S-capade: Sửa lỗi chính tả nhằm mục đích Lỗi lệch lạc đặc biệt

Emma O'Neill1 , Robert Young2 , Elsa Thiaville2 , Muireann MacCarthy2 và Julie Carson-Berndsen1 , Anthony Ventresque2

Trung tâm 1ADAPT, Trường Khoa học Máy tính, Đại học Cao đẳng Dublin, Ireland Trung tâm Nghiên cứu 2Lero, Trường Khoa học Máy tính, Đại học Cao đẳng Dublin, Ireland
{emma.l.oneill,robert.young2,elsa.thiaville,
muireann.maccarthy}@ucdconnect.ie
{julie.berndsen,anthony.ventresque}@ucd.ie

Trừu tượng. S-capade (sửa lỗi chính tả nhằm vào các lỗi sai lệch đặc biệt) là một công cụ kiểm tra chính tả dựa trên khoảng cách âm vị1 nhằm mục đích sửa lỗi chính tả do trẻ em mắc phải. Trong khi lỗi chính tả thường lệch khỏi mục tiêu chỉ một hoặc hai ký tự, thì lỗi chính tả của trẻ em lại có xu hướng thiên về ngữ âm hơn. Họ bị ảnh hưởng cả bởi cách trẻ cảm nhận cách phát âm của một từ và bởi các chữ cái mà chúng chọn để thể hiện cách phát âm đó. Do đó, những lỗi chính tả này đặc biệt sai lệch so với mục tiêu và có thể tác động tiêu cực đến hiệu suất của các trình kiểm tra chính tả thông thường. Trong bài viết này, chúng tôi chứng minh rằng S-capade có khả năng sửa một phần đáng kể lỗi chính tả của trẻ em khi các công cụ sửa lỗi thông thường không thành công.

Từ khóa: Sửa lỗi chính tả · Khoảng cách âm vị · Chính tả trẻ em

1. Giới thiệu

Lỗi chính tả thường được coi là một trong hai loại: typographic hoặc nhận thức [22]. Lỗi đánh máy là kết quả của sự phối hợp vận động; có lẽ thay thế một ký tự cho một ký tự liền kề trên bàn phím. Mặt khác, lỗi nhận thức xuất phát từ quan niệm sai lầm hoặc thiếu kiến thức về cách viết đúng chính tả của một từ. Một tập hợp con cụ thể của những lỗi nhận thức này được gọi là lỗi ngữ âm trong đó người viết viết sai chính tả, mặc dù không đúng về mặt chính tả nhưng vẫn nắm bắt được chuỗi ngữ âm của từ mục tiêu.

Những lỗi ngữ âm này đặc biệt phổ biến trong việc đánh vần của trẻ em, từ lâu đã được coi là dựa trên ngữ âm. Trong một cuộc kiểm tra khả năng viết chính tả ban đầu của trẻ , Read [29] đã thảo luận về ảnh hưởng đáng kể của âm thanh lời nói và mối quan hệ giữa chúng. Vì vậy, mặc dù một số lỗi chính tả có vẻ “kỳ lạ” và sai lệch nhiều so với từ mục tiêu nhưng chúng có xu hướng phản ánh khả năng phán đoán về mặt ngữ âm của trẻ. Ngoài ra, nó hiện đang phổ biến trong lớp học.

Cả hai tác giả đều có đóng góp như nhau cho bài viết này.

[1] Kho mã nguồn có thể tìm thấy ở phần tài liệu tham khảo [35].

cho trẻ được dạy đọc và viết bằng ngữ âm: một phương pháp tập trung vào mối quan hệ giữa chữ cái và âm thanh [33]. Vì vậy, trẻ em được khuyến khích sử dụng phương pháp 'phát âm' khi đánh vần những từ không quen thuộc - một phương pháp được những người đánh vần kém dựa vào [7].

Bất chấp sự phổ biến của các lỗi kiểu ngữ âm, các công cụ sửa lỗi chính tả thông thường không có đủ khả năng sửa các kiểu lỗi chính tả này và do đó, hiệu suất đánh vần của trẻ em kém hơn. Trong bài báo này chúng tôi trình bày một phương pháp sửa lỗi dựa trên sự tương đồng về âm vị có khả năng sửa lỗi chính tả tiếng Anh của trẻ sai lệch nhiều so với từ mục tiêu. Kukich [22] nhóm công việc sửa lỗi chính tả thành các nhiệm vụ riêng biệt; phát hiện lỗi, sửa lỗi riêng biệt và sửa lỗi phụ thuộc vào ngữ cảnh. Công việc này tập trung vào việc sửa lỗi riêng biệt, tạo ra danh sách các sửa lỗi có thể áp dụng từ thực tế dựa trên các thuộc tính ngữ âm của lỗi chính tả.

Tương tự như Hodge và Austin [17], mục tiêu của chúng tôi là tối đa hóa khả năng thu hồi thông qua việc tạo ứng viên khi chúng tôi hình dung phương pháp này như một thành phần của mô hình tổng thể sẽ xử lý việc lựa chọn ứng viên như một nhiệm vụ phụ thuộc vào ngữ cảnh.

2 công việc liên quan

Các thuật toán sửa lỗi chính tả ban đầu thường sử dụng khoảng cách chỉnh sửa ký tự giữa lỗi chính tả và sửa từ thực, dựa trên phát hiện rằng phần lớn lỗi chính tả khác nhau chỉ bằng một thao tác chỉnh sửa (chèn, xóa, thay thế hoặc chuyển vị) [8]. Những phương pháp này phù hợp với lỗi chính tả.

Tuy nhiên, lỗi chính tả về mặt ngữ âm thường sai lệch nhiều so với mục tiêu của từ thực tế [22]. Hiệu suất được cải thiện khi sử dụng các mô hình kênh nhiễu cho phép thực hiện nhiều thao tác chỉnh sửa [4]. Đặc biệt, Brill và Moore [3] đã chứng minh những cải tiến hiệu suất đáng kể đối với mô hình kênh nhiễu bằng cách tính toán xác suất của các chỉnh sửa từ chuỗi này sang chuỗi khác và kết hợp chúng khi so sánh lỗi chính tả với các sửa lỗi chính tả từ ứng viên thực tế.

Việc kết hợp thông tin ngữ âm vào các phương pháp này tỏ ra có lợi cho việc sửa lỗi chính tả. Veronis [36] đã sử dụng thuật toán khoảng cách chỉnh sửa có trọng số trong đó chi phí cho các hoạt động chỉnh sửa dựa trên sự tương tự về mặt ngữ âm giữa các biểu đồ. Người ta cũng thường chuyển đổi các từ từ dạng chính tả sang dạng thể hiện các đặc điểm ngữ âm của chúng. Ví dụ, Soundex, được mô tả bởi Kukich [22] và được cấp bằng sáng chế bởi Russel và Ordell [30], ánh xạ các từ thành mã chữ và số có độ dài cố định dựa trên các ký tự của nó. Các giá trị số được gán cho các nhóm chữ cái giống nhau về mặt ngữ âm. Do đó, các từ được phát âm giống nhau sẽ có cùng mã hóa (ví dụ: 'sure' và 'shore' đều có mã hóa S600). Các thuật toán khoảng cách chỉnh sửa có thể được áp dụng cho các mã hóa này để tìm các cách sửa từ thực tế có ngữ âm giống với lỗi chính tả. Tuy nhiên, Soundex đã bị chỉ trích là quá chung chung do tính hoán vị hạn chế của nó [17, 23]. Vì vậy, các quy tắc chuyển đổi ngữ âm, được xác định bởi kiến thức ngôn ngữ của ngôn ngữ đích, thường được sử dụng trước khi mã hóa [17, 28]. Ngoài ra, các dạng âm vị có thể được sử dụng trực tiếp bằng cách chuyển đổi một lỗi chính tả thành chuỗi âm vị tương ứng bằng cách sử dụng các quy tắc chuyển từ chữ cái sang âm thanh [9,

21, 23, 34]. Các cách tiếp cận khác để sửa lỗi chính tả bao gồm giải quyết vấn đề như một trong Dịch máy [2, 31] hoặc như một nhiệm vụ tổng hợp/nhận dạng [32].

Phương pháp được mô tả trong bài viết này kết hợp các yếu tố từ một số phương pháp này. Các lỗi chính tả được chuyển đổi thành chuỗi âm vị tương ứng bằng cách sử dụng công cụ chuyển biểu đồ thành âm vị được máy học [5] thay vì các quy tắc chuyển chữ cái thành âm thanh rõ ràng. Khoảng cách chỉnh sửa có trọng số được tính toán giữa các lỗi chính tả và các lần sửa từ thực tế bằng cách sử dụng ma trận khoảng cách giữa âm vị với âm vị dựa trên cả đặc tính âm học và phân bổ của âm vị. Phần 3 mô tả chi tiết phương pháp này, trong khi Phần 4 nêu chi tiết cách thiết lập thử nghiệm để so sánh phương pháp này với các công cụ sửa lỗi chính tả khác trên các bộ dữ liệu khác nhau. Kết quả của việc này được trình bày trong Phần 5, nơi chúng tôi chứng minh rằng S-capade có khả năng sửa một tỷ lệ đáng kể lỗi chính tả của trẻ em ngoài những lỗi được sửa bằng các công cụ khác.

Phương pháp 3 chữ S

Khi một đứa trẻ sử dụng phương pháp 'phát âm' để đánh vần, chúng đang ước chừng những âm thanh mà chúng cảm nhận được trong từ mục tiêu với các chữ cái mà chúng tin rằng đại diện cho những âm thanh đó. Như vậy, những sai lệch so với cách viết đúng xảy ra do nhận biết âm vị không chính xác, ví dụ như âm vị /V/2 được hiểu là /F/ dẫn đến lỗi chính tả 'gif' (give) và do chọn sai các chữ cái, ví dụ: đại diện cho âm vị /AY/ bằng chữ 'i' trong lỗi chính tả 'ciber' (cyber). Phần lớn các lỗi chính tả do trường hợp sau được xử lý bằng cách chuyển đổi lỗi chính tả dạng chữ sang dạng âm vị của nó. Trong những trường hợp này, hình thức âm vị thường khớp với hình thức viết đúng chính tả. Tuy nhiên, lỗi chính tả của biến thể trước đây ánh xạ tới các chuỗi âm vị tương tự như cách viết đúng nhưng không nhất thiết phải giống hệt nhau. Trong những trường hợp này, chúng tôi yêu cầu một số thước đo về sự tương đồng ở cấp độ âm vị, chẳng hạn, để chúng tôi có thể dự đoán rằng 'gif' có nhiều khả năng là 'give' hơn là 'gig' do âm vị /F/ giống với / hơn V/ hơn là /G/.

Trong tác phẩm này, sự giống nhau được mô hình hóa bằng hai đặc điểm đã được chứng minh là có ảnh hưởng đến nhận thức của người bản ngữ về sự giống nhau về âm vị; cụ thể là các đặc tính âm thanh và phân bố của âm vị. Sự giống nhau về âm vị được coi là một chức năng gây nhầm lẫn [13] - hai âm vị có thể được coi là giống nhau nếu một âm vị thường bị xác định nhầm là âm vị kia. Công việc trước đây của Kane và Carson-Berndsen [19] đã điều tra tính dễ nhầm lẫn của âm vị bằng cách sử dụng một hệ thống nhận dạng chưa được xác định cụ thể. Một âm vị mục tiêu đã bị loại bỏ trong quá trình huấn luyện để tại thời điểm thử nghiệm, hệ thống buộc phải chọn một âm vị thay thế - một âm vị có âm thanh tương tự với mục tiêu. Tần số mà một âm vị được xác định là âm vị khác được sử dụng làm thước đo mức độ tương tự về âm thanh của chúng. Ảnh hưởng tiềm tàng của các thuộc tính phân bố của âm vị đến sự tương tự được cảm nhận đã được chứng minh trong nghiên cứu trước đây của O'Neill và Carson- Berndsen [27]. Ở đây, những âm vị thường xuất hiện trong cùng một môi trường

[2] Trong suốt bài viết này, chúng tôi sử dụng ký hiệu ARPAbet khi đề cập đến âm vị.

(có cùng âm vị trước và sau) được cho là giống nhau hơn. Một mô hình word2vec, được huấn luyện trên Brown Corpus [11], đã được áp dụng ở cấp độ âm vị và được sử dụng để tạo ra các phần nhúng âm vị. Khoảng cách giữa các phần nhúng này, hoặc các biểu diễn vectơ, thể hiện sự tương đồng về phân bố giữa các âm vị tương ứng. Cả hai đặc tính âm thanh và phân bố đều được kết hợp để tạo thành ma trận khoảng cách giữa các âm vị trong đó các giá trị khoảng cách nhỏ hơn biểu thị nhiều âm vị giống nhau hơn.

Đáng chú ý là ma trận khoảng cách không đối xứng, tức là khoảng cách giữa một âm vị mục tiêu X được coi là Y không nhất thiết phải giống với âm vị mục tiêu Y được coi là X. Ví dụ, có khả năng là /NG/, như trong 'đi bộ' ', sẽ được phát âm là /N/; dẫn đến lỗi chính tả 'walkin'.
Tuy nhiên, ít có khả năng âm vị /N/, như trong 'happen', sẽ được phát âm là /NG/; dẫn đến lỗi chính tả 'happeng'. Ma trận khoảng cách được sử dụng trong công việc này có thể tạo ra sự khác biệt này.

Việc sửa lỗi ứng viên có thể là mục tiêu thực tế của lỗi chính tả. Cả việc sửa lỗi ứng cử viên từ thực và lỗi chính tả lần đầu tiên được chuyển đổi sang dạng âm vị tương ứng của chúng; cái trước sử dụng Từ điển phát âm CMU [38] và cái sau sử dụng công cụ chuyển từ biểu đồ sang âm vị được đào tạo về từ điển này [5]. Để xác định mức độ giống nhau giữa hai chuỗi âm vị, điểm khoảng cách được tính giữa từ sai chính tả và từ thực bằng cách sử dụng thuật toán chỉnh sửa khoảng cách có trọng số tương tự như của Wagner và Fischer [37]. Chi phí để thực hiện thao tác thay thế được xác định là khoảng cách giữa hai âm vị trên ma trận khoảng cách. Các thao tác xóa và chèn được coi là sự thay thế một âm vị bằng chuỗi trống và ngược lại. Các giá trị khoảng cách cho các thao tác này được chọn theo phương pháp phỏng đoán dựa trên tài liệu hiện có về âm vị nào thường trải qua quá trình chèn (epenthsesis) và xóa (ellision) trong lời nói [6, 10, 15, 18, 42].

Sự so sánh giữa khoảng cách chỉnh sửa cấp độ ký tự và khoảng cách chỉnh sửa trọng số cấp độ âm vị S-capades được sử dụng trong bài viết này được đưa ra trong Bảng 1. Lỗi chính tả 'sichweshan' và 'tình huống' mục tiêu từ thực của nó có tính chất cao khoảng cách chỉnh sửa. Vì vậy, các phương pháp sửa lỗi chính tả không theo ngữ âm không thể sửa được lỗi này. Tuy nhiên, sự giống nhau về mặt ngữ âm của chúng dẫn đến khoảng cách chỉnh sửa rất nhỏ khi sử dụng S-capade, do đó khiến nó dễ sửa hơn.

Bảng 1. Khoảng cách chỉnh sửa cấp ký tự so với khoảng cách chỉnh sửa âm vị của S-capade

	Cấp độ nhân vật	S-capade
Sai chính tả	s	ichweshen S IH CH W EH SH AH N
Mục tiêu từ thực s		uation S IH CH UW EY SH AH N Nó
Chỉnh sửa-Khoảng cách 7		1.1

4 Thiết lập thử nghiệm

Trong phần này chúng tôi mô tả bố trí thử nghiệm được thiết kế để kiểm tra khả năng của S-capade để sửa các lỗi sai lệch đặc biệt. Như đã nêu trước đây, S-capade phương pháp được hình dung như một thành phần của một hệ thống lớn hơn. Nó không dành cho việc sửa lỗi chính tả mà đặc biệt nhằm mục đích sửa lỗi chính tả nằm vượt ra ngoài phạm vi của các công cụ sửa lỗi chính tả thông thường. Như vậy, S-capade là không được kỳ vọng sẽ hoạt động tốt hơn các công cụ khác mà thay vào đó nhằm mục tiêu duy nhất là đạt được mục tiêu lớn hơn tỷ lệ lỗi trên các tập dữ liệu có khả năng chứa các lỗi sai lệch đặc biệt này tức là những lỗi chính tả do trẻ em thực hiện.

4.1 Bộ dữ liệu

Các phương pháp sửa lỗi chính tả cơ bản, xem Phần 4.2, và phiên âm phương pháp khoảng cách được thảo luận trong Phần 3 đã được đánh giá và so sánh bằng cách sử dụng một bộ sưu tập gồm năm bộ dữ liệu sai chính tả khác nhau. Bốn trong số các bộ dữ liệu này là có sẵn công khai và được lấy ở định dạng được xử lý trước từ Birkbeck Đại học Luân Đôn [25]. Thứ năm được mua lại thông qua sự hợp tác với một công ty giáo dục Ireland, Zeeko [43]. Chi tiết của tất cả các bộ dữ liệu có thể được nhìn thấy trong Bảng 2. Đối với mỗi tập dữ liệu, chỉ những lỗi chính tả được sửa lại mục tiêu một từ được bao gồm. Một mục tiêu sửa một từ sẽ là lỗi chính tả 'hapen' được sửa thành 'xảy ra'. Một ví dụ về điều chỉnh mục tiêu là hai các từ sẽ viết sai chính tả 'alot' được sửa thành 'alot'.

Bảng 2. Bộ dữ liệu sai chính tả

Tập dữ liệu	Lỗi chính tả	Lỗi chính tả Từ	mục tiêu được sử dụng	Công khai
Birkbeck	36.133	33.887	6.068	Đúng
Holbrook	1.791	1.562	1.177	Đúng
Wikipedia	2.455	2.230	1.909	Đúng
Aspell	531	515	437	Đúng
Zeeko	232	232	163	KHÔNG

- Birkbeck – lỗi của người bản ngữ (Anh hoặc Mỹ) [25]. Đa số lỗi của học sinh, sinh viên đại học hoặc học sinh trưởng thành biết đọc viết [24].
- Holbrook – trích đoạn bài viết của học sinh trung học Anh ở năm học áp chót của họ [25].
- Wikipedia - lỗi chính tả phổ biến của các biên tập viên (Anh hoặc Mỹ) trên Wikipedia [25]. Lỗi chính tả phổ biến là lỗi xảy ra ít nhất một lần trong một năm trên trang web [39].
- Aspell - Bộ dữ liệu kiểm tra chính tả GNU (dạng tiếng Anh). Bao gồm chung lỗi chính tả [1]
- Zeeko - bao gồm các lỗi chính tả của học sinh tiểu học Ireland [43]. Độ tuổi của người trả lời là 8-14 tuổi.

Trên năm bộ dữ liệu này có rất nhiều thông tin nhân khẩu học biết chữ; cụ thể là học sinh tiểu học, học sinh trung học cơ sở, sinh viên đại học và biên tập viên bài viết Wikipedia. Chúng tôi đưa ra giả thuyết rằng do nỗ lực đánh vần đầu tiên của trẻ dựa trên phương pháp 'phát âm', như đã thảo luận trong Phần 1, S-capade sẽ hoạt động tốt hơn trên các tập dữ liệu chứa các nỗ lực đánh vần phiên âm của trẻ , có xu hướng có khoảng cách chỉnh sửa ký tự lớn hơn. Ví dụ: trong tập dữ liệu Holbrook, 53% lỗi chính tả có khoảng cách chỉnh sửa ký tự là 1, 31% có khoảng cách chỉnh sửa ký tự là 2 và 16% có khoảng cách chỉnh sửa ký tự lớn hơn 2.
Ngược lại, trong tập dữ liệu Wikipedia, 69% lỗi chính tả có khoảng cách chỉnh sửa là 1, 28% là 2 và chỉ 3% có khoảng cách chỉnh sửa lớn hơn 2. Trong tập dữ liệu Zeeko , 91% lỗi chính tả có khoảng cách chỉnh sửa từ hai ký tự trở xuống .

4.2 Công cụ so sánh sửa lỗi chính tả thông thường

Ba công cụ sửa lỗi chính tả thông thường khác nhau được sử dụng để so sánh trong bài viết này - PySpellChecker, SymSpell và Aspell. Tất cả ba công cụ đều dựa trên giới hạn khoảng cách chỉnh sửa ký tự là 2, sử dụng từ điển tiếng Anh Anh và tạo ra cách sửa lỗi chính tả được đề xuất cũng như danh sách các cách sửa đề xuất. S-capade bị giới hạn ở khoảng cách hai lần chèn và xóa các âm vị trong một chuỗi lỗi chính tả để tạo ứng cử viên chỉnh sửa (phỏng theo SymSpellPy [14, 40]) và tra cứu. Bất kỳ từ mục tiêu nào trong bộ dữ liệu không có trong từ điển mặc định của công cụ đều được thêm thủ công để đảm bảo tính công bằng của kết quả.

- PySpellChecker - hoán vị từ được tạo thông qua chèn, xóa, thay thế và chuyển vị [26] sau đó được so sánh với các từ đã biết trong từ điển tần số [12].

- SymSpell - tạo ra các hoán vị từ để so sánh thông qua các từ sai chính tả và các từ hợp lệ trong từ điển chỉ sử dụng các từ xóa [14]. Lựa chọn dựa trên khoảng cách chỉnh sửa nhỏ nhất và từ có tần suất cao nhất [16] [41].
- Aspell - thực hiện so sánh từ trong một từ điển nhất định và sử dụng so sánh ngữ âm với các từ khác [20]. Điều này được thực hiện thông qua mã ngữ âm được điều khiển bằng bảng cho phép so sánh và gợi ý từ 'nghe giống như'. Điều này làm cho nó trở thành công cụ phù hợp nhất để so sánh với phương pháp S-capade của bài báo này.

4.3 Số liệu

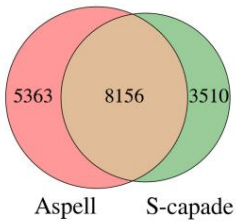
Trong Phần 5, chúng tôi so sánh độ chính xác và khả năng thu hồi của S-capade trên năm bộ dữ liệu với ba công cụ sửa lỗi chính tả thông thường là Pyspell, Symspell và Aspell. Đối với mỗi lỗi chính tả, các từ ứng cử viên thực sự được xếp hạng theo thứ tự khoảng cách và tần suất sau đó. Chúng tôi xác định độ chính xác là liệu ứng cử viên gần nhất có khớp với mục tiêu từ thực hay không và nhớ lại liệu mục tiêu từ thực có được tìm thấy trong 10 ứng cử viên gần nhất hay không. Các biểu đồ chồng chéo sửa lỗi từ, dựa trên độ chính xác, được trình bày cho từng bộ dữ liệu so sánh S-capade với trình sửa lỗi chính tả Aspell. Trong các biểu đồ này, các sửa lỗi chung giữa mỗi phương pháp và các sửa lỗi chính tả chỉ được thực hiện bằng phương pháp này hoặc phương pháp khác sẽ được hiển thị. Aspell được chọn làm đối tượng so sánh cho S-capade do nó sử dụng phương pháp sửa từ 'nghe giống', xem Phần 4.2.

5 Kết quả và thảo luận

Các kết quả của tập dữ liệu Birkbeck được trình bày trong Bảng 3. Về độ chính xác, S-capade có thể so sánh với PySpell và SymSpell, đồng thời vượt trội hơn cả về khả năng thu hồi. Aspell vượt trội hơn S-capade về độ chính xác và thu hồi. Trong số 17.029 lỗi chính tả được sửa giữa cả hai phương pháp, 48% là các sửa lỗi chính tả từ phổ biến cho cả hai phương pháp, 32% là các sửa chỉ do Aspell thực hiện và 20% là các sửa chỉ do S-capade thực hiện. Có thể xem phạm vi sửa lỗi chính tả cho tập dữ liệu giữa hai phương pháp trong Hình 1.

Phương pháp	Độ chính xác	Thu hồi
PySpell	35,3%	42,6%
SymSpell	34,74%	43,04%
capade	39,89%	66,03%
Aspell	S-	34,43%
	51,49%	

Bảng 3. Điểm hiệu chỉnh Birkbeck

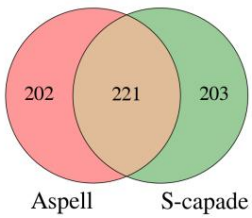


Hình 1. Birkbeck: Aspell vs S-capade

Bảng 4 hiển thị kết quả của tập dữ liệu Holbrook. So với kết quả của tập dữ liệu Birkbeck, điểm tổng thể tương tự nhau. Bạn có thể thấy kết quả thú vị nhất từ tập dữ liệu này trong Hình 2, so sánh phạm vi sửa lỗi chính tả của hai phương pháp đối với tập dữ liệu Birkbeck. Trong số 626 lỗi chính tả được sửa giữa cả hai phương pháp, 35,3% là các lỗi chính tả từ phổ biến ở cả hai phương pháp, 32,3% là các sửa chỉ do Aspell thực hiện và 32,4% là các sửa chỉ do S-capade thực hiện. Có thể thấy, S-capade có phạm vi sửa lỗi chính tả lớn hơn một chút so với Aspell. Tập dữ liệu Holbrook có nhiều khả năng bao gồm các lỗi chính tả ngữ âm do nhân khẩu học của nó, được thảo luận trong Phần 4.1, và do đó cho thấy cách tiếp cận của chúng tôi sửa các lỗi chính tả khác với các lỗi do Aspell sửa.

Phương pháp	Độ chính xác	Thu hồi
PySpell	29,32%	42,06%
SymSpell	27,46%	42,51%
Aspell	27,08%	67,93%
S-capade	27,14%	52,82%

Bảng 4. Điểm hiệu chỉnh Holbrook

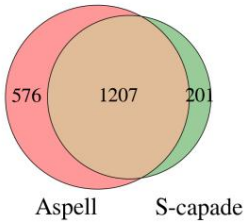


Hình 2. Holbrook: Aspell vs S-capade

Trong số năm bộ dữ liệu đang được phân tích, S-capade hoạt động kém nhất trên bộ dữ liệu Wikipedia, so với các phương pháp sửa lỗi chính tả khác. Điều này được thể hiện rõ trong Bảng 5, nơi nó đạt được điểm thấp nhất về độ chính xác và khả năng thu hồi. So với điểm trùng lặp của các bộ dữ liệu khác trong Hình 1, 2, 4 và 5, S-capade cũng có tỷ lệ sửa lỗi chính tả nhỏ nhất. Trong số 1.984 lỗi chính tả được sửa giữa cả hai phương pháp, 61% là các sửa lỗi chính tả từ phổ biến cho cả hai phương pháp, 29% là các sửa chỉ do Aspell thực hiện và 10% là các sửa chỉ do S-capade thực hiện. Như đã thảo luận trong Phần 4.1, tập dữ liệu Wikipedia được tạo thành từ các lỗi chính tả phổ biến của các biên tập viên Wikipedia. Đây thường là lỗi chính tả và như mong đợi, phương pháp ngữ âm của chúng tôi không mang lại kết quả cạnh tranh cho tập dữ liệu này.

Phương pháp	Độ chính xác	Thu hồi
PySpell	78,39%	88,48%
SymSpell	80,99%	92,11%
Aspell	79,96%	97,04%
S-capade	63,14%	77,80%

Bảng 5. Điểm hiệu chỉnh Wikipedia

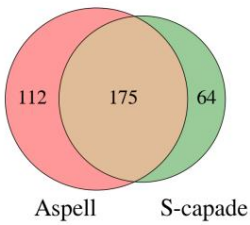


Hình 3. Wikipedia: Aspell vs S-capade

Điểm số của tập dữ liệu Aspell có thể được xem trong Bảng 6, trong đó S-capade có hiệu suất tương tự như PySpell và SymSpell. Trong số 351 lỗi chính tả được sửa giữa cả hai phương pháp, 50% là các sửa lỗi chính tả từ phổ biến cho cả hai phương pháp, 32% là các sửa chỉ do Aspell thực hiện và 18% là các sửa chỉ do S-capade thực hiện, như có thể thấy trong Hình 4 . Bộ dữ liệu Aspell tập trung vào những lần viết sai chính tả đặc biệt; những cái đi chệch khỏi mục tiêu từ thực bằng nhiều thao tác chỉnh sửa. Tuy nhiên, đây không nhất thiết là lỗi chính tả về mặt ngữ âm và do đó, S-capade hoạt động tốt trên tập dữ liệu này.

Phương pháp	Độ chính xác	Thu hồi
PySpell	49,32%	62,33%
SymSpell	53,20%	67,18%
Aspell	55,73%	85,6%
S-capade	46,41%	65,24%

Bảng 6. Điểm hiệu chỉnh Aspell



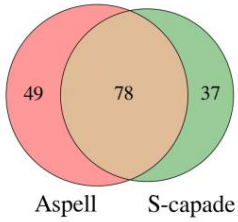
Hình 4. Aspell: Aspell vs S-capade

Điểm số của tập dữ liệu Zeeko có thể được xem trong Bảng 7. Sau Holbrook, tập dữ liệu Zeeko mang lại hiệu suất tốt thứ hai cho S-capade với

liên quan đến việc thu hồi liên quan đến các phương pháp so sánh. Hình 5 cho thấy rằng, trong số 164 lỗi chính tả được Aspell và S-capade sửa, 47,5% là các sửa thông thường, 30% là các sửa chỉ do Aspell thực hiện và 22,5% là các sửa chỉ do S-capade thực hiện. Như đã thảo luận trong Phần 4.1, 91% lỗi chính tả của tập dữ liệu Zeeko có khoảng cách chỉnh sửa từ 2 ký tự trở xuống. Chúng tôi tin rằng điều này cho thấy rằng mặc dù khoảng cách chỉnh sửa lỗi chính tả nằm trong ranh giới của các công cụ thông thường, nhưng lỗi chính tả về ngữ âm đòi hỏi một cách tiếp cận sửa lỗi khác.

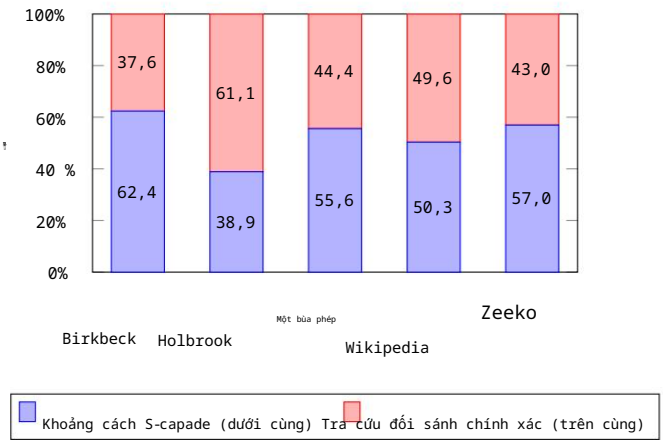
Phương pháp	Độ chính xác	Thu hồi
PySpell	56,90%	72,41%
SymSpell	54,74%	70,69%
Aspell	54,74%	86,64%
S-capade	49,57%	76,29%

Bảng 7. Điểm hiệu chỉnh Zeeko



Hình 5. Zeeko: Aspell vs S-capade

Từ điển phát âm CMU được sử dụng theo phương pháp S-capade để tra cứu chuỗi âm vị theo từ khi sửa lỗi chính tả của từ. Hình 6 hiển thị phân tích chi tiết các chỉnh sửa được thực hiện bởi S-capade, hiển thị sự phân chia giữa tra cứu khớp chính xác trong từ điển CMU bằng cách sử dụng chuỗi âm vị, trong trường hợp đó khoảng cách chỉnh sửa bằng 0 hoặc sử dụng S-capade để tính toán âm vị khoảng cách giữa lỗi chính tả và mục tiêu từ thực.
Có thể thấy rằng đối với bốn trong số năm bộ dữ liệu, phương pháp khoảng cách của S-capade chiếm hơn 50% số hiệu chỉnh ứng viên từ thực được thực hiện.



Hình 6. Tính toán khoảng cách S-capade và đối sánh từ điển chính xác

Cách tiếp cận của S-capade đã dẫn đến việc sửa chữa từ thú vị của những từ có cách đánh vần phiên âm với khoảng cách chỉnh sửa ký tự lớn mà thông thường công cụ sửa lỗi chính tả không thể sửa được. Bảng 8 cho thấy một số lựa chọn lỗi chính tả từ các bộ dữ liệu mà S-capade đã sửa và cách viết truyền thống khoảng cách chỉnh sửa ký tự so với khoảng cách chỉnh sửa âm vị do chúng tôi tạo ra tiếp cận.

Bảng 8. S-capade sửa lỗi từ thú vị

Mục tiêu	Sai chính tả	Khoảng cách ký tự	Khoảng cách S-capade
nhất thiết phải là triết học		[illegible]	1,62
tập trung vào tình		7	0,54
huống sichweshen		7	1.10
thuốc	bên ngoài	6	0,46
lắc đủ	người hầu	6	0,93
thủ tục prosiegeur huýt		5	0,59
sáo hội đồng Wisheld		4	1.14
cousall thực sự achuly		3	1,00
		3	1,28

6. Kết luận

Trong bài báo này, chúng tôi đã trình bày một cách tiếp cận dựa trên khoảng cách âm vị để đánh vần sửa lỗi có khả năng xử lý lỗi chính tả về mặt ngữ âm mà thông thường công cụ không thể sửa được. Sự sáng tạo trong việc đánh vần của trẻ đã được chứng minh là tạo ra lỗi chính tả về mặt ngữ âm và có nhiều sai lệch so với mục tiêu từ thực tế. Như vậy, chúng tôi thấy hiệu suất kém hơn của cách viết chính tả thông thường công cụ sửa lỗi trên các tập dữ liệu đặc biệt bao gồm các lỗi chính tả của trẻ em. Phương pháp được mô tả trong bài viết này được chứng minh là có thể sửa được một phần đáng kể lỗi chính tả trong các bộ dữ liệu này mà một trong những bộ dữ liệu tiếng Anh có hiệu suất hàng đầu người kiểm tra chính tả, Aspell, không thể.

Cách tiếp cận dựa trên khoảng cách âm vị được hình dung như một thành phần trong một hệ thống sửa lỗi chính tả phụ thuộc vào ngữ cảnh. Công việc trong tương lai sẽ xem xét kết hợp phương pháp này thành một trình kiểm tra chính tả có khả năng xử lý cả kiểu chữ và ngữ âm lỗi chính tả và chọn mục tiêu từ thực tế chính xác từ danh sách các ứng cử viên dựa trên ngữ cảnh của lỗi chính tả. Các kế hoạch cải tiến tiếp theo bao gồm nghiên cứu ảnh hưởng của trọng âm đến lỗi chính tả được tạo ra và lợi ích tiềm năng của hệ thống đặc biệt về giọng nói đối với độ chính xác của việc sửa lỗi chính tả.

Lời cảm ơn Công việc này được hỗ trợ bởi sự hỗ trợ tài chính của Quỹ Khoa học Ireland cấp 13/RC/2094 cho Lero - Nghiên cứu SFI Trung tâm Phần mềm (www.lero.ie). Trung tâm nội dung số ADAPT

Công nghệ (www.adaptcentre.ie) được tài trợ theo Chương trình Trung tâm Nghiên cứu SFI (Cấp 13/RC/2106). Các tác giả xin cảm ơn nhóm Zeeko (https://zeeko.ie/) đã hỗ trợ nghiên cứu của họ.

Người giới thiệu

1. Atkinson, K.: Dữ liệu kiểm tra chính tả Aspell (2002), http://aspell.net/test/cur-all/batch0.tab, truy cập lần cuối vào ngày 19 tháng 5 năm 2020
2. Aw, A., Zhang, M., Xiao, J., Su, J.: Mô hình thống kê dựa trên cụm từ để chuẩn hóa văn bản sms. Trong: COLING/ACL. trang 33–40 (2006)
3. Brill, E., Moore, RC: Một mô hình lỗi cải tiến để sửa lỗi chính tả kênh bị nhiễu. Trong: ACL. trang 286–293 (2000)
4. Church, KW, Gale, WA: Chấm điểm xác suất sửa lỗi chính tả. Thống kê và Máy tính 1(2), 93–103 (1991)
5. CMUSphinx: Công cụ đồ thị thành âm vị dựa trên việc học theo trình tự (2016), https://github.com/cmusphinx/g2p-seq2seq 6. Collins, B., Mees, IM: Ngữ âm học và âm vị học thực tế : Sách tham khảo dành cho sinh viên. Định tuyến (2013)
7. Daffern, T., Critten, S.: Quan điểm của học sinh và giáo viên về chính tả. Tạp chí Ngôn ngữ và Đọc viết Úc 42(1), 40–57 (2019)
8. Damerau, FJ: Một kỹ thuật phát hiện và sửa lỗi chính tả trên máy tính. Cộng đồng. ACM 7(3), 171–176 (tháng 3 năm 1964)
9. Fisher, WM: Chức năng thống kê chuyển văn bản sang điện thoại sử dụng ngram và quy tắc. Trong: ICASSP. tập. 2, trang 649–652 (1999)
10. Fourakis, M., Port, R.: Dừng câu chữ trong tiếng Anh. Tạp chí Ngữ âm học 14(2), 197–221 (1986)
11. Francis, WN, Kucera, H.: Cẩm nang ngữ liệu màu nâu (1979)
12. FrequencyWords: Trình tạo danh sách từ tần số (2020), https://github.com/hermitdave/FrequencyWords, truy cập lần cuối vào ngày 21 tháng 5 năm 2020
13. Gallagher, G., Graff, P.: Vai trò của sự tương đồng trong âm vị học. Ngôn ngữ 2(122), 107–111 (2012)
14. Garbe, W.: Symspell (2020), https://github.com/wolfgarbe/symspell, accham dứt vào ngày 21 tháng 5 năm 2020
15. Gimson, AC, Ramsaran, S.: Giới thiệu về cách phát âm tiếng Anh, tập. 4. Edward Arnold Luân Đôn (1970)
16. Google: Trình xem ngram sách của Google (2012), http://storage.googleapis.com/ books/ngrams/books/datasetsv2.html, truy cập lần cuối vào ngày 21 tháng 5 năm 2020 17. Hodge, VJ, Austin, J.: Đánh giá của người kiểm tra chính tả ngữ âm (2001)
18. Itˆo, J.: Một lý thuyết thi vị về câu kết. Ngôn ngữ tự nhiên & Lý thuyết ngôn ngữ học 7(2), 217–259 (1989)
19. Kane, M., Carson-Berndsen, J.: Tăng cường sự nhầm lẫn của điện thoại dựa trên dữ liệu bằng cách sử dụng nhận dạng hạn chế. Trong: INTERSPEECH. trang 3693–3697 (2016)
20. Kevin Atkinson, GA: Aspell hoạt động như thế nào) (2004), http://aspell.net/0.50-doc/man-html/8_How.html, truy cập lần cuối vào ngày 21 tháng 5 năm 2020
21. Khoury, R.: Chuẩn hóa vi văn bản bằng cách sử dụng từ có lẽ giống về mặt ngữ âm khám phá. Trong: WiMob. trang 384–391 (2015)
22. Kukich, K.: Kỹ thuật tự động sửa lỗi chữ trong văn bản. Khảo sát máy tính Acm (CSUR) 24(4), 377–439 (1992)

23. de Mendon¸ca Almeida, GA, Avan¸co, L., Duran, MS, Fonseca, ER, Nunes, MdGV, Alu'ısio, SM: Đánh giá cách đánh vần ngữ âm cho nội dung do người dùng tạo ra bằng tiếng Bồ Đào Nha ở Brazil. Trong: Hội nghị quốc tế về xử lý tính toán của tiếng Bồ Đào Nha. trang 361–373 (2016)

24. Mitton, R.: Tập dữ liệu lỗi chính tả Birkbeck (1980), https://ota.bodleian.ox.ac. uk/repository/xmlui/handle/20.500.12024/0643, truy cập lần cuối vào ngày 19 tháng 5 năm 2020 25. Mitton, R.: Corpora of badspellings for download (2007), https://www.dcs.bbk. ac.uk/~ROGER/corpora.html, truy cập lần cuối vào ngày 19 tháng 5 năm 2020
26. Norvig, P.: Pyspellchecker (2020), https://pypi.org/project/pyspellchecker/, truy cập lần cuối vào ngày 21 tháng 5 năm 2020
27. O'Neill, E., Carson-Berndsen, J.: Ảnh hưởng của sự phân bổ âm vị đến sự tương đồng về nhận thức trong tiếng Anh. INTERSPEECH trang 1941–1945 (2019)
28. Philips, L.: Thuật toán tìm kiếm siêu âm kép. Tạp chí người dùng C/C++ 18(6), 38–43 (2000)
29. Đọc, C.: Đánh vần sáng tạo của trẻ. Lộ Trình (2018)
30. Russell, R., Odell, M.: Soundex. Bằng sáng chế Hoa Kỳ 1.261.167 (1918)
31. Silfverberg, M., Kauppinen, P., Lind'en, K.: Sửa lỗi chính tả dựa trên dữ liệu bằng các phương pháp trạng thái hữu hạn có trọng số. Trong: Hội thảo SIGFSM về NLP thống kê và Automata có trọng số. trang 51–59 (2016)
32. St¨uker, S., Fay, J., Berkling, K.: Hướng tới việc sửa lỗi chính tả ngữ âm phụ thuộc vào ngữ cảnh trong văn bản được sáng tác tự do của trẻ em nhằm mục đích chẩn đoán và sư phạm. Trong: INTERSPEECH (2011)
33. Torgerson, C., Brooks, G., Hall, J.: Đánh giá có hệ thống các tài liệu nghiên cứu về việc sử dụng ngữ âm trong việc dạy đọc và đánh vần. Ấn phẩm DfES Nottingham (2006)

34. Toutanova, K., Moore, RC: Mô hình phát âm để cải thiện việc sửa lỗi chính tả . Trong: Kỷ yếu Hội nghị thường niên lần thứ 40 về Hiệp hội Ngôn ngữ học tính toán. trang 144–151 (2002)

35. Đại học Cao đẳng Dublin: Kho lưu trữ S-capade github (2020), https://github. com/ucd-csl/Scapade, truy cập lần cuối vào ngày 15 tháng 7 năm 2020
36. Veronis, J.: Sửa lỗi ghi âm bằng máy tính. Máy tính và Nhân văn 22(1), 43–56 (1988)
37. Wagner, RA, Fischer, MJ: Bài toán hiệu chỉnh chuỗi này sang chuỗi khác. JACM 21(1), 168–173 (1974)
38. Weide, RL: Từ điển phát âm CMU (1998), http://www.speech.cs. cmu.edu/cgi-bin/cmudict
39. Wikipedia: Wikipedia:danh sách các lỗi chính tả phổ biến (2020), https://en.wikipedia. org/wiki/Wikipedia:Lists_of_common_misspellings, truy cập lần cuối vào ngày 19 tháng 5 năm 2020 40. Wolf Garbe, S.: Symspellpy (2020), https://github.com/mammothb/symspellpy, truy cập lần cuối vào ngày 21 tháng 5 năm 2020
41. Wordlist, A.: Scowl (danh sách từ hướng tới trình kiểm tra chính tả) (2019), http://wordlist. aspell.net, truy cập lần cuối vào ngày 21 tháng 5 năm 2020 42. Yip, M.: nguyên âm tiếng Anh. Ngôn ngữ tự nhiên & Lý thuyết ngôn ngữ trang. 463–484 (1987)
43. Zeeko: Zeeko: Phản hồi khảo sát văn bản miễn phí (2020), https://zeeko.ie, truy cập lần cuối 19 tháng 5 năm 2020