DỊCH MÁY NHẬT-VIỆT SỬ DỤNG MÔ HÌNH MẠNG NƠRON HỌC SÂU

Ngô Thị Vinh[1,2,*], Nguyễn Phương Thái[2]
[1]*Trường Đại học Công nghệ Thông tin và Truyền thông - ĐH Thái Nguyên*
[2]*Trường Đại học Công nghệ – Đại học Quốc Gia Hà Nội*

TÓM TẮT

Mô hình mạng nơron sử dụng kỹ thuật học sâu đã đem lại thành công lớn khi áp dụng trong các lĩnh vực nhận dạng đối tượng, nhận dạng giọng nói. Khoảng ba năm trở lại đây, mô hình này được áp dụng cho các các nhiệm vụ khác nhau của xử lý ngôn ngữ tự nhiên như dịch máy, phát hiện các cụm từ, trích xuất các từ được biểu diễn bởi các vectơ. Trong dịch máy, kỹ thuật dịch thống kê sử dụng mạng nơron bước đầu đã đạt các kết quả tốt hơn so với mô hình dịch thống kê tuyến tính truyền thống. Trong bài báo, chúng tôi áp dụng mô hình mạng nơron học sâu cho cặp ngôn ngữ Nhật-Việt trên tập ngữ liệu huấn luyện nhỏ được thu thập từ wiki và đánh giá kết quả đạt được qua chỉ số BLEU là 7,7. Chúng tôi cũng chỉ ra những tồn tại cần cải tiến để nâng cao chất lượng hệ thống dịch trong tương lai.

Từ khóa: *xử lý ngôn ngữ tự nhiên; dịch máy thống kê; dịch máy Nhật-Việt; mạng nơron hồi quy; học sâu.*

GIỚI THIỆU

Trong xử lý ngôn ngữ tự nhiên, dịch máy hay còn gọi là dịch tự động được xem là một trong những nhiệm vụ được quan tâm hàng đầu. Dịch máy được nghiên cứu từ những năm 1950 của thế kỷ trước [4]. Ban đầu, các nhà nghiên cứu sử dụng các từ điển và với các luật được xây dựng bằng tay bởi các chuyên gia để xác định nghĩa đúng của từ. Cách tiếp cận này có hạn chế là không thể xây dựng một hệ luật hoàn chỉnh cho mọi ngôn ngữ. Đến những năm 1990, mô hình dịch máy thống kê ra đời và đạt được những thành tựu lớn trong lĩnh vực này [6]. Các hệ thống dịch máy thống kê sử dụng một kho ngữ liệu song ngữ được thu thập bởi các chuyên gia ngôn ngữ trong quá trình huấn luyện. Tuy nhiên, hiệu suất của các hệ thống dịch máy thống kê phụ thuộc nhiều vào cặp ngôn ngữ đưa ra. Các thực nghiệm về dịch máy thống kê cho thấy hiệu suất của hệ thống đạt được tốt nhất trên những cặp ngôn ngữ có cấu trúc ngữ pháp tương đồng, chẳng hạn như tiếng Anh và tiếng Pháp. Các cặp ngôn ngữ có sự khác biệt lớn về cấu trúc ngữ pháp, đặc biệt các ngôn ngữ không có nguồn gốc từ châu Âu như tiếng Trung Quốc, tiếng Nhật, … thì hiệu suất hệ thống dịch giảm đáng kể. Các nghiên cứu về dịch máy thống kê gần đây cho thấy các mô hình mạng nơron sử dụng kỹ thuật học sâu đạt được các kết quả tốt hơn so với mô hình dịch máy thống kê truyền thống [3], [9]. Trong đó, các thực nghiệm chủ yếu thực hiện hiện trên các cặp ngôn ngữ phổ biến Anh - Pháp, Anh - Trung, Anh - Nhật, Anh-Đức,…, còn các cặp ngôn ngữ khác không bao gồm tiếng Anh như Trung - Việt, Nhật-Việt thì chưa được nghiên cứu rộng rãi. Trong bài báo chúng tôi áp dụng mô hình mạng nơron học sâu cho cặp ngôn ngữ Nhật-Việt.

CÁC NGHIÊN CỨU LIÊN QUAN

Cách tiếp cận dịch máy sử dụng mô hình mạng nơron được phát triển từ rất sớm bởi R. Miikkulainen và M.G. Dyer năm 1991 [10]. Bengio và các cộng sự [2] đã sử dụng mô hình mạng nơron để học các phân bố của dãy các từ với quy mô lớn. Gần đây, dịch máy dựa trên mạng nơron đang là cách tiếp cận thu hút sự chú ý nhiều nhà nghiên cứu về dịch máy [3], [5], [9], [11]. Trong cách tiếp cận này, một bộ mã hóa (gọi là encoder) được sử dụng để mã hóa các câu đầu vào trong ngôn ngữ nguồn thành trình tự các các vectơ có độ dài cố định, sau đó, một bộ giải mã (gọi là decoder) thực hiện giải mã các vectơ có độ dài cố định thành các câu đầu ra trong ngôn ngữ đích. Quá trình huấn luyện là quá trình

* *Tel: 0987 706830, Email: ntvinh@ictu.edu.vn*

cực đại hóa xác suất có điều kiện của bản dịch ứng với các câu nguồn đưa ra. Các mạng nơron hồi quy (gọi tắt là RNN - Recurent Neural Network) sử dụng kỹ thuật học sâu với các đơn vị ẩn dạng LSTM (long short-term memory) hoặc GRU (gated recurrent units) đã giải quyết được vấn đề biến mất hoặc bùng nổ giá trị gradient của các mạng RNN truyền thống trong quá trình huấn luyện [3], [11].

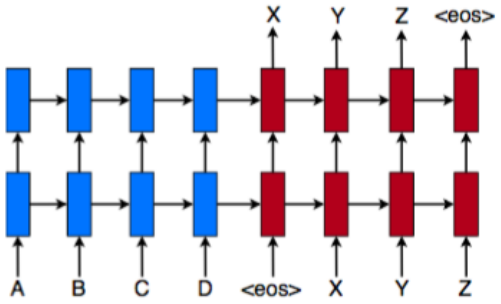
Mặc dù tiếng Việt là ngôn ngữ đứng thứ 13 thế giới về mức độ được sử dụng rộng rãi. Tuy nhiên, các nghiên cứu về tiếng Việt còn khá ít. Khoảng 10 năm trở lại đây có một số nhóm nghiên cứu về dịch máy tiếng Việt nhưng chủ yếu tập trung vào hệ dịch Anh-Việt, Pháp-Việt. Hiện nay Google là hệ thống dịch mở được sử dụng nhiều nhất trên thế giới đã tích hợp tiếng Việt vào hệ thống của họ. Hệ dịch mở Google dịch khá tốt giữa tiếng Anh với các ngôn ngữ khác, tuy nhiên với các cặp ngôn ngữ khác như Nhật-Việt Google sử dụng tiếng Anh làm trung gian nên chất lượng dịch còn khá thấp [17]. Các nghiên cứu về dịch Nhật-Việt trực tiếp còn khá khiêm tốn [7], [14] do bị hạn chế nguồn ngữ liệu song ngữ Nhật-Việt và sự khác biệt về cấu trúc ngữ pháp và đặc trưng ngôn ngữ. *Cho đến nay vẫn chưa tồn tại nguồn ngữ liệu song ngữ Nhật-Việt mở nào*. Các nghiên cứu về dịch Nhật-Việt gần đây vẫn áp dụng dụng mô hình dịch thống kê truyền thống [13].

Trong bài báo này chúng tôi áp dụng cách tiếp cận dịch sử dụng mô hình mạng nơron học sâu được đề xuất trong [11], [12] kết hợp cơ chế nhớ LSTM [15] cho cặp ngôn ngữ Nhật-Việt và khai thác hệ thống mã nguồn mở [16] trong việc xây dựng hệ thống.

MÔ HÌNH MẠNG NƠRON HỌC SÂU CHO DỊCH MÁY THỐNG KÊ NHẬT-VIỆT

Mô hình mạng nơron hồi quy

Mô hình dịch máy ở đây sử dụng một Encoder là một mạng nơron hồi quy thực hiện mã hóa các câu đầu vào của văn bản tiếng Nhật thành trình tự các vectơ có độ dài cố định và một Decoder thực hiện giải mã các trình tự các vectơ thành các từ trong văn bản tiếng Việt. Mạng RNN ở cả phía Encoder và Decoder là mạng nơron nhiều lớp.



Hình 1. *Sơ đồ cấu trúc mạng nơron sử dụng mô hình seq2seq, với A, B, C, D là các từ trong ngôn ngữ nguồn, X, Y, Z là các từ trong ngôn ngữ đích*

Gọi cặp (x, y) là cặp tương ứng giữa câu tiếng Nhật trong văn bản nguồn và câu tiếng Việt trong văn bản đích của ngữ liệu song ngữ. Với x và y chứa dãy các từ sao cho: $x = (x_1, \ldots, x_n)$ và $y = (y_1, \ldots, y_m)$. Tại thời điểm t, với mỗi đầu vào x_t RNN Encoder sẽ cập nhật trạng thái bộ nhớ của nó và sinh ra trạng thái ẩn h_t theo công thức:

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1}) \quad (1)$$

Đầu ra tại thời điểm t là $s_t(y)$ được tính theo công thức:

$$s_t(y) = W_{hy}h_t \quad (2)$$

Phân bố xác suất trên mỗi bản dịch y là:

$$p_t(y) = \frac{e^{s_t(y)}}{\Sigma_{y' \in Y} e^{s_t(y')}} \quad (3)$$

Với $y' \neq y, W_{xh}, W_{hy}$ là các ma trận vuông của số thực được khởi tạo ban đầu, trạng thái ẩn h_0 ban đầu được khởi tạo là 0.

Cơ chế nhớ dạng LSTM (Long Short-Term Memory)

Để giải quyết vấn đề biến mất hoặc bùng nổ giá trị gradient trong quá trình huấn luyện chúng tôi sử dụng các đơn vị ẩn LSTM được đề xuất trong [15] với 3 cổng đầu vào (i_t), đầu ra (o_t), khoá (f_t) để tính giá trị ô nhớ c_t và trạng thái ẩn h_t cho một đơn vi ẩn:

$$i_t = sigmoid(W_1.x_t + W_2h_{t-1}) \quad (4)$$

$$f_t = sigmoid(W_3x_t + W_4h_{t-1}) \quad (5)$$

$$o_t = W_5x_t + W_6h_{t-1} \quad (6)$$

$$i'_t = tanh(W_7x_t + W_8h_{t-1}) \quad (7)$$

$$c_t = c_{t-1} \circ f_t + i_t \circ i'_t \quad (8)$$

$$h_t = c_t \circ o_t \quad (9)$$

Ký hiệu ∘ biểu thị phép nhân giữa hai vectơ hoặc vectơ với ma trận, các ma trận từ $W_1, \ldots, W_8$ và vectơ h_0 là các tham số được khởi tạo ban đầu của mô hình.

Giải thuật huấn luyện

Quá trình huấn luyện gồm 2 pha là lan truyền thẳng và lan truyền ngược với hàm mục tiêu:

$$J = \sum_{(x,y)\in D} -logp(y \vee x) \quad (10)$$

Với D là tập ngữ liệu huấn luyện chứa các cặp câu song ngữ Nhật-Việt (x,y) tương ứng.

Quá trình huấn luyện mạng nơron để tính toán lượng tổn thất trên cặp (x,y) chính là tính toán lượng tổn thất trên các câu đích y, do x là câu nguồn không thay đổi. Gọi lượng mất mát l_t tại thời điểm t và được tính như sau:

$$l_t = -logp_t(y_t) \quad (11)$$

Pha lan truyền thẳng tính toán lượng mất mát từng câu y trong văn bản đích (tiếng Việt) và trạng thái cuối cùng của tầng ẩn trong Encoder để làm trạng thái khởi tạo ban đầu của Decoder. Giải thuật lan truyền thẳng như hình 2. Với W_e là ma trận vectơ từ (word embedding) biểu diễn từ và T_{lstm} là tập các ma trận từ $W_1, \ldots, W_8$ trong các công thức từ (4) đến (7).

Pha lan truyền ngược thực hiện ngược lại so với pha tiến bằng cách tính các gradient tương ứng tại mỗi thời điểm *t* để cập nhật lại giá trị cho từ được dự đoán trong ma trận W_e theo các công thức lần lượt được áp dụng trong các giải thuật lan truyền trong hình 2 và hình 3 như sau:

$$ds_t = 1_{yt} - p_t \quad (12)$$

$$dW_{hy} = dW_{hy} + ds_t.h_t^T \quad (13)$$

$$dh_t = dh_t + W_{hy}^T.ds_t \quad (14)$$

$$do_t = tanh(c_t) \circ dh_t \quad (15)$$

$$dc_t = dc_t + tanh'(c_t) \circ o_t \circ dh_t \quad (16)$$

$$df_t = c_{t-1} \circ dc_t \quad (17)$$

$$di_t = i'_t \circ dc_t \quad (18)$$

$$di'_t = i_t \circ dc_t \quad (19)$$

$$dc_{t-1} = f_t \circ dc_t \quad (20)$$

Gọi $du_t = [di_t; df_t; do_t; di'_t]$, ta có các gradient trung gian:

$$dT_x \leftarrow (g'(T_{lstm}z_t) \circ du_t).x_t^T \quad (21)$$

$$dT_h \leftarrow (g'(T_{lstm}z_t) \circ du_t).h_{t-1}^T \quad (22)$$

Các gradient đầu vào dx_t và dh_{t-1} được tính như sau:

$$dx_t \leftarrow T_x^T.(g'(T_{lstm}z_t) \circ du_t) \quad (23)$$

$$dh_{t-1} \leftarrow T_h^T.(g'(T_{lstm}z_t) \circ du_t) \quad (24)$$

Giải thuật 2 trong hình 3 là giải thuật huấn luyện trong pha lan truyền ngược.

Quá trình giải mã

Quá trình giải mã sẽ thực hiện chọn bản dịch tốt nhất ứng với mỗi câu nguồn đưa ra theo công thức (25).

$$\hat{T} = argmaxP(T \vee X) \quad (25)$$

với T là bản dịch trung gian, $\hat{T}$ là bản dịch tốt nhất.

Ngữ liệu song ngữ và thực nghiệm

Trong bài báo, chúng tôi sử dụng ngữ liệu song ngữ Nhật-Việt được thu thập từ wikipedia gồm 40.000 cặp câu. Ngữ liệu tiếng Nhật được tách từ sử dụng công cụ Mecab, ngữ liệu tiếng Việt được tách từ sử dụng công cụ vitk.vn. Sau khi loại bỏ các câu có độ dài trên 80 từ, chúng tôi chia dữ liệu thành tập huấn luyện gồm 35.530 cặp câu, tập phát triển gồm 500 cặp câu, tập kiểm tra gồm 1000 cặp câu.

Chúng tôi tiến hành thực nghiệm hệ thống trên máy chủ tính toán hiệu năng cao với GPU có 12GB RAM. Số lượng từ vựng có tần

số lớn nhất cho cả tập dữ liệu tiếng Nhật và Tiếng Việt là 20000 từ. Chúng tôi tiến hành thực nghiệm hệ thống với nhiều bộ tham số khác nhau như trong bảng 1. Thời gian huấn luyện 8 giờ đồng hồ với 12000 bước lặp, tốc độ học là 1,0 và số bước nhảy 0,2.

Đánh giá hiệu suất hệ thống

Để đánh giá hiệu suất của mô hình trên tập dữ liệu, chúng tôi sử dụng chỉ số đánh giá đo độ trùng khớp giữa hai văn bản song ngữ BLEU (bilingual evaluation understudy) được đề xuất bởi Kishore và các cộng sự [8]. Chỉ số này đo độ chính xác giữa bản dịch bằng máy với bản dịch được thực hiện bởi các chuyên gia bằng cách đến các cặp n-gram tương ứng và tỉ lệ của chúng trong bản dịch so với bản tham chiếu.

KẾT QUẢ VÀ BÀN LUẬN

Sau khi thực hiện huấn luyện với 4 bộ tham số như bảng 1, chúng tôi đạt được chỉ số BLEU cao nhất là 7.7 sau 12.000 bước lặp với 2 lớp ẩn và 128 chiều. Chỉ số này là khả quan và tỉ lệ với tập dữ liệu khi so sánh với hệ dịch Anh-Việt là 23.3 với 2 lớp ẩn, 512 chiều trên tập ngữ liệu 133.000 cặp câu, hệ dịch Anh-Đức là 29.9 với 4 lớp ẩn, 1024 chiều trên tập ngữ liệu 4.5 triệu câu.

Bảng 1. *Kết quả thực nghiệm với các bộ tham số khác nhau*

Lần	Số lớp ẩn	Số chiều	Số bước lặp	Điểm BLEU tốt nhất
1	4	512	12.000	2.38
2	4	128	45.000	2.7
3	2	128	12.000	7.7

Khi áp dụng phương pháp được đề xuất trong [5] và [6] trên cùng tập dữ liệu Nhật-Việt, chúng đạt được chỉ số BLEU cao nhất lần lượt là 0.54 và 1.07. Khi chúng tôi tăng số lớp ẩn và số chiều thì chỉ số điểm BLUE giảm rất lớn. Chúng tôi cũng tăng số bước lặp lên 100 và quan sát thấy chỉ số BLUE cũng không tăng lên mặc dù số bước lặp lớn tăng lên nhiều lần. Chúng tôi nhận thấy nhiều mẫu dịch khá sát so với bản dịch được thực hiện bởi con người như trong bảng 2 (chỉ số BLUE là 90). Các từ không xuất hiện trong tập từ vựng được coi là UNK. Bảng 2 cho thấy các từ tên riêng thường xuất hiện ít nên chúng không được đưa vào tập từ vựng và bị coi là UNK. Bảng 2 cũng cho thấy hệ thống dịch vẫn còn tồn tại các vấn đề như tập dữ liệu huấn luyện nhỏ dẫn đến việc thưa dữ liệu làm giảm chất lượng bản dịch, các từ UNK chưa được xử lý triệt để để tăng chất lượng dịch, các định dạng dữ liệu số, dữ liệu kiểu ngày tháng còn chưa được xử lý tốt.

KẾT LUẬN

Chúng tôi đã áp dụng phương pháp dịch máy sử dụng mạng nơron cho cặp ngôn ngữ Nhật-Việt và bước đầu thu được kết quả khả quan. Trong tương lai chúng tôi sẽ tiếp tục thu thập và bổ sung thêm ngữ liệu cho hệ thống, tối ưu các tham số đầu vào cho hệ thống, đồng thời khai thác thêm các đặc trưng ngôn ngữ và các nguồn dữ liệu mở như các từ điển nhằm tích hợp vào hệ thống để nâng cao chất lượng dịch.

Giải thuật 1: Giải thuật huấn luyện pha lan truyền thẳng

Đầu vào: tập huấn luyện gồm các cặp câu song ngữ Nhật-Việt (x,y) với x có độ dài m_x và y có độ dài m_y.

Các tham số: $W_e^{encoder}$, $T_{lstm}^{encoder}$, $W_e^{decoder}$, $T_{lstm}^{decoder}$.

Đầu ra: lượng mất mát l và các biến trung gian cho quá trình lan truyền ngược.

s ←[x,-,y-]; *//s là xâu kết hợp s= x+"-" +y+"-"*

$W_e, T_{lstm}^{(1..L)} \leftarrow W_e^{encoder}, T_{lstm}^{encoder}$; *//khởi tạo các ma trận trọng số encoder*

$h_0^{(1..L)}, c_0^{(1..L)} \leftarrow 0$;

for $t = 1 \rightarrow (m_x + 1 + m_y)$ **do**

  if $t == (m_x + 1)$ **then**

    $W_e, T_{lstm}^{(1..L)} \leftarrow W_e^{decoder}, T_{lstm}^{decoder}$; *//khởi tạo ma trận trọng số decoder*

  end

  $h_t^{(0)} \leftarrow Emb\_lookUp(s_t, W_e)$; *//gán embedding ứng với từ đầu tiên cho mỗi* s_t

  for $l = 1 \rightarrow L$ **do**

    //tính toán các trạng thái ẩn theo công thức (8), (9)

    $h_t^{(l)}, c_t^{(l)} \leftarrow LSTM(h_{t-1}^{(l)}, c_{t-1}^{(l)}, h_t^{(l-1)}, T_{lstm}^{(l)})$;

  end

  // dự đoán từ trong văn bản đích phía decoder

  if $t \geq (m_x + 1)$ **then**

    $l_t, p_t \leftarrow$ Predict($s_{t+1}, h_t^{(L)}, W_{hy}$) ; // tính theo công thức (12) và (3)

  end

end

Hình 2. *Giải thuật huấn luyện lan truyền thẳng*

Giải thuật 2: Giải thuật huấn luyện pha lan truyền ngược

$dh_{m_x+1+m_y}^{(1..L)}, dc_{m_x+1+m_y}^{(1...L)} \leftarrow 0$; *//khởi tạo gradient các trạng thái và ô nhớ*

$dT_{lstm}^{(1..L)}, dW_e, dW_{hy} \leftarrow 0$; *//khởi tạo gradient trọng số mô hình*

for $t = (m_x + 1 + m_y) \rightarrow 1$ **do**

  if $t == m_x$ **then**

    $dW_e^{decoder}, dT_{lstm}^{decoder} \leftarrow dW_e, dT_{lstm}^{(1..L)}$; *//cập nhật gradient cho decoder*

    $dT_{lstm}^{(1..L)}, dW_e \leftarrow 0$;

  end

  //dự đoán từ phía decoder

  if $t \geq (m_x + 1)$ **then**

    $dh, dW \leftarrow Predict(s_{t+1}, p_t, h_t^{(L)})$; *// tính gradient theo công thức từ (2) - (4)*

    $dh_t^{(L)} \leftarrow dh_t^{(L)} + dh$; *//tính các gradient theo chiều thẳng đứng*

    $dW_{hy} \leftarrow dW_{hy} + dW$;

  end

  // tính gradient trên nhiều lớp LSTM

  for $l = L \rightarrow 1$ **do**

    $dh_{t-1}^{(l)}, dc_{t-1}^{(l)}, dx, dT \leftarrow grad(dh_t^{(l)}, dc_t^{(l)}, h_{t-1}^{(l)}, h_t^{(l-1)})$; *//tính theo công thức từ (15) – (24)*

    $dh_t^{(l-1)} \leftarrow dh_t^{(l-1)} + dx$; *//tính các gradient theo chiều thẳng đứng*

    $dT_{lstm}^{(l)} \leftarrow dT_{lstm}^{(l)} + dT$;

  end

  Cập nhật dW_e ứng với từ được chọn

end

$dW_e^{encoder}, dT_{lstm}^{encoder} \leftarrow dW_e, dT_{lstm}^{(1..L)}$; *//Lưu lại gradients phía encoder*

Hình 3. *Giải thuật huấn luyện lan truyền ngược*

Bảng 2. *Một số mẫu câu trong bản dịch từ hệ thống thực nghiệm*

STT	Câu tiếng Nhật/Câu dịch đúng	Câu tiếng Việt dịch từ thực nghiệm
1	アイルランドは、現在、過去三度、対戦するたびにイングランドを負かした。	ireland hiện đã đánh_bại anh trong ba lần sau khi trận đấu này .
	ireland đánh_bại anh trong ba lần gần nhất mà hai bên đã gặp nhau .	
2	メキシコは、地域に熱帯暴風雨警報を発令した。	mexico đã sắp_xếp một báo_động về bão nhiệt_đới .
	mexico đã thông_báo về việc đề_phòng cơn bão nhiệt_đới trong toàn khu_vực .	
3	ハリケーン・ダニエルは、2010 の大西洋ハリケーン・シーズンの2 個目のハリケーンである。	theo bão siêu bão , có hai cơn bão lớn trong mùa bão bão nhiệt_đới của năm 2010 .
	bão_danielle là cơn bão thứ_hai trong mùa bão đại_tây_dương năm 2010 .	
4	AOL スポークスマン、アンドリュー・ウェインスタインは、これらの問題が来月利用できるアップグレードにおいて対処されると主張する。	<unk> , người thẳng_thắn , khẳng_định vấn_đề này là được giải_quyết bằng việc nâng_cấp vào tháng tới .
	phát_ngôn viên của aol andrew_weinstein tuyên_bố rằng những vấn_đề này sẽ được giải_quyết trong cuộc nâng_cấp vào tháng tới .	

TÀI LIỆU THAM KHẢO

1. Michael Auli, Michel Galley, Chris Quirk, Geoffrey Zweig (2013), "*Joint Language and Translation Modeling with Recurrent Neural Networks*" EMNLP 2013, pp.1044-1054.

2. Bengio, Yoshua, et al. (2003), "*A neural probabilistic language model*", The Journal of Machine Learning Research 3, pp.1137-1155.

3. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio (2014), "*Neural machine translation by jointly learning to align and translate*", In Proceedings of the 4th International Conference on Learning Representations (ICLR 2015), CoRR abs/1409.0473.

4. Hutchins, John (2005), "*The history of machine translation in a nutshell*",

http://www.hutchinsweb.me.uk/Nutshell-2005.pdf.
5. K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio (2014), "*Learning phrase representations using RNN encoder-decoder for statistical machine translation*", EMNLP 2014, pp. 1724-1734.
6. Koehn, Philipp (2009), "*Statistical machine translation*", *Cambridge University Press*, ISBN-13: 978-0521874151
7. Nguyen M.e., Tanaka Tomoki, Ikeda Takashi (2006), "*Translation of Structure [N I no N2] in Japanese-Vietnamese Machine Translation*", Journal of natural language processing, voU3, no.2, pp.145-169.
8. Papineni, Kishore, et al (2001), "*BLEU: a method for automatic evaluation of machine translation*", Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311–318.
9. Philip Arthur, Graham Neubig, and Satoshi Nakamura (2016), "*Incorporating discrete translation lexicons into neural machine translation*", In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1557-1567.
10. R.Miikkulainen and M.G. Dyer (1991), "*Natural language processing with modular neural networks and distributed lexicon*", Cognitive Science, vo15, no3, pp.343-399.
11. Sutskever, Ilya, Oriol Vinyals, and Quoc VV Le (2014), "*Sequence to sequence learning with neural networks*", Advances in Neural Information Processing Systems (NIPS 2014), pp.3104-3112.
12. Lương Minh Thang (2016), Ph.D. Neural Machine Translation, https://github.com/lmthang/thesis.
13. Thi-Ngoc-Diep DO, Masao UTIY AMA, Eiichiro SUMIT A (2015), "*Machine Translation from Japanese and French to Vietnamese, the Difference among Language Families*", International Conference on Asian Language Processing, IALP 2015, pp.17-20.
14. Vo Ho Bao Khanh, Shun Ishizaki (2010), "*Japanese–Vietnamese compound noun translation*", In Proceedings of the 16th Annual Meeting of the Association for Natural Language Processing, pp.896-899
15. Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals (2014), "*Recurrent neural network regularization*", CoRR abs/1409.2329.
16. https://github.com/tensorflow/nmt, truy cập ngày 15/10/2017
17. https://translate.google.com/, truy cập ngày 15/10/2017

SUMMARY

JAPANESE-VIETNAMESE MACHINE TRANSLATION USES RECURRENT NEURAL NETWORK MODEL

Ngo Thi Vinh[1*], Nguyen Phuong Thai[2]
[1]Information and Communication Technology Thai Nguyen University
[2]University of Engineering and Technology – Viet Nam University

Neural network models using deep learning has brought great success when applied in the field of object recognition, speech recognition. For the past three years, this model has been applied to the various tasks of natural language processing such as machine translation, phrase detection, and extraction of words represented by vectors. In this area, machine translation using neural network has initially achieved better results than the traditional linear statistical model. In this paper, we apply neural network model with deep learning technique for machine translation on the task of Japanese-Vietnamese translation and evaluate the results obtained through BLEU metric as 7,7 point. We also point out the shortcomings to be solved to improve the quality of translation system in the future.

Key words: *natural language processing; statistical machine translation; Japanese-Vietnamese machine translation; recurrent nerural network; deep learning*

Ngày nhận bài: 07/11/2017; Ngày phản biện: 18/12/2017; Ngày duyệt đăng: 05/3/2018

* *Tel: 0987 706830, Email: ntvinh@ictu.edu.vn*