Đồ án tốt nghiệp

Đề tài: Nghiên cứu và xây dựng mô hình học sâu trong phát hiện và sửa lỗi chính tả trong tiếng Việt (Deep Learning)

Giảng viên hướng dẫn: T.S Roãn Thị Ngân

Đỗ Đức Tiến - 1660364

Khoa học máy tính - Đại học Xây dựng Hà Nội

10 Tháng 1 năm 2024

Nội dung

Nội dung

Giới thiệu

- Bài toán sửa lỗi chính tả là một bài toán khá phức tạp, được không ít đơn vị nghiên cứu, phát triển và nó có tính ứng dụng cao, đặc biệt là trong các ứng dụng soạn thảo hay nhận dạng văn bản.
- Chương trình sửa lỗi chính tả cần có hai chức năng chính, cơ bản là chỉ ra lỗi sai và đưa ra gợi ý sửa lỗi. Tuy nhiên, các chức năng kiểm lỗi chính tả được tích hợp nhiều trong ứng dụng và phần mềm soạn thảo tiếng Việt hiện nay (Vietkey, Unikey, ...) không đưa ra gợi ý cho người dùng lựa chọn.

Figure: Một số phần mềm

Nội dung

Các lỗi tiếng Việt

Lỗi chính tả chính

- Lỗi nhận thức khi có từ đồng âm, gần âm đặc biệt là các âm người Việt bị nhầm như ch-tr, l-n, r-d-gi như các từ ví dụ như: xương - sương, làm - nàm, sẻ - xẻ, ...
- Lỗi viết tắt từ teencode như các từ dưới đây:
  - Từ "không" thành các từ "khum, ko, hok, kh,..."
  - Từ "biết rồi" thành từ "bít rùi"
  - Từ "thôi" thành các từ "hoy, thoy, thui"
  - Từ "ừ" thành các từ "uhm, uh, ừa, ỏ"

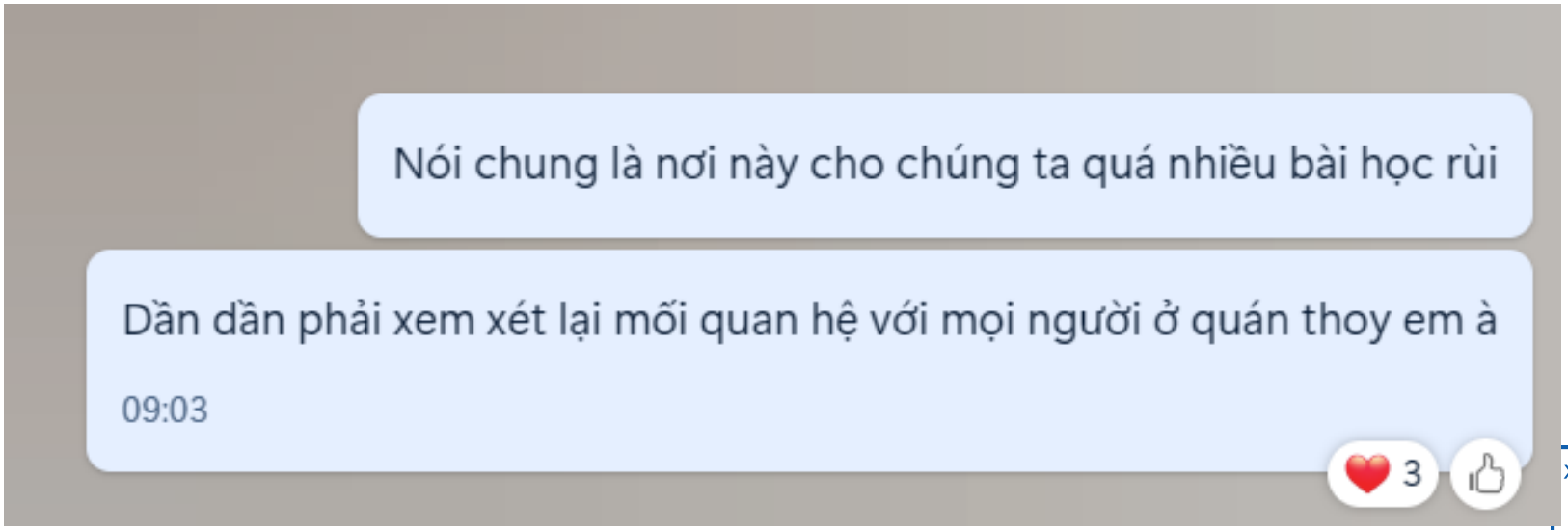


Figure: Ví dụ lỗi chính tả tiếng lóng

Các lỗi tiếng Việt

Nguyên nhân gây lỗi chính tả

- Sự bất cẩn của người viết
- Sự ảnh hưởng của ngôn ngữ mạng
- Sự ảnh hưởng của ngôn ngữ địa phương
- Không cập nhật quy tắc chính tả hiện hành

Nội dung

Mô tả bài toán

Mô tả

- **Input:**
  - Tập chuỗi đầu vào là $X = \{x_1, x_2, ..., x_n\}$ với từ sai chính tả ở một vị trí bất kì trong các giá trị từ x_1 đến x_n (được gọi là x_i)
- **Output:**
  - Tập chuỗi đầu ra là $Y = \{y_1, y_2, ..., y_n\}$ với từ sai chính tả ở vị trí bất kỳ ở vị trí tập x_1 đến x_n đã được sửa lỗi chính tả (được gọi là y_i)
- Với bài toán này với mỗi từ sai ở vị trí x_i thì cần phải sửa đúng ở vị trí y_i tức là ta phải xây một hàm $f : X \rightarrow Y$ thỏa mãn $f(x_i) = y_i$

Định hướng đề tài

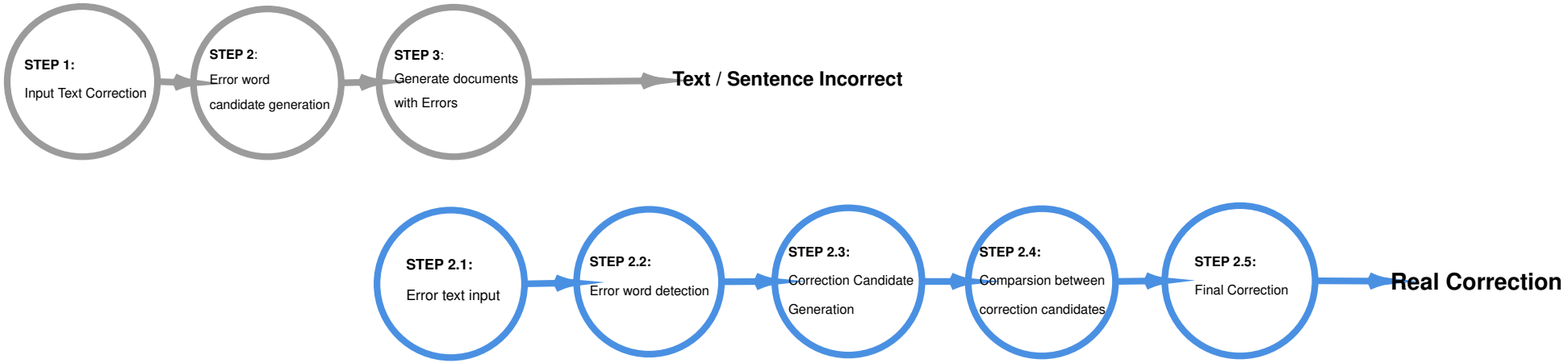
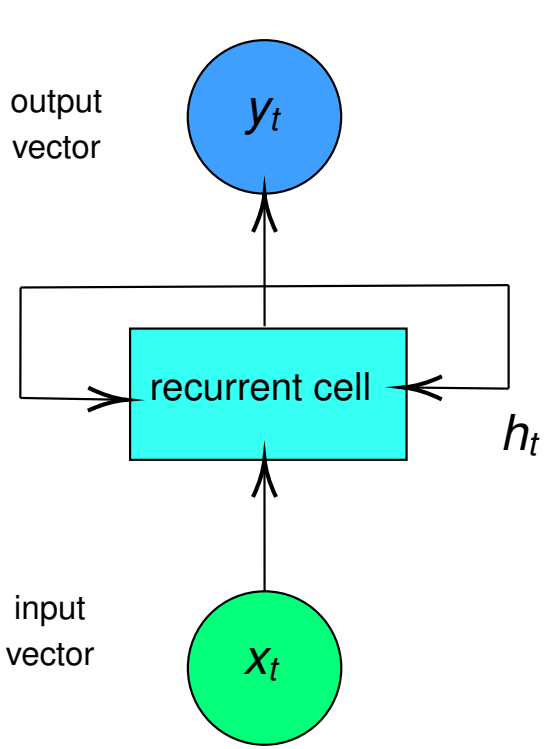
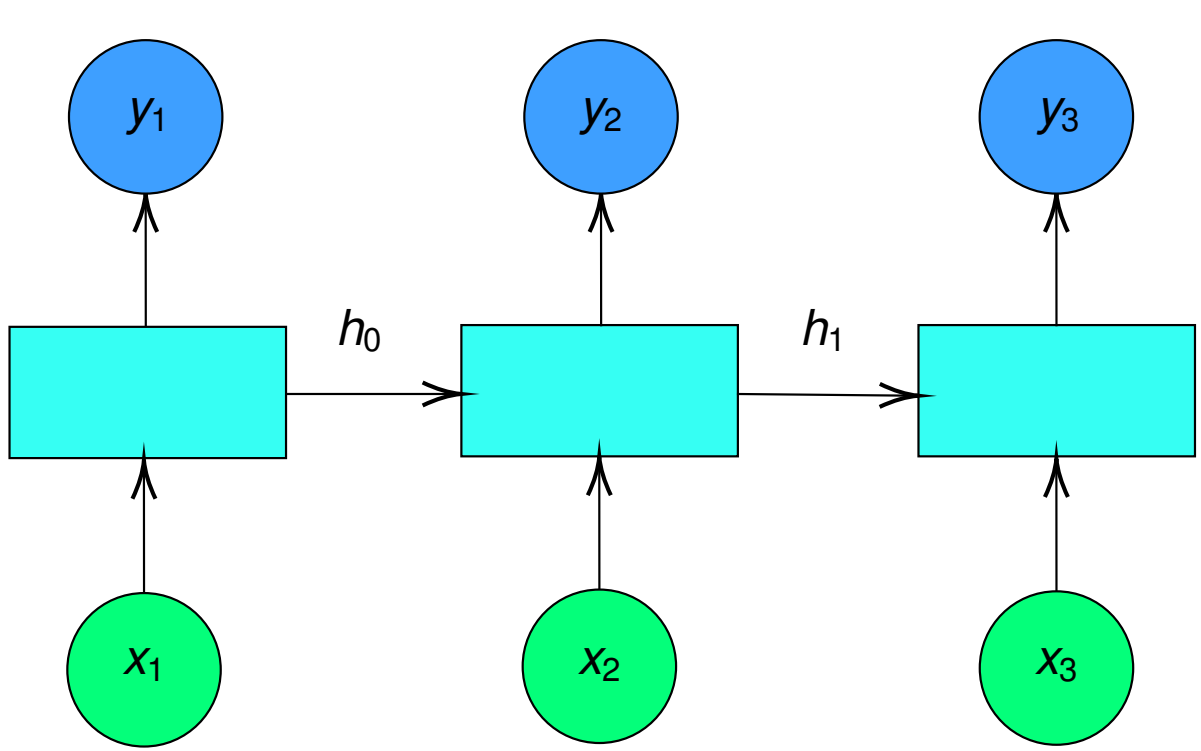


Figure: Quy trình giải quyết bài toán

Nội dung

Nhắc lại mô hình RNN





$$y_t = f(x_t,\ h_{t-1})$$

output input past memory

Các dạng của mô hình RNN

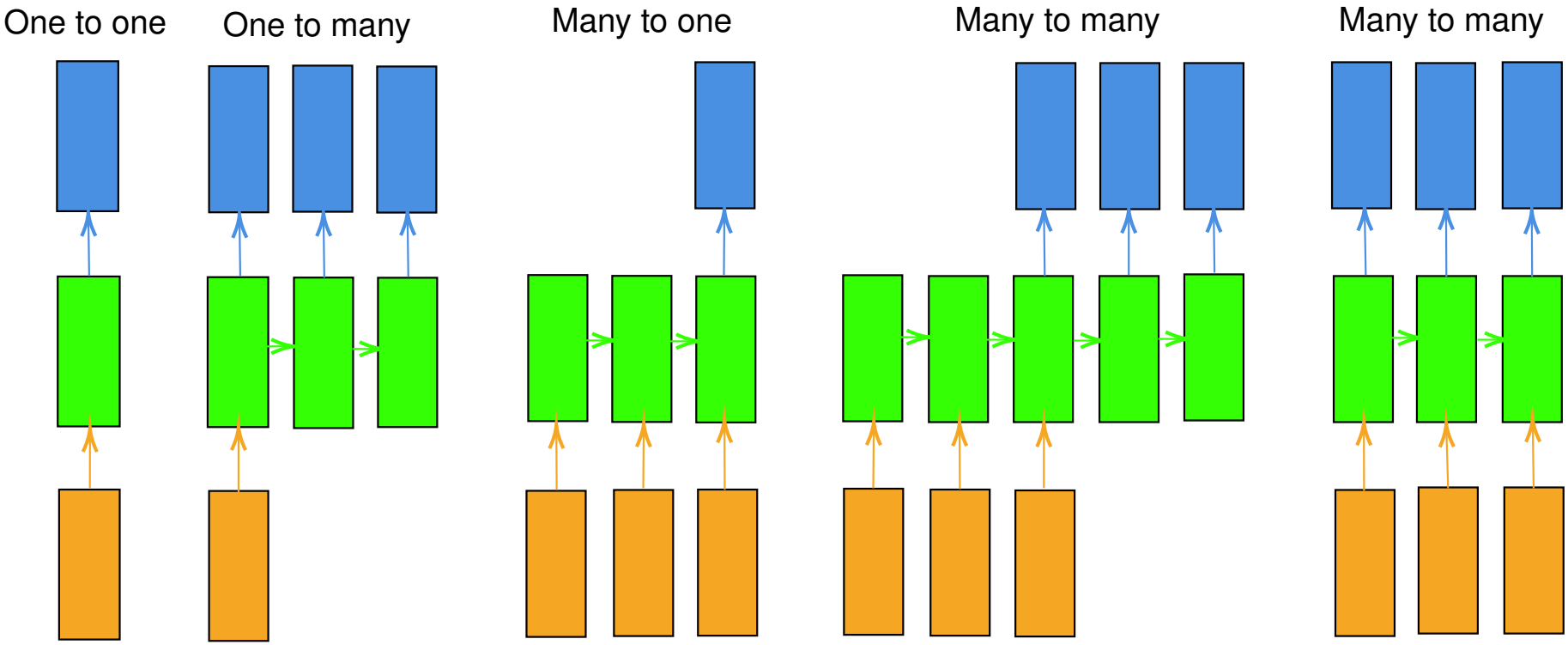


Figure: Minh họa mô hình được sử dụng trong mạng RNN

Nhắc lại mô hình LSTM

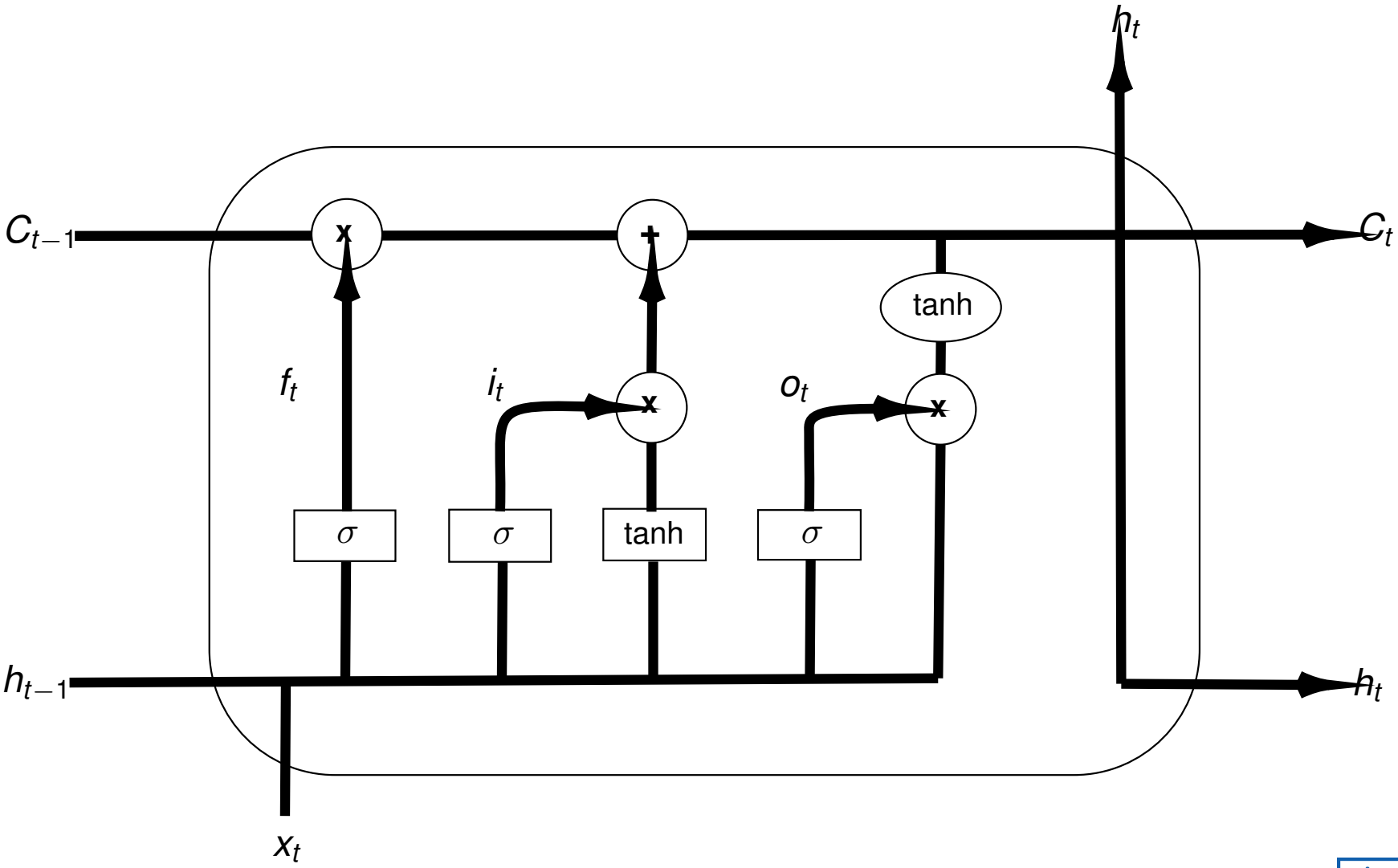


Figure: Mô hình LSTM

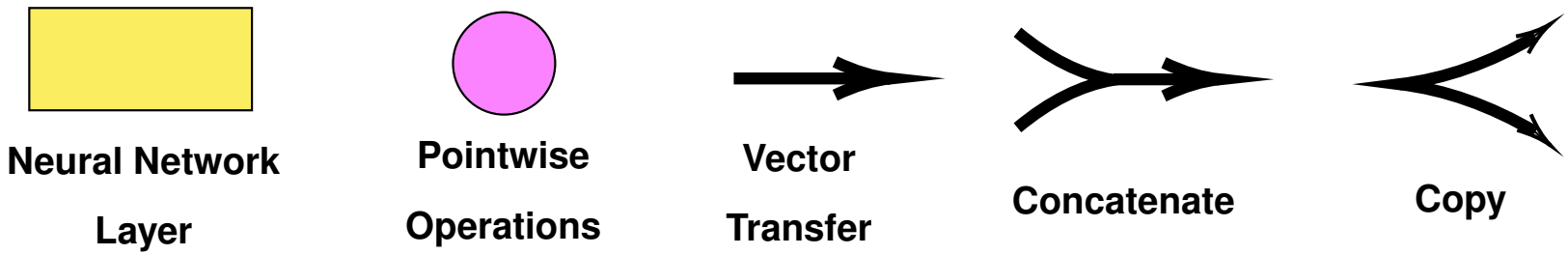
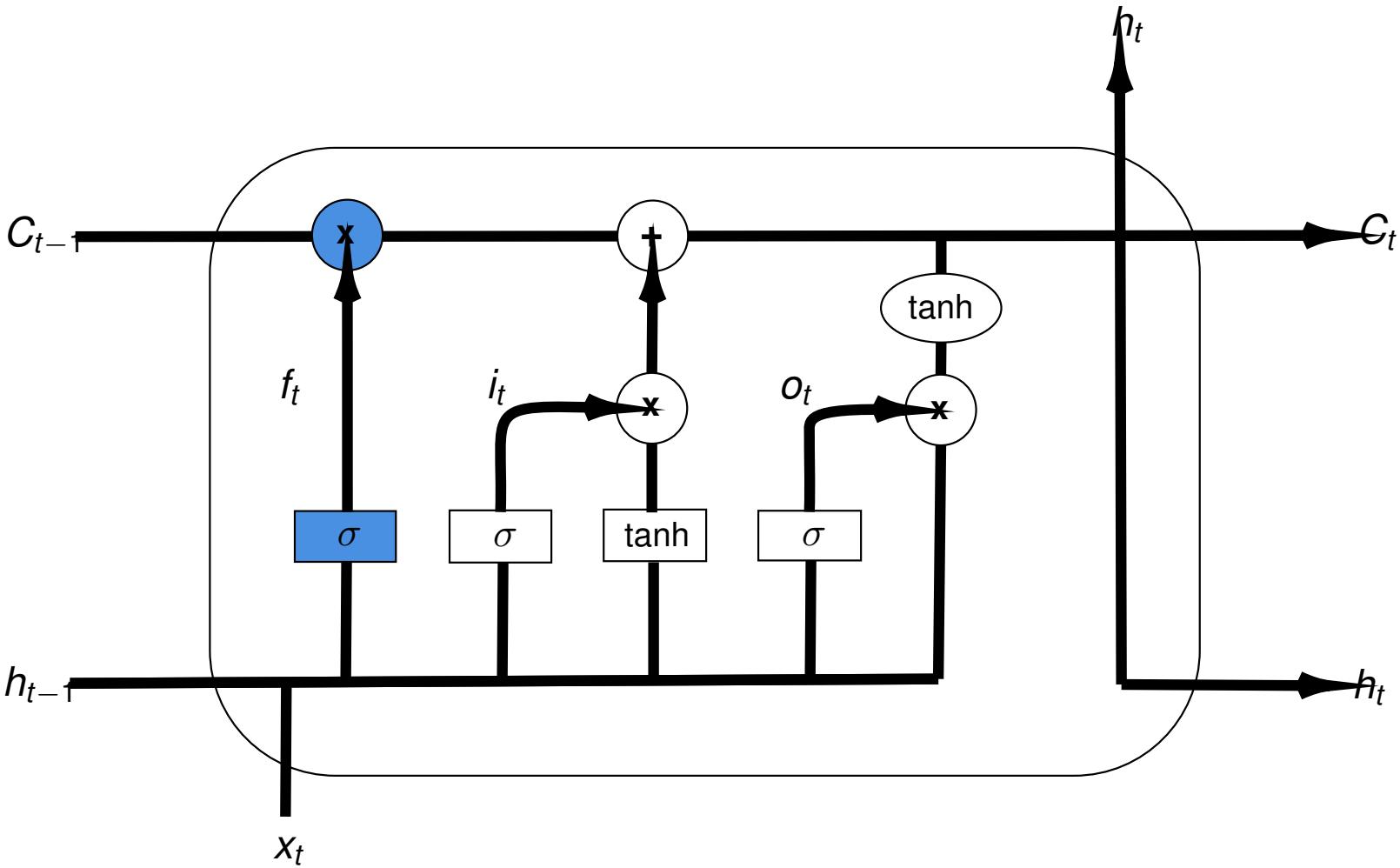


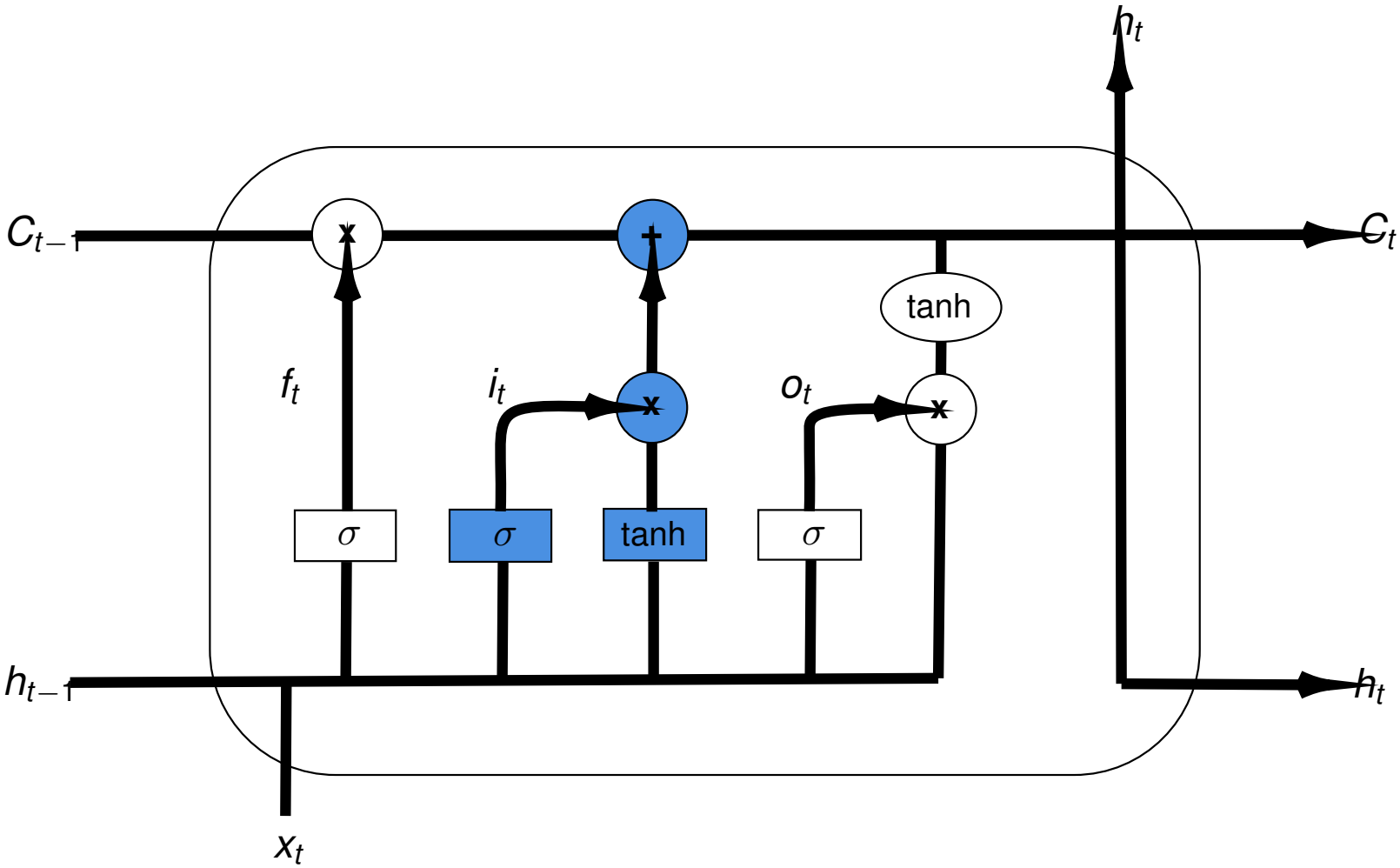
Figure: Diễn giải các kí hiệu trong đồ thị mạng nơ ron (áp dụng chung cho toàn bộ bài)

LSTM - Forget Gate



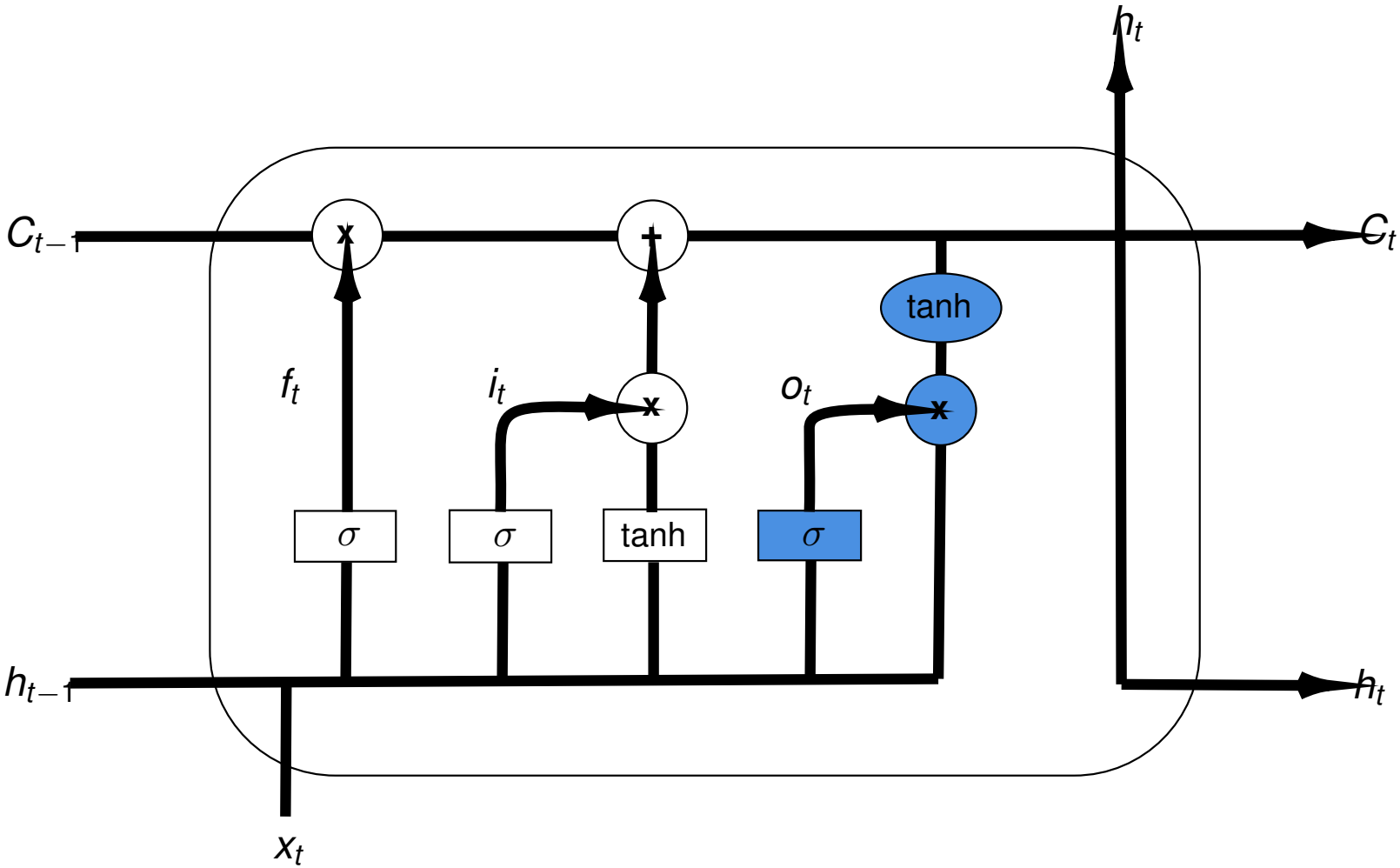
$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right)$$

LSTM - Input/Update Gate



$$i_t = \sigma\left(\mathrm{W}_i\left[h_{t-1}, x_t\right] + b_i\right)$$
$$\widetilde{C}_t = \tanh\left(\mathrm{W}_c\left[h_{t-1}, x_t\right] + b_c\right)$$

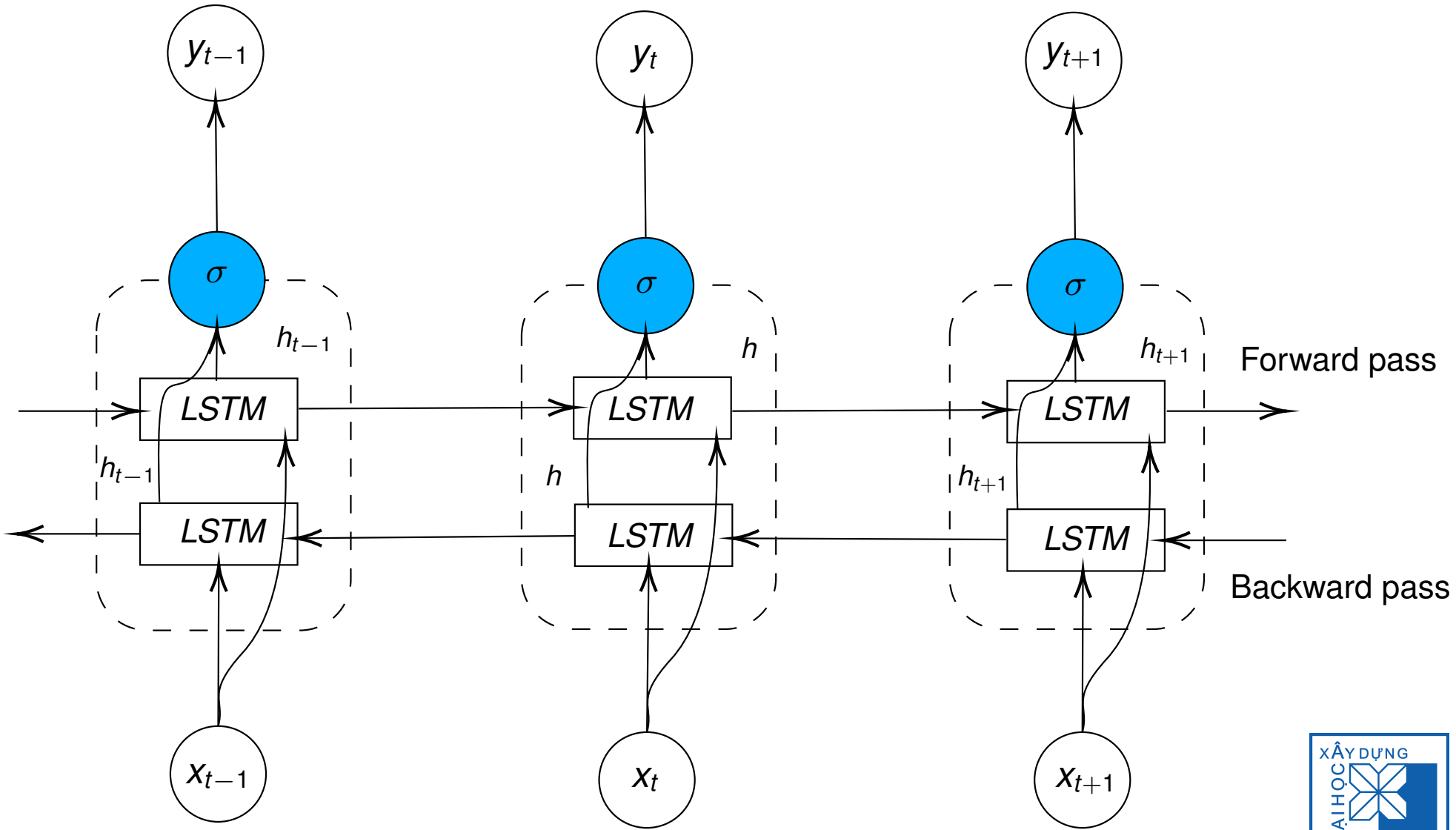
LSTM - Output Gate



$$o_t = \sigma\left(\mathrm{W}_0\left[h_{t-1}, x_t\right] + b_0\right)$$

$$h_t = o_t \otimes \tanh\left(C_t\right)$$

Mô hình LSTM 2 chiều



Nội dung

Tập dữ liệu

Nguồn dữ liệu

- **Nguồn dữ liệu:** https://github.com/duyvuleo/VNTC/tree/master/Data/10Topics/Ver1.1
- **Thông tin dữ liệu:** Dữ liệu được tổng hợp từ nhiều trang báo điện tử khác nhau như vnexpress.net,tuoitre.vn, thanhnien.vn, nld.com.vn. Các bài báo từ rất nhiều lĩnh vực trong cuộc sống như chính trị xã hội, đời sống, khoa học, kinh doanh, pháp luật, sức khoẻ, thế giới, thể thao, văn hoá, ... được chia thành các file nhỏ giúp việc huấn luyện mô hình trở nên dễ dàng và thuận tiện hơn.

Tạo nhiễu dữ liệu

Tạo lỗi khi nhập câu:

Trong phần tạo lỗi này thì em sẽ tạo nhiễu với các trường hợp như sau:

- Lỗi gõ phím (chữ cái và phần phụ âm trong từ)
- Lỗi từ lóng, từ liên quan teencode
- Lỗi các từ phát âm giống nhau
- Lỗi mất dấu câu trong từ
- Lỗi thêm, sửa, xóa chữ cái trong từ đó

Tạo nhiễu dữ liệu - Sơ đồ

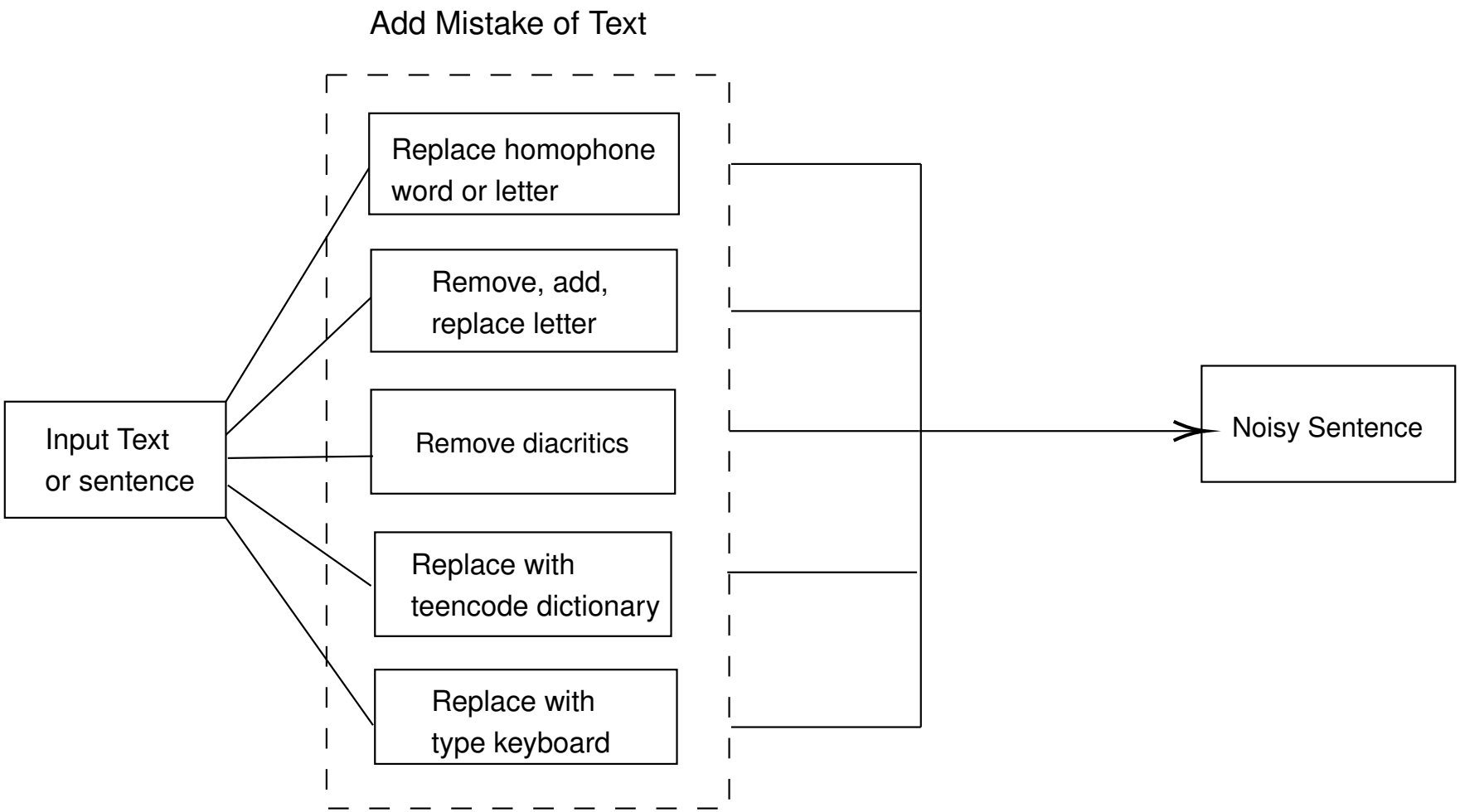


Figure: Tạo lỗi văn bản

Hướng tiếp cận

- **Làm sạch văn bản:** Loại bỏ các ký tự đặc biệt. Dữ liệu khi được lấy từ trên mạng về sẽ thường có các ký tự đặc biệt như: "@%$#" hoặc các icon và các thẻ html. Các ký tự này là các nhiễu làm giảm độ chính xác của quá trình phân loại văn bản.
- **Tách câu:** Tách văn bản thành từng câu được ngăn cách bởi dấu chấm câu “.”
- **Tách từ:** Trong tiếng Việt, dấu cách (space) không được sử dụng như 1 kí hiệu phân tách từ, nó chỉ có ý nghĩa phân tách các âm tiết với nhau. Vì thế, để xử lý tiếng Việt, công đoạn tách từ (word segmentation) là 1 trong những bài toán cơ bản và quan trọng bậc nhất.
- **Chuẩn hóa từ:** Đưa các chữ cái có viết hoa trở về thành các chữ thường để đảm bảo nhiều từ vựng huấn luyện nhất có thể.

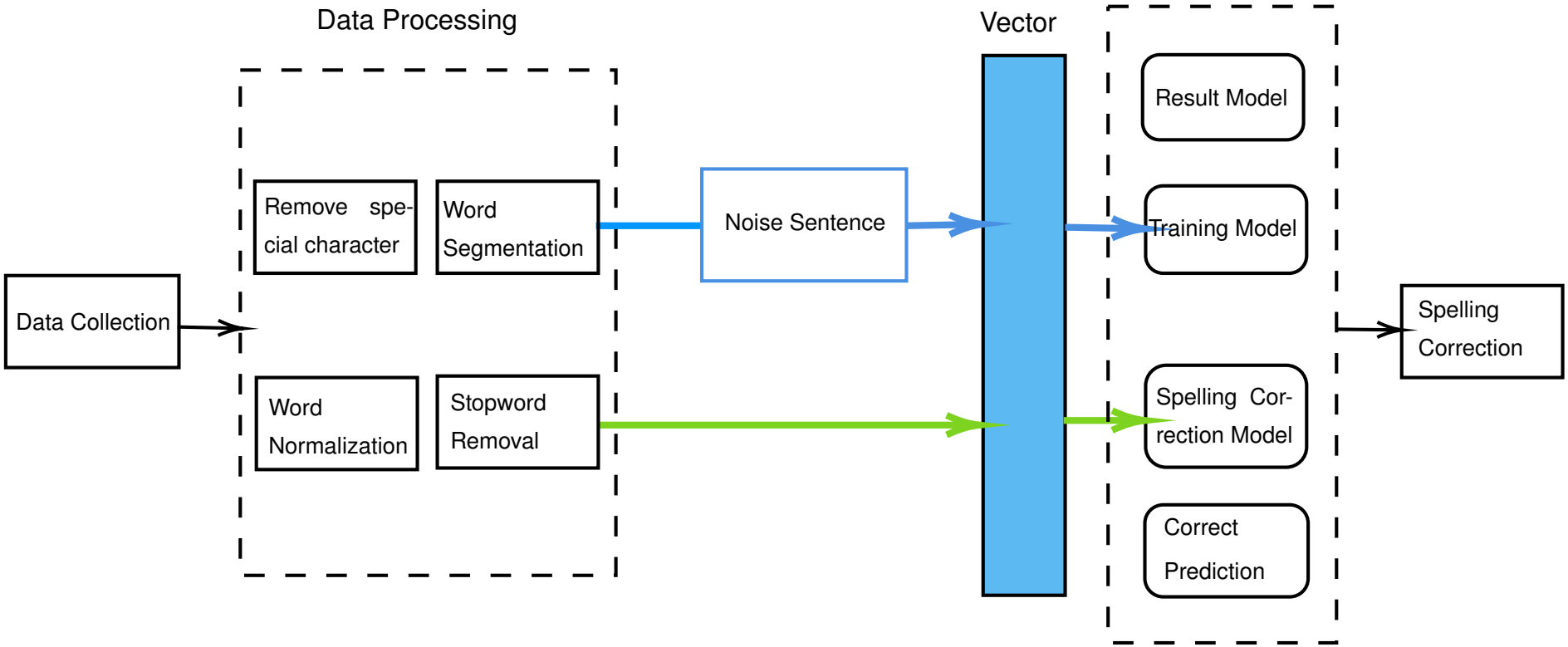


Figure: Xử lý dữ liệu

Thuật toán đề xuất

Thuật toán

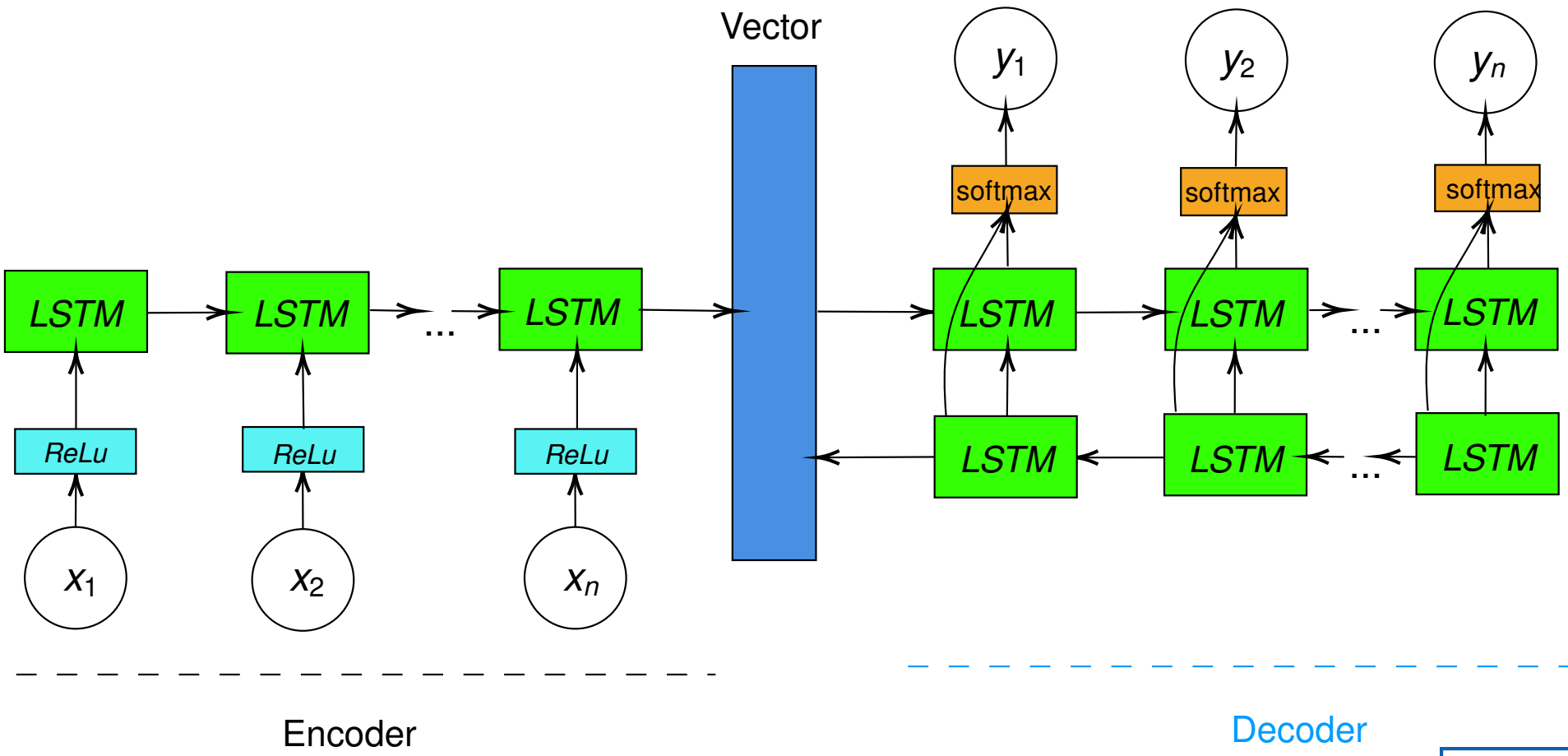
- Tách từ **N_grams**: Tách các nội dung trong tập văn bản đã thu thập với N = 2
- Tách từ **N_grams** với **TF-IDF**: Tương tự như trường hợp N_grams nhưng sẽ đề xuất thuật toán TF-IDF xem trong văn bản thu thập trên có xuất hiện từ nào nhiều nhất thì ta sẽ đề xuất từ đó vào trong kho huấn luyện mô hình rồi tiến hành xử lý tách theo N_grams
- Tách từ **N_grams** với **Stopwords**: Giống như trường hợp 1 nhưng sẽ kết hợp với từ dừng loại bỏ các từ không cần thiết trong câu rồi tiến hành xử lý và tách với N_grams trong kho dữ liệu đã được xử lý

Xây dựng từ điển

Xử lý các từ

- **Xây dựng bộ từ điển:** Biến tất cả các từ của trong văn bản của chúng ta thành dạng biểu diễn số. Cách đơn giản nhất mà chúng ta có thể làm đó chính là xây dựng một bộ từ điển rồi sau đó thay thế từ đó bằng thứ tự xuất hiện trong từ điển.
- **Vector hóa từ và văn bản:** Bước này mục đích là vector hoá từ trong từng câu. Thuật toán dùng để vector hóa từ hay dùng có thể kể đến **word2vect**: biểu thị mỗi từ thành 1 vector. Ngoài ra ta cũng có thể dùng **doc2vect**: biểu thị văn bản thành 1 vector.

Kiến trúc mô hình



Nội dung

Đánh giá thuật toán

Name Model	Accuracy	Precision	F1-Score	Recall
N_grams	0.997969	0.991461	0.748326	0.600955
TF_IDF & N_grams	0.997918	0.990713	0.740360	0.591232
Stopwords & N_grams	0.997898	0.988741	0.73778	0.588426

Biểu đồ so sánh

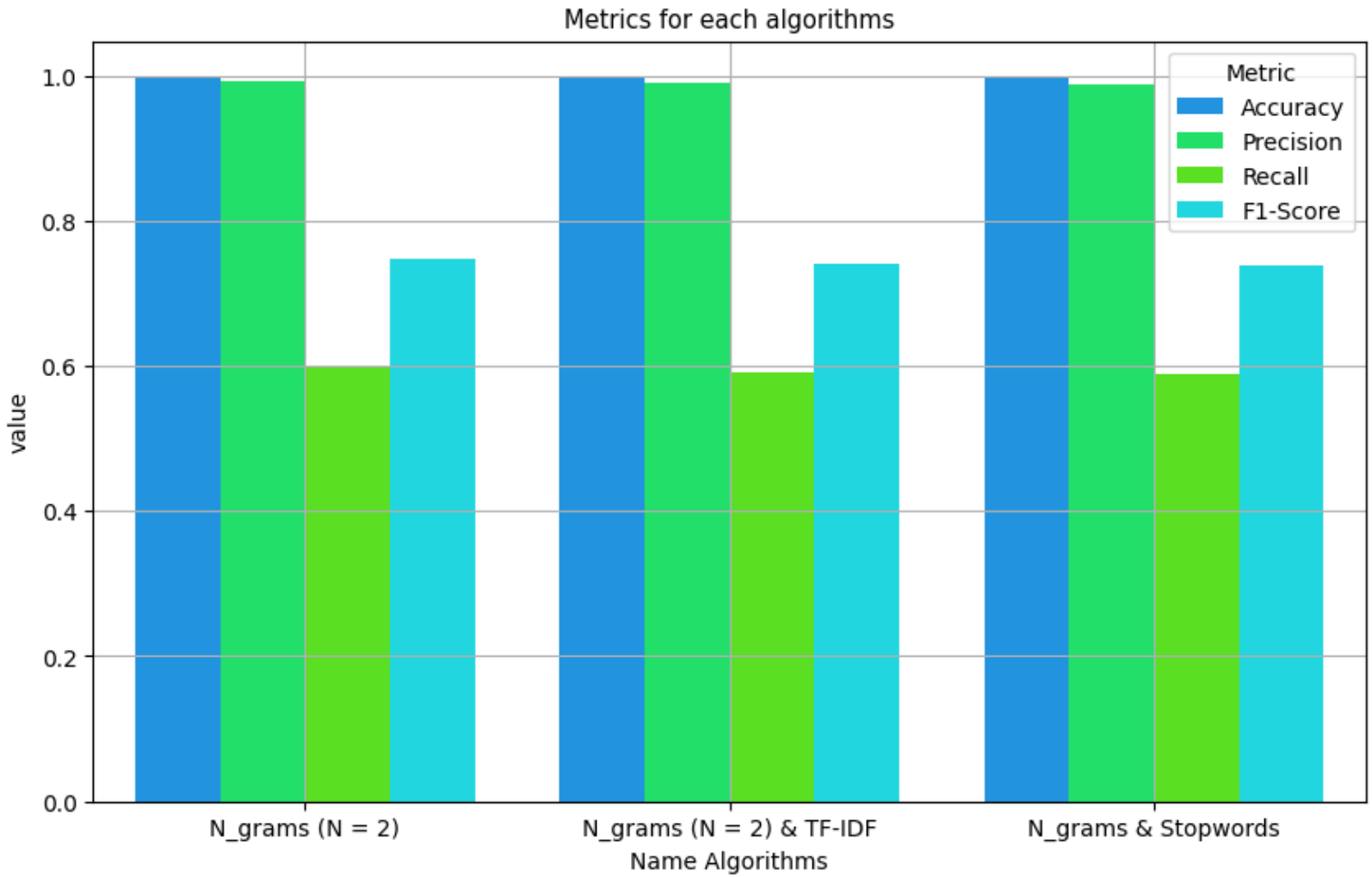


Figure: Biểu đồ so sánh giữa các thuật toán

Nội dung

Các câu sai chính tả

- Dưới đây là một số ví dụ về thực nghiệm phát hiện và sửa lỗi chính tả
- Ngoài ra em có thể thực hiện kiểm thử trên một phần mềm thiết lập bằng PyQt5 tích hợp phần huấn luyện trước đó để tiến hành thực nghiệm

Câu văn bị lỗi chính tả	**Câu văn đã sửa lỗi chính tả**
Anh vaf em	Anh và em
T iu e	Tôi yêu em
Chuyền đajt kiến thức	Truyền đạt kiến thức
Toorng cujc tình báo TocoToco	Tổng cục tình báo TocoToco
Ngọn lúi này rất ca	Ngọn núi này rất cao
Phát thank dẫn chương chình	Phát thanh dẫn chương trình

Đoạn văn sai chính tả

Đoạn văn 1

- **Input:** Các phast thanh viên chuyền hình, laf những cno nguowfi đã nổi tieesng rồi, hoj thường noori bật với cách ăn nói truyeenf tải đến vowsi khán giả car nước, nhiều quys khán giar đã rất ấn tuownjg với phong cach của họ, khán giar hâm mộ thường ấn tượng họ vì hoj đã vieest lên những câu chuyeejn để tạo nên đuowjc điểm nhấn mà khasn giả ko bao gio quên đươwjc, họ áp lực để làm seo mà giữ được hình arnh của mình trướ mắt của công chúng, chỉ là nếk có sự cố gì xary ra thôi là đi luôn car sự nghiệp và rất khó để quayy lại và được khán giả đón nhận như trước kiak
- **Output:** Các phát thanh viên truyền hình là những con người đã nổi tiếng rồi, họ thường nổi bật với cách ăn nói truyền tải đến với khán giả cả nước, nhiều quý khán giả đã rất ấn tượng với phong cách của họ, khán giả hâm mộ thường ấn tượng họ vì họ đã viết lên những câu chuyện để tạo nên được điểm nhấn mà khán giả không bao giờ quên được, họ áp lực để làm sao mà giữ được hình ảnh của mình trước mắt của công chúng chỉ là nếu có sự cố gì xảy ra thôi là đi luôn cả sự nghiệp và rất khó để quay lại và được khán giả đón nhận như trước kia

Đoạn văn sai chính tả

Đoạn văn 2

- **Input:** Chị Daaju là nguowfi có lòng yêu thuwong chồng cực đại nafng ta rất chi là duxng cảm ko hề sợ roi vọt khi chị nhinf thấy chồng mìnk là Thống Lí Pá Tra đánh đập dax man nàng hùng duxng tung chưởng bằng cú karate hết suwsc đẹp mắt rùi mang cho a moojt bát cháo hafnh để ăn cho lịa sứk
- **Output:** Chị Dậu là người có lòng yêu thương chồng cực đại nàng ta rất chi là dũng cảm không hề sợ roi vọt khi chị nhìn thấy chồng mình là Thống Lí Pá Tra đánh đập dã man nàng hùng dũng tung trưởng bằng cú karate hết sức đẹp mắt rồi mang cho anh một bát cháo hành để ăn cho lại sức

Nội dung

Tổng kết

Đồ án của em đạt được một số cơ bản như sau:

- Đồ án này đã đưa ra các lý thuyết và vấn đề trong quá trình thiết lập, huấn luyện và xây dựng một mô hình sửa lỗi chính tả cho tiếng Việt.
- Kết quả ban đầu đạt được là tiền đề để tạo ra các trợ lý ảo, xây dựng các ứng dụng thông minh có thể hiểu được ngôn ngữ tiếng Việt.
- Có khả năng áp dụng vào các bài toán thực tế, ví dụ như các gợi ý sửa lỗi chính tả trong tình soạn thảo, tự động sửa lỗi chính tả trong tìm kiếm, gợi ý từ tiếp theo trong soạn tin nhắn
- Phù hợp với yêu cầu để áp dụng thực tế