Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XVI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Đà Nẵng, ngày 28-29/9/2023
DOI: 10.15625/vap.2023.0080

DỊCH NGHĨA TỰ ĐỘNG THƠ VĂN CHỮ HÁN CỦA VIỆT NAM SANG TIẾNG VIỆT ĐƯƠNG ĐẠI SỬ DỤNG DỊCH MÁY THỐNG KÊ

Thái Hoàng Lâm, Đinh Điền

Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên, TP. Hồ Chí Minh,
Đại học Quốc gia TP. Hồ Chí Minh

22C11063@student.hcmus.edu.vn, lamhoangthai.work@gmail.com, ddien@fit.hcmus.edu.vn

TÓM TẮT: *Dịch nghĩa các tác phẩm Hán văn của nước ta sang tiếng Việt đương đại là công việc thiết yếu trong tiến trình khai thác các giá trị của kho tàng di sản Hán - Nôm, đặc biệt là đối với lĩnh vực thơ văn chữ Hán của Việt Nam. Thiết nghĩ, cùng với sự bùng nổ của trí tuệ nhân tạo, nếu chúng ta có thể xây dựng một hệ dịch nghĩa tự động đủ tốt cho lĩnh vực văn bản này thì phải chăng đây sẽ là công cụ đắc lực đồng hành cùng các nhà Hán - Nôm trong sứ mệnh tìm về với các giá trị cổ xưa cũng như hỗ trợ những người Việt ngoài lĩnh vực Hán-Nôm (hoặc các ngành có liên quan) có thể hiểu được nghĩa của các văn bản chữ Hán của nước ta khi tham quan các đền, các khu di tích cổ,... Tuy nhiên, theo hiểu biết của chúng tôi thì hiện nay chưa có một hệ dịch tự động nào đặt trọng tâm vào dịch nghĩa lĩnh vực thơ văn chữ Hán của Việt Nam hoặc có vài công cụ chỉ đơn thuần là phiên âm Hán-Việt. Vì vậy, trong bài báo này, chúng tôi đặt những viên gạch đầu tiên cho bài toán dịch tự động thơ văn chữ Hán của Việt Nam sang tiếng Việt đương đại với hướng tiếp cận dịch máy thống kê. Kết quả nghiên cứu cho thấy, bài báo đã có những bước tiến đầu tiên khi đưa ra một mô hình dịch nghĩa thơ văn chữ Hán Việt Nam tốt hơn ba công cụ có khả năng dịch chữ Hán sang tiếng Việt phổ biến khác. Đây sẽ là động lực cho các nghiên cứu dịch tự động đối với lĩnh vực thơ văn chữ Hán của dân tộc ta về sau.*

Từ khóa: *Dịch nghĩa tự động, Thơ văn chữ Hán, Tiếng Việt, Dịch máy thống kê.*

I. GIỚI THIỆU

Trong những năm gần đây, việc ứng dụng trí tuệ nhân tạo vào tiến trình khai thác các thư tịch Hán - Nôm của Việt Nam đã nhận được nhiều sự quan tâm hơn từ các nhóm nghiên cứu và đã đạt được những thành tựu nhất định, chẳng hạn các công trình "Chuyển tự từ chữ Nôm sang chữ Quốc Ngữ" [1, 2], nhận diện ký tự quang học (Optical Character Recognition - OCR) các văn bản cổ Hán – Nôm [3]. Tuy vậy, trong tiến trình giải mã các tài liệu cổ Hán - Nôm, một bước khá quan trọng bên cạnh việc chuyển tự chữ Nôm sang chữ Quốc ngữ hay OCR, đó là dịch nghĩa các văn bản chữ Hán [4] thay vì chỉ phiên âm Hán Việt. Dịch nghĩa hay "dịch theo hướng giải thích" (Interpretation), tuy có thể không giữ được hình thức của tác phẩm gốc hay vần luật của thơ, nhưng truyền tải được ý nghĩa của bài thơ chữ Hán đó [5]. Bảng 1 bên dưới là ví dụ cho tác phẩm thơ chữ Hán "Tụng giá hoàn kinh sư" (Trần Quang Khải), bao gồm bản chữ Hán, bản phiên âm Hán Việt và bản dịch nghĩa tiếng Việt. Nhiều tác phẩm thơ chữ Hán của Việt Nam cũng đã được vào sách giáo khoa phổ thông để giảng dạy cho học sinh với cả bản phiên âm, dịch nghĩa và dịch thơ như các tác phẩm "Nam quốc sơn hà", "Tụng giá hoàn kinh sư", "Tỏ lòng",… Điều đó cho thấy dịch nghĩa đóng vai trò cốt lõi trong việc hiểu ý nghĩa của các tác phẩm thơ văn chữ Hán.

Bảng 1. Bài thơ "Tụng giá hoàn kinh sư" (Trần Quang Khải) với bản dịch nghĩa[1]

Chữ Hán	Phiên âm Hán Việt	Dịch nghĩa
"奪槊章陽渡, 擒胡鹹子關。 太平須努力, 萬古此江山。"	"Đoạt sáo Chương Dương độ, Cầm Hồ Hàm Tử quan. Thái bình tu nỗ lực, Vạn cổ thử gian san."	"Cướp giáo giặc ở bến Chương Dương, Bắt quân Hồ ở cửa Hàm Tử. Buổi thái bình nên gắng hết sức, Thì muôn đời mãi có giang sơn này."

Về ngôn ngữ của bản dịch nghĩa cho các văn bản chữ Hán của Việt Nam là tiếng Việt đương đại, vì đối tượng sử dụng bản dịch nghĩa này mà chúng ta hướng đến là người Việt hiện đại, chứ không phải là các thế hệ cổ nhân[4]. Một vấn đề cần nói đến ở đây nữa là lĩnh vực văn bản chính mà chúng ta đang hướng đến dịch nghĩa là các văn bản thơ văn chữ Hán của Việt Nam chứ không phải của Trung Hoa. Dù rằng, cả 2 dân tộc đều từng dùng chữ Hán để sáng tác văn học cổ trong quá khứ và có những điểm tương đồng nhau, nhưng dưới sự ảnh hưởng của văn hoá, lịch sử, địa lý, phong tục, tập quán,… của dân tộc Việt Nam ta, chữ Hán được sử dụng ở nước ta nói chung và các tác phẩm văn học chữ Hán do người Việt ta sáng tác nói riêng, đã có những đặc điểm, "giá trị riêng biệt và độc lập" so với chữ Hán ở Trung Hoa, đặc biệt là từ thế kỷ X - khi dân tộc ta bước vào kỷ nguyên độc lập [4]. Đó cũng chính là lý do và động lực cho việc chúng tôi muốn xây dựng một hệ dịch tự động với lĩnh vực (domain) tập trung vào thơ văn chữ Hán của Việt Nam.

Tuy nhiên, khác với phiên âm Hán Việt thì việc dịch nghĩa tự động thơ ca chữ Hán là một bài toán thật sự thách thức. Bởi lẽ, muốn dịch nghĩa được một bài thơ, bài văn chữ Hán cho dễ hiểu thì không thể chỉ bó hẹp trong phạm vi tri thức về ngôn ngữ, văn tự mà còn cần nhiều tri thức từ các lĩnh vực khác như cần am hiểu về văn học, lịch

[1] Thi viện, https://www.thivien.net/Trần-Quang-Khải/Tụng-giá-hoàn-kinh-sư/poem-qrQFLWYXm2n6Cs-nv-xKow, truy cập gần nhất vào ngày 30/8/2023

sử, văn hoá, để giải mã được các điển tích điển cố chẳng hạn,… Hơn nữa, thơ ca với lớp ngôn ngữ đặc trưng có "thể thức và phong cách riêng" [6], lời ít ý nhiều, thì việc dịch nghĩa lại không đơn giản. Dù vậy, với niềm tin vào sự phát triển của lĩnh vực xử lý ngôn ngữ tự nhiên, chúng tôi vẫn muốn đặt những bước chân đầu tiên cho việc giải quyết bài toán đầy thử thách, nhưng cũng đầy ý nghĩa này. Tại sao chúng tôi gọi là những bước chân đầu tiên? Bởi lẽ, theo khảo sát của chúng tôi trên các nguồn chính thống, uy tín đến thời điểm viết bài báo này như Google Scholar[2], ACL Anthology[3], IEEE Xplore[4],…thì chúng tôi chưa thấy có một công trình nào đã công bố về dịch nghĩa tự động lĩnh vực thơ văn chữ Hán của Việt Nam. Gần với đề tài của chúng tôi nhất có thể nói là các nghiên cứu dịch tự động từ thơ Trung Hoa cổ sang tiếng Trung hiện đại. Tuy nhiên, như đã giải thích ở trên, văn chương chữ Hán của Trung Hoa có những điểm khác biệt nhất định so với thơ văn chữ Hán của Việt Nam.

Để tiếp cận bài toán dịch nghĩa tự động thơ văn chữ Hán của Việt Nam sang tiếng Việt đương đại, chúng ta cần một bộ ngữ liệu song song chữ Hán - bản dịch nghĩa tiếng Việt để huấn luyện và đánh giá mô hình. Bộ ngữ liệu này đã được xây dựng tự động từ nguồn uy tín như các trang web lưu trữ các bài thơ chữ Hán của Thi viện (thivien.net). Hơn thế nữa, thiết nghĩ để sử dụng kho ngữ liệu này một cách hiệu quả nhất, chúng tôi đề xuất một phương pháp phân chia và kết hợp một cách hợp lý các tập ngữ liệu con của kho ngữ liệu này. Trong đó, bởi lẽ lượng ngữ liệu cho các bài thơ chữ Hán của Việt Nam có bản dịch nghĩa không đủ lớn để huấn luyện mô hình, nên chúng tôi bổ sung thêm vào tập huấn luyện các bài thơ chữ Hán cổ của Trung Quốc có bản dịch nghĩa tiếng Việt với giả định rằng giữa thơ chữ cổ chữ Hán của Trung Quốc và Việt Nam có nhiều điểm tương đồng mà mô hình có thể khai thác. Tuy nhiên, vì mục tiêu là đưa ra một hệ dịch tự động cho lĩnh vực thơ văn chữ Hán của Việt Nam nên tập kiểm tra chỉ bao gồm các bài thơ chữ Hán của Việt Nam được thivien.net dẫn nguồn với độ tin cậy, chất lượng cao. Về mặt mô hình học máy, trong nghiên cứu bước đầu cho bài toán này, vì với kho ngữ liệu hiện có thì đây là cặp ngôn ngữ ít tài nguyên (low-resource language pair), nên chúng tôi chọn dịch máy thống kê (Statistical Machine Translation - SMT) để đảm bảo chất lượng thay vì dịch máy mạng Nơ-ron (Neural Machine Translation - NMT)[7].

Đóng góp của bài báo gồm những điểm chính như sau:

- Đưa ra động lực nghiên cứu, thách thức để đặt những bước chân đầu tiên trong việc giải bài toán dịch nghĩa tự động thơ văn chữ Hán của Việt Nam sang tiếng Việt đương đại.
- Kết hợp kho ngữ liệu thơ ca chữ Hán của Trung Quốc có bản dịch nghĩa tiếng Việt vào tập huấn luyện nhằm mục đích tận dụng những tri thức về chữ Hán có điểm tương đồng với thơ ca chữ Hán của Việt Nam.
- Đề xuất một mô hình dịch máy SMT với kết quả bước đầu tốt hơn các công cụ dịch tự động khác ở lĩnh vực thơ văn chữ Hán của Việt Nam (tuy các công cụ có sẵn khác này có thể dịch văn bản chữ Hán, nhưng cho kết quả không khả thi đối với văn bản thơ ca chữ Hán của Việt Nam ta).
- Khảo sát vấn đề dịch nghĩa thơ chữ Hán tự động với đơn vị dòng thơ hay đơn vị đoạn thơ/ bài thơ sẽ cho kết quả tốt hơn đối với mô hình SMT.
- Khảo sát và đưa ra câu trả lời cho việc nên dịch nghĩa thẳng từ bản chữ Hán hay bản phiên âm Hán Việt.

II. CÁC CÔNG TRÌNH LIÊN QUAN

Hiện nay, dựa trên sự hiểu biết của chúng tôi thì chưa có các công trình dịch nghĩa tự động thơ văn chữ Hán của Việt Nam sang tiếng Việt được công bố trước đây. Tuy vậy, vẫn có một số công trình có liên quan ít nhiều đến ý tưởng, nghiên cứu của bài báo này.

Đầu tiên phải kể đến công cụ dịch tự động của Microsoft (Microsoft Translator). Năm 2021, Microsoft Translator [8] đã cho ra mắt một mô-đun (module) dịch ngôn ngữ văn chương tiếng Trung (literary Chinese) sang tiếng Trung hiện đại với phương pháp dịch máy dạng nơ-ron (NMT) và từ mô hình dịch này, họ cũng hỗ trợ dịch ngôn ngữ văn chương tiếng Trung sang hơn 90 ngôn ngữ khác. Microsoft Translator cũng phát biểu rằng, với việc ra mắt module dịch này nhằm mục đích giúp khách du lịch có thể hiểu được các văn bản cổ Tiếng Trung, các bài thơ được viết ở bảo tàng và di tích lịch sử,… Ngoài ra, bài đăng này cũng nói rằng, thách thức đối với mô hình dịch cho ngôn ngữ văn chương tiếng Trung là việc thiếu lượng dữ liệu huấn luyện chất lượng. Công trình này của Microsoft chính là động lực cho nghiên cứu của chúng tôi về bài toán gần tương tự nhưng với lĩnh vực là thơ văn chữ Hán của người Việt Nam.

Một nghiên cứu khác của Lam và các cộng sự [2] về "Chuyển tự tự động từ chữ Nôm sang chữ Quốc ngữ", các tác giả có đề cập đến kho ngữ liệu song song thơ văn chữ Hán - phiên âm Hán Việt với vai trò là nguồn tri thức Hán - Việt giúp việc chuyển tự các chữ Nôm giả tá (loại mượn cả hình và âm chữ Hán) được chính xác hơn khi sử dụng mô hình dịch máy thống kê (SMT). Mặc dù nghiên cứu trên đã có mối quan tâm đến lĩnh vực thơ ca chữ Hán của Việt Nam, nhưng bộ ngữ liệu được sử dụng chưa có bản dịch nghĩa tiếng Việt mà chỉ dừng ở mức phiên âm Hán Việt. Chính vì vậy, để tiếp nối về bộ ngữ liệu của công trình trên, nghiên cứu của chúng tôi đã sử dụng số lượng ngữ liệu lớn hơn và có cả bản dịch nghĩa cho các bài thơ chữ Hán để làm thành một bộ ngữ liệu song song. Bên cạnh đó, chúng tôi

[2] Google Scholar , https://scholar.google.com/, truy cập gần nhất vào ngày 26/09/2023
[3] ACL Anthology, https://aclanthology.org/, truy cập gần nhất vào ngày 26/09/2023
[4] IEEE Xplore, https://ieeexplore.ieee.org/Xplore/home.jsp, truy cập gần nhất vào ngày 26/09/2023

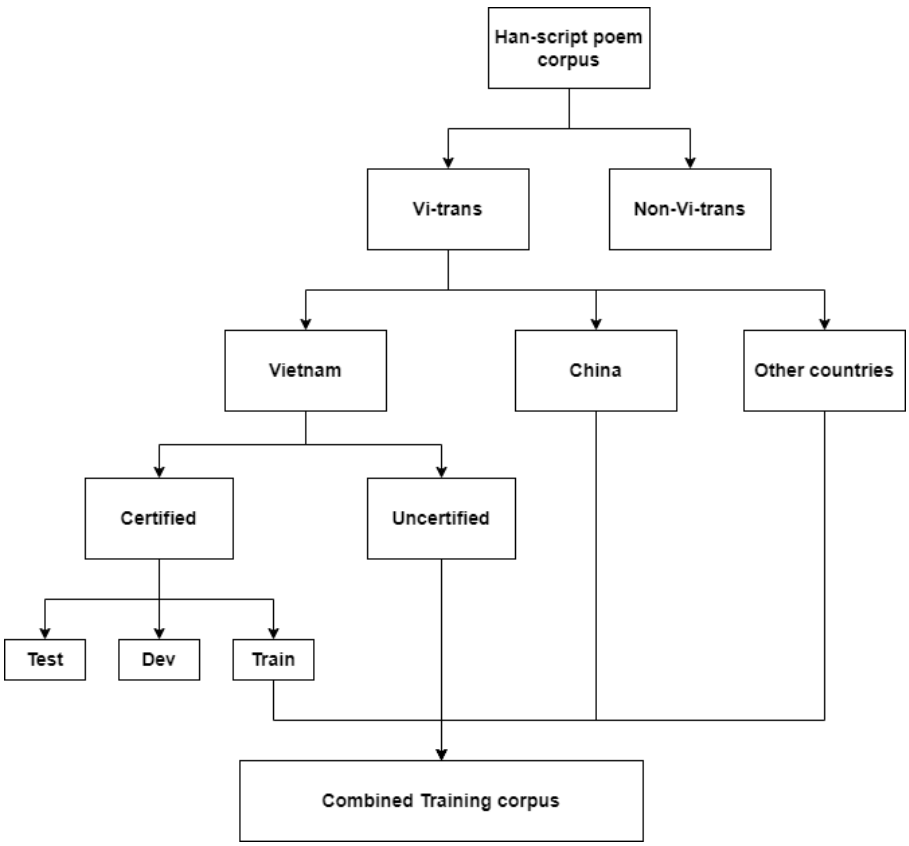
cũng tiếp tục sử dụng mô hình SMT cho nghiên cứu "Dịch nghĩa tự động thơ văn chữ Hán của Việt Nam sang tiếng Việt đương đại" của bài báo này.

III. PHƯƠNG PHÁP ĐỀ XUẤT

A. Phương pháp kết hợp kho ngữ liệu huấn luyện

Kho ngữ liệu đã xây dựng chủ yếu bao gồm thơ chữ Hán của Việt Nam và Trung Quốc có bản dịch nghĩa Tiếng Việt, ngoài ra có một lượng nhỏ các bài thơ chữ Hán của các quốc gia khác. Các bài thơ chữ Hán của Trung Quốc sẽ được sử dụng với mục đích tận dụng được các yếu tố tương đồng với thơ văn chữ Hán của Việt Nam nhằm giải quyết phần nào việc thiếu ngữ liệu huấn luyện. Trong đó, dựa trên chú thích của thivien.net cho các bài thơ chữ Hán về nguồn gốc (trong lúc thu thập chúng tôi cũng lấy thêm thông tin này), chúng tôi phân loại ra các bài thơ có trích dẫn nguồn rõ ràng và các bài thơ không có trích dẫn nguồn. Điều này có nghĩa, các bài thơ có trích dẫn nguồn là những bài thơ có độ tin cậy cao về tính chính xác của bản dịch cũng như nguyên tác. Những bài thơ chữ Hán của Việt Nam có bản dịch nghĩa tiếng Việt và được dẫn nguồn rõ ràng này sẽ được tách ra thành một tập gọi là tập ngữ liệu vàng (Golden corpus) với chủ trương rằng các tập kiểm tra sẽ sử dụng tập ngữ liệu vàng này để việc đánh giá có độ tin cậy cao hơn.

Vì để trình bày một cách ngắn gọn về cấu trúc kho ngữ liệu trong sơ đồ ở Hình 1, chúng tôi viết sơ đồ với các nhãn bằng tiếng Anh. Sơ đồ ở Hình 1 mô tả như sau: từ kho ngữ liệu thơ chữ Hán (Han-script poem corpus), chúng tôi giữ lại các bài thơ có bản dịch nghĩa tiếng Việt hiện đại (Vi-trans) để phục vụ cho kho ngữ liệu song song, còn các bài thơ chưa có bản dịch nghĩa tiếng Việt (Non-Vi-trans) sẽ đem đi lưu trữ riêng. Trong các bài thơ chữ Hán có bản dịch nghĩa tiếng Việt, chúng tôi dựa trên thông tin thu thập được từ thivien.net để phân loại ra các tập thơ bao gồm thơ văn chữ Hán của Việt Nam (Vietnam), thơ văn chữ Hán của Trung Quốc (China) và một lượng rất ít thơ văn chữ Hán của nước khác (Other countries) như Nhật Bản, Hàn Quốc,... Với các bài thơ chữ Hán của Việt Nam có bản dịch nghĩa tiếng Việt, chúng tôi tiếp tục phân ra gồm các bài thơ có nguồn trích dẫn (Certified) và bài thơ không được dẫn nguồn (Uncertified). Như trên đã trình bày, thì tập thơ có nguồn trích dẫn này được đưa vào kho ngữ liệu vàng (golden corpus). Sau cùng, chúng tôi phân chia các tập huấn luyện (Train), tập phát triển (Dev) và tập đánh giá (Test). Đối với tập Dev và tập Test, chúng tôi sẽ hoàn toàn sử dụng ngữ liệu từ tập Golden corpus để kết quả đánh giá chính xác nhất có thể - tập trung vào lĩnh vực thơ văn chữ Hán của Việt Nam. Tập ngữ liệu huấn luyện kết hợp (Combined Traning Corpus) được tạo thành từ 4 phần: khoảng 80% tập Golden Corpus, ngữ liệu thơ văn chữ Hán của Việt Nam không có nguồn trích dẫn, thơ văn chữ Hán của Trung Quốc và một ít thơ văn chữ Hán của các nước khác.



Hình 1. Phương pháp kết hợp ngữ liệu huấn luyện cho mô hình đề xuất

Lợi ích của việc phân chia các tập ngữ liệu một cách chi tiết như thế này nhằm giúp cho việc đánh giá mô hình ở các thực nghiệm đúng đối tượng văn bản chúng ta đang hướng tới là thơ văn chữ Hán của Việt Nam và kết quả sẽ có độ tin cậy cao hơn khi tập đánh giá được chọn từ tập ngữ liệu có chất lượng cao.

B. Mô hình dịch máy thống kê (Statistical Machine Translation - SMT)

Trong bài báo này, với lĩnh vực hiếm ngữ liệu song song là thơ ca chữ Hán của Việt Nam, chúng tôi đề xuất sử dụng mô hình dịch máy thống kê để đảm bảo chất lượng dịch. Các nghiên cứu trước đó về chuyển tự tự động chữ Nôm sang chữ Quốc ngữ như [1][2][9] cũng ứng dụng mô hình dịch máy thống kê trong tình huống ngữ liệu huấn luyện còn ít và cũng đã cho các kết quả khá khả quan. Vì vậy, trong nghiên cứu này, bước đầu tiếp cận với bài toán dịch nghĩa tự động thơ văn chữ Hán của Việt Nam sang tiếng Việt, chúng tôi vẫn sử dụng mô hình SMT. Đầu vào của bài toán (source language) sẽ là câu thơ/ bài thơ chữ Hán và đầu ra của bài toán (target language) là câu/bài dịch nghĩa tiếng Việt đương đại. Về đơn vị dịch là một dòng thơ, một bài thơ hay một đoạn thơ, chúng tôi sẽ thực nghiệm cả 2 trường hợp: các mẫu huấn luyện là dòng thơ hoặc bài thơ (trong một số trường hợp là đoạn thơ vì khi ấy cả bài thơ quá dài nên không thể đưa vào mô hình mà phải tách ra thành các đoạn ngắn hơn). Sở dĩ thực nghiệm với đơn vị đoạn thơ là bởi vì giữa các dòng thơ trong 1 bài thơ sẽ có mối liên hệ mật thiết với nhau.

Bài toán được phát biểu như sau: chúng ta có 1 dòng thơ/ bài thơ chữ Hán h và chúng ta cần tìm một câu dịch nghĩa tiếng Việt v sao cho xác xuất $P(v|h)$ là lớn nhất (nghĩa là câu dịch nghĩa phù hợp nhất với câu thơ chữ Hán). Theo luật Bayes, ta triển khai xác xuất $P(v|h)$ như công thức 1. [12]

$$\arg\max_v P\,(v|h) = \arg\max_v \frac{P(h|v)P(v)}{P(h)} \tag{1}$$

Từ công thức 1, ta có thể lược bỏ mẫu số $P(h)$ vì $P(h)$ không phụ thuộc vào v, ta được công thức 2.

$$\hat{v} = arg\max_v P\,(v|h) = arg\max_v P\,(h|v)P(v) \tag{2}$$

trong đó, $\hat{v}$ là bản dịch nghĩa tiếng Việt chúng ta cần tìm, $P(h|v)$ là xác suất của mô hình dịch (translation model) và $P(v)$ là xác suất của mô hình ngôn ngữ (language model) tiếng Việt [12]. Công thức 2 cũng là thể hiện của một mô hình kênh nhiễu (Noisy Channel) trong sự kết hợp giữa mô hình dịch và mô hình ngôn ngữ như trên [14]. Trong bài báo này, chúng tôi sẽ sử dụng luôn phần bản dịch nghĩa tiếng Việt của tập huấn luyện làm ngữ liệu đơn ngữ để huấn luyện mô hình ngôn ngữ.

IV. THỰC NGHIỆM VÀ KẾT QUẢ

A. Dữ liệu

Như đã trình bày ở phần "Phương pháp kết hợp kho ngữ liệu huấn luyện", tập phát triển và tập kiểm tra sẽ được lấy từ kho ngữ liệu vàng (Golden corpus) với tỉ lệ 8:1:1 (với 80% cho tập huấn luyện).

Với Tập dữ liệu "Ngữ liệu huấn luyện kết hợp" (Combined Training Corpus) sẽ là tập huấn luyện chính với sự kết hợp từ 3 tập nhỏ như đã trình bày ở Mục III.A (trong đó có bao gồm các bài thơ chữ Hán của Trung Hoa có bản dịch nghĩa tiếng Việt).

Các tập dữ liệu này đã được tiền xử lý để loại bỏ các dấu câu, HTML tag, các ký tự đặc biệt khác. Đồng thời, đối với đơn vị là bài thơ, khi này những bài thơ nào có tổng số chữ lớn hơn 80 (ở câu chữ Hán hoặc câu tiếng Việt) sẽ được tách thành các đoạn thơ nhỏ hơn để có thể đưa vào huấn luyện mô hình SMT bằng công cụ Moses [10]. Thế nên trong phần thực nghiệm này, chúng tôi tạm gọi chung đơn vị của đoạn thơ hay bài thơ là đoạn thơ. Thông tin thống kê các tập dữ liệu được thể hiện trong Bảng 2 với các đơn vị là đoạn thơ, dòng thơ, chữ Hán, chữ Việt (trong bài báo này, chúng tôi sử dụng dấu phẩy "," để phân tách phần ngàn của các con số và dấu chấm "." để phân tách phần thập phân).

Bảng 2. Thông tin các tập dữ liệu

Tập dữ liệu	Đơn vị	Huấn luyện (train)	Phát triển (Dev)	Kiểm tra (Test)
Ngữ liệu vàng (Golden corpus)	Đoạn thơ	2,509	314	315
	Dòng thơ	17,363	2,160	2,206
	Chữ Hán	115,752	14,480	14,552
	Chữ Việt	116,773	21,167	21,057
Ngữ liệu huấn luyện kết hợp (Combined Traning corpus)	Đoạn thơ	15,077		
	Dòng thơ	80,389		
	Chữ Hán	488,345		
	Chữ Việt	714,756		

B. Thiết lập và kết quả

Chúng tôi sử dụng độ đo là điểm BLEU [11] để đánh giá bài toán dịch nghĩa tự động này. Ngoài ra, chúng tôi sử dụng công cụ Moses [10] để huấn luyện mô hình SMT với các tham số cài đặt như mặc định. Và để so sánh một cách ổn định giữa những lần thực nghiệm, chúng tôi không tinh chỉnh (tuning) cho các thực nghiệm bên dưới (được trình bày trong Bảng 3). Hơn nữa, chúng tôi huấn luyện mô hình ngôn ngữ 5-gram cho toàn bộ các thực nghiệm bằng công cụ KenLM [13].

Như đã đề cập ở Mục III.B, nghiên cứu của chúng tôi sẽ huấn luyện mô hình SMT trên 2 đơn vị của tập huấn luyện là đoạn thơ và dòng thơ. Đối với đơn vị là đoạn thơ, các dòng thơ trong cùng 1 đoạn thơ được nối với nhau bằng dấu ";". Ngoài ra, để mô phỏng gần nhất với trường hợp thực tế khi đầu vào có khi là 1 đoạn thơ hay có khi là dòng thơ, chúng tôi cũng sử dụng tập phát triển (dev), tập kiểm tra (test) với cả 2 đơn vị là đoạn thơ và dòng thơ. Tuy nhiên, để tránh điểm BLEU bị ảnh hưởng bởi ký tự ";" thì sau khi đã dịch đoạn thơ đó, chúng tôi tiến hành loại bỏ ký tự này rồi tách thành dòng thơ để đánh giá với bản dịch tham khảo (reference) có đơn vị là dòng thơ tương ứng.

Đồng thời, với ngôn ngữ nguồn thay vì chỉ dịch nghĩa thẳng từ câu chữ Hán sang câu tiếng Việt, chúng tôi thử dịch từ bản phiên âm Hán Việt để so sánh kết quả với dịch thẳng từ bản chữ Hán, vì thiết nghĩ giữa bản phiên âm Hán Việt và bản tiếng Việt sẽ có nhiều từ ngữ tương đồng, đó là các yếu tố gốc Hán.

Trong Bảng 3 - kết quả thực nghiệm, cột "Đơn vị huấn luyện" biểu thị mô hình được huấn luyện với các mẫu ngữ liệu có đơn vị là đoạn thơ hay dòng thơ. Trong cột "Điểm BLEU", chúng tôi đưa ra kết quả đánh giá trên tập dữ liệu phát triển (dev), kiểm tra (test) tương ứng với đơn vị dịch là đoạn thơ và dòng thơ. Chúng tôi cũng thêm cột số thứ tự (trong Bảng 3 được viết tắt là STT) để đánh số các thực nghiệm nhằm tiện đề cập đến.

Kết quả thực nghiệm được ghi trong Bảng 3 cho thấy, đối với mô hình SMT cơ bản và bộ ngữ liệu của chúng tôi thì đơn vị huấn luyện cùng với đơn vị kiểm tra là dòng thơ sẽ cho kết quả cao hơn so với đoạn thơ. Điều này có lẽ là do, khi ta đưa vào mô hình SMT 1 đoạn thơ dài, mô hình sẽ khó khăn trong việc học và dịch hơn so với các dòng thơ ngắn.

Bảng 3. Kết quả thực nghiệm

STT	Ngôn ngữ nguồn	Đơn vị huấn luyện	Điểm BLEU			
			Tập phát triển (Dev)		Tập kiểm tra (Test)	
			Đoạn thơ	Dòng thơ	Đoạn thơ	Dòng thơ
1	Hán	Đoạn thơ	8.26 52.4/19.3/6.6/2.8	8.34 52.8/19.0/6.3/2.7	8.10 52.7/19.2/6.3/2.5	8.47 53.3/19.3/6.4/2.6
2		Dòng thơ	**8.28** **52.9/19.0/6.4/2.8**	**8.48** **52.4/19.0/6.3/2.6**	**8.52** **53.2/19.4/6.6/2.7**	**8.61** **52.3/19.1/6.4/2.5**
3	Hán Việt	Đoạn thơ	7.53 50.0/18.4/6.1/2.7	7.63 50.5/18.2/6.0/2.6	7.37 50.4/18.4/6.1/2.4	7.40 50.6/18.1/5.8/2.2
4		Dòng thơ	7.54 50.5/17.9/5.9/2.7	7.80 50.0/18.0/6.0/2.6	7.38 50.5/18.0/5.7/2.3	7.76 49.8/18.0/5.9/2.4

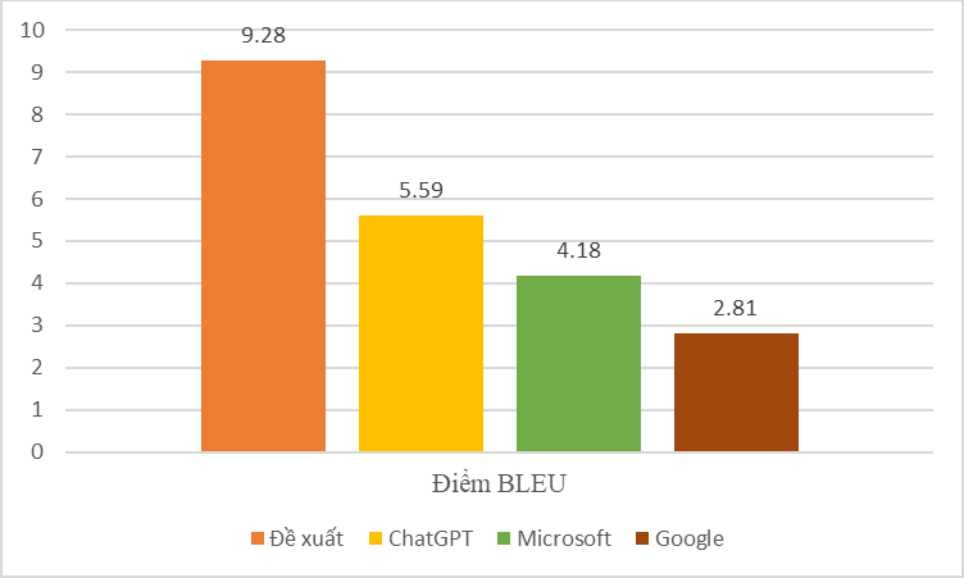
Ngoài ra, dựa trên bảng kết quả thực nghiệm, ta thấy rằng dịch nghĩa thẳng từ Hán sang Việt (thực nghiệm 1,2) sẽ cho kết quả cao hơn khi dịch từ bản phiên âm Hán - Việt (thực nghiệm 3,4). Bởi lẽ, 1 âm Hán Việt có thể tương ứng với nhiều chữ Hán, trong khi các chữ Hán ấy có thể lại có nghĩa khác nhau (ở đây chúng ta khoan đề cập về vấn đề dị thể của chữ Hán). Vì vậy, tốt nhất chúng ta nên dịch nghĩa thẳng từ câu chữ Hán hoặc tìm cách kết hợp sao cho cùng đưa vào mô hình cả bản chữ Hán và bản phiên âm Hán Việt.

C. So sánh với các công cụ có khả năng dịch chữ Hán khác

Từ kết quả thực nghiệm ở phần trên, chúng tôi thêm một bước là tinh chỉnh (tuning) mô hình tốt nhất của thực nghiệm 2 trên tập phát triển với kỹ thuật BatchMIRA [15] để mô hình đạt kết quả tối ưu nhất có thể và sau đó đem mô hình này đánh giá trên tập 50 bài thơ được trích ra từ tập đánh giá của kho ngữ liệu vàng nhằm so sánh với 3 công cụ là ChatGPT[5] và Microsoft Translator (sử dụng công cụ Translator được tích hợp trong Microsoft Excel) và Google Dịch[6]; bản dịch của ChatGPT và Microsoft Translator được lấy vào ngày 01/07/2023 và của Google Dịch được lấy vào ngày 26/09/2023. Với ChatGPT, chúng tôi đặt câu hỏi như sau: "Please translate meanings of below poem to Vietnamese with only Vietnamese output!" kết hợp với nhập vào bài thơ chữ Hán cần dịch để ChatGPT đưa ra câu trả lời là bản dịch nghĩa cho đoạn thơ chữ Hán đó. Đối với Microsoft Translator, chúng tôi chọn dịch từ Chinese (Literary) sang Vietnamese, bởi vì ngôn ngữ nguồn là văn chương cổ của Trung Hoa sẽ gần nhất với nghiên cứu của chúng tôi (như đã đề cập ở Mục II. Các công trình liên quan). Còn với Google Dịch, chúng tôi đã chọn dịch từ tiếng Trung (Phồn thể) sang tiếng Việt.

[5] OpenAI, https://chat.openai.com/, truy cập ngày 01/07/2023
[6] Google, Google Dịch, https://translate.google.com/, truy cập ngày 26/09/2023

Kết quả được thể hiện ở biểu đồ cột trong Hình 2 cho thấy, mô hình đề xuất của chúng tôi đạt được kết quả tốt hơn 3 công cụ còn lại. Đây là một tín hiệu khả quan cho việc giải quyết bài toán này với hướng tiếp cận của chúng tôi - tập trung vào lĩnh vực thơ văn chữ Hán của Việt Nam.



Hình 2. Biểu đồ so sánh điểm BLEU của mô hình đề xuất và 3 công cụ khác trên tập kiểm tra gồm 50 bài thơ chữ Hán của Việt Nam

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Với kho ngữ liệu song song đủ tốt và giải bài toán dịch nghĩa tự động thơ văn chữ Hán của Việt Nam sang tiếng Việt với mô hình SMT đã mang đến kết quả bước đầu khá tốt. Đây sẽ là động lực để chúng ta tiếp tục phát triển các mô hình dịch cho bài toán này. Ngoài ra, từ kết quả thực nghiệm chúng tôi còn rút ra được rằng, dịch nghĩa thẳng từ bản chữ Hán ra tiếng Việt đương đại sẽ cho kết quả tốt hơn khi dịch từ bản phiên âm Hán Việt, đồng thời với mô hình SMT thì đơn vị dịch là dòng thơ sẽ tốt hơn so với bài thơ.

Trong tương lai, chắc chắn chúng tôi sẽ tiếp tục cải tiến bài toán này với các hướng tiếp cận như kết hợp cả bản chữ Hán và bản phiên âm Hán Việt cho đầu vào mô hình dịch, sử dụng mô hình NMT, tích hợp thêm các tri thức ngôn ngữ,…

TÀI LIỆU THAM KHẢO

[1] Dinh, Dien, Phuong Nguyen, and Long HB Nguyen, "Transliterating Nôm Scripts into Vietnamese National Scripts using Statistical Machine Translation," *International Journal of Advanced Computer Science and Applications*, 12.2 (2021).

[2] Thai, Lam H., Long HB Nguyen, and Dien Dinh, "Improve the automatic transliteration from Nôm scripts into Vietnamese National scripts by integrating Sino-Vietnamese knowledge into Statistical Machine Translation," *Proceedings of the 4th International Conference on Information Technology and Computer Communications*, 2022.

[3] Dang, Hoang-Quan, et al., "NomNaOCR: The First Dataset for Optical Character Recognition on Han-Nom Script," *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 2022.

[4] Hoàng Lê, Mai Hải, Linh Quế, Viện nghiên cứu Hán Nôm, *Dịch từ Hán sang Việt: một khoa học - một nghệ thuật*. Nhà xuất bản Khoa học Xã hội, 1982.

[5] Lê, Thị Huyền Trang, "Chuyển dịch một số bài thơ Nôm sang chữ Hán: Cảm nhận và chia sẻ," 2019.

[6] Hữu Đạt, *Ngôn ngữ thơ Việt Nam.* Nhà xuất bản Giáo dục, 1996.

[7] Koehn, Philipp, and Rebecca Knowles, "Six challenges for neural machine translation," *arXiv preprint arXiv:1706.03872*, 2017.

[8] Microsoft Translator, "Microsoft Translator releases literary Chinese translation," 2021, https://www.microsoft.com/en-us/translator/blog/2021/08/25/microsoft-translator-releases-literary-chinese-translation/ (Truy cập 20/8/2023).

[9] Vo, Tin N., Long Nguyen, and Dien Dinh, "Improving Transliteration from Nôm Scripts into Vietnamese National Scripts using Class-based Language Model," *Proceedings of the 2023 9th International Conference on Computer Technology Applications*, 2023.

[10] Koehn, Philipp, et al., "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.

[11] Papineni, Kishore, et al., "Bleu: a method for automatic evaluation of machine translation," *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002.

[12] Đinh Điền, *Xử lý ngôn ngữ tự nhiên*. Nhà xuất bản Đại học Quốc gia TP. Hồ Chí Minh, 2006.

[13] Heafield, Kenneth, "KenLM: Faster and smaller language model queries," *Proceedings of the sixth workshop on statistical machine translation*, 2011.

[14] Koehn, Philipp, *Statistical machine translation*. Cambridge University Press, 2009.

[15] Cherry, Colin, and George Foster, "Batch tuning strategies for statistical machine translation," *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2012.

AUTOMATIC INTERPRETATION FROM VIETNAM'S HAN-SCRIPT LITERARY WORKS INTO CONTEMPORARY VIETNAMESE USING STATISTICAL MACHINE TRANSLATION

Thai Hoang Lam, Dinh Dien

ABSTRACT: *Interpretation from Vietnam's Han-script works into contemporary Vietnamese is an essential step in the process of mining values from Han-Nom heritage, especially for Vietnam's literary works written in Han script. With the booming of Artificial Intelligence, we suppose that if we can construct a good automatic interpretation model for this text domain, it will become a helpful assistant for Han-Nom researchers in the mission of exploring ancient values as well as a useful tool for Vietnamese people that are not researchers of Han-Nom or related fields when they come to visit the temple, historic places,…However, to our best knowledge, there is no automatic translation tool that pays attention to translating the meanings of Vietnam's Han-script literary works or some tools only help to transcribe Han script to Sino-Vietnamese. Therefore, in this paper, we start off solving the mentioned problem - automatic interpretation from Vietnam's Han-script literary works into contemporary Vietnamese by using Statistical Machine Translation. The experimental results show that our research made progress when proposing a Vietnam's Han-script literary works interpretation model which is better than other tools that allow to input Han-script for translation. This outcome will encourage the next research on automatic interpretation towards Vietnam's Han-script literary works domain.*

Keywords: *Automatic interpretation, Han-script literary works, Vietnamese, Statistical machine translation.*