Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XIV về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), TP. HCM, ngày 23-24/12/2021
DOI: 10.15625/vap.2021.0091

CHUYỂN TỰ CHỮ NÔM BẰNG TIẾP CẬN DỊCH MÁY MẠNG NEURAL ĐA NGỮ

Nguyễn Hồng Bửu Long[1], Trang Minh Chiến[1], Nguyễn Thế Hữu[2], Đinh Điền[1]

[1]Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh

[2]Khoa công nghệ thông tin, Trường Đại học Công nghiệp thực phẩm Thành phố Hồ Chí Minh

nhblong@fit.hcmus.edu.vn, chientrangminh@gmail.com, huunt@hufi.edu.vn, ddien@fit.hcmus.edu.vn

TÓM TẮT: *Chữ Nôm là chữ viết được sử dụng trong gần một thế kỷ để ghi chép nhiều tác phẩm văn học, lịch sử, y học,… của dân tộc ta. Để khai thác các nguồn tư liệu trên, nhiều phương pháp đã được sử dụng để xây dựng hệ thống chuyển tự tự động từ chữ Nôm sang chữ Quốc Ngữ, trong đó nổi bật nhất là phương pháp dịch máy mạng neural. Tuy nhiên, việc áp dụng các phương pháp dịch máy mạng neural được sử dụng còn gặp nhiều khó khăn bởi số lượng hạn chế của ngữ liệu song ngữ Nôm - Quốc Ngữ. Trong bài báo này, chúng tôi đề xuất phương pháp dịch máy mạng neural đa ngữ cho bài toán chuyển tự tự động chữ Nôm sang chữ Quốc Ngữ. Với phương pháp được đề xuất, hệ thống chuyển tự có thể tận dụng các đặc trưng tương đồng giữa tiếng Việt và các ngôn ngữ khác có nhiều ngữ liệu, từ đó giúp cải thiện chất lượng chuyển tự. Mô hình dịch máy mạng neural đa ngữ của chúng tôi gồm các bộ mã hóa, giải mã cho từng ngôn ngữ, được kết nối với nhau bằng một bộ liên kết ngôn ngữ với chức năng tận dụng các đặc trưng riêng của từng ngôn ngữ để phát triển thành đặc trưng độc lập với các ngôn ngữ. Kết quả thực nghiệm cho thấy mô hình đạt được sự cải thiện về chất lượng chuyển tự so với mô hình dịch máy mạng neural song ngữ.*

Từ khóa: *Chuyển tự tự động, chữ Nôm, chữ Quốc Ngữ, học sâu, dịch máy mạng neural đa ngữ.*

I. GIỚI THIỆU

Chuyển tự là bài toán thay thế các đơn vị của một hệ thống chữ viết bằng các đơn vị tương ứng của một hệ thống chữ viết khác, trong cùng một ngôn ngữ. Ví dụ, "さくら" trong hệ chữ Hiragana của tiếng Nhật có vị tự "さ" được chuyển tự thành "**sa**", vị tự "く" được chuyển thành "**ku**" và "ら" được chuyển thành "**ra**", kết quả cuối cùng trong hệ chữ Latin là "**sa-ku-ra**" (hoa anh đào). Bài toán chuyển tự thường có thể được thực hiện một cách tự động bằng phương pháp tra bảng vì sự tương đồng 1 - 1 giữa các hệ chữ viết như Hiragana - Latin của tiếng Nhật Bản hay Kirin - Latin của tiếng Nga. Tuy nhiên, phương pháp trên lại không thể được áp dụng hoàn toàn tự động giữa chữ Nôm và chữ Quốc Ngữ bởi tính đa trị của chữ Nôm, như chữ Nôm 劍có thể được chuyển tự thành sáu chữ Quốc Ngữ tương ứng: *chém, ghém, gươm, kém, kiếm* và *sớm*. Vì vậy, có nhiều nghiên cứu đã được tiến hành để tìm ra phương pháp khác.

Những năm gần đây, phương pháp dịch máy mạng neural đa ngữ được áp dụng hiệu quả cho các ngôn ngữ ít tài nguyên như các ngôn ngữ Indo-Aryan [1] và tiếng Việt [2], bằng cách tận dụng nguồn ngữ liệu từ các ngôn ngữ có nhiều tài nguyên như tiếng Anh và tiếng Đức. Vì vậy, trong bài báo, chúng tôi đề xuất sử dụng phương pháp dịch máy mạng neural đa ngữ cho bài toán chuyển tự tự động chữ Nôm sang chữ Quốc Ngữ. Mô hình được đề xuất sử dụng bộ mã hóa và giải mã riêng cho từng ngôn ngữ, kết hợp với biểu diễn ngôn ngữ và biểu diễn vị trí nhận thức ngôn ngữ để tăng cường đặc trưng của riêng từng ngôn ngữ khi huấn luyện. Đồng thời, chúng tôi sử dụng tầng liên kết ngôn ngữ được chia sẻ giữa các cặp ngôn ngữ huấn luyện để trích xuất đặc trưng độc lập ngôn ngữ. Các thực nghiệm khác nhau cũng thực hiện để phân tích ảnh hưởng của tầng liên kết ngôn ngữ lên các mô hình khác nhau. Kết quả thực nghiệm cho thấy phương pháp đề xuất hiệu quả hơn mô hình dịch máy mạng neural song ngữ về chỉ số BLEU.

Bài báo được trình bày với cấu trúc như sau: Phần II sẽ giới thiệu về tổng quan lý thuyết và các công trình liên quan. Trong Phần III, chúng tôi trình bày về hệ thống được đề xuất. Phần IV chúng tôi sẽ trình bày các thực nghiệm, so sánh kết quả giữa các hệ thống và phân tích các kết quả thu được. Cuối cùng, Phần V sẽ trình bày các kết luận.

II. TỔNG QUAN VỀ LÝ THUYẾT

A. Chuyển tự tự động từ chữ Nôm sang chữ Quốc Ngữ

Chữ Nôm bắt đầu được số hóa từ những năm 1990 bởi nhiều nhà nghiên cứu. Nhờ vào việc số hóa mà những chữ Nôm phổ biến đã được mã hóa thành công trong Unicode/ISO 10646, tạo tiền đề cho sự hình thành và phát triển các công cụ hỗ trợ cũng như nghiên cứu chữ Nôm về sau.

Một trong những công cụ đầu tiên được xây dựng nhờ vào việc số hóa chính là bộ gõ chữ Nôm với *Việt Hán Nôm 2002* và *Hanasoft 3.0* là hai công cụ nổi bật nhất. Ngoài chức năng gõ chữ Nôm, hai công cụ còn hỗ trợ tra cứu chữ Nôm và chữ Hán. Trang web của *Hội Bảo tồn di sản chữ Nôm Việt Nam* được chính các nhà nghiên cứu từng thực hiện số hóa chữ Nôm thành lập vào năm 1999, cung cấp tra cứu giữa chữ Nôm - chữ Quốc Ngữ và chữ Nôm - chữ Hán. Ngoài ra, trang web còn lưu trữ các tài liệu văn bản điện tử chữ Nôm, với phần lớn là ảnh của chữ viết tay.

Các nghiên cứu về bài toán chuyển tự tự động chữ Nôm sang chữ Quốc Ngữ hiện nay được tiếp cận theo hai hướng: dịch máy thống kê và dịch máy mạng neural. Đinh Điền [3] áp dụng phương pháp dịch máy thống kê kết hợp với ngữ liệu đơn ngữ của chữ Quốc Ngữ để huấn luyện và tinh chỉnh mô hình ngôn ngữ, nhờ đó mà đạt được kết quả rất tốt. Ngoài ra, nghiên cứu còn trình bày các phân tích, thống kê về chữ Nôm. Ngoài ra, trang web *chunom.org* cũng cung cấp một công cụ tương tự, Nôm Converter, cũng là hệ thống áp dụng dịch máy thống kê để chuyển tự tự động qua lại giữa chữ Nôm và chữ Quốc Ngữ. Tuy nhiên, số lượng dữ liệu huấn luyện cho hệ thống ít hơn dữ liệu huấn luyện của Đinh Điền [3] và tác giả cũng không có mô tả cụ thể về hệ thống. Dịch máy mạng neural cũng được sử dụng cho bài toán chuyển tự tự động với nghiên cứu của Đinh Điền [4]. Mô hình dịch máy mạng neural được huấn luyện trên lượng ngữ liệu song ngữ hơn 6.000 cặp câu và được áp dụng nhiều phương pháp chuẩn hóa để cải thiện chất lượng mô hình trong hoàn cảnh ít tài nguyên huấn luyện. Vì hệ thống dịch máy mạng neural cần dữ liệu lớn để có thể hoạt động hiệu quả, nên chất lượng của mô hình còn thấp so với hệ thống dịch máy thống kê được huấn luyện trên cùng số lượng ngữ liệu.

B. Tổng quan về chữ Nôm

Nguồn gốc của chữ Nôm đến nay vẫn còn được giải thích bởi nhiều giả thuyết khác nhau, thế nhưng các nhà nghiên cứu đều thống nhất rằng, chữ Nôm được hình thành trong giai đoạn từ thế kỷ thứ X (khoảng năm 938) đến thế kỷ thứ XII và được sử dụng cho đến năm 1945 [10]. Trong suốt giai đoạn tồn tại, chữ Nôm đã được sử dụng phổ biến với nhiều công dụng, từ việc được các sĩ tử sử dụng để làm thơ quốc âm, đến việc được dùng để ghi chép các văn kiện hành chính dưới thời Tây Sơn trong suốt hơn 20 năm.

Chữ Nôm được người Việt tạo ra dựa trên cơ sở mượn chữ Hán để ghi âm tiết của tiếng Việt và là âm tự biểu âm kiêm biểu ý [11]. Trên cơ sở chữ Nôm là âm tự biểu âm kiêm biểu ý, chữ Nôm có thể được phân làm hai loại: chữ Nôm được cấu tạo theo phương thức biểu âm và chữ Nôm được cấu tạo theo phương thức biểu ý. Chữ Nôm biểu ý là các chữ được mượn hoặc cải tiến hoặc kết hợp từ chữ Hán, bỏ qua yếu tố âm đọc, ví dụ chữ *trời* (𡗶) được tạo thành từ hai chữ *thiên* (天) và *thượng* (上) trong tiếng Hán hay chữ *chữ* (字) là sự kết hợp của hai chữ *tự* (字). Một số chữ Nôm biểu ý được hình thành từ sự kết hợp một bộ thủ với một chữ Hán (chữ *quạt,* 搧 là sự kết hợp của 扌 và 扇). Chữ Nôm biểu âm được tạo ra tương tự như chữ Nôm biểu ý, nhưng yếu tố âm đọc được giữ lại: Nôm tự biểu âm được tạo ra bằng việc kết hợp một yếu tố về nghĩa và một yếu tố về âm tự chữ Hán [12]. Nôm tự 𠄩 (số ba) được tạo từ việc kết hợp yếu tố về âm là chữ Hán 巴có âm pinyin là /bā/ và yếu tố về nghĩa là chữ Hán 三có nghĩa là số ba.

Vì phần lớn chữ Nôm là chữ biểu âm, được tạo nên từ việc kết hợp của nghĩa và âm, có nhiều trường hợp một Nôm tự có thể được ánh xạ với nhiều hơn một chữ Quốc Ngữ. Điều này có thể được giải thích bằng nhiều nguyên nhân, trong đó có nguyên nhân là sự khác nhau về số lượng thanh điệu giữa tiếng Việt (sáu thanh điệu) và tiếng Hán (bốn thanh điệu) hay nguyên nhân đến từ sự thay đổi cách sử dụng của các triều đại phong kiến khác nhau. Một ví dụ tiêu biểu là Nôm tự 味mang ý nghĩa là *mùi* vào trước thời nhà Đường nhưng đã được bổ sung thêm nghĩa là *vị* kể từ sau triều đại [11]. Bằng cách phân tích từ điển song ngữ Nôm - Quốc Ngữ với 22.264 Nôm tự, Đinh Điền [12] đã thống kê số lượng chữ Nôm có một chữ Quốc Ngữ tương ứng và số lượng chữ Nôm có từ hai chữ Quốc Ngữ tương ứng và nhận thấy, chỉ có hơn 20% chữ Nôm có nhiều hơn một chữ Quốc Ngữ.

C. Dịch máy mạng neural

Gọi $x = (x_1, \ldots, x_n)$ là câu nguồn và $y = (y_1, \ldots, y_m)$ là câu đích tương ứng, hệ thống dịch máy mạng neural hoạt động theo kiến trúc mã hóa - giải mã [13], trong đó, bộ mã hóa tạo ra biểu diễn h từ câu nguồn và sẽ được truyền cho bộ giải mã để tạo ra câu đích tương ứng. Quá trình trên phân tách trực tiếp xác suất $p(y \vee x)$:

$$p(y \vee x) = \prod_{t=1}^{m} p\left(y_t \vee y_{1 \leqslant i < t}, h\right) \quad (1)$$

với y_i là từ thứ i trong câu. Xác suất $p\left(y_t \vee y_{1 \leqslant i < t}, h\right)$ được tính qua công thức sau:

$$p\left(y_t \vee y_{1 \leqslant i < t}, h\right) = softmax\left(f(h, c_t)\right) \quad (2)$$

trong đó $h = \left(h_1, \ldots, h_n\right)$ là biểu diễn của câu nguồn được tạo bởi bộ mã hóa và c_t là véctơ ngữ cảnh được tính toán từ biểu diễn của câu nguồn và véctơ gióng hàng $a_{i,t}$:

$$a_{i,t} = softmax\left(e_{i,t}\right) = \frac{exp(e_{i,t})}{\sum_{j=1}^{n} e_{j,t}} \quad (3)$$

$$c_t = \sum_{i=1}^{n} a_{i,t}\, h_i \quad (4)$$

với véctơ điểm gióng hàng $e_{i,t}$ được dùng để đánh giá sự tương đồng giữa hai biểu diễn của câu nguồn và đích.

Mục tiêu huấn luyện của một mô hình dịch máy nơron là cực đại hóa log-likelihood có điều kiện của câu đích khi cho trước câu nguồn. Lần lượt gọi $\theta_{enc}, \theta_{dec}, \theta_{attn}$ là các tham số của bộ mã hóa, bộ giải mã và khi tính toán véctơ gióng hàng, $D = \{(x, y)\}$ là tập ngữ liệu huấn luyện và m là chiều dài câu đích, hàm mất mát của một hệ thống dịch máy mạng neural có dạng như sau:

$$L_t(D; \theta) = \sum_{d=1}^{|D|} \sum_{t=1}^{m} log\, p\left(y_t \vee y_{1 \leqslant i < t}, x; \theta_{enc}, \theta_{dec}, \theta_{attn}\right) \quad (5)$$

Bộ mã hóa và giải mã có thể được cài đặt bằng nhiều kiến trúc mạng neural khác nhau như mạng neural hồi quy (LSTM, GRU) [14], mạng neural tích chập [15] hoặc cơ chế tập trung kết hợp mạng neural lan truyền thẳng [16].

D. Mô hình Transformer

Transformer [16] là kiến trúc mã hóa - giải mã [13] chỉ bao gồm cơ chế tập trung và mạng neural lan truyền thẳng. Bộ mã hóa của Transformer bao gồm nhiều lớp giống nhau, mỗi lớp bao gồm hai lớp con. Lớp con thứ nhất là cơ chế tự tập trung nhiều đầu, và lớp thứ hai là một mạng lan truyền thẳng đơn giản. Sau mỗi lớp con là một liên kết thặng dư (residual connection), theo sau là một thao tác chuẩn hóa lớp (layer normalization). Bộ giải mã cũng bao gồm nhiều lớp tương tự nhau, mỗi lớp có ba lớp con. Ngoài hai lớp tương tự như trong bộ mã hóa, bộ giải mã còn sử dụng thêm một lớp tập trung nhiều đầu để kết hợp kết quả đầu ra của bộ mã hóa với biểu diễn ẩn của bộ giải mã. Liên kết thặng dư và chuẩn hóa lớp cũng được sử dụng sau mỗi lớp con. Có một điểm khác biệt giữa cơ chế tự tập trung ở bộ giải mã và bộ mã hóa, cơ chế tự tập trung ở bộ giải mã có sử dụng mặt nạ để ngăn không cho mô hình tập trung vào những vị trí của câu đích sau thời điểm hiện tại. Hay nói cách khác, bộ giải mã chỉ được tập trung vào những từ của câu đích xuất hiện từ thời điểm quá khứ tính đến vị trí hiện tại, vì bộ giải mã làm nhiệm vụ phát sinh từng từ trong mỗi lượt dựa trên những từ xuất hiện trước.

Mô hình Transformer không xử lý dữ liệu theo tính tuần tự nên biểu diễn vị trí được sử dụng cùng với biểu diễn câu để làm đầu vào cho mô hình, giúp Transformer kiểm soát được tính tuần tự của câu. Mô hình Transformer được sử dụng để khảo sát trong bài báo vì mặc dù tầng liên kết ngôn ngữ được đề xuất bởi Raganato [24], tác giả chỉ thực hiện các thực nghiệm trên mô hình dịch máy sử dụng mạng neural hồi quy.

E. Dịch máy mạng neural đa ngữ

Dịch máy mạng neural đa ngữ đã được nghiên cứu nhiều trong các công trình của Dong [5], Luong [6] và Johnson [7] và được kiểm chứng bằng nhiều thực nghiệm về tính hiệu quả khi áp dụng cho các bài toán dịch máy ít tài nguyên ([8], [9]).

Dịch máy mạng neural đa ngữ được xem là mô hình học máy thực hiện bài toán đa tác vụ với một mức độ chia sẻ tham số nhất định. Các tác vụ trong một mô hình dịch máy mạng neural đa ngữ chính là dịch các cặp ngôn ngữ khác nhau được huấn luyện đồng thời cùng lúc. Mô hình dịch máy mạng neural đa ngữ có thể được coi là mô hình học máy thực hiện cùng lúc nhiều tác vụ nhất với số lượng cặp ngôn ngữ có thể lên đến 100 cặp [17].

Nếu gọi L là số lượng cặp ngôn ngữ được huấn luyện đồng thời thì hàm mất mát của một mô hình dịch máy mạng neural đa ngữ có dạng như sau:

$$L_t(D;\theta) = \sum_{l=1}^{L} \sum_{d=1}^{|D|} \sum_{t=1}^{m} log\, p\left(y_t^l \vee y_{1\leqslant i<t}^l, x^l; \theta_{enc}^l, \theta_{dec}^l, \theta_{attn}^l\right) \quad (6)$$

Kiến trúc mô hình dịch máy mạng neural đa ngữ có thể được phân làm ba loại: mô hình chia sẻ hạn chế, mô hình chia sẻ có kiểm soát và mô hình chia sẻ hoàn toàn. Mô hình chia sẻ hạn chế [18] chỉ chia sẻ tầng tập trung giữa các ngôn ngữ và là hướng tiếp cận ít được chú ý nhất. Mô hình chia sẻ có kiểm soát [19], [20] là những mô hình đa ngữ được thiết kế để chia sẻ các thành phần khác nhau, đồng thời tách rời các thành phần khác để căn bằng giữa việc học đặc trưng chung giữa các ngôn ngữ và các đặc trưng riêng của mỗi ngôn ngữ. Mô hình chia sẻ hoàn toàn [7], [21], [9] là các mô hình mà trong đó, tất cả thành phần được chia sẻ giữa tất cả ngôn ngữ huấn luyện.

Mô hình dịch máy mạng neural đa ngữ được trình bày trong bài báo gồm các cấu trúc và thành phần đã được đề xuất trong các nghiên cứu trước. Tuy nhiên, việc kết hợp tầng liên kết ngôn ngữ với mô hình Transformer chưa được thực hiện bởi các bài báo trong mục Tài liệu tham khảo.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Chúng tôi đề xuất mô hình dịch máy mạng neural đa ngữ với tầng liên kết ngôn ngữ kết nối bộ mã hóa và bộ giải mã có khả năng tạo ra biểu diễn có kích thước cố định. Kiến trúc mô hình được minh họa trên Hình 1.

A. Tầng liên kết ngôn ngữ và bộ mã hóa, giải mã riêng cho từng ngôn ngữ

Tầng liên kết ngôn ngữ, đề xuất bởi Raganato [24], được chia sẻ giữa các cặp ngôn ngữ nhận biểu diễn $h = \left(h_1, \ldots, h_n\right) \in R^{d_h x n}$ được tạo ra bởi bộ mã hóa. Biểu diễn có chiều dài không cố định, tùy thuộc vào chiều dài của câu nguồn. Tầng liên kết ngôn ngữ biến đổi biểu diễn thành một biểu diễn có chiều cố định $M \in R^{d_h x k}$ tập trung vào k thành phần khác nhau của một câu ([22], [23], [24]) bằng cách sử dụng cơ chế tập trung:

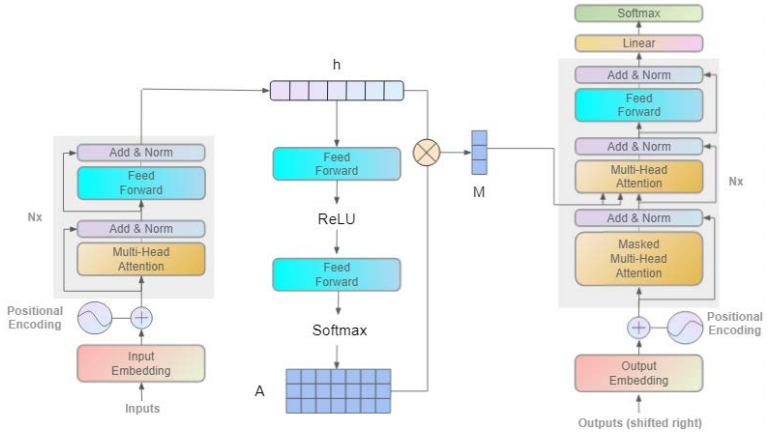
$$A = softmax\left(W_2 ReLU(W_1 h^T)\right) \quad (7)$$

$$M = Ah \quad (8)$$

trong đó, $W_1 \in R^{d_w x d_h}$ và $W_2 \in R^{k x d_w}$ là các ma trận trọng số với d_w là siêu tham số và k là số lượng đầu của tầng tập trung (attention heads). Mỗi một cột trong ma trận M tập trung vào một phần khác nhau của một câu, vì vậy mà tất cả các cột sẽ biểu diễn cho ngữ nghĩa của toàn bộ câu [22], [23]. Hai ma trận trọng số W_1 và W_2 được chia sẻ giữa các véctơ biểu diễn của các ngôn ngữ khác nhau, nhờ đó mà hai ma trận có thể trích xuất được đặc trưng của các ngôn ngữ để tạo ra đặc trưng chung.

Chúng tôi thêm tầng liên kết ngôn ngữ vào giữa bộ mã hóa và bộ giải mã. Véctơ biểu diễn có kích thước không cố định được tạo ra bởi bộ mã hóa sẽ được đưa vào tầng liên kết ngôn ngữ, để tạo ra véctơ có kích thước cố định và được sử dụng bởi bộ giải mã để tạo ra câu đích. Tầng liên kết ngôn ngữ có vai trò "học" những tri thức chung giữa các ngôn ngữ khác nhau.

Để khai thác những tri thức riêng của từng ngôn ngữ, chúng tôi sử dụng bộ mã hóa và giải mã riêng cho từng ngôn ngữ.



Hình 1. Kiến trúc mô hình đa ngữ được đề xuất. Đầu ra của bộ mã hóa, h, được đưa qua hai mạng neural lan truyền thẳng để tạo ra ma trận A. Sau đó, ma trận A và véctơ h được kết hợp để tạo ra biểu diễn có kích thước cố định M và được dùng làm đầu vào cho bộ giải mã

B. Biểu diễn ngôn ngữ và biểu diễn vị trí nhận thức ngôn ngữ

Các hệ thống dịch máy mạng neural đa ngữ thường thêm vào bộ từ vựng của mô hình một token đặc biệt để biểu diễn cho từng ngôn ngữ [7], [17], [25]. Các token được thêm vào đầu mỗi câu nguồn hoặc câu đích để giúp bộ giải mã dịch đúng ngôn ngữ, đồng thời tăng cường các đặc trưng riêng của từng ngôn ngữ cho hệ thống đa ngữ. Tuy nhiên, các đặc trưng riêng có thể bị yếu đi trong quá trình huấn luyện vì phải lan truyền qua nhiều tầng khác nhau của bộ mã hóa và giải mã. Để giải quyết vấn đề, chúng tôi xây dựng biểu diễn riêng cho từng ngôn ngữ và kết hợp với biểu diễn của câu đầu vào trước khi đưa vào bộ mã hóa.

Mô hình Transformer truyền thống sử dụng biểu diễn vị trí cố định [16] cho mọi loại ngôn ngữ khác nhau, thế nhưng, những ngôn ngữ khác nhau có thể có sự khác biệt về cấu trúc câu, vì vậy mà các ngôn ngữ cần có các biểu diễn vị trí khác nhau. Ý tưởng cũng được các nghiên cứu của Wang [26] áp dụng cho hệ thống dịch máy đa ngữ của họ. Chúng tôi thêm vào biểu diễn vị trí của Transformer một hệ số $W_L L_{emb}$ trong đó W_L là ma trận trọng số và L_{emb} là biểu diễn ngôn ngữ.

IV. CÁC THỰC NGHIỆM VÀ KẾT QUẢ

A. Dữ liệu

Chúng tôi khảo sát mô hình đa ngữ trên bốn cặp ngôn ngữ: Anh - Việt (en-vi), Anh - Hoa (en-zh), Hoa - Việt (zh-vi) và Nôm - Việt (nôm-vi). Ngữ liệu en-vi được lấy từ dữ liệu IWSLT'15 [27], chúng tôi chỉ sử dụng 50.000 cặp câu trong ngữ liệu huấn luyện. Ngữ liệu en-zh được cung cấp bởi IWSLT 2017 [28] và hai bộ ngữ liệu song ngữ còn lại, en-zh và zh-vi, được cung cấp bởi Trung tâm CLC. Dữ liệu được chia thành ba tập dữ liệu và số lượng cặp câu trong từng tập được thể hiện trong Bảng 1. Tập dữ liệu huấn luyện dùng để huấn luyện mô hình, trong khi tập dữ liệu tinh chỉnh dùng để đánh giá hiệu quả các mô hình trong quá trình huấn luyện và được dùng để so sánh, lựa chọn mô hình tốt nhất. Tập dữ liệu kiểm tra chỉ được sử dụng một lần trên mô hình tốt nhất để đánh giá chỉ số BLEU cuối cùng và số liệu được ghi vào các Bảng 2 và 3.

Bảng 1. Thống kê số lượng cặp câu được phân chia cho từng tập dữ liệu

	Huấn luyện	Tinh chỉnh	Kiểm tra
en-vi	50.000	1.552	1.267
en-zh	100.000	6.925	2.502
zh-vi	32.060	1.780	1.780
nôm-vi	6.348	786	786

Lượng dữ liệu huấn luyện giữa các cặp ngôn ngữ không cân bằng, số lượng ngữ liệu song ngữ của cặp Nôm - Việt nhỏ hơn rất nhiều so với ba cặp còn lại. Nhờ vậy, cặp Nôm - Việt có thể tận dụng được tri thức từ các ngôn ngữ có nhiều tài nguyên hơn.

Dữ liệu được tiền xử lý bằng Moses [29] và với tiếng Việt và tiếng Hoa, chúng tôi lần lượt sử dụng công cụ RDRSegmenter [30] và Jieba để tách từ. Sau đó, dữ liệu huấn luyện và tinh chỉnh của các cặp ngôn ngữ được sử dụng để học mô hình byte-pair-encoding (BPE) [31] với kích thước bộ từ vựng là 20.000. Mô hình BPE sau đó được sử dụng để mã hóa các tập dữ liệu.

B. Các thông số kỹ thuật

Mô hình được cài đặt bằng fairseq-py [32] và các kết quả được đánh giá dựa trên chỉ số BLEU [33].

Đối với mô hình song ngữ nôm-vi, vì ngữ liệu có kích thước nhỏ nên mô hình cũng cần được điều chỉnh. Chúng tôi chọn số tầng của bộ mã hóa và giải mã đều là năm, kích thước mạng lan truyền thẳng trong bộ mã hóa là 512. Số đầu của tầng tập trung ở bộ mã hóa và giải mã cũng được thu nhỏ và bằng hai. Dropout được sử dụng ở tầng giải mã, cơ chế tập trung và hàm kích hoạt với số liệu lần lượt là 0,3; 0,1 và 0,3.

Đối với các mô hình song ngữ của ba cặp ngôn ngữ còn lại và mô hình đa ngữ, chúng tôi sử dụng các siêu tham số trong [16]. Các mô hình sử dụng tầng liên kết ngôn ngữ với siêu tham số $d_w = 1024$.

Mô hình cơ sở là mô hình Transformer có các bộ mã hóa và giải mã riêng cho từng ngôn ngữ, nhưng không có các thành phần gồm tầng liên kết ngôn ngữ, biểu diễn ngôn ngữ và biểu diễn vị trí nhận thức ngôn ngữ.

C. Các thực nghiệm

Các thực nghiệm được tiến hành nhằm phân tích ảnh hưởng của siêu tham số k lên các mô hình khác nhau. Chúng tôi tiến hành ba thực nghiệm sau:

1. Thực nghiệm với mô hình song ngữ.
2. Thực nghiệm với mô hình đa ngữ N-vi (N là số lượng ngôn ngữ).
3. Thực nghiệm với mô hình đa ngữ N-N.

Trong đó, mô hình đa ngữ N-vi có các ngôn ngữ nguồn là en, nôm và zh, mô hình đa ngữ N-N sử dụng bốn cặp ngôn ngữ.

- ***Ảnh hưởng của tầng liên kết ngôn ngữ lên mô hình song ngữ***

Kết quả ở Bảng 2 cho thấy ảnh hưởng của tầng liên kết ngôn ngữ lên mô hình song ngữ. Chúng tôi thực hiện so sánh với bốn giá trị khác nhau của k: $k = 1, 10, 25, 50$. Đối với hai cặp ngôn ngữ zh-vi và nôm-vi, kết quả cao nhất thuộc về mô hình có kích thước $k = 25$, khi giá trị lớn hơn 25, hiệu quả của mô hình bị giảm đi. k là số đầu của cơ chế tập trung trong tầng liên kết ngôn ngữ, giá trị càng lớn thì ma trận M sẽ mã hóa nhiều thông tin hơn từ câu nguồn. Kết quả thực nghiệm cho thấy, kích thước số đầu chỉ nên đạt đến một mức độ nhất định, 25, nếu không có thể làm giảm tính hiệu quả của mô hình song ngữ.

Tuy nhiên, cặp ngôn ngữ en-vi lại cho kết quả khác. Số lượng đầu lớn, 50, thì sẽ giúp tăng hiệu quả của mô hình. Kết quả cũng trùng khớp với các kết quả thực nghiệm của Raganato [24] với chiều dịch là X-en và en-X. Một cách biệt lớn cũng được tìm thấy giữa kết quả của mô hình cơ sở và mô hình có số đầu là 1, điều này cũng được tìm thấy trong thực nghiệm của Raganato [24].

Qua các thực nghiệm, chúng tôi nhận thấy việc sử dụng tầng liên kết ngôn ngữ cho mô hình song ngữ làm giảm đi tính hiệu quả của mô hình. Các kết quả thực nghiệm của Raganato [24] cũng đạt được kết quả tương tự.

Bảng 2. Chỉ số BLEU của các mô hình song ngữ

Ngôn ngữ		Mô hình cơ sở	k = 1	k = 10	k = 25	k = 50
vi	en	13,85	6,06	11,95	11,95	**13,06**
	nôm	60,37	32,47	57,72	**59,84**	58,86
	zh	26,63	15,65	18,08	**18,35**	17,31
en	vi	12,21	6,84	12,56	12,19	**13,27**
nôm		73,36	33,92	70,49	**73,11**	71,79
zh		26,71	17,08	18,79	**19,28**	18,40

- ***Ảnh hưởng của tầng liên kết ngôn ngữ lên các mô hình đa ngữ***

Thực nghiệm về mô hình đa ngữ cho ta thấy được ảnh hưởng giữa các cặp ngôn ngữ khi huấn luyện đồng thời với nhau. Qua bảng kết quả ta nhận thấy, giá trị chỉ số BLEU của cặp nôm-vi được cải thiện đáng kể, trong khi ba cặp en-vi, zh-vi và en-zh lại bị giảm chỉ số BLEU. Tác động tích cực lên ngôn ngữ ít tài nguyên và tác động tiêu cực lên

ngôn ngữ nhiều tài nguyên của mô hình dịch máy mạng neural đa ngữ cũng được tìm thấy trong các thực nghiệm của Arivazhagan [9]. Chiều dịch zh-vi chịu ít ảnh hưởng tiêu cực hơn chiều dịch en-vi, một trong những lý do là vì tiếng Hán có nhiều chữ trùng với chữ Nôm hơn tiếng Anh.

Kết quả cuối cùng của chiều nôm-vi là **76,12**, cao hơn mô hình dịch máy mạng neural song ngữ 75,80 [4].

Bảng 3. Kết quả chỉ số BLEU trên các mô hình đa ngữ khác nhau

Ngôn ngữ		N-vi					N-N		
		Mô hình cơ sở	k = 1	k = 10	k = 25	k = 50	Mô hình cơ sở	k = 1	k = 50
en	vi	5,14	1,10	2,21	**2,23**	2,02	19,25	5,29	12,46
nôm		74,99	36,51	**72,58**	72,22	71,49	74,58	38,77	**76,12**
zh		12,33	8,01	9,44	**9,69**	9,65	33,99	24,42	29,70
en	zh						10,29	3,31	6,41

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày một phương pháp dịch máy mạng neural đa ngữ áp dụng cho bài toán chuyển tự chữ Nôm sang chữ Quốc Ngữ. Mô hình của chúng tôi gồm các bộ mã hóa và giải mã riêng cho từng ngôn ngữ, kết hợp với tầng liên kết ngôn ngữ để trích xuất đặc trưng độc lập ngôn ngữ. Ngoài ra, biểu diễn ngôn ngữ và biểu diễn vị trí nhận thức ngôn ngữ cũng được sử dụng để tăng cường đặc trưng riêng ngôn ngữ. Chúng tôi cũng tiến hành nhiều thực nghiệm để khảo sát ảnh hưởng của tầng liên kết ngôn ngữ lên các mô hình song ngữ và đa ngữ. Mô hình đề xuất có kết quả chỉ số BLEU cao hơn mô hình mạng neural song ngữ.

Trong tương lai, chúng tôi sẽ thực hiện các thực nghiệm với các ngữ liệu lớn hơn và khảo sát trên nhiều ngôn ngữ khác. Chúng tôi cũng sẽ tiếp tục cải tiến mô hình để có thể đạt được các kết quả tốt hơn.

VI. LỜI CẢM ƠN

Đề tài được tài trợ bởi Sở Khoa học và Công nghệ Thành phố Hồ Chí Minh.

TÀI LIỆU THAM KHẢO

[1] Vikrant Goyal, Sourav Kumar, Dipti Misra Sharma. “Efficient neural machine translation for low-resource languages via exploiting related languages”. Association for Computational Linguistics, page 162 - 168, 07/2020.

[2] Thi-Vinh Ngo, Phuong-Thai Nguyen, Thanh-Le Ha, Khac-Quy Dinh, Le-Minh Nguyen. “Improving multilingual neural machine translation for low-resource languages: French, English - Vietnamese”. Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, page 55 - 61, Suzhu, China, 2020.

[3] Dien Dinh, Phuong Nguyen, Long H. B. Nguyen. “Transliterating Nom scripts into Vietnamese national Scripts using statistical machine translation”. International Journal of Advanced Computer Science and Applications, 2020.

[4] Đinh Điền, Trang Minh Chiến, Nguyễn Thị Kim Phượng, Nguyễn Hồng Bửu Long. “Chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ theo tiếp cận dịch máy neural”. Hội nghị khoa học quốc gia về “Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin” (FAIR2020), Nha Trang, Việt Nam, 10/2020.

[5] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. “Multi-task learning for multiple language translation”. In Proceedings of ACL 2015, pages 1723 - 1732, 2015.

[6] Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals and Lukasz Kaiser. “Multi-task sequence to sequence learning”. In Proceedings of ICLR 2016.

[7] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean, “Google’s multilingual neural machine translation system: Enabling zero-shot translation”. Transactions of the Association for Computational Linguistics, 5:339-351, 2017.

[8] Vikrant Goyal, Sourav Kumar, Dipti Misra Sharma. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 162 - 168, 2020.

[9] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, Yonghui Wu. “Massively multilingual neural machine translation in the wild: Findings and challenges”. arXiv:1907.05019v1, 2019.

[10] Trần Trọng Dương. “Nguồn gốc, lịch sử và cấu trúc của chữ Nôm từ bối cảnh văn hóa Đông Á”. Viện Nghiên cứu Hán Nôm.

[11] Nguyễn Tuấn Cường. “Nghiên cứu diên cách cấu trúc chữ Nôm qua các văn bản giải âm ‘Kinh Thi’”. Luận văn Tiến sĩ. 2012.

[12] Đinh Điền, Nguyễn Thị Kim Phượng, Diệp Gia Hân, Trần Nguyễn Sơn Thanh. “Chuyển tự tự động từ chữ Nôm sang chữ Quốc Ngữ”. Hội thảo 100 năm chữ Quốc Ngữ, 2019.

[13] Ilya Sutskeyer, Oriol Vinyals, and Quoc V. Le. “Sequence to sequence learning with neural machine translation”. Conference on Neural Information Processing Systems, 2014.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. arXiv preprint arxiv: 1409.0473, 2014.

[15] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin. “Convolutional Sequence to Sequenc Learning”. arXiv:1705.03122v3, 2017.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is all you need”, Conference on Neural Information Processing Systems, 2017.

[17] Sneha Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan and Orhan Firat. "Investigating multilingual NMT representations at scale". arXiv:1909.02197v2, 2019.

[18] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. "Multi-Way, multilingual neural machine translation with a shared attention mechanism". Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, 866-875, 2016.

[19] Devendra Sachan and Graham Neubig. "Parameter sharing methods for multilingual self-attentional translation models". Proceedings of the Third Conference on Machine Translation: Research Papers. Association for Computational Linguistics, Belgium, Brussels, 261-271, 2018.

[20] Graeme Blackwood, Miguel Ballesteros, and Todd Ward. "Multilingual neural machine translation with task-specific attention". Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3112-3122, 2018.

[21] Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. "Multilingual neural machine translation with soft decoupled encoding". Proceedings of International Conference on Learning Representations. New Orleans, 2019.

[22] Lin, Zhouhan, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. "A structured self-attentive sentence embedding". 5th International Conference on Learning Representation, ICLR 2017, Conference Track (Poster), 2017.

[23] Chen, Qian, Zhen-Hua Ling, and Xiaodan Zhu. "Enhancing sentence embedding with generalized pooling". Proceedings of the 27th International Conference on Computational Linguistics, pages 1815 - 1826, Santa Fe, NM, 2018.

[24] Alessandro Raganato, Raul Vázquez, Mathias Creutz and Jorg Tiedemann. "An evaluation of language-agnostic inner-attention-based representations in machine translation". Proceedings of the 4th Workshop on Representation Learning for NLP (ReL4NLP-2019), pages 27 - 32. Florence, Italy, 2019.

[25] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer. "Multilingual Denoising Pre-training for Neural Machine Translation". arXiv:2001.08210v2, 2020.

[26] Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu and Chengquing Zong. "A compact and language-sensitive multilingual translation method". Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1213 - 1223. Florence, Italy, 2019.

[27] Minh-Thang Luong, Christopher D. Manning. "Stanford neural machine translation systems for spoken language domains". International Workshop on Spoken Language Translation. Da Nang, Vietnam, 2015.

[28] M. Cettolo, C. Girardi, and M. Federico. "WIT3: Web Inventory of Transcribed and Translated Talks". Proceeding of EAMT, pp. 261-268, Trento, Italy, 2013.

[29] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. "Moses: Open source toolkit for statistical machine translation". Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, pages 177 - 180. Association for Computational Linguistics, 2007.

[30] Dat Quoc Nguyen and Dai Quoc Nguyen and Thanh Vu and Mark Dras and Mark Johnson. "A fast and accurate Vietnamese word segmenter". Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), pages 2582 - 2587, 2018.

[31] Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units". Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1715 - 1725, Berlin, 2016.

[32] Myle Ott and Sergey Edunov and Alexei Baevski and Angela Fan and Sam Gross and Nathan Ng and David Grangier and Michael Auli. "fairseq: A fast, extensible toolkit for sequence modeling". Proceedings of NAACL-HLT 2019: Demonstrations, 2019.

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: A method for automatic evaluation of machine translation". 40th Annual meeting of the Association for Computational Linguistics. pp. 311-318, 2002.

NÔM-SCRIPTS TRANSLITERATION USING MULTILINGUAL NEURAL MACHINE TRANSLATION APPROACH

Nguyen Hong Buu Long, Trang Minh Chien, Nguyen The Huu, Dinh Dien

ABSTRACT: *Nôm-scripts were used to record many literary, historical, medical documents, etc. for nearly a decade by our ancestor and to exploit these knowledge resource, different approaches have been proposed to build an automatic transliteration system from Nôm-scripts to Vietnamese National Scripts, including neural machine translation. However, applying neural machine translation for Nôm-scripts to Vietnamese transliteration encounters many difficulties due to the constrain of bilingual Nôm-Vietnamese. In this paper, we propose a multilingual neural machine translation for Nôm-scripts to Vietnamese National scripts transliteration. With this approach, the transliteration system could use the common representation between different languages to improve the quality of the transliteration task. Our multilingual neural machine translation has language-specific encoders, decoders connected by an attention bridge that can extract language-dependent representation to create language-independent representation. Results from experiments show improvement over bilingual neural machine translation.*