HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN

THỰC TẬP CƠ SỞ

BÁO CÁO GIỮA KỲ

Đề tài: Finetune mô hình dịch máy Neural cho văn bản chuyên ngành y khoa Anh-Việt

Lớp tín chỉ:	**D23CQCN04-B**
Giảng viên hướng dẫn:	**TS. Kim Ngọc Bách**
Sinh viên thực hiện:	**Nguyễn Đức Trung**
Mã sinh viên:	**B23DCCN858**

Hà Nội 2026

MỤC LỤC

DANH MỤC VIẾT TẮT

Từ viết tắt	Tên đầy đủ	Giải thích
NMT	Neural Machine Translation	dịch máy dựa trên mạng nơ-ron
RNN	Recurrent Neural Networks	Mạng nơ-ron hồi quy
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên

I. GIỚI THIỆU DỰ ÁN

1. Lý do chọn đề tài

Dịch tự động (Machine Translation – MT) là một trong những ứng dụng quan trọng của lĩnh vực Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), cho phép chuyển đổi nội dung giữa các ngôn ngữ một cách tự động và nhanh chóng [1]. Tuy nhiên, đối với cặp ngôn ngữ Anh – Việt, chất lượng dịch trong các lĩnh vực chuyên sâu, đặc biệt là y học, vẫn còn nhiều hạn chế. Nguyên nhân chủ yếu xuất phát từ sự thiếu hụt dữ liệu song ngữ chuyên ngành chất lượng cao, cũng như tính phức tạp của hệ thống thuật ngữ y khoa.

Văn bản y học chứa nhiều thuật ngữ chuyên môn, cấu trúc câu học thuật phức tạp và các biểu đạt mang tính chính xác cao. Các công cụ dịch phổ thông thường gặp khó khăn trong việc xử lý các thuật ngữ như tên bệnh, cơ chế sinh học, phác đồ điều trị hoặc mô tả lâm sàng, dẫn đến sai lệch nghĩa hoặc dịch thiếu chính xác. Điều này có thể gây hiểu nhầm nghiêm trọng trong bối cảnh nghiên cứu và thực hành y khoa.

Xuất phát từ thực tiễn đó, đề tài hướng đến việc xây dựng một hệ thống dịch Anh – Việt chuyên ngành y học dựa trên mô hình học sâu với trọng số đã được huấn luyện sẵn (pre-trained model), sau đó tiến hành fine-tune trên tập dữ liệu song ngữ y khoa. Cách tiếp cận này nhằm tận dụng tri thức ngôn ngữ tổng quát của mô hình nền, đồng thời điều chỉnh để phù hợp hơn với đặc thù thuật ngữ và văn phong của lĩnh vực y học.

2. Ý nghĩa và ứng dụng

Việc phát triển hệ thống dịch tự động chuyên ngành y học mang ý nghĩa thực tiễn quan trọng trong bối cảnh hội nhập và phát triển khoa học hiện nay. Phần lớn các tài liệu y khoa, bài báo nghiên cứu, hướng dẫn điều trị và báo cáo lâm sàng được

công bố bằng tiếng Anh. Một hệ thống dịch có độ chính xác cao sẽ giúp sinh viên ngành y, bác sĩ, nhà nghiên cứu và cán bộ y tế tại Việt Nam tiếp cận nhanh chóng nguồn tri thức quốc tế, từ đó nâng cao chất lượng học tập, nghiên cứu và thực hành chuyên môn.

Bên cạnh ý nghĩa thực tiễn, đề tài còn có giá trị học thuật trong lĩnh vực dịch máy chuyên ngành (domain-specific machine translation). Việc fine-tune mô hình trên dữ liệu y học cho phép đánh giá mức độ cải thiện so với mô hình dịch tổng quát, qua đó kiểm chứng hiệu quả của phương pháp domain adaptation [2]. Kết quả nghiên cứu có thể đóng góp vào việc xây dựng bộ dữ liệu song ngữ y khoa phục vụ cho các nghiên cứu tiếp theo.

Ngoài ra, hệ thống dịch chuyên ngành y học còn có tiềm năng ứng dụng trong thực tế như: dịch tài liệu hướng dẫn sử dụng thuốc, báo cáo nghiên cứu lâm sàng, tài liệu đào tạo y khoa hoặc hỗ trợ dịch thuật trong môi trường bệnh viện và viện nghiên cứu. Điều này mở ra khả năng phát triển và thương mại hóa trong tương lai, đặc biệt trong bối cảnh chuyển đổi số ngành y tế.

II. CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ SỬ DỤNG

1. Cơ sở lý thuyết

a. Neural Machine Translation

Dịch máy hiện đại chủ yếu dựa trên phương pháp Neural Machine Translation , trong đó quá trình dịch được mô hình hóa bằng các mạng nơ-ron sâu để học ánh xạ trực tiếp từ câu nguồn sang câu đích. Khác với các phương pháp dịch máy thống kê trước đây, NMT cho phép mô hình học được các biểu diễn ngữ nghĩa phức tạp và tận dụng ngữ cảnh toàn bộ câu trong quá trình dịch. Phần lớn các hệ thống NMT sử dụng cấu trúc encoder–decoder, trong đó bộ mã hóa (encoder) chuyển đổi câu nguồn thành biểu diễn vector, còn bộ giải mã (decoder) sinh ra câu đích dựa trên biểu diễn này [1].

b. Kiến trúc Transformer

Một bước tiến quan trọng trong dịch máy là sự ra đời của kiến trúc Transformer, được đề xuất trong nghiên cứu *Attention Is All You Need* [3]. Khác với các mô hình trước đây dựa trên RNN [4] hoặc CNN [5], Transformer chỉ sử dụng cơ chế attention để mô hình hóa quan hệ giữa các từ trong câu.

Cơ chế multi-head attention cho phép mô hình học nhiều loại quan hệ ngữ nghĩa khác nhau giữa các từ, giúp nắm bắt ngữ cảnh toàn cục hiệu quả hơn. Ngoài ra, Transformer cho phép xử lý song song các token trong câu, giúp tăng tốc độ huấn luyện và cải thiện hiệu năng so với các kiến trúc trước đó. Nhờ các ưu điểm này,

Transformer đã trở thành kiến trúc nền tảng cho hầu hết các hệ thống dịch máy hiện đại.

c. Fine-tuning và domain adaptation

Trong nhiều trường hợp, mô hình NMT được huấn luyện trên dữ liệu tổng quát nhưng cần áp dụng cho một lĩnh vực chuyên ngành cụ thể. Khi đó, phương pháp phổ biến là fine-tuning [6], tức là tiếp tục huấn luyện mô hình đã được tiền huấn luyện trên một tập dữ liệu mới thuộc miền chuyên ngành (domain adaptation) [2]. Quá trình này giúp mô hình thích nghi tốt hơn với các thuật ngữ và cấu trúc câu đặc thù của lĩnh vực đó.

Tuy nhiên, việc tinh chỉnh mô hình cũng có thể dẫn đến hiện tượng Catastrophic Forgetting, trong đó mô hình có xu hướng quên đi các kiến thức tổng quát đã học trước đó. Vì vậy, quá trình fine-tuning cần được thực hiện với các siêu tham số phù hợp nhằm cân bằng giữa khả năng thích nghi miền và việc duy trì kiến thức ngôn ngữ chung của mô hình.

2. Công nghệ và dữ liệu

a. Mô hình EnViT5

Trong đề tài này, mô hình được sử dụng là EnViT5 [7] [8], một mô hình dịch máy Anh–Việt được xây dựng dựa trên kiến trúc Transformer. EnViT5 được huấn luyện trước trên các tập dữ liệu song ngữ Anh–Việt quy mô lớn, giúp mô hình học được các biểu diễn ngôn ngữ và khả năng dịch cơ bản giữa hai ngôn ngữ.

Trong nghiên cứu này, mô hình EnViT5 được sử dụng làm mô hình nền (pretrained model) và tiếp tục được fine-tune trên dữ liệu y khoa nhằm cải thiện khả năng dịch các thuật ngữ chuyên ngành. Việc sử dụng mô hình đã được tiền huấn luyện giúp giảm đáng kể chi phí huấn luyện và cho phép mô hình đạt hiệu quả tốt ngay cả khi dữ liệu chuyên ngành có kích thước hạn chế.

b. Bộ dữ liệu MedEV

Trong đề tài này, bộ dữ liệu được sử dụng để tinh chỉnh mô hình là MedEV [9], một tập dữ liệu song ngữ Anh–Việt trong lĩnh vực y sinh học. Bộ dữ liệu MedEV bao gồm khoảng 360.000 cặp câu song ngữ Anh–Việt, được thu thập và xây dựng từ các nguồn tài liệu y khoa như bài báo khoa học, tài liệu nghiên cứu y sinh và các nguồn học thuật liên quan đến lĩnh vực y học. Nội dung của tập dữ liệu bao phủ nhiều chủ đề chuyên ngành như bệnh học, dược học, điều trị lâm sàng và nghiên cứu y sinh học, do đó chứa nhiều thuật ngữ chuyên môn và cấu trúc câu mang tính học thuật.

Việc lựa chọn MedEV cho quá trình fine-tuning xuất phát từ đặc điểm của mô hình EnViT5. Trong giai đoạn tiền huấn luyện, EnViT5 được huấn luyện trên các tập dữ

liệu song ngữ Anh–Việt quy mô lớn thuộc nhiều lĩnh vực khác nhau như tin tức, bách khoa, hội thoại hoặc phụ đề phim. Những tập dữ liệu này giúp mô hình học được kiến thức ngôn ngữ tổng quát và khả năng dịch cơ bản giữa hai ngôn ngữ. Tuy nhiên, các dữ liệu này thường thiếu các thuật ngữ chuyên ngành và ngữ cảnh chuyên sâu của lĩnh vực y khoa.

So với các tập dữ liệu tổng quát đã được sử dụng để huấn luyện ban đầu cho EnViT5, MedEV có sự khác biệt rõ rệt về miền dữ liệu (domain). Trong khi dữ liệu huấn luyện ban đầu của EnViT5 mang tính đa miền và tập trung vào ngôn ngữ phổ thông, MedEV lại tập trung hoàn toàn vào lĩnh vực y sinh học. Điều này giúp bổ sung cho mô hình một lượng lớn các thuật ngữ y khoa, tên bệnh, tên thuốc và các mô tả lâm sàng mà dữ liệu tổng quát thường không bao phủ đầy đủ.

Ngoài ra, MedEV là một bộ dữ liệu tương đối mới trong nghiên cứu dịch máy Anh–Việt và được xây dựng nhằm giải quyết sự thiếu hụt dữ liệu song ngữ chuyên ngành y khoa trước đây. Với quy mô khoảng 360 nghìn cặp câu và nội dung chuyên biệt trong lĩnh vực y sinh học, MedEV cung cấp nguồn dữ liệu phù hợp để fine-tune mô hình EnViT5, giúp mô hình thích nghi tốt hơn với văn bản y khoa và cải thiện chất lượng dịch trong các ngữ cảnh chuyên ngành.

III. PHÂN TÍCH YÊU CẦU CỦA DỰ ÁN

Mục tiêu của dự án là xây dựng một hệ thống dịch máy Anh–Việt trong lĩnh vực y khoa bằng cách tinh chỉnh một mô hình dịch máy đã được huấn luyện trước. Hệ thống cần có khả năng tiếp nhận các văn bản tiếng Anh thuộc lĩnh vực y sinh học và sinh ra bản dịch tiếng Việt tương ứng với độ chính xác cao, đặc biệt trong việc xử lý các thuật ngữ chuyên ngành. Để đạt được mục tiêu này, mô hình nền được sử dụng trong nghiên cứu là EnViT5, một mô hình dịch máy dựa trên kiến trúc Transformer và đã được huấn luyện trước trên các tập dữ liệu song ngữ Anh–Việt quy mô lớn.

1. Yêu cầu chức năng

Về yêu cầu chức năng, hệ thống cần thực hiện quá trình dịch tự động từ tiếng Anh sang tiếng Việt đối với các văn bản thuộc lĩnh vực y khoa. Văn bản đầu vào cần được tiền xử lý nhằm chuẩn hóa dữ liệu và chuyển đổi sang định dạng phù hợp với mô hình dịch máy. Sau đó, mô hình EnViT5 được sử dụng để sinh bản dịch tiếng Việt tương ứng. Trong quá trình nghiên cứu, mô hình sẽ được fine-tune trên dữ liệu chuyên ngành để giúp mô hình thích nghi tốt hơn với các thuật ngữ và cấu trúc ngôn ngữ đặc thù của lĩnh vực y khoa.

2. Yêu cầu dữ liệu

Đối với yêu cầu dữ liệu, dự án cần sử dụng một tập dữ liệu song ngữ Anh–Việt trong lĩnh vực y sinh học để phục vụ quá trình huấn luyện và đánh giá mô hình. Trong nghiên cứu này, bộ dữ liệu được lựa chọn là MedEV, bao gồm khoảng 360.000 cặp câu song ngữ Anh–Việt. Bộ dữ liệu này cung cấp nhiều nội dung liên quan đến bệnh học, dược học và nghiên cứu y sinh, do đó giúp mô hình học được các thuật ngữ và ngữ cảnh chuyên ngành cần thiết cho bài toán dịch văn bản y khoa. Dữ liệu sẽ được chia thành các tập huấn luyện, kiểm tra và đánh giá để đảm bảo quá trình huấn luyện mô hình diễn ra hiệu quả và khách quan.

3. Yêu cầu công nghệ

Sử dụng các công cụ và thư viện hỗ trợ cho việc triển khai và huấn luyện mô hình. Thư viện Hugging Face Transformers được sử dụng để triển khai mô hình dịch máy và thực hiện quá trình fine-tuning. Ngoài ra, môi trường huấn luyện cần hỗ trợ GPU nhằm tăng tốc quá trình huấn luyện mô hình trên tập dữ liệu lớn.

4. Yêu cầu đánh giá

Chất lượng của hệ thống dịch máy cần được đánh giá thông qua các chỉ số phổ biến trong lĩnh vực dịch máy, trong đó chỉ số BLEU score được sử dụng để đo lường mức độ tương đồng giữa bản dịch do mô hình sinh ra và bản dịch tham chiếu trong tập dữ liệu. Kết quả đánh giá sẽ được sử dụng để phân tích hiệu quả của phương pháp fine-tuning đối với mô hình EnViT5 khi áp dụng cho bài toán dịch văn bản y khoa.

IV. KẾ HOẠCH THỰC HIỆN DỰ ÁN

Dự án được triển khai trong khoảng thời gian hai tháng và được chia thành các giai đoạn chính nhằm đảm bảo tiến độ cũng như chất lượng của hệ thống dịch máy. Trong giai đoạn đầu, tiến hành khảo sát và nghiên cứu các tài liệu liên quan đến bài toán dịch máy trong lĩnh vực xử lý ngôn ngữ tự nhiên. Nội dung nghiên cứu tập trung vào các phương pháp Neural Machine Translation, kiến trúc Transformer và các kỹ thuật fine-tuning mô hình ngôn ngữ đã được huấn luyện trước. Đồng thời, cũng tìm hiểu về cấu trúc và cách sử dụng mô hình EnViT5 để phục vụ cho bài toán dịch Anh–Việt.

Sau khi hoàn thành bước nghiên cứu lý thuyết, dự án chuyển sang giai đoạn chuẩn bị dữ liệu. Trong giai đoạn này, bộ dữ liệu MedEV được thu thập và xử lý để phục vụ cho quá trình huấn luyện mô hình. Các bước tiền xử lý bao gồm làm sạch dữ liệu, chuẩn hóa văn bản và thực hiện tokenization phù hợp với kiến trúc của mô hình dịch máy. Sau đó, dữ liệu được chia thành các tập huấn luyện, kiểm tra và

đánh giá nhằm đảm bảo quá trình huấn luyện và đánh giá mô hình được thực hiện một cách khách quan.

Tiếp theo là giai đoạn xây dựng và huấn luyện mô hình. Trong giai đoạn này, môi trường huấn luyện được thiết lập với các thư viện cần thiết, trong đó thư viện Hugging Face Transformers được sử dụng để triển khai và fine-tune mô hình EnViT5. Mô hình sau đó được huấn luyện trên bộ dữ liệu MedEV nhằm giúp nó thích nghi với các thuật ngữ và ngữ cảnh chuyên ngành y khoa. Trong quá trình huấn luyện, các siêu tham số của mô hình được theo dõi và điều chỉnh phù hợp để đảm bảo hiệu quả học tập của mô hình.

Sau khi quá trình huấn luyện hoàn tất, mô hình được đánh giá trên tập dữ liệu kiểm tra để đo lường chất lượng dịch. Việc đánh giá được thực hiện thông qua các chỉ số phổ biến trong dịch máy như BLEU score, đồng thời tiến hành phân tích định tính một số ví dụ dịch để quan sát khả năng xử lý các thuật ngữ y khoa của mô hình. Kết quả của mô hình sau khi fine-tune sẽ được so sánh với mô hình ban đầu để xác định mức độ cải thiện về chất lượng dịch.

Trong giai đoạn cuối cùng, các kết quả thực nghiệm được tổng hợp và phân tích nhằm rút ra các nhận xét về hiệu quả của phương pháp fine-tuning đối với bài toán dịch văn bản y khoa. Từ đó, tiến hành hoàn thiện báo cáo thuyết minh của dự án, trình bày các bước thực hiện, phương pháp nghiên cứu và kết quả đạt được. Đồng thời, các tài liệu và nội dung trình bày cũng được chuẩn bị để phục vụ cho quá trình báo cáo và đánh giá đề tài.

THAM KHẢO

[1] P. Koehn, Neural Machine Translation, Cambridge: Cambridge University Press, 2017.

[2] H. Daumé III, Domain Adaptation in Natural Language Processing, San Rafael, CA: Morgan & Claypool Publishers, 2012.

[3] A. Vaswani , N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser và I. Polosukhin, “Attention Is All You Need,” *Advances in Neural Information Processing Systems,* tập 30, p. 5998–6008, 2017.

[4] I. Sutskever, O. Vinyals và Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” *Advances in Neural Information Processing Systems,* tập 27, pp. 3104-3112, 2014.

[5] A. Krizhevsky, I. Sutskever và G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances in Neural Information Processing Systems,* tập 25, pp. 1097-1105, 2012.

[6] J. Devlin, M.-W. Chang, K. Lee và K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of NAACL-HLT,* p. 4171–4186, 2019.

[7] C. Ngo, T. H. Trinh, L. Phan, H. Tran, T. Dang, H. Nguyen, M. Nguyen và M.-T. Luong, “MTet: Multi-domain Translation for English and Vietnamese,” *arXiv preprint arXiv:2210.05610,* 2022.

[8] C. Ngo, T. H. Trinh, L. Phan, H. Tran, T. Dang, H. Nguyen, M. Nguyen và M.-T. Luong, “EnViT5: English–Vietnamese Translation Model,” VietAI, 2022. [Trực tuyến]. Available: https://huggingface.co/VietAI/envit5-translation. [Đã truy cập 27 February 2026].

[9] N. Vo, D. Q. Nguyen, D. D. Le, M. Piccardi và W. Buntine, “Improving Vietnamese-English Medical Machine Translation,” trong *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, 2024.