ỨNG DỤNG MÔ HÌNH BILSTM TRONG BÀI TOÁN GÁN NHÃN TỪ LOẠI

Đại học Bách Khoa Hà Nội
Giảng viên hướng dẫn: TS. Trần Việt Trung

05/2020

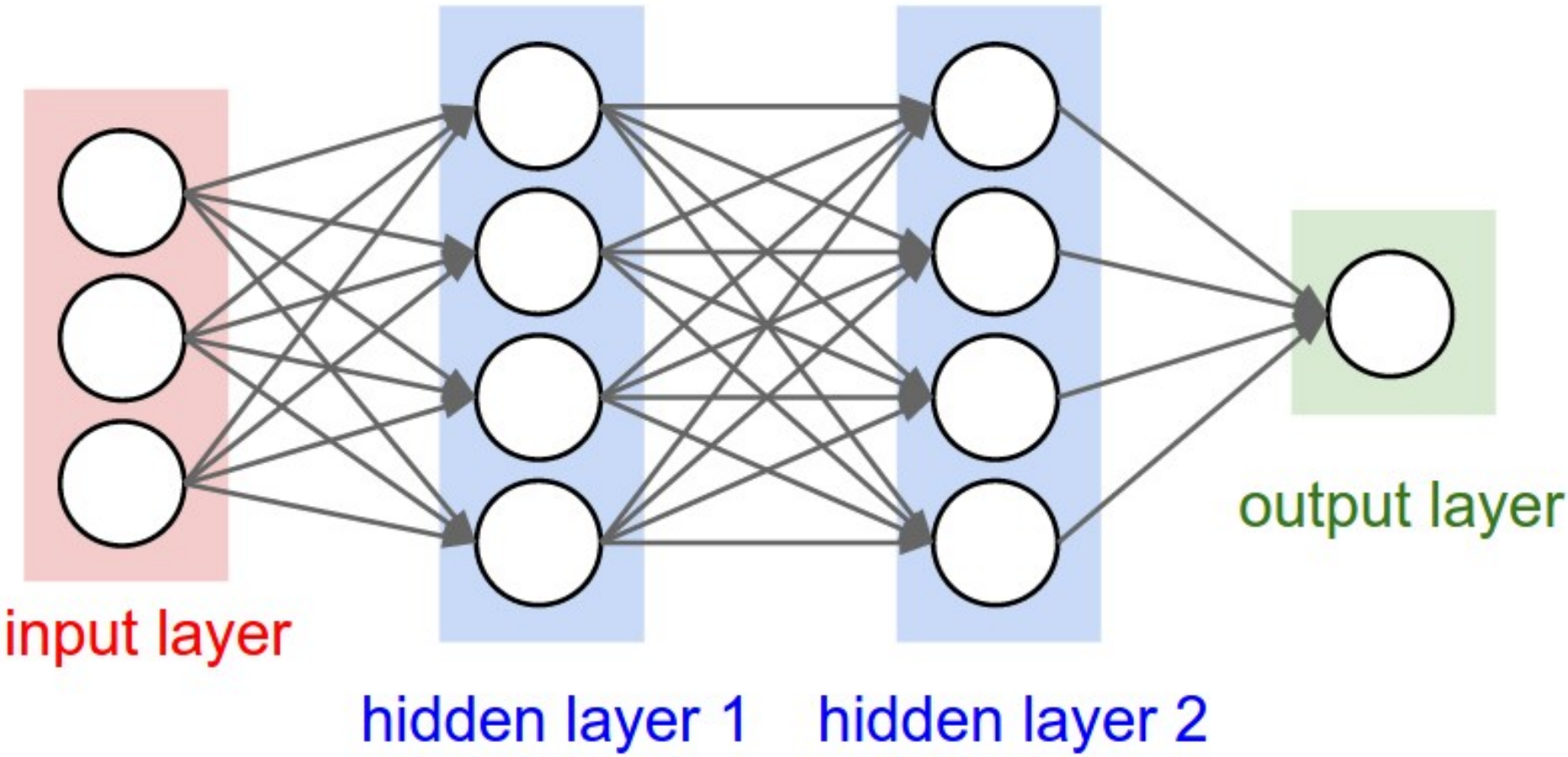
Nội dung

Giới thiệu bài toán

- Trong nhiều bài toán xử lý ngôn ngữ tự nhiên (NLP), ta mong muốn xây dựng được một mô hình mà chuỗi các quan sát (câu, từ ngữ...) đi kèm với chuỗi các nhãn đầu ra (từ loại, ranh giới từ, tên thực thể,...) gọi là pairs of sequences.
- Ví dụ về gán nhãn từ loại:

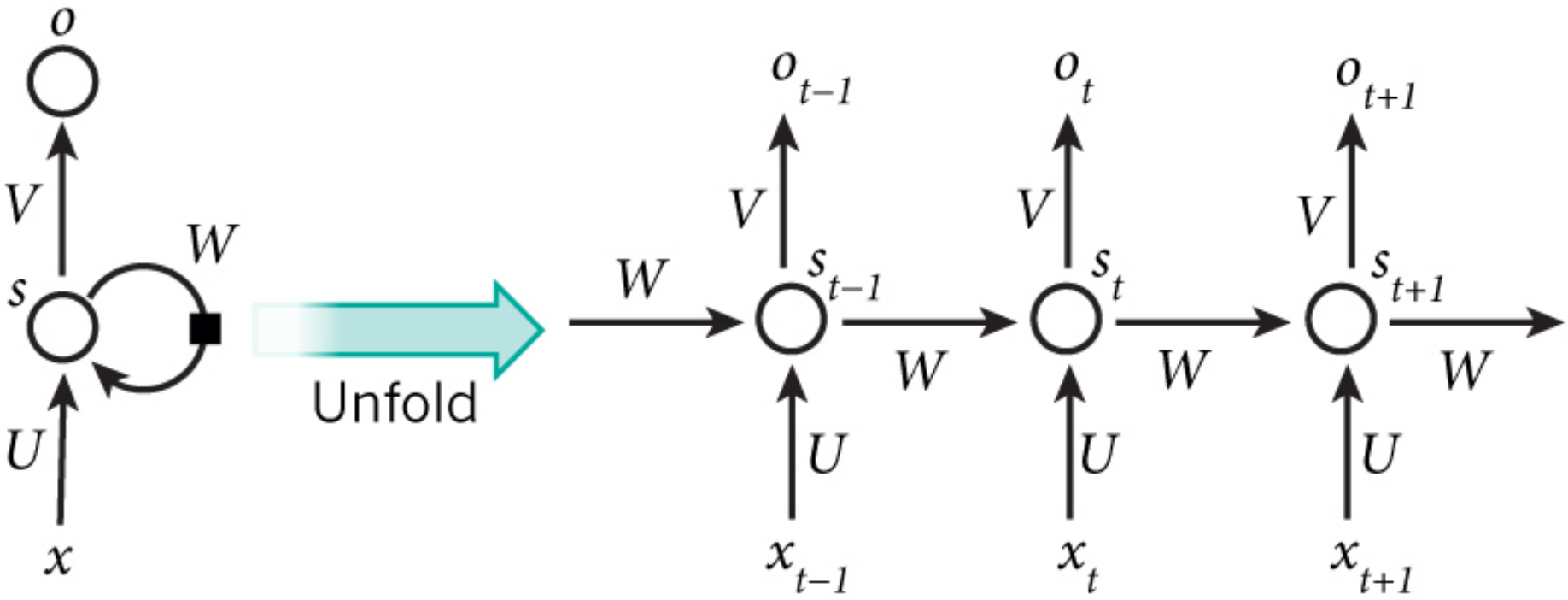
Con	*ruồi*	*đậu*	*mâm*	*xôi*	*đậu*
B-NP	*I-NP*	*B-VP*	*B-NP*	*I-NP*	*I-NP*

Kiến thức cơ sở - Mạng neural



Hình: Kiến trúc mạng Neural

Kiến thức cơ sở - Mạng RNN



Hình: Mô hình mạng Recurrent Neural Network

Kiến thức cơ sở - Mạng RNN

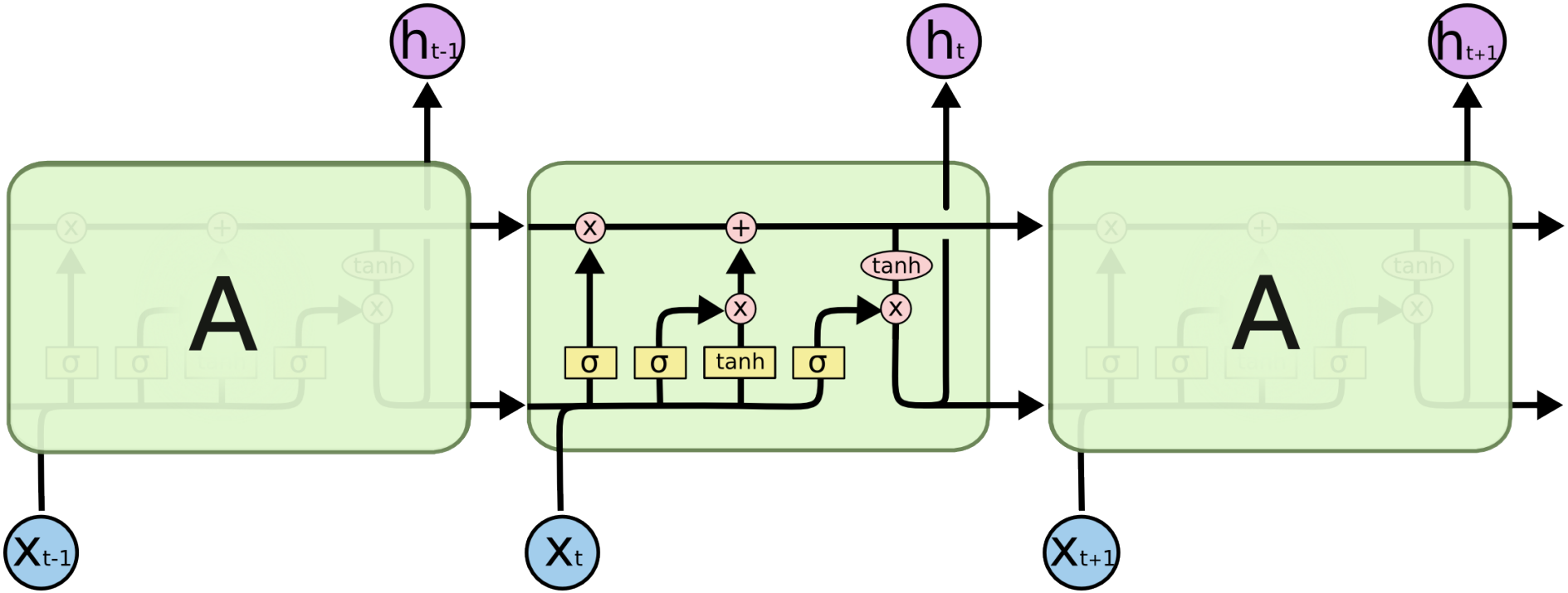
- Lan truyền tiến:

$$s_{t+1} = f(Ux_{t+1} + Ws_t)$$
$$o_{t+1} = g(Vs_{t+1})$$

- Lan truyền ngược:

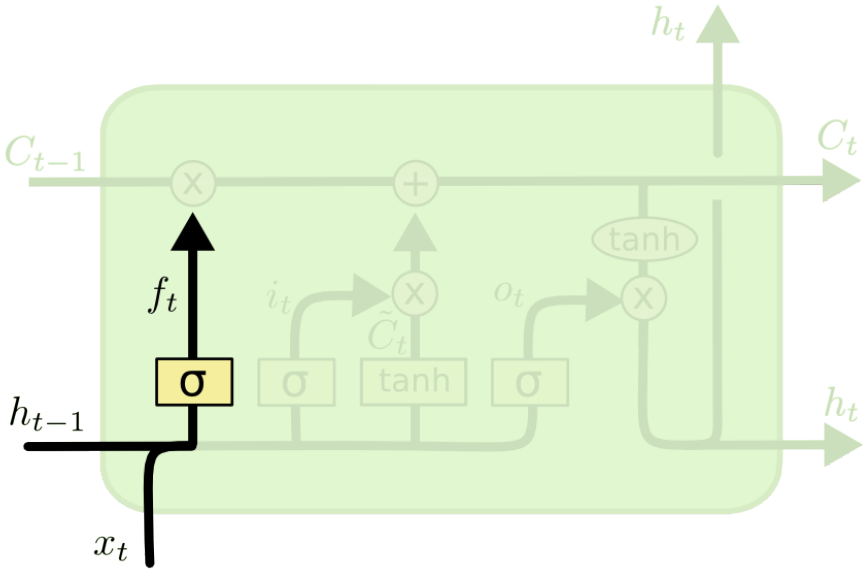
$$\frac{\partial L}{\partial V} = \frac{\partial L}{\partial o_{t+1}} * \frac{\partial o_{t+1}}{\partial V}$$
$$\frac{\partial L}{\partial U} = \frac{\partial L}{\partial o_{t+1}} * \frac{\partial o_{t+1}}{\partial s_{t+1}} * \frac{\partial s_{t+1}}{\partial U}$$
$$= \frac{\partial L}{\partial o_{t+1}} * \frac{\partial o_{t+1}}{\partial s_{t+1}} * \frac{\partial s_{t+1}}{\partial s_t} * \ldots * \frac{\partial s_1}{\partial U}$$
$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial o_{t+1}} * \frac{\partial o_{t+1}}{\partial s_{t+1}} * \frac{\partial s_{t+1}}{\partial U}$$
$$= \frac{\partial L}{\partial o_{t+1}} * \frac{\partial o_{t+1}}{\partial s_{t+1}} * \frac{\partial s_{t+1}}{\partial s_t} * \ldots * \frac{\partial s_1}{\partial W}$$

Kiến thức cơ sở - Mạng LSTM



Hình: Mô hình mạng LSTM

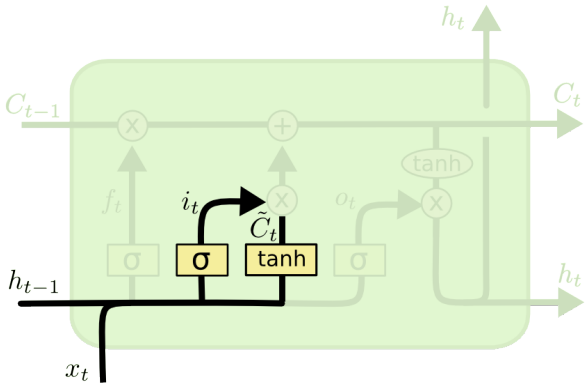
Kiến thức cơ sở - Mạng LSTM (Cổng quên)



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

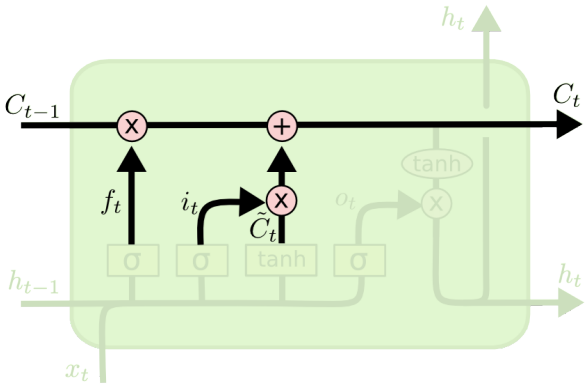
Hình: Cổng quên của mạng LSTM

Kiến thức cơ sở - Mạng LSTM (Cổng vào)



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

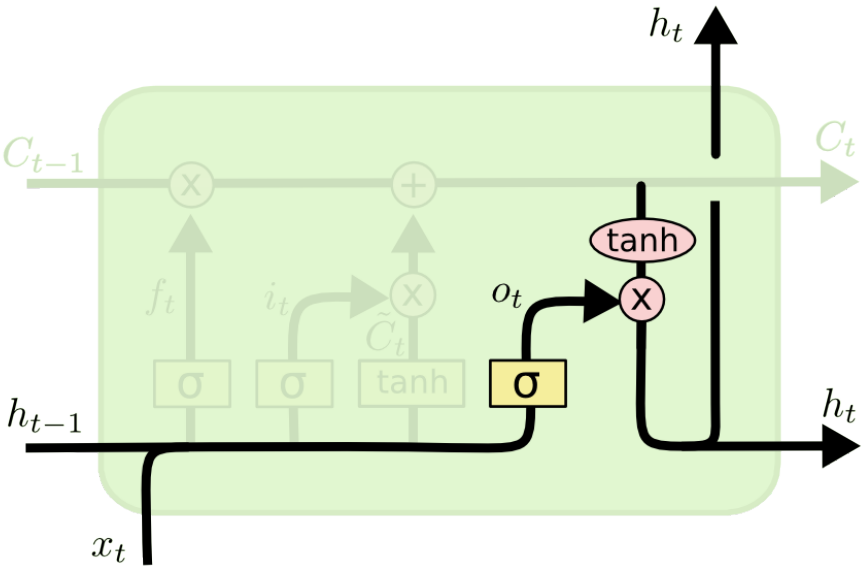
Hình: Cổng vào của mạng LSTM



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Hình: Cập nhật trạng thái tế bào từ cổng vào của mạng LSTM

Kiến thức cơ sở - Mạng LSTM (Cổng ra)

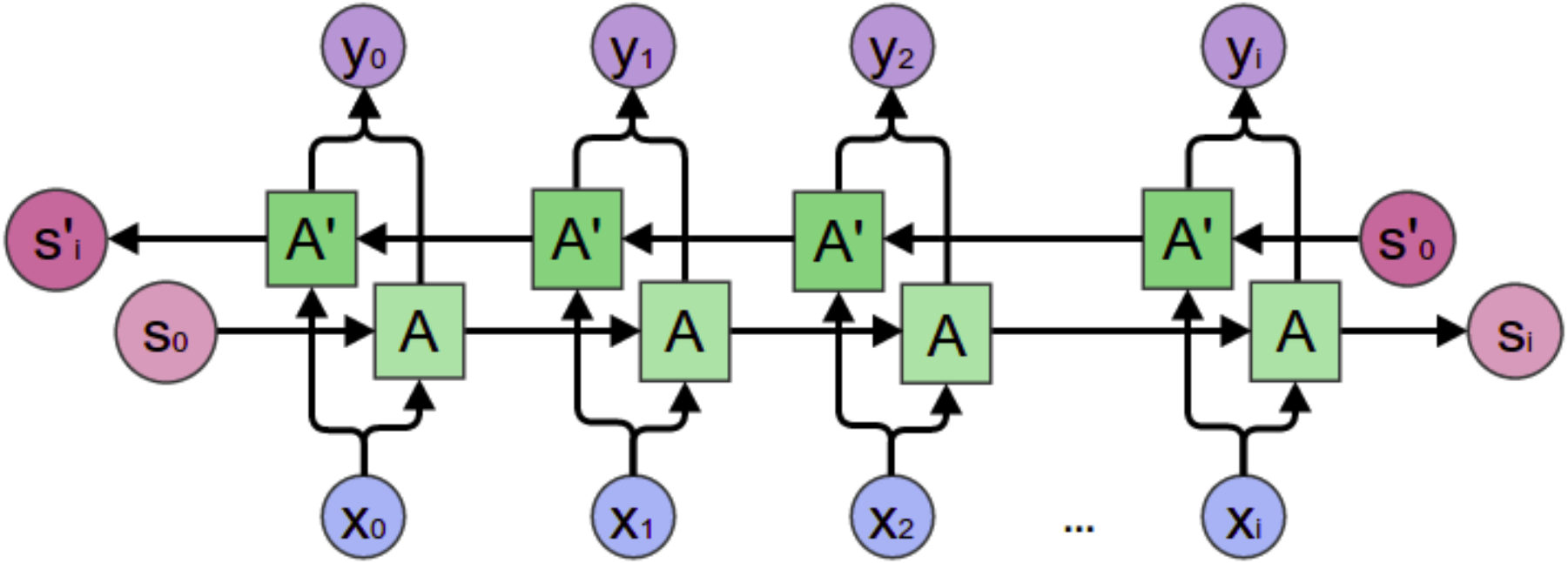


$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

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

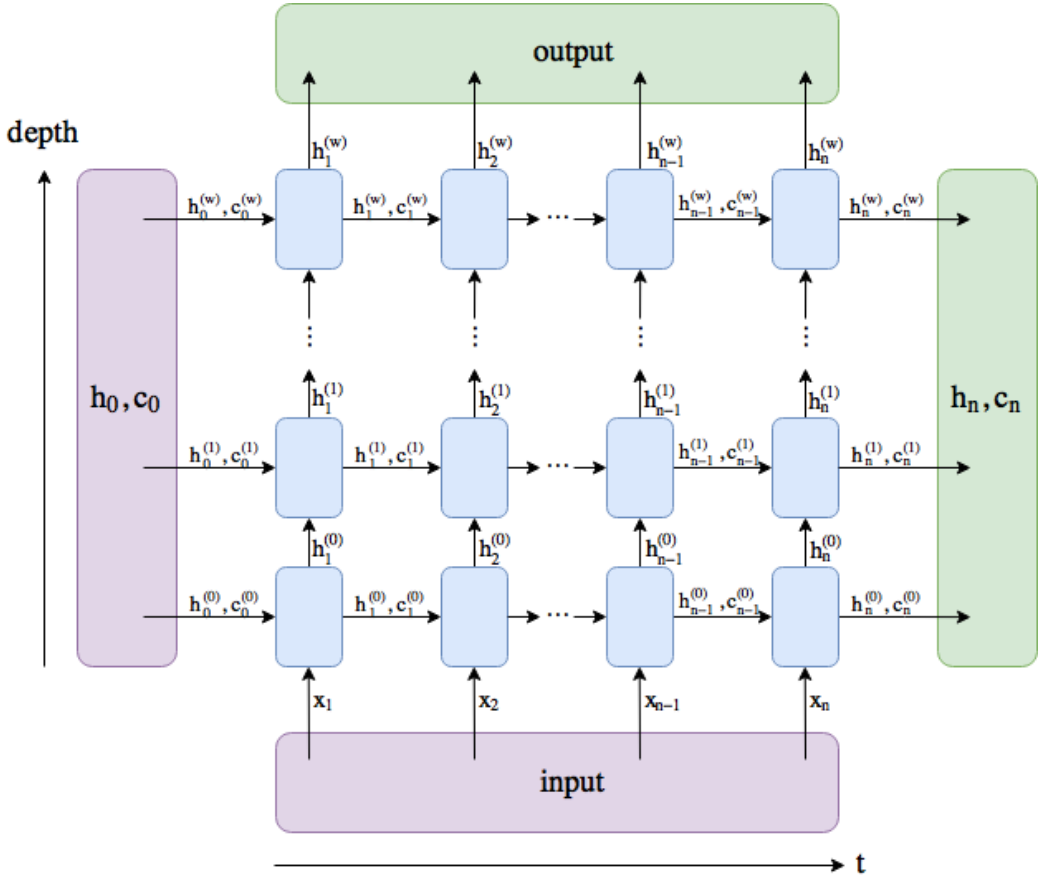
Hình: Cổng ra của mạng LSTM

Kiến thức cơ sở - Mô hình BiLSTM



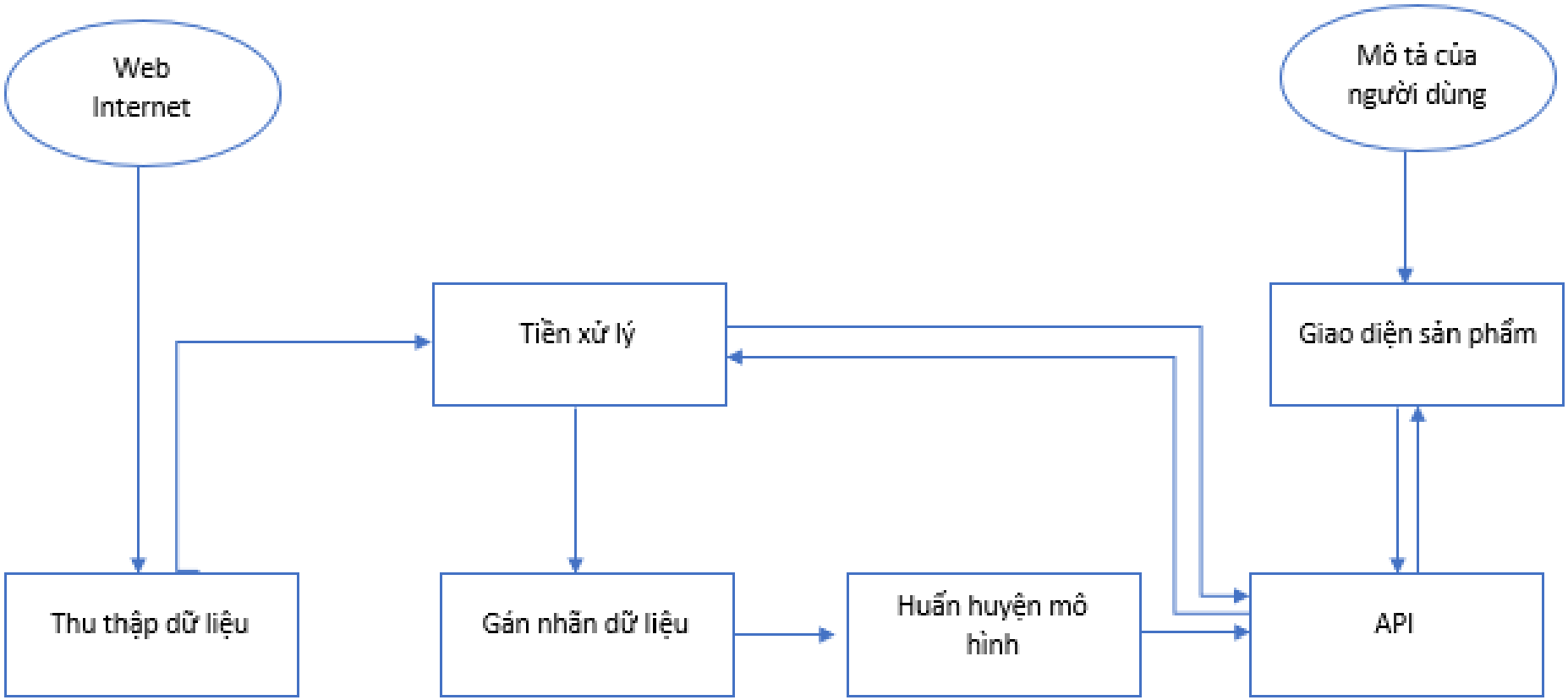
Hình: Hoạt động của Bidirectional RNN

Kiến thức cơ sở - Multi-layer RNN



Hình: Hoạt động của Multi-layer RNN

Thực nghiệm - Quy trình



Hình: Quy trình xử lý bài toán

Thực nghiệm - Các công cụ sử dụng

Công cụ	Chú thích	Pha sử dụng
Scrapy	Thư viện hỗ trợ crawl dữ liệu của Python	Thu thập dữ liệu
Spash	Kết hợp với Scrapy để crawl web chạy bằng JS	Thu thập dữ liệu
Pyvi	Công cụ tokenizer cho tiếng Việt	Tiền xử lý
Tensorflow	Framework Tensorflow machine learning	Huấn luyện mô hình và API
Flask	Thư viện của Python	API
Jquery	Thư viện của Javascript	Giao diện sản phẩm

Hình: Các công cụ sử dụng

Thực nghiệm - Gán nhãn

Quân thù	N	B-NP	O
đang	R	O	O
còn	V	B-VP	O
đó	P	B-NP	O
,	CH	O	O
bao nhiêu	P	B-NP	O
bà	Nc	B-NP	O
mẹ	N	B-NP	O
còn	R	O	O
mất	V	B-VP	O
con	N	B-NP	O
,	CH	O	O
bao nhiêu	P	B-NP	O
người	N	B-NP	O
chồng	N	B-NP	O
mất	V	B-VP	O
vợ	N	B-NP	O
.	CH	O	O

Hình: Các công cụ sử dụng

Thực nghiệm - Ý nghĩa các thẻ

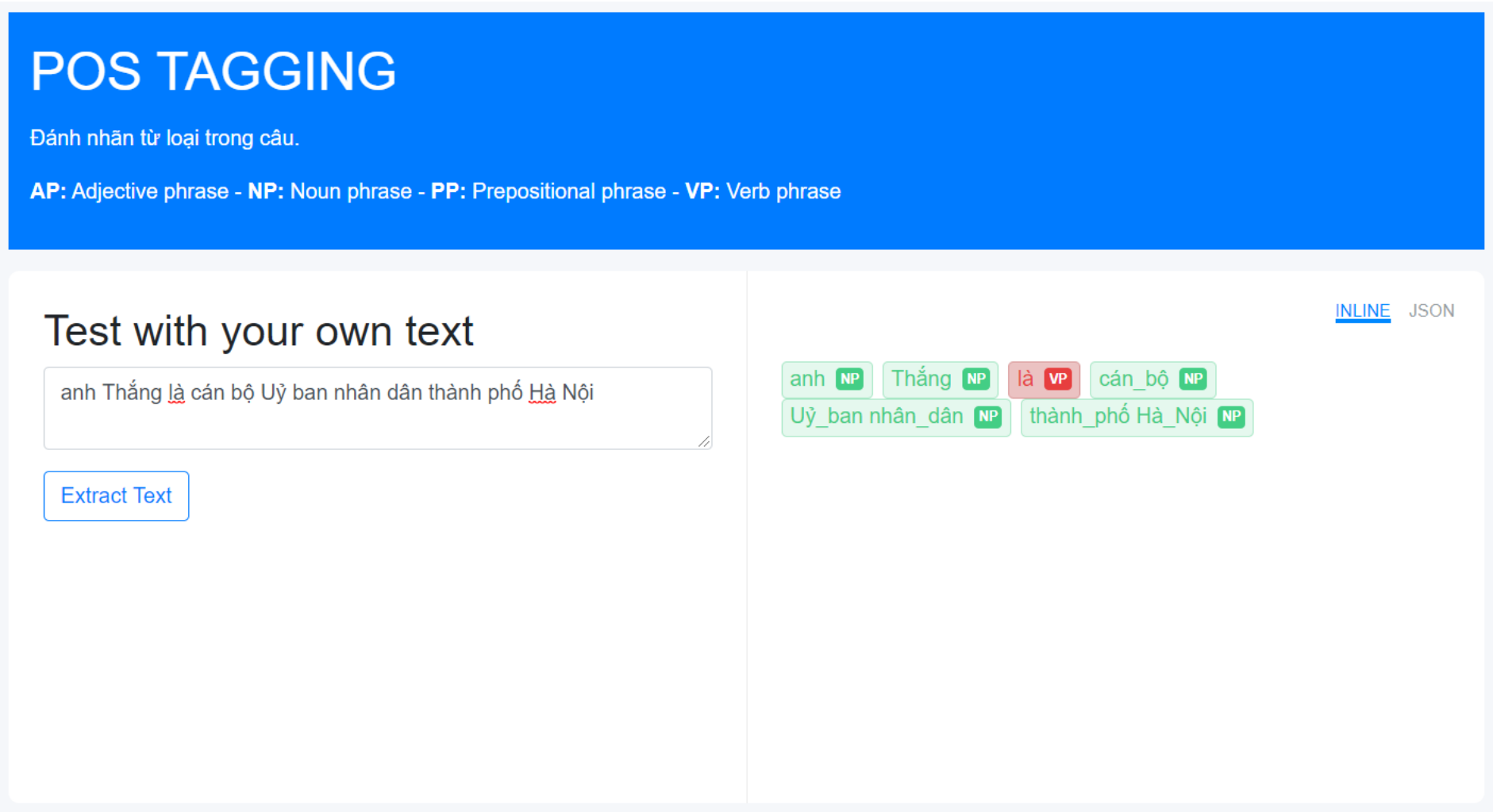
Trong đó ý nghĩa các thẻ như sau:

- **AP**: (Adjective pharse) - Cụm tính từ.
- **NP**: (Noun pharse) - Cụm danh từ.
- **PP**: (Prepositional pharse) - Cụm giới từ.
- **VP**: (Verb pharse) - Cụm động từ.
- **O**: Không xác định.
- **B-**: Thẻ bắt đầu nhãn.
- **I-**: Thực thể nằm trong nhãn nào đó.

Thực nghiệm - Mô hình

Layer	**Output shape**
Embedding	(None, 120, 100)
BiLSTM 1	(None, 120, 128)
Dropout 1	(None, 120, 128)
BiLSTM 2	(None, 120, 256)
Dropout 2	(None, 120, 256)
Dense	(None, 120, 9)

Thực nghiệm - Giao diện



Hình: Giao diện sản phẩm

Thực nghiệm - Giao diện

Hình: Giao diện sản phẩm