ĐÁNH GIÁ VAI TRÒ CỦA KHO NGỮ LIỆU ĐỐI VỚI CHẤT LƯỢNG DỊCH TỰ ĐỘNG TIẾNG VIỆT

EVALUATION OF THE ROLE OF CORPUS IN VIETNAMESE-RELATED MACHINE TRANSLATION QUALITY

Nguyễn Văn Bình[1], Huỳnh Công Pháp[1]

[1]Trường Đại học Công nghệ Thông tin và Truyền thông Việt-Hàn - Đại học Đà Nẵng
nvbinh@vku.udn.vn; hcphap@vku.udn.vn

(Nhận bài: 30/11/2020; Chấp nhận đăng: 10/01/2021)

Tóm tắt - Chất lượng của các hệ thống dịch tự động tiếng Việt hiện nay vẫn còn thấp khi so sánh với chất lượng dịch của các cặp ngôn ngữ phổ biến khác. Có nhiều yếu tố ảnh hưởng đến chất lượng của mô hình dịch, trong đó có phương pháp dịch và kho ngữ liệu. Để xây dựng một hệ thống dịch có chất lượng tốt, cần sử dụng kho ngữ liệu tốt về chất lượng và có số lượng lớn. Bài báo này tiến hành nghiên cứu thực trạng của các kho ngữ liệu song ngữ tiếng Việt hiện nay và tổ chức xây dựng các hệ thống dịch Anh-Việt từ các kho ngữ liệu có kích thước khác nhau, sử dụng các phương pháp dịch khác nhau. Kết quả đánh giá chất lượng của các hệ thống dịch thu được cho thấy, khi sử dụng kho ngữ liệu có kích thước càng lớn thì chất lượng của hệ thống dịch càng tăng.

Từ khóa - Dịch tự động; kho ngữ liệu; kho ngữ liệu song ngữ; chất lượng dịch tự động; đánh giá chất lượng hệ thống dịch.

Abstract - The quality of current Vietnamese-related automatic translation systems is still low when compared with the translation quality of other popular language pairs. There are many factors that affect the quality of the translation model, including the translation method and the corpus. To build a good quality translation system, it is necessary to use good quality and large quantity of linguistic resources. This article researches the current situation of Vietnamese bilingual corpus and builds the English-Vietnamese translation systems from corpus of different sizes, using other translation methods. The results of the quality of the translation systems obtained show that, when using the larger corpus size, the quality of the translation system is increased.

Key words - Machine translation; corpus; bilingual corpus; machine translation quality; evaluation of machine translation.

1. Đặt vấn đề

Dịch tự động hay còn gọi là dịch máy nghiên cứu việc sử dụng phần mềm để dịch văn bản từ một ngôn ngữ sang ngôn ngữ khác [1], chẳng hạn dịch một văn bản từ tiếng Anh sang tiếng Việt. Bộ máy dịch tự động là một chương trình máy tính có nhiệm vụ tiếp nhận văn bản ở ngôn ngữ nguồn, sau đó bằng các thuật toán toán của mình để đưa ra dự đoán kết quả dịch ở ngôn ngữ đích. Các thuật toán ở bài toán dịch tự động hoạt động trên cơ sở tổng hợp và xử lý tri thức từ ngôn ngữ tự nhiên, chẳng hạn thông qua từ điển, các cặp câu dịch mẫu; Các luật ngữ pháp; Thống kê từ ngữ…

Như vậy, có thể thấy rằng, để tạo nên một hệ thống dịch có chất lượng, cần có hai yếu tố then chốt là nguồn dữ liệu và phương pháp dịch. Nguồn dữ liệu sử dụng cho bộ máy dịch thuật phải đáp ứng: (1) Có chất lượng tốt, nghĩa là dữ liệu phải chính xác, ngữ nghĩa không nhập nhằng, có phân tích cú pháp, xác định ranh giới từ, xác định danh từ riêng…; (2) Có số lượng lớn, nghĩa là có đầy đủ các luật về ngữ pháp, có số lượng các cặp câu song ngữ lớn, bao phủ tất cả các lĩnh vực, có đầy đủ các từ, cụm từ trong ngôn ngữ tự nhiên.

Để giải quyết các bài toán xử lý ngôn ngữ tiếng Việt, trong đó có dịch máy, nhiều nhóm nghiên cứu đã xây dựng được các kho ngữ liệu dành riêng cho tiếng Việt, đồng thời đưa ra các giải pháp để nâng cao chất lượng của các kho ngữ liệu. Tuy nhiên, số lượng dữ liệu của các kho ngữ liệu hiện nay còn khá ít, đồng thời chưa có những đánh giá, so sánh cụ thể để có số liệu minh chứng sự ảnh hưởng của chất lượng và số lượng kho ngữ liệu đến chất lượng các hệ thống dịch.

Bài báo này sẽ nêu thực trạng của các hệ thống dịch máy hiện nay và các nghiên cứu cũng như kết quả xây dựng kho ngữ liệu. Sau đó, nghiên cứu sẽ thực hiện tổng hợp một kho ngữ liệu lớn và sử dụng để tổ chức thực nghiệm xây dựng hệ thống dịch đối với các bộ dữ liệu khác nhau và các phương pháp dịch khác nhau nhằm đánh giá vai trò của kho ngữ liệu đối với chất lượng của các hệ thống dịch tự động đối với cặp ngôn ngữ tiếng Anh và tiếng Việt. Kết quả nghiên cứu của bài báo có vai trò quan trọng trong việc cải tiến chất lượng các hệ thống dịch tự động và chất lượng các nguồn tài nguyên dữ liệu xử lý ngôn ngữ tự nhiên.

2. Thực trạng về chất lượng của các hệ thống dịch tiếng Việt hiện nay

Với các ngôn ngữ quốc tế, đã có nhiều nghiên cứu đánh giá chất lượng của các hệ thống dịch hiện nay. Khi so sánh giữa bản dịch của máy tính và bản dịch do con người thực hiện, nghiên cứu tại [2] cho thấy, các hệ thống dịch máy chỉ cho kết quả dịch tốt khi dịch các từ riêng lẻ hoặc các cụm từ, còn đối với các câu dài và phức tạp sẽ cho kết quả kém. Đối với dịch thuật trong lĩnh vực chuyên môn, nghiên cứu tại [3] tiến hành đánh giá việc sử dụng các hệ thống dịch trong lĩnh vực y tế. Kết quả cho thấy, chỉ có 57,7% câu dịch cho kết quả chính xác, nhiều câu vô nghĩa hoặc cho kết quả hoàn toàn sai với nội dung gốc. Điều đó cho thấy, sự hạn chế ở các hệ thống dịch tự động hiện nay khi dịch trong các chủ đề chuyên ngành hẹp.

Có nhiều nghiên cứu của các tác giả trong và ngoài nước trong lĩnh vực dịch tự động liên quan đến tiếng Việt. Các nhà khoa học đã đề xuất các giải pháp nhằm nâng cao chất lượng của dịch máy tiếng Việt, trong đó bao gồm các

[1] The University of Danang - Vietnam-Korea University of Information and Communication Technology (Nguyen Van Binh, Huynh Cong Phap)

giải pháp cải tiến mô hình dịch cũng như xây dựng và cải tiến kho ngữ liệu phục vụ hệ thống dịch. Bên cạnh đó, còn có nhiều thực nghiệm xây dựng hệ thống dịch tự động tiếng Anh sang tiếng Việt bằng các mô hình dịch khác nhau.

Việc phát triển một hệ thống dịch tự động từ tiếng nước ngoài ra tiếng Việt được bắt đầu nghiên cứu vào những năm 60 thế kỉ 20. Đến nay, có một số sản phẩm dịch máy được ứng dụng nhưng cho chất lượng dịch còn nhiều hạn chế do sự khác biệt về mặt cấu trúc ngữ pháp và tính nhập nhằng về ngữ nghĩa trong ngôn ngữ tiếng Việt. Một số hệ thống dịch được đưa ra làm sản phẩm thương mại như EVTran được nghiên cứu và phát triển từ năm 1989, Cồ Việt của Công ty Cổ phần Tin học Lạc Việt, Google Translation, Bing Translator… [4].

Vấn đề nâng cao chất lượng các hệ thống dịch tự động là một bài toán luôn được các nhà nghiên cứu tập trung giải quyết. Trong hơn 20 năm phát triển gần đây của lĩnh vực dịch máy, tuy đã có những bước phát triển đáng kể nhưng đến nay kết quả của các hệ thống dịch máy vẫn còn là một khoảng cách xa so với các bản dịch do con người thực hiện. Đối với các ngôn ngữ phổ biến như tiếng Anh, tiếng Pháp, các hệ thống cho ra bản dịch có thể chấp nhận được trong một số lĩnh vực thông dụng, có thể sử dụng để tham khảo nghĩa của ngôn ngữ đích mà không cần đến người phiên dịch. Tuy nhiên, đối với các ngôn ngữ ít phổ biến như tiếng Việt, chất lượng các câu dịch của hệ thống còn thấp, khó có thể áp dụng rộng rãi trong thực tế. Đặc biệt ở các lĩnh vực chuyên ngành như y tế, kỹ thuật, công nghệ, văn bản quy phạm pháp luật… các hệ thống dịch không dịch đúng các khái niệm chuyên môn nên nhiều văn bản dịch trở nên khó hiểu, không có giá trị. Dưới đây là một ví dụ được trích từ nhiều kết quả qua khảo sát thực tế đối với một số bộ dữ liệu cụ thể:

Câu nguồn (bản gốc từ Hiệp định TPP):	disputing Party means a complaining Party or a responding Party; Panel means a panel established pursuant to Article 28.7 (Establishment of a Panel);
Câu tham chiếu:	*Bên tranh chấp* là *Bên nguyên đơn* hoặc *Bên bị đơn*; *Ban hội thẩm* là ban được thành lập căn cứ theo Điều 28.7 (Thành lập Ban hội thẩm);
Câu dịch bởi hệ thống Google Translation	Bên tranh chấp có nghĩa là một *Bên khiếu nại* hoặc một *Bên đáp ứng*; Ban Hội thẩm là ủy ban được thành lập theo Điều 28.7 (Thành lập Ban Hội thẩm);
Câu dịch bởi hệ thống Bing Translator	*bên đảng* có nghĩa là một *bên khiếu nại* hoặc một *bên responding*; *Bảng điều khiển* có nghĩa là một *bảng điều khiển* được thành lập theo quy định bài 28.7 (thành lập một bảng điều khiển).

Tổ chức đánh giá chất lượng kết quả dịch từ tiếng Anh sang tiếng Việt bằng phương pháp chủ quan (sử dụng bảng đánh giá ở các mức độ khác nhau và do con người thực hiện) với một tập dữ liệu gồm 984 câu ở lĩnh vực hội thoại hàng ngày, kết quả thu được ở Bảng 1.

Bảng 1. *Kết quả đánh giá chất lượng hệ thống dịch bằng phương pháp chủ quan*

	Số câu	(1) Có hiểu	(2) Hiểu đúng	(3) Dùng được
Google	984	789	687	516
Microsoft	984	517	458	308

Với kết quả trên có thể thấy rằng, để dùng được kết quả dịch trong giao tiếp thông thường, chỉ có 516 câu (đối với Google) và 308 câu (đối với Microsoft), chiếm tỷ lệ là 52% và 30%. Quan sát cụ thể dữ liệu, có nhiều câu còn làm cho người đọc hiểu sai ý nghĩa của bản gốc.

Qua các đánh giá ở trên, có thể thấy rằng, mặc dù các hệ thống dịch tự động hiện nay đã được ứng dụng rất rộng rãi, nhưng để sử dụng được kết quả dịch cần phải tiếp tục có nhiều cải tiến, đặc biệt đối với dịch tiếng Việt. Chất lượng các hệ thống dịch tiếng Việt chưa tốt bởi một số nguyên nhân:

- Phương pháp dịch chưa phù hợp: Các mô hình dịch thống kê hoặc dịch dựa trên mạng nơ ron có nhiều ưu điểm, nhưng muốn áp dụng hiệu quả đối với dịch tiếng Việt cần có thêm các đánh giá và nghiên cứu bổ sung. Tiếng Việt khác với một số ngôn ngữ khác, mỗi từ bao gồm nhiều âm tiết, trong khi các hệ thống đều làm việc trên đơn vị từ đơn lẻ, vì vậy sẽ làm giảm hiệu quả của các mô hình dịch này. Các công cụ xử lý dành cho tiếng Việt đã được nghiên cứu và áp dụng như công cụ tách từ vnTokenizer, Đông Du, công cụ phân tích cú pháp, công cụ gán nhãn từ loại VnTagger, tuy nhiên vẫn còn một số hạn chế. Các hệ thống dịch hiện nay đang xem xét câu nguồn để tái tạo câu đích mà chưa đặt văn bản dịch vào ngữ cảnh nên nhiều câu dịch không phù hợp khi áp dụng vào thực tế. Bên cạnh đó, sự nhập nhằng về ngữ nghĩa trong tiếng Việt là một vấn đề cần nghiên cứu và có giải pháp xử lý để có được ý nghĩa rõ ràng ở các văn bản tiếng Việt trước khi được hệ thống dịch.

- Kho ngữ liệu chưa đầy đủ: Các kho ngữ liệu sử dụng để huấn luyện cho các hệ thống dịch tự động chưa đầy đủ, số lượng dữ liệu còn ít, vì vậy một số từ các hệ thống chưa nhận diện được. Đặc biệt trong các lĩnh vực chuyên ngành hẹp, như lĩnh vực y tế, kỹ thuật, văn bản hành chính… các khái niệm quan trọng nhưng các hệ thống vẫn chưa dịch đúng làm cho bản dịch trở nên khó hiểu.

3. Thực trạng các kho ngữ liệu tiếng Việt dùng trong lĩnh vực dịch tự động

Kho ngữ liệu (corpus) được hiểu là tập hợp văn bản đơn ngữ, đa ngữ hay song ngữ [5]. Kho ngữ liệu song song (Parallel Corpus) là một tập các văn bản (tài liệu) trong nhiều ngôn ngữ khác nhau, trong đó có một ngôn ngữ nguồn và một hoặc nhiều ngôn ngữ đích được dịch từ ngôn ngữ nguồn.

Kho ngữ liệu song ngữ là một tập hợp dữ liệu gồm các cặp văn bản đã được dịch tương ứng 1-1 về mặt ngữ nghĩa. Trong ngữ liệu song ngữ, các bản dịch tương ứng của mỗi ngôn ngữ phải được đặt song song với nhau hay còn được gọi là gióng hàng với nhau (alignment). Mức độ gióng hàng có thể ở cấp độ văn bản (text alignment), nghĩa là từng văn bản trong ngôn ngữ nguồn được gióng với văn bản dịch tương ứng trong ngôn ngữ đích. Tương tự cho cấp độ đoạn (paragraph alignment), cấp độ câu (sentence alignment), cấp độ ngữ (phrase alignment) và sâu nhất là cấp độ từ (word alignment).

Kho ngữ liệu song ngữ chứa các văn bản của hai ngôn ngữ khác nhau, vì vậy ngoài nội dung còn có các thông tin đã được xử lý như gióng hàng, gán nhãn từ… Về cơ bản, các kho ngữ liệu sẽ chứa những thông tin sau đây:

- Phần dữ liệu nguyên thủy/ thô (primary data);
- Thông tin về văn bản: id, title, authors…: Được gọi

là phần đầu (Header);
- Thông tin về cấu trúc và nội dung: Các phần (section), đoạn (paragraph), câu (sentence)…: Được gọi phần Text;
- Phần chú giải ngôn ngữ học (linguistic annotation);
- Ranh giới đoạn, câu, từ;
- Từ loại của từ (POS);
- Gốc từ (lemma);
- Thông tin về gióng hàng (alignment).

Trên thế giới hiện có rất nhiều kho ngữ liệu song ngữ song song được chia sẻ miễn phí cho cộng đồng nghiên cứu. Dưới đây là một vài kho ngữ liệu song ngữ song song tiêu biểu:

- Kho ngữ liệu song ngữ song song được xây dựng từ sự hỗ trợ của dự án EuroMatrix. Kho ngữ liệu này gồm các cặp ngôn ngữ khác nhau được lấy nguồn từ các kỷ yếu (proceeding) của Quốc hội Châu Âu (European Parliament) từ năm 1996 – 2006. Kho ngữ liệu song ngữ song song này gồm 10 cặp ngôn ngữ như được liệt kê trong Bảng 2.

Bảng 2. *Dữ liệu của kho ngữ liệu EuroMatrix*

Kho ngữ liệu song ngữ (L1-L2)	Số cặp câu	Số từ ở ngôn ngữ L1	Số từ ở ngôn ngữ L2
Tiếng Đan Mạch - Tiếng Anh	1.304.947	34.169.707	36.225.880
Tiếng Đức - Tiếng Anh	1.313.096	34.700.362	36.663.083
Tiếng Hy Lạp - Tiếng Anh	662.090	18.834.758	18.827.241
Tiếng Tây Ban Nha - Tiếng Anh	1.304.116	37.870.751	36.429.274
Tiếng Phần Lan - Tiếng Anh	1.257.720	24.895.790	34.802.617
Tiếng Pháp - Tiếng Anh	1.334.080	41.573.117	37.436.222
Tiếng Ý - Tiếng Anh	1.251.315	36.411.166	36.510.033
Tiếng Hà Lan - Tiếng Anh	1.326.412	36.784.168	36.690.392
Tiếng Bồ Đào Nha - Tiếng Anh	1.287.757	37.342.426	36.355.907
Tiếng Thụy Điển - Tiếng Anh	1.164.536	28.882.142	32.053.628

- Kho ngữ liệu song ngữ song song Anh-Pháp, Canadian Hansard Corpus, của hiệp hội dữ liệu ngôn ngữ học (Linguistic Data Consortium- LDC) kho ngữ liệu này gồm 2.8 triệu cặp câu [16]. Dữ liệu văn bản thuần chủ yếu được lấy từ trang web của Quốc hội Canada.

- Kho ngữ liệu song ngữ song song Hoa – Anh PKU 863 của đại học Bắc kinh gồm hơn 200.000 cặp câu thuộc nhiều lĩnh vực kinh tế xã hội khác nhau [17].

Bảng 3. *Tổng hợp các kho ngữ liệu đa ngôn ngữ*

Tên kho ngữ liệu	Số ngôn ngữ	Độ lớn dữ liệu
Europarl	21	30.32M
Wikipedia	21	25.90M
OpenSubtitles	62	3.35G
TED2013	15	3.81M
EUbookshop	48	173.20M

Ngoài ra, có một số kho ngữ liệu song ngữ với số lượng câu lớn ở nhiều ngôn ngữ khác nhau được chia sẻ cho cộng đồng nghiên cứu được cung cấp tại [13], [14] và được liệt kê trong Bảng 3.

Liên quan đến kho ngữ liệu tiếng Việt phục vụ cho bộ máy dịch tự động, có các nghiên cứu xây dựng và cải tiến kho ngữ liệu. Một số kho ngữ liệu song ngữ Anh – Việt được tổng hợp trong Bảng 4.

Bảng 4. *Tổng hợp một số kho ngữ liệu song ngữ Anh – Việt*

Đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt" (VLSP) [6]	80.000 cặp câu Kinh tế - Xã hội; 20.000 cặp câu Tin học
Xây dựng hệ thống dịch dựa trên ví dụ [7]	6.000 cặp câu song ngữ Anh-Việt
Xây dựng hệ thống dịch thích ứng miền trong dịch máy nơ ron cho cặp ngôn ngữ Anh - Việt [8]	100.000 cặp câu song ngữ Anh Việt thuộc miền pháp lý
Xây dựng hệ thống dịch 2 chiều Anh – Việt bằng mô hình dịch thống kê sử dụng Moses [9]	Kho ngữ liệu gồm 880.000 cặp câu song ngữ Anh – Việt

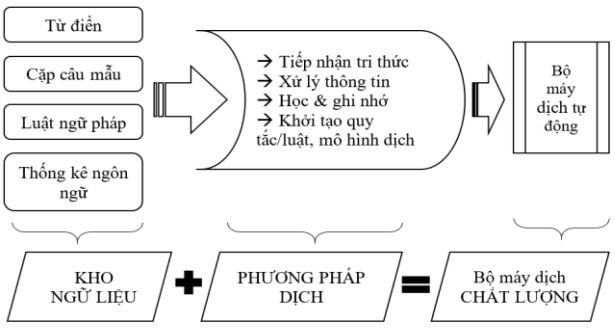
Có thể thấy rằng, các kho ngữ liệu tiếng Việt có số lượng câu rất ít khi so sánh với các kho ngữ liệu ở các ngôn ngữ phổ biến khác. Trong các kho ngữ liệu song ngữ tiếng Việt, dữ liệu được tổng hợp ở nhiều lĩnh vực khác nhau. Lượng dữ liệu đối với từng lĩnh vực chưa nhiều, đặc biệt dữ liệu thuộc các lĩnh vực hẹp, chuyên sâu như lĩnh vực y tế, văn bản quy phạm pháp luật… hầu như xuất hiện rất ít trong các kho ngữ liệu nói trên.

Ngoài ra, các kho ngữ liệu song ngữ hiện nay chủ yếu giữa cặp ngôn ngữ tiếng Anh và tiếng Việt, có rất ít kho ngữ liệu song ngữ giữa tiếng Việt với các ngôn ngữ khác được nghiên cứu và xây dựng.

Kho ngữ liệu là nền tảng để xây dựng, đánh giá và cải tiến chất lượng của các hệ thống dịch tự động. Nếu có được kho ngữ liệu đa ngữ đủ lớn về khối lượng, tốt về chất lượng thì chắc chắn chất lượng dịch của các hệ thống dịch tự động hiện nay sẽ được cải thiện đáng kể.

4. Đánh giá vai trò của kho ngữ liệu đến chất lượng hệ thống dịch Anh – Việt

Đối với bài toán xây dựng hệ thống dịch tự động và nâng cao chất lượng của hệ thống dịch, kho ngữ liệu đóng vai trò then chốt vì đó là dữ liệu đầu vào để thực hiện quá trình huấn luyện hệ thống dịch thông qua các phương pháp khác nhau. Vai trò của kho ngữ liệu trong bài toán dịch tự động được thể hiện trong Hình 1.



Hình 1. *Các thành phần quyết định chất lượng của hệ thống dịch tự động*

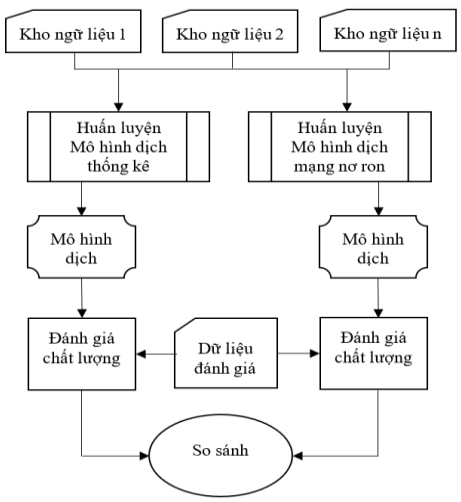
Đối với các cặp ngôn ngữ phổ biến như tiếng Anh-Pháp, đã có các công trình nghiên cứu chứng minh sự ảnh hưởng và mối quan hệ mật thiết giữa chất lượng và khối lượng của kho ngữ liệu với chất lượng dịch của các hệ thống dịch tự động [10]. Tuy nhiên, đối với tiếng Việt hiện nay vẫn chưa có các số liệu đánh giá chi tiết để thể hiện vai trò của kho ngữ liệu đối với các mô hình dịch khác nhau.

Hiện nay, các nghiên cứu liên quan đến bài toán dịch tự động chủ yếu tập trung ở hai phương pháp: (1) Phương pháp dịch thống kê; (2) Phương pháp dịch sử dụng mạng nơ ron. Các hệ thống dịch tự động được sử dụng rộng rãi như Google Translation, Bing Translate… cũng đang sử dụng các mô hình dịch này và cho kết quả dịch khá tốt so với các phương pháp dịch trước đây. Về cơ bản, các phương pháp dịch này sử dụng số lượng lớn dữ liệu về các cặp câu đã được dịch, từ đó sử dụng các mô hình học máy để huấn luyện và tạo ra mô hình dịch phù hợp.

Các nghiên cứu liên quan đến xây dựng và cải tiến hệ thống dịch tự động tiếng Việt đã có nhiều thực nghiệm trên các bộ dữ liệu khác nhau, với kích thước kho ngữ liệu ở nhiều mức độ về mặt số lượng. Chính vì vậy, khó có thể có cái nhìn tổng thể về vai trò của khối lượng kho ngữ liệu trong vấn đề chất lượng của hệ thống dịch tiếng Việt.

Trong nghiên cứu này, nhóm tác giả sẽ thực nghiệm xây dựng hệ thống dịch với kho ngữ liệu có độ lớn khác nhau, sau đó triển khai đánh giá mô hình dịch thu được trên cùng một bộ dữ liệu đánh giá để xem xét chất lượng của mô hình dịch này. Nghiên cứu thực hiện đối với cặp ngôn ngữ Anh – Việt, xây dựng bộ máy dịch từ tiếng Anh sang tiếng Việt.

Quy trình xây dựng hệ thống và triển khai đánh giá thể hiện trong Hình 2.



Hình 2. *Sơ đồ tổ chức đánh giá*

Bước 1: Chuẩn bị kho ngữ liệu.

Để chuẩn bị kho ngữ liệu phục vụ xây dựng hệ thống dịch, chúng tôi đã sử dụng các phương pháp trích rút dữ liệu từ các tài liệu song ngữ tin cậy như các website song ngữ, tài liệu học tập, các văn bản song ngữ đã được số hóa. Sau khi tổng hợp được các cặp câu song ngữ Anh – Việt, thực hiện các bước tiền xử lý văn bản, bao gồm chuyển font chữ về định dạng unicode, loại bỏ các cặp câu trùng lặp, xử lý các ký tự đặc biệt…

Kho ngữ liệu thu được để sử dụng để huấn luyện và kiểm thử: Gồm 500.000 cặp câu song ngữ Anh – Việt ở tất cả các lĩnh vực. Chi tiết về dữ liệu thể hiện trong Bảng 5.

Bảng 5. *Mô tả dữ liệu dùng cho hệ thống dịch*

	Số lượng cặp câu	Độ dài câu tiếng Anh	Độ dài câu tiếng Việt
Dữ liệu huấn luyện	500.000	22,16	23,48
Dữ liệu đánh giá chất lượng hệ thống dịch	2.000	20,70	22,14

Để huấn luyện hệ thống dịch thống kê, nhóm tác giả sử dụng tỷ lệ dữ liệu cho bộ dữ liệu huấn luyện, bộ dữ liệu điều chỉnh tham số và bộ dữ liệu đánh giá tương ứng là 70%-10%-20%. Ngoài ra, đối với hệ thống dịch thống kê Moses, nghiên cứu sử dụng 2.241.987 câu tiếng Việt được thu thập từ các trang báo điện tử để làm kho ngữ liệu đơn ngữ phục vụ huấn luyện mô hình ngôn ngữ tiếng Việt.

Bước 2: Xây dựng hệ thống dịch và huấn luyện mô hình dịch.

Nghiên cứu sử dụng hai mã nguồn nổi tiếng nhất liên quan đến phương pháp dịch thống kê và phương pháp dịch sử dụng mạng nơ ron là Moses và OpenNMT.

- Moses [11] là hệ dịch máy thống kê cho phép người dùng dễ dàng tạo ra mô hình dịch cho bất cứ một cặp ngôn ngữ nào. Moses cung cấp cả hai loại mô hình dịch là dựa trên cụm từ và dựa trên cây. Nó bao gồm đầy đủ các thành phần để tiền xử lý dữ liệu, huấn luyện mô hình ngôn ngữ và mô hình dịch. Moses thực chất là phiên bản cao hơn của Pharaoh, là phần mềm được nhiều trường đại học, nhóm nghiên cứu nổi tiếng về xử lý ngôn ngữ tự nhiên và dịch máy thống kê như Edinburgh (Scotland), RWTH Aachen (Germany)... [5] tham gia phát triển. Đây là phần mềm có chất lượng khá tốt, khả năng mở rộng cao được dùng để xây dựng nhiều hệ thống dịch thử nghiệm cho nhiều cặp ngôn ngữ như Anh-Czech, Anh-Trung, Anh-Pháp… Để triển khai hệ thống Moses, nghiên cứu sử dụng SRILM toolkit [18] để xây dựng mô hình ngôn ngữ, sử dụngGIZA++ [19] để gióng hàng trong quá trình huấn luyện mô hình và dự đoán câu dịch. Các công cụ mã nguồn mở, tài nguyên, tài liệu, kho ngữ liệu liên quan đến dịch máy thống kê được chia sẻ tại website http://statmt.org.

- OpenNMT [12] là là hệ dịch sử dụng mạng nơ ron mã nguồn mở hoàn thiện, nổi tiếng, được công bố năm 2017 của nhóm Harvard NLP và SYSTRAN, công cụ này được nhiều nhóm nghiên cứu sử dụng trong cộng đồng dịch máy. OpenNMT ứng dụng các thuật toán mới nhất trong dịch tự động và đang tiếp tục được các nhà nghiên cứu phát triển. OpenNMT được xây dựng dựa trên các nghiên cứu cải tiến mô hình NMT truyền thống, cho phép mô hình dịch tự động quan sát toàn bộ chuỗi đầu vào để khởi tạo những từ mới ở đầu ra, cho các kết quả tốt khi dịch các câu dài. Đồng thời, OpenNMT cho phép tối ưu hóa bộ nhớ, tăng tốc độ tính toán khi sử dụng bộ xử lý đồ họa GPU.

Quá trình cài đặt và huấn luyện với các mã nguồn này, nghiên cứu sử dụng các tham số mặc định đã được khuyến nghị với mục đích nhận được sự nhất quán của kết quả.

Môi trường cài đặt:

- Phần mềm: Hệ điều hành Ubuntu 16.04, 64 bit;

- Phần cứng: Intel(R) Xeon(R) CPU E3-1220 v3 @ 3.10GHz, RAM 8Gb, GPU GeForce GTX 750 Ti/PCIe/SSE2.

Kết quả của bước 2 là các mô hình dịch đã được huấn luyện theo hai phương pháp đã đề xuất.

Bước 3: Đánh giá chất lượng của mô hình dịch nhận được

Từ mô hình dịch đã nhận được ở bước 2, tiến hành đánh giá chất lượng của hệ thống dịch bằng cách sử dụng cùng một bộ dữ liệu đầu vào bằng tiếng Anh gồm 2.000 câu để nhận được bản dịch tương ứng. Bản dịch nhận được sẽ được so sánh với bản dịch chuẩn thông qua chỉ số BLEU.

Ở đây, BLEU [13] là một chỉ số dùng để đánh giá chất lượng hệ thống dịch, có giá trị từ 0. Chỉ số BLEU càng cao thì hệ thống dịch càng đạt chất lượng tốt. Ý tưởng chính của phương pháp là so sánh kết quả bản dịch tự động bằng máy với một bản dịch chuẩn dùng làm bản đối chiếu. Quá trình so sánh được thực hiện thông qua việc thống kê sự trùng khớp của các từ trong hai bản dịch có tính đến thứ tự của chúng trong câu (phương pháp n-grams theo từ).

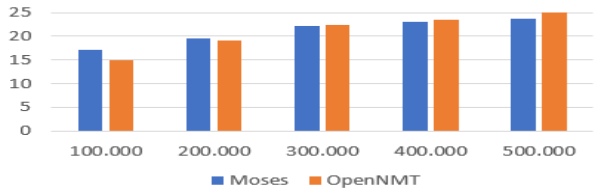
Sau khi thực nghiệm bằng bộ dữ liệu cụ thể nêu trên, chất lượng của các mô hình dịch nhận được ở Bảng 6.

Bảng 6. *Chất lượng các mô hình dịch nhận được*

Ngôn ngữ	Độ lớn kho ngữ liệu (số cặp câu)	Chất lượng (điểm BLEU)	
		Moses	OpenNMT
Anh → Việt	100.000	17,2	15,0
Anh → Việt	200.000	19,5	19,2
Anh → Việt	300.000	22,1	22,4
Anh → Việt	400.000	23,1	23,6
Anh → Việt	500.000	23,7	25,0

Từ bảng trên, chúng ta nhận được biểu đồ thể hiện các số liệu tương quan giữa độ lớn kho ngữ liệu và điểm chất lượng BLEU ở Hình 3.

Từ kết quả trên có thể nhận xét, khi xây dựng hệ thống dịch với kho ngữ liệu càng lớn thì chất lượng dịch càng tăng. Như vậy, rõ ràng chất lượng của kho ngữ liệu ảnh hưởng đến chất lượng của hệ thống dịch tự động Anh – Việt. Quan sát thực tế dữ liệu cũng có thể thấy rằng, khi số lượng lớn các cặp câu song ngữ làm dữ liệu đầu vào để huấn luyện mô hình dịch càng ít, thì kết quả dịch sẽ không đầy đủ và nhiều từ không được dịch, vì vậy chất lượng kết quả dịch sẽ giảm.



Hình 3. *So sánh tương quan giữa số lượng kho ngữ liệu và chất lượng hệ thống dịch*

5. Kết luận

Kết quả thực hiện đánh giá bằng phương pháp chủ quan cho thấy, chất lượng của các hệ thống dịch tiếng Việt hiện nay vẫn còn nhiều hạn chế. Qua thực nghiệm đánh giá đối với phương pháp dịch thống kê và phương pháp dịch sử dụng mạng nơ ron trên các kho ngữ liệu có kích thước khác nhau, có thể thấy rằng khối lượng của kho ngữ liệu đóng vai trò quan trọng ảnh hưởng đến chất lượng của kết quả hệ thống dịch tự động tiếng Việt. Khối lượng kho ngữ liệu càng lớn, chất lượng dịch sẽ càng tốt. Chính vì vậy, vấn đề nâng cao chất lượng và khối lượng của các kho ngữ liệu tiếng Việt hiện nay cần được quan tâm nghiên cứu nhằm góp phần xây dựng được các hệ thống dịch mà sản phẩm có thể ứng dụng được vào thực tiễn.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi Quỹ Phát triển Khoa học và Công nghệ - Đại học Đà Nẵng trong đề tài có mã số B2019-DN07-05.

TÀI LIỆU THAM KHẢO

[1] Hutchins, William John and Somers, Harold L, "An introduction to machine translation", *Academic Press London*, vol. 362, 1992.

[2] Haiying Li, Arthur C. Graesser and Zhiqiang Cai, "Comparison of Google Translation with Human Translation", *Proceedings of the Twenty - Seventh International Florida Artificial Intelligence Research Society Conference*, 2014.

[3] Sumant Patil, Patrick Davies, "Use of Google Translate in medical communication: Evaluation of accuracy", *BMJ: British medical journal*, December 2014.

[4] Đào Hồng Thu, "Xây dựng hệ thống dịch tự động tiếng Việt", *Tạp chí Ngôn ngữ và Đời sống*, vol. 11(157), 2008.

[5] Hồ Bảo Quốc, Đinh Điền, Đặng Bác Văn, Lương Vỹ Minh, Phạm Đào Duy Vũ, *Báo cáo kỹ thuật Xây dựng kho ngữ liệu song ngữ Anh – Việt*, Đề tài nhánh SP.74, Đề tài cấp nhà nước mã số KC.01.01.04/06-10, p. 46, 2009.

[6] Hệ thống trình diễn một số sản phẩm của nhánh đề tài "Xử lí văn bản" là một phần của đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt" (VLSP), https://vlsp.hpda.vn/demo/

[7] Minh Quang Nguyen, Dang Hung Tran and Thi Anh Le Pham, "Using example-based Machine Translation for English-Vietnamese Translation", *Faculty of IT Hanoi National University of Education*. Link: http://www.academia.edu/ download/ 20676214/MQEBMT .pdf, truy cập ngày 01/06/2020.

[8] Luan, Pham Nghia, Vinh, Nguyen Van, and Hoang, Nguyen Huy, "Thích ứng miền trong dịch máy nơ ron cho cặp ngôn ngữ Anh-Việt", *Kỷ yếu Hội nghị Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công Nghệ thông tin (FAIR)*, 2019.

[9] Phuoc, Nguyen Quang and Quan, Yingxiu and Ock, Cheol-Young, "Building a bidirectional English-Vietnamese statistical machine translation system by using MOSES", *International Journal of Computer and Electrical Engineering*, IACSIT Press, vol. 8(2), 2016, pp. 161-168.

[10] Boitet C., "Corpus pour la TA: types, tailles, et problèmes associés, selon leur usage et le type de système", *Revue française de linguistique appliquée*, vol. XII–2007, 2007, pp. 25-38.

[11] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N. & Dyer, C., "Moses: Open source toolkit for statistical machine translation", Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, 2007, pp. 177-180.

[12] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation", arXiv preprint arXiv:1701.02810, 2017

[13] Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing, "BLEU: a method for automatic evaluation of machine translation", *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, pp. 311-318.

[14] "Corpus-Based Language Studies", https://www.lancaster.ac.uk/fass/projects/corpus/, truy cập tháng 12/2020.

[15] "Linguistic Data Consortium", https://www.ldc.upenn.edu/, truy cập tháng 12/2020.

[16] Salim Roukos, David Graff, Dan Melamed, "Hansard French/English", https://catalog.ldc.upenn.edu/LDC95T20, truy cập tháng 12/2020.

[17] "Corpus-Based Language Studies", http://www.ling.lancs.ac.uk/corplang/863parallel/, truy cập tháng 12/2020.

[18] "SRILM - The SRI Language Modeling Toolkit", http://www.speech.sri.com/projects/srilm/, truy cập tháng 12/2020.

[19] "GIZA++: Training of statistical translation models.", http://www.statmt.org/moses/giza/GIZA++.html, truy cập tháng 12/2020.