Thêm dấu cho tiếng việt không dấu

Đề tài dự thi CodeWar tại Atmarkcafe Việt Nam

Thí sinh: Phạm Văn Khánh
Team: AI

Tại sao phải thêm dấu cho tiếng Việt không dấu?

* Có nhiều lí do khiến chúng ta viết tiếng Việt không dấu:
  - không có phần mềm gõ tiếng việt
  - font chữ lỗi
  - do người dùng lười gõ có dấu.

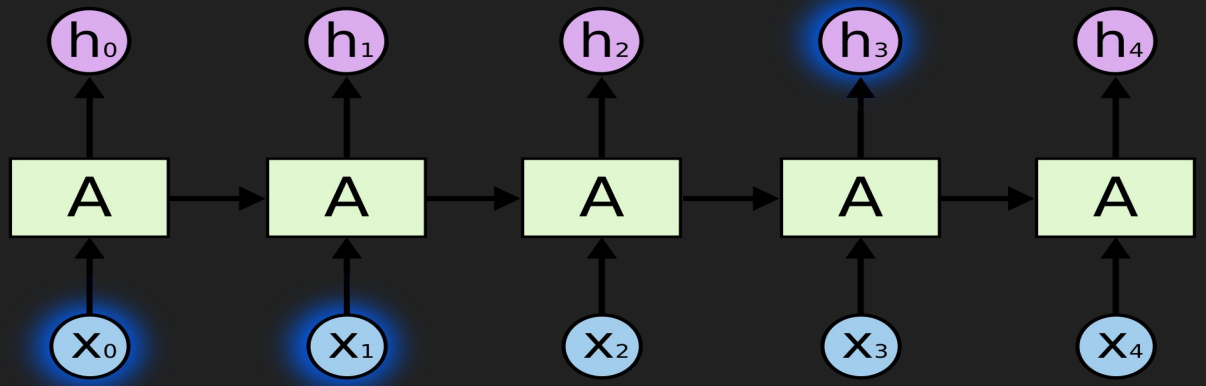
* Tiếng Việt không dấu gây khó hiểu cho người đọc, nội dung không rõ ràng.

Ví dụ: "an com chua con cho?"

Thêm dấu cho tiếng Việt sử dụng deep learning

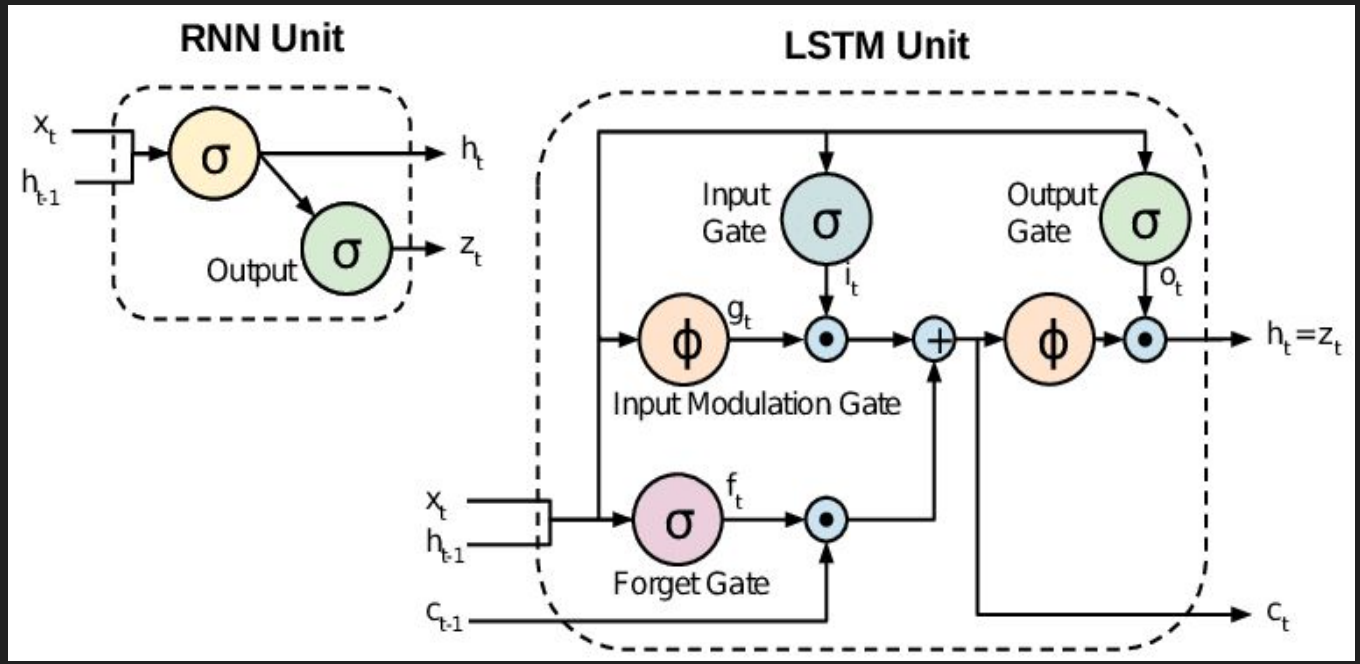
- Sử dụng mạng LSTM(Long Short Term Memory) là một biến thể của mạng RNN(recurrent neural network)

- mạng RNN



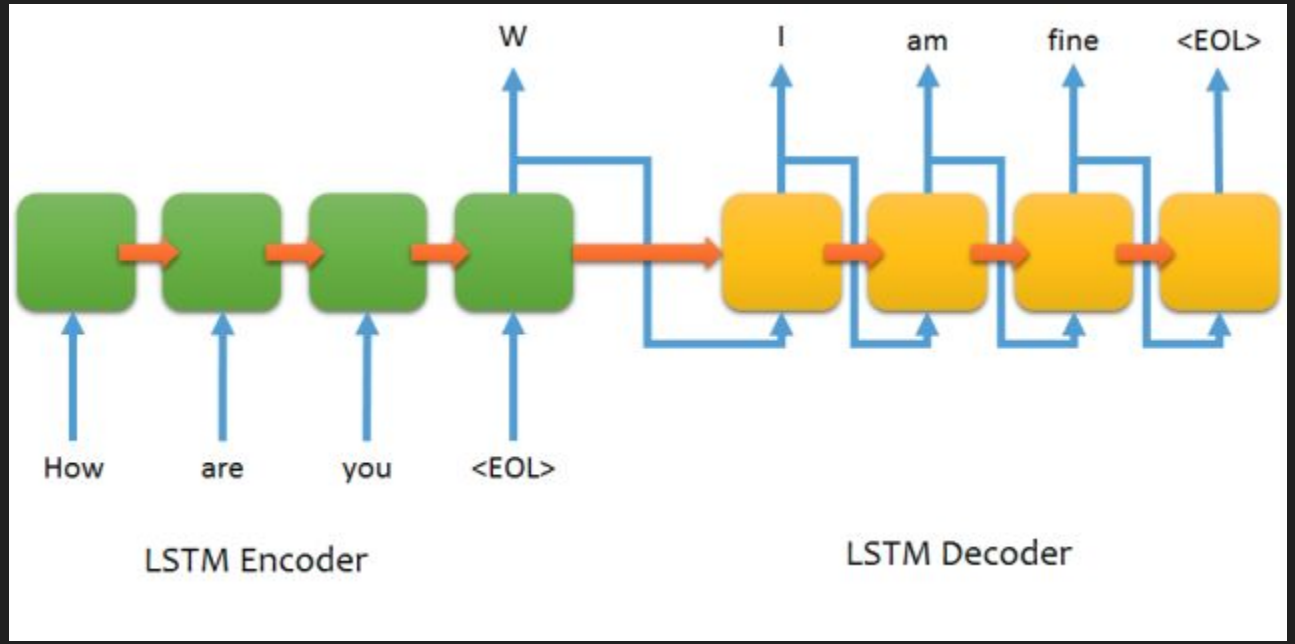
Thêm dấu cho tiếng Việt sử dụng deep learning

- Mạng RNN truyền thống vs LSTM



Thêm dấu cho tiếng Việt sử dụng deep learning

- Mạng encoder-decoder



Thêm dấu cho tiếng Việt sử dụng deep learning

- Input đầu vào được biểu diễn dưới dạng một matrix 2 chiều với 1 dòng là một vector onehot biểu diễn 1 kí tự của chuỗi không dấu với độ dài là 39
- Ví dụ: xin chao:

input: x:[1,0,0,0,0,0,0,0,0,...],
i:[0,1,0,0,0,0,0,0,0,...]
n[0,0,1,0,0,0,0,0,0,...]...

Thêm dấu cho tiếng Việt sử dụng deep learning

- Output biểu diễn tương tự input nhưng mỗi vector độ dài là 109
- Loại bỏ dấu từ bộ dữ liệu có dấu để được dữ liệu không dấu.
- Bộ dữ liệu được thu thập từ vnexpress.net
- Đưa dữ liệu vào mạng để học.

DEMO KẾT QUẢ:

Cảm ơn mọi người đã lắng nghe.