

CẢI TIẾN CHẤT LƯỢNG DỊCH MÁY KẾT HỢP GIẢI PHÁP XÂY DỰNG KHO NGỮ LIỆU PHỤC VỤ ĐÁNH GIÁ CHẤT LƯỢNG HỆ THỐNG DỊCH TỰ ĐỘNG TIẾNG VIỆT

IMPROVING QUALITY OF MACHINE TRANSLATION COMBINING SOLUTIONS OF CREATING CORPORA FOR MACHINE TRANSLATION EVALUATION IN VIETNAMESE

Huỳnh Công Pháp, Nguyễn Văn Bình

Trường Cao đẳng Công nghệ Thông tin, Đại học Đà Nẵng; hcphap@cit.udn.vn, nvbinh@cit.udn.vn

Tóm tắt - Đánh giá và cải tiến chất lượng dịch máy, đặc biệt các hệ thống dịch tự động tiếng Việt ngày càng trở nên cấp bách bởi chất lượng dịch tự động hiện nay đạt chất lượng quá thấp so với yêu cầu thực tế trong thời kỳ hội nhập và “thế giới phẳng”. Hiện nay có nhiều phương pháp và độ đo khác nhau để đánh giá chất lượng các hệ thống dịch tự động, nhưng quá trình đánh giá chưa đóng góp dữ liệu để giúp cải thiện hệ thống dịch. Chất lượng của các hệ thống dịch tự động phụ thuộc rất lớn vào nguồn dữ liệu mà hệ thống sử dụng, cả về số lượng và chất lượng. Trong bài báo này, nhóm tác giả sẽ đề xuất giải pháp kết hợp việc đánh giá chất lượng các hệ thống dịch tự động với quá trình cải tiến chất lượng bản dịch máy và xây dựng kho ngữ liệu phục vụ đánh giá chất lượng các hệ thống dịch tự động tiếng Anh – tiếng Việt hiện nay.

Từ khóa - dịch máy; cải tiến chất lượng dịch máy; kho ngữ liệu; đánh giá chất lượng dịch máy; hiệu đính bản dịch

Abstract - Evaluation and improvement of machine translation quality, particularly in Vietnamese become increasingly urgent, because the current machine translation quality is too low compared to the actual requirements of the integration period and “flat world.” Currently there are many methods and scale to evaluate quality of machine translation systems, but the evaluation process does not contribute data to helping improve the system quality. The quality of machine translation systems highly depends on the data resources serving them in two aspects: quality and quantity. In this paper, we will propose measures to combine improving quality of machine translation and measures to create corpora for machine translation evaluation in Vietnamese.

Key words - machine translation; improve quality of machine translation; corpus; machine translation evaluation; post-editing

1. Đặt vấn đề

Hiện nay, dịch tự động hay dịch máy đã được sử dụng phổ biến trong cuộc sống, thậm chí có thể trợ giúp một cách hiệu quả cho quá trình dịch thuật.

Dịch tự động bằng máy tính nếu cho kết quả dịch tốt sẽ mang lại hiệu quả với chi phí bỏ ra ít, có thể dịch nhanh với khối lượng tài liệu lớn thuộc các lĩnh vực chuyên môn khác nhau. Khi đó các hệ dịch máy sẽ trở thành công cụ giúp con người tiếp cận kho tri thức khổng lồ viết bằng các ngôn ngữ khác.

Khi sử dụng một hệ thống dịch tự động, người dùng quan tâm đến chất lượng của bản dịch. Tuy nhiên, hiện nay, chất lượng dịch tự động giữa tiếng Việt với các ngôn ngữ khác khá thấp [11] nên kết quả dịch chủ yếu để tham khảo, nắm đại ý của văn bản. Trong một số trường hợp, bản dịch làm cho người đọc hiểu sai nội dung một phần hoặc toàn bộ nội dung chính của văn bản.

Trong lĩnh vực dịch tự động tiếng Việt, có nhiều nghiên cứu về các hệ thống dịch tự động [1] [3], tuy nhiên chưa có nghiên cứu cụ thể nào về việc đánh giá chất lượng các hệ thống dịch đang hoạt động hiện nay, trong đó có các hệ thống dịch phổ biến như Google Translator, Microsoft, EVTRANS, VDict... [4] đang được nhiều người sử dụng.

Để đánh giá chất lượng dịch tự động, đã có nhiều giải pháp và số đo được đề xuất và nghiên cứu [15] [13]. Trong đa số các giải pháp đánh giá chất lượng dịch, điều kiện cần thiết là phải có kho ngữ liệu với các đặc trưng riêng phục vụ cho mục đích đánh giá. Việc phụ thuộc vào các tham chiếu trong kho ngữ liệu để đánh giá kết quả dịch có thể không chính xác và mềm dẻo, đồng thời các chỉ số này vẫn chưa thể hiện được thời gian và sức lực con người cần sử dụng để chỉnh sửa bản dịch.

Trong khi đó, quá trình hiệu đính bản dịch (post-editing) được thực hiện trên hầu hết các kết quả dịch máy để cải tiến chất lượng, nhưng quá trình này thực hiện riêng biệt, không được tận dụng để đánh giá và nâng cao hiệu quả dịch tự động.

Do đó, ý tưởng mà chúng tôi đề xuất trong bài báo này là kết hợp đánh giá chất lượng các hệ thống dịch tự động với quá trình cải tiến chất lượng bản dịch máy và xây dựng kho ngữ liệu phục vụ đánh giá chất lượng các hệ thống dịch tự động tiếng Anh – tiếng Việt hiện nay.

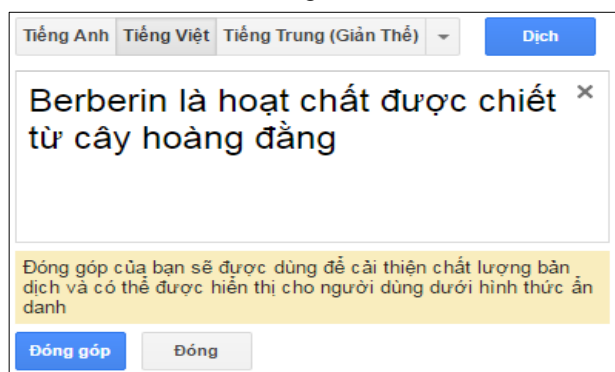
2. Các nghiên cứu liên quan

2.1. Hiệu đính bản dịch máy (Post-editing)

Trong lĩnh vực dịch tự động, hiệu đính là quá trình con người chỉnh sửa, hoàn thiện bản dịch của máy tính để cải tiến chất lượng và xây dựng các bản dịch chuẩn. Quá trình hiệu đính cũng thường được chia thành các mức độ khác nhau. Hiệu đính sơ bộ (Light post-editing) chỉ nhằm mục đích chỉnh sửa bản dịch để có thể hiểu được và tiết kiệm thời gian nhất có thể. Hiệu đính tổng quát (Full post-editing) sẽ cho ra bản dịch đúng văn phong, ngữ pháp và ngữ nghĩa. Trong thực tiễn và cả các nghiên cứu, hiệu đính kết quả dịch máy sẽ giúp quá trình dịch thuật tiết kiệm được nhiều thời gian và công sức hơn so với công việc dịch từ đầu, không sử dụng kết quả dịch máy [10]. Tuy nhiên quá trình hiệu đính giúp tiết kiệm được bao nhiêu thời gian còn phụ thuộc vào nhiều yếu tố. Một số thống kê thực tiễn khẳng định 40% [14], trong khi một số nghiên cứu khác đưa ra kết quả 15-20% thời gian tiết kiệm được từ quá trình hiệu đính [12].

Hiệu đính bản dịch máy cũng đã được ứng dụng trong một số hệ thống dịch tự động trực tuyến, như hệ thống

Google Translator. Hệ thống dịch tự động của Google đã có chức năng sửa chữa bản dịch và gửi lên hệ thống nhằm mục đích cải thiện chất lượng của bản dịch [Hình 1].



Hình 1. Giao diện chức năng chỉnh sửa bản dịch của hệ thống dịch tự động trực tuyến Google Translation

2.2. Kho ngữ liệu trong dịch tự động

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 [9].

Đã có nhiều kho ngữ liệu đa ngữ được nghiên cứu và công bố có số lượng ngôn ngữ và khối lượng dữ liệu tương đối lớn như EuroParl (11 ngôn ngữ, 34-55 triệu từ), JRC-Acquis (22 ngôn ngữ, 11-22 triệu từ), Xinhua News (2 ngôn ngữ, 12-14 triệu từ), EuroMatrix (9 ngôn ngữ lấy nguồn từ các kỳ yếu của Quốc hội châu Âu từ năm 1996–2006), Canadian Hansard (song ngữ Anh-Pháp, 2,8 triệu cặp câu), WaCky (hơn 1 tỷ từ được thu thập từ Internet)...

Kho ngữ liệu song ngữ tiếng Anh – tiếng Việt cũng đã được nhiều tổ chức, nhà khoa học nghiên cứu. Kho ngữ liệu tiếng Việt của Vietlex (Vietlex Corpus) chứa khoảng 80.000.000 âm tiết (tương đương gần 4 triệu câu), được thu thập từ các tác phẩm văn học và báo chí, các tác phẩm khoa học, các văn bản pháp luật, các bài viết chuyên ngành [5]. Kho ngữ liệu song ngữ EVC (5 triệu từ) do các nhà nghiên cứu của Trung tâm Ngôn ngữ học Tính toán của Trường ĐH Khoa học Tự nhiên – Tp.HCM xây dựng [6] chứa dữ liệu ở tất cả các lĩnh vực khoa học, xã hội, đời sống... Kho ngữ liệu song ngữ Bitext-PTB chứa 100.000 cặp câu song ngữ được xây dựng trong nội dung của nhánh đề tài "Xử lý văn bản tiếng Việt" thuộc 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" của nhóm tác giả Hồ Tú Bảo và Lương Chi Mai...

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. Các kho ngữ liệu đã được xây dựng trong lĩnh vực xử lý ngôn ngữ tiếng Việt hiện nay phần lớn thuộc các nhóm sau:

- Phục vụ nghiên cứu từ điển tiếng Việt: nội dung văn bản trong kho ngữ liệu đã được tách từ, phân tích từ loại phục vụ cho việc xây dựng từ điển song ngữ [6].

- Phục vụ nghiên cứu ngôn ngữ: gồm các kho ngữ liệu được chú giải ngữ pháp, gán nhãn ngữ pháp, phân cụm và phân tích câu tiếng Việt; xây dựng tập các quy tắc ngữ pháp tiếng Việt dùng cho xử lý tự động ngôn ngữ; nghiên cứu và xây dựng bộ phân tích cú pháp, câu tiếng Việt [2].

- Phục vụ các nghiên cứu xử lý ngôn ngữ tự nhiên, trong đó có dịch tự động tiếng Việt: kho ngữ liệu dùng cho các hệ thống dịch, cải tiến chất lượng hệ thống dịch... Tuy nhiên chưa có công trình nào nghiên cứu các giải pháp xây dựng kho ngữ liệu phục vụ cho việc đánh giá dịch tự động tiếng Việt.

2.3. Các giải pháp đánh giá chất lượng dịch

Đánh giá chất lượng dịch là hoạt động nhằm xác định mức độ hoàn thiện của bản dịch do máy tính đưa ra hoặc so sánh chất lượng dịch giữa các hệ thống dịch tự động khác nhau.

Hiện nay, có nhiều phương pháp và độ đo khác nhau để đánh giá chất lượng dịch tự động, có thể nhóm thành hai loại chính là đánh giá chủ quan (subjective evaluation) và đánh giá khách quan (objective evaluation).

Đánh giá chủ quan do con người trực tiếp thực hiện, dựa trên việc đánh giá thang điểm cho các tiêu chí được xây dựng sẵn: đánh giá tính chính xác, đầy đủ thông tin và đánh giá sự trôi chảy của câu dịch. Cách đánh giá chủ quan cho kết quả tin cậy nhưng tốn nhiều thời gian và chi phí, có phụ thuộc vào khả năng của người đánh giá [11].

Đánh giá khách quan là sử dụng các chương trình thay cho con người để đánh giá. Các chương trình sẽ so khớp hoặc đo tỉ lệ lỗi của kết quả từ hệ thống dịch với câu dịch tham khảo đã có sẵn. Một số phương pháp đánh giá phổ biến như [8][8]:

WER (Word Error Rate): đo số lượng các từ khác biệt giữa bản dịch của máy tính với bản dịch tham khảo. WER được tính toán như là khoảng cách Levenshtein giữa các từ của bản dịch với những từ của bản dịch tham khảo chia cho chiều dài của bản dịch tham khảo.

TER (Translation Edit Rate): đếm số bước sửa đổi để thay đổi bản dịch của máy tính thành một trong những bản dịch tham khảo có sẵn.

BLEU (BiLingual Evaluation Understudy): sử dụng bản dịch tự động so sánh với một bản dịch chuẩn của con người để tính điểm dựa trên việc thống kê sự trùng khớp của các từ trong hai bản dịch có tính đến thứ tự của chúng trong câu, sử dụng n-grams theo từ [15].

NIST (National Institute of Standards and Technology): dựa trên phương pháp BLEU nhưng có một số thay đổi là ngoài việc so sánh và tính toán trên số lượng thì còn tính đến sự thay đổi về vị trí các phần tử trên cùng một n-grams. Sự thay đổi này sẽ ảnh hưởng đến kết quả đánh giá dựa trên sự tương ứng về vị trí của các n-grams trên phân đoạn.

Như vậy, với các phương pháp đánh giá này, cần phải có một kho ngữ liệu song ngữ chuẩn để làm cơ sở đánh giá. Việc so khớp và đánh giá chất lượng bản dịch hoàn toàn phụ thuộc vào các bản dịch tham khảo có sẵn trong kho ngữ liệu này.

2.4. Hạn chế

Đánh giá chất lượng các hệ thống dịch tự động bằng các phương pháp và số đo nêu trên đã được nghiên cứu và áp dụng rộng rãi. Tuy nhiên trong một số trường hợp vẫn còn có những hạn chế khi tổ chức đánh giá.

- Trong cả phương pháp đánh giá chủ quan và khách quan, cần phải có kho ngữ liệu song ngữ mới có thể tổ chức đánh giá. Việc thu thập kho ngữ liệu trong các chủ đề thông dụng có thể thực hiện được, tuy nhiên đối với các lĩnh vực chuyên ngành hẹp, các dữ liệu song ngữ khó có thể thu thập được đầy đủ và tổng quan. Chẳng hạn trong lĩnh vực y học, các thông tin liên quan đến một số loại cây thuốc nam thường chỉ được viết dưới dạng tiếng Việt, hầu như không có tài liệu bằng tiếng Anh. Vì vậy, sử dụng các phương pháp so khớp không thể đánh giá đầy đủ chất lượng hệ thống dịch, vì kho ngữ liệu dùng để đánh giá không bao quát.

- Kết quả đánh giá không khách quan vì chỉ so sánh bản dịch của máy tính với tập dữ liệu giả định là những câu dịch tham chiếu có sẵn trong kho ngữ liệu. Trong thực tế, một câu ở ngôn ngữ này có thể có nhiều cách dịch khác nhau ở ngôn ngữ khác tùy theo ngữ cảnh. Ví dụ lấy một cặp câu đơn giản từ kho ngữ liệu song ngữ tiếng Anh – tiếng Việt [6] để đưa vào kho ngữ liệu đánh giá dịch:

Câu nguồn: *I've visited a few times before; it's a beautiful city.*

Câu tham chiếu (bản dịch): *Tôi đã đến thăm vài lần trước đây; đó là một thành phố xinh đẹp.*

Hệ thống Google Translate sẽ dịch câu nguồn thành: *Tôi đã đến thăm một vài lần trước; đó là một thành phố xinh đẹp.*

Đây là một kết quả dịch chính xác, kể cả nội dung và sự trôi chảy. Tuy nhiên khi đánh giá kết quả dịch [7][7] bằng các chỉ số trên thì chất lượng khá thấp: BLEU=0,7, TER=0,1, NIST=3,8.

- Đối với đánh giá chủ quan do con người trực tiếp thực hiện: tốn chi phí về thời gian và nhân lực để thực hiện quá trình đánh giá. Việc đánh giá bằng các tiêu chí ước lượng nên kết quả không cụ thể. Ngoài ra không tận dụng được kết quả phục vụ cho nâng cao chất lượng của hệ thống dịch.

- Nhiều hệ thống cho phép người dùng tham gia chỉnh sửa kết quả dịch để cải thiện chất lượng hệ thống dịch như hệ thống Google Translator, quá trình chỉnh sửa của con người chưa được đo lường cụ thể để đánh giá chi phí về thời gian và sức lao động của con người sử dụng khi hoàn thiện bản dịch.

3. Đề xuất giải pháp

Hiện nay các hệ thống dịch tự động chưa thể cho các kết quả dịch hoàn toàn chính xác để có thể sử dụng ngay, mà cần có quá trình can thiệp của con người thông qua quá trình hiệu đính để hoàn thiện bản dịch, bao gồm việc kiểm tra từ ngữ, ngữ pháp, chính tả, tên riêng, thuật ngữ chuyên ngành... mà hệ thống dịch chưa xử lý chính xác. Vì vậy việc đo chi phí về thời gian, công sức lao động của con người để hoàn thiện bản dịch là điều cần thiết và thể hiện được chất lượng và hiệu quả của hệ thống dịch. Một hệ

thống dịch tự động tốt thì con người bỏ ra ít thời gian và công sức để chỉnh sửa hơn. Tuy nhiên các phương pháp đánh giá nêu trên mới chỉ so sánh mức độ tương đương giữa bản dịch máy và bản dịch tham chiếu chứ chưa đo được chi phí thời gian và sức lao động. Bên cạnh đó, quá trình hiệu đính cũng giúp xây dựng kho ngữ liệu phục vụ đánh giá và nâng cao chất lượng hệ thống dịch.

3.1. Vấn đề xây dựng kho ngữ liệu phục vụ đánh giá

Để phục vụ cho các nghiên cứu trong lĩnh vực dịch tự động tiếng Việt, chúng ta thường sử dụng các kho ngữ liệu song ngữ, chẳng hạn kho ngữ liệu bao gồm các cặp câu tiếng Anh – tiếng Việt. Có nhiều giải pháp để xây dựng kho ngữ liệu song ngữ này:

- Trích rút tự động từ các trang web song ngữ: sử dụng các thuật toán tìm kiếm các trang web song ngữ, sau đó trích rút các cặp câu tương ứng.

- Lấy nguồn từ các sách song ngữ: sử dụng các tài liệu song ngữ, trong đó có các tài liệu học tiếng Anh để xây dựng kho ngữ liệu [2].

- Sử dụng từ điển và trích các ví dụ về các từ: trong các bộ từ điển, tương ứng với mỗi từ luôn có các ví dụ sử dụng và bản dịch chuẩn, có thể sử dụng các ví dụ này để xây dựng kho ngữ liệu song ngữ [1].

- Tổ chức dịch các câu tiếng Anh sang tiếng Việt: con người trực tiếp tạo bộ dữ liệu tiếng Việt hoặc thu thập các câu tiếng Việt từ nhiều nguồn, sau đó tổ chức dịch và nhập vào dữ liệu để có kho ngữ liệu song ngữ.

Quá trình hiệu đính bản dịch máy cũng tạo ra các văn bản song ngữ có giá trị, tuy nhiên vẫn chưa được đề cập như là một trong những phương pháp để xây dựng kho ngữ liệu song ngữ.

Như đã trình bày ở trên, vấn đề tìm kiếm các nguồn dữ liệu để xây dựng kho ngữ liệu ở các chủ đề phổ biến khá dễ dàng. Tuy nhiên, ở nhiều lĩnh vực chuyên ngành, hầu như có rất ít tài liệu song ngữ.

Để giải quyết vấn đề này, kho ngữ liệu phục vụ đánh giá dịch tự động chỉ cần chứa các câu ở ngôn ngữ nguồn. Từ các câu nguồn này, sử dụng các hệ thống dịch tự động để dịch và lấy kết quả dịch làm câu đích, tạo thành một cặp câu song ngữ. Như vậy, việc xây dựng dữ liệu đánh giá chỉ cần thu thập các văn bản có sẵn ở một ngôn ngữ mà không cần dịch sang các ngôn ngữ khác. Điều này khá dễ dàng và không tốn nhiều chi phí cho kho ngữ liệu phục vụ đánh giá.

Tuy nhiên cặp câu song ngữ này có thể chưa phải là một bản dịch chính xác. Thông qua quá trình hiệu đính với sự trợ giúp của con người, hệ thống đánh giá sẽ đo các chỉ số và phân tích mức độ chính xác của bản dịch đã có. Bên cạnh đó, khi tích hợp chức năng đánh giá này vào các hệ thống dịch tự động, quá trình hiệu đính sẽ giúp đánh giá được chất lượng dịch mà không phải tiến hành các giải pháp bổ sung nào khác.

Sau quá trình hiệu đính và đánh giá chất lượng dịch, chúng ta thu được bản dịch chính xác, từ đó xây dựng được kho ngữ liệu song ngữ hoàn chỉnh. Kho ngữ liệu này có thể được sử dụng trong những phương pháp đánh giá khác và đồng thời có thể sử dụng để cải thiện chất lượng của hệ

thống dịch. Kết quả đánh giá dựa trên kho ngữ liệu này sẽ giúp quá trình đánh giá được chính xác, không phụ thuộc vào các bản dịch tham chiếu có sẵn mà đánh giá trực tiếp trên bản dịch do máy tính đưa ra.

3.2. Giải pháp đánh giá chất lượng dịch

Từ các phân tích trên, nhóm tác giả đề xuất sử dụng thêm một số chỉ số để đo chi phí của con người khi tham gia chỉnh sửa từ bản dịch của máy tính sang bản dịch hoàn chỉnh.

Chỉ số thời gian: đo thời gian cần thiết để con người chỉnh sửa kết quả dịch từ hệ thống dịch tự động thành bản dịch đúng. Khi có kết quả dịch từ hệ thống dịch tự động, con người cần tiếp tục kiểm tra, rà soát và sửa lỗi để cho ra bản dịch chính xác. Thời gian sửa lỗi càng ít thì hệ thống dịch càng chính xác. Hệ thống đánh giá chất lượng dịch tự động sẽ có nhiệm vụ đo và tính toán chỉ số thời gian này.

$$T_{pe} = T/N$$

T: Thời gian sửa bản dịch được tính từ khi người dùng chọn chức năng “Sửa bản dịch” cho đến khi người dùng xác nhận “Gửi kết quả”.

N: Số lượng ký tự ở bản dịch chính xác sau khi người dùng đã chỉnh sửa xong, không tính đến các ký tự trống (khoảng trắng, tab, ký tự xuống dòng) và các dấu câu.

Chỉ số thao tác: đo số lượng ký tự cần thay đổi để sửa kết quả dịch thành bản dịch đúng.

$$O_{pe} = (D + I) / N$$

Trong đó:

(D + I) là số lượng ký tự mà người dùng đã thay đổi, bao gồm thao tác xóa, thêm, sửa để hiệu chỉnh bản dịch.

D: số lượng ký tự bị người dùng xóa

I: số lượng ký tự được người dùng thêm mới

Mỗi thao tác ghi đề được tính bằng một thao tác xóa và một thao tác thêm mới.

N: Số lượng ký tự ở bản dịch chính xác sau khi người dùng đã chỉnh sửa xong, không tính đến các ký tự trống (khoảng trắng, tab, ký tự xuống dòng) và các dấu câu.

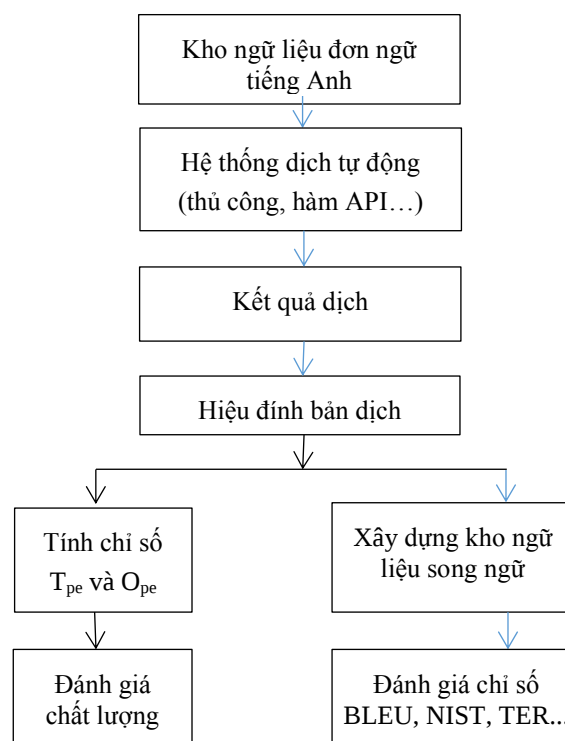
Như vậy, Chỉ số thời gian T_{pe} là thời gian trung bình để chỉnh sửa một ký tự tính trên bản dịch kết quả. Chỉ số thao tác O_{pe} là số lượng ký tự trung bình cần chỉnh sửa tính trên mỗi ký tự ở bản dịch đúng. Các chỉ số này càng thấp thì hệ thống dịch càng chất lượng. $T_{pe} = 0$ và $O_{pe} = 0$ khi bản dịch do máy tính đưa ra hoàn toàn chính xác, không cần sửa chữa (do người dùng xác nhận).

Từ các chỉ số T_{pe} và O_{pe} , chúng ta có thể đánh giá được mức độ chính xác của bản dịch do các hệ thống dịch tự động đưa ra. Kết quả đánh giá này hoàn toàn dựa trên bản dịch gốc của máy tính thông qua sự chỉnh sửa của con người, nên không cần các bản dịch tham khảo mà vẫn cho ra kết quả đánh giá chính xác và phù hợp nhất. Ngoài ra có thể so sánh chất lượng của các hệ thống dịch tự động (chẳng hạn giữa Google và Microsoft) một cách chính xác và khách quan mà không phụ thuộc vào bản dịch chuẩn của dữ liệu mẫu.

3.3. Giải pháp kết hợp hiệu đính bản dịch máy và đánh giá chất lượng

Hiện nay, không phải hệ thống nào cũng hỗ trợ thao tác hiệu đính để giúp người dùng chỉnh sửa kết quả bản dịch máy thành bản dịch chính xác và có thể sử dụng được. Ở các hệ thống này, người dùng chỉ có thể xem kết quả bản dịch, sau đó có thể sao chép sang các trình soạn thảo văn bản khác để tiếp tục chỉnh sửa hoàn thiện bản dịch này. Như vậy các hoạt động hiệu đính chỉ phục vụ cho mục đích của người sử dụng, chứ chưa được ứng dụng để cải thiện và đánh giá hệ thống dịch.

Các hoạt động hiệu đính và đánh giá chất lượng được thực hiện tách rời nhau và không kế thừa kết quả của nhau. Vì vậy, nhóm tác giả đề xuất kết hợp các hoạt động hiệu đính bản dịch máy với đánh giá chất lượng, giúp giảm chi phí và nâng cao độ chính xác. Để làm được điều này, cần xây dựng hệ thống đánh giá chất lượng dịch tiếng Anh – tiếng Việt của các hệ thống dịch tự động trực tuyến (như Google, Microsoft) thông qua các thao tác hậu xử lý với các đặc điểm sau:



Hình 2. Sơ đồ quy trình kết hợp hậu xử lý với đánh giá chất lượng hệ thống dịch tự động

- Kho ngữ liệu nguồn: là kho ngữ liệu đơn ngữ (tiếng Anh), chưa được dịch sang tiếng Việt.

- Dịch thô: Từ câu nguồn tiếng Anh (cần dịch), sử dụng các hệ thống dịch tự động trực tuyến để lấy kết quả bản dịch (thực hiện thủ công, sử dụng các hàm API hoặc các kỹ thuật khác) là câu đích ở tiếng Việt. Đây là kết quả “dịch thô”, chưa chính xác và cần quá trình sửa chữa của con người.

- Hiệu đính bản dịch: Cho phép người dùng kiểm tra và chỉnh sửa bản dịch tiếng Việt này để có được bản dịch chính xác mà không cần dựa trên các bản dịch có sẵn.

- Đánh giá chất lượng: theo dõi và đo các chỉ số về thời gian (T_{pe}) và thao tác (O_{pe}) để đánh giá chất lượng của các bản dịch.

- Xây dựng kho ngữ liệu: sau quá trình hiệu đính, hệ thống thu được các cặp câu nguồn tiếng Anh và bản dịch tiếng Việt đã được chỉnh sửa. Các cặp câu song ngữ này được sử dụng để tạo kho ngữ liệu tương ứng.

- Đánh giá chất lượng bằng các phương pháp khác: khi đã có kho ngữ liệu song ngữ, chúng ta có thể tiếp tục hoạt động đánh giá bằng các số đo khác như đã trình bày ở trên, từ đó có thể đối chiếu, so sánh giữa các thang đo.

- Cải thiện chất lượng hệ thống dịch: bản dịch sau khi được con người sửa chữa sẽ được sử dụng để cải thiện chất lượng của hệ thống dịch. Ở các lần dịch tiếp theo, kết quả của hệ thống dịch chắc chắn sẽ tốt hơn.

3.4. Thực nghiệm

Để triển khai thực nghiệm quá trình hiệu đính bản dịch máy kết hợp đánh giá chất lượng bản dịch và xây dựng kho ngữ liệu, nhóm tác giả xây dựng ngữ cảnh như sau:

- Lấy 500 câu song ngữ tiếng Anh – tiếng Việt từ các đoạn hội thoại đã được dịch trong tài liệu học tiếng Anh ở chủ đề thông dụng là “Cuộc sống hàng ngày”. Dữ liệu này được chia thành 5 bộ để thực nghiệm. Câu tiếng Anh (E) là câu nguồn cần dịch, câu tiếng Việt được xem là bản dịch tham chiếu (V).

- Xây dựng chương trình sử dụng các dịch vụ được cung cấp của hệ thống dịch Google, Microsoft để tự động gửi các câu nguồn tiếng Anh (E) và lấy về các bản dịch tiếng Việt (V1) từ các hệ thống dịch đó. Dùng bản dịch của kho ngữ liệu (V) để làm câu tham chiếu, đo các chỉ số đánh giá (D1) theo thuật toán Edit Distance (ED), Word Error Rate (WER), BLEU, NIST để so sánh sự tương đồng giữa hai câu V và V1. Chỉ số Edit Distance cho biết số lượng từ và ký tự khác nhau giữa 2 câu, Word Error Rate đo tỷ lệ lỗi giữa câu dịch và câu tham chiếu, còn chỉ số BLEU và NIST tính toán mức độ khác nhau dựa trên n-grams. Kết quả thu được ở Bảng 1.

Bảng 1. Trung bình các chỉ số trên 5 bộ dữ liệu

	BLEU	NIST	ED	WER
Test 1	0,242	1,563	6,739	0,886
Test 2	0,542	2,876	5,684	0,561
Test 3	0,292	1,972	7,000	0,814
Test 4	0,273	2,450	9,286	0,757
Test 5	0,380	2,546	8,067	0,634

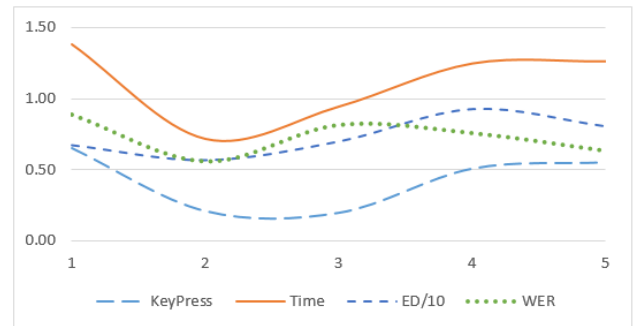
- Xây dựng hệ thống cho phép người dùng hiệu chỉnh các bản dịch. Quá trình này được tự động đo thời gian và đếm số lượng thao tác mà người dùng thực hiện để hiệu chỉnh. Từ đó tính toán chỉ số T_{pe} và O_{pe} để đánh giá chất lượng bản dịch. Bản dịch hiệu chỉnh (V2) được đưa vào kho ngữ liệu để làm câu tham chiếu.

- Tiếp tục sử dụng hệ thống để tính toán lại các chỉ số Edit Distance, Word Error Rate, BLEU, NIST (D2) giữa hai câu V1 và V2. So sánh D1 và D2. Kết quả được thể hiện ở Bảng 2.

Bảng 2. Kết quả sau khi hiệu chỉnh bản dịch

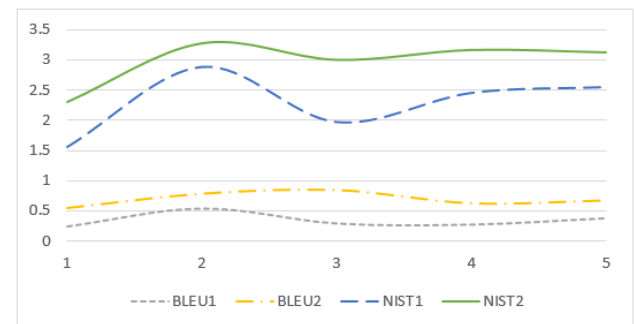
	T_{pe}	O_{pe}	BLEU	NIST	ED	WER
Test 1	0,649	1,385	0,553	2,304	3,913	0,403
Test 2	0,211	0,717	0,789	3,281	1,632	0,142
Test 3	0,199	0,945	0,849	3,006	1,071	0,102
Test 4	0,506	1,250	0,631	3,170	3,500	0,301
Test 5	0,550	1,265	0,677	3,130	4,133	0,257

Từ kết quả này, chúng ta có thể thấy được sự tương đồng giữa chỉ số T_{pe} , O_{pe} với các chỉ số Edit Distance và Word Error Rate thông qua đồ thị ở Hình 3.



Hình 3. Sự tương đồng giữa T_{pe} , O_{pe} và ED, WER

Khi so sánh kết quả bản dịch với câu tham chiếu thu được từ quá trình hiệu chỉnh, các chỉ số BLEU và NIST đã tốt hơn (Hình 4) và việc đánh giá không phụ thuộc vào các bản dịch có sẵn, thể hiện được tính khách quan khi đánh giá kết quả của hệ thống dịch.



Hình 4. Chỉ số BLEU, NIST trước và sau khi hiệu chỉnh

4. Kết luận

Chất lượng của kho ngữ liệu trong hệ thống dịch máy có ảnh hưởng đến chất lượng của kết quả dịch. Việc kết hợp quá trình hiệu đính bản dịch máy với đánh giá chất lượng của hệ thống dịch tự động và xây dựng kho ngữ liệu nhằm cải tiến chất lượng dịch tự động là một giải pháp khả thi và tiết kiệm tài nguyên, công sức.

Đánh giá các hệ thống dịch thông qua các chỉ số về thời gian và số lượng ký tự mà người dùng cần chỉnh sửa sẽ phản ánh đúng chất lượng của bản dịch so với yêu cầu của người dùng, đồng thời mang tính khách quan vì điểm số này dựa trên kết quả dịch chứ không phụ thuộc các đáp án có sẵn.

Bài báo đã tiến hành thực nghiệm và thu được kết quả trên tập dữ liệu nhỏ. Việc đo chỉ số thời gian cần có các nghiên cứu chi tiết hơn để kiểm soát quá trình hiệu đính bản dịch.

TÀI LIỆU THAM KHẢO

- [1] Đào Ngọc Tú, “Nghiên cứu về dịch thống kê dựa vào cụm từ và thử nghiệm với cặp ngôn ngữ Anh – Việt”, *Luận văn thạc sỹ Khoa học máy tính*, Học viện Công nghệ Bưu chính viễn thông, 2012.
- [2] Đinh Điền, Lý Ngọc Minh, “Ứng dụng Ngữ liệu Song ngữ Anh-Việt trong Giảng dạy Ngôn ngữ”, hội thảo Liên ngành NNH Ứng dụng & Giảng dạy Ngôn ngữ, 11/2015, Huế, tr.559-567.
- [3] Nguyễn Hữu Siêu, Lâm Tùng Giang, Võ Trung Hùng (2010), “Nghiên cứu xây dựng từ điển cho hệ thống dịch tự động UNL - Tiếng Việt”, *Tạp chí KH&CN Đại học Đà Nẵng* số 4(39) 2010.
- [4] Huỳnh Công Pháp, Đặng Đại Thọ, Nguyễn Văn Bình (2015), “Cải tiến chất lượng dịch tự động bằng giải pháp mở rộng kho ngữ liệu”, *Kỷ yếu Hội thảo Khoa học Quốc gia: Nghiên cứu cơ bản và ứng dụng công nghệ thông tin*, FAIR 2015, Số: ISBN: 978-604-913-397-8, Trang: 315–320.
- [5] Trung tâm từ điển học Vietlex, *Giới thiệu Kho ngữ liệu tiếng Việt (Vietnamese Corpus)*, truy cập ngày 27/09/2016, http://www.vietlex.com/help/about_corpus.htm
- [6] Trung tâm Ngôn ngữ học Tính toán của Trường ĐH Khoa học Tự nhiên – Tp.HCM, *Kho ngữ liệu song ngữ tiếng Anh – tiếng Việt EVC*, truy cập ngày 27/09/2016, http://www.clc.hcmus.edu.vn/wp-content/uploads/resources/Corpus/CLC_EVC.zip
- [7] Đánh giá trực tuyến các kết quả dịch tự động, http://asiya.cs.upc.edu/demo/asiya_online.php
- [8] Arne Mauser, Hermann Ney (2008), “Automatic evaluation measures for statistical machine translation system optimization”, International Conference on Language Resources and Evaluation.
- [9] Boitet C. (2007), *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, pp. 25-38.
- [10] Green, Spence, Jeffrey Heer, and Christopher D. Manning (2013), *The Efficacy of Human Post-Editing for Language Translation*, ACM Human Factors in Computing Systems.
- [11] Huynh C-P. (2010), “Des suites de test pour la TA à un système d’exploitation de corpus alignés de documents et métadocuments multilingues, multiannotés et multimedia”, *PhD thesis-National Polytechnic Institute of Grenoble*, 228 p.
- [12] Läubli, Samuel, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk (2013), *Assessing post-editing efficiency in a realistic translation environment*, Proceedings of the 2nd Workshop on Post-editing Technology and Practice (WPTP), p. 83–91.
- [13] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul, *A Study of Translation Edit Rate with Targeted Human Annotation*.
- [14] Plitt, Mirko and Francois Masselot (2010), “A Productivity Test of Statistical Machine Translation Post-Editing in A Typical Localisation Context”, *Prague Bulletin of Mathematical Linguistics*, 93:7–16.
- [15] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002), “BLEU: a method for automatic evaluation of machine translation” in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311–318.

(BBT nhận bài: 01/03/2017, hoàn tất thủ tục phản biện: 25/04/2017)