

BƯỚC ĐẦU XÂY DỰNG TỰ ĐỘNG KHO NGỮ LIỆU SONG SONG THƠ CA CHỮ HÁN CỦA VIỆT NAM CÓ BẢN DỊCH NGHĨA TIẾNG VIỆT HIỆN ĐẠI

THÁI HOÀNG LÂM¹

DƯƠNG THI AN²

ĐÌNH ĐIỀN³

Abstract: In the progress of excavating the treasures of Vietnamese Han-script documents, translating their meanings into contemporary Vietnamese is a crucial step in assisting modern Vietnamese people in learning many priceless lessons from our ancestors. To support the interpretation of a large number of these writings, we should create or enhance machine translation tools. However, due to a shortage of high-quality parallel corpora, the existing machine translation systems have struggled to do this task effectively, especially for the poetic domain. In light of these challenges, we propose the utilization of computer software to automatically create a Vietnamese Han-script corpus with contemporary Vietnamese interpretation, concentrating on the domain of Vietnamese Han script poetry. Our research shows that we constructed a parallel corpus with more than 4000 Vietnamese Han-script poems, each having a Vietnamese interpretation and including information about the author and the relevant dynasties, etc. This corpus encourages automatic interpretation of Vietnam's literary works written in Han script.

Keywords: *Han-script, Vietnamese, poetry, interpretation, corpus, ...*

1. Giới thiệu

Trong suốt hơn 20 thế kỉ trước, chữ Hán được sử dụng rộng rãi và đóng vai trò quan trọng trong đời sống văn hoá của dân tộc ta. Suốt ngàn ấy năm, kể cả khi có chữ Nôm xuất hiện, chúng ta vẫn tiếp tục sử dụng chữ Hán để ghi chép tài liệu trong nhiều lĩnh vực đời sống khác nhau và nhiều tài liệu còn được lưu trữ đến hiện tại. Đó là những sản phẩm văn hoá đáng quý cần được bảo tồn, khai thác và phát huy giá trị. Trong đó, không thể không kể đến kho tàng thơ ca chữ Hán đồ sộ mang những giá trị bất hủ đối với nền văn học nước nhà. Chúng ta có thể thấy các tác phẩm thơ chữ Hán như “Bạch Đằng Giang phú” (Trương Hán Siêu), “Tỏ lòng”

¹ Cử nhân, Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên, ĐHQG Thành phố Hồ Chí Minh

² Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên, ĐHQG Thành phố Hồ Chí Minh

³ PGS.TS, Trung tâm Ngôn ngữ học tính toán, Trường Đại học Khoa học Tự nhiên, ĐHQG Thành phố Hồ Chí Minh

Chúng tôi xin chân thành cảm ơn trang web Thi Viện (thivien.net) đã tạo điều kiện cho cộng đồng đóng góp, chia sẻ các bài thơ văn cổ được viết bằng chữ Hán để kiến tạo nên một kho lưu trữ thi ca đồ sộ. Điều này giúp việc sưu tập dữ liệu có tính sáng tạo cho các nghiên cứu khoa học như nghiên cứu của chúng tôi có nguồn tư liệu dồi dào hơn.

(Phạm Ngũ Lão), “Độc Tiểu Thanh kí” (Nguyễn Du),... đã được phiên âm, dịch nghĩa, dịch thơ và thậm chí còn được đưa vào sách giáo khoa Ngữ văn của chương trình phổ thông. Điều này cho thấy được giá trị to lớn của việc lưu trữ, số hoá, dịch thuật kho tàng thơ ca này sang tiếng Việt hiện đại nhằm giúp cho các hệ trẻ ngày nay và mai sau dễ dàng tiếp thu hơn các tri thức vô giá của cha ông đang được mã hoá trong những vần thơ ấy.

Hiện nay, chữ Hán là hệ chữ ghi ý và cũng là loại văn tự “có nguồn gốc tượng hình duy nhất trên thế giới còn được sử dụng và phát triển cho đến ngày nay” [1, tr. 27]. Vì vậy, việc nhập liệu các tài liệu chữ Hán lên máy tính không quá khó như các văn bản chữ Nôm của dân tộc ta, đặc biệt ở mặt phonetic. Tuy nhiên, ở nước ta ngày nay, không phải ai cũng biết chữ Hán, vì chữ Hán không còn được dùng như một hệ chữ viết chính trong đời sống mà thay vào đó là chữ Quốc ngữ. Vì vậy, để mọi người ở các lĩnh vực, độ tuổi, trình độ,... khác nhau có thể hiểu được các văn bản chữ Hán, đặc biệt là văn bản chữ Hán ở thể loại thơ ca không phải là điều đơn giản. Bởi lẽ, ngôn ngữ thơ là một lớp ngôn từ đặc biệt, hay còn gọi là “ý tại ngôn ngoại”. Do đó, việc dịch thuật thơ ca chữ Hán mà đặc biệt là thơ ca chữ Hán của Việt Nam chưa bao giờ là dễ dàng, vì cần phải vận dụng và kết hợp rất nhiều các tri thức về ngôn ngữ và cả ngoài ngôn ngữ. Trong đó, việc chuyển dịch thơ chữ Hán sang tiếng Việt hiện đại thường bao gồm các bước: phiên âm Hán Việt, dịch nghĩa, dịch thơ. Tùy mục đích mà chúng ta chọn sử dụng phương pháp dịch nào, không có phương pháp dịch nào biểu đạt được hết những đặc trưng của tác phẩm gốc [2]. Nhưng một cách rõ ràng nhất, điều chúng ta hướng đến trước hết là hiểu được ý nghĩa bài thơ, thế nên các bài thơ chữ Hán được in trong sách giáo khoa không chỉ có bản dịch thơ mà còn kèm theo phần dịch nghĩa. Vì vậy, có thể nói việc dịch nghĩa các bài thơ chữ Hán là bước thiết yếu trong việc truyền tải được ý nghĩa, thông điệp tiềm ẩn trong những vần thơ ấy.

Chính vì lí do trên cùng với sự phát triển vượt bậc của lĩnh vực Trí tuệ nhân tạo nói chung và Xử lí ngôn ngữ tự nhiên nói riêng, thiết nghĩ chúng ta nên phát triển các hệ thống có khả năng dịch tự động nghĩa của các tác phẩm thơ ca chữ Hán của Việt Nam sang tiếng Việt hiện đại. Mục tiêu của việc này là để những người không có chuyên môn trong lĩnh vực Hán-Nôm cũng có thể tiếp cận được các văn bản chữ Hán chưa từng được học hoặc đơn giản là khi họ tham quan các đền, chùa có thể dùng một ứng dụng dịch tự động để phân nào hiểu được các câu đối, bài thơ, liễn viết bằng chữ Hán. Một câu hỏi đặt ra là tại sao chúng ta không sử dụng ứng dụng dịch tiếng Trung để dịch những bài thơ chữ Hán của Việt Nam? Chúng tôi có thể trả lời câu hỏi này như sau: bởi lẽ các ứng dụng dịch tự động ấy thường được huấn luyện bằng ngữ liệu tổng quát, không thiên về lĩnh vực thơ ca, đặc biệt là thơ ca ở các thể kỉ trước. Hơn nữa, sự khác biệt về văn hoá, lịch sử,... giữa 2 quốc gia Trung Quốc và Việt Nam cũng sẽ ảnh hưởng nhiều đến ngôn ngữ cho dù trước đó cả 2 dân tộc đều có sử dụng chữ Hán trong đời sống văn hoá cũng như sáng tác văn học. Ngoài ra, chữ Hán trong thơ văn cổ Việt Nam có những điểm khác với chữ Hán hiện đại. Thế nên, chúng ta vẫn có thể sử dụng những ứng dụng dịch tiếng

Trung ấy để dịch thơ ca chữ Hán của Việt Nam nhưng khó có thể đạt được bản dịch chất lượng cao hoặc bản dịch rất khó hiểu.

Nguyên nhân cốt lõi của việc các hệ dịch tự động hiện hữu trên mạng chưa cho kết quả dịch tốt ở lĩnh vực (domain) thơ ca chữ Hán của Việt Nam là do những các công cụ này vẫn chưa được trang bị một bộ ngữ liệu huấn luyện đủ tốt riêng cho lĩnh vực thơ ca chữ Hán của nước ta. Để giải quyết vấn đề này, chúng tôi bước đầu xây dựng một bộ ngữ liệu song song chữ Hán - bản dịch nghĩa tiếng Việt hiện đại ở lĩnh vực thơ ca chữ Hán của Việt Nam. Chúng tôi đề xuất phương pháp sử dụng chương trình máy tính để có thể xây dựng tự động số lượng lớn các mẫu ngữ liệu song song và đồng thời việc xây dựng cũng tuân theo các tiêu chí được định ra dựa trên cơ sở lí luận ngôn ngữ học - ngữ liệu.

Các đóng góp chính của nghiên cứu này:

1) Đề xuất phương pháp xây dựng tự động kho ngữ liệu thơ ca chữ Hán - bản dịch nghĩa tiếng Việt hiện đại với tỉ lệ thu thập thành công hơn 80% số lượng trang web có nội dung cần sưu tập.

2) Xây dựng tự động kho ngữ liệu song song văn bản chữ Hán - bản dịch nghĩa tiếng Việt ở lĩnh vực thơ ca chữ Hán của Việt Nam với hơn 4.000 bài thơ.

3) Thống kê kho ngữ liệu theo số dòng thơ, tác giả, thời kì và thể loại. Điều này góp phần cho công tác khai thác kho ngữ liệu ở các nghiên cứu tiếp theo được thuận tiện hơn.

2. Các công trình liên quan

Theo hiểu biết của chúng tôi, hiện nay chưa có một công trình nghiên cứu nào chú trọng vào việc xây dựng một kho ngữ liệu song song chữ Hán - tiếng Việt ở lĩnh vực thơ ca chữ Hán của Việt Nam (trong sự phân biệt với thơ ca chữ Hán của Trung Quốc) bằng phương pháp tự động và dựa trên các cơ sở lí luận của ngôn ngữ học - ngữ liệu. Ấy vậy, vẫn có một số công trình có liên quan ít nhiều đến ngữ liệu chữ Hán hoặc thơ ca chữ Hán của Việt Nam.

Đầu tiên, chúng ta cần kể đến trang web Thi Viện⁴ được lập ra từ năm 2004 và được duy trì đến ngày nay [7]. Trang web đã lưu trữ một lượng lớn thơ ca, trong đó thơ ca được viết bằng chữ Hán lên đến 17.209 bài thơ (số liệu đến ngày 30/7/2023 tại trang web thống kê của Thi Viện⁵). Bộ sưu tập đồ sộ những tác phẩm thơ này đã giúp cho người học, quan tâm đến thơ ca chữ Hán có được một nguồn tư liệu dồi dào để tham khảo, tra cứu. Hơn thế, việc trang web tổng hợp các bài thơ chữ Hán ở nhiều quốc gia (Trung Quốc, Việt Nam, Nhật Bản,...) còn tạo tiền đề cho chúng tôi sưu tập, chọn lọc và xây dựng tự động được kho ngữ liệu song song chữ Hán - bản dịch nghĩa tiếng Việt nhằm thực hiện các nghiên cứu xa hơn như: ứng dụng tin học để dịch nghĩa tự động các tác phẩm thơ chữ Hán của Việt Nam mà vẫn chưa có bản dịch,...

⁴ <https://www.thivien.net/>, Truy cập ngày 30/7/2023

⁵ <https://www.thivien.net/statistics.php?Mode=Poems>, Truy cập ngày 30/7/2023.

Tiếp theo, trong tiến trình khai thác và phát huy giá trị của kho tàng thư tịch Hán Nôm, đã có nhiều công trình khoa học xây dựng những bộ ngữ liệu song song chữ Nôm (loại văn tự được hình thành dựa trên chữ Hán) - chữ Quốc ngữ để phục vụ cho bài toán chuyển tự tự động. Điển hình là công trình của PGS. TS. Đinh Điền cùng các cộng sự năm 2021 [3] và đề tài nghiên cứu cấp sở do PGS. TS. Đinh Điền làm chủ nhiệm đã cho ra mắt một công cụ chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ trực tuyến⁶, trong đó phần ngữ liệu huấn luyện có sử dụng các tài nguyên ngôn ngữ liên quan đến Hán - Việt như tự - từ điển Hán Việt [4]. Điều này cho phép ta nhập một câu thơ chữ Hán vào công cụ này, hệ thống có thể chuyển tự sang âm Hán - Việt cho ta. Tuy nhiên, với câu chữ Nôm, khi chuyển tự sang chữ Quốc ngữ thì vẫn giữ được trật tự từ và ngữ pháp của tiếng Việt. Ngược lại, đối với các câu thơ chữ Hán khi được chuyển sang âm Hán Việt thì chung quy vẫn rất khó hiểu đối với người không biết tiếng Hán hoặc có chuyên môn ngoài lĩnh vực ngôn ngữ học.

Một công trình trước đây của Thái Hoàng Lâm và các cộng sự (2022) [8], có quan tâm đến các kho ngữ liệu về chữ Hán. Cụ thể, tác giả đã xây dựng kho ngữ liệu song song chữ Hán - phiên âm Hán Việt, bao gồm 26.957 dòng thơ chữ Hán của Việt Nam thời kì trung đại và 21.676 câu văn xuôi chữ Hán từ tác phẩm “Đại Việt sử kí toàn thư” để tích hợp vào mô hình chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ. Không dừng lại ở mức độ phiên âm Hán Việt, trong nghiên cứu lần này, chúng tôi chú trọng vào xây dựng kho ngữ liệu thơ ca chữ Hán của Việt Nam có bản dịch nghĩa tiếng Việt hiện đại.

3. Phương pháp đề xuất

Dựa trên định nghĩa về ngữ liệu của Sinclair [9] (được dẫn lại qua chuyên khảo “Ngôn ngữ học ngữ liệu” [5, tr.2] bởi PGS. TS. Đinh Điền) như sau: “ngữ liệu” là “Tập hợp các mẫu văn bản được lựa chọn một cách có hệ thống theo những tiêu chí nhất định nhằm đại diện cho một thể loại ngôn ngữ cụ thể với mục đích nghiên cứu ngôn ngữ”, chúng tôi đã định ra một số tiêu chí để làm cơ sở cho việc xây dựng kho ngữ liệu song song của nghiên cứu này (trình bày ở phía sau) với phương pháp sử dụng phần mềm máy tính để thu thập và sắp xếp một cách có hệ thống các mẫu văn bản song song; trong đó, ngôn ngữ nguồn đại diện cho thể loại thơ ca chữ Hán của Việt Nam thuộc nền văn học trung đại (từ thế kỉ X đến đầu thế kỉ XX) và ngôn ngữ đích là bản dịch nghĩa tiếng Việt hiện đại nhằm mục đích nghiên cứu dịch nghĩa tự động. Đồng thời, tham khảo các tiêu chí lựa chọn mẫu ngữ liệu theo công trình “Xử lí ngôn ngữ tự nhiên” [6], chúng tôi đã đưa ra 5 tiêu chí cho việc xây dựng kho ngữ liệu của nghiên cứu này. Thứ nhất, do chúng ta cần xây dựng một kho ngữ liệu song song để phục vụ cho bài toán dịch nghĩa tự động từ thơ ca chữ Hán của Việt Nam sang tiếng Việt hiện đại, nên các văn bản được thu thập phải bao gồm ít nhất 2 phần là bản chữ Hán và bản tiếng Việt hiện đại, trong đó bản tiếng Việt hiện đại phải là bản dịch nghĩa của bản chữ Hán (nếu có thể thì thêm một bản thứ 3 là phần phiên âm Hán - Việt của bản chữ Hán). Thứ hai, các bài thơ chữ Hán chúng tôi thu

⁶ <https://tools.clc.hcmus.edu.vn/>, Truy cập ngày 30/7/2023

thập phải là con người sáng tác chứ không phải do máy tạo ra. Vì hiện nay, với sự phát triển của trí tuệ nhân tạo tạo sinh (Generative AI) thì việc dùng các công cụ này sản sinh ra các bài thơ chữ Hán không phải là không thể, nhưng các bài thơ này chưa chắc đạt được chuẩn mực ngôn ngữ và cũng không phù hợp với mục đích thực tiễn như đã giới thiệu ở trên, đó là chúng ta cần các công cụ dịch nghĩa tự động để hỗ trợ chuyển ngữ kho tàng thơ ca chữ Hán của ông cha ta đã sáng tác trong những thế kỉ trước. Ngoài ra, nếu chúng ta sử dụng các văn bản do máy tạo ra này cũng không phù hợp với đối tượng nghiên cứu của ngôn ngữ học - ngữ liệu, đó là các “cứ liệu ngôn ngữ” được sử dụng trong thực tế [5, tr.11]. Thế nên, các mẫu ngữ liệu song song mà chúng tôi sưu tập phải từ các nguồn uy tín (hoặc văn bản được dẫn nguồn gốc, xuất xứ một cách rõ ràng), được nhiều người biết đến và sử dụng, bao gồm các bài thơ chữ Hán được các bậc tiền nhân của nước ta sáng tác và được lưu giữ đến ngày nay, cùng với bản dịch nghĩa do dịch giả người Việt hiện đại thực hiện. Thứ ba, chúng tôi ưu tiên xây dựng kho ngữ liệu từ các văn bản ở dạng điện tử để có thể tận dụng các chương trình máy tính trong việc thu thập, xử lý số lượng lớn một cách nhanh chóng. Thứ tư, ngoài phần văn bản song song mà chúng ta cần thu thập, chúng tôi cũng lưu trữ thêm các thông tin khác có thể có của mẫu ngữ liệu đó như tác giả, thời đại, thể loại,... một cách tường minh trong cấu trúc dữ liệu lưu trữ thông qua các trường “author” (tác giả), “age” (thời đại),... được thể hiện trong hình 3 bên dưới. Điều này giúp các nghiên cứu trên kho ngữ liệu như thống kê theo thể loại, so sánh các văn bản giữa các thời đại khác nhau,... một cách thuận tiện hơn. Một tiêu chí nữa đó là kho ngữ liệu ở dạng điện tử mà chúng tôi xây dựng phải được lưu trữ một cách có hệ thống, đồng nhất về cấu trúc, định dạng lưu trữ để dễ dàng cho việc truy vấn. Đồng thời, kho ngữ liệu phải được tiền xử lý, chuẩn hoá hay còn được gọi trong tin học là làm sạch dữ liệu. Chẳng hạn nếu kho ngữ liệu không được chuẩn hoá về mặt bỏ dấu thì cùng một chữ máy tính có thể sẽ “hiểu” thành hai hay nhiều chữ khác nhau, ví dụ trường hợp của chữ “thỏa” và “thỏa”. Điều này sẽ gây ra một số ảnh hưởng không tốt đến kết quả của các bài toán về xử lý ngôn ngữ tự nhiên nói chung và dịch tự động nói riêng nếu sử dụng một kho dữ liệu không được chuẩn hoá.

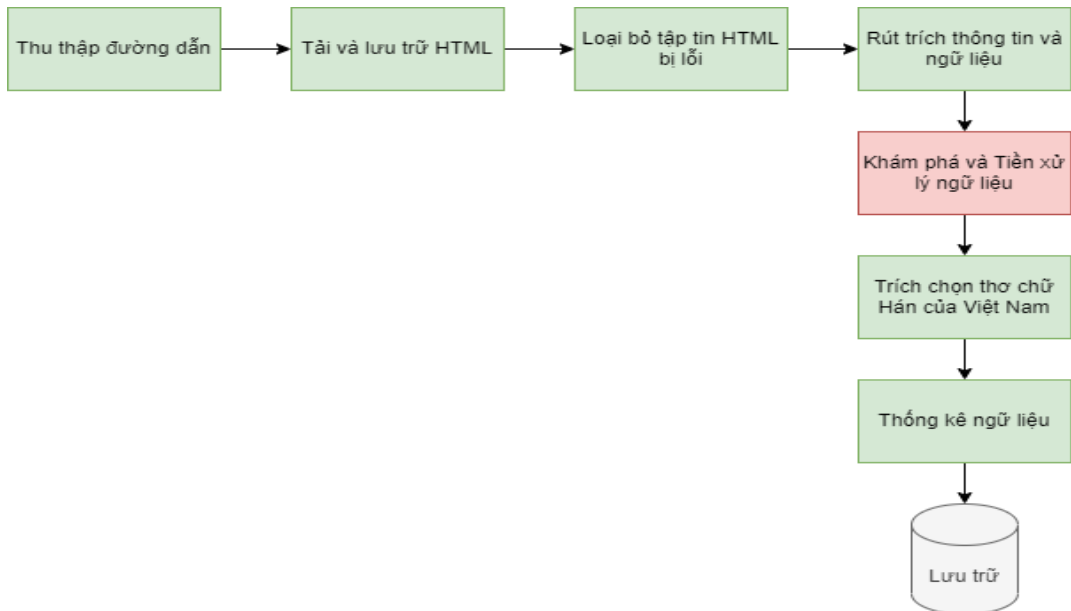
Dựa vào những tiêu chí đã đề ra ở trên cũng như để thu thập, xử lý và lưu trữ số lượng lớn ngữ liệu song song thơ ca chữ Hán của Việt Nam có bản dịch nghĩa tiếng Việt, chúng tôi đề xuất phương pháp sử dụng chương trình máy tính để thực hiện. Một cách tổng quan, chúng tôi dùng phần mềm máy tính để trích xuất các bài thơ chữ Hán và bản dịch nghĩa tương ứng kèm với các thông tin khác như tác giả, thời đại, thể thơ (nếu có),... từ các trang web có lưu trữ thơ ca chữ Hán với các thông tin về các văn bản thơ đó một cách tường minh nhất có thể, điển hình là trang web thivien.net. Sau đó lưu trữ dưới dạng tập tin .json (JSON là JavaScript Object Notation) và .xlsx (Excel). Quy trình thực hiện bao gồm các bước được mô tả trong Hình 1. Cách tiếp cận này lấy việc sưu tập thông qua Thi viện để minh hoạ tuy nhiên vẫn có thể tham khảo áp dụng cho các nguồn đáng tin cậy khác trên internet như Wikisource⁷,...

⁷ https://vi.wikisource.org/wiki/Trang_Chính, (tác giả truy cập lần sau cùng ngày 1/11/2023)

Bước 1: Trước tiên, chúng tôi cần thu thập một danh sách đường dẫn (url) đến các trang web thơ chữ Hán của Thi viện, bởi lẽ Thi viện lưu trữ mỗi bài thơ tương ứng với một trang web chứa đầy đủ các thông tin về bài thơ đó nhất có thể. Để có được danh sách các đường dẫn này, chúng tôi không thu thập thủ công mà rút trích từ các tập tin .html (HTML là HyperText Markup Language - Ngôn ngữ Đánh dấu Siêu văn bản) của các trang web về các mục chữ Hán trên trang web Từ điển Hán Nôm⁸ mà đã sưu tập từ nghiên cứu của công trình trước đó [8]. Mỗi trang web trình bày 1 mục chữ Hán thường sẽ có thêm phần “Một số bài thơ có sử dụng”. Trong phần này, trang web sẽ liệt kê ra danh sách các bài thơ chữ Hán có sử dụng Hán tự đó. Ví dụ như trang web tra chữ “槩” (Âm Hán Việt: sáo, sóc) của Thi viện có liệt kê ra các bài thơ được viết bằng chữ Hán mà có sử dụng Hán tự này (Hình 2), trong đó bao gồm bài thơ “Thuật hoài” của Phạm Ngũ Lão.

Bước 2: Sau khi có được các url, chúng tôi tiến hành viết chương trình máy tính (sử dụng ngôn ngữ lập trình Python) để tải tự động các file .html của trang web tương ứng với từng url. Mỗi tập tin .html sẽ chứa cấu trúc và nội dung của trang web về bài thơ chữ Hán đó.

Bước 3: Với các tập tin .html đã tải về, chúng tôi tiến hành loại bỏ các trang web bị tải lỗi hoặc không phải là các trang web của các bài thơ chữ Hán (sử dụng chương trình máy tính).



Hình 1: Quy trình xây dựng tự động kho ngữ liệu song song thơ ca chữ Hán của Việt Nam có bản dịch nghĩa tiếng Việt

Bước 4: Chúng tôi tiến hành rút trích tự động văn bản thơ bao gồm: bản chữ Hán, bản phiên âm Hán - Việt và bản dịch nghĩa bằng cách viết chương trình dựa trên cấu trúc của tập

⁸ <https://hvdic.thivien.net/>, Truy cập ngày 30/7/2023

tin .html đó. Ta phải đảm bảo rằng các bài thơ sau khi được rút trích phải được giống hàng đến mức dòng thơ (dựa trên cấu trúc tập tin .html, vì nếu rút trích xong bị mất tính song song sẽ làm kho ngữ liệu bị nhiễu dẫn đến hệ lụy không tốt cho các bài toán tiếp theo có sử dụng kho ngữ liệu này, chẳng hạn như dịch máy). Ngoài ra, chúng tôi cũng rút trích thêm các thông tin về tác giả, đất nước, thời kỳ, nguồn,... nếu có trong tập tin .html đó. Các thông tin và văn bản đã được rút trích này sẽ được lưu vào 1 tập tin .json. Cấu trúc của tập tin này gồm các thẻ chính (còn gọi là “key”) và các giá trị (“value”) tương ứng như sau:

- “file_name”: tên của tập tin html đã được phần mềm thu thập.

- “author”: tác giả của bài thơ.

- “parallel_text”: bao gồm các cặp câu song song. Trong đó, các thẻ con là “title” (tên bài thơ), “content” (nội dung bài thơ) và “paragraph” (nếu bài thơ được chia thành các đoạn thơ, thì phần này sẽ đánh dấu thứ tự và nội dung của từng đoạn thơ). Tính song song của ngữ liệu được thể hiện qua việc chúng ta sẽ đặt song song bản chữ Hán, bản Hán Việt và bản dịch nghĩa tiếng Việt hiện đại của từng dòng thơ trong cấu trúc lưu trữ bằng các “key”: “cn” (Chinese script) cho bản chữ Hán, “sv” (Sino-Vietnamese) cho bản Hán Việt và “vi” (Vietnamese) cho bản dịch nghĩa tiếng Việt hiện đại.

- “country”: quốc gia - nơi xuất xứ của tác phẩm.

- “authentication”: chứng chặn rằng bài thơ được trích dẫn từ nguồn đáng tin cậy (được cung cấp bởi trang web - nơi thu thập ngữ liệu).

- “reference_source”: nguồn tham khảo đáng tin cậy - nơi mà trang web đã trích dẫn bài thơ.

- “additional_information”: bao gồm thẻ “lang” lưu thông tin về ngôn ngữ nguồn, thẻ “type” lưu thể loại/ dạng thức của văn bản, thẻ “age” thể hiện giai đoạn / triều đại mà bài thơ ra đời.

- Một số thẻ khác như: “vie_trans_exist” chỉ ra rằng bài thơ có bản dịch nghĩa tiếng Việt hiện đại, “line_count” thể hiện số dòng thơ của bài thơ và “paragraph_count” cho biết số đoạn của bài thơ.

Hình 3 là một phần của tập tin .json đã lưu với cấu trúc dữ liệu đã mô tả, minh họa cho tác phẩm “Thuật hoài” của Phạm Ngũ Lão sau khi được rút trích.

Bước 5: Tiếp theo, chúng tôi tiến hành khám phá để tiền xử lý các văn bản và thông tin đã được trích xuất từ bước 4. Đây là bước cực kỳ quan trọng, vì bước này có ảnh hưởng lớn đến chất lượng của kho ngữ liệu. Giả sử, trong bước này, chương trình máy tính bị lập trình sai sẽ dẫn đến thông tin hoặc nội dung các bài thơ bị ánh xạ sai hoặc chứa các ký tự gây nhiễu cho kho ngữ liệu. Một số bước tiền xử lý được thực hiện như sau:

- 1) Loại bỏ các con số, cả số Latin và La Mã (vì hầu như các con số trong văn bản thơ sẽ thể hiện dưới dạng con chữ), dấu câu hoặc các ký tự đặc biệt không cần sử dụng khi dùng kho

ngữ liệu song song này cho mục đích cho vào các mô hình dịch máy tự động. Ví dụ: /, @, &, #, %, \, <, >, ...

2) Chuẩn hoá tiếng Việt về việc bỏ dấu, ví dụ “hóa” thành “hoá”,...

3) Loại bỏ các khoảng trắng thừa. Ví dụ: “sơn hà” => “sơn hà”

4) Một số bước tiền xử lí khác khi cần sử dụng kho ngữ liệu cho bài toán dịch máy sau này. Ví dụ như: viết thường văn bản phiên âm và dịch nghĩa, tách từ,...

Bước 6: Ở bước này, chúng tôi trích chọn ra những bài thơ chữ Hán của Việt Nam. Bởi lẽ, trong lúc thu thập tự động, có thể sẽ có rất nhiều bài thơ chữ Hán của nước khác lẫn vào, trong khi đối tượng chúng ta cần quan tâm là các bài thơ chữ Hán của Việt Nam. Việc lựa chọn này dựa trên thông tin về đất nước và tác giả mà trang web Thi viện có đính kèm trong từng bài thơ.

Bước 7: Cuối cùng chúng tôi tiến hành thống kê ngữ liệu theo: tác giả, thời kì, thể loại và nguồn trích dẫn. Trong đó, thông tin về nguồn trích dẫn đặc biệt quan trọng vì thông tin này cho chúng ta biết bài thơ được đăng trên Thi Viện có được trích dẫn từ các nguồn uy tín hay không, nghĩa là chất lượng bản dịch có tin tưởng được hay không.

Các bước trên trong quy trình xây dựng kho ngữ liệu song song thơ ca chữ Hán có bản dịch nghĩa tiếng Việt chủ yếu dựa vào sức mạnh của phần mềm máy tính. Điều này giúp tiết kiệm thời gian, công sức, chi phí và giảm được vấn đề lỗi do con người (human error, ví dụ như ghi thông tin của bài thơ này nhầm vào bài thơ khác).



Hình 2. Danh sách bài thơ chữ Hán có sử dụng Hán tự “樂”⁹

⁹ Từ điển Hán Nôm, <https://hvdic.thivien.net/whv/樂>, Truy cập ngày 30/7/2023

```
{'file_name': 'thivien_Pham-Ngũ-Lão_Thuật-hoài_raw.html',
'author': 'Phạm Ngũ Lão',
'parallel_text': {'title': {'cn': '述懷', 'sv': 'Thuật hoài', 'vi': 'Tò lòng'},
'content': [{'cn': '橫梁江山恰幾秋',
'sv': 'Hoành sóc giang sơn cấp kỳ thu,',
'vi': 'Cáp ngang ngọn giáo gìn giữ non sông đã mấy thu,'},
{'cn': '三軍獵虎氣吞牛',
'sv': 'Tam quân tỳ hổ khí thôn Ngưu.',
'vi': 'Ba quân như gấu hổ, át cả sao Ngưu Đẩu.'},
{'cn': '男兒未了功名債',
'sv': 'Nam nhi vị liễu công danh trái,',
'vi': 'Thân nam nhi nếu chưa trả xong nợ công danh,'},
{'cn': '羞聽人間說武侯',
'sv': 'Tu thính nhân gian thuyết Vũ hầu.',
'vi': 'Át thẹn thùng khi nghe người đời kể chuyện Vũ hầu.'}],
'paragraph': {'1': [{'cn': '橫梁江山恰幾秋',
'sv': 'Hoành sóc giang sơn cấp kỳ thu,',
'vi': 'Cáp ngang ngọn giáo gìn giữ non sông đã mấy thu,'},
{'cn': '三軍獵虎氣吞牛',
'sv': 'Tam quân tỳ hổ khí thôn Ngưu.',
'vi': 'Ba quân như gấu hổ, át cả sao Ngưu Đẩu.'},
{'cn': '男兒未了功名債',
'sv': 'Nam nhi vị liễu công danh trái,',
'vi': 'Thân nam nhi nếu chưa trả xong nợ công danh,'},
{'cn': '羞聽人間說武侯',
'sv': 'Tu thính nhân gian thuyết Vũ hầu.',
'vi': 'Át thẹn thùng khi nghe người đời kể chuyện Vũ hầu.'}]},
'country': 'Việt Nam',
'authentication': 'Bài thơ này được tham khảo từ ấn phẩm, nguồn tin chính thống hoặc có thể kiểm chứng',
'reference_source': 'Nguồn:1. Hợp tuyển thơ văn Việt Nam (tập II), NXB Văn học, Hà Nội, 19762. Thơ văn Lý Trần (tập II), NXB K
hoa học xã hội, 1988',
'additional_information': {'lang': 'Chữ Hán',
'type': 'Thất ngôn tứ tuyệt',
'age': 'Trần'},
'vie_trans_exist': True,
'line_count': 4,
'paragraph_count': 1}
```

Hình 3. Bài thơ chữ Hán “Thuật hoài” sau khi được rút trích nội dung từ tập tin .html tương ứng

4. Kết quả nghiên cứu

Trong phần này chúng tôi trình bày kết quả của quá trình xây dựng tự động kho ngữ liệu song song thơ ca chữ Hán của Việt Nam có bản dịch nghĩa tiếng Việt.

Chúng tôi thu thập được 5.249 bài thơ chữ Hán của Việt Nam, tỉ lệ thu thập tự động được là 5.249/6.095 tương ứng 86,12% lượng bài thơ chữ Hán của Việt Nam đang có mặt trên trang web Thi Viện¹⁰. Trong đó có 4.231 bài thơ có bản dịch nghĩa tiếng Việt, chiếm khoảng 80,6% tổng số lượng bài thơ thu thập được (4.231/5.249).

Vậy nên, kho ngữ liệu song song thơ ca chữ Hán của Việt Nam có bản dịch nghĩa tiếng Việt mà chúng tôi muốn xây dựng gồm 4.231 bài thơ. Chúng tôi đã thống kê chi tiết về các khía cạnh: tổng số dòng thơ, số lượng tác giả, thời kì, thể loại và số lượng bài thơ có trích dẫn nguồn, đối với 4.231 bài thơ chữ Hán có bản dịch nghĩa tiếng Việt hiện đại. Các số liệu này được thể hiện trong Bảng 1.

¹⁰ <https://www.thivien.net/searchpoem.php?ViewType=2&Country=2>, Truy cập gần nhất vào ngày 30/7/2023

Bảng 1: Thống kê chi tiết cho 4.231 bài thơ chữ Hán của Việt Nam có bản dịch nghĩa tiếng Việt hiện đại

	Số lượng
Dòng thơ	30.053
Tác giả	406
Thời kì	11
Thể loại	16
Bài thơ được trích dẫn nguồn	3.070

Theo kết quả thống kê ngữ liệu, ta thấy rằng trong 4.231 có đến 3.070 bài thơ được trích dẫn nguồn gốc, xuất xứ đầy đủ. Đây sẽ là kho ngữ liệu vàng nếu chúng ta sử dụng để huấn luyện và đánh giá mô hình dịch nghĩa tự động từ thơ văn chữ Hán của Việt Nam sang tiếng Việt hiện đại.

Trong đó, tổng 3 số dòng thơ trong 1 tác phẩm, 3 tác giả, 3 thời kì và 3 thể loại có nhiều tác phẩm nhất trong kho ngữ liệu được thể hiện trong Bảng 2.

Bảng 2. Thống kê tổng 3 số dòng thơ có trong 1 tác phẩm, 3 tác giả, 3 thời kì và 3 thể loại có mặt nhiều nhất trong kho ngữ liệu

	Tổng 3	Số lượng (Bài thơ)
Số dòng thơ	8	2.593
	4	1.431
	12	33
Tác giả	Nguyễn Khuyến	262
	Nguyễn Du	220
	Nguyễn Phúc Ứng Bình	216
Thời kì	Nguyễn	1.337
	Hậu Lê, Mạc, Trịnh - Nguyễn	1.056
	Cận đại	709
Thể loại	Thất ngôn bát cú	2.321
	Thất ngôn tứ tuyệt	1.249
	Ngũ ngôn bát cú	238

Từ kết quả thống kê trên, ta thấy rằng, lượng lớn các bài thơ chữ Hán của Việt Nam mà chúng tôi thu thập được có số dòng thơ là 8 và 4 chiếm đa số, hơn nữa 2 thể loại chiếm phần lớn là: thất ngôn bát cú và thất ngôn tứ tuyệt, cả 2 đều bắt nguồn từ thể thơ của Trung Quốc. Từ quan sát này, chúng ta có thể chú ý thêm một yếu tố đó là: thơ chữ Hán của Việt Nam chịu nhiều ảnh hưởng của các thể loại thơ Trung Quốc. Do đó tương lai chúng ta có thể khai thác thêm kho ngữ liệu thơ chữ Hán của Trung Quốc có bản dịch tiếng Việt để làm giàu thêm cho kho ngữ liệu song song dùng để huấn luyện mô hình dịch nghĩa tự động thơ chữ Hán của Việt Nam sang tiếng Việt hiện đại của chúng ta.

5. Kết luận và hướng phát triển

Vận dụng khả năng của máy tính thông qua việc viết các chương trình tự động và cơ sở lý luận của ngôn ngữ học ngữ liệu, chúng tôi đã xây dựng thành công một kho ngữ liệu song song chữ Hán - bản dịch nghĩa tiếng Việt ở lĩnh vực thơ ca chữ Hán của Việt Nam với số lượng hơn 4000 bài thơ. Đây sẽ là nguồn tài nguyên hữu dụng đối với nhiều nghiên cứu trong tương lai như dịch nghĩa tự động thơ ca chữ Hán của Việt Nam sang tiếng Việt, sử dụng phần chữ Hán và phiên âm Hán Việt để bổ sung tri thức Hán - Việt cho bài toán chuyển tự tự động chữ Nôm sang chữ Quốc ngữ, phân tích thơ, so sánh đối chiếu thơ chữ Hán của Việt Nam và Trung Quốc tự động, các nghiên cứu về văn tự Hán,...

Ngoài ra, bộ ngữ liệu chúng tôi xây dựng cũng mang ý nghĩa tích cực trong việc tích hợp liên ngành khoa học: tin học - ngôn ngữ học, Hán - Nôm. Chúng tôi đã vận dụng sức mạnh của máy tính để làm giàu kho ngữ liệu một cách tự động trong bối cảnh công nghệ thông tin đang bùng nổ như hiện nay. Điều này giúp tiết kiệm được nhiều thời gian và chi phí.

Trong tương lai, chúng tôi sẽ tiếp tục khai thác phần ngữ liệu thơ ca chữ Hán của Trung Quốc mà có nhiều mối tương quan với kho ngữ liệu thơ ca chữ Hán của Việt Nam để mở rộng kho ngữ liệu huấn luyện. Hơn thế nữa, chúng tôi cũng sẽ xây dựng mô hình dịch nghĩa tự động thơ ca chữ Hán của Việt Nam sang tiếng Việt hiện đại bằng cách tận dụng kho ngữ liệu song song đã xây dựng và các công nghệ dịch máy hiện đại.

TÀI LIỆU THAM KHẢO

1. Lê Đình Khản, *Từ vựng gốc Hán trong tiếng Việt*, NXB Đà Nẵng, 2010.
2. Lê Thị Huyền Trang, *Chuyển dịch một số bài thơ Nôm sang chữ Hán: Cảm nhận và chia sẻ*, TC Nghiên cứu nước ngoài, 155, 2019.
3. Dinh, D., Nguyen, P., & Nguyen, L. H., *Transliterating Nôm scripts into Vietnamese national scripts using statistical machine translation*, International Journal of Advanced Computer Science and Applications, 12(2), 2021.
4. *Tiến sĩ làm công cụ chuyển chữ Nôm sang chữ Quốc ngữ*, Truy xuất từ <https://tuoitre.vn/tien-si-lam-cong-cu-chuyen-chu-nom-sang-chu-quoc-ngu-20230501095100967.htm>. Truy cập: 30/07/2023.
5. Đinh Điền, *Ngôn ngữ học ngữ liệu* (Sách chuyên khảo), NXB ĐHQG Thành phố Hồ Chí Minh, 2018.
6. Đinh Điền, *Xử lý ngôn ngữ tự nhiên*, NXB ĐHQG Thành phố Hồ Chí Minh, 2006.
7. Thi Viện, <https://www.thivien.net/>. Truy cập: 30/7/2023
8. Lam H. Thai, Long H. B. Nguyen, & Dien Dinh, D., *Improve the automatic transliteration from Nôm scripts into Vietnamese National scripts by integrating Sino-Vietnamese knowledge into Statistical Machine Translation*, In *Proceedings of the 4th International Conference on Information Technology and Computer Communications*, pp. 69-77, 2022, June.
9. Sinclair, J., *Corpus concordance collocation*, Oxford University Press, HK, 1991.