CÁC BIỆN PHÁP ĐÁNH GIÁ MỨC ĐỘ NGHIÊM TRỌNG ĐỂ ĐÁNH GIÁ LỖI TRONG NHẬN DẠNG GIỌNG NÓI TỰ ĐỘNG

qua

Ryan Whetten

Một luận án

nộp một phần để hoàn thành

của các yêu cầu về mức độ

Thạc sĩ Khoa học Máy tính

Đại học Boise State

Tháng 5 năm 2023

TRƯỜNG CAO ĐẲNG ĐẠI HỌC BANG BOISE

ỦY BAN QUỐC PHÒNG VÀ PHÊ DUYỆT ĐỌC CUỐI CÙNG

của luận án được nộp bởi

Ryan Whetten

Tiêu đề luận án: Các biện pháp đánh giá mức độ nghiêm trọng để đánh giá lỗi trong nhận dạng giọng nói tự động

Ngày thi vấn đáp cuối kỳ : 01 tháng 5 năm 2023

Những cá nhân sau đây đã đọc và thảo luận luận án do sinh viên Ryan nộp Whetten và họ đã đánh giá bài thuyết trình và phản hồi các câu hỏi trong kỳ thi vấn đáp cuối cùng. Họ thấy rằng sinh viên đã vượt qua kỳ thi vấn đáp cuối cùng.

Casey Kennington, Tiến sĩ	Chủ tịch, Ủy ban giám sát
Tiến sĩ Tim Andersen	Thành viên, Ủy ban giám sát
Tiến sĩ Michael Ekstrand	Thành viên, Ủy ban giám sát

Sự chấp thuận đọc cuối cùng của luận án đã được chấp thuận bởi Casey Kennington, Tiến sĩ, Chủ tịch Ủy ban giám sát. Luận án đã được Hội đồng sau đại học chấp thuận

Trường cao đẳng.

dành tặng cho anh trai tôi, Andrew Whetten

LỜI CẢM ƠN

Trước tiên tôi muốn bày tỏ lòng biết ơn tới cố vấn của tôi, Casey Kennington, người đã là tấm gương cho tôi trong suốt thời gian học Thạc sĩ. Qua nhiều cuộc họp, email, bài giảng và cuộc trò chuyện Casey đã định hình tôi theo hướng tốt hơn, không chỉ là giáo viên và cố vấn, mà còn là tấm gương tuyệt vời bên ngoài khuôn viên trường đại học. Casey đã giúp tôi tăng cường sự hiểu biết của mình trong xử lý ngôn ngữ tự nhiên, biến tôi từ một sinh viên đầy tham vọng nhưng không biết gì, trở thành một người có thể hiểu được và bắt đầu tiến hành nghiên cứu trong lĩnh vực này.

Tôi cũng muốn cảm ơn Tim Andersen và Michael Ekstrand đã dành thời gian và nỗ lực trong việc tham gia ủy ban của tôi và trở thành giáo viên chính của tôi mặc dù năm và 8 tháng qua. Tất cả các thành viên trong ủy ban của tôi đã đóng một vai trò trong dạy tôi cách suy nghĩ sâu sắc hơn, cách học độc lập và cách ứng xử nghiên cứu.

Tôi vô cùng biết ơn Kavan Hess, Rachael Powell và Ryan Roper đã tham gia thời gian nghỉ học y khoa đã giúp tôi thực hiện nghiên cứu này.

Cuối cùng, tôi muốn cảm ơn gia đình tôi, đặc biệt là vợ và anh trai tôi, những người đã là đội cổ vũ của tôi và cũng là nguồn động lực cho những ý tưởng và nguồn cảm hứng của tôi để tiếp tục đang đi. Thực tế, đó là trong một cuộc trò chuyện với anh trai tôi khi ý tưởng đó xuất hiện vào luận án này đã ra đời.

TÓM TẮT

Một số liệu chung để đánh giá Nhận dạng giọng nói tự động (ASR) là Word
Tỷ lệ lỗi (WER) chỉ tính đến sự khác biệt ở cấp độ từ.
Mặc dù WER hữu ích, nhưng không đảm bảo nó sẽ tương quan tốt với khả năng hiểu được
hoặc hiệu suất trên các tác vụ hạ nguồn sử dụng ASR. Đánh giá có ý nghĩa-
việc khắc phục lỗi ASR trở nên quan trọng hơn trong các tình huống có rủi ro cao như
như chăm sóc sức khỏe. Tôi đề xuất 2 biện pháp chung để đánh giá chất lượng hoặc mức độ nghiêm trọng của
những sai lầm do hệ thống ASR gây ra, một dựa trên phân tích tình cảm và một dựa trên
về nhúng văn bản. Cả hai đều có khả năng khắc phục những hạn chế của WER.
Tôi đánh giá các biện pháp này trên các cuộc trò chuyện mô phỏng giữa bệnh nhân và bác sĩ. Các biện pháp
mức độ nghiêm trọng dựa trên xếp hạng tình cảm và nhúng văn bản tương quan với xếp hạng của con người
mức độ nghiêm trọng. Các biện pháp dựa trên nhúng văn bản có khả năng dự đoán con người
xếp hạng mức độ nghiêm trọng tốt hơn WER. Các biện pháp này được sử dụng trong số liệu thống kê trong tổng thể
đánh giá 5 động cơ ASR cùng với WER. Kết quả cho thấy các số liệu này nắm bắt
đặc điểm của lỗi ASR mà WER không có. Hơn nữa, tôi đào tạo một hệ thống ASR
sử dụng mức độ nghiêm trọng như một hình phạt trong hàm mất mát và chứng minh tiềm năng sử dụng
mức độ nghiêm trọng không chỉ trong việc đánh giá mà còn trong quá trình phát triển ASR. Ưu điểm và
Những hạn chế của phương pháp này được phân tích và thảo luận.

MỤC LỤC

DANH SÁCH CÁC BẢNG

DANH SÁCH CÁC HÌNH ẢNH

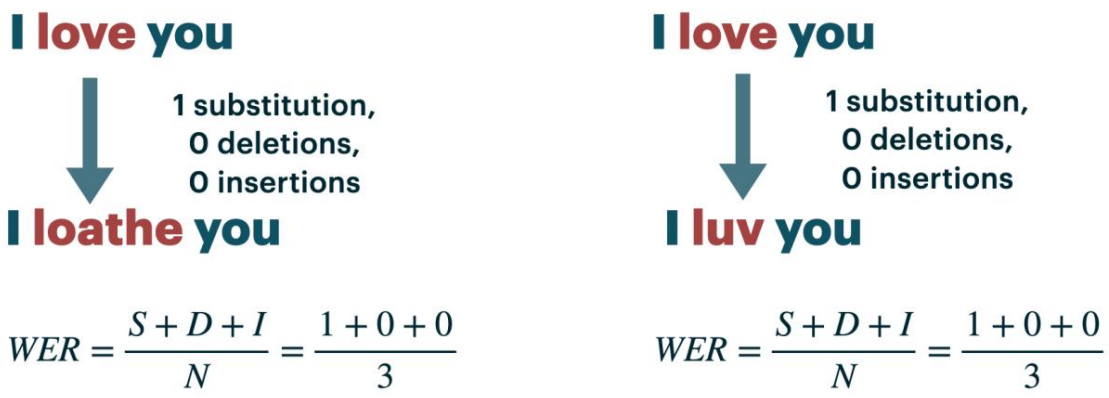
CHƯƠNG 1

GIỚI THIỆU

Nhận dạng giọng nói tự động (ASR) là nhiệm vụ xử lý giọng nói của con người thành văn bản. ASR đã được cải thiện đáng kể trong thập kỷ qua và đã cách mạng hóa cách nhiều người tương tác với máy tính bằng các ứng dụng như tìm kiếm bằng giọng nói, đọc chính tả và trợ lý ảo (ví dụ Siri của Apple, Bixby của Samsung, Google Now) [56, 2]. Thực hành phổ biến là đánh giá các hệ thống ASR bằng cách tính toán lỗi từ tỷ lệ (WER). WER có thể được tính bằng cách đếm số lượng từ cần phải thay thế (S), xóa (D) và chèn (I) để đi từ một sự thật cơ bản phiên mã của con người thành đầu ra của ASR. Số đếm này sau đó được chia cho tổng số số lượng từ trong bản ghi chép sự thật cơ bản (N) [32]. WER thường được viết là (S + I + D)/N. Về cơ bản, WER xử lý từng sự khác biệt giữa sự thật cơ bản phiên mã và đầu ra từ ASR là như nhau.

Tuy nhiên, một vấn đề là không phải tất cả các lỗi ASR đều như nhau. Ví dụ, hãy lấy câu “Tôi yêu bạn,” và giả sử hệ thống ASR tạo ra “Tôi ghét bạn.” Điều này sẽ dẫn đến WER là 0,33. Bây giờ chúng ta hãy giả sử một hệ thống ASR khác dự đoán “Tôi yêu bạn.” Điều này cũng sẽ dẫn đến WER là 0,33 (xem Hình 1.1). Mặc dù WERs là bình đẳng, so với sự thật cơ bản "Anh yêu em", lỗi "luv" là có thể nói là ít nghiêm trọng hơn lỗi “ghê tởm”, có nghĩa ngược lại từ câu sự thật cơ bản. Ví dụ này, với hai lỗi khác nhau và cùng một

WER, cho thấy cách con người có thể nhận thức được mức độ nghiêm trọng của lỗi trong quá trình phiên mã sẽ không phải lúc nào cũng phù hợp với WER.



Hình 1.1: Tính toán WER cho ví dụ “Anh yêu em”. S, D, I biểu diễn số lượng thay thế, xóa và chèn để chuyển từ bản ghi chép thực tế sang đầu ra của ASR. N biểu diễn tổng số từ trong bản ghi chép thực tế.

Các nhà nghiên cứu đã nghiên cứu cách con người nhận thức và xếp hạng các lỗi ASR. Họ phát hiện ra rằng có nhiều sự đồng thuận hơn giữa những người đánh giá ở mức cực đoan (tức là mức cao nhất và lỗi ít nghiêm trọng nhất) [37]. Sự đồng thuận này ở mức cực đoan cho thấy rằng có khả năng tồn tại các mô hình lỏng lẻo hoặc một số phương pháp mà con người sử dụng để đánh giá mức độ nghiêm trọng của một lỗi trong phiên âm. Hơn nữa, khi so sánh khoảng cách chỉnh sửa hoặc số lượng thao tác tối thiểu cần thiết để chuyển đổi một chuỗi thành chuỗi khác, để mức độ nghiêm trọng của các lỗi trong quá trình phiên âm, họ thấy rằng có nhiều lỗi nghiêm trọng lỗi có khoảng cách chỉnh sửa tương đối thấp. Điều này phù hợp với các nghiên cứu khác, tương tự như vậy cho thấy WER không phải lúc nào cũng có mối tương quan tốt với khả năng hiểu được hoặc hiệu suất trong một nhiệm vụ hạ nguồn nhất định, chẳng hạn như trong việc hiểu ngôn ngữ tự nhiên (NLU) hoặc nhận dạng thực thể được đặt tên (NER) [18, 54, 48].

Có thể hiểu được mức độ nghiêm trọng hoặc mức độ nghiêm trọng của các lỗi ASR trở nên thậm chí còn quan trọng hơn trong các tình huống rủi ro cao như chăm sóc sức khỏe. ASR đã được sử dụng trong chăm sóc sức khỏe từ những năm 1970 và đã được chứng minh là có lợi, giúp giảm chi phí và thời gian cần thiết trong việc báo cáo [26]. Trong nghiên cứu chăm sóc sức khỏe, phiên âm được sử dụng rộng rãi nhiều nhiệm vụ khác nhau như trong việc phát hiện tự động chứng mất trí nhớ [14], trong việc ước tính điểm số của các bài kiểm tra sàng lọc sức khỏe nhận thức được chuẩn hóa [15], và trong dự đoán và giải thích về chẩn đoán [41]. Tất cả các tác phẩm này đều hoạt động trên cơ sở có một bản sao dễ hiểu và chính xác. Mục đích của nghiên cứu này là phát triển một phương pháp đo lường và hiểu biết một cách có hệ thống về chất lượng của các hệ thống ASR, đặc biệt là trong các bối cảnh có rủi ro cao như chăm sóc sức khỏe, vượt ra ngoài WER bằng cách xem xét tại sự khác biệt về ý nghĩa giữa kết quả thực tế và kết quả đầu ra ASR.

Không cần phải nói, trong những môi trường có rủi ro cao như chăm sóc sức khỏe, nếu không có biện pháp hữu ích chắc chắn hiểu được mức độ nghiêm trọng và tác động tiềm tàng của lỗi trong phiên âm, trở nên khó khăn để đánh giá chất lượng của động cơ ASR và do đó, tìm thấy các lĩnh vực cải tiến chính. Trong nghiên cứu này, tôi đề xuất hai phương pháp để tự động đánh giá mức độ nghiêm trọng của lỗi trong phiên mã ASR bằng cách sử dụng 1) sự khác biệt trong xếp hạng tình cảm và 2) khoảng cách cosin giữa các nhúng văn bản của đầu ra của ASR và bản ghi chép chân thực của con người về một bản âm thanh nhất định. Tình cảm xếp hạng sẽ nắm bắt được tính cực của một văn bản nhất định và sẽ hoạt động như một biện pháp đơn giản để nắm bắt ý nghĩa của văn bản đó. Việc nhúng văn bản có tiềm năng tốt hơn nắm bắt ý nghĩa của một văn bản nhất định và độ tương đồng cosin của các văn bản nhúng là một phương pháp chung để so sánh ngữ nghĩa của văn bản [47]. Bằng cách sử dụng khoảng cách cosin,

1 cos-similarity, chúng ta nhận được một giá trị, giống như WER, giá trị càng cao thì càng xa ngoài ra sự thật cơ bản là từ đầu ra ASR. Tôi kiểm tra tính khả thi và tính hữu ích của những phương pháp này trên dữ liệu y tế mô phỏng. Tôi thực hiện điều này bằng cách 1) so sánh kết quả của

các biện pháp này đối với nhãn mức độ nghiêm trọng của con người (Mục 5.1) và bằng cách 2) kết hợp các biện pháp này đo lường thành số liệu để đánh giá tổng thể động cơ ASR (Phần 5.2). kết quả của các số liệu này trên 5 kiến trúc ASR khác nhau được so sánh.

Kết quả cho thấy có sự đồng thuận đáng tin cậy giữa những người đánh giá dựa trên Fleiss Các biện pháp tương quan của Kappa, Cohen' Kappa và Kendall [17, 9, 27]. Sự khác biệt trong xếp hạng tình cảm tương quan với xếp hạng mức độ nghiêm trọng của con người, nhưng không tốt bằng WER hoặc khoảng cách cosin của nhúng văn bản. Nhúng văn bản chứng minh là tốt hơn dự đoán nhãn của con người về mức độ nghiêm trọng hơn WER. Công trình này cũng cho thấy tình cảm xếp hạng, nhúng văn bản và WER nắm bắt các khía cạnh khác nhau của lỗi trong quá trình truyền tải các bài viết và cho thấy có những ưu điểm và hạn chế của từng phương pháp. Trong thí nghiệm cuối cùng (Phần 5.3), tôi chứng minh tiềm năng của mức độ nghiêm trọng được sử dụng trong sự phát triển và cải thiện các hệ thống ASR. Tôi kết luận bằng cách thảo luận về tương lai lĩnh vực nghiên cứu.

CHƯƠNG 2

PHÁT BIỂU LUẬN VĂN

Trước khi phát triển một phương pháp để ước tính mức độ nghiêm trọng của lỗi ASR bằng cách sử dụng tình cảm phân tích và/hoặc nhúng văn bản, trước tiên tôi tìm cách trả lời câu hỏi sau: để làm gì phạm vi thông tin được thu thập bởi các máy phân tích tình cảm và/hoặc nhúng văn bản có tương quan với đánh giá của con người về mức độ nghiêm trọng của lỗi ASR trong môi trường chăm sóc sức khỏe không?

Nếu câu hỏi đầu tiên này cho thấy rằng các trình phân tích tình cảm và/hoặc nhúng văn bản thực hiện tương quan với đánh giá của con người ở một mức độ nào đó, thì chúng ta có bằng chứng rằng chúng ta có thể sử dụng phân tích tình cảm hoặc nhúng văn bản trong đánh giá có hệ thống các lỗi ASR.
Điều này dẫn đến những câu hỏi sau: chúng ta có thể sử dụng một trình phân tích tình cảm và/hoặc văn bản không? nhúng để phát triển một trình ước tính tự động hữu ích cho việc đánh giá mức độ nghiêm trọng của Lỗi ASR? Nếu vậy, điều này có thể được sử dụng trong đánh giá tổng thể của hệ thống ASR không? Làm thế nào đánh giá này có thể so sánh với WER không? Ưu điểm và hạn chế của Những phương pháp này? Và cuối cùng, liệu chúng có thể được sử dụng trong quá trình phát triển hệ thống ASR không?

Tôi cho rằng phân tích tình cảm và nhúng văn bản có thể nắm bắt đủ thông tin để phát triển một phương pháp tự động để đánh giá mức độ nghiêm trọng của lỗi ASR trong một thiết lập chăm sóc sức khỏe. Tôi cũng dự đoán rằng chúng ta cũng có thể sử dụng các phương pháp này kết hợp với WER để đánh giá tốt hơn hiệu suất tổng thể của động cơ ASR.

CHƯƠNG 3

LÝ LỊCH

Trong phần này, tôi sẽ cung cấp một cái nhìn tổng quan ngắn gọn về phân tích tình cảm và nhúng văn bản, bao gồm các mô hình và phân tích tình cảm cụ thể cho nhúng văn bản được sử dụng trong công việc này. Sau đó, tôi giới thiệu các động cơ ASR mà tôi sử dụng cho các thí nghiệm.

3.1 Phân tích tình cảm và nhúng

Khi nói đến việc hiểu và tự động đánh giá mức độ nghiêm trọng của lỗi trong ASR, người ta cần phải có một phương pháp để phân tích một cách có hệ thống sự khác biệt trong nghĩa giữa hai cụm từ hoặc câu. Trong khi về mặt triết học, một cơ thể văn bản thực sự có nghĩa là một câu hỏi khó trả lời, một cách đơn giản để nắm bắt một số bản chất của ý nghĩa của một lời nói là thực hiện phân tích tình cảm về đầu ra của công cụ ASR và bản ghi chép thực tế.

Trong lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP), phân tích tình cảm là nhiệm vụ của việc phát hiện thái độ, cảm xúc hoặc cực tính của một văn bản nhất định. Nó là phổ biến cho các thuật toán này lấy một chuỗi làm đầu vào và đưa ra dự đoán từ -1 đến 1 dựa trên mức độ tiêu cực hay tích cực của văn bản. Bởi vì các thuật toán này có thể thay đổi và có những hạn chế riêng, tôi sử dụng 3 bộ phân tích tình cảm khác nhau từ 3 các thư viện NLP được sử dụng rộng rãi NLTK, FLAIR và TextBlob (TB) [24, 1, 36]. Đây là một phương pháp ngây thơ để nắm bắt ý nghĩa của một văn bản nhất định bởi vì rõ ràng có hai văn bản

có thể có nhiều ý nghĩa khác nhau nhưng cả hai đều có cùng một tình cảm. Mặc dù có lẽ
quá đơn giản, mục đích của việc sử dụng tình cảm là để tạo ra một đường cơ sở
biện pháp nắm bắt một cái gì đó ngoài sự khác biệt trong chính tả và để kiểm tra như thế nào
xếp hạng tình cảm cũng có thể thực hiện tốt.

Một phương pháp phổ biến khác để nắm bắt ý nghĩa của ngôn ngữ tự nhiên là sử dụng
nhúng văn bản. Tạo nhúng từ là quá trình chuyển đổi từng từ
các từ thành các vectơ n chiều, thường là nhằm mục đích chuyển đổi văn bản thành
thứ gì đó có thể được xử lý bằng thuật toán Học máy hoặc Học sâu.
Có nhiều phương pháp nhúng từ khác nhau, từ phương pháp dựa trên quy tắc đơn giản
phương pháp đến các phương pháp phức tạp hơn liên quan đến các kỹ thuật học máy [38,
43, 44]. Tương tự như vậy, các phương pháp đã được phát triển để nhúng nhiều hơn chỉ một
từ [31, 47]. Cho dù nhúng từng từ riêng lẻ hay toàn bộ câu, nói chung,
với các nhúng tốt, các từ hoặc cụm từ càng giống nhau về mặt ngữ nghĩa thì càng gần nhau
chúng phải nằm trong không gian vectơ n chiều [38, 39].

Đối với dự án này, tôi sử dụng 4 mô hình được đào tạo sẵn có do Sentence- cung cấp.
Transformers1 để tính toán nhúng câu. Dưới đây là mô tả ngắn gọn về từng
một.

bert-base-nli-mean-tokens (BertNLI) là một sửa đổi của BERT được đào tạo trước
mô hình “sử dụng cấu trúc mạng ba và mạng Xiêm để suy ra ngữ nghĩa
nhúng câu có ý nghĩa” [47]. Sử dụng BERT thô cho ngữ nghĩa quy mô lớn
tìm kiếm hoặc so sánh tốn kém về mặt tính toán, tuy nhiên với phương pháp này
mạng lưới lớn có thể được tinh chỉnh để có sự tương đồng về mặt ngữ nghĩa bằng cách sử dụng suy luận ngôn ngữ tự nhiên

1https://www.sbert.net/docs/pretrain_models.html

dữ liệu. Tất cả các mô hình đều dựa trên cùng một quy trình này để lấy một dữ liệu lớn được đào tạo trước
mô hình ngôn ngữ và tinh chỉnh nó trên dữ liệu để có sự tương đồng về mặt ngữ nghĩa hiệu quả.

all-MiniLM-L6-v2 (MiniLM) dựa trên mô hình MiniLM của Microsoft [53]. Vì
mô hình cơ sở, các nhà nghiên cứu đã phát triển một phương pháp chưng cất kiến thức được gọi là sâu
chưng cất tự chú ý, trong đó mục đích là chưng cất hoặc thu nhỏ một lượng lớn
mô hình, thường chứa hàng trăm triệu tham số, thành một mô hình nhỏ hơn
thường duy trì hiệu suất của mô hình lớn hơn và có thể được sử dụng rộng rãi hơn
đã sử dụng. Mô hình cơ sở này sau đó được tinh chỉnh để có sự tương đồng về mặt ngữ nghĩa.

all-mpnet-base-v2 (MPNET) dựa trên mô hình MPNet của Microsoft [51], trong đó
bao gồm sự kết hợp của mô hình ngôn ngữ được che giấu (một phương pháp đào tạo trước được sử dụng
trong các mô hình như BERT [10]) và mô hình ngôn ngữ hoán vị trong quá trình đào tạo trước (một
phương pháp đào tạo được sử dụng trong XLR [55]), tìm cách tận dụng lợi thế của cả hai phương pháp. Điều này
mô hình cơ sở sau đó được tinh chỉnh để có sự tương đồng về mặt ngữ nghĩa.

all-distilroberta-v1 (DisRob) là mẫu cuối cùng tôi sử dụng, dựa trên DisilRoBERTa
là quá trình chưng cất tuân theo cùng một quy trình như DistilBERT [49], ngoại trừ
RoBERTa [34] làm cơ sở thay thế cho BERT. Mục đích chưng cất BERTA của
RoBERTA sẽ thu nhỏ kích thước mô hình để tăng tốc độ và giảm bộ nhớ
yêu cầu trong khi vẫn duy trì hiệu suất. Giống như tất cả các mô hình trước đó, cơ sở này
mô hình sau đó được tinh chỉnh để có sự tương đồng về mặt ngữ nghĩa.

3.2 Động cơ ASR

Đối với dự án này, tôi sử dụng năm động cơ ASR để thử nghiệm nhằm thu thập và thu được kết quả từ nhiều kiến trúc khác nhau. Tôi chọn các kiến trúc sau vì tính khả dụng, hiệu suất và vì chúng có thể chạy cục bộ (mà có nghĩa là người ta sẽ không phải giải quyết các vấn đề tiềm ẩn khi gửi dữ liệu nhạy cảm qua internet đến hệ thống ASR đám mây). Trong phần này, tôi sẽ mô tả ngắn gọn của năm kiến trúc này.

DeepSpeech2 (DS2) của Mozilla là một triển khai của [3]. Trong kiến trúc này, Mạng nơ-ron hồi quy lấy các spetrogram từ một tệp âm thanh và được đào tạo để đầu ra văn bản[2] .

Wav2Vec2 (W2V2) của Meta là một mô hình được đề xuất bởi [4]. Không giống như DeepSpeech, mô hình này kiến trúc hoạt động trực tiếp trên dữ liệu âm thanh thô thay vì spretrogram. mô hình được đào tạo đầu tiên theo phương pháp bán giám sát trên nhiều giờ lời nói không có nhãn dữ liệu và sau đó được tinh chỉnh trên dữ liệu được gắn nhãn. Mô hình này được thực hiện dễ dàng truy cập bằng Ôm mặt [3].

PocketSphinx (PS) của CMU là một trong những ASR nhẹ hơn mà tôi sử dụng [23]. PS là một trọng lượng ASR là một phần của bộ công cụ nhận dạng giọng nói nguồn mở được gọi là Dự án CMUSphinx. Mô hình này được đào tạo trên 1.600 câu nói từ RM-1 ngữ liệu đào tạo độc lập với người nói. Không giống như các mô hình đã đề cập trước đó, PS không sử dụng mạng nơ-ron và thay vào đó dựa trên các phương pháp truyền thống

2https://deepspeech.readthedocs.io/en/latest/index.html
3https://huggingface.co/docs/transformers/model_doc/wav2vec2

nhận dạng giọng nói bằng cách sử dụng Mô hình Markov ẩn, mô hình ngôn ngữ và ngữ âm
từ điển.4

Vosk của Alpha Cephei (với mô hình vosk-model-en-us-0.22) được xây dựng bằng Kaldi
[45], và giống như PS, sử dụng mô hình âm thanh, mô hình ngôn ngữ và từ điển ngữ âm.
Tuy nhiên, không giống như PS, Vosk sử dụng mạng nơ-ron cho phần mô hình âm thanh của
động cơ.5

Không giống như Wave2Vec2, Whisper của OpenAI sử dụng phương pháp đào tạo được giám sát hoàn toàn
đang thu thập 680.000 giờ nội dung được phiên âm từ internet ở 99 ngôn ngữ khác nhau
ngôn ngữ [46]. Theo các kiến trúc khác như DeepSpeech2, mô hình này sử dụng
phổ âm thanh làm đầu vào, nhưng thay vì Mạng nơ-ron hồi quy, mô hình này
sử dụng kiến trúc Biến đổi mã hóa-giải mã dựa trên [52] với nhiều loại
các mã thông báo đặc biệt được sử dụng để chỉ ra nhiệm vụ nào đang được thực hiện (ví dụ: phiên âm hoặc
Đối với các thí nghiệm của mình, tôi sử dụng mô hình cơ sở6 (bao gồm 74 triệu
tham số).

4https://github.com/cmusphinx/pocketsphinx-python
5https://alphacephei.com/vosk/
6https://huggingface.co/openai/whisper-base

CHƯƠNG 4

DỮ LIỆU

4.1 Thu thập dữ liệu

Với mục đích thử nghiệm trong một kịch bản chăm sóc sức khỏe, một tập dữ liệu được công bố vào năm 2022 của các cuộc phỏng vấn y tế giữa bệnh nhân và bác sĩ được mô phỏng được sử dụng [13]. Bộ dữ liệu này chứa 272 tệp âm thanh có bản ghi chép. Các tệp này có độ dài từ khoảng 7 đến 20 phút hoặc từ 800 đến 2200 từ.

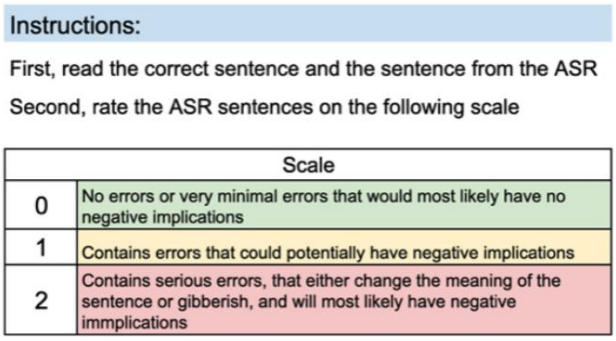
Các tập tin được chia thành các khoảng thời gian không im lặng bằng cách sử dụng librosa1 thiết lập ngưỡng của im lặng đến 60 decibel. Với ngưỡng 60 decibel, các tập tin được chia thành hơn 39.600 khoảng không im lặng, mà tôi sẽ gọi là lời phát biểu vì mỗi tệp chứa một lời phát biểu nhỏ của bài phát biểu. Trong số này, tôi lấy một mẫu gồm 110 phát ngôn và chạy chúng qua Các công cụ ASR cũng như lấy thủ công bản ghi chép các lời nói từ các tệp phiên âm tương ứng đi kèm với tập dữ liệu. Bởi vì chúng đã được chạy thông qua năm công cụ ASR, kết quả là danh sách 550 cặp bản sao trong đó một đến từ một động cơ ASR và cái còn lại là bản ghi chép sự thật. Một trong những những lời nói thực sự không có lời nói nào và đã bị xóa bỏ với tổng số cuối cùng là 545 cặp.

150 cặp bảng điểm này đã được trao cho 3 sinh viên trường y được yêu cầu đánh giá từng cặp bằng 0, 1 hoặc 2 (2 là lỗi nghiêm trọng, 1 là không

1https://librosa.org/doc/main/generated/librosa.effects.split.html

lỗi rất nghiêm trọng và 0 là lỗi rất nhỏ hoặc bản sao hoàn hảo). Chính xác
hướng dẫn được đưa ra và một số ví dụ về dữ liệu được cung cấp trong Hình 4.1.

Các cặp được chuẩn hóa bằng cách loại bỏ các ghi chú nhận dạng người nói “P:” và “D:”
đối với bệnh nhân và bác sĩ, làm cho tất cả các chữ cái thường và bằng cách loại bỏ bất kỳ ký tự đặc biệt nào
các ký tự và dấu câu ngoại trừ dấu nháy đơn (vì chúng có thể quan trọng trong
phân biệt các từ như “its” và “it's” hoặc “they're” và “their”).



Instructions:

First, read the correct sentence and the sentence from the ASR

Second, rate the ASR sentences on the following scale

Scale	
0	No errors or very minimal errors that would most likely have no negative implications
1	Contains errors that could potentially have negative implications
2	Contains serious errors, that either change the meaning of the sentence or gibberish, and will most likely have negative immplications

Correct	Output of ASR system	Rating
okay	propane	2
sorry yeah the pain has been there this whole time and it's gotten worse ever since it started	si yet the pad has been there this wholl time and it's gotten worse iave ever since it started ocet	2
it made it a bit worse but	it made it a big worse by ta	2
no	no	0
a multivitamin	a a multy biteman	2
when did this pain start	when do this pain start	1

Hình 4.1: Hình ảnh hiển thị hướng dẫn dành cho người đánh giá và một vài cặp câu ví dụ với phiên âm đúng ở bên trái, kết quả đầu ra của công cụ ASR ở giữa và đánh giá mức độ nghiêm trọng của con người ở bên phải.

4.2 Chứng chỉ của người đánh giá

Cả ba người đánh giá hiện đang theo học chương trình tiến sĩ tại Cao đẳng Idaho
Y học nắn xương (ICOM). Kinh nghiệm của các thành viên bao gồm nghiên cứu y khoa

tại các địa điểm như Phòng khám Mayo và Đại học Utah, làm việc như người Tây Ban Nha-
Người ngắt lời tiếng Anh tại các phòng khám y tế, làm việc như một kỹ thuật viên gây mê và giữ
các vị trí như đại diện sinh viên trong ủy ban nghiên cứu của ICOM.

4.3 Xác thực dữ liệu: Người đánh giá có đồng ý không?

Các công trình nghiên cứu trước đây cho thấy rằng mức độ nghiêm trọng của lỗi trong quá trình phiên mã là một nhiệm vụ khó khăn
nơi không có sự đồng thuận tốt giữa những người đánh giá [37]. Trước khi phát triển một
biện pháp đánh giá lỗi theo cùng cách mà con người sẽ làm, trước tiên nó cần phải được hiển thị
rằng con người có một số phương pháp luận hoặc sự nhất quán với nhau khi nó
đến việc đánh giá mức độ nghiêm trọng của lỗi.

Theo các số liệu đánh giá được sử dụng trong [37], tôi sử dụng Kappa của Cohen [9] và Fleiss
Kappa [17], để đo lường sự đồng thuận giữa người chú thích bên trong. Tuy nhiên, các số liệu này không
lưu ý rằng dữ liệu là thứ tự (tức là sự khác biệt trong xếp hạng các giá trị 0
và 1 được xử lý giống hệt như sự khác biệt giữa giá trị 0 và 2 mặc dù
sự khác biệt sau lớn hơn sự khác biệt trước [12]). Do đó, vì bản chất của
những xếp hạng này là thứ tự, tôi cũng xem xét hệ số tương quan thứ hạng của Kendall [27]
để đo lường chất lượng của mối liên hệ thứ tự giữa hai người đánh giá nhất định.

Tôi tính toán giá trị Kappa của Fleiss là 0,452 và điểm Kappa của Cohen trong phạm vi
từ 0,420 đến 0,567, cho thấy sự đồng thuận vừa phải giữa những người đánh giá (xem Bảng 4.1).
Hệ số tương quan Kendall giữa những người đánh giá cho thấy có mối tương quan mạnh mẽ
người đánh giá tween dao động từ 0,662 đến 0,727 (xem Bảng 4.2). Xem xét tính chủ quan
của nhiệm vụ, các giá trị Kappa vừa phải và các giá trị tương quan cao cho thấy rằng
có sự nhất quán đáng tin cậy giữa những người đánh giá.

Bảng 4.1: Thỏa thuận giữa người chú thích bên trong ma trận nhầm lẫn.

	Người đánh giá 1	Người đánh giá 2	Người đánh giá 3
Người đánh giá 1	-	0,416	0,567
Người đánh giá 2	0,416	-	0,440
Người đánh giá 3	0,567	0,440	-

Bảng 4.2: Hệ số tương quan của Kendall khoa học

	Người đánh giá 1	Người đánh giá 2	Người đánh giá 3
Người đánh giá 1	-	0,727	0,718
Người đánh giá 2	0,727	-	0,662
Người đánh giá 3	0,718	0,662	-

CHƯƠNG 5

PHƯƠNG PHÁP

Chương này mô tả các thí nghiệm và phương pháp được sử dụng để kiểm tra các giả thuyết được mô tả trong Chương 2.

5.1 Thí nghiệm 1: Kiểm tra Điểm mức độ nghiêm trọng

Trong thí nghiệm này, mục tiêu là kiểm tra xem các trình phân tích tình cảm và/hoặc nhúng văn bản có thể đánh giá lỗi tương tự như cách con người đánh giá trong môi trường chăm sóc sức khỏe. Tôi sử dụng 150 phát ngôn có nhãn mức độ nghiêm trọng của con người được mô tả trong Phần 4.1 và kể từ đó có sự đồng thuận khá tốt, tôi sử dụng chế độ xếp hạng làm mục tiêu.

Tôi tính toán WER và nhiều điểm nghiêm trọng khác nhau (được định nghĩa bên dưới trong Phần 5.1.1) sử dụng các trình phân tích tình cảm và các mô hình ngôn ngữ để nhúng văn bản được mô tả trong Phần 3.1. Tôi so sánh các điểm nghiêm trọng với nhau bằng cách đo lường sự tương quan mối quan hệ giữa điểm số nghiêm trọng và chế độ đánh giá của con người. Một mối tương quan cao giữa những điểm số nghiêm trọng này và đánh giá của con người hỗ trợ cho giả thuyết rằng phân tích tình cảm và/hoặc nhúng văn bản thu thập đủ thông tin để sử dụng đánh giá mức độ nghiêm trọng của lỗi ASR trong môi trường chăm sóc sức khỏe.

Để đánh giá thêm tính hữu ích của các điểm số nghiêm trọng này, tôi tạo nhiều Thứ tự Mô hình hồi quy logistic với một điểm số nghiêm trọng duy nhất là một biến độc lập trong mỗi mô hình và chế độ đánh giá của con người như là biến mục tiêu và so sánh

hiệu suất của các mô hình. Để so sánh, tôi thực hiện xác thực chéo 10 lần cho mỗi
mô hình và xem xét độ chính xác trung bình trên dữ liệu thử nghiệm và so sánh chúng với từng
khác và với phân loại đa số.

5.1.1 Từ Phân tích tình cảm và nhúng đến Điểm nghiêm trọng

Trong phần này tôi giải thích cách tôi tính toán các điểm nghiêm trọng khác nhau bằng cách sử dụng tình cảm
trình phân tích và nhúng văn bản.

Khi đánh giá mức độ nghiêm trọng của lỗi với tư cách là con người, chúng ta có thể cố gắng nhìn nhận một cách khách quan
tại sự khác biệt về ý nghĩa giữa sự thật cơ bản và kết quả đầu ra của ASR
động cơ. Người ta có thể tưởng tượng quá trình này có khả năng liên quan đến việc sử dụng một số mô hình trong
bộ não nơi tình yêu gần với tình yêu hơn là sự căm ghét về mặt ý nghĩa, và do đó tình yêu sẽ được đánh giá
như một lỗi ít nghiêm trọng hơn. Tôi tìm cách mô phỏng quá trình này bằng cách sử dụng các trình phân tích tình cảm và văn bản
nhúng.

Cho một bộ phân tích tình cảm s(x) đưa ra giá trị giữa -1 và 1 (chẳng hạn như
SentimentIntensityAnalyzer của NLTK [6]), chúng ta có thể thử mô phỏng quá trình này bằng cách thực hiện
giá trị tuyệt đối của sự khác biệt trong tình cảm và sử dụng điều này như một mô hình để biểu diễn
sự khác biệt về ý nghĩa hoặc mức độ nghiêm trọng. Điều này có thể được thể hiện bằng những điều sau:

$$\text{Mức độ nghiêm trọng}(x, y) = |s(x) \quad s(y)|$$

trong đó x và y là một cặp đầu ra thực tế và ASR. Điều này sẽ dẫn đến
một đánh giá trong phạm vi [0, 2], trong đó 0 sẽ là hai cụm từ có cùng một giá trị chính xác
đánh giá tình cảm, nói cách khác là lỗi không nghiêm trọng, nếu có, và đánh giá 2 sẽ
thể hiện lỗi nghiêm trọng nhất có thể xảy ra do có cảm xúc và thái cực đối lập.

Thực hiện theo cùng một logic với nhúng văn bản, biết rằng nhúng chặt chẽ hơn

nên gần gũi hơn về mặt ngữ nghĩa, chúng ta có thể biểu diễn sự khác biệt về ý nghĩa như
sự khác biệt giữa chúng, nói cách khác, là một trừ đi cosin của phép nhúng của chúng.

$$\text{Mức độ nghiêm trọng}(x, y) = 1 \quad \cos(x, y)$$

Điều này sẽ dẫn đến xếp hạng trong phạm vi [-1, 1]. Tuy nhiên, thông lệ chung là
ràng buộc các vectơ trong không gian dương sẽ dẫn đến phạm vi [0, 1]. Bởi vì
chúng ta đang xem xét sự khác biệt và giá trị 0 sẽ biểu diễn hai chuỗi
giống nhau và một giá trị gần bằng 1 sẽ biểu diễn hai chuỗi rất khác nhau
về mặt ngữ nghĩa.

5.1.2 Kết quả

Các mối tương quan được thể hiện trong Bảng 5.1 cho thấy WER có mối tương quan với xếp hạng của con người
mức độ nghiêm trọng là 0,43. Tất cả các điểm số nghiêm trọng dựa trên nhúng văn bản có tương quan với
đánh giá của con người tốt hơn WER, với mức tăng tương quan từ 28%
lên đến 36% trên WER. Ngược lại, tất cả các điểm số nghiêm trọng dựa trên tình cảm đều
ít tương quan hơn WER. Bộ phân tích tình cảm từ FLAIR tương quan nhiều nhất
với xếp hạng của con người có giá trị là 0,34 (xem Bảng 5.1). Kết quả cho thấy văn bản
nhúng có thể phù hợp hơn cho việc đánh giá tự động mức độ nghiêm trọng trong ASR
lỗi nhiều hơn WER.

Bảng 5.1: Sự tương quan giữa đánh giá mức độ nghiêm trọng của con người đối với WER và các mức độ nghiêm trọng khác nhau
biện pháp. Ba chữ in nghiêng là điểm số nghiêm trọng dựa trên phân tích tình cảm,
bốn chữ in đậm là điểm số nghiêm trọng dựa trên nhúng câu. Nhúng câu
điểm số có mối tương quan cao nhất với đánh giá mức độ nghiêm trọng của con người.

WER	NLTK	FLAIR	TB	MiniLM	BertNLI	MPNET	DisRob			
0,43		0,29		0,34	0,55	0,29		0,53	0,56	0,59

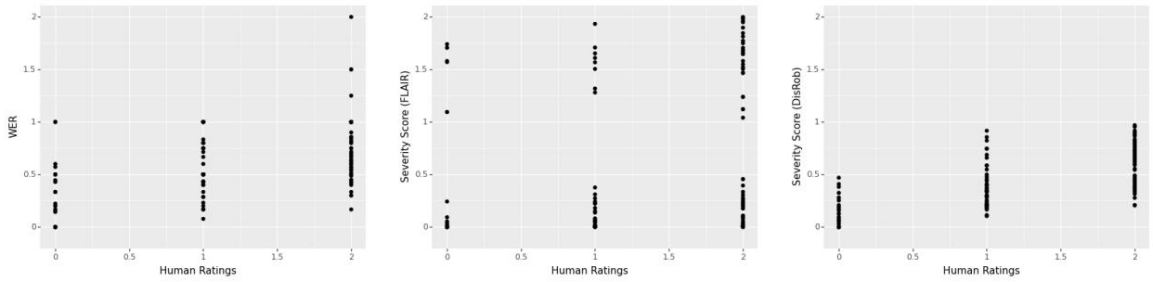
Biểu đồ so sánh WER, FLAIR và all-distilroberta-v1 (DisRob) với chuột thí nghiệm của con người các thông tin trong Hình 5.1. Biểu đồ phụ 5.1c, so sánh xếp hạng của con người với DisRob, cho thấy rằng nhúng thực hiện tốt nhất công việc nhóm các lỗi ASR với cùng một đánh giá của con người cùng nhau. Nói cách khác, nói chung, những lỗi nghiêm trọng nhất (với đánh giá của con người là 2) được đẩy về phía 1 (mức tối đa cho biện pháp này), các lỗi trung bình-nghiêm trọng (với đánh giá của con người là 1) được nhóm lại quanh mức 0,5 và các lỗi ít nghiêm trọng nhất (với con người xếp hạng 0) nằm giữa 0 và 0,5.

Ngược lại, các WER được hiển thị trong sơ đồ phụ 5.1a được phân tán rộng hơn, có một số lỗi nghiêm trọng với WER tương đối thấp và một số lỗi không nghiêm trọng với WER tương đối thấp WER cao.

Biểu đồ ở giữa, biểu đồ phụ 5.1b, cho thấy các lỗi nghiêm trọng có xu hướng bị đẩy đến điểm số cao hơn theo máy phân tích tình cảm FLAIR, nhưng giống như WER, FLAIR đánh giá một số lỗi nghiêm trọng với điểm số thấp. Nhìn sâu hơn, điểm số nghiêm trọng dựa trên tình cảm có xu hướng làm tốt trong việc phát hiện ra những điều đối lập hoặc phủ định như "tình yêu" và “ghét” hoặc “không nằm xuống” so với “nằm xuống” phù hợp với mục đích của họ (tức là để phát hiện cực tính). Do đó, tôi tin rằng vẫn còn một số tiềm năng trong việc sử dụng tình cảm để đánh giá mức độ nghiêm trọng, đặc biệt là để phát hiện các mặt đối lập hoặc bỏ sót trong phủ định như trong các ví dụ về “yêu thương” và “không nằm xuống”, nhưng dựa trên những kết quả này, nó sẽ nhất có thể không linh hoạt bằng nhúng văn bản.

Bảng 5.2: Ví dụ về máy phân tích tình cảm, FLAIR, đưa ra xếp hạng cao (gần 2) cho các mặt đối lập và phủ định. Nhiều ví dụ hơn được hiển thị trong 5.5

Sự thật thực tế Đầu ra ASR ừm	SỰ TINH TẾ	ASR
không nằm xuống giúp nằm xuống giúp tôi yêu bạn tôi	1.943	DeepSpeech2
ghét bạn	1.989	Ví dụ về đồ chơi



(a) NGƯỜI (b) Tình cảm (Flair) (c) Nhúng (DisRob)

Hình 5.1: Biểu đồ so sánh đánh giá mức độ nghiêm trọng của con người (trục x) với WER và hai mức độ nghiêm trọng đánh giá một dựa trên điểm số tình cảm và một dựa trên câu nhúng (trục y). Lưu ý rằng một số lượng lớn lỗi có mức độ nghiêm trọng cao của con người xếp hạng có WER tương đối thấp.

Sử dụng Điểm số Mức độ Nghiêm trọng như một Công cụ Dự đoán

Đối với kết quả của các mô hình hồi quy Logist thứ tự, tôi tạo các mô hình bằng cách sử dụng tất cả các các biện pháp nghiêm trọng đã đề cập trước đó, cũng như phần lớn được phân loại là mức cơ sở. Mô hình có hiệu suất kém nhất là mô hình sử dụng điểm số dựa trên FLAIR với độ chính xác trung bình là 50% tương đương với đường cơ sở. Mô hình dựa trên WER là tốt nhất tiếp theo với độ chính xác trung bình là 62%. Tất cả các mô hình dựa trên văn bản nhúng điểm thực hiện tốt hơn mô hình WER với mô hình tốt nhất có độ chính xác trung bình là 70,67%. Kết quả được tóm tắt trong Bảng 5.3.

Bảng 5.3: Độ chính xác trung bình của Mô hình hồi quy logistic thứ tự với xác thực chéo 10 lần. Tất cả các mô hình dựa trên nhúng văn bản đều có độ chính xác trung bình cao hơn WER.

Thiếu tá lớp.	WER	FLAIR	MiniLM	BertNLI	MPNET	DisRob		
50,00	62,00	50,00	66,00	70,67	63,33	66,67		

Kết quả từ nhiệm vụ Hồi quy Logistic Thứ tự đưa ra thêm bằng chứng rằng nhúng văn bản phù hợp hơn với việc đánh giá mức độ nghiêm trọng tự động trong ASR lỗi nhiều hơn WER.

Trong thí nghiệm này, tôi chỉ nghiên cứu mối tương quan giữa mức độ nghiêm trọng được đề xuất điểm số và xếp hạng tương ứng của con người. Tuy nhiên, WER thường là trung bình trên tất cả các phát ngôn trong một tập dữ liệu thử nghiệm. WER trung bình trở thành giá trị duy nhất được sử dụng làm thước đo để đánh giá hiệu suất tổng thể của ASR động cơ. Vì điểm số nghiêm trọng có mối tương quan tốt với nhãn của con người, nên điều đó đáng giá thử nghiệm để xem liệu những điểm số này có thể được sử dụng để tính toán một số liệu hay không, theo cách tương tự cách tính WER trung bình, có thể được sử dụng để đánh giá hiệu suất của ASR động cơ.

5.2 Thí nghiệm 2: Sử dụng số liệu để đánh giá ASR

Thí nghiệm này chứng minh tính hữu ích tiềm tàng của việc sử dụng điểm số nghiêm trọng (từ Thí nghiệm 1) trong các số liệu để đánh giá và so sánh tổng thể ASR hệ thống. Để kiểm tra điều này, tôi đề xuất ba số liệu (được liệt kê bên dưới, Mục 5.2.1) và sử dụng mỗi số liệu để đánh giá và so sánh hiệu suất của năm động cơ ASR từ Mục 3.2.

Để thực hiện đánh giá này, tôi lấy 110 câu nói đã được ghi chép thủ công được phiên âm và chạy chúng qua tất cả năm động cơ ASR. Như đã đề cập trong Phần 4, một trong những tập tin âm thanh không có lời thoại nào trong đó và đã bị xóa. Kết quả này trong tổng số 109 5 = 545 bản sao chép. Như trước đây, tôi tính toán mức độ nghiêm trọng khác nhau điểm cho mỗi đầu ra ASR, cặp thực tế. Những điểm nghiêm trọng này sau đó được sử dụng làm đầu vào cho các số liệu được mô tả trong phần sau.

Các số liệu này được so sánh với WER và với nhau trên năm ASR công cụ để đánh giá cách thức các số liệu này hoạt động.

5.2.1 Đo lường 1: MAE của sự khác biệt trong tình cảm

Chỉ số đầu tiên tôi đề xuất là sai số tuyệt đối trung bình của sự khác biệt trong tình cảm (MAE-DS). Về mặt hình thức, đưa ra một bộ phân tích tình cảm s(x) đưa ra một giá trị giữa -1 và 1 (chẳng hạn như SentimentIntensityAnalyzer của NLTK [6, 24]), chúng ta có thể biểu thị MAE-DS trong công thức dưới đây:

$$\frac{1}{n} \sum_{x,y \in C} |s(x) - s(y)|$$

trong đó C là một tập hợp các cặp phiên âm thực tế và dự đoán ASR và n là cặp số. x và y là tập hợp các phát biểu thực tế và dự đoán từ C.

Đầu ra của số liệu này sẽ nằm trong khoảng từ 0 đến 2 và sẽ dễ dàng diễn giải. Đối với ví dụ, MAE-DS 0,5, chỉ ra rằng, xét về mặt tình cảm, đầu ra của ASR là, trung bình, sai lệch 0,5 so với thực tế.

5.2.2 Đo lường 2: MSE của sự khác biệt trong tình cảm

Chỉ số thứ hai tôi đề xuất là sai số bình phương trung bình của sự khác biệt trong tình cảm (MSE-DS). Tương tự như số liệu đầu tiên, với một máy phân tích tình cảm s(x), chúng ta có thể viết điều này trong công thức dưới đây:

$$\frac{1}{n} \sum_{x,y \in C} (s(x) - s(y))^2$$

trong đó, C là một tập hợp các cặp phiên âm thực tế và dự đoán ASR và n là cặp số. x và y là tập hợp các phát biểu thực tế và dự đoán từ C.

Phạm vi của số liệu này là từ 0 đến 4. Tính hữu ích tiềm tàng của số liệu này nằm ở thực tế là MSE nhạy cảm hơn với các giá trị ngoại lệ so với MAE. Do đó, điều này sẽ phạt nặng hơn các lỗi ASR có khoảng cách lớn hơn về mặt tình cảm so với sự thật cơ bản.

5.2.3 Đo lường 3: Độ tương đồng của câu sử dụng Mô hình ngôn ngữ

Chỉ số thứ ba mà tôi đề xuất dựa trên các văn bản nhúng được phân loại theo ngôn ngữ mô hình. Trong trường hợp này, tôi tính toán độ tương đồng cosin giữa các nhúng của sự thật cơ bản và đầu ra ASR như một sự thể hiện về mức độ khác nhau của các câu trong ý nghĩa. Trong mô hình mà tôi sẽ sử dụng, độ tương tự cosin của nhúng sẽ thường nằm trong khoảng từ 0 đến 1 trong đó một là câu chính xác giống nhau và các giá trị gần hơn 0 có khoảng cách ngữ nghĩa xa so với câu mục tiêu [47].

Với điều này, tôi đề xuất sử dụng giá trị trung bình của khoảng cách cosin, hoặc một trừ đi cosin sự tương đồng. Điều này có thể được viết bằng công thức sau:

$$\frac{1}{n} \sum_{x,y \in C} 1 \quad \text{cosin}(x, y)$$

trong đó C là một tập hợp các cặp phiên âm thực tế và dự đoán ASR và n là cặp số. x và y là một tập hợp các nhúng của một giá trị thực tế nhất định và dự đoán những phát biểu từ C.

Tính hữu ích tiềm tàng của số liệu này là nhúng văn bản có thể nắm bắt được nhiều hơn thông tin hơn là xếp hạng tình cảm và các mô hình ngôn ngữ tạo ra nhúng cũng có thể được tinh chỉnh cho một tập dữ liệu nhất định (tức là trong lĩnh vực y tế, một có thể làm cho tên thuốc và chẩn đoán cách xa nhau hơn).

Tuy nhiên, không có sự điều chỉnh tinh tế nào được thực hiện trong công việc này. Một lợi thế tiềm năng khác của

phương pháp này rất giống với hàm mất mát Cosine và có khả năng
có thể dễ dàng kết hợp vào hàm mất mát như một hình phạt cho quá trình đào tạo mạng nơ-ron.

5.2.4 Kết quả

Kết quả được tóm tắt trong Bảng 5.4. Đối với tất cả các số liệu, giá trị càng thấp thì tốt hơn. Nhìn chung, kết quả đều nhất quán, bất kể chúng ta sử dụng số liệu nào, phần lớn cho thấy Whisper có hiệu suất tốt nhất tiếp theo là Vosk. Theo sau những điều này trong hiệu suất là DeepSpeech2 (DS2) và Wav2Vec2 (W2V2). ASR với
hiệu suất thấp nhất là PocketShpinx (PS). Mặc dù các số liệu thống kê đồng ý với hầu hết một phần, đi sâu hơn và nghiên cứu mức độ các số liệu này đồng ý và
khi các số liệu không thống nhất, người ta có thể có được những hiểu biết hữu ích về thông tin nào số liệu đề xuất đang được thu thập.

Bảng 5.4: Kết quả của Thí nghiệm 2. Hàng trên cùng hiển thị từng động cơ ASR trong số 5 động cơ. Phần sau đây hiển thị WER. Các nhãn trong cột đầu tiên kết thúc bằng mae và mse lần lượt là sai số tuyệt đối trung bình và sai số bình phương trung bình của sự khác biệt trong điểm số tình cảm. Cuối cùng cho các hàng là khoảng cách cosin trung bình.

Cơ sở Đo lường Số liệu Chỉnh		DS2 PS	Vosk Whis.	W2V2		
sửa Khoảng cách WER	0,482	0,910	0,307	0,273	0,525	
Tình cảm	NLTK mae	0,127	0,241	0,062	0,056	0,127
	FLAIR mae	0,620	0,700	0,324	0,322	0,516
	TB mae	0,111	0,181	0,050	0,029	0,120
Tình cảm	NLTK mse	0,057	0,141	0,022	0,020	0,048
	FLAIR mse	0,981	1,132	0,459	0,473	0,788
	TB mse	0,051	0,086	0,026	0,010	0,044
Cô sin Khoảng cách	MiniLM	0,361	0,649	0,171	0,153	0,403
	BertNLI	0,188	0,398	0,079	0,093	0,181
	MPNET	0,400	0,688	0,193	0,180	0,400
	Loại bỏ	0,388	0,676	0,189	0,172	0,406

Từ Vosk đến Whisper, tỷ lệ WER giảm khoảng 11,07%.

Phần trăm giảm trung bình của khoảng cách cosin trên 4 mô hình ngôn ngữ

khá nhỏ ở mức 2,13%. Trong một ví dụ khác, phần trăm giảm trong WER từ
DeepSpeech2 đến Vosk là 36,31% trong khi phần trăm giảm trung bình khoảng cách cosin
trên 4 mô hình ngôn ngữ lớn hơn ở mức 53,41%. Sự khác biệt trong các phần trăm này
chứng minh rằng tốc độ cải thiện mức độ nghiêm trọng (khoảng cách cosin) không nhất thiết là
liên quan đến tốc độ cải thiện trong WER (tức là một số liệu có thể cải thiện đáng kể trong khi
cái kia thì không nhiều lắm và ngược lại).

Để chứng minh thêm sự khác biệt của các số liệu này, tôi xem xét các ví dụ cụ thể
nơi WER và các biện pháp về mức độ nghiêm trọng không đồng nhất. Tôi thực hiện điều này bằng cách phân tích mức độ nghiêm trọng nhất
lỗi đưa ra một biện pháp nhất định trong khi biện pháp khác được giữ ở mức tương đối thấp. Đầu tiên
xem xét mức độ nghiêm trọng nhất theo điểm số tình cảm của FLAIR trong khi vẫn giữ WER
dưới 0,5 (các ví dụ từ đây được hiển thị ở 6 hàng đầu tiên trong Bảng 5.5). Sau đó, tôi
nhìn vào khoảng cách cosin nghiêm trọng nhất trong khi vẫn giữ WER ở dưới
0,5 (hiển thị ở nhóm giữa gồm 6 trong Bảng 5.5). Cuối cùng, tôi xem xét mức nghiêm trọng nhất
theo WER trong khi giữ khoảng cách cosin dưới 0,5 (6 ví dụ cuối cùng
trong Bảng 5.5). Những ví dụ về trường hợp ngoại lệ này cho thấy những lợi thế và hạn chế tiềm ẩn
của WER, điểm tình cảm và điểm dựa trên nhúng văn bản.

5.2.5 Ưu điểm và hạn chế

Tất cả các ví dụ sau đây trong phần này đều được lấy từ Bảng 5.5.

WER có ưu điểm chính là đơn giản và nhất quán; nó chỉ là bản chỉnh sửa
khoảng cách được chuẩn hóa theo tổng số từ. Không có nhiều mô hình
giống như có nhiều mô hình ngôn ngữ khác nhau để phân tích tình cảm và để tạo ra
nhúng văn bản. Hạn chế chính của WER là, vì nó dựa trên chỉnh sửa
khoảng cách và không có bất kỳ sự hiểu biết hoặc mô hình ngôn ngữ nào, có những lỗi nghiêm trọng

Bảng 5.5: Ví dụ về lỗi nghiêm trọng. Nhóm 6 đầu tiên và nhóm 6 thứ hai dựa trên tình cảm và nhúng văn bản tương ứng trong khi WER được giữ dưới 0,5. Cuối cùng 6 dựa trên WER trong đó khoảng cách cosin của nhúng văn bản được giữ dưới 0,5.

Sự thật thực tế Đầu ra ASR uh	FLAIR MiniLM WER ASR		
tôi hút khoảng một gói một ngày uh hút khoảng một gói một ngày và	1.929	0,104	0,250 Whis.
bạn sử dụng ma túy đá thường xuyên như thế nào và bạn sử dụng bùn pha lê thường xuyên như thế	1.858	0,371	0,125 Whis.
nào? Nghe có vẻ như là một công việc khá căng thẳng và giống như một công việc khá căng	1.850	0,298	0,375 DS2
thẳng uhm nó bắt đầu từ đêm qua và nó bắt đầu từ đêm qua	1.707	0,138	0,200 W2V2
những gì họ đã làm cho cơn đau tim của bạn những gì họ đã làm cho đàn của bạn tấn công	1.617	0,546	0,143 W2V2
bất kỳ cuộc phẫu thuật trước đó bất kỳ cuộc phẫu thuật	1.580	0,111	0,333 DS2
nào trước đó dường như không có tác dụng gì... dorthins dường như làm cho anh ta bất	0,003	0,692	0,364 W2V2
kỳ ... những gì họ đã làm cho cơn đau tim của bạn họ đã làm gì cho cuộc tấn công của bầy đàn của	1.617	0,546	0,143 W2V2
bạn và bạn sử dụng ma túy đá thường xuyên như thế nào và bạn sử dụng ánh sáng mặt trời thường	0,010	0,512	0,250 Vosk
xuyên như thế nào mà bạn lại bị đau ngực rằng bạn đang trải qua một số thử nghiệm về	0,049	0,469	0,333 Tiếng rít.
cùng một điều ổn chứ và nó đã trở nên... cùng một moqe và nó đã trở nên nhiều hơn...	0,028	0,461	0,200 W2V2
rằng bạn đang bị đau ngực rằng bạn đang trải qua một số cuộc trò chuyện	1.889	0,456	0,333 Vosk
được rồi được	1.094	0,061	1.000 DS2
rồi một loại vitamin tổng hợp một loại vitamin tổng hợp	0,000	0,150	1.000 DS2
cha mẹ tôi bạn bè của chúng tôi	0,005	0,370	1,00PS
tôi đã thử rồi tôi đã thử thêm	1.733	0,451	1.000 Vosk
uh ba mươi tám độ 38 độ uh ba	0,161	0,177	0,750 Whis.
mươi tám độ các cấp độ	0,007	0,324	0,750 DS2

có WER tương đối thấp và ngược lại, có những lỗi không nghiêm trọng
WER cao như multivitamin so với multivitamin.

Tình cảm có những hạn chế lớn do thực tế là các thuật toán được sử dụng để
tạo ra điểm số tình cảm được thiết kế chỉ để nắm bắt mức độ tích cực hay tiêu cực
văn bản đã cho là. Điểm tình cảm cũng có thể rất nhạy cảm với những sai sót trong sự trôi chảy
như um hoặc uh. Điều này được đánh dấu trong ví dụ, uhm nó bắt đầu vào đêm qua vs
và nó bắt đầu từ đêm qua, khi có sự khác biệt lớn về tâm lý là 1,707.
Điều này có thể là một lợi thế hoặc một hạn chế tùy thuộc vào tình huống. Nhiều ASR
động cơ bỏ qua sự thiếu lưu loát, nhưng, ví dụ trong tương tác giữa người và robot hoặc nói
hệ thống đối thoại, sự thiếu lưu loát có thể rất quan trọng đối với sự hiểu biết và hiệu suất [5, 8].
Bên cạnh sự tương tác giữa người và máy tính, sự thiếu lưu loát trong bản ghi chép cũng có thể rất quan trọng
trong bối cảnh chăm sóc sức khỏe, nơi sự khác biệt về khả năng nói trôi chảy được sử dụng như là yếu tố dự đoán
tình trạng mất trí nhớ [14, 35, 40].

Ngoài ra còn có giới hạn về hiệu suất của mô hình, trong đó mô hình
phân loại sai cảm xúc. Trong các ví dụ bất kỳ ca phẫu thuật trước đó so với bất kỳ
phẫu thuật trước đó hoặc uh tôi hút khoảng một gói một ngày so với uh hút khoảng một gói một ngày,
có một sự khác biệt lớn về tình cảm nhưng sự khác biệt duy nhất là thiếu đại từ
i hoặc số nhiều của phẫu thuật, điều này không ảnh hưởng nhiều đến tình cảm.

Tuy nhiên, bất chấp những hạn chế này, tình cảm có thể phát hiện ra những lỗi nghiêm trọng khi
WER tương đối thấp. Trong ví dụ mà ma túy đá trở thành bùn tinh thể
hoặc nơi cơn đau ngực trở thành trò chuyện thì WER lần lượt là 0,125 và 0,333 nhưng
sự khác biệt về tình cảm rất cao ở mức 1,858 và 1,889.

Việc nhúng văn bản bị giới hạn bởi hiệu suất của mô hình, giống như tình cảm,
tuy nhiên, có thể nắm bắt được nhiều hơn là chỉ cực tính của một văn bản nhất định. Biết rằng nhiều trong số này
các mô hình được đào tạo theo cách tự giám sát bằng cách sử dụng ngữ cảnh trong văn bản đào tạo,
chúng ta có thể tưởng tượng ra sự gắn kết của cha mẹ tôi và bạn bè chúng tôi có thể giống nhau như thế nào.
Cả hai cụm từ này đều có thể xuất hiện trong các văn bản xung quanh tương tự; chúng có
cấu trúc ngữ pháp giống nhau (một tính từ sở hữu theo sau là một danh từ) và cha mẹ
và bạn bè đều là mối quan hệ của con người.

Một hạn chế khác của các mô hình này là lượng văn bản mà chúng có thể xử lý. Bất kỳ -
điều gì đó vượt quá giới hạn của mô hình sẽ bị cắt bớt và do đó mất đi ý nghĩa của
văn bản bị cắt bớt. Mặc dù có nhiều câu nói ngắn trong dữ liệu đào tạo ASR.
giới hạn ký tự trên các mô hình này có thể ảnh hưởng đến hiệu suất của các câu nói dài hơn.

Mặc dù có những hạn chế, nhúng văn bản vẫn có thể nắm bắt tốt sự khác biệt
về ý nghĩa. Việc nhúng văn bản có thể đưa ra điểm số cao cho các ví dụ trong đó
ma túy đá trở thành ánh sáng mặt trời và nơi cơn đau ngực trở thành thử thách khi WER
và điểm số tình cảm tương đối thấp. Nhúng văn bản cũng có thể đưa ra mức thấp
xếp hạng cho các cách viết khác nhau của từ okay và các con số (ok so với okay, hoặc uh thirty
tám độ so với 38 độ) khi WER cao.

5.3 Thí nghiệm 3: Sử dụng mức độ nghiêm trọng để cải thiện ASR

Cho đến thời điểm này, kết quả cho thấy rằng 1) có sự nhất quán đáng tin cậy giữa con người
người đánh giá, 2) khoảng cách cosin của nhúng văn bản tương quan tốt hơn với nhãn của con người
mức độ nghiêm trọng hơn WER và 3) sử dụng cảm xúc hoặc nhúng văn bản trong một số liệu để
đánh giá tổng thể của ASR nắm bắt thông tin khác với WER. Với những
kết quả đã được thiết lập, mục đích của thí nghiệm này là để kiểm tra xem một biện pháp tự động của

mức độ nghiêm trọng có thể được sử dụng không chỉ trong việc đánh giá ASR mà còn trong quá trình phát triển cũng như vậy.

Công việc trước đây được thực hiện trong nghiên cứu về lỗi ASR liên quan đến các phương pháp tự động hóa có thể phát hiện lỗi bằng cách sử dụng nhúng từ và văn bản (và thậm chí các tính năng khác như đặc điểm âm thanh/ngữ điệu), [19, 20, 21] và tự động sửa lỗi trong các các trường hợp (chẳng hạn như trong một số từ đồng âm trong tiếng Pháp) [11]. Tuy nhiên, thay vì lỗi ASR phát hiện hoặc sửa chữa xảy ra sau khi dự đoán, cách tiếp cận của tôi sẽ bao gồm mức độ nghiêm trọng trong việc tạo ra hệ thống ASR nhằm mục đích giảm số lượng lỗi (được đo bằng WER) và để giảm mức độ nghiêm trọng chung của các lỗi được tạo ra (được đo bằng khoảng cách cosin trung bình được đề xuất trong Mục 5.3.3). Để thực hiện điều này, Tôi kết hợp mức độ nghiêm trọng vào chức năng mất mát trong quá trình đào tạo ASR. Chính xác phương pháp để kết hợp mức độ nghiêm trọng vào hàm mất mát, mô hình được sử dụng và Các thí nghiệm tôi thực hiện được mô tả trong các phần sau.

5.3.1 Thêm mức độ nghiêm trọng vào hàm mất mát

Các công cụ ASR liên quan đến mạng nơ-ron thường được đào tạo bằng cách sử dụng Hàm mất mát của Phân loại thời gian kết nối (CTC) [22]. Thuật toán này cho phép một để làm việc với dữ liệu trong đó cả đầu vào và đầu ra có thể thay đổi về độ dài như trong nhận dạng chữ viết tay và nhận dạng giọng nói. Tổng CTC trên các xác suất của tất cả các sự sắp xếp có thể có giữa đầu vào và đầu ra. Đương nhiên, điều này có thể là khá tốn kém. Để khắc phục điều này, thuật toán CTC tận dụng lợi thế của động phương pháp lập trình để tính toán hiệu quả xác suất của mỗi đầu ra. Trong nói cách khác, đưa ra và đầu vào của âm thanh X và bản ghi chép thực tế Y , CTC có thể tính toán hiệu quả $p(Y|X)$.

Khi đào tạo một mạng lưới nơ-ron, lý tưởng nhất là chúng ta muốn tối đa hóa khả năng giá trị thực tế, Y, cho âm thanh tương ứng, X, càng gần 1 càng tốt. Vì khả năng có thể cực kỳ nhỏ nên các tham số của mô hình thường là được điều chỉnh, điều chỉnh, giảm thiểu khả năng log-likehood tiêu cực, $(X,Y)\ D\ \log p(Y|X)$, trong đó D là một tập huấn luyện nhất định.

Để kết hợp mức độ nghiêm trọng vào hàm mất mát, khoảng cách cosin được sử dụng như một trọng lượng trong hàm mất mát. Để tính trọng lượng này, w, khoảng cách cosin bị giới hạn trong phạm vi từ một số gần bằng không, $1{,}0\times10^{-7}$, đến 1. Điều này được thể hiện trong Phương trình 5.1, trong đó W là trọng số biểu thị mức độ nghiêm trọng giữa giá trị thực tế, Ytruth, và đầu ra của ASR, Ypred. Trọng số này được nhân với giá trị mất mát CTC để có được tổn thất cuối cùng (Phương trình 5.2).

Điều này dẫn đến một chức năng trong đó tổn thất CTC ban đầu được thu nhỏ lại, cùng với các gradient của mạng nơ-ron, dựa trên sự tương đồng về mặt ngữ nghĩa giữa đầu ra thực tế và ASR.

$$w = 1 \quad \max(1{,}0 \times 10^{-7}, \cos(Ytruth, Ypred)) \tag{5.1}$$

$$L = w \quad CTC \tag{5.2}$$

Tôi sẽ gọi hàm mất mát được đề xuất này là hàm mất mát CTC-by-Cosine do việc sử dụng CTC và độ tương tự cosin. Đối với thí nghiệm này, tôi sử dụng all-MiniLM-L6-v (MiniLM) để tạo ra các nhúng cho dự đoán thực tế và ASR.

5.3.2 Mô hình

Với kế hoạch về cách kết hợp mức độ nghiêm trọng vào quá trình phát triển hệ thống ASR đã sửa, bây giờ tôi sẽ mô tả mô hình tôi sử dụng trong thí nghiệm này.

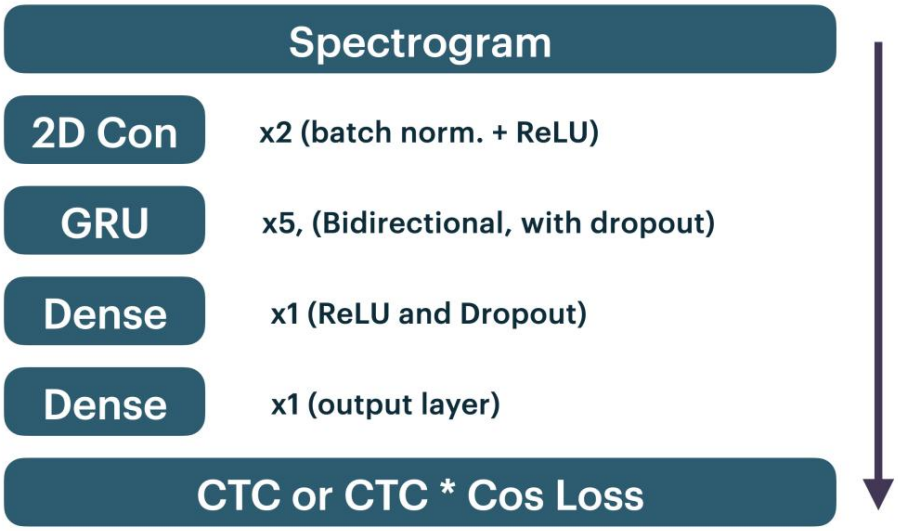
Mô hình dựa trên DeepSpeech2 [3], trong đó đầu vào là quang phổ từ các tập tin âm thanh và đầu ra là phân phối xác suất của trên một tập hợp các ký tự tại mỗi bước thời gian. Bộ ký tự bao gồm tất cả các chữ cái của bảng chữ cái tiếng Anh cùng với các ký tự sau: dấu nháy đơn, dấu hỏi, dấu chấm than, và ký hiệu trống.

Mô hình bắt đầu với hai lớp tích chập 2D, cả hai đều có 32 bộ lọc và hàng loạt chuẩn hóa và trải qua hàm kích hoạt ReLU sau mỗi lớp. hạt nhân cho các lớp tích chập là [11, 41] và [11, 21]. Sau các phép tích chập, có năm lớp hồi quy có cổng song hướng (GRU), mỗi lớp có 512 đơn vị với lớp bỏ qua với tỷ lệ 0,5 sau mỗi lớp lặp lại ngoại trừ lớp cuối cùng.

Sau lớp tái diễn cuối cùng có hai lớp dày đặc. Lớp đầu tiên duy trì có cùng kích thước với các lớp hồi quy và được truyền qua hàm kích hoạt ReLU và một lớp bỏ qua với tỷ lệ 0,5. Lớp dày đặc thứ hai là lớp đầu ra với softmax là hàm kích hoạt. Adam được sử dụng để tối ưu hóa với một học tập tỷ lệ $1{,}0 \times 10^{-4}$. Hình 5.2 mô tả các thành phần cốt lõi của mô hình này.

Điều này dẫn đến một mô hình có khoảng 26M tham số. Con số này tương đối nhỏ so với đến các hệ thống ASR khác. Ví dụ, DeepSpeech2 có 38M tham số (khoảng 1,5 lần nhiều thông số hơn so với mô hình được sử dụng trong công trình này), phiên bản cơ sở của Wav2Vec2 có 95M tham số (nhiều hơn khoảng 3,7 lần tham số) và cơ sở Phiên bản Whisper chứa 74 triệu tham số (nhiều hơn khoảng 1,2 lần). Tuy nhiên, mục đích của thí nghiệm này không phải là để đạt được hiệu suất hiện đại

với một kiến trúc mới lạ, đó là để kiểm tra trên quy mô tương đối nhỏ khả năng xảy ra
sử dụng tính nghiêm ngặt trong quá trình phát triển hệ thống ASR.



Hình 5.2: Các thành phần chính của mô hình được triển khai dựa trên DeepSpeech2.

5.3.3 Chế độ dữ liệu và đào tạo

Bởi vì 109 lời nói được ghi chép lại của cuộc trò chuyện giữa bác sĩ và bệnh nhân được mô phỏng
các tập tin không đủ để truyền dữ liệu từ một công cụ ASR, đối với thí nghiệm này tôi sử dụng
Bộ dữ liệu giọng nói LJ bao gồm “13.100 đoạn âm thanh ngắn của một người nói
đọc các đoạn trích từ 7 cuốn sách phi hư cấu” [25].

Tôi đào tạo 2 mô hình trên 90% dữ liệu đầu tiên, giữ lại 10% dữ liệu cuối cùng
để xác thực. Mô hình đầu tiên có kiến trúc được mô tả ở trên và chỉ sử dụng
hàm mất mát CTC. Mô hình này sẽ đóng vai trò là đường cơ sở. Mô hình thứ hai sử dụng
cùng một kiến trúc và chế độ đào tạo chính xác, nhưng sử dụng tổn thất CTC-by-Cosine
chức năng trong 5 kỷ nguyên cuối cùng.

Bảng 5.6: Tỷ lệ phần trăm tăng hiệu suất từ mô hình cơ sở sang mô hình CTC-by-Cos trên cả tập dữ liệu đào tạo và xác thực. Mức độ nghiêm trọng được đo bằng cosin trung bình khoảng cách được đề xuất trong Mục 5.3.3.

	Tàu hỏa WER	Tàu hỏa Cos	Val WER	Val Cos
Căn cứ	0,058	0,051	0,268	0,249
CTC-by-Cos	0,008	0,006	0,219	0,201

Để đánh giá, tôi sẽ xem xét WER và mức độ nghiêm trọng (được đo bằng giá trị trung bình khoảng cách cosin từ Phần) trên các tập dữ liệu đào tạo và xác thực trong suốt 50 thời kỳ đào tạo.

5.3.4 Kết quả

Kết quả cho thấy sự cải thiện về cả mức độ nghiêm trọng (khoảng cách cosin trung bình) và WER khi kết hợp mức độ nghiêm trọng vào chức năng mất mát. Mô hình CTC-by-Cos, được hiển thị ở trên cải thiện 85% về mức độ nghiêm trọng và WER trên tập dữ liệu đào tạo và trên 18% cải thiện về mức độ nghiêm trọng và WER trên tập dữ liệu xác thực (xem Bảng 5.6). Điều này cải thiện hiệu suất cho thấy có tiềm năng sử dụng mức độ nghiêm trọng trong phát triển ASR để giảm cả mức độ nghiêm trọng tổng thể và WER.

CHƯƠNG 6

PHẦN KẾT LUẬN

6.1 Thảo luận: Ý nghĩa và công việc trong tương lai

Do những hạn chế trong WER và dựa trên kết quả của công trình này, tôi đề xuất
nhìn vào khoảng cách cosin trung bình giữa các nhúng thực tế truyền
ghi chú và đầu ra ASR kết hợp với WER (Số liệu 3 trong 5.3.3). Tuy nhiên,
khi sử dụng phương pháp này người ta phải nhận thức rằng các mô hình ngôn ngữ đó
tạo nhúng văn bản không hoàn hảo và có thể thay đổi. Các trình phân tích tình cảm có
được chứng minh là rất hạn chế vì chúng chỉ nắm bắt được một khía cạnh hoặc chất lượng của một lời nói,
tuy nhiên, chúng vẫn có thể có vai trò trong việc phát hiện lỗi trong phủ định hoặc một số từ nhất định
(tức là "là" so với "không phải" hoặc "đau đớn" so với "trả tiền").

Trong Thí nghiệm 2, WER nhìn chung đồng ý với các số liệu khác, nhưng dựa trên
nghiên cứu trước đây đã thực hiện cho thấy những hạn chế của WER, một lĩnh vực nghiên cứu tiếp theo
sẽ nghiên cứu xem các số liệu được đề xuất trong công trình này có tốt hơn hay không
chỉ số về khả năng hiểu hoặc hiệu suất trên nhiệm vụ NLU hơn WER [18, 54, 48].

Công trình này cũng cung cấp cơ sở lý thuyết cho việc sử dụng nhúng văn bản trong
đào tạo ASR. Công trình gần đây đã xuất hiện cho thấy rằng sử dụng căn chỉnh ngữ nghĩa
(sử dụng nhúng văn bản) để Hiểu ngôn ngữ nói trong quá trình đào tạo là rất
hứa hẹn [28, 30]. Kết hợp mức độ nghiêm trọng với hàm mất mát chung, Connectionist
Phân loại theo thời gian, trong chế độ đào tạo của động cơ ASR trong Thí nghiệm

3 cho thấy có tiềm năng tối ưu hóa WER và mức độ nghiêm trọng cùng một lúc để tốt hơn
kết quả. Với nhiều phương pháp học máy, những kết quả này có thể thay đổi rất nhiều tùy thuộc vào
về kiến trúc, dữ liệu và chế độ đào tạo, do đó cần thử nghiệm thêm
là cần thiết để kiểm tra đầy đủ hơn tiềm năng của phương pháp này.

Tôi cũng thấy tiềm năng nhúng văn bản được sử dụng rộng rãi hơn trong đánh giá
vượt ra ngoài lĩnh vực ASR và vào bất kỳ lĩnh vực nào liên quan đến việc tạo ngôn ngữ tự nhiên
chẳng hạn như trong tóm tắt văn bản tự động hoặc thậm chí dịch máy với
phát triển nhiều mô hình ngôn ngữ đa ngôn ngữ hơn [16]. Tạo ngôn ngữ tự nhiên
cũng quan trọng trong chăm sóc sức khỏe để tạo ra những lời giải thích dễ hiểu cho AI
mô hình [7], tuy nhiên, tương tự như WER, trong lĩnh vực tạo ngôn ngữ tự nhiên, phổ biến
các số liệu như ROGUE [33] và BLEU [42] đã được chứng minh là có mối tương quan kém với
đánh giá của con người [7, 29]. Dựa trên công trình này, tôi tin rằng công trình trong tương lai có thể liên quan đến
kết hợp các biện pháp nghiêm ngặt này vào số liệu để tạo ngôn ngữ tự nhiên
nhiệm vụ và nghiên cứu cách thức chúng tương quan với các đánh giá của con người theo cách tương tự
đối với các số liệu như BERTScore hoặc BLEURT [57, 50].

6.2 Tóm tắt

Trong công trình này, tôi thử nghiệm với các trình phân tích tình cảm và nhúng văn bản trong
phân tích lỗi ASR. Cụ thể hơn, tôi muốn trả lời những câu hỏi sau:
1) các nhà phân tích tình cảm và/hoặc nhúng văn bản có tương quan với xếp hạng của con người không
mức độ nghiêm trọng, và nếu có, ở mức độ nào? 2) chúng ta có thể sử dụng các công cụ phân tích tình cảm và/hoặc văn bản
nhúng như một biện pháp hữu ích hoặc ước tính mức độ nghiêm trọng của lỗi ASR, 3) chúng ta có thể
sử dụng các biện pháp này để đánh giá chất lượng chung về mức độ nghiêm trọng của động cơ ASR?,
và 4) mức độ nghiêm trọng có thể hữu ích trong việc phát triển hệ thống ASR không.

Đầu tiên tôi tạo bộ dữ liệu của riêng mình gồm 150 lỗi ASR và 3 đánh giá của con người về mức độ nghiêm trọng sử dụng bộ dữ liệu âm thanh mô phỏng các cuộc trò chuyện giữa bệnh nhân và bác sĩ có kèm theo bản ghi chép và 3 người đánh giá trong lĩnh vực y tế. Là một bước đầu tiên, tôi chỉ ra rằng có sự nhất quán giữa những người đánh giá.

Để trả lời hai câu hỏi đầu tiên, tôi sử dụng sự khác biệt trong điểm số tình cảm từ 3 bộ phân tích tình cảm và khoảng cách cosin của nhúng văn bản từ 4 ngôn ngữ các mô hình như các biện pháp về mức độ nghiêm trọng. Tôi xem xét mối tương quan giữa các biện pháp này và đánh giá của con người về mức độ nghiêm trọng cũng như xem xét khả năng dự đoán mức độ nghiêm trọng của họ bằng cách sử dụng hồi quy logistic thứ tự đơn giản. Chúng được so sánh với WER, một số liệu phổ biến để đánh giá ASR. Trong khi điểm số tình cảm không thể dự đoán mức độ nghiêm trọng cũng như WER, đối với tất cả các mô hình nhúng văn bản, khoảng cách cosin từ nhúng văn bản đến Đầu ra ASR dự đoán mức độ nghiêm trọng tốt hơn WER.

Tôi đề xuất một phương pháp đơn giản để kết hợp các biện pháp này vào số liệu để đánh giá đã đánh giá chất lượng tổng thể của động cơ ASR. Nhìn chung, kết quả phù hợp với WER, (tức là ASR với WER tốt nhất hoạt động tốt nhất về số liệu dựa trên tình cảm điểm số hoặc khoảng cách cosin của nhúng văn bản). Tôi chỉ ra rằng, khi xem xét kỹ hơn, các số liệu này đang nắm bắt những phẩm chất khác nhau trong lỗi ASR và có thể khắc phục một số về những hạn chế của WER.

Cuối cùng, tôi thử nghiệm bằng cách kết hợp mức độ nghiêm trọng vào quá trình phát triển ASR hệ thống. Kết quả cho thấy có khả năng nghiêm trọng để giúp cải thiện hiệu suất.

TÀI LIỆU THAM KHẢO

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter và Roland Vollgraf. FLAIR: Một khuôn khổ dễ sử dụng cho NLP hiện đại.
Trong NAACL 2019, Hội nghị thường niên năm 2019 của Chi nhánh Bắc Mỹ thuộc Hiệp hội Ngôn ngữ học tính toán (Trình bày), trang 54–59, 2019.

[2] Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri và Maha Almojil. Nhận dạng giọng nói tự động: Đánh giá tài liệu có hệ thống.
Truy cập IEEE, 9:131858–131876, 2021.

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: Nhận dạng giọng nói đầu cuối bằng tiếng Anh và tiếng Quan Thoại. Trong Hội nghị quốc tế về máy học, trang 173–182.
PMLR, 2016.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed và Michael Auli. wav2vec 2.0: Một khuôn khổ cho việc học tự giám sát các biểu diễn giọng nói.
Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 33:12449–12460, 2020.

[5] Timo Baumann, Casey Kennington, Julian Hough và David Schlangen. Nhận dạng lời nói đàm thoại: Một asr gia tăng nên làm gì cho một hệ thống đối thoại và cách thực hiện. Đối thoại với Robot xã hội: Khả năng, Phân tích và Đánh giá, trang 421–432, 2017.

[6] Steven Bird và Edward Loper. NLTK: Bộ công cụ ngôn ngữ tự nhiên. Trong Biên bản báo cáo của ACL Interactive Poster and Demonstration Sessions, trang 214–217, Barcelona, Tây Ban Nha, tháng 7 năm 2004. Hiệp hội Ngôn ngữ học tính toán.

[7] Alodie Boissonnet, Marzieh Saeidi, Vassilis Plachouras và Andreas Vlachos.
Đánh giá có thể giải thích được các bài viết về chăm sóc sức khỏe với QA. Trong Biên bản Hội thảo lần thứ 21 về Xử lý ngôn ngữ y sinh, trang 1–9, Dublin, Ireland, tháng 5 năm 2022. Hiệp hội Ngôn ngữ học tính toán.

[8] Herbert H Clark và Jean E Fox Tree. Sử dụng uh và um trong lời nói tự phát.
Nhận thức, 84(1):73–111, 2002.

[9] Jacob Cohen. Hệ số đồng thuận cho thang đo danh nghĩa. Giáo dục và đo lường tâm lý, 20(1):37–46, 1960.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee và Kristina Toutanova. Bert: Đào tạo trước các bộ biến đổi song hướng sâu để hiểu ngôn ngữ. Bản in trước arXiv arXiv:1810.04805, 2018.

[11] Richard Dufour và Yannick Est`eve. Sửa lỗi đầu ra asr: giải pháp cụ thể cho các lỗi cụ thể trong tiếng Pháp. Trong Hội thảo công nghệ ngôn ngữ nói IEEE năm 2008, trang 213–216. IEEE, 2008.

[12] Rosa Falotico và Piero Quatto. Thống kê kappa của Fleiss không có nghịch lý. Chất lượng và Số lượng, 49(2):463–470, 2015.

[13] Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Cheva-lier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. Một tập dữ liệu phỏng vấn y tế giữa bệnh nhân và bác sĩ mô phỏng tập trung vào các trường hợp hô hấp. Dữ liệu khoa học, 9(1):1–7, 2022.

[14] Shahla Farzana, Ashwin Deshpande và Natalie Parde. Bạn nói điều đó quan trọng như thế nào: Đo lường tác động của các thẻ nói lắp bắp đối với việc phát hiện chứng mất trí tự động. Trong Biên bản Hội thảo lần thứ 21 về Xử lý ngôn ngữ y sinh, trang 37–48, Dublin, Ireland, tháng 5 năm 2022. Hiệp hội Ngôn ngữ học tính toán.

[15] Shahla Farzana và Natalie Parde. Khám phá dự đoán điểm mmse bằng cách sử dụng tín hiệu bằng lời và không bằng lời. Trong INTERSPEECH, trang 2207–2211, 2020.

[16] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan và Wei Wang. Nhúng câu BERT không phụ thuộc vào ngôn ngữ. Trong Biên bản báo cáo của Hội nghị thường niên lần thứ 60 của Hiệp hội Ngôn ngữ học tính toán (Tập 1: Bài báo dài), trang 878–891, Dublin, Ireland, tháng 5 năm 2022. Hiệp hội Ngôn ngữ học tính toán.

[17] Joseph L Fleiss. Đo lường sự đồng thuận về thang đo danh nghĩa giữa nhiều người đánh giá. Bản tin tâm lý, 76(5):378, 1971.

[18] Olivier Galibert, Mohamed Ameur Ben Jannet, Juliette Kahn và Sophie Rosset. Tạo danh sách lỗi được sắp xếp theo nhiệm vụ để nhận dạng giọng nói. Trong Biên bản Hội nghị quốc tế lần thứ mười về Tài nguyên ngôn ngữ và Đánh giá (LREC'16), trang 1883–1889, Portoroˇz, Slovenia, tháng 5 năm 2016. Hiệp hội tài nguyên ngôn ngữ châu Âu (ELRA).

[19] Sahar Ghannay, Yannick Esteve và Nathalie Camelin. Kết hợp nhúng từ và mạng nơ-ron để tăng cường độ mạnh mẽ trong phát hiện lỗi asr. Trong

Hội nghị xử lý tín hiệu châu Âu lần thứ 23 năm 2015 (EUSIPCO), trang 1671–1675. IEEE, 2015.

[20] Sahar Ghannay, Yannick Est`eve và Nathalie Camelin. Nhúng câu cụ thể cho nhiệm vụ để phát hiện lỗi ASR. Trong Interspeech 2018, Hyderabad, Ấn Độ, tháng 9 năm 2018. ISCA.

[21] Sahar Ghannay, Yannick Est`eve và Nathalie Camelin. Một nghiên cứu về biểu diễn từ và câu không gian liên tục được áp dụng để phát hiện lỗi asr. Giao tiếp bằng lời nói, 120:31 – 41, 2020.

[22] Alex Graves, Santiago Fern'andez, Faustino Gomez, và J¨urgen Schmidhuber. Phân loại thời gian theo chủ nghĩa kết nối: gắn nhãn dữ liệu chuỗi chưa phân đoạn bằng mạng nơ-ron hồi quy. Trong Kỷ yếu hội nghị quốc tế lần thứ 23 về Học máy, trang 369–376, 2006.

[23] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Rav-ishankar và Alexander I Rudnicky. Pocketsphinx: Hệ thống nhận dạng giọng nói liên tục, thời gian thực miễn phí cho các thiết bị cầm tay. Năm 2006, Hội nghị quốc tế IEEE về âm học Biên bản xử lý tín hiệu và giọng nói, tập 1, trang I–I. IEEE, 2006.

[24] Clayton Hutto và Eric Gilbert. Vader: Một mô hình dựa trên quy tắc tiết kiệm để phân tích tình cảm của văn bản phương tiện truyền thông xã hội. Trong Biên bản báo cáo của hội nghị quốc tế AAAI về web và phương tiện truyền thông xã hội, tập 8, trang 216–225, 2014.

[25] Keith Ito và Linda Johnson. Bộ dữ liệu giọng nói lj. https://keithito.com/ Bộ dữ liệu giọng nói LJ/, 2017.

[26] Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis và Linda Dawson. Một đánh giá có hệ thống về công nghệ nhận dạng giọng nói trong chăm sóc sức khỏe. Tin học y tế BMC và ra quyết định, 14(1):1–14, 2014.

[27] Maurice G Kendall. Một thước đo mới về tương quan thứ hạng. Biometrika, 30(1/2):81–93, 1938.

[28] Sameer Khurana, Antoine Laurent và James Glass. Samu-xlsr: Biểu diễn giọng nói đa phương thức liên kết ngữ nghĩa ở cấp độ phát ngôn. Tạp chí IEEE về các chủ đề được chọn trong xử lý tín hiệu, 16(6):1493–1504, 2022.

[29] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong và Richard Socher. Tóm tắt văn bản thần kinh: Một đánh giá quan trọng. Trong Biên bản báo cáo Hội nghị năm 2019 về Phương pháp thực nghiệm trong Ngôn ngữ tự nhiên

Xử lý và Hội nghị chung quốc tế lần thứ 9 về Xử lý ngôn ngữ tự nhiên (EMNLP-IJCNLP), trang 540–551, Hồng Kông, Trung Quốc, tháng 11 năm 2019. Hiệp hội ngôn ngữ học tính toán.

[30] Ga¨elle Laperri`ere, Valentin Pelloin, Micka¨el Rouvier, Themos Stafylakis và Yannick Est`eve. Về việc sử dụng các biểu diễn lời nói được căn chỉnh ngữ nghĩa để hiểu ngôn ngữ nói. Trong Hội thảo Công nghệ Ngôn ngữ Nói (SLT) của IEEE năm 2022, trang 361–368, 2023.

[31] Quốc Lê và Tomas Mikolov. Biểu diễn phân tán của câu và tài liệu. Trong Hội nghị quốc tế về học máy, trang 1188–1196.
PMLR, 2014.

[32] Vladimir I Levenshtein và cộng sự. Mã nhị phân có khả năng sửa lỗi xóa, chèn và đảo ngược. Trong Vật lý Liên Xô doklady, tập 10, trang 707–710.
Liên Xô, 1966.

[33] Chin-Yew Lin. ROUGE: Một gói để đánh giá tự động các bản tóm tắt. Trong Text Summarization Branches Out, trang 74–81, Barcelona, Tây Ban Nha, tháng 7 năm 2004. Hiệp hội Ngôn ngữ học tính toán.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer và Veselin Stoyanov. Roberta: Một phương pháp đào tạo trước bert được tối ưu hóa mạnh mẽ. Bản in trước arXiv arXiv:1907.11692, 2019.

[35] Karmele Lopez-de Ipi˜na, Unai Martinez-de Lizarduy, Pilar M Calvo, Blanca Beitia, Joseba Garcia-Melero, Miriam Ecay-Torres, Ainara Estanga và Marcos Faundez-Zanuy. Phân tích sự thiếu trôi chảy để tự động phát hiện sự truyền đạt nhận thức nhẹ: một phương pháp học sâu. Trong Hội nghị và Hội thảo quốc tế năm 2017 về Trí thông minh lấy cảm hứng từ sinh học (IWOBI), trang 1–4. IEEE, 2017.

[36] Steven Loria et al. tài liệu textblob. Phiên bản 0.15, 2(8), 2018.

[37] Daniel Luzzati, Cyril Grouin, Ioana Vasilescu, Martine Adda-Decker, Eric Bilin-ski, Nathalie Camelin, Juliette Kahn, Carole Lailler, Lori Lamel và Sophie Rosset. Chú thích của con người về vùng lỗi asr: “Trọng lực” có phải là một khái niệm có thể chia sẻ cho những người chú thích không? Trong Biên bản báo cáo Hội nghị quốc tế lần thứ chín về tài nguyên ngôn ngữ và đánh giá (LREC'14), trang 3050–3056, 2014.

[38] Tomas Mikolov, Kai Chen, Greg Corrado và Jeffrey Dean. Ước tính hiệu quả các biểu diễn từ trong không gian vectơ. Bản in trước arXiv arXiv:1301.3781, 2013.

[39] Tomas Mikolov, Wen-tau Yih và Geoffrey Zweig. Các quy luật ngôn ngữ trong các biểu diễn từ không gian liên tục. Trong Biên bản báo cáo Hội nghị năm 2013 của Chi hội Bắc Mỹ thuộc Hiệp hội Ngôn ngữ học tính toán: Công nghệ ngôn ngữ của con người, trang 746–751, Atlanta, Georgia, tháng 6 năm 2013.
Hiệp hội Ngôn ngữ học tính toán.

[40] Kimberly D Mueller, Rebecca L Koscik, Bruce P Hermann, Sterling C John-son và Lyn S Turkstra. Sự suy giảm trong ngôn ngữ kết nối có liên quan đến suy giảm nhận thức nhẹ rất sớm: Kết quả từ sổ đăng ký wisconsin để phòng ngừa bệnh Alzheimer.
Frontiers in Aging Neuroscience, 9:437, 2018.

[41] Hillary Ngai và Frank Rudzicz. Bác sĩ XAvIer: Chẩn đoán có thể giải thích được về các cuộc đối thoại giữa bác sĩ và bệnh nhân và đánh giá XAI. Trong Biên bản Hội thảo lần thứ 21 về Xử lý ngôn ngữ y sinh, trang 337–344, Dublin, Ireland, tháng 5 năm 2022. Hiệp hội Ngôn ngữ học tính toán.

[42] Kishore Papineni, Salim Roukos, Todd Ward, và Wei-Jing Zhu. Bleu: một phương pháp đánh giá tự động bản dịch máy. Trong Biên bản báo cáo của cuộc họp thường niên lần thứ 40 của Hiệp hội Ngôn ngữ học tính toán, trang 311–318, 2002.

[43] Jeffrey Pennington, Richard Socher và Christopher D Manning. Glove: Các vectơ toàn cục để biểu diễn từ. Trong Biên bản báo cáo hội nghị năm 2014 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên (EMNLP), trang 1532–1543, 2014.

[44] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee và Luke Zettlemoyer. Biểu diễn từ ngữ theo ngữ cảnh sâu. Trong Biên bản báo cáo Hội nghị năm 2018 của Chi hội Bắc Mỹ thuộc Hiệp hội Ngôn ngữ học tính toán: Công nghệ ngôn ngữ của con người, Tập 1 (Bài báo dài), trang 2227–2237, New Orleans, Louisiana, tháng 6 năm 2018.
Hiệp hội Ngôn ngữ học tính toán.

[45] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. Bộ công cụ nhận dạng giọng nói kaldi. Trong hội thảo IEEE 2011 về nhận dạng và hiểu giọng nói tự động, số CONF. IEEE Signal Processing Society, 2011.

[46] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey và Ilya Sutskever. Nhận dạng giọng nói mạnh mẽ thông qua giám sát yếu quy mô lớn. Bản in trước arXiv arXiv:2212.04356, 2022.

[47] Nils Reimers và Iryna Gurevych. Sentence-bert: Nhúng câu sử dụng mạng siamese bert. Trong Biên bản báo cáo Hội nghị năm 2019 về Phương pháp thực nghiệm trong Xử lý ngôn ngữ tự nhiên. Hiệp hội Ngôn ngữ học tính toán, 11 2019.

[48] Giuseppe Riccardi và Allen L Gorin. Các mô hình ngôn ngữ ngẫu nhiên để nhận dạng và hiểu giọng nói. Trong ICSLP, 1998.

[49] Victor Sanh, Lysandre Debut, Julien Chaumond và Thomas Wolf. Distilbert, phiên bản tinh chế của bert: nhỏ hơn, nhanh hơn, rẻ hơn và nhẹ hơn. Bản in trước arXiv arXiv:1910.01108, 2019.

[50] Thibault Sellam, Dipanjan Das và Ankur P Parikh. Bleurt: Học các số liệu mạnh mẽ để tạo văn bản. Bản in trước arXiv arXiv:2004.04696, 2020.

[51] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu và Tie-Yan Liu. Mpnet: Tiền đào tạo được che dấu và hoán vị để hiểu ngôn ngữ. Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, 33:16857–16867, 2020.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser và Illia Polosukhin. Sự chú ý là tất cả những gì bạn cần. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 30, 2017.

[53] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang và Ming Chu. Minilm: Chưng cất sự chú ý sâu sắc cho việc nén không phụ thuộc vào tác vụ của các bộ biến áp được đào tạo trước. Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, 33:5776–5788, 2020.

[54] Ye-Yi Wang, Alex Acero và Ciprian Chelba. Tỷ lệ lỗi từ có phải là một chỉ báo tốt cho độ chính xác của việc hiểu ngôn ngữ nói không. Trong hội thảo IEEE năm 2003 về nhận dạng và hiểu giọng nói tự động (IEEE Cat. No. 03EX721), trang 577–582. IEEE, 2003.

[55] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov và Quoc V Le. Xlnet: Tiền huấn luyện tự hồi quy tổng quát để hiểu ngôn ngữ. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 32, 2019.

[56] Dong Yu và Li Deng. Nhận dạng giọng nói tự động, tập 1. Springer, 2016.

[57] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger và Yoav Artzi. Bertscore: Đánh giá việc tạo văn bản bằng bert. Bản in trước arXiv arXiv:1904.09675, 2019.