Xây dựng kho ngữ liệu du lịch song ngữ Việt–Anh gióng hàng mức câu cho dịch máy

Nguyễn Tiến Hà[1], Nguyễn Thị Minh Huyền[2], Nguyễn Minh Hải[2]
[1]Trung tâm Giáo dục Thường xuyên tỉnh Phú Thọ
[2]Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội

Tác giả liên hệ: Nguyễn Tiến Hà, tienhapt@gmail.com
Ngày nhận bài: 11/08/2017, ngày sửa chữa: 03/05/2018, ngày duyệt đăng: 25/07/2018
Xem sớm trực tuyến: 08/11/2018, định danh DOI: 10.32913/rd-ict.vol1.no39.550
Biên tập lĩnh vực điều phối phản biện và quyết định nhận đăng: PGS. TS. Nguyễn Lê Minh

Tóm tắt: Kho ngữ liệu song ngữ được gióng hàng mức câu là một dạng tài nguyên ngôn ngữ quan trọng được sử dụng trong nhiều ứng dụng của xử lý ngôn ngữ tự nhiên, như: nghiên cứu ngôn ngữ học so sánh, tìm kiếm thông tin xuyên ngữ, xây dựng từ điển song ngữ. Đặc biệt trong lĩnh vực dịch máy, chất lượng và độ lớn của kho ngữ liệu song ngữ có vai trò quyết định đến chất lượng dịch. Các hệ thống dịch máy hiện nay vẫn cần được cải tiến để xử lý nhiều hiện tượng ngôn ngữ. Các hệ thống dịch máy huấn luyện trên miền tổng quát thường có chất lượng kém khi ứng dụng vào văn bản trên miền hạn chế. Một giải pháp cho vấn đề này là kết hợp mô hình dịch trên miền tổng quát và miền hạn chế. Để làm được điều đó, việc xây dựng được kho ngữ liệu trên miền hạn chế là rất cần thiết. Bài báo này trình bày việc xây dựng một kho ngữ liệu song ngữ Việt–Anh trong lĩnh vực du lịch và cải thiện một công cụ gióng hàng ở mức câu đã có cho văn bản song ngữ Việt–Anh, đạt được độ chính xác trên 90% cho các tập dữ liệu của chúng tôi. Với sự trợ giúp của công cụ này, chúng tôi đã xây dựng được kho ngữ liệu song ngữ Việt–Anh miền du lịch có gióng hàng mức câu, cho phép huấn luyện mô hình dịch máy Việt–Anh tăng được khoảng 8,79 điểm BLEU so với các mô hình được huấn luyện trên miền tổng quát.

Từ khóa: *Dịch máy thống kê, dịch máy Việt–Anh, dữ liệu song ngữ, gióng hàng song ngữ, kho văn bản du lịch.*

Title: **Building a sentence-aligned Vietnamese–English bilingual corpus in tourism domain for machine translation**

Abstract: Sentence-aligned bilingual corpora constitute an important language resource for many applications in natural language processing, such as comparative linguistics, cross-language information retrieval, bilingual dictionary construction. In machine translation, in particular, the quality and the size of bilingual corpora plays a crucial role in translation quality. Present machine translation systems still need to be improved to handle many linguistic phenomena. Translation systems trained on general-domain corpora usually perform poorly on texts from a specific domain. A solution is to combine the general-domain translation model with a specific-domain translation model. Consequently, the construction of annotated bilingual corpora in specific domains is important. In this paper, we present our work on the construction of a Vietnamese–English bilingual corpus in the field of tourism, and the improvement of an existing sentence alignment tool for Vietnamese–English bilingual texts, with the accuracy of above 90% on our different datasets. With the help of this tool, we build a sentence-aligned tourism domain corpus which, when used to train a Vietnamese–English translation model, allows an improvement of about 8.79 BLEU scores in comparison with the models trained with only parallel general domain texts.

Keywords: *Bilingual data, bilingual alignment, statistical machine translation, tourism domain corpus, Vietnamese–English machine translation.*

I. GIỚI THIỆU

Kho ngữ liệu song ngữ gióng hàng ở mức câu là tài nguyên ngôn ngữ quan trọng cho nhiều ứng dụng của xử lí ngôn ngữ tự nhiên, như nghiên cứu ngôn ngữ học so sánh, tìm kiếm thông tin xuyên ngữ, xây dựng từ điển song ngữ, đặc biệt là để huấn luyện các hệ thống dịch máy dựa vào thống kê, ví dụ như hệ thống MOSES [1]. Chất lượng dịch của một hệ thống dịch máy dựa vào thống kê chịu ảnh hưởng rất nhiều bởi kích thước và chất lượng của kho ngữ liệu song ngữ. Bên cạnh đó, các hệ dịch máy huấn luyện trên miền tổng quát có chất lượng giảm đi rõ rệt khi ứng dụng vào dịch văn bản trên miền hạn chế. Do vậy, khi triển khai hệ thống dịch máy thống kê trên một miền hạn chế, việc xây dựng kho ngữ liệu phù hợp là một nhiệm vụ thiết yếu.

Vấn đề dịch máy Anh–Việt trên miền tổng quát đã được nhiều nhóm nghiên cứu quan tâm. Đặc biệt đã có những kho ngữ liệu song ngữ Anh–Việt được xây dựng cho dịch máy trên miền tổng quát như kho VLSP gồm 100000 cặp câu được xây dựng bởi các nhóm nghiên cứu trong khuôn khổ đề tài VLSP KC01/06-10[1], hay kho ngữ liệu EVBCorpus gồm 800000 cặp câu [2].

Trong bài báo này, chúng tôi quan tâm tới bài toán dịch trên miền văn bản du lịch. Hiện nay, lượng khách du lịch nước ngoài đến du lịch tại Việt Nam là khá lớn, theo thống kê của Tổng cục Du lịch Việt Nam, 9 tháng đầu năm 2016, lượng khách quốc tế đến Việt Nam là 7.265.380 lượt khách[2]. Nhu cầu tra cứu thông tin du lịch hầu hết bằng tiếng Anh của du khách rất lớn. Gần đây cũng đã có đề tài về dịch tiếng nói cho các hội thoại nhằm phục vụ khách du lịch[3]. Bài báo này tập trung vào chủ đề dịch máy Việt–Anh cho văn bản trong lĩnh vực du lịch, nhằm hỗ trợ cho việc truyền bá các thông tin du lịch của các địa phương. Cụ thể, chúng tôi đặt mục tiêu cải thiện chất lượng dịch văn bản du lịch bằng việc thực hiện xây dựng kho ngữ liệu song ngữ Việt–Anh gióng hàng ở mức câu với kích thước lớn trên miền hạn chế là thông tin du lịch. Việc xây dựng này bao gồm hai nhiệm vụ: thứ nhất là thu thập văn bản song ngữ Việt–Anh về chủ đề du lịch, thứ hai là phát triển nâng cấp một phần mềm gióng hàng câu hiệu quả cho văn bản song ngữ Việt–Anh nhằm hỗ trợ việc gióng hàng kho văn bản song ngữ. Chúng tôi cũng chỉ ra rằng kho ngữ liệu song ngữ có gióng hàng đã xây dựng thực sự có giá trị nâng cao chất lượng dịch văn bản Việt–Anh thuộc lĩnh vực du lịch.

Trong mục II của bài báo chúng tôi trình bày các bước xây dựng một kho ngữ liệu song ngữ có gióng hàng câu. Mục III trình bày việc xây dựng kho ngữ liệu du lịch song ngữ Việt–Anh, việc cải tiến một công cụ tự động gióng hàng mức câu và ứng dụng công cụ đó trong việc gióng hàng kho ngữ liệu thu được. Mục IV trình bày kết quả thực nghiệm ứng dụng kho ngữ liệu đã xây dựng vào hệ thống dịch máy Việt–Anh cho dữ liệu văn bản du lịch. Mục V đưa ra kết luận và định hướng nghiên cứu tiếp theo.

II. PHƯƠNG PHÁP THU THẬP NGỮ LIỆU SONG NGỮ VÀ GIÓNG HÀNG CÂU

Giai đoạn đầu tiên trong tiến trình xây dựng kho ngữ liệu song ngữ có gióng hàng mức câu là thu thập văn bản song ngữ. Có hai phương pháp cơ bản thu thập văn bản song ngữ, đó là phương pháp thủ công [3] và phương pháp tự động [4].

[1]https://vlsp.hpda.vn/demo/?page=resources.

[2]http://vietnamtourism.gov.vn/index.php/items/21541.

[3]Đề tài Nhà nước KC01.03/11-15: Nghiên cứu phát triển hệ thống dịch tiếng nói hai chiều Việt–Anh, Anh–Việt có định hướng lĩnh vực.

Phương pháp thủ công [3]: Chụp ảnh hoặc scan các sách báo, tài liệu, bảng thông tin, v.v. rồi dùng các phầm mềm xử lí để tách văn bản; gõ lại văn bản; hoặc tìm các dữ liệu đã số hóa, chẳng hạn như các trang web, rồi trích xuất văn bản ra. Ưu điểm của phương pháp này là cho phép thu thập được từ nhiều nguồn văn bản song ngữ khác nhau trong đó có cả những nguồn chưa được số hóa, nhưng nhược điểm là tốn rất nhiều công sức, tiền bạc và thời gian.

Phương pháp tự động [4]: Chủ yếu dùng các chương trình gom tự động các dữ liệu trên mạng Internet rồi trích chọn văn bản song ngữ có sự tương đương dịch. Sau đó, cần kiểm tra lại bằng phương pháp thủ công để loại bỏ các kết quả không như ý. Ưu điểm của phương pháp này là cho phép thu thập văn bản song ngữ nhanh và tốn ít chi phí, nhưng nhược điểm là nguồn dữ liệu song ngữ thu thập bị hạn chế. Trong thực tế, phương pháp này chỉ có thể áp dụng để thu thập văn bản song ngữ từ các trang web song ngữ.

Mỗi phương pháp thu thập văn bản song ngữ đều có ưu và nhược điểm của nó. Qua tìm hiểu, chúng tôi nhận thấy ngữ liệu du lịch song ngữ Việt–Anh khá ít và phân tán ở nhiều nguồn khác nhau, như sách, sổ tay, bảng thông báo, hay website song ngữ, nên phương pháp thu thập tự động thu được ít dữ liệu. Do vậy, chúng tôi chủ yếu dùng phương pháp thu thập dữ liệu du lịch song ngữ một cách thủ công.

Giai đoạn thứ hai là gióng hàng mức câu các văn bản song ngữ thu thập được. Phương pháp gióng hàng văn bản song ngữ mức câu đầu tiên dựa trên độ dài câu được Brown và cộng sự đề xuất năm 1991 [5]. Độ dài câu được tính bằng số lượng từ (token) có trong câu. Thuật toán giả thiết rằng độ dài của một câu bất kì và bản dịch của nó có sự phụ thuộc chặt chẽ. Thuật toán gióng hàng hai văn bản dựa vào mô hình Markov ẩn. Gale và Church [6] cũng có hướng tiếp cận tương tự nhưng hai ông đo độ dài câu bằng số lượng kí tự và áp dụng thuật toán quy hoạch động.

Kay và Röscheisen [7] giả định nếu hai câu là gióng hàng của nhau thì các từ của chúng cũng phải tương ứng. Ban đầu, một ma trận các cặp ứng viên câu gióng hàng với nhau được khởi tạo với cặp câu đầu, cuối văn bản và mỗi cặp câu ở giữa phân bố gần đường chéo cũng được giả định gióng với nhau. Sau đó, tính toán tần suất của các cặp từ xuất hiện đồng thời trong cặp câu ứng viên. Bảng các cặp câu ứng viên được cập nhật dựa trên số lượng cặp từ có tần suất cao mà cặp câu đó chứa. Các cặp từ với tần suất rất cao tạo thành các điểm neo mới để cập nhật giả định gióng hàng các câu ở giữa. Thuật toán lặp đi lặp lại đến khi hội tụ. Thuật toán đạt độ chính xác cao nhưng chậm.

Chen [8] đề xuất thuật toán gióng hàng dựa trên việc tính toán xác suất cặp từ có mặt trong cặp câu gióng hàng với nhau trong văn bản huấn luyện. Sau đó, áp dụng mô hình Markov ẩn tương tự như của Brown và cộng sự để gióng hàng câu.

Simard và Plamondon [9] đề xuất dùng các từ cùng gốc (cognate) như ngày, tháng, tên riêng, một số dấu câu để tạo thành các điểm neo chia 2 văn bản thành các khối tương ứng nhỏ hơn. Các cognate được định nghĩa là cặp từ tố trong 2 văn bản có 4 kí tự đầu giống nhau.

Romary và Bonhomme [10] đề xuất phương pháp gióng hàng dựa vào cấu trúc văn bản kết hợp với gióng hàng dựa vào độ dài văn bản theo ký tự của Gale và Church [6]. Huyen và Rossignol [11] đề xuất cải tiến công cụ gióng hàng XAlign theo cách tiếp cận này bằng cách cho phép chương trình ước lượng tự động các tỉ lệ độ dài trung bình của văn bản trong hai ngôn ngữ bất kì. Tuy nhiên, đánh giá kết quả gióng hàng của công cụ XAlign trên cặp ngôn ngữ Anh–Việt và Pháp–Việt cho thấy độ chính xác thấp hơn đáng kể so với kết quả gióng hàng các cặp ngôn ngữ Ấn Âu như Anh–Pháp.

Cho đến thời điểm hiện tại các phương pháp gióng hàng câu song ngữ Việt–Anh đều cho độ chính xác chưa cao [12], nên cần được tiếp tục nghiên cứu và cải tiến.

III. XÂY DỰNG KHO NGỮ LIỆU DU LỊCH SONG NGỮ VIỆT–ANH GIÓNG HÀNG MỨC CÂU

Việc xây dựng kho ngữ liệu du lịch song ngữ Việt–Anh có gióng hàng câu được tiến hành theo ba bước sau.

1. Nguồn thu thập dữ liệu

Các văn bản song ngữ trong lĩnh vực du lịch không nhiều và khá tản mát. Việc thu thập tự động dữ liệu song ngữ du lịch trên các trang web không khả thi do số lượng trang web trong lĩnh vực này không nhiều, và nếu có thì số lượng bài cũng rất ít. Nếu có hai bài về cùng một chủ đề thì thường lại viết khác nhau nên không thể coi là bản dịch của nhau. Do đó, chúng tôi thu thập dữ liệu du lịch song ngữ Việt–Anh bằng phương pháp thủ công là chủ yếu.

Các nguồn thu thập chủ yếu gồm có:

- Sổ tay du lịch của các địa danh du lịch, sổ hướng dẫn sử dụng khách sạn, sách dạy hội thoại tiếng Anh với khách du lịch;
- Lời giới thiệu song ngữ trên các bảng gắn tại các di tích, địa điểm du lịch do tác giả đi du lịch chụp lại và nhờ bạn bè đi du lịch chụp và gửi cho (Theo cách thu thập này tác giả đã thu được 36 trang văn bản song ngữ Việt–Anh, tương ứng với 741 câu Tiếng Việt và 756 câu Tiếng Anh);
- Tờ rơi, tờ gấp quảng cáo du lịch;
- Sách Luật du lịch;
- Văn bản trong hồ sơ đề nghị công nhận di sản văn hóa của Việt Nam;
- Văn bản hợp tác du lịch với các quốc gia;
- Trang web song ngữ giới thiệu về du lịch Việt Nam.

2. Chuyển dữ liệu song ngữ thu thập thành dữ liệu số có cấu trúc thống nhất

Đối với các tài liệu như sách, sổ tay, tờ rơi, bảng thông báo, nếu chỉ có bản cứng (văn bản trên giấy), không có bản mềm (văn bản lưu trên máy tính), thì chúng tôi tiến hành công việc như sau:

Bước 1: Dùng máy quét ảnh hoặc máy ảnh để chụp ảnh;

Bước 2: Dùng phần mềm chuyển file ảnh văn bản thành văn bản;

Bước 3: Chỉnh sửa các lỗi văn bản do phần mềm nhận dạng văn bản nhận dạng sai để thu được văn bản song ngữ chính xác bằng bản mềm.

Các dữ liệu dạng mềm được làm sạch thành phần không phải chữ như ảnh, các thẻ, các bảng biểu (nếu có).

Tất cả các văn bản mềm sau đó được tách thành chương, đoạn theo một định dạng thống nhất. Dữ liệu mới sau đó trải qua quá trình tách câu bằng công cụ tự động. Chúng tôi dùng công cụ tách câu vnSentDetector[4] cho văn bản tiếng Việt và Stanford NLP cho văn bản tiếng Anh[5]. Văn bản sau khi tách câu được kiểm tra lại một lần nữa để loại bỏ lỗi sai.

Chúng tôi cũng xây dựng một công cụ tự động gắn thẻ cấu trúc văn bản (các khối văn bản như chương, đoạn văn và câu).

3. Gióng hàng câu văn bản song ngữ

Các văn bản đã tách đoạn và câu được tiến hành gióng hàng. Chúng tôi cải tiến công cụ XAlign [11] để gióng hàng văn bản. Việc lựa chọn công cụ XAlign có hai lí do sau. Thứ nhất là công cụ được phát triển bởi một thành viên trong nhóm tác giả. Thứ hai đây là một trong các công cụ có độ chính xác cao và ổn định trong các công cụ tham gia dự án đánh giá các công cụ gióng hàng ARCADE II [13]. Trong mục này, phương pháp cải tiến của chúng tôi là mở rộng khả năng gióng hàng và đề xuất giá trị phạt (penalty) phù hợp đối với từng loại gióng hàng cho cặp ngôn ngữ Việt–Anh. Dưới đây, trước hết, chúng tôi trình bày lí do phải mở rộng khả năng gióng hàng và đề xuất công thức mở rộng đối với giải thuật DTW (Dynamic Time Warping). Sau đó, chúng tôi đề xuất công thức để tính giá trị hàm phạt *pen* phù hợp đối với từng loại gióng hàng cho cặp ngôn ngữ Việt–Anh. Cuối cùng, chúng tôi trình bày kết quả thực nghiệm.

1) Mở rộng khả năng gióng hàng:

Cũng như tất cả các phương pháp gióng hàng câu đã trình bày trong mục II, phương pháp gióng hàng cài đặt trong XAlign chỉ xét đến các kiểu gióng hàng *n-m*, với *n*

[4]http://mim.hus.vnu.edu.vn/phuonglh/softwares.

[5]https://stanfordnlp.github.io/CoreNLP/download.html.

câu văn bản gốc và m câu văn bản dịch, như sau: 0-1, 1-0, 1-1, 1-2, 2-1, 2-2 (gióng hàng đến cấp độ 2). Chúng tôi thống kê trên kho ngữ liệu du lịch có khoảng 5000 cặp câu song ngữ du lịch Việt–Anh đã được gióng hàng chính xác (bằng cách gióng hàng tự động rồi chỉnh sửa thủ công), chúng tôi nhận thấy rằng các gióng hàng 3-1, 1-3, 2-3, 3-2, 3-3 (gióng hàng đến cấp độ 3) chiếm khoảng $1,7\%$. Gióng hàng từ cấp độ 4 trở lên chiếm tỉ lệ nhỏ hơn nhiều, khoảng $0,42\%$.

Phương pháp gióng hàng sử dụng thuật toán DTW trả lại kết quả là phép gióng hàng tất cả các câu trên hai văn bản mà có tổng chi phí gióng hàng các câu theo trật tự tuyến tính là nhỏ nhất. Với thống kê kể trên, việc không tính đến các gióng hàng cấp độ 3 ảnh hưởng khá lớn tới chất lượng gióng hàng, do sự lan truyền lỗi. Vì thế chúng tôi quyết định mở rộng phương pháp gióng hàng câu được đề xuất trong [11] đến cấp độ 3, tạm thời không xét cấp độ 4 có tỉ lệ thấp.

Với việc mở rộng sang các phép gióng hàng đến cấp độ 3, công thức của giải thuật DTW trong [11] được đề xuất mở rộng thêm 5 khả năng gióng hàng so với công thức cũ, như sau:

$$m_{ij} = \min\{a_{11}, a_{10}, a_{01}, a_{21}, a_{12}, a_{22}, a_{13}, a_{31}, a_{23}, a_{32}, a_{33}\}, \quad (1)$$

trong đó

$$\begin{aligned}
a_{11} &= m_{i-1,j-1} + c(a_{i-1}, b_j)\\
a_{10} &= m_{i-1,j} + c(a_{i-1}, 0) + pen_{10}\\
a_{01} &= m_{i,j-1} + c(0, b_{i-1}) + pen_{01}\\
a_{21} &= m_{i-2,j-1} + c(a_{i-1} + a_{i-2}, b_{i-1}) + pen_{21}\\
a_{12} &= m_{i-1,j-2} + c(a_{i-1}, b_{i-1} + b_{i-2}) + pen_{12}\\
a_{22} &= m_{i-2,j-2} + c(a_{i-1} + a_{i-2}, b_{i-1} + b_{i-2}) + pen_{22}\\
a_{13} &= m_{i-1,j-3} + c(a_{i-1}, b_{i-1} + b_{i-2} + b_{i-3}) + pen_{13}\\
a_{23} &= m_{i-2,j-3} + c(a_{i-1} + a_{i-2}, b_{i-1} + b_{i-2} + b_{i-3}) + pen_{23}\\
a_{32} &= m_{i-3,j-2} + c(a_{i-1} + a_{i-2} + a_{i-3}, b_{i-1} + b_{i-2}) + pen_{32}\\
a_{33} &= m_{i-3,j-3} + c(a_{i-1} + a_{i-2} + a_{i-3}, b_{i-1} + b_{i-2} + b_{i-3}) + pen_{33}
\end{aligned}$$

Giả sử trong hai văn bản song song cần gióng hàng có n câu ở ngôn ngữ nguồn và p câu ở ngôn ngữ đích. Khi đó gọi a_i $(1 \le i \le n)$ và b_j $(1 \le j \le p)$ lần lượt là mảng chứa số kí tự từng câu trong văn bản nguồn và văn bản đích. Giá trị m_{ij} $(1 \le i \le n,\ 1 \le j \le p)$ lưu giữ chi phí gióng hàng nhỏ nhất khi gióng khớp i câu nguồn với j câu đích. Như vậy, m_{np} chính là chi phí nhỏ nhất khi gióng hàng n câu nguồn này với p câu đích kia.

Trong phần giải thích các ký hiệu của biểu thức (1), giá trị $c(l_s, l_t)$ là hàm chi phí định nghĩa dựa trên mức độ chênh lệch về độ dài giữa hai đoạn văn bản tương đương dịch khi thực hiện gióng hàng hai đoạn văn bản nguồn và đích có độ dài tương ứng là l_s và l_t. Giá trị pen_{ij} là giá trị hàm phạt cho mỗi kiểu gióng hàng khác với kiểu gióng hàng phổ biến nhất là 1-1. Giá trị này tỉ lệ nghịch với xác suất của kiểu gióng hàng tương ứng.

Bảng I
GIÁ TRỊ *pen* CHO MỖI KIỂU GIÓNG HÀNG

Gióng hàng	0-1	1-0	1-1	2-1	1-2	2-2
penalty	482	547	0	200	-177	44
Gióng hàng	2-3	3-2	3-1	1-3	3-3	
penalty	795	657	426	-265	4691	

2) Tính giá trị pen phù hợp cho cặp ngôn ngữ Việt–Anh:

Giá trị pen_{ij} được tính theo số lượng gióng hàng i-j so với gióng hàng 1-1 trên kho ngữ liệu mà ta lựa chọn. Trong công cụ gióng hàng viXAlign, ngoại trừ gióng hàng kiểu 1-1 không xét giá trị phạt, pen_{ij} được tính theo công thức sau dựa trên kho ngữ liệu 5000 cặp câu song ngữ Anh–Việt lĩnh vực du lịch đã được gióng hàng chính xác:

$$pen_{ij} = -100\left(\frac{P(\text{match}(i,j))}{P(\text{match}(1-1))}\right) + 177, \quad (2)$$

trong đó các cặp chỉ số i, j trong (1) thỏa mãn $0 \le i, j \le 3$, $P(\text{match}(i-j))$ là xác suất gióng hàng kiểu i-j. Xác suất này được ước lượng dựa trên kho ngữ liệu gióng hàng mẫu 5000 gióng hàng dùng làm khảo sát.

Giá trị *pen* thu được đối với từng loại gióng hàng như trong Bảng I.

3) Kết quả thực nghiệm: Chúng tôi sử dụng các độ đo độ chính xác (Prec), độ phủ (Rec), độ đo F (F-mea) để đánh giá công cụ gióng hàng câu.

$$\text{Prec} = \frac{\text{CorS}}{\text{AliS}}, \quad (3)$$

$$\text{Rec} = \frac{\text{CortS}}{\text{HanS}}, \quad (4)$$

$$\text{F-mea} = 2 \times \text{Rec}\frac{\text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}, \quad (5)$$

trong đó, CorS là số gióng hàng câu đúng bởi giải thuật, AliS là tổng số gióng hàng câu bởi giải thuật và HanS là tổng số gióng hàng câu thủ công làm chuẩn tham chiếu.

Thực hiện chạy công cụ XAlign được cải tiến trên kho ngữ liệu song ngữ Việt–Anh "Le Petit prince" có 1663 câu tiếng Việt và 1660 câu tiếng Anh và kho ngữ liệu du lịch song ngữ Việt–Anh có 12457 câu tiếng Anh và 12286 câu tiếng Việt so với khi chưa cải tiến chúng tôi thu được kết quả trong Bảng II và Bảng III.

Như vậy, nhờ có việc bổ sung các phép gióng hàng cấp độ 3, cùng với việc tính các giá trị phạt phù hợp, chất lượng gióng hàng đã được tăng lên đáng kể trên cả văn bản trong lĩnh vực du lịch và văn bản trên miền văn học.

Bảng II
GIÓNG HÀNG TRÊN KHO NGỮ LIỆU VIỆT–ANH "LE PETIT PRINCE"

	Precision	Recall	F-measure
Chưa cải tiến	81,42%	76,21%	78,73%
Đã cải tiến	89,15%	88,18%	88,66%

Bảng III
GIÓNG HÀNG TRÊN KHO NGỮ LIỆU DU LỊCH VIỆT–ANH

	Precision	Recall	F-measure
Chưa cải tiến	80,61%	84,99%	82,74%
Đã cải tiến	90,60%	89,77%	90,18%

Phần mềm đã cải tiến được chia sẻ tại địa chỉ sau: https://github.com/viXAlign/viXAlign-project.

IV. ỨNG DỤNG KHO NGỮ LIỆU DU LỊCH SONG NGỮ VIỆT–ANH CHO HỆ THỐNG DỊCH MÁY

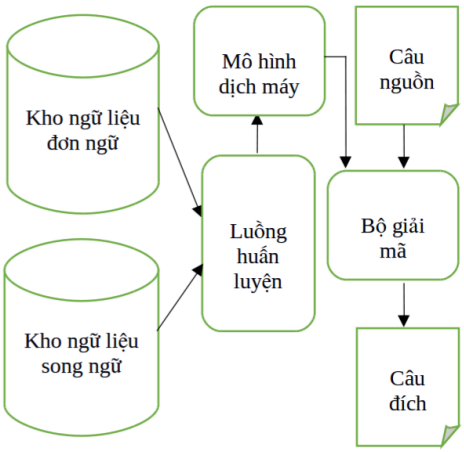
Phương pháp tiếp cận của chúng tôi hướng vào việc xây dựng kho ngữ liệu song ngữ huấn luyện hệ thống dịch máy phân chia theo các lĩnh vực. Kho ngữ liệu song ngữ Việt–Anh đã gióng hàng ở bước trên được sử dụng để cải thiện chất lượng của hệ thống dịch máy thống kê cho các văn bản thuộc lĩnh vực du lịch. Trong phần này chúng tôi trình bày kết quả thực nghiệm trên hệ thống dịch máy MOSES.

Cụ thể, chúng tôi sử dụng MOSES để huấn luyện hệ thống dịch máy trên kho ngữ liệu du lịch song ngữ Việt–Anh mà chúng tôi thu thập được, sau đó sử dụng hệ thống dịch máy này để dịch văn bản du lịch rồi so sánh chất lượng bản dịch với hệ thống dịch máy huấn luyện trên kho ngữ liệu không chia theo lĩnh vực, sử dụng phương pháp đánh giá chất lượng dịch máy theo điểm BLEU.

1. Hệ thông dịch máy MOSES

MOSES (Hình 1) là một hệ thống dịch máy thống kê. Trong dịch máy thống kê, các hệ thống dịch máy được huấn luyện trên kho ngữ liệu song ngữ lớn (để hệ thống học cách dịch các đoạn nhỏ) và kho ngữ liệu đơn ngữ (để học cách đưa ra đầu ra trôi chảy).

MOSES có hai thành phần chính, là luồng huấn luyện và bộ giải mã. Luồng huấn luyện là một tập các công cụ nhận dữ liệu thô (song ngữ và đơn ngữ) và biến nó thành một mô hình dịch máy. Bộ giải mã là một ứng dụng C^{++} đơn giản, với một mô hình dịch máy được huấn luyện và một câu nguồn cho trước, bộ giải mã sẽ dịch câu nguồn thành câu đích.



Hình 1. Hệ thống dịch máy MOSES.

2. Độ đo đánh giá: điểm BLEU

Điểm BLEU (BiLingual Evaluation Understudy), được đề xuất bởi Papineni và cộng sự vào năm 2002 [14], là thước đo tự động đầu tiên được chấp thuận dùng để đánh giá các bản dịch, được định nghĩa như sau:

$$\text{BLEU} = BP \cdot \mathrm{e}^{\sum_{n=1}^{n} w_n \log pn}, \tag{6}$$

trong đó pn là số n-gram của bản dịch máy mà xuất hiện trong tập bản dịch tham chiếu chia cho tổng n-gram của bản dịch máy, w_i là trọng số tích cực và BP là phạt ngắn dùng để phạt các bản dịch "quá ngắn". Phạt ngắn được tính toán trên toàn bộ kho ngữ liệu và được lựa chọn như là hàm số mũ giảm ở "r/c", với "c" là độ dài của bản dịch ứng viên và r là độ dài của bản dịch tham chiếu, theo công thức sau:

$$BP = \begin{cases} 1, & \text{nếu } c > r, \\ e^{1-\frac{r}{c}}, & \text{nếu } c < r. \end{cases} \tag{7}$$

3. Kết quả thực nghiệm

1) Đánh giá hiệu quả ứng dụng kho ngữ liệu du lịch vào hệ thống dịch trên miền du lịch:

Trong mục này, chúng tôi thực hiện đánh giá hiệu quả của việc ứng dụng kho ngữ liệu song ngữ Việt–Anh vào huấn luyện các hệ thống dịch máy văn bản trong miền du lịch. Để làm điều này, chúng tôi thực nghiệm so sánh kết quả dịch của một hệ thống không được huấn luyện với dữ liệu song ngữ trong miền du lịch (hệ thống 1) với 6 hệ thống được huấn luyện với dữ liệu miền du lịch theo nguyên tắc đánh giá chéo. Chia 12000 cặp câu song ngữ du lịch Việt–Anh thành 6 phần độc lập, lần lượt giữ lại 1 phần (2000 cặp câu) để làm dữ liệu đánh giá, 5 phần còn lại (10000 cặp câu) đưa thêm vào kho ngữ liệu để huấn luyện Hệ thống dịch máy. Cách thức huấn luyện các hệ thống dịch như sau.

Bảng IV
ĐIỂM BLEU CỦA CÁC HỆ THỐNG DỊCH

	Hệ thống dịch máy	Hệ thống dịch máy 1	(+)
2	16,75	4,16	12,59
3	20,05	5,24	14,81
4	11,59	4,42	7,17
5	10,42	3,59	6,8
6	10,89	2,88	8,01
7	7,16	3,85	3,31
TB	**12,81**	**4,02**	**8,79**

Trước hết là huấn luyện Hệ thống dịch máy 1. Chúng tôi sử dụng kho ngữ liệu 165678 cặp câu song ngữ Việt–Anh thuộc nhiều lĩnh vực khác nhau (từ nguồn đề tài VLSP[1] và một số dữ liệu khác mà chúng tôi thu thập, gióng hàng và cung cấp cùng với phần mềm gióng hàng). Sử dụng MOSES để huấn luyện hệ thống dịch máy Việt–Anh trên kho ngữ liệu này chúng tôi thu được hệ thống dịch máy 1.

Tiếp theo là huấn luyện Hệ thống dịch máy 2, 3, 4, 5, 6 và 7. Chúng tôi sử dụng hệ thống MOSES lần lượt huấn luyện để thu được 6 hệ thống dịch máy Việt–Anh trên kho ngữ liệu 165678 cặp câu song ngữ Việt–Anh mà chúng tôi đã huấn luyện ra Hệ thống dịch máy 1 nhưng thay thế 10000 cặp câu của kho ngữ liệu này bằng 10000 cặp câu song ngữ du lịch Việt–Anh ở mỗi lượt đánh giá chéo. Lần 1 thay thế từ cặp câu thứ 1 đến cặp câu thứ 10000. Lần 2 thay thế từ cặp câu thứ 30001 đến cặp câu thứ 40000. Lần 3 thay thế từ cặp câu thứ 50001 đến cặp câu thứ 60000. Lần 4 thay thế từ cặp câu thứ 90001 đến cặp câu thứ 100000. Lần 5 thay thế từ cặp câu thứ 120001 đến cặp câu thứ 130000. Lần 6 thay thế từ cặp câu thứ 150001 đến cặp câu thứ 160000.

Chúng tôi áp dụng lần lượt các cặp hệ thống dịch máy Việt–Anh (1,2), (1,3), (1,4), (1,5), (1,6) và (1,7) cho việc dịch 2000 câu tiếng Việt trong miền du lịch được giữ lại làm dữ liệu kiểm tra. Sau đó dùng công cụ tính điểm BLEU của MOSES [1] để tính điểm cho từng hệ thống dịch này và so sánh kết quả tính được. Kết quả cho thấy cả 6 hệ thống 2, 3, 4, 5, 6 và 7 đều cải thiện điểm BLEU so với hệ thống 1 như trong Bảng IV. Kết quả qua 6 lần thực nghiệm điểm BLEU tăng trung bình là 8,79.

Các kết quả thu được cho phép chúng tôi khẳng định được ý nghĩa của việc xây dựng dữ liệu huấn luyện trên một miền hạn chế để tăng chất lượng của các hệ thống dịch máy trên miền này.

2) So sánh kết quả dịch giữa hệ thống huấn luyện trên kho ngữ liệu với Google Translate:

Chúng tôi làm thực nghiệm trên kho ngữ liệu văn bản có tổng cộng 177688 cặp câu, bao gồm các lĩnh vực sau:

Bảng V
ĐIỂM BLEU CỦA 17 HỆ THỐNG DỊCH MÁY KHI DỊCH CÁC TẬP KIỂM TRA GỒM 10000 CÂU TIẾNG VIỆT SANG TIẾNG ANH, SO VỚI HỆ THỐNG DỊCH MÁY GOOGLE

	Hệ thống dịch máy	Google Translate	(+)
1	21,78	16,83	4,95
2	21,46	17,77	3,69
3	23,14	18,75	4,39
4	21,25	17,22	4,03
5	20,29	16,30	3,99
6	21,67	17,92	3.75
7	21,58	16,92	4,66
8	21,66	18,93	2,73
9	21,38	18,72	2,66
10	21,60	18,41	3,19
11	23,65	18,40	5,61
12	22,06	18,63	3.43
13	24,99	20,08	4,91
14	24,20	18,43	5,77
15	23,50	17,97	5,53
16	25,18	17,77	7,41
17	24,45	17,57	6,88
TB	**22,58**	**18,02**	**4,56**

- Văn bản lĩnh vực Luật: 30258 cặp câu[1];
- Văn bản lĩnh vực Tin học: 19705 cặp câu[1];
- Văn bản lĩnh vực Xã hội: 84613 cặp câu[1];
- Văn bản lĩnh vực Kinh thánh: 31102 cặp câu[1];
- Văn bản lĩnh vực Du lịch: 12010 cặp câu (do tác giả thu thập được).

Chúng tôi thực hiện kiểm tra chéo như sau. Lần lượt giữ lại 10000 cặp câu để làm tệp kiểm tra, lấy trải đều trên mỗi lĩnh vực văn bản, cụ thể là: 1700 cặp câu văn bản lĩnh vực Luật; 1100 cặp câu văn bản lĩnh vực Tin học; 4700 cặp câu văn vản lĩnh vực Xã hội; 1700 cặp câu văn bản lĩnh vực Kinh thánh; 800 cặp câu văn bản lĩnh vực Du lịch. Chúng tôi thu được 17 tệp kiểm tra, mỗi tệp gồm 10000 cặp câu, và 17 tệp dùng huấn luyện Hệ thống dịch, mỗi tệp gồm 167688 cặp câu.

Dùng MOSES huấn luyện 17 hệ thống dịch máy trên 17 kho ngữ liệu với 167688 cặp câu còn lại. Kết quả điểm BLEU của 17 hệ thống dịch máy khi dịch tệp kiểm tra, so với hệ thống dịch máy Google được thể hiện trong Bảng V.

Hệ thống huấn luyện trên kho ngữ liệu 167688 cặp câu song ngữ Anh–Việt trung bình đạt cao hơn 4,6 điểm BLEU so với hệ thống dịch máy Google hiện nay.

3) Phân tích kết quả hệ thống dịch:
Hệ thống dịch có một số hạn chế sau:

1) Hệ thống không dịch được những từ không biết;
2) Cấu trúc ngữ pháp của một số câu dịch chưa đúng;
3) Không dịch được theo ngữ cảnh của văn bản dịch;
4) Tách từ tiếng Việt bị sai;
5) Đa số các câu được dịch ở đầu ra chưa được trôi chảy.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi đã xây dựng được kho ngữ liệu du lịch song ngữ Việt–Anh được gióng hàng câu chính xác với trên 12000 cặp câu, sẵn sàng chia sẻ cho cộng đồng nghiên cứu xử lí ngôn ngữ tự nhiên. Chúng tôi cũng đã cải tiến công cụ gióng hàng XAlign sẵn có và thu được được công cụ gióng hàng câu tự động Việt–Anh viXAlign đạt mức độ chính xác khoảng 90%, cao hơn khoảng 7% so với các công cụ gióng hàng câu Việt–Anh hiện nay. Công cụ này được chia sẻ tại địa chỉ https://github.com/viXAlign/viXAlign-project. Chúng tôi cũng đã dùng kho ngữ liệu này để nâng cao chất lượng hệ thống dịch máy thống kê, thực nghiệm cho điểm BLEU đã tăng lên 8,79 so với hệ thống chỉ dùng ngữ liệu trên miền tổng quát gồm 165678 cặp câu để huấn luyện.

Chúng tôi cũng đã tiến hành đánh giá hệ thống dịch máy được huấn luyện trên kho ngữ liệu song ngữ Việt–Anh với 167688 cặp câu. Kết quả cho thấy, hệ thống dịch máy mà chúng tôi huấn luyện đạt cao hơn 4,6 điểm BLEU so với hệ thống dịch máy Google hiện nay. Mặc dù sự so sánh này có thể nói là thiếu công bằng vì hai hệ thống sử dụng nguồn tài nguyên khác nhau, nhưng kết quả cũng giúp chúng ta thấy rằng kết quả dịch của Google Translate còn phải cải thiện khá nhiều.

Trong thời gian tới, chúng tôi tập trung vào việc xây dựng kho ngữ liệu du lịch song ngữ Việt–Anh để có được kho ngữ liệu chất lượng, lớn về số lượng và đa dạng về chủ đề, đồng thời chia sẻ kho ngữ liệu này cùng với công cụ gióng hàng câu tự động cho cộng đồng nghiên cứu. Chúng tôi cũng sẽ tiếp tục nghiên cứu cải tiến công cụ gióng hàng câu tự động Việt–Anh để tăng mức độ chính xác. Song song với việc xây dựng tài nguyên, chúng tôi thực hiện phân tích lỗi của hệ thống dịch để đưa ra giải pháp khắc phục đồng thời nghiên cứu đề xuất các giải pháp nhằm nâng cao hơn nữa chất lượng dịch của hệ thống dịch máy Việt–Anh trên miền du lịch.

TÀI LIỆU THAM KHẢO

[1] P. Koehn, *MOSES Statistical Machine Translation System User Manual and Code Guide*. references, September 19, 2016. [Online]. Available: https://vlsp.hpda.vn/demo/?page=resources

[2] N. Quoc-Hung and W. Winiwarter, "Building an english-vietnamese bilingual corpus for machine translation," *International Conference on Asian Language Processing*, pp. 157–160, 2012.

[3] Đinh Điền and L. N. Minh, "Ứng dụng ngữ liệu song ngữ anh-việt trong giảng dạy ngôn ngữ," *hội thảo Liên ngành NNH Ứng dụng và Giảng dạy Ngôn ngữ*, pp. 559–567, 11/2015.

[4] M. M.Sakre, M. M.Kouta, and A. M.N.Allam, "automated construction of arabic-english parallel corpus," *Arab World English Journal (AWEJ) Special Issue on Translation*, vol. No.5, May, 2016.

[5] P. F. Brown, J. C. Lai, and R. L.Mercer, "Aligning sentences in parallel corpora," *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics (ACL)*, 1991.

[6] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," in *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics (ACL)*, 1991.

[7] M. Kay and M. Röscheisen, "Text-translation alignment," in *Computational Linguistics*, 1993.

[8] S. F. Chen, "Aligning sentences in bilingual corpora using lexical information," *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, 1993.

[9] M. Simard and P. Plamondon, "Bilingual sentence alignment: Balancing robustness and accuracy," *Proceedings of the Conference of the Association for Machine Translation in the Americas*, 1998.

[10] L. Romary and P. Bonhomme, "Parallel alignment of structured documents," *Jean Véronis. Parallel Text Processing, Kluwer Academic Publisher*, pp. 233–253, 2000.

[11] N. T. M. Huyền and M. Rossignol, "A language-independent method for the alignement of parallel corpora," *Proceedings of 20th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 2006.

[12] H.-L. Trieu, P.-T. Nguyen, and L.-M. Nguyen, "A new feature to improve moore's sentence alignment method," *VNU Journal of Science: Comp. Science & Com*, vol. Eng. Vol. 31. No. 1, p. 32–44, 2015.

[13] Y.-C. Chiao, O. Kraif, D. Laurent, T. M. H. Nguyen, and e. a. Nasredine Semmar, "Evaluation of multilingual text alignment systems: the arcade ii project," *5th international Conference on Language Resources and Evaluation - LREC'06, May 2006, Genoa/Italy*, 2006.

[14] K. Papineni, S. Roukos, T. Ward, , and W.-J. Zhu, "Leu: a method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia*, pp. 311–318, July 2002.

Nguyễn Tiến Hà sinh năm 1977 tại Vĩnh Phúc. Tác giả tốt nghiệp Trường Đại học Sư phạm Hà Nội năm 2005; nhận bằng Thạc sĩ tại Trường Đại học Khoa học Tự Nhiên, Đại học Quốc gia Hà Nội, năm 2010. Hiện nay, tác giả đang công tác tại Trung tâm Giáo dục Thường xuyên tỉnh Phú Thọ và là nghiên cứu sinh tiến sĩ tại Trường Đại học Khoa học Tự Nhiên, Đại học Quốc gia Hà Nội. Lĩnh vực nghiên cứu của tác giả là Xử lý ngôn ngữ tự nhiên.

Nguyễn Thị Minh Huyền sinh năm 1973 tại Hà Nội. Tác giả tốt nghiệp Trường Đại học Tổng hợp Hà Nội năm 1994; nhận bằng Thạc sĩ và Tiến sĩ tại Trường Đại học Nancy 1, Cộng hòa Pháp vào các năm 1999 và 2006. Hiện nay, tác giả đang công tác tại Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội. Lĩnh vực nghiên cứu của tác giả là Xử lý ngôn ngữ tự nhiên.

Nguyễn Minh Hải sinh năm 1986 tại Ninh Bình. Tác giả nhận bằng Cử nhân và Thạc sĩ tại Trường Đại học Khoa học Tự Nhiên, Đại học Quốc gia Hà Nội vào các năm 2013 và 2016. Hiện nay, tác giả đang công tác tại Trường Đại học Khoa học Tự Nhiên, Đại học Quốc gia Hà Nội. Lĩnh vực nghiên cứu của tác giả là Xử lý ngôn ngữ tự nhiên.