THÍCH ỨNG MIỀN TRONG DỊCH MÁY NƠ RON CHO CẶP NGÔN NGỮ ANH - VIỆT

Phạm Nghĩa Luân[1], Nguyễn Văn Vinh[2], Nguyễn Huy Hoàng[3]

[1] Trường Đại học Hải Phòng
[2] Trường Đại học Công nghệ, Đại học Quốc Gia Hà Nội
[3] Học viện Quản lý giáo dục

luanpn@dhhp.edu.vn, vinhnv@vnu.edu.vn, huyhoangpvn@gmail.com

TÓM TẮT— *Dịch máy là một trong những hướng nghiên cứu quan trọng trong xử lý ngôn ngữ tự nhiên. Trong những năm gần đây, dịch máy nơ ron đã và đang được nghiên cứu phổ biến hơn trong cộng đồng dịch máy vì hiện tại nó cho chất lượng dịch tốt hơn so với phương pháp dịch máy thống kê truyền thống. Tuy nhiên, dịch máy nơ ron lại cần lượng lớn dữ liệu song ngữ để huấn luyện. Hệ dịch sẽ cho chất lượng bản dịch tốt hơn khi nó được thử nghiệm trong cùng miền với miền dữ liệu mà nó được huấn luyện, ngược lại thì chất lượng bản dịch sẽ bị sụt giảm, mức độ sụt giảm phụ thuộc vào mức độ khác biệt giữa dữ liệu miền huấn luyện và dữ liệu miền thử nghiệm. Hiện nay, các kĩ thuật thích ứng miền cho dịch máy nơ ron đã được công bố chủ yếu được thực hiện trên một số cặp ngôn ngữ phổ biến giàu tài nguyên, và chưa có nhiều nghiên cứu đã được công bố về thích ứng miền trong dịch máy nơ ron cho cặp ngôn ngữ Anh - Việt.*

Trong bài báo này, chúng tôi đề xuất một phương pháp thích ứng miền mới cho dịch máy nơ ron, áp dụng cho cặp ngôn ngữ Anh - Việt. Ý tưởng chính của bài báo là kết hợp dữ liệu đơn ngữ ngoài miền ở ngôn ngữ nguồn (tiếng Anh) với bản dịch của nó ở ngôn ngữ đích (tiếng Việt) để làm dữ liệu huấn luyện hệ dịch. Các thực nghiệm đã chứng minh rằng phương pháp chúng tôi đề xuất dễ thực hiện, khai thác được những ưu điểm của dữ liệu đơn ngữ như luôn có sẵn, chi phí xây dựng thấp và đặc biệt là chất lượng của hệ dịch được và tăng 2,21 điểm BLEU trong thử nghiệm của chúng tôi.

Từ khóa— *Thích ứng miền, dịch máy nơ ron, dịch máy.*

I. GIỚI THIỆU

Mục tiêu của dịch máy là nghiên cứu các phương pháp, kĩ thuật để xây dựng được một hệ thống có thể dịch tự động các câu từ một ngôn ngữ tự nhiên này sang ngôn ngữ khác, đây là một trong những hướng nghiên cứu quan trọng trong trí tuệ nhân tạo, đặc biệt trong xử lý ngôn ngữ tự nhiên. Dịch máy là một nhánh nhỏ của xử lý ngôn ngữ tự nhiên, và vì xử lý ngôn ngữ tự nhiên là lĩnh vực liên ngành giữa khoa học máy tính và ngôn ngữ học, chính đặc điểm đó nên các nghiên cứu về dịch máy có thể chia thành hai nhóm phương pháp chính là các phương pháp dựa trên luật và các phương pháp dựa trên ngữ liệu. Trong số đó, các phương pháp dựa trên ngữ liệu có thể được chia thành các phương pháp dựa trên thống kê và các phương pháp dựa trên ví dụ. Trong những năm gần đây, với sự phát triển của internet, dịch máy đã đạt được những kết quả tốt cả về học thuật và trong công nghiệp.

Gần đây, các nghiên cứu về dịch máy đã dịch chuyển dần từ các phương pháp dịch thống kê *(Statistical Machine Translation)* sang dịch máy nơ ron *(Neural Machine Translation)*, hiện tại đây được coi là một hệ dịch cho chất lượng dịch vượt trội so với các phương pháp truyền thống trước đây. Tuy nhiên, các hệ dịch nơ ron lại yêu cầu nhiều dữ liệu song ngữ hơn để huấn luyện hệ dịch, điều này ít ảnh hưởng tới chất lượng bản dịch của hệ dịch dành cho các cặp ngôn ngữ phổ biến và giàu tài nguyên nhưng nó lại là thách thức lớn đối với các cặp ngôn ngữ có ít tài nguyên.

Thông thường, hệ dịch được huấn luyện trên lượng lớn dữ liệu song ngữ và dữ liệu đơn ngữ của ngôn ngữ đích đối với dịch máy thống kê và dữ liệu song ngữ đối với dịch máy nơ ron, trong bản thân những dữ liệu huấn luyện này có thể bao gồm các chủ đề đồng nhất hoặc không đồng nhất và thường thì mỗi chủ đề đó sẽ có tập các từ thuật ngữ riêng biệt. Chất lượng của bản dịch phụ thuộc rất lớn vào dữ liệu huấn luyện, nếu miền dữ liệu huấn luyện và miền thử nghiệm giống nhau hoặc có sự tương đồng càng lớn thì chất lượng bản dịch thu được sẽ càng tốt so với việc miền dữ liệu dùng để huấn luyện và miền thử nghiệm đặc biệt khác nhau hoặc có ít sự tương đồng hơn. Ví dụ, nếu hệ dịch được huấn luyện với dữ liệu thuộc miền tin tức thì khi dịch các văn bản cũng thuộc miền tin tức sẽ cho chất lượng bản dịch tốt, nhưng nếu đem hệ dịch đó để dịch các văn bản thuộc miền khác với miền tin tức như miền y tế, tin học, luật, v.v... thì chất lượng của bản dịch sẽ bị giảm đột ngột, mức độ giảm tùy thuộc vào mức độ tương đồng của miền dữ liệu dùng để huấn luyện hệ dịch so với miền dữ liệu dùng để thử nghiệm.

Các miền dữ liệu song ngữ trong thực tế thường rất hiếm hoặc bị giới hạn về số lượng, đặc biệt đối với các cặp ngôn ngữ ít phổ biến như ngôn ngữ Anh - Việt, nhất là các miền dữ liệu đặc thù. Để đạt được chất lượng bản dịch tốt nhất thì dữ liệu huấn luyện phải thuộc cùng một miền, cùng một thể loại và cùng một phong cách với miền mà hệ dịch được áp dụng nhưng tế để có được lượng dữ liệu huấn luyện đủ lớn trong mỗi miền mà thỏa mãn những đặc điểm trên là rất khó, hoặc cần phải trả một chi phí rất lớn để xây dựng dữ liệu huấn luyện. Vì vậy, trong bài báo này chúng tôi trình bày một phương pháp thích ứng miền mới cho dịch máy nơ ron, áp dụng cho cặp ngôn ngữ Anh - Việt với chiều dịch từ tiếng Anh sang tiếng Việt. Các thử nghiệm được tiến hành trên hai miền dữ liệu là miền tổng quan và miền pháp lý, chất lượng dịch trên miền tổng quan làm cơ sở để so sánh, đánh giá chất lượng hệ dịch khi được áp dụng trong miền pháp lý cũng như đánh giá hiệu quả của phương pháp được đề xuất. Qua thử nghiệm cho thấy, phương pháp này dễ thực hiện,

tận dụng được lượng lớn dữ liệu đơn ngữ luôn có sẵn với chi phí thấp và khả quan khi đã cải tiến được chất lượng bản dịch tăng 2,21 điểm BLEU [6] *(từ 22,17 điểm lên 24,38 điểm)*.

Bài báo này được trình bày cấu trúc như sau: Tiếp theo, phần 2 sẽ giới thiệu các nghiên cứu trước đây có liên quan; phần 3 trình bày tổng quan phương pháp chúng tôi đề xuất; phần 4 trình bày các thử nghiệm và các kết quả; phần 5 là kết luận và hướng phát triển; và cuối cùng phần 6 là một số tài liệu tham khảo.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Những năm gần đây, thích ứng miền là một trong những chủ đề đã giành được rất nhiều sự quan tâm của các nhà khoa học trên thế giới. Hiện nay, đã có nhiều phương pháp được đề xuất để thích ứng cho dịch máy thống kê cũng như dịch máy nơ ron, nhưng các đề xuất đó chủ yếu áp dụng cho một số cặp ngôn ngữ phổ biến trên thế giới như Anh - Pháp, Anh - Nhật, Anh - Tây Ban Nha,... Các phương pháp đã đề xuất được công bố đều thuộc một trong ba hướng chính, đó là: (1) bổ sung thêm nhiều dữ liệu hơn; (2) các kĩ thuật để có dữ liệu chất lượng hơn và (3) các kĩ thuật để có mô hình chất lượng hơn. Với hướng tiếp cận (1) và (2), đã có nhiều công bố đề xuất sử dụng dữ liệu đơn ngữ để cải tiến chất lượng hệ dịch khi dịch trong miền mới, các đề xuất này chủ yếu đã được chứng minh bằng thực nghiệm trong dịch máy thống kê, và chưa có nhiều đề xuất đối với dịch máy nơ ron.

Trong [2], kỹ thuật thích ứng giữa các miền được đề xuất để áp dụng cho dịch máy thống kê dựa vào cụm từ về nhiệm vụ Europarl[1] [3], để dịch các bình luận tin tức từ tiếng Pháp sang tiếng Anh. Cụ thể, một phần nhỏ dữ liệu song ngữ miền được khai thác để thích ứng mô hình ngôn ngữ và mô hình dịch bằng kỹ thuật nội suy tuyến tính. Việc thích ứng các mô hình dịch, mô hình đảo trật tự từ được thực hiện qua việc sinh thêm dữ liệu song ngữ từ dữ liệu đơn ngữ.

Công bố [9] đã đề xuất một số phương pháp thích ứng khá phức tạp dựa trên việc bổ sung thêm dữ liệu song ngữ được tổng hợp từ các tập dữ liệu dùng để tối ưu tham số và thử nghiệm. Ngoài ra, trong [10], đề xuất một phương pháp nhằm khai thác nguồn tài nguyên dữ liệu đơn ngữ miền bằng cách tổng hợp dữ liệu song ngữ từ việc dịch dữ liệu đơn ngữ miền sang ngôn ngữ đích. Phương pháp này chủ yếu liên quan đến kĩ thuật được đề xuất trong [2] nhưng khác nhau ở dữ liệu dùng để thích ứng miền, cụ thể ở [10] chỉ sử dụng dữ liệu đơn ngữ miền.

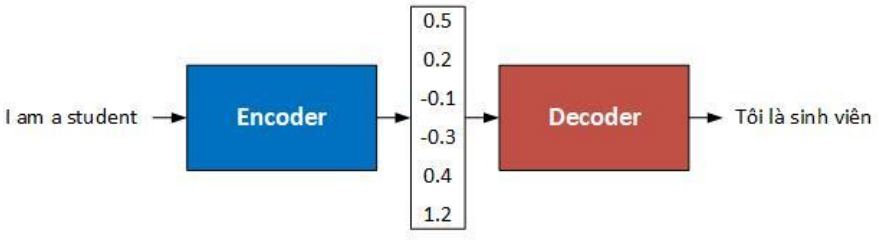
Các đề xuất trên được công bố cho dịch máy thống kê. Tuy nhiên, năm 2016 có công bố [11] đã đề xuất thích ứng miền cho dịch máy nơ ron dựa vào sinh dữ liệu song ngữ cho hệ dịch bằng việc dịch ngược các dữ liệu đơn ngữ trong miền đích. Trong bài báo này, phương pháp chúng tôi đề xuất có phần giống với phương pháp [9] vì chúng tôi có sử dụng thêm một tập dữ liệu miền pháp lý để tối ưu tham số của hệ dịch cơ sở theo định hướng miền đích, nhưng cũng liên quan nhiều đến phương pháp được đề xuất trong [10] và [11].

Nhìn chung, các phương pháp về thích ứng miền nói chung cho dịch máy đã được công bố khá phức tạp, thử nghiệm công phu và sử dụng nhiều mô hình toán học. Tuy nhiên, các thử nghiệm mới chỉ áp dụng cho một số cặp ngôn ngữ phổ biến như Anh - Pháp, Anh - Nhật, Anh - Tây Ban Nha,... Hiện vẫn chưa có công bố nào áp dụng cho cặp ngôn ngữ Anh - Việt.

III. PHƯƠNG PHÁP ĐỀ XUẤT

A. *Tổng quan về dịch máy nơ ron*

Đối với phương pháp dịch máy truyền thống như dịch máy thống kê dựa vào cụm thì hệ dịch thực hiện phân tách câu nguồn thành nhiều từ hoặc cụm từ riêng biệt, sau đó dịch tuần tự từng từ hoặc cụm từ một rồi sắp xếp lại trật tự các từ theo đúng trật tự trong ngôn ngữ đích. Vì thế, nên bản dịch không được trôi chảy và các dịch này không giống như cách con người dịch, để dịch, chúng ta sẽ đọc trọn vẹn một câu nguồn, hiểu ý nghĩa của nó rồi mới tiến hành dịch câu đó sang ngôn ngữ đích. Dịch máy nơ ron thực hiện dịch tương tự như cách của con người.



Hình 1. *Kiến trúc Encoder - Decoder*

Cụ thể, đầu tiên hệ dịch nơ ron sử dụng bộ mã hóa *(Encoder)* để đọc toàn bộ câu nguồn và mã hóa nó dưới dạng một vectơ biểu diễn ý nghĩa. Sau đó, bộ giải mã *(Decoder)* sẽ đọc và giải mã vec tơ biểu diễn câu nguồn này để sinh ra bản dịch tương ứng sang ngôn ngữ đích, quá trình mã hóa - giải mã được minh họa như ở hình 1 và hình 2 [5]. Theo cách dịch này, hệ dịch nơ ron có thể giải quyết được vấn đề dịch cục bộ trong phương pháp dịch dựa vào cụm truyền thống,

[1] http://www.statmt.org/europarl/

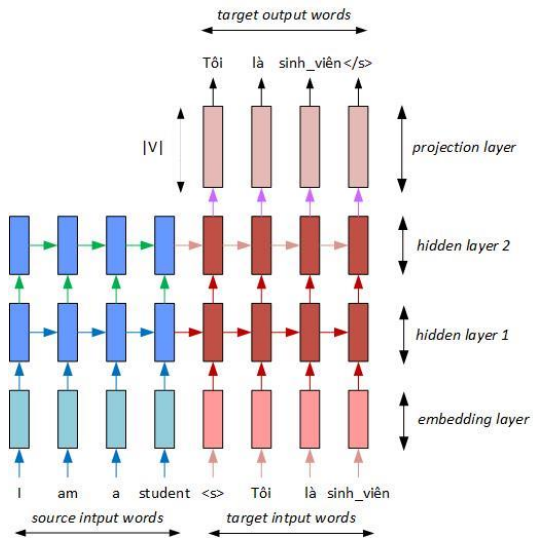
đó là: nó có thể nắm bắt được các phụ thuộc xa hơn trong các ngôn ngữ và tạo ra các bản dịch trôi chảy hơn nhiều so với hệ dịch thống kê dựa vào cụm truyền thống.

Bộ mã hóa - Bộ mã hóa đọc câu nguồn $X = (x_1, x_2, ..., x_T)$ và chuyển đổi nó thành một chuỗi các trạng thái ẩn $h = (h_1, h_2, ..., h_T)$ sử dụng mạng nơ ron hồi quy hai chiều *(bi-directional RNN)*. Tại mỗi thời điểm t, trạng thái ẩn h_t được xác định như là một kết hợp các trạng thái ẩn của mạng nơ ron hồi quy theo chiều xuôi *(forward RNN)* và theo chiều ngược *(backward RNN)* $[\overrightarrow{h_t}; \overleftarrow{h_t}]$ với điều kiện $\overrightarrow{h_t} = \text{RNN}(x_t, \overrightarrow{h_{t-1}}), \overleftarrow{h_t} = \text{RNN}(x_t, \overleftarrow{h_{t+1}})$

Bộ giải mã - Bộ giải mã sử dụng mạng nơ ron hồi quy khác để sinh ra bản dịch $Y = (y_1, y_2, ..., y_{T'})$ dựa trên các trạng thái ẩn h được sinh bởi bộ mã hóa. Tại mỗi thời điểm i, xác suất có điều kiện của mỗi từ y_i trong tập từ vựng V_y của ngôn ngữ đích được tính bởi công thức:

$$P(y_i|y_{<i}, h) = g(y_{i-1}, z_i, c_i),$$

với điều kiện z_i là trạng thái ẩn i_{th} của bộ giải mã, và được tính dựa vào trạng thái ẩn trước z_{i-1}, từ trước y_{i-1} và vectơ ngữ cảnh nguồn c_i: $Z_i = RNN(z_{i-1}, y_{i-1}, c_i)$.



Hình 2. *Kiến trúc tổng quát của hệ dịch nơ ron*

B. Phương pháp đề xuất

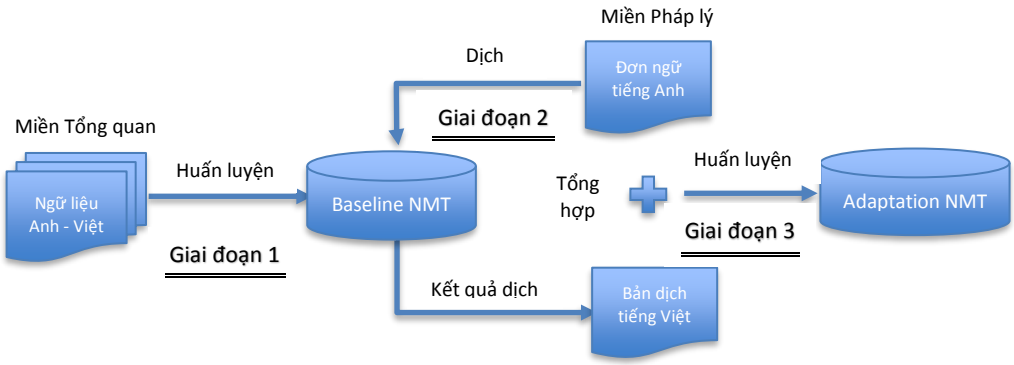
Trong thực tế, dữ liệu song ngữ thường không có sẵn, đặc biệt đối với các miền dữ liệu thuộc các lĩnh vực, chuyên ngành đặc thù, còn nếu muốn xây dựng dữ liệu song ngữ cho từng miền thì chi phí phải trả sẽ rất cao nhưng dữ liệu đơn ngữ thì lại luôn có sẵn với bất cứ miền dữ liệu nào. Trong dịch máy, dữ liệu đơn ngữ thường được dùng để làm mịn câu, khiến bản dịch của câu trôi chảy hơn và đọc lên thấy tự nhiên nhất. Dữ liệu đơn ngữ cũng đã được chứng minh có nhiều lợi ích trong việc cải tiến chất lượng dịch của cả hệ dịch máy thống kê và dịch máy nơ ron, đặc biệt trong nhiệm vụ thích ứng trong trường hợp nguồn tài nguyên bị hạn chế, nguồn dữ liệu song ngữ không đủ lớn. Hiện nay, cũng đã có một số đề xuất sử dụng dữ liệu đơn ngữ cho việc cải tiến chất lượng dịch, trong đó có đề xuất sinh dữ liệu song ngữ từ dữ liệu đơn ngữ cho dịch máy nhưng chưa có đề xuất, thử nghiệm hay khảo sát nào được công bố về sử dụng dữ liệu đơn ngữ để thích ứng miền áp dụng cho cặp ngôn ngữ Anh – Việt.

Như đã trình ở phần 2, phương pháp chúng tôi đề xuất có liên quan tới các công bố [9]; [10] và [11]. Theo [11], để sinh dữ liệu song ngữ thì việc dịch theo chiều ngược là cũng một giải pháp để có thể tận dụng được nguồn dữ liệu đơn ngữ miền. Để dịch theo chiều ngược hay theo chiều xuôi thì khá đơn giản và dễ áp dụng vì nó không yêu cầu phải thay đổi các thuật toán huấn luyện của hệ dịch.

Xuất phát từ ý tưởng trên, chúng tôi đề xuất một phương pháp mới để sinh dữ liệu song ngữ cho nhiệm vụ thích ứng miền áp dụng cho cặp ngôn ngữ Anh - Việt với chiều dịch từ Anh sang Việt, phương pháp của chúng tôi chỉ sử dụng dữ liệu đơn ngữ trong miền đích của ngôn ngữ đích. Phương pháp của chúng tôi khác với công bố trong [9]; [10] vì các công bố này chỉ thực nghiệm, áp dụng cho dịch máy thống kê dựa vào cụm còn phương pháp của chúng tôi là áp dụng cho dịch máy nơ ron. Ngoài ra, công bố [11] cũng khá liên quan tới phương pháp của chúng tôi khi cũng áp dụng cho dịch máy nơ ron, nhưng sử dụng kĩ thuật dịch ngược. Còn phương pháp của chúng tôi, cùng với các thử nghiệm, đánh giá hệ dịch dựa trên cách dịch xuôi dữ liệu đơn ngữ trong miền đích của ngôn ngữ đích. Phương pháp chúng tôi đề xuất được mô tả như hình 3, gồm 3 giai đoạn:

- **Giai đoạn 1**: Giai đoạn này chúng tôi sử dụng dữ liệu song ngữ Anh – Việt thuộc miền tổng quan để huấn luyện một hệ dịch nơ ron làm cơ sở để so sánh, đánh giá hiệu quả của phương pháp chúng tôi đề xuất *(đặt tên là **Baseline NMT** như mô tả trong Hình 3 **,** trong các thử nghiệm gồm các hệ dịch Baseline_L và Baseline_G)*;

- **Giai đoạn 2**: Sau khi đã có hệ dịch Baseline NMT ở giai đoạn 1, chúng tôi sử dụng hệ dịch này để dịch các văn bản đơn ngữ thuộc miền pháp lý trong tiếng Anh sang ngôn ngữ đích là tiếng Việt;
- **Giai đoạn 3**: Sau khi có kết quả dịch ở giai đoạn 2, chúng tôi sử dụng kết quả dịch này kết hợp với các văn bản đơn ngữ bằng tiếng Anh ở giai đoạn 2 để huấn luyện một hệ dịch nơ ron khác *(đặt tên là **Adaptation NMT** như mô tả trong Hình 3, trong các thử nghiệm là hệ dịch **Adapt_System**)*, hệ dịch này được sử dụng để cải tiến chất lượng dịch của các văn bản thuộc miền pháp lý.



***Hình 3.** Minh họa phương pháp đề xuất*

Bằng thực nghiệm, các kết quả so sánh thông qua cách đánh giá bằng điểm BLEU [6] đã chỉ ra rằng phương pháp chúng tôi đề xuất là cách tiếp cận khả quan, dễ thực hiện và đã cho kết quả dịch cải tiến hơn so với hệ dịch cơ sở ban đầu.

IV. THỰC NGHIỆM VÀ KẾT QUẢ

Để so sánh, đánh giá phương pháp đề xuất, chúng tôi tiến hành huấn luyện ba hệ dịch nơ ron, lần lượt *là **(1) Baseline_G*** - là hệ dịch cơ sở được huấn luyện với tập dữ liệu huấn luyện và tập tối ưu tham số *(tập dữ liệu G_train và tập dữ liệu G_val)* cùng thuộc miền tổng quan; ***(2) Baseline_L*** - là hệ dịch được huấn luyện với tập dữ liệu huấn luyện thuộc miền tổng quan *(G_train)*, còn tập tối ưu tham số thuộc miền luật *(L_val)*; ***(3) Adapt_System*** - là hệ dịch được huấn luyện với dữ liệu song ngữ được tổng hợp ở giai đoạn 2 của hình 3 và dữ liệu tối ưu tham số thuộc miền luật *(L_val)*.

Tiếp theo, chúng tôi sẽ mô tả về các tập dữ liệu, các bước tiền xử lý đối với dữ liệu huấn luyện của từng hệ dịch trên, đồng thời chúng tôi cũng trình bày cụ thể các bước thực nghiệm và kết quả tương ứng.

A. Dữ liệu

Để huấn luyện hệ dịch, trong các thử nghiệm của chúng tôi có hai loại dữ liệu miền khác nhau, ở góc độ bài toán mà chúng tôi giải quyết đó là tận dụng dữ liệu đơn ngữ thuộc miền cần dịch và một hệ dịch có sẵn thuộc miền tổng quan để nâng cao chất lượng dịch theo miền *(miền pháp lý trong các thực nghiệm của chúng tôi)*. Để thống nhất, chúng tôi gọi dữ liệu thuộc miền tổng quan để huấn luyện hệ dịch là dữ liệu trong miền và dữ liệu không thuộc miền huấn luyện là dữ liệu ngoài miền.

1. Thống kê dữ liệu

a) Dữ liệu trong miền: Chúng tôi sử dụng tập dữ liệu được cung cấp bởi hội nghị IWSLT 2015[2], tập dữ liệu này thuộc miền tổng quan gồm 131.000 cặp câu song ngữ tiếng Anh - tiếng Việt dành cho nhiệm vụ về dịch máy, tập dữ liệu này được gọi là tập ***G_train*** và được sử dụng để huấn luyện các hệ dịch cơ sở *(Baseline_G và Baseline_L)*. Để tối ưu các tham số của hệ dịch trong miền tổng quan, chúng tôi sử dụng tập dữ liệu gồm 745 cặp câu song ngữ thuộc miền tổng quan và gọi là tập ***G_val***. Để đánh giá chất lượng của các hệ dịch khi dịch trong miền tổng quan, chúng tôi sử dụng 1.046 cặp câu song ngữ Anh – Việt thuộc miền tổng quan.

b) Dữ liệu ngoài miền: Chúng tôi sử dụng 100.000 câu đơn ngữ tiếng Anh thuộc miền pháp lý và dùng hệ dịch cơ sở Basline_NMT theo mô tả ở giai đoạn 2 của hình 3 để dịch nhằm tạo ra bản dịch gồm 100.000 câu tiếng Việt tương ứng. Để đánh giá chất lượng của các hệ dịch trong miền pháp lý, chúng tôi sử dụng 2.000 cặp câu song ngữ Anh - Việt cùng thuộc miền pháp lý.

Một số thống kê về đặc điểm dữ liệu sử dụng trong các thực nghiệm được mô tả như Bảng 1.

2. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước xử lý không thể thiếu trong các bài toán dịch. Sau khi thu thập được đầy đủ các tập dữ

[2] http://workshop2015.iwslt.org/

liệu, chúng tôi tiến hành chuẩn hóa. Đầu tiên, chúng tôi thực hiện tách từ trong văn bản, đối với văn bản tiếng Anh thì cần quan tâm tới việc tách các dấu ". , ' ; ? " và các kí tự đặc biệt khác ra khỏi các từ trong văn bản. Để thực hiện việc này, chúng tôi sử dụng công cụ tách từ Tokenizer có sẵn trong hệ dịch mã nguồn mở Moses [4] do Koehn và cộng sự phát triển (2007). Đối với tiếng Việt, vì dấu cách không phải là dấu hiệu để phân biệt các từ, mà một từ trong tiếng Việt được cấu tạo bởi một hoặc nhiều âm tiết. Chính vì vậy, để tiến hành tách từ cho văn bản tiếng Việt, chúng tôi sử dụng công cụ tách từ dành riêng cho tiếng Việt khá phổ biến là VnTokenizer [1].

Sau đó, chúng tôi thực hiện chuyển tất cả các kí tự hoa trong các tập dữ liệu về dạng kí tự thường và loại bỏ những cặp câu có độ dài quá lớn trong dữ liệu, trong các thực nghiệm này chúng tôi chỉ chọn những câu có độ dài nhỏ hơn 80.

Bảng 1. *Một số thống kê về các tập dữ liệu thử nghiệm*

Miền dữ liệu	Dữ liệu thử nghiệm		Tiếng Anh	Tiếng Việt
Tổng quan	**G_train**	Số câu	**131.000**	
		Độ dài trung bình	18,91	17,98
	G_val	Số câu	**745**	
		Độ dài trung bình	22,73	21,41
	G_test	Số câu	**1.046**	
		Độ dài trung bình	22,70	21,42
Pháp lý	**L_val**	Số câu	**2.003**	
		Độ dài trung bình	25,48	26,11
	L_test	Số câu	**2.000**	
		Độ dài trung bình	28	27,82

B. Các thực nghiệm

Để huấn luyện các hệ dịch nơ ron, chúng tôi sử dụng công cụ OpenNMT[3] [7], đây là hệ dịch mã nguồn mở hoàn thiện, nổi tiếng, được công bố năm 2017 của nhóm Harvard NLP và SYSTRAN, công cụ này được nhiều người nghiên cứu trong cộng đồng dịch máy sử dụng. Các hệ dịch được huấn luyện với cùng các tham số mặc định, bao gồm hai tầng mạng LSTM với 500 nút ẩn và có sử dụng mô hình attention theo kiến trúc của Thang Luong [8]. Để so sánh, đánh giá chất lượng của các hệ dịch với nhau, chúng tôi sử dụng cách đánh giá tự động dựa vào điểm BLEU [6], đây cũng là cách đánh giá phổ biến trong bài toán dịch máy. Như mô tả ở hình 3:

Giai đoạn 1: Chúng tôi huấn luyện các hệ dịch cơ sở Baseline NMT, các hệ dịch này được huấn luyện với dữ liệu song ngữ thuộc miền tổng quan, nhưng được tối ưu tham số trong các miền dữ liệu khác nhau, cụ thể:

- Hệ dịch **Baseline_G:** Sử dụng tập dữ liệu G_train và G_val *(mô tả trong bảng 1)* để huấn luyện, hệ dịch cơ sở này được huấn luyện với dữ liệu song ngữ và tối ưu các tham số trong cùng một miền tổng quan.
- Hệ dịch **Baseline_L:** Sử dụng tập dữ liệu G_train và L_val *(mô tả trong bảng 1*) để huấn luyện, hệ dịch cơ sở này được huấn luyện với dữ liệu song ngữ thuộc miền tổng quan nhưng các tham số của hệ dịch được tối ưu trong miền pháp lý.

Việc lựa chọn hệ dịch có chất lượng bản dịch tốt, để từ đó tiến hành dịch xuôi và tổng hợp được dữ liệu song ngữ có chất lượng tốt. Chúng tôi tiến hành đánh giá, so sánh chất lượng bản dịch của hai hệ dịch cơ sở này khi dịch trong cùng một miền dữ liệu tổng quan và miền dữ liệu pháp lý. Kết quả thử nghiệm được đánh giá thông qua điểm BLEU được thể hiện như bảng 2. Ở bảng 2, ta thấy:

- Khi dịch với cùng tập dữ liệu là G_test thuộc miền tổng quan, hệ dịch Baseline_G cho điểm BLEU = 29,34 trong khi Baseline_L có điểm BLEU = 29,56.
- Khi dịch với cùng tập dữ liệu L_test thuộc miền pháp lý thì hệ dịch Baseline_G cho điểm BLEU = 22,17 và hệ dịch Baseline_L cho điểm BLEU = 23,01.

Như vậy, khi hệ dịch cơ sở Baseline_L được tối ưu tham số trong miền pháp lý đã cải tiến được chất lượng của bản dịch khi dịch trong miền pháp lý, cụ thể đã tăng 0.84 điểm BLEU *(điểm BLEU = 23,01 so với 22,17 của hệ dịch Baseline_G)*. Căn cứ vào kết quả so sánh này, chúng tôi lựa chọn hệ dịch cơ sở Baseline_L để thực hiện các bước trong giai đoạn 2.

Giai đoạn 2: Chúng tôi dùng hệ dịch Baseline_L ở trên để dịch tập dữ liệu đơn ngữ gồm 100.000 câu tiếng Anh thuộc miền pháp lý sinh ra bản dịch tương ứng gồm 100.000 câu tiếng Việt.

[3] http://opennmt.net/

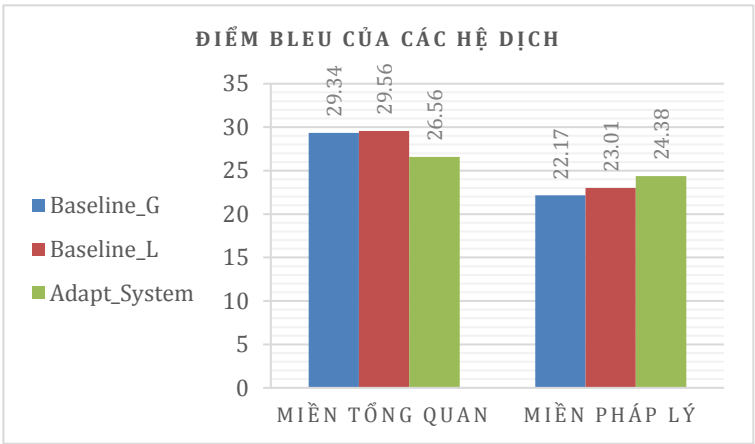
Bảng 2. *Điểm BLEU của các hệ dịch*

Miền thử nghiệm	Hệ dịch	BLEU Score
Tổng quan	**Baseline_G** *(G_train + G_val)*	29,34
	Baseline_L *(G_train + L_val)*	29,56
	Adapt_System	26,56
Pháp lý	**Baseline_G** *(G_train + G_val)*	22,17
	Baseline_L *(G_train + L_val)*	23,01
	Adapt_System	24,38

Giai đoạn 3: Chúng tôi sử dụng cặp dữ liệu đơn ngữ ở giai đoạn 2 *(gồm 100.000 tiếng Anh và bản dịch của nó gồm 100.000 tiếng Việt)* để huấn luyện hệ dịch Adapt_System, hệ dịch này được tối ưu tham số với tập dữ liệu L_val thuộc miền pháp lý. Các thử nghiệm cho kết quả điểm BLEU = 26,56 khi dịch tập dữ liệu G_test thuộc miền tổng quan, và điểm BLEU = 24,38 khi dịch tập dữ liệu L_test thuộc miền pháp lý.

Như vậy, hệ dịch Adapt_System cho chất lượng dịch trong miền pháp lý cao hơn so với các hệ dịch cơ sở Baseline_G và Baseline_L. Cụ thể, điểm BLEU cao hơn 2,21 điểm so với Baseline_G *(cải tiến từ 22,17 điểm tăng lên 24,38 điểm)* và cao hơn 1,37 điểm so với Baseline_L *(cải tiến từ 23,01 điểm tăng lên 24,38 điểm)*. Các kết quả thử nghiệm được thể hiện trong bảng 2 và sự biến đổi về chất lượng của bản dịch được thể hiện như biểu đồ trong hình 4.

Các kết quả thử nghiệm đã cho thấy phương pháp mà chúng tôi đề xuất là cách tiếp cận khả quan, dễ thực hiện và đã cho kết quả dịch khi dịch trong miền pháp lý cải tiến hơn so với hệ dịch cơ sở ban đầu.



Hình 4*. Biểu đồ so sánh điểm BLEU giữa các hệ thống thử nghiệm*

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất một phương pháp thích ứng miền mới cho dịch máy nơ ron, phương pháp này đặc biệt hiệu quả đối với các miền dữ liệu có ít tài nguyên của cặp ngôn ngữ Anh - Việt, trong các thử nghiệm của chúng tôi, chúng tôi sử dụng dữ liệu thuộc miền pháp lý. Qua thực nghiệm cho thấy, cách tiếp cận này là khả quan, dễ thực hiện và đã cho kết quả dịch có điểm BLEU tăng 2,21 điểm *(từ 22,17 điểm lên 24,38 điểm)*. Như vậy, chất lượng dịch sau khi thích ứng đã có cải tiến hơn so với hệ dịch cơ sở ban đầu.

Trong tương lai, chúng tôi sẽ tiến hành thử nghiệm mở rộng thêm trên cả hai chiều dịch đối với một số miền dữ liệu khác, và khảo sát với các tình huống khi dữ liệu đơn ngữ theo miền có sự thay đổi về lượng thì chất lượng dịch của hệ thống lúc này sẽ thay đổi như thế nào, và lượng dữ liệu đơn ngữ này thay đổi như thế nào là vừa đủ đối với từng miền dữ liệu.

VI. TÀI LIỆU THAM KHẢO

[1] Phuong-Le Hong, Huyen-Nguyen Thi Minh, Azim Roussanaly and Vinh-Ho Tuong (2008). A Hybrid Approach to Word Segmentation of Vietnamese Texts.In Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, Springer, LNCS 5196.

[2] Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 224–227, Prague, Czech Republic.

[3] Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished, http://www.isi.edu/∼koehn/europarl.

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177– 180, Prague, Czech Republic.

[5] Philipp Koehn. 2017. Neural machine translation. CoRR, abs/1709.07809.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL), pages 311–318, Philadelphia, PA.

[7] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proceedings of AMTA 2018, vol. 1: MT Research Track.

[8] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. EMNLP.

[9] Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Semi-supervised model adaptation for statistical machine translation. Machine Translation, 21(2):77–94.

[10] Nicola Bertoldi, Marcello Federico. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 182–189.

[11] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

DOMAIN ADAPTATION IN NEURON MACHINE TRANSLATION FOR ENGLISH – VIETNAMESE

Pham Nghia Luan, Nguyen Van Vinh, Nguyen Huy Hoang

ABSTRACT— *Machine translation is an important field of research in natural language processing. In recent years, neuron machine translation has been studied more popularly because the translation is better quality than the traditional Statistical Machine Translation. However, Neural Machine Translation has obtained state-of-the art performance for several language pairs, while only using parallel data for training. Target side monolingual data plays an important role in boosting fluency for phrase-based statistical machine translation, and In this paper, we investigate the use of monolingual data for Neural Machine Translation.*

In this paper, we propose a new domain adaptation method in neural machine translation for English-Vietnamese language pairs. We explore to train with monolingual data without changing the neural network architecture. The main idea of the paper is to combine monolingual data in-domain in source side (English) with its translation in the target side (Vietnamese) to training neural machine translation system. Experiments showed that the our method is simple, exploiting the advantages of monolingual data, especially the quality of the translation system increase 2.21 BLEU points.