BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐỀ CƯƠNG LUẬN VĂN THẠC SỸ

Chuyên ngành: Công Nghệ Thông Tin

Mã ngành: 60480201

PHÁT HIỆN VÀ SỬA LỖI CHÍNH TẢ TIẾNG VIỆT

GVHD : TS. LÊ THỊ NGỌC THƠ
HVTH : TRẦN VĂN BẮC
MSHV : 2041860001
LỚP : 20SCT11

TP. HCM, tháng 10/2021

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. HỒ CHÍ MINH, ngày … tháng … năm 2021

Giáo viên hướng dẫn

TS. LÊ THỊ NGỌC THƠ

NHẬN XÉT CỦA HỘI ĐỒNG XÉT DUYỆT

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. HỒ CHÍ MINH, ngày … tháng … năm 2021

Hội đồng xét duyệt

MỤC LỤC

1 Giới thiệu

1.1 Đặt vấn đề

Như chúng ta đã biết trong thời đại cách mạng công nghiệp 4.0 hiện nay, lĩnh vực trí tuệ nhân tạo nói chung và xử lý ngôn ngữ tự nhiên nói riêng có vai trò đặc biệt quan trọng và tác động to lớn đến đời sống kinh tế, xã hội. Những nghiên cứu về lĩnh vực này đã được ứng dụng triển khai trong thực tế và từng bước cải thiện đời sống chúng ta ngày càng tốt hơn.

Tính năng phát hiện và sửa lỗi chính tả có mặt trong hầu hết các ứng dụng liên quan đến văn bản, từ máy tính cho đến các thiết bị di động. Khi nhập một từ không chính xác, hệ thống sẽ phát hiện lỗi, có thể đó là lỗi chính tả hoặc từ không phù hợp ngữ cảnh. Sau đó, hệ thống tự động sửa bằng một từ khác hoặc đề xuất một danh sách các từ có thể thay thế.

1.2 Lý do chọn đề tài

Ngày nay, tính năng phát hiện và sửa lỗi chính tả có mặt trong hầu hết các ứng dụng liên quan đến văn bản, từ máy tính cho đến các thiết bị di động. Khi nhập một từ không chính xác, hệ thống sẽ phát hiện lỗi, có thể đó là lỗi chính tả hoặc từ không phù hợp ngữ cảnh. Sau đó, hệ thống tự động sửa bằng một từ khác hoặc đề xuất một danh sách các từ có thể thay thế.

Bài toán phát hiện và sữa lỗi chính tả trước đây đã có nhiều đề xuất giải quyết bằng các phương pháp dựa vào bản chất ngôn ngữ. Tuy nhiên, do sự phức tạp của ngôn ngữ tự nhiên, sự nhập nhằng nghĩa của từ và cụm từ, sự phụ thuộc cú pháp và ngữ nghĩa của các từ vào ngữ cảnh, nên các phương pháp nêu trên còn hạn chế về tính chính xác. Đối với tiếng Việt lỗi chính tả càng phức tạp hơn do cách đọc một số chữ, âm tiết gần giống nhau, ví dụ sai dấu, chữ x hay s, ch hay tr, l hay n. Ngôn ngữ nói ở các vùng miền khác nhau làm phát sinh nhiều lỗi chính tả trong văn viết. Bảng 1 cho thấy sự đa dạng các loại lỗi khác nhau trong tiếng Việt về từ vựng. Do đó cần phải có một mô hình phát hiện và sửa lỗi chính tả tốt hơn các phương pháp dựa trên ngôn ngữ.

Bảng 1: Một số lỗi chính tả tiếng Việt [9]

	Lỗi	**Sửa**
LỖI ĐÁNH MÁY		
Lỗi trùng ký tự	chuung	chung
Lỗi do dùng kiểu gõ VNI	chu7ng	chúng
Bỏ dấu không đúng chỗ	qũy	quỹ
Thêm dấu	nien	niên
Thêm ký tự	kông	không
LỖI CHÍNH TẢ		
Giọng Bắc		
Bắt đầu với ch/tr	chăn chối	trăn trối
Bắt đầu với x/s	năng xuất	năng suất
Bắt đầu với gi/d	giang giở	dang dở
Giọng Nam và Trung		
Bắt đầu với d/gi/v	dang danh	vang danh
Bắt đầu với qu/h/u	quy quyền	uy quyền
Phụ âm cuối c/t	hờn mác	hờn mát
Phụ âm cuối n/ng	khốn cùn	khốn cùng
Phụ âm cuối i/y	trình bài	trình bày
Vần uo/u	buồi nguồi	bùi ngùi
Vần iê/uyê	tiệt tình	tuyệt tình
Vần iê/i	đương kiêm	đương kim
Dấu hỏi/ngã	vui vẽ	vui vẻ
Vần ă/â	chăm bẳm	châm bẩm

➔ Từ cá nhân nhận thấy rằng hầu hết các bài nghiên cứu về phát hiện và sửa lỗi chính hầu như chưa xem xét về trường các trường hợp lỗi câu có chứa từ ngoại lai. Ví dụ như sau:

Lỗi chính tả	**Sửa lại**
em nốt lại nha.	em note lại nha.
Tạo linh gu gồ mít chưa?	Tạo link google meeting chưa?
bạn đã tét cô vít chưa?	bạn đã test Covid chưa?
tài khoản da lô và phây búc mình bị khóa rồi.	tài khoản Zalo với Facebook mình bị khóa rồi.
Một số câu là câu hỏi [?] em ăn cơm chưa? | em ăn cơm trưa.	

Từ những vấn đề nêu trên, em chọn đề tài: "**phát hiện và sửa lỗi chính tả tiếng việt**" để làm luận văn tốt nghiệp, nhằm áp dụng kỹ thuật dịch máy vào bài toán bắt lỗi chính tả (dịch từ câu sai chính tả sang câu đúng chính tả) và phát triển để ứng dụng thử nghiệm cho tiếng Việt cho những trường hợp lỗi này.

2 Mục tiêu, nội dung và phương pháp nghiên cứu

2.1 Mục tiêu của đề tài

Phát hiện ra được lỗi chính tả trong các câu hay những văn bản tiếng việt rồi từ đó sửa những lỗi chính tả đó để kết quả trả về là những câu hay văn bản đúng chính tả. Hay nói một cách cụ thể là đầu vào là những câu hay văn bản có thể bị sai chính tả, thông qua kĩ thuật xử lý đầu ra là một câu, văn bản đúng chính tả.

2.2 Nội dung nghiên cứu

Trong khuôn khổ của luận văn, em giới hạn nghiên cứu hai phương pháp phát hiện và sửa các loại lỗi chính tả cơ bản trong tiếng Việt theo hướng học sâu (Deep Learning-DL) sử dụng mô hình Sequence-to-Sequence kết hợp kỹ thuật Attention và Trie rồi đứa ra nhận xét đánh giá. Dữ liệu huấn luyện và kiểm tra cho mô hình được tự động sinh ra từ văn bản đúng chính tả. Đầu vào hệ thống là văn bản tiếng Việt chuẩn unicode có thể sai chính tả ở nhiều vị trí, đầu ra là văn bản đúng chính tả hoặc danh sách các câu đúng chính tả đề xuất.

2.3 Phương pháp luận và phương pháp nghiên cứu

Trên cơ sở những thành công của các công trình nghiên cứu, áp dụng kỹ thuật mạng nơ-ron vào xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP), luận văn nghiên cứu đề xuất mô hình học sâu ứng dụng cho phát hiện và sửa lỗi chính tả. Luận văn tập trung vào nghiên cứu các mô hình Sequenceto-Sequence kết hợp kỹ thuật Attention và kĩ thuật Trie để so sánh đánh giá giữa 2 phương pháp đưa ra kết quả tối ưu và giải quyết bài toán trên, nhằm áp dụng và phát triển để ứng dụng thử nghiệm cho tiếng Việt.

3 Tổng quan lĩnh vực nghiên cứu

Dựa trên các nghiên cứu về học sâu, đề tài đã ứng dụng và kết hợp các phương pháp này để cho ra mô hình giúp bắt một số lỗi chính tả mức đơn giản cho tiếng Việt. Đây

là hướng nghiên cứu khá mới hiện nay và có những thuận lợi, khó khăn riêng so với phương pháp truyền thống.

Những thuận lợi có thể kể tới như việc rút trích đặc trưng không còn làm một cách thủ công mà đã tích hợp hoàn toàn bằng mô hình học máy giúp giảm đi sự phức tạp và tăng độ hiệu quả. Đặc biệt với các nghiên cứu gần đây về mô hình sequence-to-sequence, kỹ thuật Attention[3], kiến trúc Transformer [10] và mô hình BERT[11] cho chất lượng vượt trội so với những phương pháp trước kia. Hơn nữa các phương pháp học sâu phù hợp với các phần cứng tính toán song song như GPU giúp tăng tốc trong quá trình xử lý. Bên cạnh đó, vẫn tồn tại những khó khăn, thách thức lớn nhất là vấn đề dữ liệu. Các phương pháp học sâu đòi hỏi một lượng dữ liệu khá lớn để có thể làm việc tốt.

4 Tiến độ thực hiện đề tài

Tháng thứ **Dự kiến nội dung thực hiện**	**1**	**2**	**3**	**4**	**5**	**6**
- Tìm hiểu phương pháp phát hiện lỗi chính tả - Tìm hiểu phương pháp sửa lỗi chính tả -Thu Thập dữ liệu						
-Tìm hiểu kiến trúc Transformer -Tìm hiểu mô hình BERT						
-Tìm hiểu Phương pháp Attention , Trie và N-grams						
-Đưa ra mô hình phù hợp với bài toán. -Thực nghiệm triển khai ứng dụng. Song song là đưa ra những đánh giá.						
Hoàn thiện luận văn.						

5 Bố cục dự kiến của luận văn

*Luận văn dự kiến bao gồm 6 chương

*Cấu trúc của luận văn:

6.3 Hướng phát triển

5 Tài liệu tham khảo

[1] Andrew McCallum, Kedar Bellare, and Fernando Pereira (2012), "A conditional random field for discriminatively-trained finite-state string edit distance", arXiv preprint, arXiv:1207.1406.

[2] Denny Britz (2015), "Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs", http://www.wildml.com/2015/09/recurrent-neural-networkstutorial-part-1-introduction-to-rnns/

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio (2014), "Neural machine translation by jointly learning to align and translate", arXiv preprint, arXiv:1409.0473.

[4] Hochreiter S., Schmidhuber J. (1997), "Long Short-Term Memory", Neural Computation 9(8), 1735-1780.34

[5] Julian R. Ullmann (1977), "A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words", The Computer Journal 20(2), 141–147.

[6] Krenker A., Bester J., Kos A. (2011), "Introduction to the Artificial Neural Networks", Artificial Neural Networks - Methodological Advances and Biomedical Applications, ISBN: 978-953-307-243-2, InTech.

[7] Leon Davidson (1962), "Retrieval of misspelled names in an airlines passenger record system", Communications of the ACM 5(3), 169–171.

[8] Minh-Thang Luong, Hieu Pham, and Christopher D Manning (2015) "Effective approaches to attention-based neural machine translation", arXiv preprint, arXiv:1508.04025.

[9] VSpell TDK Development (2019), "Một số lỗi chính tả Tiếng Việt", http://vspell.com/home/specifications.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin (2017), “Attention Is All You Need”, arXiv preprint, arXiv: 1706.03762.36

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019), “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv preprint, arXiv: 1810.04805.