (Câu)
	TỔNG CỘNG	116.0	546	536



Hình 4.4. Số lượng câu theo Danh mục của ba bộ dữ liệu (Gốc, ChatGPT, Gemini)

Xét về mở rộng miền dữ liệu, dữ liệu gốc tập chung chủ yếu ở cà phê (31% và nước ép (34%) và ở danh mục macchiato và đá xay là 0 câu. Trong khi tăng cường, ChatGPT và Gemini đã tạo mới được 2 danh mục sản phẩm này vì do lúc tăng cường, có bổ sung đầy đủ menu với 7 danh mục – hơn 50 món. Cụ thể, ChatGPT tạo số câu về macchiato và đá xay lần lượt là 101 và 12 câu, trong khi Gemini lần lượt là 8 và 20 câu. Có thể thấy rằng, trong lúc tạo ra miền mới, ChatGPT và Gemini đưa ra trọng số khác nhau. Nhìn chung, ChatGPT và Gemini tốt trong mở rộng miền trong tăng cường dữ liệu.

Xét về phân bổ số lượng câu trong danh mục, dữ liệu gốc bị tập trung về cà phê (31%) và nước ép (34%). Khi tăng cường dữ liệu, Gemini có hướng bị ảnh hưởng bởi dữ liệu gốc, cà phê chiếm 201 câu (37%). Ngược lại, ChatGPT có xu hướng cân bằng danh mục, cụ thể số lượng câu trong cà phê, nước ép, trà, sữa chua lần lượt dao động trong 110 – 135 câu. Tóm lại, Gemini học sự phân bổ trong dữ liệu gốc, ChatGPT có xu hướng tạo dữ liệu cân bằng hơn.

4.2.2. Đánh giá về chất lượng dữ liệu

Tiếp theo, để đánh giá chất lượng dữ liệu mới tạo thành của ChatGPT và Gemini, sử dụng hai chỉ số là Type Token Ratio (TTR) – đo sự đa dạng từ vựng và Cosine Similarity – đo sự giống nghĩa với dữ liệu gốc. - TTR (Type-Token Ratio): Càng cao → Từ vựng càng phong phú/đa dạng.

Cosine Similarity: Càng gần 1 → Ngữ nghĩa càng sát với tập dữ liệu gốc.

Bảng 4-4. Đánh giá chất lượng dữ liệu tăng cường (TTR và Consine similarity)

	Danh mục	TTR Gốc	TTR Gemini	TTR ChatGPT	Cosine Sim (Gemini vs Gốc)	Cosine Sim (ChatGPT vs Gốc)
0	cà phê	0.1541	0.0427	0.0835	0.1520	0.1834
1	macchiato	0.0000	0.0510	0.4154	0.0000	0.0000
2	nước ép	0.2069	0.0661	0.0922	0.1696	0.1798
3	sữa chua	0.2111	0.0628	0.0918	0.1859	0.1911
4	trà	0.3465	0.1257	0.0880	0.1700	0.1687
5	Món khác	0.2199	0.2877	0.1261	0.3383	0.2910
6	đá xay	0.0000	0.3049	0.2332	0.0000	0.0000

Xét khía cạnh đa dạng tự vựng (TTR), nhìn chung bộ dữ liệu gốc cao hơn 2 bộ dữ liệu tăng cường vì chỉ số TTR ở bộ dữ liệu gốc cao hơn 2 bộ dữ liệu còn lại. Điều này có nghĩa là bộ dữ liệu gốc có số lượng ít nhưng lượng từ vựng phong phú, còn ChatGPT và Gemini hơi bị lặp về từ vựng.

Ở hai danh mục macchiato và đá xay, nơi dữ liệu gốc không có thì ChatGPT và Gemini tạo rất tốt. Xét về độ đa dạng từ vựng – chỉ số TTR, ChatGPT tạo đa dạng từ vựng trong macchiato hơn Gemini (0.41 > 0.05), tuy nhiên ở danh mục đá xay thì ngược lại, Gemini đa dạng hơn ChatGPT (0.3 > 0.2).

Xét về ngữ nghĩa so với câu gốc (Cosine similarity), chỉ số hai bên ChatGPT và Gemini khá thấp, dao động 0 đến 0.34, có nghĩa là dữ liệu được tạo mới này hơi xa ngữ nghĩa với câu gốc. Tuy nhiên, điều này là hợp lý vì các câu trong bộ dữ liệu gốc do tự thu thập tại địa điểm khảo sát nên chất lượng sẽ kém, còn bộ dữ liệu tăng cường được thêm danh mục menu quán nên cách xa dữ liệu gốc sẽ là hợp lý.

4.2.3. Tổng kết quá trình tăng cường dữ liệu

Bảng 4-5. Kết quả số lượng câu sau khi tăng cường bằng ChatGPT và Gemini

	Dữ liệu coffee ban đầu	**Dữ liệu cuối cùng sau khi tăng cường**	
Số lượng câu	146	1691 câu	Tăng gấp 10 lần

Tóm lại, ChatGPT và Gemini đều tạo ra dữ liệu tăng cường tốt. Từ 146 câu ban đầu đã tăng cường lên tổng cộng là 1691 câu liên quan đến coffee, tăng gấp 10 lần so dữ liệu đầu.

4.3. Kết quả dịch máy (Machine translation - module 2)

Trong nhiệm vụ này, mục tiêu là xây dựng mô hình có khả năng chuyển đổi câu tiếng Việt sang câu theo đúng cú pháp của NNKH.

4.3.1. Kết quả thực nghiệm mô hình ViT5-base với phương pháp Transfer learning

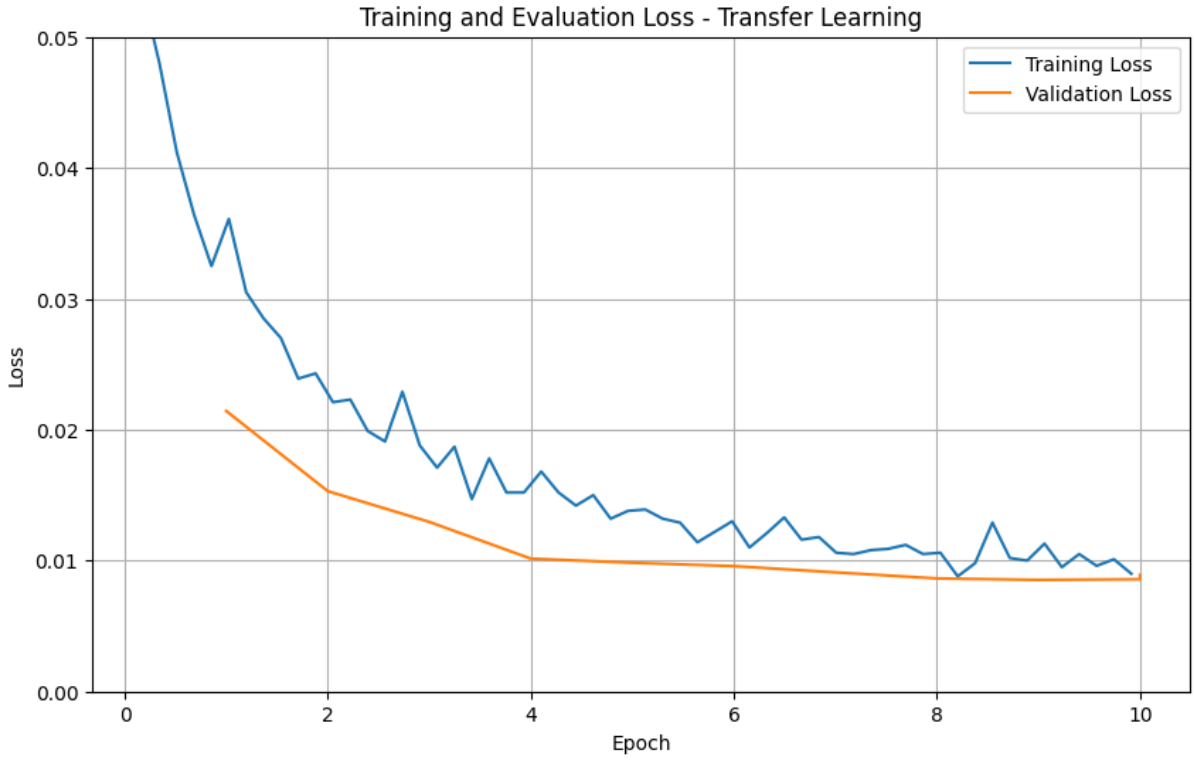
Phương pháp này sử dụng chiến lược Transfer learning trên tập dữ liệu gồm 10.000 cặp câu dữ liệu nền tảng (được kế thừa) và hơn 1.000 cặp câu chuyên sâu (cà phê, đã tăng cường).

4.3.1.1. Phân tích quá trình huấn luyện

Thời gian huấn luyện 10 epochs: 1 tiếng

Bảng 4-6 . Kết quả huấn luyện mô hình ViT5 trong phương pháp transfer learning trên bộ dữ liệu kết hợp giữa 10.000 dòng dữ liệu kết thừa và hơn 1500 dòng dữ liệu về coffee đã tăng cường

Epoch	Training Loss	Validation Loss	Bleu	Wer
1	0.032500	0.021430	85.381443	0.101308
2	0.024300	0.015305	89.403203	0.076484
3	0.018800	0.012955	91.709917	0.059935
4	0.015200	0.010153	93.321704	0.048753
5	0.013800	0.009833	93.777770	0.046852
6	0.013000	0.009579	93.177276	0.050095
7	0.011800	0.009102	94.095039	0.043610
8	0.010500	0.008638	94.479384	0.040479
9	0.010000	0.008526	94.654089	0.039808
10	0.009000	0.008575	94.651093	0.039025



Hình 4.5. Kết quả huấn luyện mô hình ViT5 trong phương pháp Transfer learning trên bộ dữ liệu kết hợp giữa 10.000 dòng dữ liệu kết thừa và hơn 1000 dòng dữ liệu về coffee đã tăng cường

Kết quả cho thấy:

Ngay từ Epoch 1, chỉ số BLEU đã đạt 85.38, cao hơn cả kết quả tốt nhất của phương pháp fine-tuning trước đó (chỉ đạt ~84 ở epoch 9). Điều này chứng minh việc duy trì 10.000 dòng dữ liệu nền tảng đã giúp mô hình giữ được kiến thức ngôn ngữ tổng quát khi chuyển sang miền dữ liệu mới.

Xét về BLEU, tăng đều và đạt đỉnh ở 94.65 (Epoch 9, 10). Đây là mức điểm gần như hoàn hảo cho các tác vụ sinh văn bản. WER (Tỷ lệ lỗi từ): Giảm xuống mức cực thấp 0.039 (tương đương sai số chưa đến 4%).

Giá trị loss giảm xuống mức **0.0085.** Mặc dù mô hình đạt độ chính xác rất cao, nhưng khoảng cách giữa *Training Loss* và *Validation Loss* rất nhỏ và song hành với nhau. Việc bổ sung 10.000 mẫu dữ liệu đa dạng đã giúp mô hình học được cấu trúc ngôn ngữ thực sự thay vì học vẹt các mẫu câu về cà phê.

4.3.1.2. Đánh giá kết quả trên tập kiểm tra

Bảng 4-7. Kết quả của mô hình ViT5 trong phương pháp transfer learning được đánh giá trên tập Test

Chỉ số	Giá trị
eval_loss	0.0089
eval_bleu	93.5689
eval_wer	0.0454

Xét kết quả đạt được trên tập test, **BLEU Score ~93.57,** chứng tỏ mô hình có khả năng chuyển đổi sang ngôn ngữ ký hiệu chính xác gần như tuyệt đối so với nhãn tham chiếu trên tập dữ liệu kiểm thử. Về **WER ~0.045,** tỷ lệ lỗi cực thấp cho thấy mô hình không chỉ bắt đúng từ khóa chuyên ngành (cà phê) mà còn xử lý cực tốt các từ ngữ ngữ pháp thông thường nhờ vào vốn kiến thức từ 10.000 câu dữ liệu cũ.

Minh họa trực quan với một số câu cầu vào thực tế:

```
Tiếng Việt: Tôi muốn mua một ly cà phê đen không đường.
VSL (dịch): Tôi cà-phê đen một đường không muốn
```

```
Input:   dịch tiếng Việt sang VSL: Mít thì ngọt .
Dự đoán: Mít ngọt .
Nhãn:    Mít ngọt .
=> ĐÁNH GIÁ: CHÍNH XÁC
```

Hình 4.6. Kết quả thử nghiệm dịch chuyển đổi văn bản theo cấu trúc ngữ pháp tiếng Việt sang cấu trúc ngữ pháp của ngôn ngữ kí hiệu (Transfer learning)

4.3.1. Kết quả thực nghiệm với phương pháp Fine-tuning

4.3.1.1. Kết quả huấn luyện finetuning lần 1

Thực hiện huấn luyện ba mô hình lớn: BARTpho, ViT5-base, ViT5-large. Cả ba mô hình đều được huấn luyện trên cùng tập dữ liệu 10.000 dòng tổng quát với cơ chế dừng sớm, early stopping patience = 3.

a. Phân tích quá trình huấn luyện của các mô hình

Bảng 4-8. Thời gian huấn luyện của 3 mô hình BARTpho, ViT5-base, ViT5-large trong fine-tuning lần 1

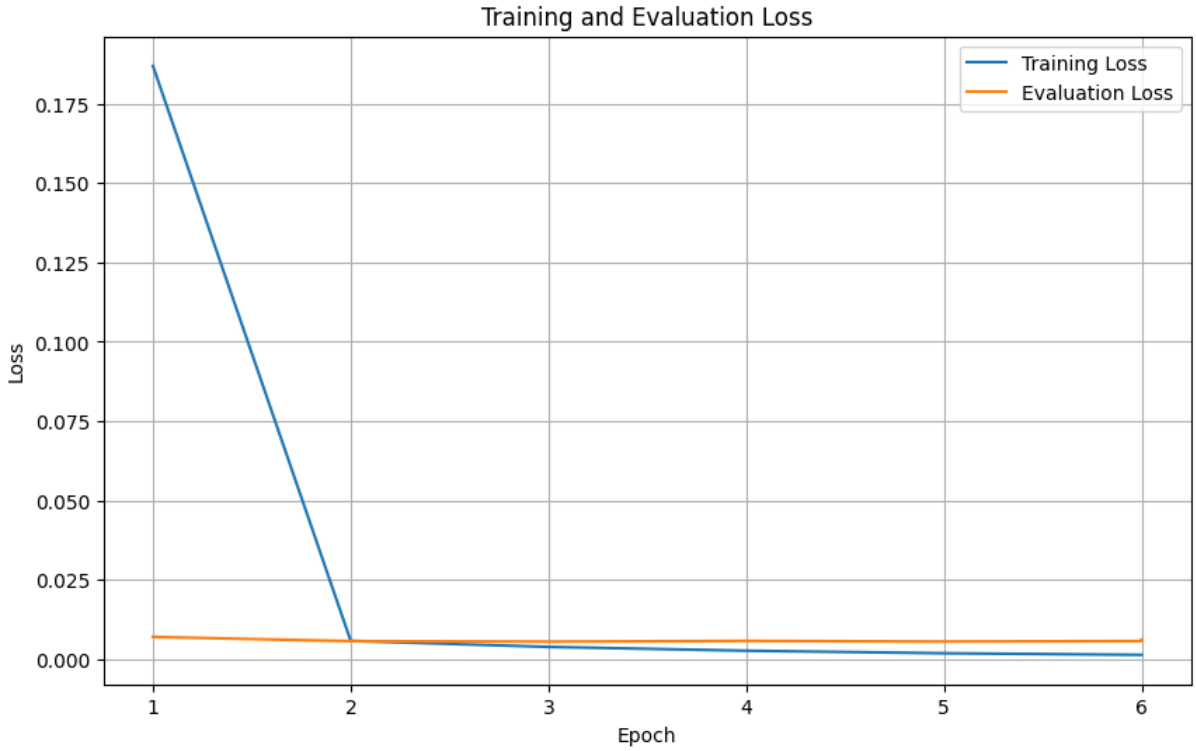
Mô hình	Thời gian huấn luyện	Số epochs thực hiện	Trạng thái dừng
ViT5-base	1 tiếng	6/10	Early stopping
ViT5-large	1 tiếng	5/10	Early stopping
BARTpho	36 phút	6/10	Early stopping

Xét về tốc độ, BARTpho nhanh hơn đáng kể so với mô hình ViT5.

Cả ba mô hình đều kích hoạt cơ chế dừng sớm ở epoch thứ 5 hoặc 6. Điều này cho thấy mô hình học nhanh, và đạt bão hòa sớm. Training loss đã có giá trị nhỏ rất nhiều (0.001 – 0.002), và có xu hướng đi ngang, hoặc tăng nhẹ. Do đó nên dừng sớm để tránh overfiting.

Bảng 4-9. Quá trình huấn luyện luyện BARTpho với phương pháp fine-tuning, thực hiện 6/10 epochs, dừng theo cơ chế early stopping patience = 3

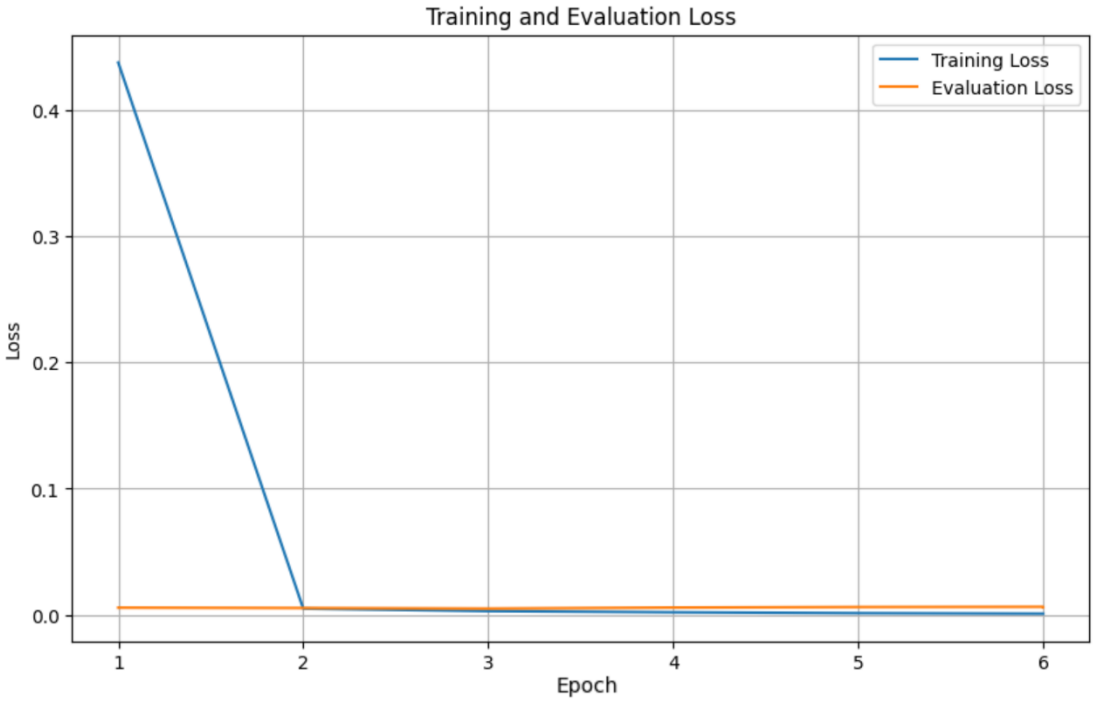
Epoch	Training Loss	Validation Loss	Bleu	Wer
1	0.1869	0.0071	95.2570	0.0356
2	0.0058	0.0057	96.2513	0.0306
3	0.0039	0.0056	95.9864	0.0324
4	0.0027	0.0058	96.1198	0.0317
5	0.0019	0.0056	96.3064	0.0305
6	0.0014	0.0057	96.5823	0.0288



Hình 4.7. Quá trình huấn luyện luyện BARTpho với phương pháp fine-tuning, thực hiện 6/10 epochs, dừng theo cơ chế early stopping patience = 3

Bảng 4-10. Quá trình huấn luyện luyện ViT5-base với phương pháp fine-tuning, thực hiện 6/10 epochs, dừng theo cơ chế early stopping patience = 3

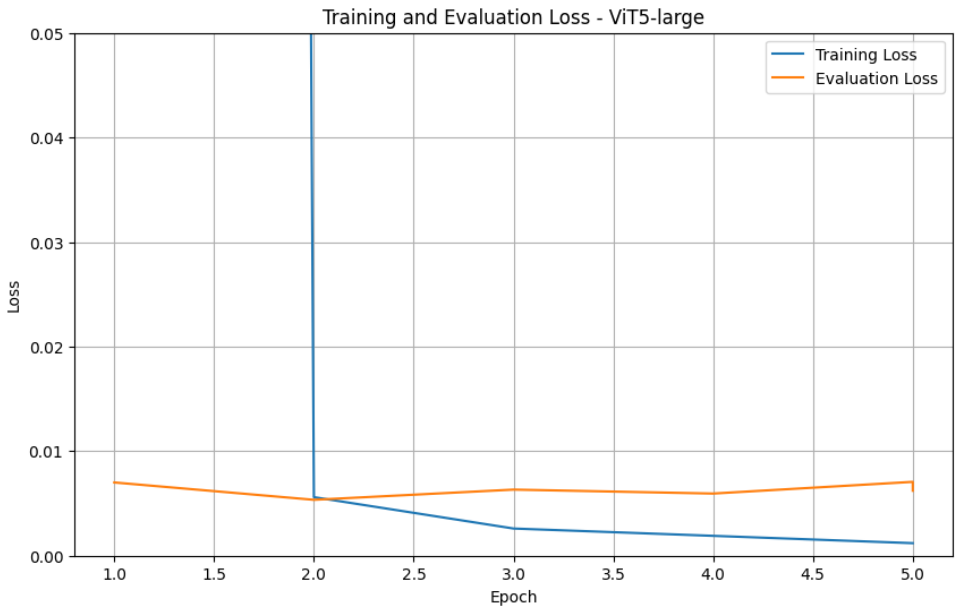
Epoch	Training Loss	Validation Loss	Bleu	Wer
1	0.438	0.006	95.932	0.029
2	0.005	0.005	96.924	0.024
3	0.003	0.005	97.187	0.021
4	0.002	0.006	97.255	0.022
5	0.001	0.006	97.171	0.022
6	0.001	0.006	97.419	0.021



Hình 4.8. Quá trình huấn luyện luyện ViT5-base với phương pháp fine-tuning, thực hiện 6/10 epochs, dừng theo cơ chế early stopping patience = 3

Bảng 4-11. Quá trình huấn luyện luyện ViT5-base với phương pháp fine-tuning, thực hiện 6/10 epochs, dừng theo cơ chế early stopping patience = 3

Epoch	Training Loss	Validation Loss	Bleu	Wer
1	3.083	0.007	95.521	0.036
2	0.006	0.005	96.612	0.025
3	0.003	0.006	97.206	0.022
4	0.002	0.006	97.315	0.021
5	0.001	0.007	97.119	0.022



Hình 4.9. Quá trình huấn luyện luyện ViT5-large với phương pháp fine-tuning, thực hiện 5/10 epochs, dừng theo cơ chế early stopping patience = 3

b. So sánh 3 mô hình BARTpho, ViT5-base, ViT5-large trên tập test

Trên tập Test, các mô hình có kết quả như sau:

Bảng 4-12. So sánh 3 mô hình BARTpho, ViT5-base, ViT5-large trong fine-tuning lần 1 trên tập test

Mô hình / Giá trị	BARTpho	ViT5-base	ViT5-large
Eval loss	0.0062	0.0061	0.0062
BLEU (%)	96.05	97.23	96.88
WER	0.033	0.017	0.019

Về độ chính xác, ViT5-base là mô hình dẫn đầu ở mọi chỉ số. Nó đạt BLEU cao nhất, tại 97,23 và tỷ lệ lỗi từ thấp nhất (0.017). Mặc dù ViT5-large có kiến trúc lớn hơn nhưng không có sự cải thiện so với phiên bản base. BARTpho tuy tốt nhưng nhìn chung vẫn kém hơn 2 phiên bản của ViT5.

Do đó, luận văn này quyết định lựa chọn ViT5 base làm mô hình tiếp tục giai đoạn Fine-tuning lần 2 trên bộ dữ liệu hơn 1.500 dòng coffee.

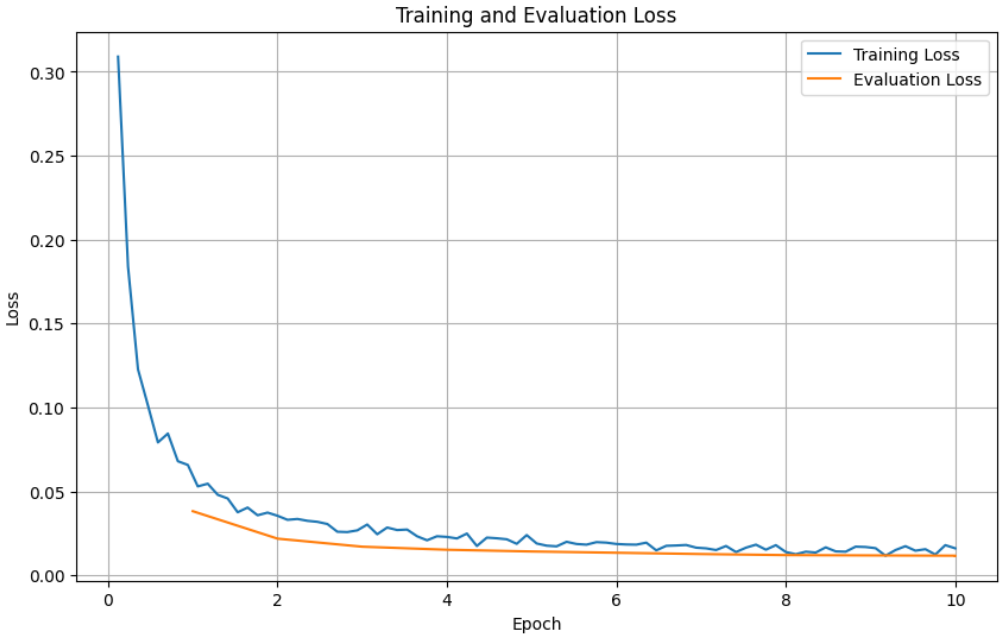
4.3.1.2. Kết quả thực nghiệm finetuning lần 2 với mô hình ViT5-base

Sau khi chọn mô hình tốt nhất từ finetuning lần 1 là ViT5, sẽ tiếp tục sử dụng mô hình này cho finetung lần hai.

Quá trình huấn luyện được thực hiện trên bộ dữ liệu Coffee đã tăng cường 1.691 câu với Learning rate = 0.00001 và batch size = 16.

Bảng 4-13. Kết quả huấn luyện mô hình ViT5 trong fine-tuning lần 2 trên bộ dữ liệu coffee đã tăng cường

Epoch	Training Loss	Validation Loss	Bleu	Wer
1	0.065800	0.038228	66.424247	0.245370
2	0.035500	0.021880	74.906515	0.180556
3	0.026800	0.017081	77.258115	0.173611
4	0.022900	0.015269	79.073533	0.152006
5	0.024000	0.014229	80.182104	0.152006
6	0.018700	0.013477	80.817070	0.146605
7	0.016500	0.012702	82.371072	0.138889
8	0.013900	0.012081	82.630784	0.135031
9	0.016900	0.011892	84.616674	0.119599
10	0.016100	0.011729	83.728810	0.128086



Hình 4.10. Kết quả huấn luyện mô hình ViT5 trong fine-tuning lần 2 trên bộ dữ liệu coffee đã tăng cường

Cả training loss và validation loss giảm liên tục qua từng epoch và chưa có dấu hiệu validation tăng.

Điểm BLEU tăng từ 66.42 lên 84.62 (epoch 1 lên epoch 9). Mặc dù hơi thấp hơn so với kết quả trên tập dữ liệu tổng quát nhưng 84.6 là kết quả ổn. Tỷ lệ lỗi từ cũng giảm mạnh từ 24.5% xuống 11.96% tại epoch 9. Điều này, mô hình đã giảm kể lỗi sai về từ vựng.

Tóm lại, giai đoạn fine-tuning lần 2 đã huấn luyện mô hình ViT5 thích nghi với tập dữ liệu liên quan đến cà phê.

4.3.2. So sánh hai mô hình sau khi Transfer learning và Fine-tuning 2 giai đoạn

Để xác định mô hình tối ưu nhất cho dịch máy trong phạm vi liên quan đến dịch vụ cà phê, luận văn so sánh kết quả thực nghiệm giữa hai phương phát tiếp cận.

+ Transfer learning: huấn luyện trực tiếp tập dữ liệu kết hợp (dữ liệu tổng quát + dữ liệu chuyên ngành) với ViT5-base

+ Fine-tuning 2 giai đoạn: Huấn luyện tuần tự, giai đoạn 1 ở dữ liệu nền tảng, giai đoạn 2 tinh chỉnh trên dữ liệu chuyên ngành. Giai đoạn đầu so sánh 3 mô hình

BARTpho, ViT5-base, ViT5-large rồi chọn mô hình tốt nhất. Mô hình tốt nhất sẽ được fine tuned lần 2.

Bảng 4-14. So sánh kết quả mô hình ViT5-base (Transfer learning) và ViT5-base (Fine-tuning 2 giai đoạn)

Mô hình	Eval loss	BLEU	WER	Thời gian huấn luyện
ViT5-base (Transfer learning)	0.0091	90.4240	0.0612	1 tiếng 10ph
ViT5-base (Fine-tuning 2 giai đoạn)	0.0045	95.7539	0.0272	1 tiếng 30ph

Về độ chính xác dịch thuật (BLEU và WER), phương pháp Fine-tuning 2 giai đoạn thể hiện sự đúng cao. Chỉ số BLEU đạt 95.75, cao hơn tầm 5 điểm so với phương pháp Transfer learning. Chỉ số WER khi fine-tuning 2 lần giảm xuống mức thấp, sai 2.7%, trong khi phương pháp còn lại tỷ lệ lỗi gấp đôi (6.12%).

Có thể giải thích rằng, phương pháp fine-tuning 2 giai đoạn cho phép mô hình hình thành cấu trúc ngữ pháp vững chắc ở giai đoạn đầu, sau đó tập trung tối ưu hóa từ vựng ở giai đoạn 2. Trong khi, phương pháp Transfer learning gập chung dữ liệu khiến mô hình bị loãng sự chú ý đến dữ liệu chuyên ngành.

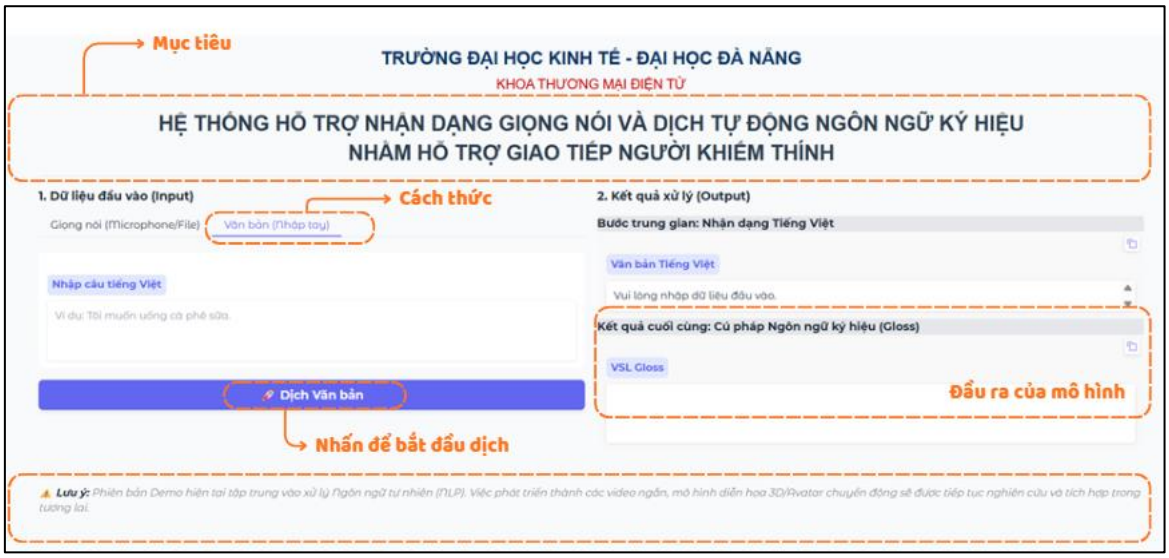
Dựa vào kết quả trên, ***ViT5-base với phương pháp Fine-tuning 2 giai đoạn*** là mô hình tối ưu nhất. Đây là mô hình cuối cùng được lựa chọn để tích hợp vào hệ thống dịch tự động hỗ trợ giao tiếp ở người khiếm thính.

4.4. Triển khai mô hình nhận dạng giọng nói và dịch máy

Tích hợp PhoWhisper và mô hình dịch ViT5, xây dựng giao diện người dùng như sau:

Hình 4.11. Giao diện hệ thống khi triển khai, với đầu vào có thể audio file, ghi âm trực tiếp hoặc văn bản

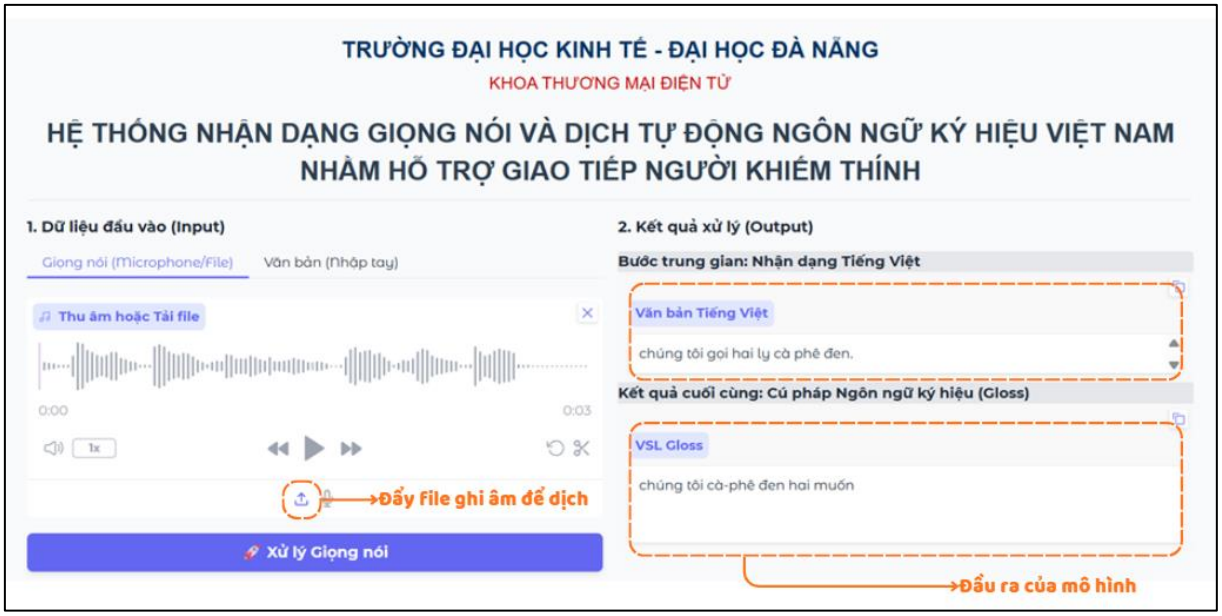
Hình 4.12. Khi ghi âm trực tiếp, hệ thống nhận dạng giọng nói, dịch và đưa ra kết quả



Hình 4.13. Giao diện hệ thống khi người dùng muốn nhập văn bản

Hình 4.14 Ví dụ khi nhập văn bản trên hệ thống

Hình 4.15. Khi ghi âm trực tiếp một câu nói dài hơn, gọi hai món thì hệ thống vẫn ghi nhận và dịch đúng



Hình 4.16. Khi sử dụng một audio file để dịch thì hệ thống vẫn ghi nhận và dịch đúng

4.5. Phân tích tình huống lỗi

Trong quá trình triển khai trong môi trường giả lập thực tế, mô hình đã bộc lộ một số hạn chế nhất định. Dưới đây là phân tích chi tiết hai nhóm lỗi điển hình:

4.5.1. Lỗi nhận dạng giọng nói trong môi trường có nhiều tiếng ồn

Tình huống: Trong môi trường quán cà phê, thường xuyên có các tạp âm như tiếng nhạc, tiếng ồn từ khách hàng. Khi thử nghiệm với đầu vào "*Cho tôi ly nước ép **cam***" trong điều kiện có tiếng nhạc, mô hình PhoWhisper đôi khi ghi nhận sai thành "*Cho tôi một ly ép **cà***".

Nguyên nhân có thể là các từ đơn trong tiếng Việt như cam, cà có cấu trúc phát âm gần giống nhau và có thêm tiếng ồn thì dẫn đến mô hình dự đoán sai từ vựng.

Đề xuất hướng khắc phục: Có thể tích hợp module tiền xử lý âm thanh để lọc tạp âm trước khi đưa vào mô hình.

4.5.2. Lỗi mất thông tin trong dịch máy

Tình huống: Đối với module dịch máy (ViT5-base), khi chạm đến xử lý các yêu cầu thuộc nhóm "đá xay", hệ thống có xu hướng lược bỏ hoặc dịch thiếu từ khóa quan trọng.

Nguyên phân có thể là, mất cân bằng dữ liệu. Như phân tích ở chương 4, ở mục kết quả tăng cường dữ liệu, mặc dù đã tăng cường nhưng số lượng mẫu câu về "Đá xây" vẫn thấp đáng kể so với nhóm "Cà phê" hay "Nước ép." Do đó, cụm từ "đá xay" xuất hiện với tần suất ít trong tập dữ liệu, khiến mô hình chưa học nhiều về thông tin này.

Đề xuất hướng khắc phục: Cần thu thập, tăng cường các dữ liệu trong nhóm có sự cân bằng phù hợp.

4.6. So sánh với công trình khác

Khi đặt tổng bối cảnh các nghiên cứu ứng dụng mô hình ngôn ngữ lớn cho bài toán dịch sang ngôn ngữ ký hiệu Việt Nam, nghiên cứu này có một số điểm cải thiện hơn.

Đa phần các nghiên cứu hiện có thường áp dụng fine-tuning đơn giai đoạn như nghiên cứu của Trân Vũ Hoàng và cộng sự (Trần Vũ Hoàng, 2025). Các mô hình này thường chỉ huấn luyện một lần trên tập dữ liệu phổ biến (10.000 cặp câu). Hạn chế của phương pháp này là mô hình có thể nắm bắt tốt ngữ pháp chung nhưng thường gặp khó khăn khi gặp các từ ngữ chuyên môn chưa xuất hiện.

Do đó, nghiên cứu này đã đề xuất và áp dụng thành công chiến lược Fine-tuning 2 giai đoạn.

+ Giai đoạn 1: Kế thừa tri thức ngữ pháp NNKH từ dữ liệu nền tăng (10.000 cặp câu)

+ Giai đoạn 2: Tinh chỉnh trên dữ liệu cụ thể (coffee)

Cách tiếp cận này giúp mô hình duy trì khả năng cấu trúc chuẩn, hiểu hiểu ngữ gọi món.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Luận văn được thực hiện xuất phát từ động lực cần thiết để xóa bỏ rào cản ngôn ngữ và hỗ trợ hòa nhập cho cộng đồng người khiếm thính tại Đà Nẵng nói riêng và Việt Nam nói chung. Trong bối cảnh dịch vụ hay môi trường giáo dục, đời sống, nhu cầu giao tiếp hai chiều là vô cùng lớn. Để giải quyết vấn đề này, đề tài tập trung nghiên cứu hệ thống tích hợp, có khả năng nhận dạng giọng nói và chuyển đổi tự động sang văn bản đúng cấu trúc ngữ pháp ngôn ngữ ký hiệu Việt Nam.

Kết quả thực nghiệm đã chứng minh tính khả thi và độ tin cậy hệ thống, cung cấp chỉ số cao, góp phần mở rộng cơ hội tiếp cận thông tin (kênh truyền hình,...) tri thức xã hội cho cộng đồng người khiếm thính.

1. Đánh giá mức độ hoàn thành mục tiêu đặt ra và kết quả đạt được

Dựa trên những mục tiêu ở phần Mở đầu, kết quả đạt được của luận văn cụ thể như sau:

Về mục tiêu xây dựng tài nguyên dữ liệu:

+ Mục tiêu: Xây dựng bộ dữ liệu về lĩnh vực dịch vụ cà phê để phục vụ huấn luyện mô hình.

+ Kết quả: Đã xây dựng bộ dữ liệu. Đặc biệt, luận văn còn giải quyết bài toán khan hiếm dữ liệu bằng cách áp dụng kỹ thuật tăng cường dữ liệu thông qua LLMs như ChatGPT, Gemini.

Về phương pháp:

+ Mục tiêu: Áp dung và kế thừa thành tựu nghiên cứu trong, ngoài nước

+ Kết quả: Đã áp dụng thành công mô hình hiện đại cho tiếng Việt như PhoWhisper, ViT5, BARTpho. Đồng thời nghiên cứu, đánh giá hai phương pháp học Transfer learning và Fine-tuning 2 giai đoạn.

Về mô hình dịch:

+ Mục tiêu: Phát triển mô hình chuyển đổi văn bản tiếng Việt sang văn bản đúng cú pháp ngôn ngữ ký hiệu có tỷ lệ lỗi từ WER < 10%.

+ Kết quả: Đã huấn luyện, so sánh và tìm ra mô hình tối ưu là ViT5-base (Fine-tuning 2 giai đoạn). Mô hình đạt hiệu suất trên tập test là BLEU – 95.75% và tỷ lệ lỗi từ 2,7% .

Về hệ thống:

+ Mục tiêu: Tích hợp hai module nhận dạng giọng nói và dịch máy thành luồng thống nhất

+ Kết quả: Đã xây dựng thành công pipeline: Đầu vào là âm thanh → Văn bản theo cấu trúc tiếng Việt → Văn bản đúng cú pháp của ngôn ngữ ký hiệu.

2. Đóng góp chính của luận văn

Về mặt học thuật, luận văn đã xây dựng bộ dữ liệu song ngữ ngôn ngữ tiếng Việt và ngôn ngữ ký hiệu Việt Nam chuyên sâu về lĩnh vực F&B. Ngoài ra, còn sử dụng phương pháp tăng cường dữ liệu bằng LLMs và Fine-tuning 2 giai đoạn. Bên cạnh đó, lận văn còn nghiên cứu, so sánh định lượng giữa các mô hình ngôn ngữ tiên tiến như ViT5-base, ViT5-large, BARTpho.

Về mặt thực tiễn: Hệ thống cung cấp một công cụ tự động, hiệu quả, đáng tin cho việc hỗ trợ người khiếm thính. Cụ thể, nó giúp người khiếm thính hiểu ý định của khách hàng mà không phụ thuộc vào người khác, đặc biệt trong môi trường đòi hỏi tốc độ như quán cà phê. Hệ thống sẽ giúp bình đẳng hóa cơ hội làm việc của người khiếm thính trong cộng đồng.

3. Hạn chế và hướng phát triển

Mặc dù đạt được kết quả khả quan, luận văn vẫn còn tồn tại một số hạn chế nhất định cần khắc phục.

Về hạn chế:

+ Dữ liệu còn hạn chế. Dù đã áp dụng tăng cường dữ liệu, bộ dữ liệu về chuyên môn cà phê (hơn 1.600) vẫn là con số khiêm tốn so với sự phức tạp ngôn ngữ.

+ Môi trường có tiếng ồn: Mô hình nhận dạng giọng nói đôi khi gặp khó khăn khi hoạt động trong môi trường có tiếng ồn lớn, nên sẽ ảnh hưởng ít nhiều đến độ chính xác của đầu vào module dịch.

+ Phạm vi ứng dụng hẹp: Hiện tại hệ thống mới chỉ tối ưu cho phạm vi quán cà phê, chưa áp dụng các lĩnh khác cần thiết hơn như y tế hay hành chính.

Về hướng phát triển trong tương lai: Bài toán mở rộng hướng tới phát triển mô hình diễn họa ngôn ngữ ký hiệu 3D

Việc diễn họa NNKH dưới dạng 3D/Avatar, thường gọi Text to sign language animation, yêu cầu đầu vào phải là một dạng biểu diễn ngôn ngữ chuẩn hóa, có khả năng mã hóa các yếu tố hình thái và cú pháp của ngôn ngữ ký hiệu

Xét tổng quan các nghiên cứu mô hình chuyển đổi văn bản thành 3D/Avatar trên thế giới

Trong nghiên cứu toàn cầu về dịch máy và diễn họa NNKH, đầu vào thường là các định dạnh sau

Gloss (Văn bản biểu diễn ngôn ngữ ký hiệu)

Glosss là hình thức biểu diễn văn bản của một câu NNKH, tuân thủ đúng cấu trúc ngữ pháp của NNKH đó (ví dụ như American Sign Language – ASL). Gloss là bước trung gian thiết yếu trong lĩnh vực dịch máy ngôn ngữ ký hiệu, đặc biệt khi chuyển đổi từ ngôn ngữ nói/viết thông thường (có ngữ pháp khác biệt) sang cú pháp của ngôn ngữ ký hiệu.

Các mô hình dịch máy tiên tiến thường được huấn luyện để chuyển đổi văn bản nguồn như tiếng Anh sang dạng Gloss. Các nghiên cứu về dịch thuật văn bản sang NNKH thường được đánh giá bằng các tiêu chuẩn như ROUGE và BLEU.

Mã hóa hình thái (Notation Systems):

HamNoSys là một hệ thống phiên âm phổ biến, bao gồm hơn 200 ký tự, dùng để dịch chuyển các ký hiệu riêng lẻ và đoạn văn trong NNKH. HamNoSys được sử dụng trong các dự án quốc tế như ViSiCAST và eSIGN.

SiGML – Sign Language Markup Language là định dạng XML cơ bản dựa trên các ký hiệu HamNoSys, được sử dụng để định nghĩa mô tả cho hoạt hình 3D, bao gồm vị trí bàn tay, vận tốc và biên độ cử chỉ.

Các công cụ tổng hợp NNKH như JASigning – được phát triển tại đại học East Anglia, cho phép chuyển đổi ký hiệu đại diện trong SiGML sang dạng hoạt hình 3D của NNKH.

Dữ liệu tư thế (Pose Data)

Các nghiên cứu tiên tiến hơn trong việc tạo video ký hiệu sử dụng trực tiếp dữ liệu tư thế. Ví dụ, một số hệ thống tạo video lấy các tư thế mô hình cơ thể 3D làm đầu vào cho module diễn họa.

Xét các nghiên cứu tại Việt Nam

Các nghiên cứu diễn họa NNKH Việt Nam cũng tập trung vào dạng mã hóa chuẩn:

Sử dụng HamNoSys và Avatar 3D: Các nghiên cứu trước đây tại Việt Nam đã ứng dụng hệ thống HamNoSys và công cụ Avatar 3D JASigning để xây dựng bộ từ điển NNKH Việt Nam. Nghiên cứu của Nguyễn Chí Ngôn và cộng sự (2017) đã xây dựng 2.558 từ dựa trên mã HamNoSys)

VSL Gloss: Gần đây, các hệ thống phiên dịch đầy đủ tại Việt Nam xác định đầu ra của quá trình dịch thuật phải là Văn bản biểu diễn đúng cú pháp ngôn ngữ ký hiệu Việt Nam, và dạng văn bản này được coi là đầu vào chuẩn hóa cho các mô hình diễn họa 3D/Avatar trong tương lai.

Một nghiên cứu của Trần Vũ Hoàng và cộng sự (2025) đã thiết kế phần mềm phiên dịch NNKH đề xuất sử dụng văn bản được chuyển đổi ngữ pháp từ mô hình ViT5 làm đầu vào cho khối dựng tổng hợp ngõ ra, khối này sau đó truy xuất vào từ điển mô hình hóa và sử dụng Blender Python API để tạo ra video.

So sánh đầu ra của nghiên cứu và đầu vào mô hình diễn họa

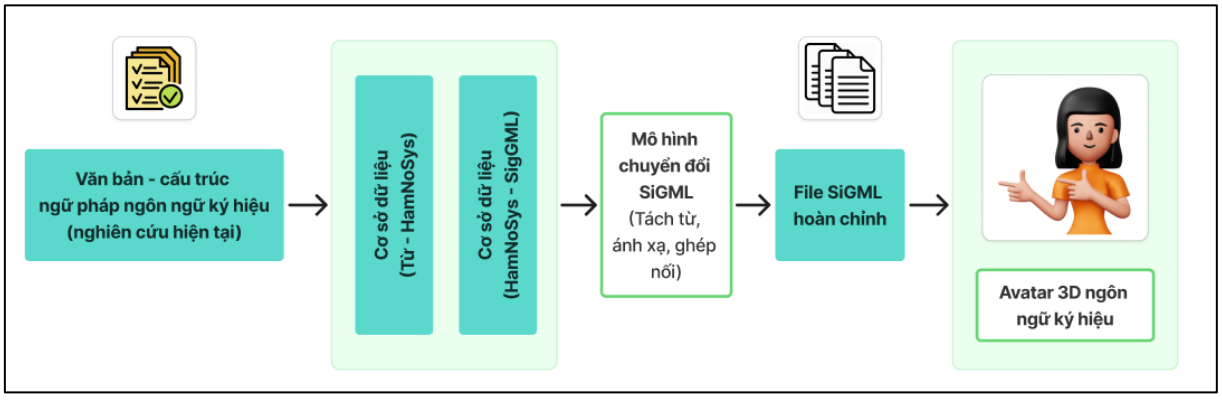
Đầu ra của nghiên cứu

Đâu ra cuối cùng của mô hình dịch máy trong nghiên cứu là Văn bản biểu diễn đúng cú pháp của NNKH Việt Nam.

Bảng 0-1. So sánh và kết luận về trùng khớp và khớp biệt của đầu ra nghiên cứu và mô hình diễn họa 3D

Đặc điểm	Kết quả nghiên cứu hiện tại	Đầu vào tiêu chuẩn cho diễn họa 3D (hướng phát triển)	Đánh giá
Định dạng	Văn bản tuân theo NNKH Việt Nam	Văn bản tuân theo đúng cú pháp NNKH, hoặc mã hóa như HamNoSys	Trùng hợp
Vai trò	Đầu ra này được thiết để là đầu vào chuẩn hóa cho các mô hình diễn họa 3D trong tương lai	Cung cấp ngữ pháp và ngữ nghĩa cần thiết để mô hình 3D tạo ra hành động ký hiệu	Có liên kết

Mục tiêu chính của luận văn là tập trung vào bài toán nhận diện giọng nói và dịch tự động ngôn ngữ ký hiệu theo đúng cú pháp, và các vấn đề liên quan đến tổng hợp hay biểu hiễn NNKH dưới dạng hình ảnh, video, avatar nằm ngoài phạm vi nghiên cứu. Do đó***, việc đầu ra và văn bản đúng cú pháp NNKH đã hoàn thành nhiệm vụ của module này là cung cấp dữ liệu đầu vào chất lượng cho bước tiếp theo xây dựng mô hình chuyển đổi từ văn bản sang hình ảnh 3D được phát triển trong tương lai***.



Hình. Quy trình xây dựng diễn họa Avatar 3D ngôn ngữ ký hiệu trong hướng phát triển

TÀI LIỆU THAM KHẢO

Beidas, A., Ghaddar, F., Mohi, K., Ahmad, I., & Abed, S. E. J. F. i. A. I. (2025). Cross-dialectal Arabic translation: comparative analysis on large language models. *8*, 1661789.

Diep, N. T. B. (2023). *Nghiên cứu và phát triển phương pháp tiếp cận dựa trên cấu trúc và thống kê trong dịch tự động ngôn ngữ ký hiệu Việt Nam.* (Tiến sĩ Khoa học máy tính), Hà Nội. Retrieved from https://gust.edu.vn/media/30/uftai-ve-tai-day30075.pdf

GeeksforGeeks. (2025). Text Augmentation Techniques in NLP. Retrieved from https://www.geeksforgeeks.org/nlp/text-augmentation-techniques-in-nlp/

Gruetzemacher, R., & Paradice, D. J. A. C. S. (2022). Deep transfer learning & beyond: Transformer language models in information systems research. *54*(10s), 1-35.

Kowsher, M., Sami, A. A., Prottasha, N. J., Arefin, M. S., Dhar, P. K., & Koshiba, T. J. I. A. (2022). Bangla-bert: transformer-based efficient model for transfer learning and language understanding. *10*, 91855-91870.

Le, T.-T., Nguyen, L. T., & Nguyen, D. Q. J. a. p. a. (2024). Phowhisper: Automatic speech recognition for vietnamese.

Luu, S. T., Van Nguyen, K., Nguyen, N. L.-T. J. M. T., & Applications. (2024). An approach of data augmentation to improve the performance of BERTology models for Vietnamese hate speech detection. *83*(19), 56763-56783.

Nguyen, D. Q., & Nguyen, A. T. J. a. p. a. (2020). PhoBERT: Pre-trained language models for Vietnamese.

Nguyễn, K. Đ., Nguyễn, T. T., Trần, N. A., & Võ, V. N. (2024). Xây dựng và phát triển mô hình nhận dạng phát âm sai trong Tiếng Việt.

Patel, B. D., Patel, H. B., Khanvilkar, M. A., Patel, N. R., & Akilan, T. (2020). *ES2ISL: an advancement in speech to sign language translation using 3D avatar animator*. Paper presented at the 2020 IEEE Canadian conference on electrical and computer engineering (CCECE).

Phan, L., Tran, H., Nguyen, H., & Trinh, T. H. J. a. p. a. (2022). ViT5: Pretrained text-to-text transformer for Vietnamese language generation.

Reviriego, P., Conde, J., Merino-Gómez, E., Martínez, G., & Hernández, J. A. J. M. L. w. A. (2024). Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. *18*, 100602.

Saunders, B., Camgoz, N. C., & Bowden, R. J. I. j. o. c. v. (2021). Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *129*(7), 2113-2135.

Stefanovič, P., Radvilaitė, U., Pliuskuvienė, B., & Ramanauskaitė, S. J. S. R. (2025). The influence of Gen-AI tools application for text data augmentation: case of Lithuanian educational context data classification. *15*(1), 26010.

Tauqeer, S., Muhammad, K., Babar, A., & Muhammad, S. J. R. (2021). A Real-Time Automatic Translation of Text to Sign Language.

Thu, M. V. T., Trần Nguyễn Duy Nghĩa, Phùng Trung. (2018). KỸ THUẬT TỐI ƯU LƯỚI TỨ GIÁC TRONG ĐIỀU KHIỂN ĐỐI TƯỢNG BA CHIỀU ÁP DỤNG DIỄN HỌA NGÔN NGỮ KÝ HIỆU VIỆT NAM. *TNU Journal of Science Technology, 178*(02), 91-96.

Tran, L. T. T., Kim, H.-G., La, H. M., & Van Pham, S. J. E. (2024). Automatic Speech Recognition of Vietnamese for a New Large-Scale Corpus. *13*(5), 977.

Tran, N. L., Le, D. M., & Nguyen, D. Q. J. a. p. a. (2021). BARTpho: pre-trained sequence-to-sequence models for Vietnamese.

Trần Vũ Hoàng, L. Q. Đ., Huỳnh Đình Hiệp, Đoàn Mạnh Cường. (2025). Thiết kế xây dựng phần mềm phiên dịch ngôn ngữ ký hiệu tiếng Việt. *TNU Journal of Science and Technology*. doi:https://doi.org/10.34238/tnu-jst.12232

Uda, H., Matsumoto, K., & Yoshida, M. (2024). *Text Data Augmentation Method Using Filtering Indicators based on Multiple Perspectives.* Paper presented at the Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation.

Uyên, L. T. T. (2020). Một số phương thức hình thành kí hiệu của người điếc Việt Nam *Tạp chí Khoa học Giáo dục Việt Nam*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. J. A. i. n. i. p. s. (2017). Attention is all you need. *30*.

xuân Mỹ, C. T. J. T. c. K. h. T. Đ. h. S. p. T. H. C. M. (2019). Quá trình hình thành và phát triển ngôn ngữ kí hiệu. (46), 181-181.

Zhang, H., Shalev-Arkushin, R., Baltatzis, V., Gillis, C., Laput, G., Kushalnagar, R., . . . Lea, C. (2025). *Towards AI-driven Sign Language Generation with Non-manual Markers.* Paper presented at the Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.