TỔNG QUAN VỀ MỘT SỐ PHƯƠNG PHÁP DỊCH MÁY CHO CẶP NGÔN NGỮ NGHÈO TÀI NGUYÊN

Phạm Nghĩa Luân
Trung tâm Thông tin Thư viện
Email: luanpn@dhhp.edu.vn

Ngày nhận bài: 26/4/2024
Ngày PB đánh giá: 07/5/2024
Ngày duyệt đăng: 31/5/2024

Tóm tắt: Dữ liệu song ngữ là rất quan trọng, không thể thiếu khi xây dựng một mô hình dịch máy. Tuy nhiên, khó khăn lớn nhất là lượng dữ liệu này thường rất ít, đặc biệt là đối với nhiều cặp ngôn ngữ ít phổ biến trên thế giới. Việc thu thập và xử lý dữ liệu song ngữ không chỉ tốn kém về mặt tài chính mà còn đòi hỏi sự đầu tư lớn về thời gian và nhân lực. Do đó, việc tạo ra một bộ dữ liệu đủ lớn và chất lượng để huấn luyện các mô hình dịch máy là một thách thức đáng kể. Để giải quyết vấn đề thiếu dữ liệu cho dịch máy, nhiều phương pháp đã được nghiên cứu và áp dụng như sử dụng dữ liệu đơn ngữ, học chuyển đổi và sử dụng ngôn ngữ trục. Mỗi phương pháp đều có ưu điểm và hạn chế riêng. Việc lựa chọn phương pháp phù hợp phụ thuộc vào đặc điểm của cặp ngôn ngữ cần dịch và mục tiêu sử dụng. Bài báo này sẽ giới thiệu tổng quan về một cách tiếp cận chính cho hướng nghiên cứu này.

Từ khóa: Dịch máy; xây dựng ngữ liệu song ngữ; ngôn ngữ nghèo tài nguyên.

OVERVIEW OF SOME MACHINE TRANSLATION METHODS FOR LOW-RESOURCE LANGUAGE PAIRS

Abstract: Parallel corpus is crucial to build a machine translation model. However, the biggest challenge lies in the scarcity of the corpus, especially for the one in less common language pairs. Collecting and processing such data requires a significant investment of not only finance but time and human resources. Therefore, creating a sufficiently large and high-quality data set to train machine translation models is challenging. Various methods have been researched and applied to address this challenge, including monolingual data methods, transfer learning, and pivot languages. Each method has its advantages and disadvantages. Choosing an appropriate solution depends on the language pair and the intended use. This paper introduces an overview of the approach in this research direction.

Keywords: Machine translation; buiding parallel corpus; low-resource language

1. ĐẶT VẤN ĐỀ

Lượng dữ liệu song ngữ để huấn luyện mô hình dịch máy theo phương pháp thống kê sẽ ảnh hưởng trực tiếp tới chất lượng của mô hình. Thông thường, chất lượng của mô hình ở mức chấp nhận được thì cần hàng triệu cặp câu song ngữ. Tuy nhiên, loại dữ liệu này thường có nhiều đối với các cặp ngôn ngữ phổ biến (ví dụ tiếng Anh - tiếng Pháp, tiếng Anh - tiếng Tây Ban Nha) và có ít đối với các cặp ít phổ biến (ví dụ tiếng Việt - tiếng Campuchia, tiếng Việt - tiếng Lào), lượng dữ liệu có sẵn của các cặp ngôn ngữ ít phổ biến này thường chỉ vài nghìn hoặc vài chục nghìn cặp câu.

Mô hình dịch máy thống kê do Keohn và cộng sự giới thiệu năm 2007 [5] đã đạt được những thành công đáng kể trong việc dịch các cặp ngôn ngữ nghèo tài nguyên chỉ với một lượng dữ liệu nhỏ. Tuy nhiên, các bản dịch từ mô hình này thường không trôi chảy và không phản ánh chính xác ngữ cảnh của từ vựng. Điều này có thể gây khó khăn trong việc hiểu rõ ý nghĩa của văn bản và giao tiếp hiệu quả giữa các ngôn ngữ. Do đó, việc nâng cao chất lượng dịch từ mô hình dịch thống kê vẫn còn là một thách thức đối với cộng đồng nghiên cứu và người sử dụng.

Năm 2014, việc phát triển mô hình dịch máy sử dụng mạng nơron (NMT - Neural Machine Translation) được tiếp tục đẩy mạnh [1, 3, 12]. Điều này đã mở ra một bước tiến mới đáng kể cho sự phát triển chất lượng các mô hình dịch máy. Ban đầu, các mô hình dịch máy sử dụng kiến trúc mạng nơron hồi quy (RNN - Recurrent Neural Network) hoặc mạng nơron tích chập (CNN - Convolutional Neural Network) để biểu diễn mối quan hệ giữa các từ trong câu và giữa các từ trong ngôn ngữ nguồn và ngôn ngữ đích. Tuy nhiên, các kiến trúc RNN và CNN có tốc độ huấn luyện chậm, chưa đáp ứng được yêu cầu về song song hóa của các ứng dụng cũng như bảo toàn thông tin trong các ngữ cảnh dài. Năm 2017, Vaswani và cộng sự đã đề xuất kiến trúc Transformer cho dịch máy nhằm khắc phục các nhược điểm trên [13]. Kiến trúc này đã mang lại những cải tiến đáng kể, với khả năng song song hóa tốt hơn và bảo toàn thông tin trong các ngữ cảnh dài, điều này đã tạo ra sự thay đổi lớn trong dịch máy, mở ra tiềm năng lớn cho ứng dụng trong thực tế.

Kiến trúc dịch dựa trên mạng nơron mang lại nhiều lợi thế cho các cặp ngôn ngữ giàu tài nguyên. Tuy nhiên, đối với các cặp ngôn ngữ nghèo tài nguyên, đây lại là một thách thức lớn do sự khan hiếm dữ liệu song ngữ. Tuy vậy, nhiều nghiên cứu đã chỉ ra rằng các mô hình dịch dựa trên mạng nơron vẫn cho kết quả tốt hơn so với các mô hình sử dụng mô hình dịch thống kê với các tập dữ liệu nhỏ [7, 10]. Để đáp ứng sự phát triển của các mô hình dịch máy dựa trên mạng nơron, nhiều nghiên cứu đã đề xuất nhiều giải pháp khác nhau để nâng cao chất lượng dịch cho các cặp ngôn ngữ nghèo tài nguyên. Trong bối cảnh ngày nay, việc phát triển các phương pháp dịch máy cho các ngôn ngữ nghèo tài nguyên đang trở nên ngày càng quan trọng. Những nỗ lực này giúp mở rộng ứng dụng của mô hình dịch máy sử dụng mạng nơron và nâng cao khả năng dịch của chúng đối với các cặp ngôn ngữ nghèo tài nguyên, góp phần quan trọng trong việc nhanh chóng cải thiện chất lượng của dịch vụ dịch thuật tự động, từ đó giúp tạo ra sự thuận lợi và hiệu quả trong giao tiếp và trao đổi văn hóa giữa các quốc gia.

Bài báo này trình bày, cung cấp một cách tổng quan về một số phương pháp dịch máy phổ biến, đơn giản để thử nghiệm nhưng hiệu quả đối với các cặp ngôn ngữ nghèo tài nguyên.

2. TỔNG QUAN VẤN ĐỀ NGHIÊN CỨU

Một trong những vấn đề lớn nhất mà các mô hình dịch máy gặp phải là việc thiếu

hụt dữ liệu song ngữ. Dữ liệu này thông thường được tạo thủ công hoặc thu thập tự động từ internet nhưng vẫn không thể đủ để huấn luyện các mô hình dịch cho kết quả tốt, đặc biệt với những cặp ngôn ngữ ít phổ biến như Việt - Lào, Việt - Campuchia.

Hơn nữa, các mô hình dịch máy thường cho kết quả tốt trên miền dữ liệu được huấn luyện, nếu đem dịch dữ liệu thuộc miền khác, kết quả sẽ rất kém. Ví dụ, dữ liệu song ngữ được thu thập từ các trang báo song ngữ, chủ đề thường về thể thao, giải trí, thì mô hình dịch được huấn luyện trên dữ liệu này sẽ cho kết quả không tốt khi dịch các câu thuộc lĩnh vực thương mại, du lịch. Bản chất đây cũng là vấn đề thiếu tài nguyên song ngữ cho dịch máy của bất kỳ cặp ngôn ngữ nào. Vì vậy, dịch máy cho các cặp ngôn ngữ nghèo tài nguyên song ngữ là rất quan trọng, được cộng đồng quan tâm nghiên cứu. Để giải quyết vấn đề này, rất nhiều ý tưởng đã được đề xuất, điển hình như sau:

(1)Sử dụng thêm dữ liệu đơn ngữ [6], ngược với việc thiếu hụt dữ liệu song ngữ thì dữ liệu đơn ngữ lại có sẵn với số lượng lớn cho hầu hết các ngôn ngữ. Một số phương pháp tận dụng tri thức từ dữ liệu đơn ngữ để làm tăng chất lượng dịch như tích hợp thêm mô hình ngôn ngữ được học từ dữ liệu đơn ngữ để câu được sinh ra hợp lý hơn, hay kĩ thuật back-translation sử dụng dữ liệu đơn ngữ để làm giàu thêm dữ liệu song ngữ.

(2)Kĩ thuật học chuyển [16], kĩ thuật này khá hiệu quả để nâng cao chất lượng dịch cho những cặp ngôn ngữ có tài nguyên hạn chế *(kích thước dữ liệu song ngữ khoảng vài chục hoặc vài trăm nghìn câu)*. Ý tưởng chính của kĩ thuật này là tận dụng tri thức đã học được từ những cặp ngôn ngữ có tài nguyên lớn như Anh - Pháp *(vài chục triệu cặp câu)*, sử dụng các tham số đã học được đó làm giá trị khởi tạo cho mô hình được huấn luyện với cặp ngôn ngữ có dữ liệu nhỏ.

(3)Sử dụng ngôn ngữ trục (ngôn ngữ trung gian), phương pháp này dựa trên ý tưởng các cặp ngôn ngữ giàu tài nguyên có thể được sử dụng làm cầu nối trung gian trong việc dịch qua lại giữa các ngôn ngữ nghèo tài nguyên.

(4)Sử dụng từ điển song ngữ [15], mô hình dịch có thể được tích hợp thêm tri thức trong từ điển song ngữ để nâng cao chất lượng dịch. Trong mô hình dịch thống kê SMT, các cặp từ song ngữ trong từ điển sẽ được thêm trực tiếp vào bảng cụm từ *(*phrase table) để làm giàu thêm bảng cụm từ. Trong mô hình dịch nơron, mô hình sử dụng thêm một bộ nhớ để chứa các từ trong từ điển nhằm hướng dẫn mô hình đưa ra dự đoán chính xác hơn. Trong thực tế, phương pháp này không đem lại nhiều hiệu quả như các phương pháp trên.

Mặc dù nhiều phương pháp được đưa ra để khắc phục việc thiếu hụt dữ liệu song ngữ và nhiều thực nghiệm đã cho thấy tính hiệu quả. Tuy nhiên, không có phương pháp nào tỏ ra ưu việt hơn cả, nó phụ thuộc vào đặc điểm của các cặp ngôn ngữ khác nhau. Hiện nay trên thế giới, có nhiều nghiên cứu cho tập trung cho các cặp ngôn ngữ với tài nguyên hạn chế, điển hình là nhóm dịch máy của Facebook [(1)].

3. PHƯƠNG PHÁP DỊCH MÁY CHO NGÔN NGỮ NGHÈO TÀI NGUYÊN

3.1 Phương pháp sử dụng dữ liệu đơn ngữ

Trên thế giới, có nhiều cặp ngôn ngữ mà nguồn dữ liệu song ngữ rất khan hiếm, thậm chí không tồn tại, điều này tạo ra những thách thức lớn đối với việc phát triển và cải tiến mô hình dịch máy cho những cặp ngôn ngữ này. Tuy nhiên, nguồn dữ liệu đơn ngữ lại rất phong phú và có thể dễ dàng thu thập từ

(1) https://ai.facebook.com/blog/announcing-new-research-awards-in-nlp-and-machine-translation/

các trang web và nguồn thông tin khác. Do đó, nhiều nghiên cứu đã đề xuất các phương pháp khác nhau nhằm tận dụng nguồn dữ liệu đơn ngữ này để nâng cao chất lượng dịch máy cho những cặp ngôn ngữ nghèo tài nguyên.

3.1.1 Kĩ thuật dịch ngược (Back Translation)

Năm 2016, nhóm tác giả Sennrich và cộng sự đã giới thiệu thuật dịch ngược (back-translation) [11], phương pháp này sử dụng dữ liệu đơn ngữ trong ngôn ngữ đích để tạo ra dữ liệu song ngữ, sau đó sử dụng để huấn luyện mô hình dịch máy mà không cần thay đổi kiến trúc mạng. Phương pháp này có thể được áp dụng một cách tổng quát và hiệu quả trong việc cải thiện chất lượng dịch máy. Bằng cách tận dụng nguồn dữ liệu đơn ngữ có sẵn trong ngôn ngữ đích, kĩ thuật dịch ngược giúp tối ưu hóa quá trình huấn luyện, giúp tiết kiệm thời gian và công sức, đồng thời nâng cao hiệu suất của mô hình dịch máy. Kĩ thuật dịch ngược có thể phát biểu tổng quát như sau:

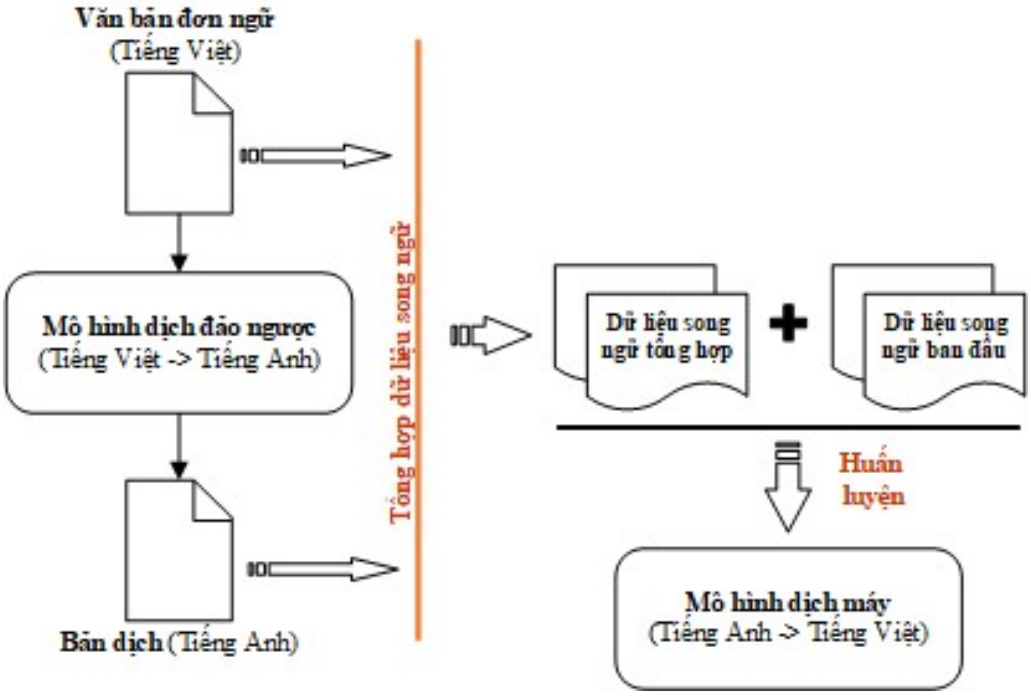
Cho một tập dữ liệu song ngữ đã được gióng hàng câu $D = \{(X_n, Y_n)\}N$ và một tập dữ liệu đơn ngữ trong ngôn ngữ đích $T = (Y_m)$ M, các bước lần lượt được thực hiện như sau:

1. Đầu tiên, một mô hình dịch ngược $NMT_{Y \to X}$ được huấn luyện với tập dữ liệu song ngữ D.

2. Thứ hai, với mô hình dịch $NMT_{Y \to X}$, tập dữ liệu đơn ngữ trong ngôn ngữ đích T được dịch ngược lại thành các bản dịch trong ngôn ngữ nguồn $S = (X_m)\ M_{m=1}$, sau đó tập dữ liệu S được ghép nối với T, tạo thành một tập dữ liệu giả song ngữ $D_{syn} = \{(X_m, Y_m)\}\ M_{m=1}$.

3. Thứ ba, tập dữ liệu giả song D_{syn} và tập dữ liệu song ngữ ban đầu D được kết hợp để huấn luyện một mô hình dịch máy $NMT_{Y \to X}$.

Các bước dịch ngược được mô tả như Hình 1 để xây dựng mô hình dịch máy Anh-Việt:



Hình 1. Minh họa kĩ thuật dịch ngược

Sennrich đã báo cáo những kết quả thực nghiệm thực sự ấn tượng khi áp dụng kĩ thuật này trên cặp ngôn ngữ tiếng Anh - tiếng Đức. Các mô hình dịch được đánh giá trên hai bộ dữ liệu chuẩn do cộng đồng cung cấp là test2014 và test2015. Kết quả thử nghiệm cho thấy điểm BLEU tăng từ +2.8 (test2014) đến +2.9 (test2015), cho thấy việc sử dụng dịch ngược có thể cải thiện chất lượng dịch giữa hai ngôn ngữ này. Chi tiết kết quả thử nghiệm như Bảng 1.

Bảng 1. Kết quả BLEU khi sử dụng dịch ngược cho cặp tiếng Anh - tiếng Đức.

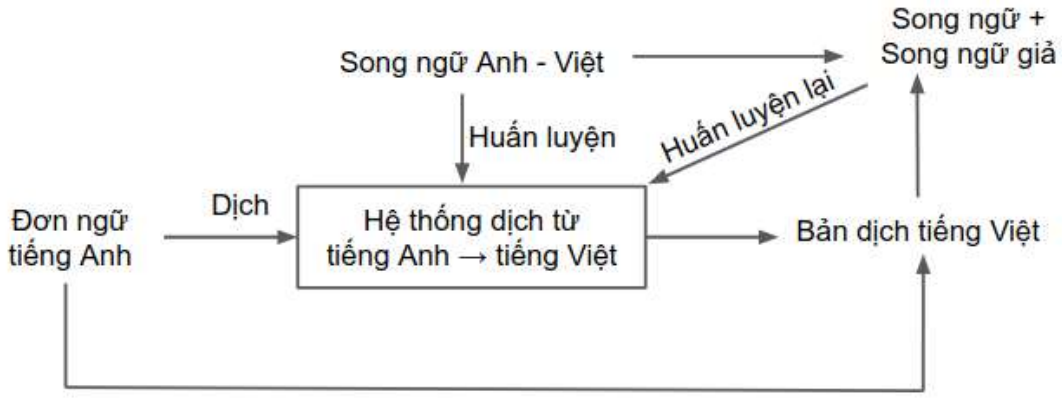
Mô hình	Dữ liệu huấn luyện	Điểm BLEU	
		test2014	test2015
Baseline	37 triệu cặp câu chuẩn	19.9	22.8
+ Dữ liệu song ngữ tổng hợp	37 triệu cặp câu chuẩn + 36 triệu cặp tổng hợp	**22.7**	**25.7**

Trong thực nghiệm, các tác giả sử dụng dữ liệu giả song ngữ có kích thước nhỏ hơn hoặc tương đương với dữ liệu song ngữ thật. Thông thường, nếu mô hình dịch tốt thì lượng dữ liệu đơn ngữ sử dụng có thể lớn nhưng tốt nhất cũng chỉ gấp đôi hoặc bằng lượng dữ liệu song ngữ chuẩn để đảm bảo mô hình có thể học tốt nhất.

3.1.2 Kĩ thuật tự học (Self Learning)

Năm 2016, nhóm tác giả Zhang và cộng sự đã giới thiệu kĩ thuật tự học (self-learning) [14], kĩ thuật này tương tự như kĩ thuật dịch ngược nhưng sử dụng dữ liệu đơn ngữ phía ngôn ngữ nguồn để sinh ra dữ liệu song ngữ tổng hợp từ chính mô hình dịch từ ngôn ngữ nguồn sang ngôn ngữ đích.

Cụ thể, với kho ngữ liệu ban đầu chứa N cặp song ngữ (X_1, Y_1) cho cặp ngôn ngữ bất kỳ, dữ liệu đơn ngữ được sử dụng gồm M câu đơn ngữ X_2. Từ tập N cặp câu song ngữ, ta huấn luyện mô hình dịch máy từ $X_1 \rightarrow Y_1$, mô hình được sử dụng để dịch M câu đơn ngữ X_2 sang Y_2 để tạo thành dữ liệu song ngữ tổng hợp (X_2, Y_2). Dữ liệu song ngữ tổng hợp này được kết hợp với dữ liệu song ngữ chuẩn ban đầu để tạo thành một kho ngữ liệu lớn hơn cho việc huấn luyện một mô hình dịch máy mới. Kĩ thuật này được minh họa như Hình 2.



Hình 2. Minh họa kĩ thuật self-learning

Nhóm tác giả thực nghiệm trên mô hình dịch tiếng Trung - tiếng Anh với lượng dữ liệu huấn luyện khác nhau. Chi tiết về dữ liệu thực nghiệm (đơn ngữ và song ngữ) được sử dụng trong các mô hình như Bảng 2.

Bảng 2. Dữ liệu thực nghiệm phương pháp tự học cho dịch máy trong nghiên cứu của Zhang và cộng sự (2016)

Dữ liệu	Số lượng	Nguồn
Tập song ngữ nhỏ	600.000 cặp câu	LDC
Tập song ngữ lớn	2.100.000 cặp câu	LDC
Dữ liệu đơn ngữ	6.5 triệu cặp cho tập dữ liệu song ngữ nhỏ và 12 triệu cặp cho tập song ngữ lớn (Gấp 10 lần dữ liệu huấn luyện)	LDC
Tập đánh giá	MT04, MT05, MT06	NIST
Tập phát triển	MT03	NIST

Các tác giả đã đánh giá ảnh hưởng của lượng dữ liệu đơn ngữ được sử dụng trong các thực nghiệm với các mức khác nhau, theo tỷ lệ 25%, 50%, 75%, 100% của tập song ngữ nhỏ (600.000 cặp câu). Nhóm tác giả đã chỉ ra với mô hình thực nghiệm của họ, tỷ lệ dữ liệu đơn ngữ sử dụng bằng 50% lượng dữ liệu đơn ngữ thì mô hình đạt kết quả tốt nhất (gấp 5 lần dữ liệu song ngữ). Nếu tiếp tục tăng dữ liệu đơn ngữ thì chất lượng của mô hình bắt đầu giảm dần. Chi tiết kết quả như Bảng 2.

Bảng 2. Kết quả thực nghiệm sử dụng kĩ thuật tự học trong nghiên cứu của Zhang và cộng sự (2016) với tập huấn luyện 600 nghìn cặp câu Trung - Anh

Mô hình dịch	MT03	MT04	MT05	MT06
Moses	30.30	31.04	28.19	30.04
RNNSearch	28.38	30.85	26.78	29.27
RNNSearch-Mono-SL (25%)	29.65	31.92	28.65	29.86
RNNSearch-Mono-SL (50%)	**32.43**	**33.16**	**30.43**	**32.35**
RNNSearch-Mono-SL (75%)	30.24	31.18	29.33	28.82
RNNSearch-Mono-SL (100%)	29.97	30.78	26.45	28.06

Trên tập huấn luyện lớn 2,1 triệu cặp câu, nhóm tác giả lại tiếp tục chỉ ra nếu sử dụng 50% lượng dữ liệu đơn ngữ (gấp 5 lần dữ liệu song ngữ) thì mô hình cho kết quả tốt nhất. Chi tiết kết quả như Bảng 3.

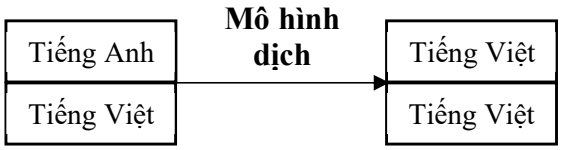
Bảng 3. Kết quả thực nghiệm sử dụng kĩ thuật tự học trong nghiên cứu của Zhang và cộng sự (2016) với tập huấn luyện 2.1 triệu cặp câu Trung - Anh.

Mô hình dịch	MT03	MT04	MT05	MT06
RNNSearch	35.18	36.20	33.21	32.86
RNNSearch-Mono-MTL (50%)	**36.32**	**37.51**	**35.08**	**34.26**
RNNSearch-Mono-MTL (100%)	35.75	36.74	34.23	33.52

3.1.3 Kĩ thuật dịch trộn lẫn (Mix-source)

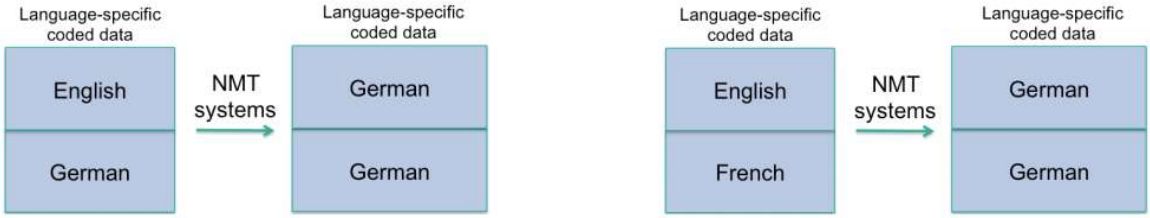
Năm 2017, nhóm tác giả Ha và cộng sự đã giới thiệu kĩ thuật dịch trộn lẫn (mix-source) nhằm tạo ra song ngữ tổng hợp (giả song ngữ) bằng cách tạo ra bản sao của dữ liệu đơn ngữ phía ngôn ngữ đích [8, 9]. Kĩ thuật này giúp cho mô hình NMT dịch tốt hơn với các thuật ngữ hoặc các tên riêng tồn tại cả phía ngôn ngữ nguồn và ngôn ngữ đích. Ví dụ, một tên riêng Alibaba trong văn bản tiếng Anh khi được dịch sang tiếng Việt thì giữ nguyên là Alibaba mà không cần dịch. Những cặp ngôn ngữ kiểu này sẽ có nhiều lợi thế từ kĩ thuật dịch trộn lẫn.

Dịch trộn lẫn có thể được minh họa như Hình 3, được mô tả như sau: Cho cặp ngôn ngữ gồm N cặp câu (X_1, Y_1), ta lấy M cặp câu song ngữ giả (X_2, X'_2) với X'_2 là bản sao của X_2, rồi kết hợp cả hai tập dữ liệu để tạo thành tập dữ liệu mới gồm N+M cặp câu. Sau đó, mô hình dịch sẽ được huấn luyện trên tập dữ liệu mới này, dữ liệu đơn ngữ có thể sử dụng ở phía ngôn ngữ nguồn hoặc ngôn ngữ đích hoặc kết hợp cả hai. Kĩ thuật dịch trộn lẫn không đòi hỏi sử dụng một mô hình dịch từ nguồn sang đích hay từ đích sang nguồn như kĩ thuật dịch đảo ngược hay kĩ thuật tự học.



Hình 3. Minh họa kĩ thuật Mix-source

Kĩ thuật dịch trộn lẫn tương đối đơn giản và dễ tích hợp vào các mô hình NMT nhưng lại cải thiện đáng kể về chất lượng mô hình cho các cặp ngôn ngữ nghèo tài nguyên. Trong kịch bản thực nghiệm, nhóm tác giả đã thực hiện trộn lẫn các ngữ liệu song ngữ từ nguồn TED (khoảng 13 đến 17 ngàn cặp câu) với một lượng lớn dữ liệu đơn ngữ được trích xuất từ nhiều nguồn khác nhau, được công bố bởi hội nghị WMT để huấn luyện mô hình (khoảng 3 triệu cặp câu), thực hiện đánh giá trong hai tình huống mix-source và mix-multi-source minh họa như Hình 4, được mô tả như sau:



(a). Mô hình Mix-Source (b). Mô hình Mix-Multi-source

Hình 4. Mô hình Mix-source và Mix-multi-source trong nghiên cứu của Ha và cộng sự

- Với mô hình Mix-Source (a), nhóm tác giả chỉ sử dụng một cặp song ngữ chuẩn tiếng Anh - tiếng Đức và bổ sung dữ liệu song ngữ giả từ tiếng Đức với câu nguồn là bản sao của câu đích.
- Với mô hình Mix-multi-source (b), nhóm tác giả sử dụng hai cặp song ngữ chuẩn tiếng Anh - tiếng Đức và tiếng Pháp - tiếng Đức. Mục đích của mô hình (b) là để đánh giá hiệu quả của phương pháp đề xuất trong mô hình (a).

Kết quả thực nghiệm cho thấy, mô hình (a) vẫn cho kết quả tốt hơn mô hình (b), mặc dù chỉ sử dụng một lượng dữ liệu nhỏ song ngữ thực sự. Chi tiết kết quả điểm BLEU như Bảng 4.

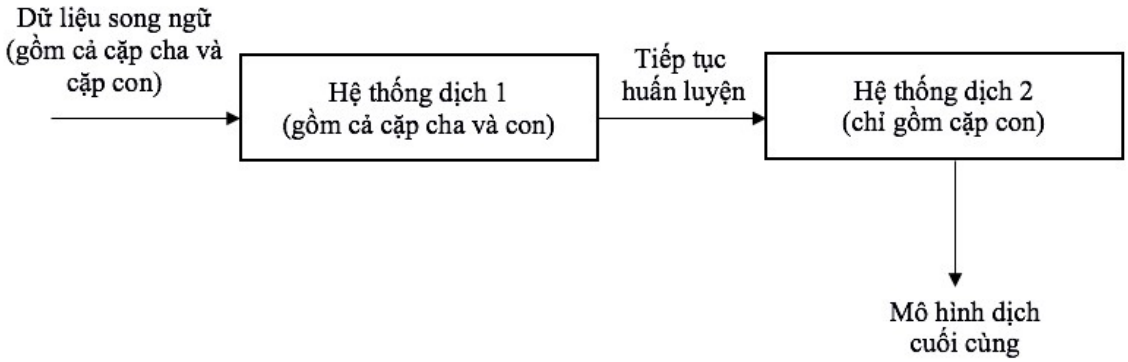
Bảng 4. Kết quả thực nghiệm mô hình dịch Anh - Đức sử dụng Mix-Source và Mix-Multi-Source trong nghiên cứu của Ha (2016).

Mô hình dịch	tst2013		tst2014	
	BLEU	ΔBLEU	BLEU	ΔBLEU
Baseline (En→De)	24.35	-	20.62	-
Mix-source (En,De→De,De)	**26.99**	**+2.64**	**22.71**	**+2.09**
Mix-multi-source (En,Fr→De,De)	26.64	+2.21	22.21	+1.59

3.2 Phương pháp học chuyển đổi (Transfer Learning)

Năm 2016, nhóm tác giả Barret Zoph và cộng sự đã chỉ ra chất lượng mô hình dịch máy với các cặp ngôn ngữ nghèo tài nguyên sẽ đạt được sự cải thiện lớn khi sử dụng phương pháp học chuyển đổi (transfer learning) [16].

Ý tưởng chính của phương pháp học chuyển đổi là dịch chuyển, kế thừa các tri thức của mô hình đã được huấn luyện trên các cặp ngôn ngữ giàu tài nguyên (gọi là mô hình cha) sang các cặp ngôn ngữ nghèo tài nguyên hơn (gọi là mô hình con), có thể được minh họa như Hình 9.



Hình 9. Minh hoạ phương pháp học chuyển đổi

Ví dụ về phương pháp học chuyển đổi:

- Giả sử có hai cặp ngôn ngữ:

(1) Tiếng Pháp - Tiếng Anh: cặp ngôn ngữ giàu tài nguyên (vài triệu cặp câu).

(2) Tiếng Thổ Nhĩ Kỳ - Tiếng Anh: cặp ngôn ngữ nghèo tài nguyên (vài nghìn cặp câu).

- Ban đầu, mô hình dịch được huấn luyện với dữ liệu là sự trộn lẫn của cả hai cặp ngôn ngữ trên theo chiều đích là tiếng Anh. Sau đó, mô hình tiếp tục được huấn luyện chỉ sử dụng dữ liệu của cặp ngôn ngữ nghèo tài nguyên (tiếng Thổ Nhĩ Kỳ - tiếng Anh). Mô hình cuối cùng này cho kết quả dịch tốt hơn hẳn so với mô hình nếu chỉ được huấn luyện với dữ liệu song ngữ của cặp ngôn ngữ nghèo tài nguyên.

Nhóm tác giả cũng đã chứng minh rằng nếu ngôn ngữ nguồn của cặp ngôn ngữ nghèo tài nguyên có hình thái, cấu trúc càng gần với ngôn ngữ nguồn của cặp giàu tài nguyên thì sự cải thiện về chất lượng của mô hình dịch càng lớn. Ngược lại, nếu các ngôn ngữ nguồn có sự khác biệt lớn về mặt từ vựng và trật tự câu thì sự cải thiện đạt được ít hơn hoặc không đáng kể [16]. Phương pháp này đơn giản để thực nghiệm, không cần phải thay đổi kiến trúc mạng mô hình dịch NMT. Tuy nhiên, phương pháp này yêu cầu cần phải có sẵn các cặp ngôn ngữ giàu tài nguyên.

Trong báo cáo [16], tác giả thực nghiệm và đánh giá trên 4 cặp ngôn ngữ con với ngôn ngữ nguồn là: Hausa, Turkish, Uzbek, Urdu và ngôn ngữ đích là tiếng Anh; cặp ngôn ngữ cha là tiếng Pháp - tiếng Anh. Kích thước các tập dữ liệu và điểm BLEU tương ứng của các mô hình dịch cơ bản dựa trên thống kê (SBMT) và mô hình cơ bản dựa trên kiến trúc mạng nơron (NMT) đối với từng cặp được chi tiết trong Bảng 6.

Bảng 6. Kích thước tập dữ liệu và điểm BLEU của các mô hình dịch cơ bản (1K tương đương 1000 cặp câu, 1M tương đương 1 triệu cặp câu)

Cặp ngôn ngữ	Dữ liệu huấn luyện	Test	SBMT BLEU	NMT BLEU
Hausa - Anh	1.0M	11.3K	23.7	16.8
Turkish - Anh	1.4M	11.6K	20.4	11.4
Uzbek - Anh	1.8M	11.5K	17.9	10.7
Urdu - Anh	0.2M	11.4K	17.9	5.2

Cặp ngôn ngữ cha là Pháp - Anh được sử dụng để nâng cao chất lượng dịch cho 4 cặp ngôn ngữ con bao gồm: Hausa → Anh, Turkish → Anh, Uzbek → Anh và Urdu → Anh. Từng cặp ngôn ngữ con được huấn luyện kết hợp với cặp ngôn ngữ cha, sau đó tiếp tục được làm mịn đối với lần lượt từng cặp ngôn ngữ con. Kết quả thực nghiệm được trình bày trong Bảng 7.

Bảng 7. Kết quả thực nghiệm sử dụng phương pháp học chuyển đổi với ngôn ngữ đích là tiếng Anh.

Mô hình dịch	Hausa	Turkish	Uzbek	Urdu
NMT	16,8	11,4	10,7	5,2
Xfer	21,3	17,0	14,4	13,8
Final	24,0	18,7	16,8	14,5
SBMT	23,7	20,4	17,9	17,9

Trong Bảng 7, mô hình Xfer là kết quả sử dụng phương pháp học chuyển đổi từ mô hình cha sang mô hình con, kết quả cho thấy điểm BLEU đều cao hơn so với các mô hình con cơ bản NMT. Mô hình Final là kết quả đánh giá khi sử dụng kĩ thuật kết hợp đầu ra từ 8 mô hình của mô hình, mô hình SBMT là kết quả từ mô hình dịch thống kê.

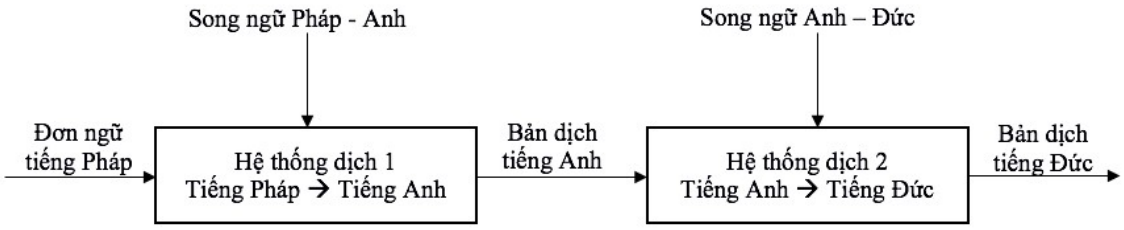
3.3 Phương pháp sử dụng ngôn ngữ trục (Pivot Language)

Năm 2017, nhóm tác giả Cheng và cộng sự giới thiệu phương pháp sử dụng ngôn ngữ trục (Pivot Language) cho dịch máy [2], sau đó có nhiều nghiên cứu tập trung theo hướng tiếp cận này [4]. Phương pháp này dựa trên ý tưởng các cặp ngôn ngữ giàu tài nguyên có thể được sử dụng làm cầu nối trung gian trong việc dịch giữa các cặp ngôn ngữ nghèo tài nguyên.

Ví dụ, hai cặp ngôn ngữ giàu tài nguyên tiếng Anh - tiếng Pháp và tiếng Anh - tiếng Đức được sử dụng để bắc cầu cho việc dịch giữa cặp ngôn ngữ tiếng Pháp - tiếng Đức.

- Trước hết, cặp ngôn ngữ tiếng Pháp - tiếng Anh được huấn luyện theo chiều từ tiếng Pháp sang tiếng Anh để sinh mô hình dịch thứ nhất (mô hình A). Sau đó, cặp ngôn ngữ tiếng Anh - tiếng Đức được huấn luyện theo chiều từ tiếng Anh sang tiếng Đức để sinh mô hình dịch thứ hai (mô hình B).

- Để dịch các câu trong văn bản từ tiếng Pháp sang tiếng Đức thì đưa văn bản lần lượt qua hai mô hình dịch trên (mô hình A và B) để thu được bản dịch tiếng Đức. Phương pháp này có thể được minh họa như Hình 10.



Hình 10. Minh hoạ kiến trúc mô hình dịch sử dụng ngôn ngữ trục

Trong Hình 10, dữ liệu song ngữ được sử dụng để huấn luyện mô hình dịch máy, dữ liệu đơn ngữ được sử dụng cho quá trình suy luận, tiếng Anh được sử dụng làm ngôn ngữ trục (ngôn ngữ trung gian) cho việc dịch văn bản từ tiếng Pháp sang tiếng Đức. Kết quả thực nghiệm cho thấy sự cải thiện đáng kể trên tập đánh giá giữa tiếng Tây Ban Nha → tiếng Pháp và tiếng Đức → tiếng Pháp từ +3 đến +13 điểm BLEU, chi tiết như trong Bảng 8 (các tập dữ liệu được lấy từ kho Europarl và WMT).

Bảng 8. Kết quả thực nghiệm từ tiếng Tây Ban Nha → tiếng Pháp, tiếng Đức → tiếng Pháp sử dụng tiếng Anh làm ngôn ngữ trục

Kho ngữ liệu	Cặp ngôn ngữ	Điểm BLEU	
		Không sử dụng ngôn ngữ trục	Sử dụng ngôn ngữ trục
Europarl	es -> fr	26,37	29,79
	de -> fr	14,02	23,70
WMT	es -> fr	11,75	24,60

Mặc dù việc sử dụng ngôn ngữ trục giúp cải thiện đáng kể chất lượng mô hình của các cặp ngôn ngữ nghèo tài nguyên nhưng phương pháp này đòi hỏi phải huấn luyện nhiều mô hình dịch riêng rẽ giữa các cặp ngôn ngữ khác nhau khiến cho mô hình dịch trở nên cồng kềnh và khó khả thi trong tình huống mô hình có nhiều cặp ngôn ngữ khác nhau.

4. KẾT LUẬN

Bài báo đã giới thiệu, cung cấp một cái nhìn tổng quan về một số phương pháp phổ biến được sử dụng cho dịch máy để cải thiện chất lượng dịch đối với các cặp ngôn ngữ hiếm và ít tài nguyên song ngữ. Trong ba cách tiếp cận được trình bày ở trên thì: (1) Các phương pháp khai thác dữ liệu đơn ngữ được sử dụng phổ biến nhất do có ưu điểm là đơn giản nhưng rất hiệu quả, có thể dễ dàng tích hợp vào các mô hình dịch máy mà không làm thay đổi kiến trúc mô hình. Tuy nhiên, lượng dữ liệu đơn ngữ tăng đến một ngưỡng thì chất lượng mô hình không tăng được nữa; (2) Phương pháp học chuyển đổi cũng mang lại hiệu quả cao, đặc biệt hiệu quả đối với các cặp ngôn ngữ có sự tương đồng lớn về từ vựng và cú pháp, vì vậy hiệu quả sẽ không cao nếu các cặp ngôn ngữ khác nhau quá nhiều; (3) Phương pháp sử dụng ngôn ngữ trục có ưu điểm rất hiệu quả nhưng nhược điểm là yêu cầu nhiều mô hình huấn luyện riêng rẽ gây lãng phí tài nguyên, bộ nhớ.

Trong tương lai, chúng tôi tiếp tục thử nghiệm thêm các phương pháp trên kết hợp với một số phương pháp để xử lý, khai phá dữ liệu

nhằm mục tiêu có được lượng dữ liệu song ngữ chất lượng tốt hơn cho dịch máy, đồng thời tiến hành thực nghiệm giữa cặp ngôn ngữ tiếng Việt với một số ngôn ngữ ít phổ biến khác.

TÀI LIỆU THAM KHẢO

1. Kyunghyun Cho, Bart van Merrienboer, C¸alar Gulc¸ehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014), Learning phrase representations using rnn encoder-decoder for statistical machine translation, *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*), pages 1724-1734, Doha, Qatar, October. Association for Computational Linguistics.

2. Yong Cheng, Yang Liu, Qian Yang, Maosong Sun and Wei Xu, *Joint Training for Pivot-based Neural Machine Translation*, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).

3. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Ben-gio (2015), *Neural machine translation by jointlylearning to align and translate*, Proceedings of Inter-national Conference on Learning Representations.

4. Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi2 Hermann Ney (2019), *Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 866-876, Hong Kong, China, November 3-7, 2019 Association for Computational Linguistics.

5. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan 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, June. Association for Computational Linguistics.

6. Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, Marc'Aurelio Ranzato (2019), ”*Two New Evaluation Datasets for Low-Resource Machine Translation: Nepali-English and Sinhala-English*”. CoRR abs/1902.01382.

7. Gu, Jiatao and Hassan, Hany and Devlin, Jacob and Li, Victor O.K (2018), Universal Neural Machine Translation for Extremely Low Resource Languages, *Proceedings of the 2018 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344-354, New Orleans, Louisiana, 2018, Association for Computational Linguistics.

8. Thanh-Le Ha, Jan Niehues, and Alexander Waibel (2017), Effective Strategies in Zero-Shot Neural Machine Translation.

9. Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel (2016), Toward multilingual neural machine translation with universal encoder and decoder. CoRR,abs/1611.04798.

10. Thi-Vinh Ngo, Thanh-Le Ha, Phuong-Thai Nguyen,and Le-Minh Nguyen (2018), *Combining advancedmethods in japanese-vietnamese neural machinetranslation*, 2018 10th International Conference onKnowledge and Systems Engineering (KSE), pages318-322.

11. Rico Sennrich, Barry Haddow, and Alexandra Birch (2016a), Improving neural machine translation modelswith monolingual data. CoRR, abs/1511.06709.

12. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le (2014), Sequence to sequence learning with neural networks, *In Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3104-3112, Cambridge, MA, USA. MIT Press

13. Ashish Vaswani, Noam Shazeer, Niki Parmar, JakobUszkoreit, Llion Jones, Aidan N. Gomez, LukaszKaiser, and Illia Polosukhin (2017), *Attention is allyou need*, CoRR, abs/1706.03762.

14. Jiajun Zhang and Chengqing Zong (2016), Exploiting source-side monolingual data in neural machinetranslation, *In Proceedings of the 2016 Conferenceon Empirical Methods in Natural Language Process-ing*, pages 1535-1545, Austin, Texas. Associationfor Computational Linguistics.

15. Jiajun Zhang and Chengqing Zong (2016), “*Bridging neural machine translation and bilingual dictionaries*”, arXiv preprint arXiv:1610.07272.

16. Zoph, Barret, Yuret, Deniz, May, Jonathan, and Knight, Kevin (2016), *Transfer learning for low-resource neural machine translation*, In EMNLP.