THÁCH THỨC VÀ GIẢI PHÁP ỨNG DỤNG AI TRONG NGHIÊN CỨU VÀ GIẢNG DẠY NGÔN NGỮ HỌC

NGUYỄN NGỌC BÌNH[1]
ĐẶNG THỊ MAI[2]

Abstract: Artificial Intelligence (AI) has undergone a significant transformation in language analysis, providing powerful tools for understanding and interpreting various aspects of language. Through techniques such as natural language processing (NLP), machine learning and deep learning, AI enables the analysis of linguistic data with unprecedented levels of accuracy, efficiency, and depth. This research will point out the challenges in applying AI to linguistic research and teaching and propose potential solutions to address these challenges. In addition, the article presents a specific application of speech-to-text recognition to demonstrate the contribution of linguistic knowledge to the development of AI -based tools.

Keywords: *linguistics, artificial intelligence, challenges, solutions, application*

1. Mở đầu

1.1. Ngôn ngữ từ lâu đã là đối tượng quan tâm của nhiều ngành khoa học do bản thân nó có vai trò nền tảng trong giao tiếp và nhận thức của con người. Ngôn ngữ học nghiên cứu cấu trúc, sự phát triển và cách thức sử dụng ngôn ngữ trong xã hội; ở mức sâu hơn là những tiếp cận bản chất, quy luật và cơ chế vận hành của ngôn ngữ. Truyền thống nghiên cứu dựa trên khung lí thuyết và dữ liệu thực nghiệm để lí giải cấu trúc, tiếp thu và sử dụng ngôn ngữ. Sự xuất hiện của Trí tuệ Nhân tạo (AI) đã tạo ra bước chuyển biến đáng kể đối với ngôn ngữ học, đặc biệt ở khía cạnh công cụ và phương pháp, mở ra hướng tiếp cận mới trong nghiên cứu ngôn ngữ học. AI là lĩnh vực phát triển mạnh với tập hợp kĩ thuật mô phỏng chức năng nhận thức, đóng vai trò tác nhân chuyển đổi trong nghiên cứu ngôn ngữ. AI không chỉ là công nghệ mà còn là phương pháp luận và hệ sinh thái ứng dụng, cho phép hệ thống máy thực hiện nhiều nhiệm vụ vốn thuộc năng lực của con người. Thông qua tích hợp các kĩ thuật như Xử lí Ngôn ngữ Tự nhiên (NLP), Học máy (ML) và Học sâu (DL), các nhà nghiên cứu có thể phân tích nhiều phương diện của ngôn ngữ với độ chính xác và hiệu quả cao. Việc tích hợp AI đã hiện đại hoá diện mạo nghiên cứu ngôn ngữ học, đặc biệt ở khả năng tăng tốc và tự động hoá các khâu xử lí. NLP giúp máy hiểu, diễn giải và sinh ngôn ngữ tự nhiên, hỗ trợ các tác vụ như gán nhãn từ loại, phân tích cú pháp, nhận dạng thực thể, phân tích cảm xúc. ML cho phép xây dựng mô hình dự đoán về quá trình tiếp thu và sử dụng ngôn ngữ từ dữ liệu quy mô lớn. DL (các mạng nơ-ron) thể hiện năng lực đáng chú ý trong việc nắm bắt mẫu hình phức tạp và cấu trúc ngữ nghĩa.

1.2. Trong nghiên cứu tiếp thu ngôn ngữ, AI đặc biệt hữu ích cho mô hình hoá quá trình phát triển (Graves, A. và cộng sự, 2006). Thông qua mô hình tính toán và mô phỏng, các quá trình nhận thức liên quan đến học ngôn ngữ có thể được tái hiện và kiểm định bằng dữ liệu. Các kĩ thuật AI cũng hỗ trợ phân tích tiếp thu ở nhiều bình diện như âm vị, hình thái, cú pháp và ngữ nghĩa. ML đã được ứng dụng để dự đoán quỹ đạo phát triển và phân tích sản phẩm ngôn ngữ cùng lỗi của người học, qua đó gợi mở cơ chế nền tảng của tiếp thu. Công cụ dựa trên AI cũng tự động hoá đánh giá năng lực, hỗ trợ giáo viên chuẩn đoán năng lực và cung cấp phản hồi cá nhân hoá. Các hệ thống này sử dụng NLP và ML để đánh

[1] Trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Hà Nội; Email: ngocbinh@vnu.edu.vn
[2] Trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Hà Nội.

giá độ chính xác ngữ pháp, vốn từ và độ lưu loát, nâng cao hiệu quả dạy học và đánh giá. Bên cạnh triển vọng, tích hợp AI cũng đặt ra thách thức: phụ thuộc vào dữ liệu gán nhãn, thiên lệch mô hình, và các vấn đề đạo đức liên quan đến quyền riêng tư và đồng thuận. Giải quyết các thách thức này, cùng thúc đẩy hợp tác liên ngành, là điều kiện cần để phát triển lí thuyết ngôn ngữ học và khai thác đầy đủ tiềm năng của AI trong nghiên cứu ngôn ngữ (Elma Sibonghanoy Groenewald, 2024).

2. Nội dung

2.1. Những vấn đề về lí thuyết

2.1.1. Sự giao thoa giữa AI và ngôn ngữ học thu hút sự quan tâm ngày càng lớn khi các kĩ thuật AI được huy động để nâng cao hiểu biết về tiếp thu và phân tích ngôn ngữ. Phần này điểm lược những đóng góp tiêu biểu về ứng dụng AI trong nghiên cứu ngôn ngữ, tập trung vào hai trục: tiếp thu và phân tích. Nhiều công trình khảo sát vai trò của NLP trong nghiên cứu tiếp thu ngôn ngữ. Lí thuyết ngữ pháp biến đổi - tạo sinh của Chomsky đặt nền tảng cho tiếp cận ngôn ngữ học tính toán đối với tiếp thu ngôn ngữ, gợi mở các nghiên cứu về mô hình học ngôn ngữ dựa trên AI. Các nhà tiên phong như Terry Winograd (Winograd, T., 1971) và Roger Schank phát triển những hệ thống AI sớm có khả năng hiểu và tạo sinh ngôn ngữ tự nhiên, tạo tiền đề cho kĩ thuật NLP hiện đại. Trong mảng tiếp thu ngôn ngữ, AI được dùng để mô hình hoá phát triển và phân tích quỹ đạo tiếp thu. Mạng nơ-ron hồi tiếp đơn giản (SRNs) của Elman và Phân loại theo Thời gian kết nối (CTC) được sử dụng mô phỏng các quá trình nhận thức trong học ngôn ngữ và dự đoán kết quả phát triển. Đồng thời, ML và DL hỗ trợ phân tích kho dữ liệu lớn và khám phá mẫu hình trong quá trình tiếp thu (Sutskever, I. và cộng sự, 2014).

2.1.2. Trong phân tích ngôn ngữ học, AI làm thay đổi cách nghiên cứu các bình diện cú pháp, ngữ nghĩa và ngữ dụng. Các hệ thống kí hiệu sớm như SHRDLU chứng minh tiềm năng suy luận trong phân tích cú pháp và hiểu câu. Gần đây, các mô hình học sâu trong AI (RNNs và transformer) đạt hiệu suất ấn tượng trong dịch máy (Vaswani, A. và cộng sự, 2017) và phân tích cảm xúc (Socher, R. và cộng sự, 2023), mở rộng hiểu biết về cấu trúc và ý nghĩa ngôn ngữ. Nhìn chung, tài liệu cho thấy tác động chuyển đổi của AI đối với ngôn ngữ học: Từ mô hình hoá tiếp thu đến phân tích và hiểu ngôn ngữ. Nhờ các kĩ thuật AI, nghiên cứu có thể đi sâu hơn vào cơ chế phức tạp của tiếp thu và sử dụng ngôn ngữ, đặt nền cho các bước tiến tiếp theo.

2.1.3. AI trở thành công cụ trọng yếu trong nghiên cứu ngôn ngữ học, tái định hình cách khảo sát, phân tích và tiếp thu ngôn ngữ. Sự kết hợp của NLP, ML, DL cho phép tiếp cận sâu các khía cạnh phức tạp của mô hình và cấu trúc ngữ nghĩa với độ sâu và độ chính xác cao. Trong nghiên cứu tiếp thu ngôn ngữ, AI hỗ trợ mô hình hoá và lí giải các quá trình nhận thức nền tảng của học ngôn ngữ theo thời gian. Các mô hình nơ-ron mô phỏng tiếp thu cú pháp, hình thái và ngữ nghĩa, qua đó làm rõ cơ chế cốt lõi của học ngôn ngữ (Elma Sibonghanoy Groenewald, 2024). Bên cạnh đó, AI cho phép phân tích dữ liệu quy mô lớn để nhận diện mô hình phát triển và biến thể giữa nhóm người dùng và bối cảnh. Các tác vụ như mô hình hoá ngôn ngữ, phân loại văn bản, truy xuất thông tin được thực hiện hiệu quả bằng ML, qua đó xây dựng các mô hình dự đoán nắm bắt cấu trúc và mẫu sử dụng ngôn ngữ.

2.1.4. DL (Học sâu) là nhánh ML dùng mạng nơ-ron nhiều lớp để học biểu diễn phức tạp. Trong ngôn ngữ học, DL được áp dụng cho dịch máy, nhận dạng tiếng nói và phân tích cảm xúc. Đặc biệt, RNNs và transformer thể hiện năng lực vượt trội trong học mô hình ngôn ngữ và ngữ nghĩa; NLP cho phép máy thực hiện các tác vụ gán nhãn từ loại, phân tích cú pháp, nhận dạng thực thể và phân tích cảm xúc với độ chính xác cao (Cho, K. và cộng sự, 2014) Các kĩ thuật này hỗ trợ dịch, truy xuất thông tin và tóm tắt tự động, khai thác tri thức từ dữ liệu ngôn ngữ lớn (Mikolov, T. và cộng sự, 2013; Vaswani, A. và cộng sự, 2017). Dù vậy, thách thức vẫn tồn tại ở gán nhãn dữ liệu, khả năng giải thích

của mô hình và vấn đề đạo đức. Giải quyết các thách thức này đòi hỏi hợp tác liên ngành giữa ngôn ngữ học, khoa học máy tính và đạo đức học để triển khai AI một cách có trách nhiệm.

2.1.5. Lợi ích và hạn chế của AI trong nghiên cứu ngôn ngữ học

2.1.5.1. Lợi ích

- Hiệu quả và khả năng mở rộng: AI cho phép xử lí và phân tích kho ngữ liệu lớn hiệu quả; NLP tự động hoá trích xuất đặc trưng, rút ngắn đáng kể thời gian so với phương pháp thủ công.

- Độ chính xác và độ tin cậy: Mô hình ML huấn luyện trên dữ liệu gán nhãn học được mẫu hình phức tạp, đạt độ chính xác cao ở gán nhãn từ loại, phân tích cú pháp và ngữ nghĩa, cung cấp bằng chứng đáng tin cậy về cấu trúc và ý nghĩa (Vaswani, A. và cộng sự, 2017; Devlin, J. và cộng sự, 2019).

- Khai phá mẫu hình ẩn: AI phát hiện các quan hệ tinh vi, xu hướng tiềm ẩn trong biến thể, biến đổi ngôn ngữ và diễn ngôn mà phân tích thủ công khó nhận ra.

- Tự động hoá tác vụ: Công cụ AI hỗ trợ nhận dạng tiếng nói, dịch máy, kiểm tra ngữ pháp, tăng năng suất nghiên cứu và thực hành phân tích.

- Hợp tác liên ngành: Việc tích hợp AI thúc đẩy kết nối giữa ngôn ngữ học, khoa học máy tính, tâm lí học và các lĩnh vực liên quan (Roy, B., 2015) qua đó hình thành phương pháp mới.

2.1.5.2. Hạn chế

- Phụ thuộc dữ liệu gán nhãn: ML/DL yêu cầu dữ liệu đã gán nhãn quy mô lớn; quá trình này tốn kém và khó triển khai cho nhiều ngôn ngữ/hiện tượng, có thể gây thiên lệch và hạn chế khả năng khái quát (Ryant, N. và cộng sự, 2019)

- Thiên lệch mô hình: Dữ liệu không đại diện có thể dẫn đến mô hình tái tạo và khuếch đại định kiến, làm sai lệch kết quả.

- Vấn đề đạo đức: Dữ liệu ngôn ngữ có thể chứa thông tin nhạy cảm; cần quy trình bảo vệ quyền riêng tư, đồng thuận và đánh giá tác động xã hội của công nghệ (Pang, B., & Lee, L., 2008)

- Khó diễn giải: Mô hình sâu thiếu minh bạch, gây khó khăn khi lí giải quyết định và liên hệ với giả thuyết lí thuyết.

- Thách thức theo lĩnh vực: Kĩ thuật phát triển cho một ngôn ngữ/lĩnh vực có thể không chuyển giao trực tiếp; hiện tượng mơ hồ, ẩn dụ, sắc thái văn hoá tiếp tục là thách thức.

2.1.6. Ứng dụng AI trong phân tích ngôn ngữ học

AI cung cấp bộ công cụ mạnh để phân tích ngôn ngữ với độ chính xác, hiệu quả và chiều sâu cao thông qua NLP, ML và DL.

2.1.6.1. Phân tích ngữ pháp và cú pháp

NLP cho phép phân tích cấu trúc câu, xác định quan hệ cú pháp và mô hình hoá ngữ pháp. Các tác vụ gán nhãn từ loại, phân tích cú pháp cấu trúc và phân tích phụ thuộc đạt hiệu quả cao, hỗ trợ kiểm tra ngữ pháp tự động và phát hiện lỗi.

2.1.6.2. Phân tích ngữ nghĩa và phân biệt nghĩa từ

ML/DL hỗ trợ trích xuất thông tin ngữ nghĩa, suy luận nghĩa và xác lập quan hệ ngữ nghĩa. Các biểu diễn từ phân tán như Word2Vec, GloVe cho phép nắm bắt tương đồng và quan hệ ngữ nghĩa (Mikolov, T. và cộng sự, 2013; Pennington, J. và cộng sự, 2014). Các mô hình transformer đạt kết quả vượt trội trong gán vai ngữ nghĩa, phân tích cảm xúc, phân biệt nghĩa từ (Đinh Điền, 2019, tr.4171 - 4186).

2.1.6.3. Phân tích diễn ngôn

AI hỗ trợ phân tích cấu trúc, mạch lạc và kết nối diễn ngôn. Các tác vụ như phân giải đồng tham chiếu, phân đoạn diễn ngôn, phân tích cấu trúc diễn ngôn giúp nhận diện đơn vị và quan hệ. Phân tích cảm xúc và khai phá quan điểm hỗ trợ nhận diện thái độ, cảm xúc trong diễn ngôn (Devlin, J. và cộng sự, 2019).

2.1.7. Định hướng tương lai của AI trong nghiên cứu ngôn ngữ học

- Chất lượng và khả năng tiếp cận dữ liệu: Ưu tiên thu thập dữ liệu đa dạng, chất lượng cao; phát triển tăng cường và tổng hợp dữ liệu để khắc phục thiếu hụt.

- Tính diễn giải và minh bạch: Xây dựng kĩ thuật giải thích quyết định của mô hình, hỗ trợ kiểm nghiệm giả thuyết, chẩn đoán lỗi và hiểu hiện tượng.

- Vấn đề đạo đức: Tuân thủ quy trình bảo vệ quyền riêng tư và bảo mật; xác định và giảm thiểu thiên lệch để tránh hệ quả tiêu cực.

- Đa ngôn ngữ và biến thể: Phát triển mô hình đa ngôn ngữ và học chuyển giao để chuyển tri thức giữa ngôn ngữ/phương ngữ; nắm bắt cấu trúc và mẫu hình đa dạng.

- Hợp tác liên ngành: Khuyến khích hợp tác để tích hợp hiệu quả kĩ thuật AI vào nghiên cứu và thực hành ngôn ngữ học.

2.2. Một ứng dụng cụ thể trong nhận diện giọng nói thành văn bản

Khoa Ngôn ngữ học, Trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Hà Nội đã triển khai thực hiện chuỗi dự án liên quan đến tạo ngữ liệu và ghi âm tiếng Việt để huấn luyện cho công cụ chuyển đổi giọng nói thành văn bản trên thiết bị điện thoại di động. Dự án tập trung vào xây dựng kho ngữ liệu tiếng Việt (trước tiên tập trung vào các đặc điểm từ vựng, ngữ pháp, ngữ âm của phương ngữ Bắc), ghi âm ngữ liệu từ giọng nói của nhiều người khác nhau, cả nam và nữ, sau đó đánh giá hiệu suất của hệ thống ứng dụng nhận diện giọng nói trên thiết bị điện thoại S24 Ultra của Samsung. Chúng tôi sẽ trình bày chi tiết phương pháp, bao gồm tạo ngữ liệu, quy trình ghi âm, nhận diện giọng nói thành văn bản, nhận diện - phân tích lỗi sai, giải pháp giảm thiểu tỉ lệ lỗi.

2.2.1. Tạo dữ liệu ngôn ngữ

Dữ liệu sử dụng để ghi âm và huấn luyện cho ứng dụng là các câu của tiếng Việt. Trong dự án này, khoa Ngôn ngữ học đã tạo 200.000 câu, mỗi câu có đặc điểm như sau: độ dài mỗi câu từ 5 đến 30 hình vị, nên bao gồm ít nhất một trong số các từ như tên riêng, từ đồng âm, từ theo trào lưu mới, đơn vị tiền tệ, đơn vị đo lường, từ viết tắt, từ mượn, ngày tháng năm, từ láy… Tổng 200.000 câu sẽ được chia nhỏ thành các gói câu theo các chủ đề: chính trị, sức khỏe, triết học, kinh tế, trí tuệ nhân tạo, văn học, công nghệ, môi trường, lịch sử, địa lí, tâm lí, thể thao, ẩm thực, tên toà nhà, tên đường, tên người nổi tiếng,…

Dữ liệu ngôn ngữ đến từ 2 nguồn khác nhau:

- Lấy từ kho dữ liệu có sẵn như: sách, báo, truyện, tạp chí…

- Tạo sinh câu mới từ các ứng dụng trí tuệ nhân tạo như Chat GPT, Gemini. Cung cấp cho ứng dụng từ khoá và cấu trúc của các mẫu câu để ứng dụng tạo ra câu mới theo yêu cầu.

2.2.2. Chuẩn hoá ngữ liệu

Chuẩn hoá ngữ liệu đang ở dạng chữ viết thành dạng lời nói, bao gồm:

- Kí hiệu "%, &, @..." chuyển thành chữ dạng nói: phần trăm, và, a còng…

- Chữ số, số La mã "5, V, X…" chuyển thành chữ dạng nói: năm, năm, mười…

- Từ viết tắt trong tiếng Việt "TW, CP, NĐ, PCCC…" chuyển thành chữ dạng nói: trung ương, chính phủ, nghị định, phòng cháy chữa cháy…

- Ngày, tháng, năm, chuyển dạng chữ thành dạng lời nói.

2.2.3. Ghi âm ngữ liệu trên phần mềm Audacity

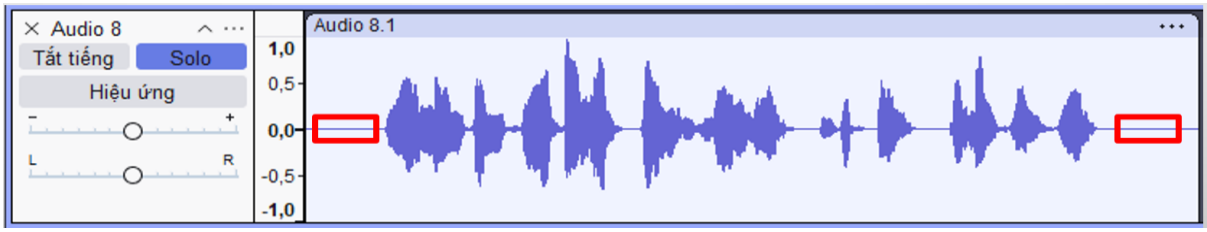
- Người ghi âm: sinh viên của trường Đại học Khoa học Xã hội và Nhân văn; nói giọng phương ngữ Bắc, không ngọng; bao gồm 15 giọng nam và 15 giọng nữ.

- Phần mềm ghi âm Audacity: đây là một phần mềm ghi âm miễn phí, phổ biến, dễ sử dụng, được lựa chọn để ghi âm từng câu của dự án nà y.

Hình 1. Biểu tượng phần mềm Audacity

- Cách ghi âm: người đọc sử dụng phần mềm ghi âm từng câu trọn vẹn, mỗi câu là một bản ghi, phần đầu và phần cuối để trống 1 giây để máy nhận diện được trọn vẹn tất cả các tiếng trong câu đó.

- Sản phẩm sau ghi âm: 50.000 câu có giọng nam hoặc nữ, 150.000 câu có cả giọng nam và giọng nữ. Tổng số lượng dữ liệu thu được là 350.000 bản ghi âm.



Hình 2. Sóng âm thu được khi ghi âm hoàn thiện 1 câu

2.2.4. Nhận diện giọng nói thành văn bản

- 350.000 bản ghi âm từng câu sẽ được nhập vào công cụ để nhận diện giọng nói thành văn bản. Ứng dụng này có tên là Speech to Text - chuyển giọng nói thành văn bản, là một nhánh của công nghệ ASR - Automatic Speech Recognition - Nhận dạng giọng nói tự động.

- Kết quả nhận diện giọng nói thành văn bản sẽ được đặt bên cạnh câu gốc đã chuẩn hoá để phục vụ cho việc đối chiếu và đánh dấu lỗi sai.

TC ID	Input voice script	Output VR ASR Text
	đặc biệt ở những nước gam hai mươi trong năm qua số vụ bảo hộ thương mại gia tăng kỉ lục	đặc biệt ở những nước gờ hai mươi trong năm qua số vụ bảo hộ thương mại gia tăng kỷ lục
	chưa dừng lại ở đó năm hai nghìn không trăm linh hai là năm bong bóng dotcom xì hơi	chưa dừng lại ở đó năm hai nghìn không trăm linh hai là năm bong bóng đét com xì hơi
	sốt tiền ảo bitcoin giống bong bóng dotcom năm xưa	sốt tiền ảo bitcoin giống bong bóng đốt con năm xưa
	mặc dù tăng trưởng cả xuất và nhập khẩu đều không nhiều nhưng cán cân thương mại vẫn	mặc dù tăng trưởng sản xuất và nhập khẩu đều không nhiều nhưng cán cân thương mại vẫ
	riêng về chi đầu tư phát triển tổng số kế hoạch vốn năm hai nghìn không trăm hai ba thủ tu	riêng về chi đầu tư phát triển tổng số kế hoạch vốn năm hai nghìn không trăm hai ba thủ tu
	phần đấu đến năm hai nghìn không trăm hai mươi việt nam cắt giảm được năm mươi phần	phần đầu đến năm hai nghìn không trăm hai mươi việt nam cắt giảm được năm mươi phần
	làn sóng lãi suất âm chính sách lãi suất âm được đưa vào áp dụng lần đầu tiên tại thuỵ điể	làn sóng lãi suất âm chính sách loại xuất âm được đưa vào áp dụng lần đầu tiên tại thụy đi
	cổ phiếu trong ngày giao dịch không hưởng quyền để trả cổ tức hoặc thưởng bằng cổ phiếu	cổ phiếu trong ngành giao dịch không ngừng quyền để trả cổ tức hoặc thưởng bằng cổ phiế
	anh cũng không nhận tiền thưởng hay cổ phiếu thưởng từ công ty	anh cũng không nhận tiền thường hãy cổ phiếu thưởng từ công ty
	cổ phiếu này chốt phiên ở mức kịch trần sau thông tin chia cổ phiếu thưởng	cổ phiếu này chốt phiên ở mức kịch trần do thông tin chưa cổ phiếu thưởng
	sẽ xuất hiện cơn lũ cổ phiếu thưởng	sẽ xuất hiện cơn lũ của phiếu thưởng
	một là từ các khoản cổ tức hoặc lãi từ các khoản đầu tư của quỹ	một là từ các khoản cổ tức hoặc lãi được các khoản đầu tư của quỹ
	cụ thể công ty chuyển hình thức trả cổ tức cho năm hai nghìn không trăm hai mươi từ tiền	cụ thể công ty truyền hình nước trả cổ tức cho năm hai nghìn không trăm hai mươi từ tiền
	ông cho rằng việc các nước gam bảy cùng cung cấp dòng tiền dưới cùng một mái nhà sẽ ma	ông cho rằng việc các nước g bảy cùng cung cấp dòng tiền dưới cùng một mái nhà sẽ mang
	mặt khác nhà đầu tư không nên chọn mục đích đầu tư ngắn hạn mà nên cân nhắc cho mục	mặt khác nhà đầu tư không nên chọn mục đích đầu tư ngắn hạn mà nên cân nhắc cho mục
	vì thế giới đầu tư quốc tế đổ đến	vì thế giới đầu tư quốc tế đồ đến
	doanh nghiệp khó khăn vì giá hàng hoá đầu vào tăng cao đầu tư xây dựng cơ bản không có	doanh nghiệp khó khăn vì giá hàng hóa đầu vào tăng cao đầu tư xây dựng cơ bản không cò
	tính từ giữa tháng tư đến nay đồng nhân dân tệ đã mất gần mười phần trăm giá trị so với đ	tính từ giữa tháng tư đến nay đồng nhân dân tệ đã mất gần mười phần trăm giá trị so với đ
	như vậy uỷ ban giám sát tài chính quốc gia có bốn phó chủ tịch chia sẻ trách nhiệm cùng ch	như vậy ủy ban giám sát tài chính quốc gia có bốn phó chủ tịch chia sẻ trách nhiệm cùng ch
	một chỉ số quan trọng khác đo lường sức mạnh của một loại tiền tệ là khối lượng giao dịch	một chỉ số quan trọng khác đơ lường sinh mệnh của một loại tiền tệ là khối lượng giao dịch
	giá vàng biến động theo diễn biến đàm phán gói kích thích kinh tế tại mĩ một chuyên gia nh	giá vàng biến động theo diễn biến đàm phán có kích thích kinh tế tại mỹ một chuyên gia nh
	nguyên nhân các mặt hàng xuất khẩu chủ lực của thành phố đều giảm nhất là gạo và cá tra	nguyên nhân các mặt hàng xuất khẩu chủ lực của thành phố đều giảm nhất là gạo và cá ch
	vay ngắn hạn là năm nghìn hai trăm bốn mươi hai tỉ đồng các khoản phải trả giãn hoãn tha	vay ngắn hạn là năm nghìn hai trăm bốn mươi hai tỷ đồng các khoản phải trả hoãn mức bố
	các khu vực mậu dịch tự do việt nam tham gia	các khu vực mộ dịch tự do việt nam tham gia
	giảng viên giảng dạy kinh tế học vi mô kinh tế học vĩ mô lịch sử học thuyết kinh tế	giảng viên giảng dạy kinh tế học vi mô kinh tế học vĩ mô lịch sử học kinh tế

Hình 3. Văn bản gốc (bên trái) đối chiếu với kết quả nhận diện giọng nói của máy (bên phải)

2.2.5. Nhận diện lỗi sai và phân loại lỗi

- Số lượng câu có lỗi sai nhận diện được: 50.009/350.000 bản ghi, chiếm 14,2%.

- Sau khi đối chiếu câu gốc và câu nhận diện được do máy, chúng tôi sẽ đánh dấu những chỗ nhận diện sai trước khi tiến hành phân loại lỗi sai.

Input voice script	Output VR ASR Text	Diffirence (miss by Input) - (reduntdance by output)
đặc biệt ở những nước gam hai mươi trong năm qua số vụ bảo hộ thương mại gia tăng kỉ lục	đặc biệt ở những nước gờ hai mươi trong năm qua số vụ bảo hộ thương mại gia tăng kỷ lục	đặc biệt ở những nước gam gờ hai mươi trong năm qua số vụ bảo hộ thương mại gia tăng kỉ kỷ lục
chưa dừng lại ở đó năm hai nghìn không trăm linh hai là năm bong bóng dotcom xì hơi	chưa dừng lại ở đó năm hai nghìn không trăm linh hai là năm bong bóng đét com xì hơi	chưa dừng lại ở đó năm hai nghìn không trăm linh hai là năm bong bóng dotcom đét com xì hơi
sốt tiền ảo bitcoin giống bong bóng dotcom năm xưa	sốt tiền ảo bitcoin giống bong bóng đốt con năm xưa	sốt tiền ảo bitcoin giống bong bóng dotcom đốt con năm xưa
mặc dù tăng trưởng cả xuất và nhập khẩu đều không nhiều nhưng cán cân thương mại vẫn	mặc dù tăng trưởng sản xuất và nhập khẩu đều không nhiều nhưng cán cân thương mại vẫ	mặc dù tăng trưởng cả sản xuất và nhập khẩu đều không nhiều nhưng cán cân thương mại vẫn đi ngang trong tháng bảy
riêng về chi đầu tư phát triển tổng số kế hoạch vốn năm hai nghìn không trăm hai ba thủ tu	riêng về chi đầu tư phát triển tổng số kế hoạch vốn năm hai nghìn không trăm hai ba thủ tu	riêng về chi đầu tư phát triển tổng số kế hoạch vốn năm hai nghìn không trăm hai ba thủ tướng chính phủ đã giao khoảng bảy trăm linh bảy nghìn tỉ tỷ đồng
phấn đấu đến năm hai nghìn không trăm hai mươi việt nam cắt giảm được năm mươi phần	phần đấu đến năm hai nghìn không trăm hai mươi việt nam cắt giảm được năm mươi phần	phấn phần đấu đến năm hai nghìn không trăm hai mươi việt nam cắt giảm được năm mươi phần trăm chi phí sử dụng thuốc bảo vệ thực vật
làn sóng lãi suất âm chính sách lãi suất âm được đưa vào áp dụng lần đầu tiên tại thuy đi	làn sóng lãi suất âm chính sách loại xuất âm được đưa vào áp dụng lần đầu tiên tại thuy đi	làn sóng lãi suất âm chính sách lãi loại suất xuất âm được đưa vào áp dụng lần đầu tiên tại thuy thụy điển vào năm hai nghìn không trăm linh chín
cổ phiếu trong ngày giao dịch không hưởng quyền để trả cổ tức hoặc thưởng bằng cổ phiếu	cổ phiếu trong ngành giao dịch không ngừng quyền để trả cổ tức hoặc thưởng bằng cổ phiế	cổ phiếu trong ngày ngành giao dịch không hưởng ngừng quyền để trả cổ tức hoặc thưởng bằng cổ phiếu quỹ cho cổ đông hiện hữu
anh cũng không nhận tiền thưởng hay cổ phiếu thưởng từ công ty	anh cũng không nhận tiền thưởng hãy cổ phiếu thưởng từ công ty	anh cũng không nhận tiền thưởng hay hãy cổ phiếu thưởng từ công ty
cổ phiếu này chốt phiên ở mức kịch trần sau thông tin chia cổ phiếu thưởng	cổ phiếu này chốt phiên ở mức kịch trần do thông tin chưa cổ phiếu thưởng	cổ phiếu này chốt phiên ở mức kịch trần sau do thông tin chia chưa cổ phiếu thưởng
sẽ xuất hiện cơm lũ cổ phiếu thưởng	sẽ xuất hiện cơm lũ của phiếu thưởng	sẽ xuất hiện cơm lũ cổ của phiếu thưởng
một là từ các khoản cổ tức hoặc lãi từ các khoản đầu tư của quỹ	một là từ các khoản cổ tức hoặc lãi được các khoản đầu tư của quỹ	một là từ các khoản cổ tức hoặc lãi từ được các khoản đầu tư của quỹ
cụ thể công ty chuyển hình thức trả cổ tức cho năm hai nghìn không trăm hai mươi từ tiền	cụ thể công ty truyền hình nước trả cổ tức cho năm hai nghìn không trăm hai mươi từ tiền	cụ thể công ty chuyển truyền hình thức nước trả cổ tức cho năm hai nghìn không trăm hai mươi từ tiền mặt sang cổ phiếu với tỉ tỷ lệ ba mươi phần trăm
ông cho rằng việc các nước gam bảy cùng cung cấp dòng tiền dưới cùng một mái nhà sẽ ma	ông cho rằng việc các nước g bảy cùng cung cấp dòng tiền dưới cùng một mái nhà sẽ mang	ông cho rằng việc các nước gam g bảy cùng cung cấp dòng tiền dưới cùng một mái nhà sẽ mang lại lợi thế
mặt khác nhà đầu tư không nên chọn mục đích đầu tư ngắn hạn mà nên cân nhắc cho mục	mặt khác nhà đầu tư không nên chọn mục đích đầu tư ngắn hạn mà nên cân nhắc cho mục	mặt khác nhà đầu tư không nên chọn mục đích đầu tư ngắn hạn mà nên cân nhắc cho mục đích đích trung chung và dài hạn
vì thế giới đầu tư quốc tế đổ đến	vì thế giới đầu tư quốc tế đổ đến	vì thế giới đầu tư quốc tế đổ đổ đến
doanh nghiệp khó khăn vì giá hàng hoá đầu vào tăng cao đầu tư xây dựng cơ bản không có	doanh nghiệp khó khăn vì giá hàng hóa đầu vào tăng cao đầu tư xây dựng cơ bản không còn	doanh nghiệp khó khăn vì giá hàng hoá hóa đầu vào tăng cao đầu tư xây dựng cơ bản không có còn lãi vận vẫn tải thua lỗ
tính từ giữa tháng tư đến nay đồng nhân dân tệ đã mất gần mười phần trăm giá trị so với đ	tính từ giữa tháng tư đến nay đồng nhân dân tệ đã mất gần mười phần trăm giá trị so với đ	tính từ giữa tháng tư đến nay đồng nhân dân tệ đã mất gần mười phần trăm giá trị so với đồng bạc bằng xanh của mĩ mỹ
như vậy uỷ ban giám sát tài chính quốc gia có bốn phó chủ tịch chia sẻ trách nhiệm cùng cl	như vậy ủy ban giám sát tài chính quốc gia có bốn phó chủ tịch chia sẻ trách nhiệm cùng cl	như vậy uỷ ủy ban giám sát tài chính quốc gia có bốn phó chủ tịch chia sẻ trách nhiệm cùng chủ tịch vũ viết ngoạn ngoại
một chỉ số quan trọng khác đo lường sức mạnh của một loại tiền tệ là khối lượng giao dịch	một chỉ số quan trọng khác đơ lường sinh mệnh của một loại tiền tệ là khối lượng giao dịch	một chỉ số quan trọng khác đo đơ lường sức sinh mạnh mệnh của một loại tiền tệ là khối lượng giao dịch ngoại hối
giá vàng biến động theo diễn biến đàm phán gói kích thích kinh tế tại mĩ một chuyên gia nh	giá vàng biến động theo diễn biến đàm phán có kích thích kinh tế tại mỹ một chuyên gia nh	giá vàng biến động theo diễn biến đàm phán gói có kích thích kinh tế tại mĩ mỹ một chuyên gia nhận định
nguyên nhân các mặt hàng xuất khẩu chủ lực của thành phố đều giảm nhất là gạo và cá tra	nguyên nhân các mặt hàng xuất khẩu chủ lực của thành phố đều giảm nhất là gạo và cá ch	nguyên nhân các mặt hàng xuất khẩu chủ lực của thành phố đều giảm nhất là gạo và cá tra cá cha
vay ngắn hạn là năm nghìn hai trăm bốn mươi hai tỉ đồng các khoản phải trả giãn hoãn tha	vay ngắn hạn là năm nghìn hai trăm bốn mươi hai tỷ đồng các khoản phải trả hoãn mức bố	vay ngắn hạn là năm nghìn hai trăm bốn mươi hai tỉ tỷ đồng các khoản phải trả giãn hoãn thanh toán ở mức bốn nghìn hai trăm sáu tám tỉ tỷ đồng
các khu vực mậu dịch tự do việt nam tham gia	các khu vực mộ dịch tự do việt nam tham gia	các khu vực mậu mộ dịch tự do việt nam tham gia
giảng viên giảng dạy kinh tế học vĩ mô kinh tế học vĩ mô lịch sử học thuyết kinh tế	giảng viên giảng dạy kinh tế học vĩ mô kinh tế học vĩ mô lịch sử học kinh tế	giảng viên giảng dạy kinh tế học vĩ mô lịch sử học thuyết kinh tế

Hình 4. Các lỗi nhận diện giọng nói được đánh dấu và phân loại theo màu

- Lỗi sai được phân loại thành các kiểu sau: Thừa chữ; Thiếu chữ; Tiếng nhận sai thanh điệu; Tiếng phát âm tương tự; Đồng âm; Nhận sai từ viết tắt; Từ nhận sai vị trí; Nhận sai từ là từ chỉ đơn vị.

2.2.6. Giải pháp giảm thiểu tỉ lệ lỗi

Từ những lỗi sai nhận được, dự án chủ động tạo ra các ngữ liệu có chứa lỗi nhận diện sai trước đó để máy học đi học lại.

Ví dụ máy nhận diện từ "tinh giản" thành "tinh giảm", dự án tạo ra loạt 30 câu có từ "tinh giản" để máy học, phân tích văn cảnh của câu có chứa từ này và nghe đi nghe lại cách phát âm của âm tiết "giản" trong từ "tinh giản":

- *Quy trình sản xuất được tinh giản để tiết kiệm thời gian.*
- *Công ty áp dụng phương pháp tinh giản nhân sự hiệu quả.*
- *Tinh giản thủ tục hành chính giúp dân chúng thuận tiện.*
- *Hệ thống quản lí đã được tinh giản đáng kể.*
- *Tinh giản chi phí vận hành là mục tiêu chính.*
- *Cô ấy trình bày ý tưởng một cách tinh giản.*
- *Phần mềm mới giúp tinh giản quy trình làm việc.*
- *Tinh giản nội dung bài giảng tăng hiệu quả học tập.*
- *Chúng tôi cần tinh giản kế hoạch để thực hiện nhanh.*
- *Tinh giản thiết kế sản phẩm giúp giảm giá thành.*

- *Chính sách mới nhằm tinh giản các quy định phức tạp.*
- *Đội ngũ lãnh đạo quyết định tinh giản bộ máy tổ chức.*
- *Tinh giản dữ liệu giúp hệ thống hoạt động mượt mà.*
- *Bài thuyết trình được tinh giản để dễ hiểu hơn.*
- *Tinh giản quy trình kiểm tra nâng cao chất lượng dịch vụ.*
- *Tinh giản quy trình sản xuất giúp tiết kiệm thời gian và chi phí vận hành.*
- *Công ty áp dụng phương pháp tinh giản nhân sự để nâng cao hiệu quả quản lí.*
- *Tinh giản thủ tục hành chính giúp người dân tiếp cận dịch vụ nhanh chóng hơn.*
- *Hệ thống quản lí được tinh giản để tối ưu hoá hiệu suất và giảm chi phí.*
- *Tinh giản chi phí vận hành là mục tiêu trọng tâm của doanh nghiệp hiện nay.*
- *Cô ấy trình bày ý tưởng một cách tinh giản để thu hút sự chú ý của mọi người.*
- *Phần mềm mới hỗ trợ tinh giản quy trình làm việc và tăng năng suất lao động.*
- *Tinh giản nội dung bài giảng giúp học sinh tiếp thu kiến thức hiệu quả hơn.*
- *Chúng tôi cần tinh giản kế hoạch dự án để đảm bảo hoàn thành đúng thời hạn.*
- *Tinh giản thiết kế sản phẩm giúp giảm chi phí và cải thiện trải nghiệm người dùng.*
- *Chính sách mới tập trung tinh giản các quy định để thúc đẩy hoạt động kinh doanh.*
- *Đội ngũ lãnh đạo quyết định tinh giản bộ máy tổ chức để tăng hiệu quả quản trị.*
- *Tinh giản dữ liệu hệ thống giúp cải thiện tốc độ xử lí và độ chính xác cao.*
- *Bài thuyết trình được tinh giản để truyền tải thông điệp rõ ràng và dễ hiểu hơn.*
- *Tinh giản quy trình kiểm tra chất lượng giúp nâng cao hiệu suất sản xuất nhanh.*

Quy trình của dự án cho thấy sự kết hợp hiệu quả giữa tri thức ngôn ngữ học và công nghệ hiện đại trong việc phát triển ứng dụng nhận diện giọng nói tiếng Việt. Quan trọng hơn, quá trình này khẳng định vai trò thiết yếu của việc xử lí và chuẩn hoá ngữ liệu trong huấn luyện mô hình trí tuệ nhân tạo. Việc phát hiện và phân tích lỗi không chỉ giúp cải thiện hiệu suất của hệ thống mà còn góp phần làm rõ những đặc trưng ngữ âm - ngữ nghĩa riêng của tiếng Việt, vốn là thách thức lớn đối với công nghệ nhận dạng giọng nói. Từ đây, dự án mở ra hướng nghiên cứu mới, nơi ngôn ngữ học không chỉ dừng ở lí thuyết mà còn trực tiếp đóng góp vào các ứng dụng công nghệ, tạo nền tảng cho những sản phẩm công nghệ có liên quan đến ngôn ngữ, có tính ứng dụng cao và phản ánh được đặc trưng vốn có của ngôn ngữ dân tộc trong kỉ nguyên số.

3. Kết luận

Có thể nói, việc tích hợp AI vào ngôn ngữ học mở ra kỉ nguyên khám phá mới, đổi mới cách tiếp cận với tiếp thu và phân tích ngôn ngữ. Thông qua NLP, ML, DL, các nghiên cứu ngôn ngữ học đạt mức độ chính xác và chiều sâu, từ mô hình hoá phát triển ngôn ngữ đến phân tích cấu trúc cú pháp, ngữ nghĩa, ngữ dụng (Elma Sibonghanoy Groenewald, 2024). AI chứng tỏ bản thân nó là công cụ mạnh để mô phỏng cơ chế nhận thức trong tiếp thu và kiểm nghiệm khung lí thuyết bằng dữ liệu. Đồng thời, AI cách mạng hoá phân tích ngôn ngữ, giúp khai phá mô hình ẩn và cấu trúc ngữ nghĩa khó nhận biết bằng thủ công. Tuy nhiên, một số thách thức vẫn đang hiện hữu như chất lượng dữ liệu, khả năng diễn giải, đạo đức, đa ngôn ngữ và nhu cầu hợp tác liên ngành. Việc khắc phục các vấn đề này cần sự phối hợp đa lĩnh vực, tuân thủ chuẩn mực đạo đức và đổi mới phương pháp. Những tiến bộ của AI cùng hợp tác liên ngành sẽ mở ra triển vọng tương lai nhằm tiếp tục làm sâu sắc hiểu biết về tiếp thu, phân tích và sử dụng ngôn ngữ.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Đinh Điền (2006). *Xử lí ngôn ngữ tự nhiên*. NXB Đại học Quốc gia thành phố Hồ Chí Minh.
2. Đinh Điền và Hồ Xuân Vinh. (2016). *Ứng dụng Kho Ngữ liệu trong việc dạy tiếng Việt cho người nước ngoài*. Kỉ yếu hội thảo quốc tế Giảng dạy, Nghiên cứu Việt Nam học và tiếng Việt, 172-180.
3. Đinh Điền (2018). *Ngôn ngữ học ngữ liệu.* NXB Đại học Quốc gia Thành phố Hồ Chí Minh.
4. Đinh Điền (2019). *Ứng dụng ngôn ngữ học tính toán trong nghiên cứu và giảng dạy ngôn ngữ.* Hội thảo Khoa học Quốc gia, Hà Nội.

Tiếng nước ngoài

5. Attali, Y., & Burstein, J. (2006). *Đánh giá bài viết tự động với e-rater® V. 2.0*. The Journal of Technology, Learning and Assessment, 4(3).
6. Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). *Về các đặc tính của dịch máy thần kinh: Các phương pháp Encoder-decoder*. arXiv preprint arXiv:1409.1259.
7. Chomsky, N. (1957). *Cấu trúc cú pháp*. Mouton de Gruyter.
8. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Tiền huấn luyện mô hình biến đổi hai chiều sâu cho việc hiểu ngôn ngữ tự nhiên*. Trong Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).
9. Elma Sibonghanoy Groenewald et al. (2024). *Trí tuệ nhân tạo trong nghiên cứu Ngôn ngữ học: Ứng dụng trong việc tiếp thu và phân tích ngôn ngữ*. Naturalista Campano. Volume 28 Issue 1.
10. Elman, J. L. (1990). *Tìm cấu trúc trong thời gian*. Cognitive Science, 14(2), 179-211.
11. Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). *Một khuôn khổ Bayesian cho phân đoạn từ: Khám phá tác động của ngữ cảnh*. Cognition, 112(1), 21-54.
12. Graves, A., Schmidhuber, J., & Mohamed, A. (2006). *Nhận dạng chữ viết tay ngoại tuyến với mạng nơ-ron hồi quy đa chiều*. Advances in neural information processing systems, 18, 545-552.
13. Graves, A., Wayne, G., & Danihelka, I. (2014). *Các máy Turing thần kinh*. arXiv preprint arXiv:1410.5401.
14. Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). *Mạng hồi quy xanh không màu mơ mộng theo cách phân cấp*. Trong Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 1195-1205).
15. Hinton, G. E., & Salakhutdinov, R. R. (2006). *Giảm chiều của dữ liệu bằng mạng nơ-ron*. Science, 313(5786), 504-507.
16. Jurafsky, D., & Martin, J. H. (2019). *Xử lí tiếng nói và ngôn ngữ: Giới thiệu về xử lí ngôn ngữ tự nhiên, ngôn ngữ học tính toán, và nhận dạng giọng nói (tái bản lần thứ 3)*. Pearson.
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Đại diện phân tán của từ và cụm từ và tính khả thi của chúng*. Trong Advances in neural information processing systems (pp. 3111-3119).
18. Pang, B., & Lee, L. (2008). *Khai thác ý kiến và phân tích cảm xúc*. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.
19. Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Các vectơ toàn cầu cho đại diện từ.* Trong Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532 -1543).
20. Pinker, S. (1984). *Khả năng học và phát triển ngôn ngữ*. Harvard University Press.
21. Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). *Dự đoán sự ra đời của một từ khẩu ngữ*. Proceedings of the National Academy of Sciences, 112(41), 12663-12668.