DỊCH MÁY HAI CHIỀU THÁI – VIỆT CHO CHỮ THÁI BẰNG MÔ HÌNH NLLB MỞ RỘNG TỪ VỰNG KHỞI TẠO TỪ TỪ ĐIỂN SONG NGỮ

Ngô Hoàng Long[1], Phạm Mạnh Phong[1], Giáp Bình Nga[2]
[1] Trường THPT Chu Văn An, Hà Nội
[2] Trường Đại học Sư phạm Hà Nội
Email: ngohoanglong052010@gmail.com

Tóm tắt: *Nghiên cứu đề xuất phương pháp dịch máy hai chiều Thái – Việt dựa trên mô hình NLLB-600M, khắc phục hạn chế không nhận diện được chữ Thái (Tai Viet). Chúng tôi mở rộng bộ từ vựng bằng 5.536 token theo âm tiết, khởi tạo vectơ nhúng từ từ điển Thái – Việt và tinh chỉnh bằng LoRA. Với 16.924 cặp huấn luyện, phương pháp khởi tạo từ từ điển nâng chrF++ từ 11,4 lên 34,9. Mô hình Thái → Việt đạt 56,3 chrF++ trên tập phát triển và 41,7 trên tập kiểm tra độc lập 445 câu; chiều Việt → Thái đạt 46,2, so với 28,6 của phương pháp tra từ điển. Hai mô hình được huấn luyện trên một GPU phổ thông và triển khai thành dịch vụ trực tuyến.*

Từ khoá: *chữ Thái (Tai Viet), dịch máy, mở rộng từ vựng, NLLB, ngôn ngữ ít tài nguyên, tiếng Thái Đen.*

BIDIRECTIONAL TAI VIET – VIETNAMESE MACHINE TRANSLATION WITH DICTIONARY-INITIALISED VOCABULARY EXTENSION OF NLLB

Abstract: *This study proposes a bidirectional Tai Viet–Vietnamese machine translation method based on NLLB-600M, addressing the model's inability to recognize the Tai Viet script. We extend its vocabulary with 5,536 syllable-level tokens, initialize their embeddings using a Tai–Vietnamese dictionary, and fine-tune the model with LoRA. Using 16,924 training pairs, dictionary initialization improves chrF++ from 11.4 to 34.9. The Tai → Vietnamese model achieves 56.3 chrF++ on the development set and 41.7 on an independent 445-sentence test set. The Vietnamese → Tai model scores 46.2, compared with 28.6 for dictionary lookup. Both models are trained on a single consumer GPU and deployed as an online translation service.*

Keywords: *Black Tai, low-resource languages, machine translation, NLLB, Tai Viet script, vocabulary extension.*

Nhận bài: 28/09/2026 *Phản biện: 28/10/2026* *Duyệt đăng: 01/11/2026*

I. ĐẶT VẤN ĐỀ

Các mô hình dịch đa ngữ như NLLB-200 (NLLB Team, 2022) hỗ trợ khoảng 200 ngôn ngữ, nhưng khả năng dịch phụ thuộc vào việc nhận diện chữ viết của ngôn ngữ nguồn. Tiếng Thái Đen (Tai Dam, mã ISO 639-3: blt) có hệ chữ Thái (Tai Viet), được mã hóa trong Unicode 5.2 thuộc khối U+AA80–U+AADF (The Unicode Consortium, 2009). Tại Việt Nam, dân tộc Thái có khoảng 1,8 triệu người, gồm cả Thái Đen và Thái Trắng; số liệu này không phản ánh số người sử dụng chữ Thái (Tổng cục Thống kê, 2020). Nhiều văn bản chữ Thái còn được lưu bằng phông chữ cũ sử dụng mã ASCII, gây khó khăn cho các công cụ xử lý Unicode.

Khảo sát cho thấy bộ từ vựng khoảng 256 nghìn token của NLLB không chứa ký tự thuộc khối chữ Thái và không hỗ trợ phân rã byte. Do đó, toàn bộ âm tiết chữ Thái bị chuyển thành token không xác định <unk>, khiến mô hình không nhận diện được đầu vào. Theo phạm vi khảo sát, chưa có hệ thống dịch máy nơ-ron được công bố hỗ trợ trực tiếp chữ Thái Đen.

Nghiên cứu đề xuất mở rộng từ vựng NLLB-600M theo âm tiết, khởi tạo vectơ nhúng từ nghĩa tiếng Việt trong từ điển song ngữ và tinh chỉnh bằng LoRA (Hu et al., 2022). Phương pháp kế thừa các nghiên cứu về khởi tạo vectơ nhúng cho token mới (Hewitt, 2021; Minixhofer et al., 2022; Dobler & de Melo, 2023), đồng thời giải quyết trường hợp chữ viết mới không có token con chung với mô hình gốc.

Các đóng góp chính gồm phân tích sự thất bại của dịch zero-shot; đề xuất phương pháp mở rộng từ vựng và kiểm chứng hiệu quả từng thành phần; xây dựng khoảng 16,9 nghìn cặp song ngữ, trong đó khoảng 5,9 nghìn câu và đoạn văn do người bản ngữ dịch; đánh giá độc lập trên 445 câu theo độ dài và hai chiều dịch; đồng thời triển khai hệ thống thành dịch vụ trực tuyến.

II. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Dữ liệu

Bảng 1 tóm tắt dữ liệu dùng cho các mô hình cuối cùng. Toàn bộ văn bản tiếng Thái được chuẩn hoá Unicode NFC và sắp xếp lại các dấu kết hợp theo một quy tắc chính tắc (mục 2.4).

Bảng 1. Dữ liệu dùng cho các mô hình cuối (tổng dữ liệu huấn luyện 16.924 cặp)

Nguồn	Số cặp	Mục đích
Cụm từ ví dụ trong từ điển Thái – Việt	10.576	huấn luyện
Câu đời thường do người bản ngữ dịch (phiếu 1–3)	3.264	huấn luyện
Câu dài và đoạn văn do người bản ngữ dịch (đợt 5)	2.261 + 239	huấn luyện
Đoạn văn do người bản ngữ dịch, 12 lĩnh vực	115	huấn luyện
Văn bản song ngữ có sẵn (sử thi, diễn văn, bài văn)	469	huấn luyện
Câu đời thường (tập phát triển câu)	226	phát triển
Đoạn văn giữ lại (tập phát triển đoạn)	30	phát triển
Giáo trình tiếng Thái, phân tầng theo độ dài (tập kiểm tra)	445	kiểm tra

Ghi chú: Số liệu đợt 5 không gồm 22 dòng đang trả lại người dịch để sửa.

Từ điển song ngữ. Từ điển Thái – Việt gồm 13.428 dòng nghĩa trên 9.610 mục từ, đóng hai vai trò: các trường ví dụ cho 10.576 cặp cụm từ Thái – Việt (trung vị 4 âm tiết)các nghĩa tiếng Việt được dùng để khởi tạo vectơ nhúng. Vì từ điển đã dùng để khởi tạo mô hình, các ví dụ của nó chỉ dùng để huấn luyện, không bao giờ dùng để đánh giá. 8,5% dòng nghĩa không có nghĩa tiếng Việt và bị loại.

Bản dịch của người bản ngữ. Câu nguồn tiếng Việt được soạn thành các phiếu dịch theo chủ đề, sau đó theo lĩnh vực và dải độ dài và được người bản ngữ dịch sang chữ Thái. Các phiếu đầu tiên chủ yếu là câu ngắn; sau khi đo được chất lượng giảm mạnh theo độ dài câu, đợt 5 gồm 5.000 câu nguồn được thiết kế riêng để lấp các dải 11–35 âm tiết và đoạn văn, trong đó 2.500 câu đã được dịch và sử dụng tại thời điểm huấn luyện. Phiếu đoạn văn có hướng dẫn riêng: dịch cả đoạn như một văn bản liền mạch, giữ nhất quán cách xưng hô, giữ ranh giới câu; 91% đoạn văn giữ đúng số câu của bản gốc. Vì câu nguồn tiếng Việt được soạn cho mục đích dịch, dữ liệu chỉ dùng để huấn luyện hoặc phát triển, không dùng để kiểm tra.

Kiểm soát chất lượng. Mỗi phiếu nhận về được kiểm tra tự động trước khi sử dụng: chuẩn hoá Unicode, chữ Latin lẫn trong văn bản chữ Thái (thường là phím dấu của kiểu gõ Telex tiếng Việt bị gõ nhầm), dấu kết hợp bị lặp, cùng một câu tiếng Thái cho hai câu tiếng Việt khác nhautỉ lệ độ dài bất thường. Các dòng bị gắn cờ được trả lại người dịch để sửa, nhóm nghiên cứu không tự sửa. Tỉ lệ số âm tiết Thái/Việt có trung vị 1,03–1,11 tuỳ phiếuđộ dài theo từng dòng tương quan ở mức $r \approx 0{,}98$, nhờ đó phát hiện được cả lỗi lệch dòng.

2.2. Khôi phục văn bản gõ bằng phông chữ cũ

Tập kiểm tra được trích từ giáo trình tiếng Thái lưu dưới dạng PDF, sử dụng phông chữ cũ mã hóa bằng byte ASCII. Chúng tôi khôi phục bảng ánh xạ byte → Unicode bằng cách đối chiếu hình dạng ký tự với phông chữ Thái Unicode và kiểm tra bằng danh sách âm tiết trong từ điển. Do chữ hoa/chữ thường biểu diễn các cặp phụ âm cao/thấp và dấu kết hợp được gõ trước phụ âm gốc, cần sắp xếp lại các dấu theo thứ tự chính tắc. Bước này nâng tỷ lệ âm tiết giải mã có trong từ điển từ 76,7% lên 96,1%. Bộ giải mã cuối đạt 97,6% token thuộc danh sách âm tiết trên văn bản dùng để tinh chỉnh và 99,1% trên văn bản độc lập; kết quả được người bản ngữ kiểm tra và phát hiện một nét chữ ghép bị ánh xạ sai.

2.3. Tập đánh giá

Tập kiểm tra từ giáo trình. Chúng tôi chọn 493 câu từ giáo trình đã giải mã, khoảng 100 câu cho mỗi dải độ dài tiếng Thái (1–5, 6–10, 11–20, 21–35, 36–80 âm tiết; dải cuối có 93 câu), loại các câu có lỗi giải mã. Ba người bản ngữ dịch các câu này sang tiếng Việt. Quá trình kiểm tra phát hiện, cùng các lỗi khác, một số bản dịch bị dán nhầm dòng; sau khi sửa và loại bỏ, còn 445 câu dùng được. Câu kiểm tra là văn bản tiếng Thái gốc, không phải bản dịch từ tiếng Việt, nên chiều Thái → Việt được đánh giá trên văn bản nguồn thậtchiều Việt → Thái được chấm với tham chiếu tiếng Thái gốc. Toàn bộ cuốn sách được giữ riêng: mọi mục dữ liệu huấn luyện có chung một chuỗi 8 âm tiết với sách đều bị loại. Tập kiểm tra chỉ dùng để báo cáo các mô hình đã được chọn trên tập phát triển, không dùng để chọn mô hình hay tinh chỉnh.

Tập phát triển. Tập phát triển câu gồm 226 câu đời thường (trung vị 8 âm tiết); tập phát triển đoạn gồm 30 đoạn văn giữ lại từ phiếu đoạn văn, cân đối theo lĩnh vực. Việc chọn mô hình và dừng sớm chỉ dựa vào tập phát triển câu.

2.4. Chuẩn hoá Unicode

Các ký tự chữ Thái không có phân rã chính

tắc; 8/9 dấu kết hợp có cùng lớp kết hợp 230, nên chuẩn hóa NFC không tự sắp xếp lại thứ tự dấu thanh và dấu nguyên âm. Điều này khiến những chuỗi hiển thị giống nhau nhưng khác mã ký tự, ảnh hưởng đến huấn luyện và tra từ điển. Chúng tôi áp dụng quy tắc sắp xếp dấu theo lớp kết hợp và mã ký tự, thống nhất trong huấn luyện, đánh giá và dịch vụ triển khai.

2.5. Mở rộng từ vựng theo âm tiết

Văn bản chữ Thái tách âm tiết bằng dấu cách, nên chúng tôi thêm mỗi loại âm tiết thành một token: 4.955 âm tiết từ từ điển và kho ngữ liệu ban đầu, 520 âm tiết chỉ xuất hiện trong dữ liệu huấn luyệncác ký tự đơn của khối làm phương án dự phòng, tổng cộng 5.536 token mới (bộ từ vựng tăng từ 256.204 lên 261.740). Token được thêm với chế độ hấp thụ dấu cách đứng trước, để mỗi âm tiết chiếm đúng một vị trí đầu vào. Trên tập phát triển câu, cách làm này loại bỏ hoàn toàn token <unk> (trước đó chiếm 47% số token) và giảm độ dài từ 2,13 xuống 1,13 token/âm tiết. Việc tách token được kiểm tra bằng cách so chuỗi token với chuỗi âm tiết nguồn, không so qua chuỗi văn bản giải mã lại vốn làm mất dấu cách giữa các âm tiết.

2.6. Khởi tạo vectơ nhúng từ từ điển

Với mỗi âm tiết mới s, chúng tôi lấy các nghĩa tiếng Việt từ từ điển. Nghĩa trực tiếp lấy từ các mục từ chỉ gồm s; nếu không có, nghĩa gián tiếp lấy từ các mục từ nhiều âm tiết có chứa s. Mỗi nghĩa được tách token bằng chính bộ tách token của NLLB và biểu diễn bằng trung bình các vectơ nhúng của nó; vectơ của âm tiết là trung bình trên tất cả các nghĩa (tức là lấy trung bình mọi nghĩa của từ đa nghĩa). Âm tiết không có nghĩa nào nhận vectơ trung bình của các vectơ đã khởi tạo cộng một nhiễu Gauss nhỏ. Mọi vectơ được chuẩn hoá về độ dài trung vị của các vectơ nhúng gốc. Trong mô hình cuối, 2.959 token được khởi tạo từ nghĩa trực tiếp, 1.175 từ nghĩa gián tiếp và 1.402 từ vectơ trung bình.

2.7. Huấn luyện

Chúng tôi tinh chỉnh mô hình nllb-200-distilled-600M bằng LoRA (Hu et al., 2022), với r = 16, α = 32, dropout 0,05, áp dụng trên các phép chiếu chú ý và lớp truyền thẳng của bộ mã hóa, giải mã. Ma trận vectơ nhúng dùng chung với lớp chiếu đầu ra được huấn luyện có chọn lọc: gradient của các hàng gốc được đặt bằng 0 qua hàm móc, đồng thời hệ số suy giảm trọng số bằng 0. Kiểm tra sau huấn luyện xác nhận các vectơ nhúng gốc không thay đổi.

Ở chiều Thái → Việt, token chữ Thái bị chặn khi sinh câu và loại khỏi hàm mất mát. Các siêu tham số gồm tốc độ học 2·10-4, khởi động 5%, giảm cosin, làm trơn nhãn 0,1, lô hiệu dụng 32 (8 × 4 bước tích lũy), cắt gradient 1,0, độ chính xác bf16, tối đa 20 epoch, dừng sớm theo chrF++ trên tập phát triển câu với kiên nhẫn 5 epoch và giải mã chùm độ rộng 4; ngôn ngữ đích cố định là `vie_Latn`.

Ở chiều Việt → Thái, token đích blt_Tavt được khởi tạo từ trung bình vectơ nhúng tha_Thai và lao_Laoo. Không gian đầu ra giới hạn ở 5.289 token phía đích, dấu câu và chữ số; loại các token gốc chứa chữ cái để tránh sinh chữ Latin. Dấu cách giữa các âm tiết được khôi phục theo quy tắc xác định và áp dụng thống nhất khi đánh giá.

2.8. Đánh giá

Chúng tôi đánh giá chất lượng dịch bằng chrF++ ở mức toàn tập (Popović, 2015, 2017), tính bằng sacreBLEU (Post, 2018). Điểm hai chiều dịch không so sánh trực tiếp được do khác biệt chữ viết và cách tách từ. Số bản dịch khác nhau cũng được thống kê để phát hiện suy biến đầu ra. Hai lần huấn luyện cùng cấu hình đạt 49,65 và 48,82 chrF++ trên tập phát triển câu, cho thấy dao động khoảng ±1 điểm; chênh lệch dưới 1,5 điểm được xem là chưa phân định.

III. KẾT QUẢ VÀ BÀN LUẬN

3.1. Hành vi zero-shot

Trên 57 câu thuộc tập phát triển, mô hình NLLB-600M chưa chỉnh sửa chỉ đạt 4,2 chrF++ khi dịch zero-shot. Kết quả gần như không thay đổi với các mã ngôn ngữ nguồn tha_Thai, lao_Laoo và vie_Latn. Mô hình chỉ tạo ra 8 bản dịch khác nhau, trong đó 17/57 câu có đầu ra phổ biến nhất là chuỗi dấu cách và dấu chấm. Hiện tượng này cho thấy mô hình không nhận diện được chữ Thái trong đầu vào, dẫn đến suy biến kết quả dịch. Vì vậy, mở rộng từ vựng là điều kiện cần thiết trước khi tinh chỉnh mô hình.

3.2. Thí nghiệm loại bỏ thành phần

Bảng 2 tách riêng tác dụng của hai thành phần, khi chỉ huấn luyện trên các cặp từ điển và đánh giá trên tập phát triển câu. Không mở rộng từ vựng, mô hình học được phân bố câu đầu ra nhưng không dịch: 7,5 chrF++ với chỉ 13 bản dịch khác nhau cho 226 câu. Mở rộng từ vựng với khởi tạo ngẫu nhiên (cùng độ dài vectơ) đạt 11,4; khởi tạo từ từ điển đạt 34,9, với bản dịch khác nhau cho mọi câu.

Bảng 2. Thí nghiệm loại bỏ thành phần trên tập phát triển câu, chỉ huấn luyện trên 10.576 cặp từ điển

Cấu hình	chrF++	Số bản dịch khác nhau
NLLB zero-shot (57 câu)	4,2	8 / 57
LoRA, không mở rộng từ vựng	7,5	13 / 226
+ mở rộng từ vựng, khởi tạo ngẫu nhiên	11,4	–
+ mở rộng từ vựng, khởi tạo từ từ điển	34,9	226 / 226

Ghi chú: Mỗi cấu hình chạy một lần; lần chạy lại độc lập của cấu hình đầy đủ chênh 0,19 điểm.

Chúng tôi đặt giả thuyết vectơ nhúng khởi tạo ngẫu nhiên sẽ phải dịch chuyển xa hơn nhiều trong quá trình huấn luyện. Thực tế kết quả chứng minh không phải vậy: độ dịch chuyển trung vị là 0,19 so với 0,17, khoảng 3% độ dài vectơ ở cả hai trường hợp. Trong huấn luyện thực tế, vectơ nhúng gần như không di chuyển, nên điểm xuất phát quyết định phần lớn điểm kết thúc; một điểm xuất phát kém không thể bù đắp chỉ bằng huấn luyện. Liệu một lịch huấn luyện dài hơn nhiều có giúp khởi tạo ngẫu nhiên bắt kịp hay không vẫn cần được kiểm chứng.

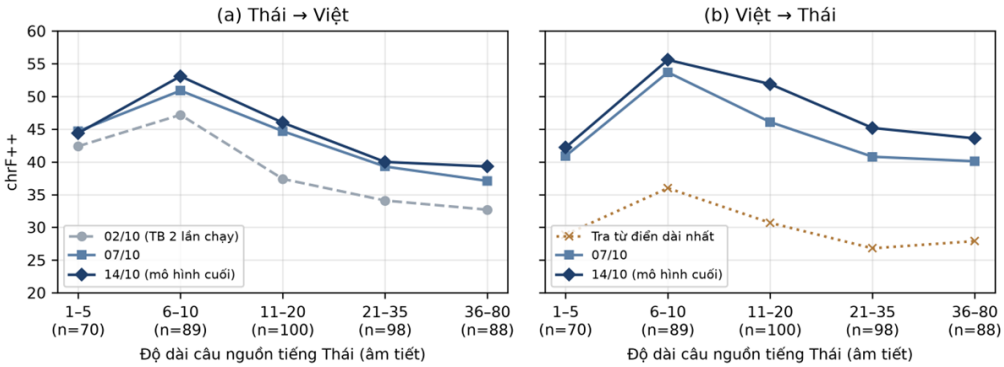
3.3. Vai trò của dữ liệu do người bản ngữ dịch

Bổ sung dữ liệu do người bản ngữ dịch giúp cải thiện chất lượng Thái → Việt. Mô hình chỉ sử dụng 10.576 cặp từ điển đạt 34,9 chrF++ trên tập phát triển câu; thêm 996 câu và 115 đoạn văn nâng điểm lên 42,9 trên tập phát triển câu và 42,2 trên tập phát triển đoạn. Với 3.263 câu, kết quả tương ứng đạt 47,1 và 46,1; cấu hình gồm 3.731 câu và 120 đoạn văn đạt trung bình 49,2 và 46,3 qua hai lần huấn luyện.

Khi bổ sung tiếp 982 câu dài và đoạn văn, điểm tăng lên 54,0 và 50,6; với 1.503 câu dài và đoạn văn, đạt 56,3 và 53,4. Kết quả cho thấy giá trị của dữ liệu câu và đoạn văn hoàn chỉnh đối với chất lượng dịch. Tuy nhiên, ba cấu hình đầu còn khác biệt về quy trình huấn luyện, GPU và độ chính xác số; chỉ ba cấu hình cuối cho phép xem xét riêng tác động của dữ liệu.

3.4. Chất lượng theo độ dài câu

Tập phát triển chưa phản ánh đầy đủ ảnh hưởng của độ dài vì 200/256 mục dài 6–10 âm tiết. Tập kiểm tra độc lập được phân tầng theo năm dải độ dài (Hình 1 và Bảng 3). Ở chiều Thái → Việt, chất lượng cao nhất tại dải 6–10 âm tiết (53,1 chrF++) và giảm xuống 39,3 ở dải 36–80 âm tiết. Chênh lệch giữa tập phát triển (56,3) và tập kiểm tra (41,7) chủ yếu liên quan đến độ dài câu trung bình tương ứng 8 và 20,8 âm tiết. Việc bổ sung có chủ đích các câu dài đã nâng kết quả ở các dải 11–80 âm tiết thêm 6–9 điểm so với mô hình trước đó.



Hình 1. chrF++ trên tập kiểm tra theo độ dài câu nguồn tiếng Thái: (a) các mô hình Thái → Việt kế tiếp nhau; (b) Việt → Thái so với phương pháp tra từ điển. Mỗi dải có 70–100 câu

Bảng 3. chrF++ trên tập kiểm tra (445 câu) theo độ dài câu tiếng Thái (âm tiết); số câu trong ngoặc. Điểm của hai chiều dịch không so sánh được với nhau

Mô hình	1–5 (70)	6–10 (89)	11–20 (100)	21–35 (98)	36–80 (88)	Toàn tập (445)
Thái → Việt 02/10 (TB 2 lần chạy)	42,4	47,2	37,4	34,1	32,7	35,1
07/10	44,7	50,9	44,7	39,3	37,1	40,1
14/10 (mô hình cuối)	44,4	53,1	46,0	40,0	39,3	41,7

Việt → Thái Tra từ điển dài nhất	28,9	36,0	30,7	26,8	27,9	28,6
07/10	40,9	53,7	46,1	40,8	40,1	42,2
14/10 (mô hình cuối)	42,2	55,6	51,9	45,2	43,6	46,2

3.5. Chiều Việt → Thái

Phương pháp cơ sở dịch bằng cách duyệt câu tiếng Việt từ trái sang phải, chọn cụm dài nhất trong 14.223 mục từ điển, thay bằng biểu thức tiếng Thái tương ứng và bỏ qua từ không tra được. Trên tập kiểm tra 445 câu, mô hình Việt → Thái đạt 46,2 chrF++, cao hơn phương pháp tra từ điển (28,6) 17,6 điểm và vượt 13–21 điểm ở các dải độ dài (Bảng 3). Trên tập phát triển, mô hình đạt 50,4 đối với câu và 51,0 đối với đoạn, so với 32,1 và 30,2 của phương pháp cơ sở. Không bản dịch nào trong 445 câu kiểm tra chứa chữ Latin. Bổ sung dữ liệu câu dài giúp tăng 4,0 điểm ở chiều Việt → Thái, so với 1,6 điểm ở chiều Thái → Việt, cho thấy mô hình học được trật tự từ và cách diễn đạt thay vì chỉ thay thế từ.

3.6. Bổ sung danh sách từ vựng theo chủ đề không cải thiện mô hình

Nhóm nghiên cứu bổ sung 1.552 từ tiếng Thái thuộc 11 chủ đề, tạo thêm 1.496 cặp từ điển (10.576 → 12.072 cặp) và 33 âm tiết mới, sau đó huấn luyện lại hai chiều dịch với cùng cấu hình.

Mô hình Thái → Việt đạt 55,2 và 54,9 chrF++ trên hai tập phát triển (so với 56,3 và 53,4 trước khi bổ sung), trong khi điểm kiểm tra tăng từ 41,7 lên 42,1. Chiều Việt → Thái giảm từ 46,2 xuống 45,8; ngược lại, phương pháp tra từ điển tăng từ 28,6 lên 31,4. Các chênh lệch của mô hình nơ-ron đều nằm trong phạm vi dao động giữa các lần huấn luyện. Kết quả cho thấy bổ sung từ đơn lẻ chưa cải thiện rõ rệt mô hình nơ-ron khi đã có dữ liệu câu do người bản ngữ dịch, nhưng vẫn có ích đối với phương pháp tra từ điển.

3.7. Văn bản thơ ca: một phép thử ngoài miền

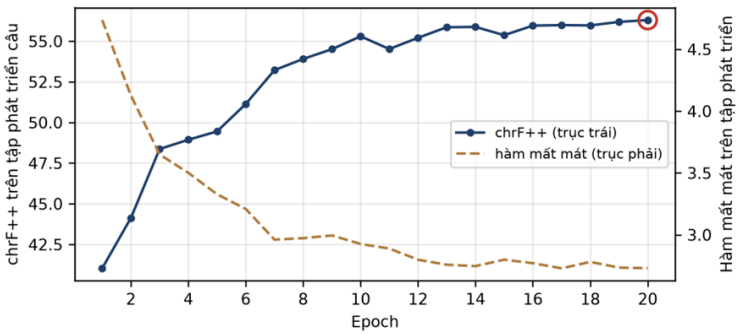
Nhóm nghiên cứu sử dụng 466 câu thơ trong sử thi Xống chụ xon xao có bản dịch tiếng Việt làm tập đánh giá ngoài miền. Do bản dịch mang tính thơ tự do, văn bản được loại khỏi dữ liệu huấn luyện và hai mô hình được huấn luyện lại.

Kết quả cho thấy mô hình chưa tiếp xúc với bài thơ đạt 19,5 chrF++ (Thái → Việt) và 18,5 (Việt → Thái), trong khi phương pháp tra từ điển đạt 16,5. Các mô hình đã sử dụng bài thơ để huấn luyện đạt lần lượt 30,4 và 29,3, cao hơn khoảng 11 điểm. Chênh lệch này phản ánh ảnh hưởng của việc ghi nhớ dữ liệu, đồng thời cho thấy hạn chế trong dịch văn vần cổ và sự cần thiết phải tách dữ liệu huấn luyện, đánh giá theo văn bản nguồn.

3.8. Quá trình huấn luyện và chi phí

Kết quả huấn luyện mô hình Thái → Việt cuối cùng (Hình 2) cho thấy chrF++ trên tập phát triển tăng nhanh trong 7 epoch đầu và tiếp tục tăng nhẹ đến epoch cuối; điều kiện dừng sớm không được kích hoạt. Hàm mất mát trên tập phát triển dần ổn định. Gradient của các vectơ nhúng chữ Thái mới lớn nhất ở giai đoạn đầu, sau đó giảm còn khoảng một nửa, trong khi các mô-đun LoRA ở lớp truyền thẳng đảm nhận phần lớn quá trình cập nhật.

Với 16.924 cặp dữ liệu và 20 epoch, mô hình Thái → Việt được huấn luyện trong 52 phút trên GPU NVIDIA RTX 5070 Ti 16 GB; chiều Việt → Thái mất 39 phút. Một lần huấn luyện tương đương trên GPU T4 miễn phí mất 90 phút. Kết quả cho thấy phương pháp có thể triển khai với tài nguyên tính toán phổ thông.



Hình 2. Lần huấn luyện Thái → Việt cuối: chrF++ trên tập phát triển câu theo epoch (đường liền; epoch được chọn khoanh tròn) và hàm mất mát trên tập phát triển (đường đứt).

3.9. Triển khai thực tế

Hai mô hình được triển khai trên máy chủ GPU thông qua API có khóa truy cập, tích hợp tại `vi-ettrans.info`; chiều Việt → Thái được ghi là bản thử nghiệm. Hệ thống sử dụng quy trình chuẩn hóa, giải mã thống nhất với đánh giá, tự tách đoạn

dài và hỗ trợ văn bản phông chữ cũ. Hai mô hình chiếm khoảng 4,4 GB bộ nhớ GPU, xử lý 45–60 ms/câu khi dịch theo lô. Cả 1.638 yêu cầu thử tải đều thành công, thời gian phản hồi trung vị 302 ms, phân vị 95% là 589 ms. Năng lượng GPU ước tính 3,3–4,2 J/câu ngắn (khoảng 0,001 Wh), chưa tính các thiết bị khác. Nhật ký chỉ lưu độ dài và thời gian xử lý, không lưu nội dung văn bản.

3.10. Hạn chế

Nghiên cứu mới sử dụng đánh giá tự động, chưa có đánh giá của con người về độ chính xác và độ trôi chảy; chiều Việt → Thái chưa được người bản ngữ thẩm định. Tập kiểm tra chỉ lấy từ một giáo trình hành chính thuộc một vùng phương ngữ, nên khả năng khái quát còn hạn chế; văn bản phông cũ tại Thanh Hóa chỉ khớp 70% token với danh sách âm tiết. Tập phát triển và phần lớn dữ liệu huấn luyện được dịch từ tiếng Việt sang tiếng Thái nên có thể mang đặc điểm của văn dịch. Phần lớn so sánh chỉ thực hiện một lần huấn luyện, với dao động ước tính ±1 chrF++; thí nghiệm loại bỏ thành phần chỉ dùng dữ liệu từ điển. Mô hình chưa xử lý riêng tên riêng; các hướng tiền huấn luyện bằng dữ liệu tổng hợp và dịch ngược chưa được thực hiện.

3.11. Quản trị dữ liệu

Kho ngữ liệu song ngữ được xây dựng và quản lý theo các nguyên tắc CARE về quản trị dữ liệu bản địa (Carroll et al., 2020), với sự tham gia và đồng thuận của cộng đồng. Dữ liệu được lưu trữ trên hạ tầng do nhóm nghiên cứu quản lý, không chuyển đến dịch vụ mô hình ngôn ngữ bên thứ ba để làm sạch hoặc dịch. Các thực nghiệm ban đầu sử dụng GPU đám mây miễn phí; mô hình cuối được huấn luyện trên phần cứng của nhóm.

Nhóm dự kiến công bố bộ điều hợp LoRA, bộ tách token mở rộng, các vectơ nhúng mới và mã nguồn theo giấy phép CC BY-NC 4.0 của mô hình gốc, với điều kiện được người dịch và đối tác cộng đồng đồng ý. Kho ngữ liệu không công bố công khai nhưng có thể được cung cấp theo yêu cầu nghiên cứu. Bản dịch máy chỉ đóng vai trò hỗ trợ, không thay thế người dịch đối với văn bản quan trọng.

IV. KẾT LUẬN

Nghiên cứu cho thấy trở ngại cơ bản của dịch máy Thái – Việt bằng NLLB là khả năng nhận diện chữ viết. Phương pháp mở rộng từ vựng theo âm tiết, khởi tạo vectơ nhúng từ nghĩa tiếng Việt trong từ điển song ngữ giúp nâng chrF++ từ 11,4 lên 34,9. Việc bổ sung dữ liệu do người bản ngữ dịch, đặc biệt là câu dài và đoạn văn, tiếp tục cải thiện chất lượng dịch.

Với 16.924 cặp huấn luyện, mô hình Thái → Việt đạt 56,3 chrF++ trên tập phát triển và 41,7 trên tập kiểm tra độc lập 445 câu; chiều Việt → Thái đạt 46,2, cao hơn phương pháp tra từ điển (28,6). Hệ thống được huấn luyện trên một GPU phổ thông và triển khai thành dịch vụ trực tuyến, cho thấy tính khả thi trong điều kiện tài nguyên hạn chế. Tuy nhiên, chất lượng dịch đối với câu dài, văn vần cổ và dữ liệu ngoài miền còn hạn chế. Các nghiên cứu tiếp theo cần bổ sung đánh giá của người bản ngữ, xử lý tên riêng, mở rộng dữ liệu theo văn phong, phương ngữ và xây dựng tập kiểm tra từ nhiều nguồn độc lập.

TÀI LIỆU THAM KHẢO

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. Data Science Journal, 19(1), Article 43. https://doi.org/10.5334/dsj-2020-043

Dobler, K., & de Melo, G. (2023). FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 13440–13454).

Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., & Birch, A. (2022). Survey of low-resource machine translation. Computational Linguistics, 48(3), 673–732. https://doi.org/10.1162/coli_a_00446

Hewitt, J. (2021). Initializing new word embeddings for pretrained language models. https://nlp.stanford.edu/~johnhew/vocab-expansion.html

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations. https://arxiv.org/abs/2106.09685

Minixhofer, B., Paischer, F., & Rekabsaz, N. (2022). WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Proceedings of NAACL-HLT 2022 (pp. 3992–4006).

NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. arXiv. https://doi.org/10.48550/arXiv.2207.04672

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation (pp. 392–395).

Popović, M. (2017). chrF++: Words helping character n-grams. In Proceedings of the Second Conference on Machine Translation (pp. 612–618).

Post, M. (2018). A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation (pp. 186–191).

The Unicode Consortium. (2009). The Unicode Standard, Version 5.2.0.

Tổng cục Thống kê. (2020). Kết quả toàn bộ Tổng điều tra dân số và nhà ở năm 2019. Nhà xuất bản Thống kê.